UNIVERSITY OF BELGRADE
SCHOOL OF ELECTRICAL ENGINEERING

Ivan Stojković

# FUNCTIONAL NORM REGULARIZATION FOR MARGIN-BASED RANKING ON TEMPORAL DATA

doctoral dissertation

Belgrade, 2018

UNIVERZITET U BEOGRADU
ELEKTROTEHNIČKI FAKULTET

Ivan Stojković

# PRIMENA FUNKCIONALNIH NORMI ZA REGULARIZACIJU RANGIRANJA NAD TEMPORALNIM PODACIMA

doktorska disertacija

Beograd, 2018

*COMMITTEE FOR THESIS EXAMINATION, EVALUATION AND DEFENSE*

Supervisor: dr Zoran Obradović, L.H. Carnell professor, Temple University, College of Science and Technology

Supervisor: dr Branko Kovačević, Full professor, University of Belgrade, School of Electrical Engineering

dr Slobodan Vučetić, Full professor, Temple University, College of Science and Technology

dr Željko Đurović, Full professor, University of Belgrade, School of Electrical Engineering

dr Kai Zhang, Associate professor, Temple University, College of Science and Technology

dr Carmen Sapienza, Full professor, Temple University, Fox Chase Cancer Center

Defense date:

# ACKNOWLEDGEMENTS

My deepest gratitude goes to my advisor Dr Zoran Obradović for providing me with all the necessary resources to complete this dissertation: an opportunity to join his lab; his expertise in data science and machine learning; his time, advice, support and patience; and last but not least, for funding my research and education.

I am also very grateful to my co-advisor Dr Branko Kovačević, my initial co-advisor Dr Stevica Graovac, and especially to Dr Željko Đurović for teaching me well and for invaluable help in navigating administrative procedures. I wouldn't be able to complete this PhD program without their support.

I am also thankful to Dr Carmen Sapienza and Dr Adam Davey for working with me and sharing their great technical knowledge and domain expertise, their persistence and passion for science is truly inspiring.

Thanks to Dr Slobodan Vučetić and Dr Kai Zhang for serving on my dissertation committee and providing valuable feedback and insights that improved my dissertation.

Special thanks to Dr Mohamed Ghalwash who introduced me to the projects, and helping me out when it was needed the most.

Thanks to my dear collaborators Vladisav Jelisavčić, Đorđe Gligorijević, Jelena Stojanović, Xi Hang Cao, Martin Pavlovski and Fang Zhou for working on joint problems and papers, I am looking forward to continue collaborations in the future.

Thanks to my lab-mates Shoumik, Dušan, Jesse, Chao, Nancy, Alexey, Nouf, Nour, Miloš, Milan, Ana, Tijana, Tom, Miloš, Miloš, Ana, Milan, Branko, Nima, Ali, Branimir, Marija and Jumana,(and other people I might left out) for sharing their views on the lab meetings and for insightful discussions at the office.

I would also like to thank my ex-colleagues from Mihajlo Pupin Institute, Dr Aleksandar Rodić for showing me the way, and to late Dr Duško Katić for long discussions on scientific and also on not-so-scientific topics.

I particularly appreciate help from my friends at the Robotics Laboratory, who struggled together with me at the beginning of this journey - Branko Miloradović, Sofija Spasojević, Aleksandar Ćosić and Vladimir Petrović. With them everything was much easier.

I must not forget to express gratitude to my friends from Philadelphia Tijana, Vladan, Jovana, Omar, Stefan, Vlada and Jelena, that were always there for hanging-out and fun, and to help me shift the focus from PhD, at least for a moment.

Finally, I would also like to thank all my friends and family back home in Serbia for being on my side for all these years, their encouragement and support were precious.

*To unexpected reader.*

**Dissertation title:** Functional norm regularization for margin-based ranking on temporal data

**Abstract:** Quantifying the properties of interest is an important problem in many domains, e.g., assessing the condition of a patient, estimating the risk of an investment or relevance of the search result. However, the properties of interest are often latent and hard to assess directly, making it difficult to obtain classification or regression labels, which are needed to learn a predictive models from observable features. In such cases, it is typically much easier to obtain relative comparison of two instances, i.e. to assess which one is more intense (with respect to the property of interest). One framework able to learn from such kind of supervised information is ranking SVM, and it will make a basis of our approach.

Applications in bio-medical datasets typically have specific additional challenges. First, and the major one, is the limited amount of data examples, due to an expensive measuring technology, and/or infrequency of conditions of interest. Such limited number of examples makes both identification of patterns/models and their validation less useful and reliable. Repeated samples from the same subject are collected on multiple occasions over time, which breaks IID sample assumption and introduces dependency structure that needs to be taken into account more appropriately. Also, feature vectors are highdimensional, and typically of much higher cardinality than the number of samples, making models less useful and their learning less efficient.

Hypothesis of this dissertation is that use of the functional norm regularization can help alleviating mentioned challenges, by improving generalization abilities and/or learning efficiency of predictive models, in this case specifically of the approaches based on the ranking SVM framework.

The temporal nature of data was addressed with loss that fosters temporal smoothness of functional mapping, thus accounting for assumption that temporally

proximate samples are more correlated. Large number of feature variables was handled using the sparsity inducing $L_1$ norm, such that most of the features have zero effect in learned functional mapping. Proposed sparse (temporal) ranking objective is convex but non-differentiable, therefore smooth dual form is derived, taking the form of quadratic function with box constraints, which allows efficient optimization. For the case where there are multiple similar tasks, joint learning approach based on matrix norm regularization, using trace norm $L_*$ and sparse row $L_{21}$ norm was also proposed. Alternate minimization with proximal optimization algorithm was developed to solve the mentioned multi-task objective.

Generalization potentials of the proposed high-dimensional and multi-task ranking formulations were assessed in series of evaluations on synthetically generated and real datasets. The high-dimensional approach was applied to disease severity score learning from gene expression data in human influenza cases, and compared against several alternative approaches. Application resulted in scoring function with improved predictive performance, as measured by fraction of correctly ordered testing pairs, and a set of selected features of high robustness, according to three similarity measures. The multi-task approach was applied to three human viral infection problems, and for learning the exam scores in Math and English. Proposed formulation with mixed matrix norm was overall more accurate than formulations with single norm regularization.

**Key words:** SVM ranking, scoring function learning, functional norm regularization, proximal algorithms for optimization, temporal data
**Scientiffic field:** Electrical Engineering and Computer Sciences
**Scientiffic subfield:** Data analysis and machine learning
**UDC number:** 004.8

**Naslov teze:** Primena funkcionalnih normi za reularizaciju rangiranja nad temporalnim podacima

**Sažetak:** Kvantifikovanje osobina (karakteristika) od interesa je važan problem u mnogim domenima, npr. utvrđivanje težine bolesti kod pacijenata, ocena rizika investicije ili relevantnost vraćenih rezultata pretrage. Međutim, osobine od interesa su često latentne i teško se mogu izmeriti direktno, što otežava dobijanje klasifikacionih oznaka (labela) ili ciljeva za regresiju, koji su potrebni za učenje prediktivnih modela iz merljivih karakteristika. U takvim slučajevima obično je mnogo lakše pribaviti relativno poređenje dva slučaja, tj. proceniti koji od dva je intenzivniji (iz ugla karakteristike od interesa). Jedna klasa algoritama koji mogu učiti iz ovakvih informacija je SVM za rangiranje i on će biti osnova ovde predloženog pristupa.

Aplikacije na biomedicinskim skupovima podataka obično imaju dodatne (specifične) izazove. Prvi, i najvažniji, je ograničena količina primera u podacima. To se najčešće dešava zbog skupih tehnologija merenja i / ili retkosti stanja od interesa (na primer oblik raka koji pogađa jako mali broj pacijenata). Takav ograničeni broj primera čini i identifikaciju obrazaca / modela i njihovu validaciju manje korisnim i pouzdanim. Ponovljeni uzorci (od istog procesa / subjekta) prikupljaju se u više navrata tokom vremena, što razbija pretpostavku o identičnoj i nezavisnoj rspodeli (IID) uzorka i uvodi strukturu zavisnosti koju je potrebno uzeti u obzir. Takođe, vektori obeležja su visokodimenzionalni i obično imaju mnogo veću kardinalnost u odnosu na broj uzoraka, čineći modele manje korisnim a njihovo obučavanje manje efikasnim.

Hipoteza ove disertacije je da korišćenje funkcionalnih normi za regularizaciju može pomoći ublažavanju prethodno pomenutih izazova, pritom poboljšavajući generalizacione sposobnosti i / ili efikasnost učenja prediktivnih modela, u ovom scučaju konkretno o pristupima zasnovanim na rangiranju pomoću SVM-a.

Vremenski karakter podataka adresiran je korišćenjem objektiva koji podstiče

vremensku glatkost funkcionalnog mapiranja, čime se uzima u obzir pretpostavka da su vremenski bliski uzorci više korelisani, te trebaju imati sličnije vrednosti mapirane funkcije. Problem velikog broja promenljivih je adresiran korišćenjem $L_1$ norme koja indukuje proređenost, tako da većina varijabli nema efekat na naučeno funkcionalno mapiranje. Predloženi objektiv za proređeno (vremensko) rangiranje je konveksan, ali ne-diferencijabilan, stoga se izvodi glatka dvojna forma, koja ima oblik kvadratne funkcije sa konstantnim ograničenjima, što omogućava efikasnu optimizaciju. U slučajevima gde postoji više sličnih zadataka, predložen je i zajednički pristup učenja zasnovan na normativnoj regularizaciji matrice, korišćenjem "tragovne" norme $L_*$ i norme za proređenost po redovima $L_{21}$. Pomenuti višestruki objektiv je rešen predloženim metodom naizmenične minimizacije upotrebom algoritama proksimalne optimizacije.

Generalizacioni potencijal predloženih formulacija za rešavanje visokodimenzionalnih i višestrukih problema rangiranja procenjen je u nizu evaluacija na sintetički generisanim i realnim podacima. Visoko-dimenzionalni pristup primenjen je na učenje funkcije bodovanja težine bolesti iz podataka o ekspresiji gena kod slučajeva ljudskog gripa i upoređivan je sa nekoliko alternativnih pristupa. Aplikacija je rezultirala funkcijom bodovanja sa poboljšanim prediktivnim performansama, mereno udelom ispravno poređanih test parova i skupom odabranih obeležja visoke robusnosti, prema tri mere sličnosti. Višestruki pristup je primenjen na problemima sa ispitivanjem tolerancije ljudi na tri virusine respiratorne infekcije, kao i za bodovanje ispita iz matematike i engleskog jezika. Predložena formulacija sa mešovitom matričnom normom se ispostavila superiornijom u odnosu na formulacije sa regulacijom pomoću pojedinačnih normi.

**Ključne reči:** SVM rangiranje, učenje funkcija za bodovanje, funkcionalna regularizacija normama, proksimalni algoritmi za optimizaciju, temporalni podaci

**Naučna oblast:** Elektrotehnika i Računarske Nauke

**Uža naučna oblast:** Analiza podataka i mašinsko učenje

**UDC broj:** 004.8

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# List of Abbreviations and Symbols

## Symbols

| | |
|---|---|
| $Y$ | $n$-dimensional target vector |
| $X_{n \times d}$ | $d$-dimensional measurement matrix of $n$ examples |
| $w$ | parameter (weight) vector |
| $\lambda$ | sparsity hyperparameter |
| $\|.\|$ | functional norm |
| $\|.\|_1$ | sparse norm $L_1$ |
| $\|.\|_2$ | Euclidean distance norm $L_2$ |
| $\|.\|_{2,1}$ | sparse row norm $L_{2,1}$ |
| $\|.\|_*$ | low rank (trace) norm $L_*$ |
| $L_h$ | Hubber loss |
| **prox** | proximal operator |
| $\sigma$ | singular value |
| $exp(.)$ | exponential function |
| $log(.)$ | natural logarithm |
| $max(.,.)$ | maximum function |
| **diag** | diagonal of a matrix |
| $\underset{w}{\text{argmin}}$ | minimizer of the function |
| $\mathcal{L}_1$ | loss |

| | |
|---|---|
| $\mathbb{R}$ | set of real numbers |
| $\Theta$ | set of parameter matrices |
| $L$ | Lipschitz constant |
| $\mathbf{0}$ | zero vector |
| $\mathbf{1}$ | vector of ones |
| $\nabla$ | vector differential operator |
| $\infty$ | infinite scalar |

## Abbreviations

| | |
|---|---|
| SOFA | Sequential Organ Failure Assessment score |
| ICU | An Intensive Care Unit |
| APACHE | Acute Physiology and Chronic Health Evaluation |
| SVM | Support Vector Machine |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| L1 Log Reg | L1 regularized Logistic Regression |
| SLDSS | Sparse Linear Disease Severity Score |
| DSSL | Disease Severity Score Learning |
| MRF | Markov Random Field |
| CRF | Conditional Random Field |
| GCRF | Gaussian Conditional Random Field |
| H3N2 | Hemagglutinin 3 - Neuraminidase 2 |
| HRV | Human Rhino-Virus |
| RSV | Respiratory Syncytial Virus |
| PANTHER | Protein Analysis Through Evolutionary Relationships |
| ID | Identification number |
| IID | Independent and Identically Distributed |

# CHAPTER 1

## INTRODUCTION

Quantifying properties of interest is integral to many domains, e.g., assessing the condition of a patient (Vincent et al., 1996), estimating the risk of an investment (Anderson, 2007), or predicting binding affinity of a ligand (Ashtawy and Mahapatra, 2015) when developing new drugs. For example, diseases and other health conditions require continuous monitoring and assessment of the subject's state. The severity of the condition needs to be quantified, such that it can subsequently be used to guide medical decisions and allow appropriate and timely interventions. Hence, various measuring technologies and sensors are devised to quantify such properties of interest, which are in turn utilized for informing decisions and making appropriate actions.

However, very often, the properties of interest are not easy to obtain, whether they are difficult to measure directly or completely unobservable. This is usually the case when the properties are conceptual, i.e. they are latent constructs, such as health, satisfaction, or intelligence, and are notoriously difficult to capture physically. Under these circumstances, other measurable characteristics, considered related and informative of the true underlying target, are observed and used as surrogate vari-

ables. For example, in clinical settings, variables like temperature, blood pressure and various biomarkers measured from tissues are commonly tracked and considered when determining the health of the patient.

Typically, some heuristic rules are decided to map these surrogate variables into the desired score. The process of deciding these heuristic rules (or scoring functions) is usually long and tedious. For example, disease severity scores that are needed in clinical practices for patient diagnostics require years of effort and consensus of the medical community before the scoring functions can become part of the protocols. Fortunately, developments in machine learning and increasing amounts of the collected data allowed for an alternative and complementary way for engineering the scoring functions by extracting rules automatically from the data, which facilitates and complements traditional approaches.

Algorithms for learning scoring functions from data were previously proposed, mainly in the medical domain, with the objective to learn disease severity scores (Yang et al., 2012; Santolino and Boucher, 2009; Dyagilev and Saria, 2015b,a; Zhou et al., 2012). Initial approaches posed the problem as traditional supervised learning tasks of classification (Yang et al., 2012; Santolino and Boucher, 2009) and regression (Zhou et al., 2012). However, classification and regression approaches require scores to already be accessible up front, which limits their applicability to problems with a good surrogate. The approach in (Dyagilev and Saria, 2015a,b) suggests the very appealing idea that there is a more convenient alternative form of supervised information to learn the scoring function from. Namely, ranked pairs are much easier to obtain than direct score estimates, and moreover, learning from pairs of ranked examples may result in more reliable and robust scoring functions.

First, we extend the suggested ranking-based approach (Dyagilev and Saria, 2015a) for score learning in multi-task settings. These efforts are motivated by applications in which there are multiple related tasks, with a limited amount of data

for each task. Related tasks commonly share underlying regularities which could be learned more accurately by modeling all tasks together. For example, in education, scores on different subjects (e.g. Math and English) are dependent on the same characteristics of a particular student and a particular school. In the medical domain, disease severity scores for related illnesses (e.g. various respiratory viral infections) are expected to share common underlying biological mechanisms. Consequently, we propose a novel multi-task formulation for learning scoring functions from pairwise comparisons, by enforcing structural regularities on joint parameter space, using a matrix norm regularization. In addition, we provide another contribution by developing an optimization algorithm in the form of an alternate minimization scheme based on a proximal gradient method.

Second, we propose an approach to the problem of learning disease severity scores in presence of irrelevant or large number of variables. We build on top of existing efforts by simultaneously performing feature selections that are most relevant for severity score learning. In particular, we are introducing the $L_1$ norm in the formulation of ranking SVM (Joachims, 2002) along with the temporal smoothness loss (Dyagilev and Saria, 2015a). Attractive regularization properties of $L_1$ norm are already well acknowledged and exploited in a number of statistical learning methods since its introduction (Tibshirani, 1996; Shaobing and Donoho, 1994). The proposed formulation of sparse severity score learning forces weights of (most of) the features to be exactly zero, therefore effectively performing feature selection by learning the sparse linear scoring function. This novel severity score objective function is convex and non-smooth and it precludes the direct use of convenient optimization tools like gradient-based methods. Therefore, we are also providing the reformulation of the problem into its dual that is smooth and that allows efficient optimization. Other than learning the severity score from the data, which is an important instrument for assessing severity, the methodology may also be used to discover the most relevant

variables/features for the disease severity phenotype. Such findings might be further used to suggest novel (testable) hypotheses about causal relations leading to disease manifestation, and also to inspire novel therapeutic approaches.

The reminder of the thesis is structured as follows:

First, a current state of the art in scoring function learning, high-dimensional and multi-task methods, as well as proximal algorithms is reviewed in Related Work Chapter.

Next, the proposed methodology is described in two Method Chapters, one for multi-task method, and another for high-dimensional approach. The two Method Chapters details the proposed new formulations of multi-task, and of high-dimensional ranking in temporal conditions, along with the derivation of their solutions. The results of application and evaluation of the two approaches are also divided in two Chapters.

In the first Results Chapter we evaluated generalization ability of multi-task framework. Initially, the evaluation is performed on synthetic data and subsequently in two real-world applications. The objective of the first application is learning exam scores of elementary school pupils, while the objective of the second application is learning the tolerance to respiratory viral infections in humans. The results suggests increased prediction accuracy of the proposed approach over the alternatives that are trained on individual tasks.

The following Results section is dedicated to evaluation of high-dimensional ranking model on a set of intuitive synthetic examples, where the advantages of sparse severity score framework over the non-sparse one are apparent. The results section continues with the assessment on a real-life applications, a gene expression dataset of H3N2 viral infection responses in humans, and gene expression data of (frog) tadpole bacteria infection. Efficacy, as well as the robustness of the proposed method, are compared favorably against multiple alternative methods. The analysis is followed

with gene ontology over-representation analysis of the discovered subset of genes most relevant for the scoring function.

Finally, contributions and limitations of the work presented in this thesis, are summarized and discussed in the Conclusion Chapter. Two additional appendices details used human respiratory virus datasets and part of the results regarding the feature selection stability analysis. At the end of the dissertation, there are Bibliography and author Biography details.

# CHAPTER 2

# RELATED WORK

Early efforts to learn scoring functions were dependent on complete supervised information (e.g. classification and regression tasks). In the classification settings, where the discrete class labels are provided, the classification methods were used to estimate the probability of a sample belonging to a certain class; these probabilities were used as a scoring function. For example, the method in (Yang et al., 2012) uses sparsity inducing $L_1$ norm in combination with a classical logistic loss function to learn the disease severity scoring function for assessing the abnormality of the skull in craniosynostosis cases.

## 2.1 Ranking Based Methods

The problem with such completely-supervised methods is the necessity of providing direct values of scores for training purposes, which render the approaches as less powerful in settings where characteristics of interest are latent and not directly accessible. However, rather than giving direct estimates of the score, an easier task seems to be comparing two samples and asserting whether one has a higher score than the other. Ranking SVM (Joachims, 2002) was the first approach that recognized the benefits of

learning desired functional mappings from ordered pairs of examples. This method was applied to learn an improved relevance function for documents retrieval from click-through data. Main insight was that clicked links are certainly more relevant for the search, as compared to non-clicked ones. And such kind of data is much more abundant than the user provided rankings. Sparse versions were proposed afterwards like (Bi et al., 2003; Lai et al., 2013). Recently, the ranking SVM-based method was adopted for Sepsis severity score learning (Dyagilev and Saria, 2015a) and extended for temporal applications by introducing a term that ensures gradual score change over consecutive time points. Another ranking method with addition of fused lasso regularization was proposed, which simultaneously performs supervised binning to discretize the continuous features and aid in model interpretability (Sokolovska et al., 2017).

## 2.2 Multi-task Learning

Multi-task learning is based on the idea that generalization (predictive performance) can be increased by accounting for the intrinsic relationships among multiple tasks. Multi-task approach is perceived particularly effective when the number of samples for each particular task is small.

One of the approaches is structured regression, which seeks to discover and exploit the relatedness structure among the tasks. Common class of modeling approaches to structured regression are undirected Probabilistic Graphical Models named Markov Networks, or Markov Random Fields (MRF). Discriminative models are often preferred, over the generative ones, as more accurate due to relaxations of independence assumptions (Sutton and McCallum, 2006). That is why Conditional Random Fields (Lafferty et al., 2001) are extensively applied in various domains, including Computer Vision (Peng and McCallum, 2006), Natural Language Processing problems (Kumar and Hebert, 2004) and Bioinformatics (Sato and Sakakibara, 2005), and

various formulations of CRF models named Gaussian Conditional Random Fields (GCRF) were proposed (Tappen et al., 2007; Radosavljevic et al., 2010; Stojanovic et al., 2015, 2016; Gligorijevic et al., 2015, 2016; Vujicic et al., 2017).

Other multi-task regression methods exist that learn the structure among the tasks using norm regularization (Wytock and Kolter, 2013; Zhou et al., 2011), or methods that utilize fixed relatedness structure (Stojkovic et al., 2016a) obtained from domain knowledge (Stojkovic and Obradovic, 2017a) or learned from a statistical correlation (Stojkovic et al., 2017a). However, since they are not directly proposed for ranking-based learning of the scoring functions, we will not consider them, nor will compare with them in this work.

To the best of our knowledge, there are no published multi-task formulations for ranking-based scoring functions, that is, for methods that learn from pairwise comparisons. And especially the ones that handle temporal data. Hence, we propose such formulation and provided solution for its training. The closest approaches are the multi-task regression-based models for Alzheimer's disease progression (Zhou et al., 2012) and search results ranking (Bai et al., 2009).

## 2.3 Functional Norm Regularization

The main problem in multi-task learning is finding the most appropriate assumption on how the tasks are related and incorporating such assumption into the model. Typically, in linear models, such structural assumptions are imposed on the joint parameter matrix, where rows correspond to features and columns correspond to different tasks. Kernel methods assume that all tasks are related and similar (Evgeniou et al., 2005), but some methods enforce tasks to be grouped into clusters (Jacob et al., 2009). For example, "Dirty method" (Jalali et al., 2010) encourages block-structured row-sparsity in the joint parameter matrix by $\|.\|_{1,1}$ norm, and element-wise sparsity with $\|.\|_{1,\infty}$. The robust approach (Gong et al., 2012) selects sparse rows of features

for related tasks with $\|.\|_{2,1}$ and dense columns for outlier tasks with $\|.\|_{1,2}$, in order to discern between related and unrelated tasks. Other approaches assume some shared common set of features (Argyriou et al., 2008) or shared common subspace (Ando and Zhang, 2005; Chen et al., 2009). The approach proposed in (Chen et al., 2011) attempts to learn such relatedness subspace with trace (nuclear) norm $\|.\|_*$ by encouraging the parameter matrix to have low rank, and finding outlier tasks with additional sparse group norm $\|.\|_{1,2}$. While, the approach in (Ghalwash et al., 2016) combines $\|.\|_1$ and $\|.\|_\infty$ to perform structured feature selection.

## 2.4 Proximal Algorithms

Proximal algorithms are a general class of algorithms for solving nonsmooth, constrained and/or high dimensional cases of convex optimization problems (Parikh and Boyd, 2014). Elementary operation in such algorithms is evaluation of the proximal operator of a function, which boils down to solving another (usually simpler) convex optimization problem, that often can be solved very efficiently (for example having a closed form solution). There are number of optimization algorithms that belong to this class, like proximal minimization algorithm (Bertsekas and Tseng, 1994), alternating direction method of multipliers (Boyd, 2011) and proximal gradient methods (Schmidt et al., 2011). Such algorithms have been applied in number of problems with norm induced regularization, like sparse inverse covariance selection (Scheinberg et al., 2010), sparse linear models (Bach et al., 2012) and nuclear-norm regularized models (Toh and Yun, 2010). The optimization approach we propose for multi-task ranking framework is an instance of the projected gradient method, although augmented with the alternating minimization scheme in the outer loop. And even the method of Lagrangian multipliers proposed for solving the dual formulation of high-dimensional ranking framework can be interpreted as special case of alternating direction method of multipliers (Parikh and Boyd, 2014).

## 2.5 Scoring Function Learning

As mentioned in the Introduction section, some of the first proposed severity score learning methods are supervised approaches that solve classification or regression tasks, and whose solution provides a way to calculate a severity score.

For example, in (Zhou et al., 2012) the Alzheimer's Disease severity, as measured by cognitive scores, was modeled as (temporal) multi-task regression using the fused sparse group lasso approach. The approach was more concerned with the progression of the disease, hence the multi-task formulation. However, as we are mostly interested in severity score mapping from a single time-point set of measurements, here we are presenting its more influential ancestor, the LASSO model (Tibshirani, 1996):

$$\underset{w}{\operatorname{argmin}} LASSO(w) = \frac{1}{2}\|Y - Xw\|_2^2 + \lambda\|w\|_1 \tag{2.1}$$

Here, $Y$ is column vector of $n$ given numeric scores, associated with $d$ dimensional measurement matrix $X_{n\times d}$, while $w$ denotes the solution in form of a $d$-dimensional column weight vector. We will use this model as one of the baselines for comparison as it is one of the main workhorses of biomarker selection (Ghosh and Chinnaiyan, 2005) and even statistical learning in general.

Another approach used sparsity-inducing $L_1$ norm in combination with classical loss function for learning disease severity scoring function (Yang et al., 2012). They proposed using $L_1$ regularized Logistic regression model (among others), to model the severity scores for the abnormality of the skull in craniosynostosis cases:

$$\underset{w}{\operatorname{argmin}} L_1 LogReg(w) = \sum_{i=1}^{n} log(1 + exp(-Y_i(X_i w))) + \lambda\|w\|_1 \tag{2.2}$$

This Sparse Logistic Regression formulation is another related model, as it also results in a sparse vector of feature weights $w$ that essentially regress the decision

boundary between the severity classes and might be used as a mapping function for severity scores. In eq. 2.2, $Y_i \in \{-1, 1\}$ is a binary label for $i$-th row of data matrix $X$.

As outlined previously, these forms of supervision where estimates of severity score functions (or severity classes) are needed, might be hard to obtain in order to be utilized for training the severity score automatically. On the other hand, obtaining the pairs of comparisons is an easier task. Seminal work of learning the scoring functions from the comparison labels is proposed in (Joachims, 2002). In that work, the ranking SVM formulation (eq. 2.3) is developed to learn better document retrieval from click-through data. This great insight came from noticing that the clicked links automatically have greater ranks compared to the ones not clicked. And such kind of data is much more abundant than the user provided rankings.

$$\underset{w}{\text{argmin}} rankingSVM(w) = \frac{1}{2}\|w\|_2^2 + c \sum_{\{p,q\} \in O} max(0, 1 - (X_p - X_q)w) \tag{2.3}$$

Set $O$ is composed of comparison of ordered pairs $\{p, q\}$, where $p$ has a higher rank than $q$ and which corresponds to rows of measurement matrix $X_p$ and $X_q$, respectively. More recently the approach was adopted for learning the Sepsis Disease Severity Score (Dyagilev and Saria, 2015b). In it (eq. 2.4), the constraint that scoring function should gradually evolve over the time was introduced and hence a temporal smoothness term is added. In addition, nonsmooth Hinge loss ($max(0, 1 - Xw)$) is replaced with its smooth approximation, Huber loss ($L_h$), to obtain the formulation of (linear) Disease Severity Score Learning (DSSL) framework:

$$\underset{w}{\operatorname{argmin}} DSSL(w) = \frac{1}{2}\|w\|_2^2 + c \sum_{\{p,q\}\in O} L_h(1 - (X_p - X_q)w) + b \sum_{\{i,i+1\}_s \in S} \left( \frac{(X_{i+1}^s - X_i^s)w}{(t_{i+1}^s - t_i^s)} \right)^2 \tag{2.4}$$

Temporal smoothness term in eq. 2.4 penalizes high rates of change in severity in consecutive time steps $t_i$ and $t_{i+1}$ of a single subject $s$. Set of all consecutive pairs in all subjects is denoted $S$ and constants $c$ and $b$ are hyperparameters determining the cost of respective loss terms.

Nonlinear version of DSSL framework, and its solution in form of gradient boosted regression trees, was also proposed in (Dyagilev and Saria, 2015a). Nevertheless, mentioned DSSL approaches are dense in a sense that they operate on all variables (in case of a linear version all coefficients are typically nonzero). The utility of the approaches in (Dyagilev and Saria, 2015a) was presented on an application with a moderately small number of different clinical information, vitals and laboratory analysis variables and it is not clear how the approach would perform in situations with high-dimensional data common in high-throughput techniques like genetic, genomic, epigenetic, proteomic, etc.

Yet, high-throughput data is also a very rich source of useful biomarkers that could be used for diagnostic and prognostic purposes, as well as for obtaining insight into causal relations (Colburn et al., 2001). Therefore we are proposing an approach that is able to learn a (temporally smooth) scoring function from comparison data while simultaneously performing the selection of most relevant (important) variables.

# CHAPTER 3

## METHODOLOGY: MULTI-TASK FORMULATION

This chapter is based on the work presented in (Stojkovic et al., 2017b), and here we are going to outline the methodology behind the proposed multi-task framework for learning the maximum-margin ranking functions from several distinct but related tasks. We start by formulating the problem.

Let us assume that we have $N$ samples (examples), where each sample $i$ is represented as $X_i \in \mathbb{R}^d$, and where $X_{ij}$ is the value of the feature $j = \{1, 2, \ldots, d\}$ for the sample $i = \{1, 2, \ldots, N\}$. Let us assume that $y_i \in \mathbb{R}$ represents the property of interest (outcome variable) for the sample $i$. Scoring function $score : \mathbb{R}^d \rightarrow \mathbb{R}$ is then a mapping $X_i \mapsto y'_i$ that provides a close estimate $y'_i$ of the true score $y_i$.

However, in many cases the values of the true scoring function are difficult to obtain. In such situations, it is easier to assess the ranking between the scores of two samples $p$ and $q$, i.e. to assert that one has perceived higher score than the other: $score(X_p) > score(X_q)$. Therefore, a set of multiple such ordered pairs can be used to find a projection in the space of measured features, that will preserve the orders in the best possible way, and that might be used as a scoring function.

Moreover, measurements collected on multiple occasions over time might belong to the same subject; In this case, the measurements at each time step will be considered as a sample. We assume that the outcome variable changes gradually (smoothly) over time for the same subject, e.g. the disease severity score changes smoothly over consecutive time points for the same patient. This assumption will lead to improving the quality of the scoring function. We assume that $X_p$ represents the feature vector for the sample $p$ (which could be one particular subject at one particular time point).

## 3.1 Model

In this work, we constrain such functional mapping *score* to the linear case, where the score estimate is computed as a weighted sum of the measured characteristics: $score(X) = w^T X$. Therefore, the problem of learning the scoring function becomes finding the appropriate weight (or parameter) vector $w \in \mathbb{R}^d$.

### *3.1.1 Single task model formulation*

Maximizing the number of correctly ordered training pairs can be performed using the soft max-margin framework expressed in a Hinge loss form (3.1), as suggested in (Joachims, 2002).

$$max(0, 1 - (X_p - X_q)w) \tag{3.1}$$

If sample $p$ should have higher score compared to sample $q$, the formulation (3.1) will favor the weighted difference $(X_p - X_q)w$ that is positive and greater than 1, thus even achieving some margin in the score difference.

The $L_2$ norm on the weight vector $||w||^2$, is introduced to regularize the magnitude of the weights, and to turn the problem into simultaneous maximization of correct ordering and maximization of normalized margin.

Gradual (smooth) change of the scoring function over time can be obtained by penalizing high changes of the score (e.g. for two samples $X^s_{i+1}, X^s_i$ of the same subject $s$), over short time intervals. In (Dyagilev and Saria, 2015b) such effect is achieved by using the temporal smoothness term:

$$\left( \frac{(X^s_{i+1} - X^s_i)w}{(t^s_{i+1} - t^s_i)} \right)^2 \tag{3.2}$$

, which essentially ensures that squared magnitude in difference, normalized with the time interval length, is kept low.

Therefore, for single task formulation of ranking-based scoring function learning, we adopted the Linear Disease Severity Score Learning formulation (Dyagilev and Saria, 2015a) which combines attractive properties of ranking SVM (Joachims, 2002), with temporal smoothness term (3.2) that enforces the gradual change of the scoring function over time:

$$\begin{aligned} \hat{w} = \underset{w}{\operatorname{argmin}} \frac{1}{2}\|w\|_2^2 + c \sum_{\{p,q\} \in O} max(0, 1 - (X_p - X_q)w) \\ + b \sum_{\{i,i+1\}_s \in S} \left( \frac{(X^s_{i+1} - X^s_i)w}{(t^s_{i+1} - t^s_i)} \right)^2 \end{aligned} \tag{3.3}$$

Every measurement (row) vector $X_i$, $i = \{1, 2, \ldots, N\}$ has associated time-stamp $t$, while $\hat{w} \in \mathbb{R}^d$ denotes the solution of the objective 3.3.

Set $O$ is composed of ordered pairs $\{p, q\}$, where $p$ has a higher rank than $q$ ($p$ is perceived to have a higher score than $q$), and which corresponds to the measurement vectors $X_p$ and $X_q$, respectively. Sum of the Hinge loss terms over all pairs from the $O$ set, serves to reduce the extent of incorrectly ordered pairs.

Set of all consecutive pairs in all subjects is denoted $S$ and the sum of the Temporal smoothness terms in eq. (3.3) penalizes high rates of change in score values in

consecutive time steps $t_i$ and $t_{i+1}$ for all subjects $s \in S$. Scalar constants $c$ and $b$ are hyperparameters that determine the cost of the respective loss terms, the Hinge loss and the Temporal loss.

We aggregate the differences of measurements in the Hinge loss term into a single data matrix $D_{k\times d}$, where $k$ is the number of pairs in the comparison set $O$. Similarly, measurement and temporal difference ratios in the Temporal loss term we write as matrix $R_{l\times d}$, where $l$ is a number of pairs in the consecutive measurements set $S$. We aggregate the $L_2$ norm and temporal smoothness terms (they are essentially weighting the square of optimization parameters) into a single weighted quadratic term $\frac{1}{2}w^TQw$, where $Q$ is constant square matrix defined in eq. (3.4):

$$Q = I + 2bR^TR \tag{3.4}$$

, $I$ being the $d$-dimensional identity matrix.

The formulation (3.3) can now be rewritten more concisely as (3.5):

$$\hat{w} = \underset{w}{\operatorname{argmin}} \frac{1}{2}w^TQw + c\sum_i max(0, 1 - D^iw) \tag{3.5}$$

### *3.1.2 Multi-task model formulation*

As mentioned before, in case of a limited amount of data for training the scoring function for a single task (3.5), it is beneficial to exploit the relatedness among the multiple similar tasks, by learning them together, as illustrated in Figure 3.1.

For $m$ different tasks, individual parameter vectors $w_i$ are aligned into a matrix $W_{d\times m}$, and a joint objective is obtained as a superposition of individual losses (eq. (3.5)) over the multiple tasks $i \in \{1, 2, ..., m\}$:

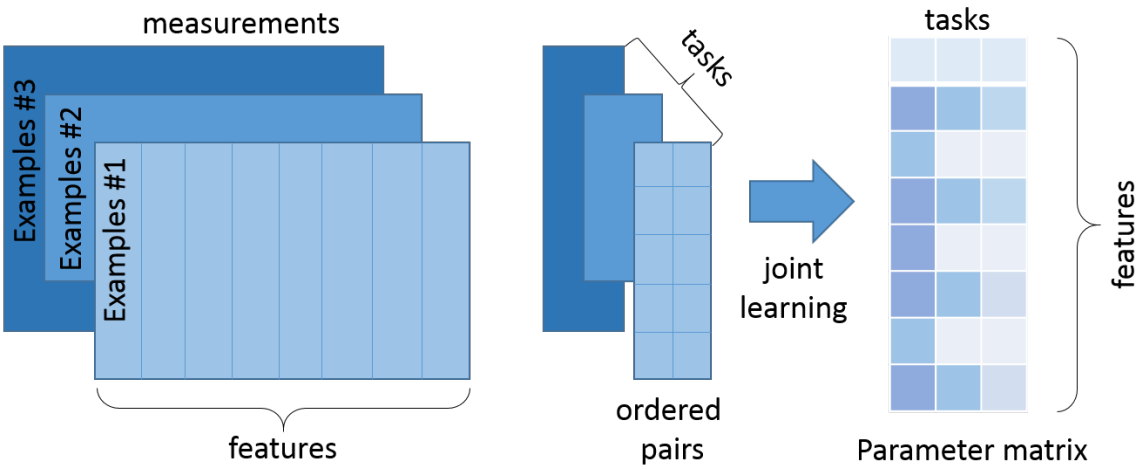


FIGURE 3.1: Illustration of joint training of multiple ranking based score learning tasks. Three distinct task are depicted, where measured data in combination with supervision in form of ordered pairs, are jointly optimized to obtain the scoring function parameters, represented as parameter matrix. Parameter matrix is typically regularized to encode the structural assumptions regarding the task relatedness.

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{m} \left( \frac{1}{2} W_i^T Q_i W_i + c \sum_j max(0, 1 - D_i^j W_i) \right) \tag{3.6}$$

Instead of the non-smooth Hinge loss $L(a) = max(0, a)$ in eq. (3.6), we work with the twice differentiable approximation in the form of Huber loss (Dyagilev and Saria, 2015a):

$$L_h(a) = \begin{cases} 0 & \text{, if } a < -h \\ \frac{(a+h)^2}{4h} & \text{, if } |a| \leqslant h \\ a & \text{, if } a > h. \end{cases} \tag{3.7}$$

, where the approximation threshold $h$ can be chosen arbitrarily small.

Further, we regularize the objective in eq. (3.6) with a joint norm on parameter matrix $\|W\|_{p,q} = (\sum_i ((\sum_j (W_{ij}^q)^{\frac{1}{q}})^p)^{\frac{1}{p}}$. For $p = 2$ and $q = 1$, this approach is known as a group Lasso penalty on the row groups (of $W$), which forces sparsity in the parameter weights corresponding to certain features (Argyriou et al., 2008). Addi-

tionally, we introduce the trace norm $L_*$ in order to get the low rank component, or in other words, the parameter weight pattern common among all the tasks. To accommodate such a setup, which will be further clarified in the Optimization section, the parameter matrix $W$ was split into two distinct matrices $A$ and $B$, where $W = A + B$.

Multitask Ranking Based Scoring Function Learning (MultiRBSFL) objective is now given in eq. (3.8), and it takes as an input two matrices (per task $i$) obtained from the data: $Q^i_{d\times d}$ and $D^i_{k\times d}$; hyperparameters $b$, $c$, $\lambda_1$ and $\lambda_2$ weighting the influence of Temporal loss, Huber loss, trace norm and sparse group norm, respectively.

$$\underset{W=A+B}{\operatorname{argmin}} \mathcal{L}_1 + \lambda_1\|A\|_* + \lambda_2\|B\|_{2,1} \tag{3.8}$$

where

$$\mathcal{L}_1 = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}(A^i + B^i)^T Q^i (A^i + B^i) + c\sum_{j=1}^{k} L_h(1 - D^i_j(A^i + B^i))\right) \tag{3.9}$$

$A^i$ and $B^i$ are column vectors $\mathbb{R}^{d\times 1}$, and $D^i_j$ is $\mathbb{R}^{1\times k}$ row-vector.

## 3.2 Optimization

The optimization (3.8) is composed of smooth and non-smooth terms. However, although the regularization terms are separable in $A$ and $B$, the loss term $\mathcal{L}_1$ is not separable. Therefore, we solve the problem by using the alternating minimization scheme, where, in each iteration, we fix $A$ and minimize (3.8) with respect to $B$, and then fix $B$ and minimize (3.8) w.r.t $A$. In this case, each subproblem can be decomposed into two different optimizations. This will be explained in the next section.

**Fix** $A$

$$\underset{B}{\operatorname{argmin}} \mathcal{L}_1 + \lambda_2 \|B\|_{2,1} \tag{3.10}$$

**Fix** $B$

$$\underset{A}{\operatorname{argmin}}\ \mathcal{L}_1 + \lambda_1 \|A\|_* \tag{3.11}$$

In general, problem (3.10) and (3.11) can be written as:

$$\underset{\mathbf{\Theta}}{\operatorname{argmin}}\ \mathcal{L}_1 + \gamma \|\mathbf{\Theta}\|_p \tag{3.12}$$

, where $\mathbf{\Theta} = \{A, B\}$ and $p = \{*, \{2, 1\}\}$.

The optimization (3.12) is convex. The expression $\mathcal{L}_1$ is smooth and the regulariation term (either group lasso or trace norm) is non-smooth. Therefore, we solve (3.12) using the proximal methods.

### *3.2.1 Proximal Algorithm*

We solve (3.12) using the proximal gradient method (Parikh and Boyd, 2014).

$$\begin{aligned}\mathbf{\Theta}^{k+1} &:= \mathbf{prox}_{\lambda\|\mathbf{\Theta}\|_p}(\mathbf{\Theta}^k - \lambda\nabla\mathcal{L}_1(\mathbf{\Theta}^k)) \\ &= \underset{\mathbf{\Theta}}{\operatorname{argmin}} \left( \|\mathbf{\Theta}\|_p + \frac{1}{2\lambda} \left\|\mathbf{\Theta} - (\mathbf{\Theta}^k - \lambda\nabla\mathcal{L}_1(\mathbf{\Theta}^k))\right\|_2^2 \right)\end{aligned} \tag{3.13}$$

, where $\mathbf{prox}_{\lambda\|\mathbf{\Theta}\|_p}$ is the proximal operator of the scaled function $\|\mathbf{\Theta}\|_p$, and $\lambda \in (0, 1/L]$ is a *constant* step size, and $L$ is a Lipschitz constant of $\nabla\mathcal{L}_1$. Problem (3.12) can be solved analytically, where the proximal operator associated with the norm can be obtained as in (Bach et al., 2011).

**Trace norm.** Let us assume that $M = U\Sigma V$ is the singular value decompoistion of $M$, where $\Sigma$ is a diagonal matrix and its entries $\sigma_i$ are the singluar values of the matrix $M$. The proximal operator of the trace norm is defined as (Cai et al., 2010):

$$\mathbf{prox}_{\lambda\|.\|_*}(M) = U\mathbf{diag}(\mathbf{prox}_{\lambda\|.\|_1}(\sigma(M)))V$$

i.e., the proximal operator of $\|.\|_*$ can be calculated by carrying out a singular value decomposition of $Z$ and evaluating the proximal operator of the corresponding absolutely symmetric function at the singular values $\sigma(M)$. Therefore,

$$\mathbf{prox}_{\lambda\|.\|_*}(M) = U\mathbf{diag}(\overline{\sigma}_1, \overline{\sigma}_2, \ldots, \overline{\sigma}_n)V \tag{3.14}$$

, where:

$$\overline{\sigma}_i = \begin{cases} \sigma_i - \lambda & \sigma_i \geqslant \lambda \\ 0 & -\lambda \leqslant \sigma_i \leqslant \lambda \\ \sigma_i + \lambda & \sigma_i \leqslant -\lambda \end{cases}$$

Equation (3.14) is sometimes called the singular value thresholding operator.

**Group lasso norm.** The proximal operator associated with the group lasso norm is defined as:

$$\left[\mathbf{prox}_{\lambda\|\cdot\|_{1,2}}(u)\right]_g = \begin{cases} (1 - \frac{\lambda}{\|u_g\|_2})u_g & \|u_g\|_2 > \lambda \\ 0 & \text{otherwise} \end{cases}$$

### *3.2.2 Step size*

In order to find an adaptive step size $\lambda^k$ in each iteration $k$, we employ the backtracking line search algorithm (Beck and Teboulle, 2009), which requires computing an upper bound for $\mathcal{L}_1$. Since $\mathcal{L}_1$ is convex and smooth, and $\nabla\mathcal{L}_1$ is $L$-Lipschitz continuous, it follows that:

$$\mathcal{L}_1(\mathbf{\Theta}) \leqslant \underbrace{\mathcal{L}_1(\mathbf{\Theta}^k) + \nabla\mathcal{L}_1(\mathbf{\Theta}^k)^T(\mathbf{\Theta} - \mathbf{\Theta}^k) + \frac{L}{2}\left\|\mathbf{\Theta} - \mathbf{\Theta}^k\right\|_2^2}_{\widehat{\mathcal{L}_1}_{\frac{1}{L}}(\mathbf{\Theta}, \mathbf{\Theta}^k)} \tag{3.15}$$

By utilizing (3.15), it can be shown that the optimization (3.13) is equivalent to (Parikh and Boyd, 2014):

$$\mathbf{\Theta}^{k+1} := \underset{\mathbf{\Theta}}{\operatorname{argmin}}\ \widehat{\mathcal{L}_1}_{\lambda^k}(\mathbf{\Theta}, \mathbf{\Theta}^k) + \|\mathbf{\Theta}\|_p \tag{3.16}$$

where $\lambda^k = \frac{1}{L}$. So at each iteration, the function $\mathcal{L}_1$ is linearized around the current point and the problem (3.16) is solved. The final fast proximal gradient method with backtracking is shown in Algorithm 1. The final alternate minimization algorithm is shown in Algorithm (2).

**Algorithm 1** Fast Gradient Proximal Method with Backtracking Step Size

**Input:** $\mathbf{\Theta}^0$ (random), $\eta$ (usually 1/2), $L > 0$
$\lambda = \frac{1}{L}$, $\mathbf{z}^1 = \mathbf{\Theta}^0$, $t_1 = 1$, $k = 0$
**repeat**
    $k \longleftarrow k + 1$
    **while** true **do**
        $\mathbf{z} \longleftarrow$Solve (3.12) ▷ use $\lambda$ and $\mathbf{z}^k$
        **if** $\mathcal{L}_1(\mathbf{z}) \leqslant \widehat{\mathcal{L}_1}(\mathbf{z}, \mathbf{z}^k)$ **then**
            break
        **end if**
        $\lambda \longleftarrow \eta\lambda$
    **end while**
    $\mathbf{\Theta}^k \longleftarrow \mathbf{z}$
    $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
    $\mathbf{z}^{k+1} = \mathbf{\Theta}^k + (\frac{t_k - 1}{t_{k+1}})(\mathbf{\Theta}^k - \mathbf{\Theta}^{k-1})$
**until** Convergence

**Algorithm 2** Alternate Minimization

**Input:** $A^0, B^0$ (random)
**repeat**
    Fix $A$, solve (3.10) using Algorithm (1).
    Fix $B$, solve (3.11) using Algorithm (1).
**until** Convergence

# CHAPTER 4

# METHODOLOGY:HIGH-DIMENSIONAL FORMULATION

This chapter is based on the material presented in (Stojkovic and Obradovic, 2017b), and here we present the methodological ground behind the proposed high-dimensional framework for maximum-margin ranking on temporal data, dubbed Sparse Learning of Disease Severity Score formulation (SLDSS).

## 4.1 Model

In SLDSS we combine attractive properties (and terms) of previously mentioned approaches, ranking SVM (eq. 2.3) (Joachims, 2002), temporal smoothness constraint (eq. 2.4) (Dyagilev and Saria, 2015a) and $L_1$ norm from sparse methods (eqs. 2.1 and 2.2) (Tibshirani, 1996; Yang et al., 2012):

$$\min_{w} SLDSS(w) = \frac{1}{2}\|w\|_2^2 + c \sum_{\{p,q\}\in O} max(0, 1 - (X_p - X_q)w) + b \sum_{\{i,i+1\}_s \in S} \left( \frac{(X_{i+1}^s - X_i^s)w}{(t_{i+1}^s - t_i^s)} \right)^2 + \lambda\|w\|_1 \tag{4.1}$$

In fact, since the model imposes both $L_1$ and $L_2$ norms on the feature vector $w$, it resembles the elastic net regularization (Zou and Hastie, 2005), which has an advantage of achieving higher stability with respect to random sampling (De Mol et al., 2009). Similar model was previously proposed in (Wang et al., 2006), just without the temporal component.

The solution $w^*$ of the optimization objective defined in eq. 4.1 serves as a sparse linear function $f(X) = Xw^*$ that may be applied on measurements from the new patient, to obtain a scalar value of severity that might be compared to previously assessed cases and inform further actions. The sparse vector $w^*$ may also serve as an indicator of which features are the most influential for pairwise comparison. The formulation contains two nonsmooth terms, $L_1$ and Hinge loss, and therefore it is not directly solvable using off-the-shelf gradient methods. In DSSL formulation (Dyagilev and Saria, 2015a) the (non-differentiable) Hinge loss is approximated with twice differentiable Huber loss, thus making the optimization criterion solvable using the second order gradient methods (eg. Newton, Quasi-Newton). In order to provide an efficient solution for the proposed nonsmooth objective, we will solve the smooth dual problem instead of relying on smooth approximation or nonsmooth optimization tools.

First we rewrite eq. 4.1 into a more suitable form for which we will later provide the smooth dual problem. We aggregate the differences of measurements into single

data matrix $D_{k\times d}$, where $k$ is a number of pairs in the comparison set $O$. Similarly, we express measurement and temporal difference ratios as matrix $R_{l\times d}$, where rows are $R_i = \frac{(X^s_{i+1}-X^s_i)}{(t^s_{i+1}-t^s_i)}$ and $l$ is a number of pairs in the consecutive measurements set $S$. We aggregate the $L_2$ norm and temporal smoothness terms (they are essentially weighting the square of optimization parameters) into a single weighted quadratic term $\frac{1}{2}w^TQw$, where $Q = I + 2bR^TR$, $I$ being $d$-dimensional identity matrix. The first two terms, weighted quadratic norm and Hinge loss, resemble the well-known SVM criterion function that we will rewrite in its "soft" form with additional slack variables $z_i$ and their associated constraints. Additional set of "dummy variables" $y$ is introduced in $L_1$ term, with trivial constraints $w = y$. The equation of the rewritten SLDSS now reads:

$$\begin{aligned} &\min_w SLDSS(w,z,y) = \frac{1}{2}w^TQw + c\sum_{i=1}^{k} z_i + \lambda\|y\|_1 \\ &s.t. \quad D_iw >= 1 - z_i, \quad z_i >= 0, \quad \forall i \in \{1,...,k\}, \quad w = y \end{aligned} \tag{4.2}$$

Now we turn this constrained problem with inequalities and equalities into its Lagrangian dual. Constraints are moved to the criterion function as penal terms weighted by Lagrangian multipliers $\alpha$, $\beta$ and $\gamma$. The equation of the SLDSS dual problem is:

$$\begin{aligned} &\min_{w,y,z\geqslant 0} \quad \max_{\alpha\geqslant 0,\beta\geqslant 0} \quad Dual(w,y,z,\alpha,\beta,\gamma) \quad = \\ &\frac{1}{2}w^TQw + c\mathbf{1}^Tz + \alpha^T(\mathbf{1} - z - Dw) - \beta^Tz + \lambda\|y\|_1 + \gamma^T(w-y) \end{aligned} \tag{4.3}$$

Given that optimization criterion is convex and feasible (Slater's condition holds (Boyd and Vandenberghe, 2004)), strong duality allows switching the order of maximization and minimization in eq. 4.3, and minimization in primal variables can be

safely performed first. Now we analyze the expression according to primal variables $w$, $y$ and $z$ and find the minimizing conditions for each of them.

The dual formulation is the quadratic function of parameters $w$ and we can find its optimal form as a function of new free parameters introduced in dual (by equating its gradient with zero):

$$\begin{aligned} \min_w DUAL(w) &= \min_w \frac{1}{2}(w^T Q - \alpha^T D + \gamma^T)w \\ \nabla_w DUAL(w) &= w^T Q - \alpha^T D + \gamma^T = \mathbf{0} \\ &\Rightarrow w^* = Q^{-1}(\alpha^T D - \gamma^T) \end{aligned} \tag{4.4}$$

Similarly, the expression for slack variables $z$ is a linear combination of dual variables and it is minimal when the directional gradient is equated to zero vector, giving the optimality condition in a form of an equality constraint:

$$\begin{aligned} \min_z DUAL(z) &= \min_z \frac{1}{2}(c\mathbf{1}^T - \alpha^T - \beta^T)z \\ \nabla_z DUAL(z) &= c\mathbf{1}^T - \alpha^T - \beta^T = \mathbf{0} \\ &\Rightarrow \beta = c\mathbf{1} - \alpha \end{aligned} \tag{4.5}$$

Resulting equality constraint $\beta = c\mathbf{1} - \alpha$ in combination with inequality $\beta \geqslant \mathbf{0}$ can be reduced to just one constraint $\alpha \leqslant c\mathbf{1}$, which removes $\beta$ from further consideration.

For minimization over dummy variables $y$ we use the convex (Fenchel) conjugate function of the expression (Boyd and Vandenberghe, 2004), and obtain optimality condition as inequality constraint over the infinity norm of the dual variable:

$$\begin{aligned} \min_y DUAL(y) &= \min_y \lambda\|y\|_1 - \gamma^T y = -\max_y \gamma^T y - \lambda\|y\|_1 \\ &= 0 \quad if \quad \|\gamma\|_\infty \leqslant \lambda \quad , \quad or \quad = -\infty \quad otherwise \\ &\Rightarrow \|\gamma\|_\infty \leqslant \lambda \end{aligned} \tag{4.6}$$

When optimal (minimizing) conditions (eqs. 4.4, 4.5 and 4.6) are replaced in dual formulation eq. 4.3, it becomes:

$$\begin{aligned}\max_{\alpha\geqslant 0,\alpha\leqslant c\mathbf{1},\|\gamma\|_\infty\leqslant\lambda} DUAL(\alpha,\gamma) = \frac{1}{2}(D^T\alpha-\gamma)^T Q^{-1}QQ^{-1}(D^T\alpha-\gamma)\\ -\alpha^T DQ^{-1}(D^T\alpha-\gamma)+\gamma^T Q^{-1}(D^T\alpha-\gamma)+\mathbf{1}^T\alpha\end{aligned} \tag{4.7}$$

After negating the equation 4.7 to turn it into minimization problem and after simplification of the expression, final problem formulation is:

$$\begin{aligned}min\frac{1}{2}(D^T\alpha-\gamma)^T Q^{-1}(D^T\alpha-\gamma)-\mathbf{1}^T\alpha\\ s.t.\quad \mathbf{0}\leqslant\alpha\leqslant c\mathbf{1},\quad -\lambda\mathbf{1}\leqslant\gamma\leqslant\lambda\mathbf{1}\end{aligned} \tag{4.8}$$

The original nonsmooth problem is turned into the smooth dual problem, which can be solved for its two sets of parameters $\alpha$ and $\gamma$. Since the strong duality holds, a solution to the dual is a solution to the original problem, and optimal weight vector $w^*$ can be retrieved after plugging in the solution of dual, $\alpha^*$ and $\gamma^*$, into equation 4.4.

Similar dual formulation, just without the dummy variables $y$ and associated multipliers $\gamma$, might be used for DSSL with the exact Hinge loss, instead of the originally proposed DSSL which uses Hubber loss approximation (Dyagilev and Saria, 2015a).

## 4.2 Optimization algorithm

The differentiable dual from eq. 4.8 is, in fact, a quadratic optimization problem with box constraints:

$$\min_x \frac{1}{2}x^T Hx+\mathbf{f}^T x \quad s.t. \quad \mathbf{lb}\leqslant x\leqslant\mathbf{ub} \tag{4.9}$$

$$x = \begin{bmatrix} \alpha \\ \gamma \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} \mathbf{1}_{k\times 1} \\ \mathbf{0}_{d\times 1} \end{bmatrix}, \quad \mathbf{lb} = \begin{bmatrix} \mathbf{0}_{k\times 1} \\ -\lambda \mathbf{1}_{d\times 1} \end{bmatrix}, \quad \mathbf{ub} = \begin{bmatrix} c\mathbf{1}_{k\times 1} \\ \lambda \mathbf{1}_{d\times 1} \end{bmatrix}$$

$$H = \begin{bmatrix} D & \mathbf{0}_{k\times d} \\ \mathbf{0}_{d\times d} & -I \end{bmatrix} \begin{bmatrix} Q^{-1} & Q^{-1} \\ Q^{-1} & Q^{-1} \end{bmatrix} \begin{bmatrix} D^T & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times k} & -I \end{bmatrix}$$

There are ready to use tools for solving the problem in eq. 4.9, and we utilized the built-in Matlab "quadprog" solver, which is implemented as a projection method with the active set.

# CHAPTER 5

## RESULTS: MULTI-TASK FRAMEWORK

This chapter is based on the work presented in (Stojkovic et al., 2017b), and here we are going to present empirical evaluation of the proposed multi-task framework for learning the maximum-margin ranking functions from several distinct but related tasks, dubbed MultiRBSFL.

The proposed approach for multitask learning of ranking-based scoring functions is tested on one synthetic and two real-world datasets. We compared our MultiRBSFL approach against the following baseline approaches:

1. $L_2$ - independently learning ($L_2$ regularized) scoring functions for each task (objective (3.3));

2. $L_1$ - independently learning sparse ($L_1$ regularized) scoring functions for each task;

3. $L_*$ - learning multiple scoring functions by imposing low rank regularization on their joint parameter matrix ($L_*$ regularized objective (3.6));

4. $L_{2,1}$ - joint objective (3.6), regularized by mixed $\|.\|_{2,1}$ norm.

Our MultiRBSFL approach, which uses composite low rank and mixed norm regularized joint objective (3.8), we will denote as $L_* + L_{2,1}$ for consistency in naming the alternative approaches.

We measured the predictive performance in terms of accuracy, which is the number of correctly ordered test pairs. As the pairwise ranking relation is antisymmetric, it is sufficient to use only the positive training instances (i.e. where the first sample in a pair has the larger score). Test pairs are exclusively generated from examples not contained in the training set. Accuracy values that we report in this study are obtained by doing 5-fold cross-validation experiments.

## 5.1 Experiments on Synthetic Data

In this settings, a Gaussian processes model (Rasmussen and Williams, 2006) with an exponential kernel was used to generate the temporal data (as visualized in Figure 5.1). We compiled 250 such processes to mimic $d = 250$ measured variables (features) per subject. Each single process was used to generate a time series with 10 time points (10 samples). We followed the same principle to generate 10 different multivariate time series (subjects) for training and 10 subjects for test, resulting in 100 samples $X^{train}_{100\times250}$ for training, and 100 samples $X^{test}_{100\times250}$ for test.

Four different tasks were created by randomly generating the weight matrix $W_{250\times4}$, with only 5 nonzero rows, which corresponds to the $L_{2,1}$ assumption (row-sparsity). This row-wise sparse matrix was then superimposed with a dense rank-1 matrix, generated by multiplication of two random vectors, which suits the $L_*$ trace norm part of the objective. True underlying scores on four tasks, for each of the 250-dimensional samples (one time point of one patient), are calculated as the weighted sum of the feature values $X * W$. Zero mean random vector was subsequently superimposed to input $X$ data, before using it to fit the scoring function, in order to simulate the measurement noise.

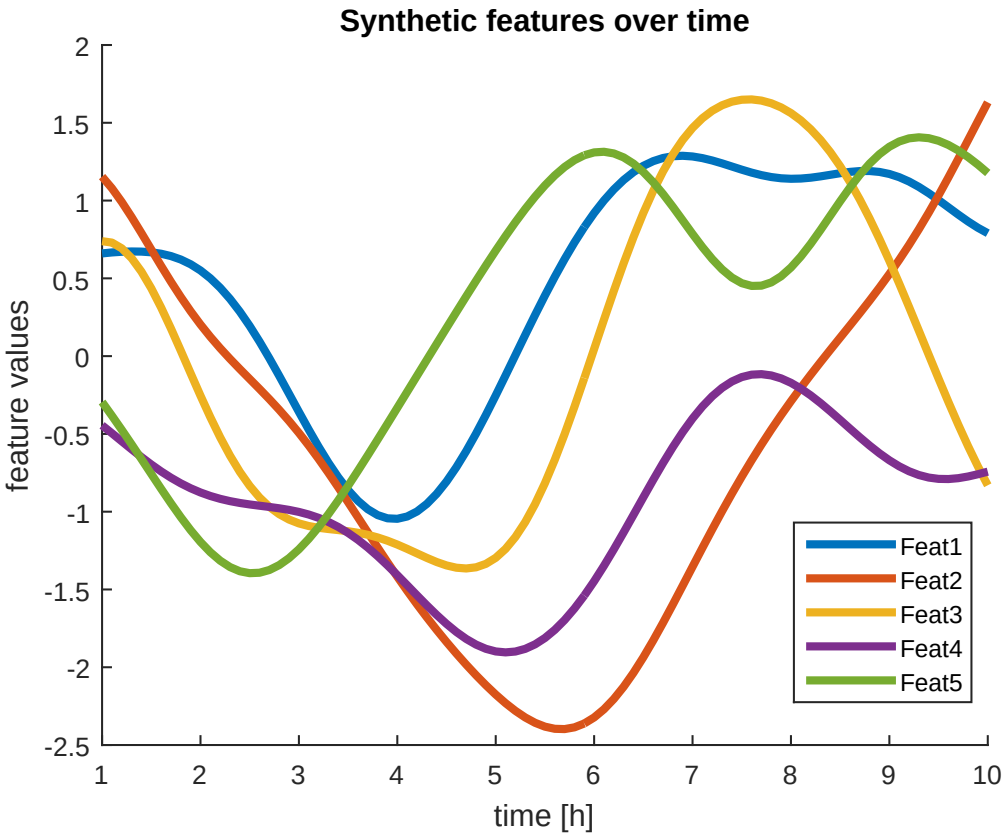


FIGURE 5.1: Example of 5 temporal features obtained from Gaussian Processes, of one fictitious subject, with enforced assumption that temporally close points have similar intensities.

A training set is then obtained by making pairs out of samples whose scores are sufficiently different (in our case we set the threshold to 1). Pairs of examples were generated independently for each task based on their scores, totaling 14,187 pairs for all four tasks jointly. Test set pairs were generated in the same fashion, but with a smaller threshold and consisted out of 19,390 pairs. Training pairs were used to learn the weight matrix $\hat{W}$, which was used to estimate the testing scores from the test samples. The obtained estimates were used to infer the relative order of the testing pairs. The accuracy (percentage of correct guesses) is reported in the Table 5.1. It is no surprise that the proposed $L_* + L_{1,2}$ approach achieves the highest accuracy on all four tasks, as the underlying assumptions are explicitly built into synthetic data.

Table 5.1: Comparison of accuracy indicators (fraction of correctly ordered pairs) for alternative score learning methods on the synthetic data of four related tasks.

| Task | $L_2$ | $L_1$ | $L_*$ | $L_{1,2}$ | $L_* + L_{1,2}$ |
|---|---|---|---|---|---|
| TASK1 | 0.538 | 0.745 | 0.680 | 0.744 | **0.757** |
| TASK2 | 0.556 | 0.707 | 0.763 | 0.782 | **0.795** |
| TASK3 | 0.592 | 0.765 | 0.744 | 0.821 | **0.837** |
| TASK4 | 0.466 | 0.864 | 0.700 | 0.874 | **0.885** |
| AVG | 0.538 | 0.770 | 0.722 | 0.805 | **0.818** |

## 5.2 School Exam Score

Intelligence as well as the capacity for understanding and using mathematics or languages are all examples of properties that are latent - yet important and often evaluated (estimated). We have tested the multitask score learning framework on data from an elementary school study (Mortimore et al., 1988), which contains longitudinal data on performance in Math and English language for pupils in 50 inner London schools [1]. In total there are scores for 3,236 exams (Math and English each), taken by 1,402 students over three consecutive school years. The goal is to rank the students' performances on Math and English test based on known score from Ravens ability test and additional information like demographics, social status, gender, class and school type. Distributions of scores for two tasks are given in the Figure 5.2 and Figure 5.3, respectively.

According to results depicted in Table 5.2, our $L_* + L_{1,2}$ approach achieved the best predictive performance in both tasks.

## 5.3 Tolerance to Infections Score

Tolerance is the host's behavior that arises from interactions with a pathogen, which describes the ability of the host to preserve fitness despite the presence of a large amount of pathogen. Therefore, it is defined as changes in host fitness (health)

[1] http://www.bristol.ac.uk/cmm/media/migrated/jsp.zip

Table 5.2: Comparison of accuracy indicators (fraction of correctly ordered pairs) for alternative score learning methods on the task of learning the performance on Math and English tests.

| task | $L_2$ | $L_1$ | $L_*$ | $L_{1,2}$ | $L_* + L_{1,2}$ |
|---|---|---|---|---|---|
| MATH | 0.780 | 0.794 | 0.725 | 0.789 | **0.812** |
| ENGLISH | 0.820 | 0.863 | 0.717 | 0.857 | **0.870** |
| AVG | 0.800 | 0.828 | 0.721 | 0.823 | **0.841** |

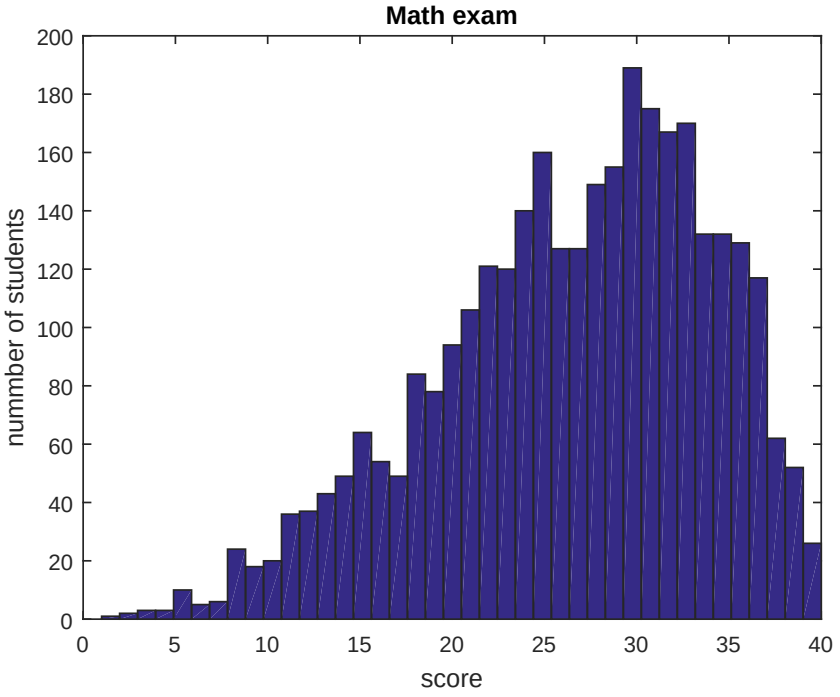


FIGURE 5.2: Distribution of test scores for Math exam.

with respect to changes in pathogen load (Simms, 2000). However, tolerance is a very understudied topic, where there is no established scoring function, despite the necessity.

We analyzed three publicly available datasets [2] that allows characterization of the tolerance behavior in humans. The data comes from the human viral challenge studies (Zaas et al., 2009) where human volunteers were infected with H3N2 influenza, rhinovirus (HRV) and respiratory syncytial virus (RSV), respectively, and which is

[2] http://people.ee.duke.edu/ lcarin/reproduce.html

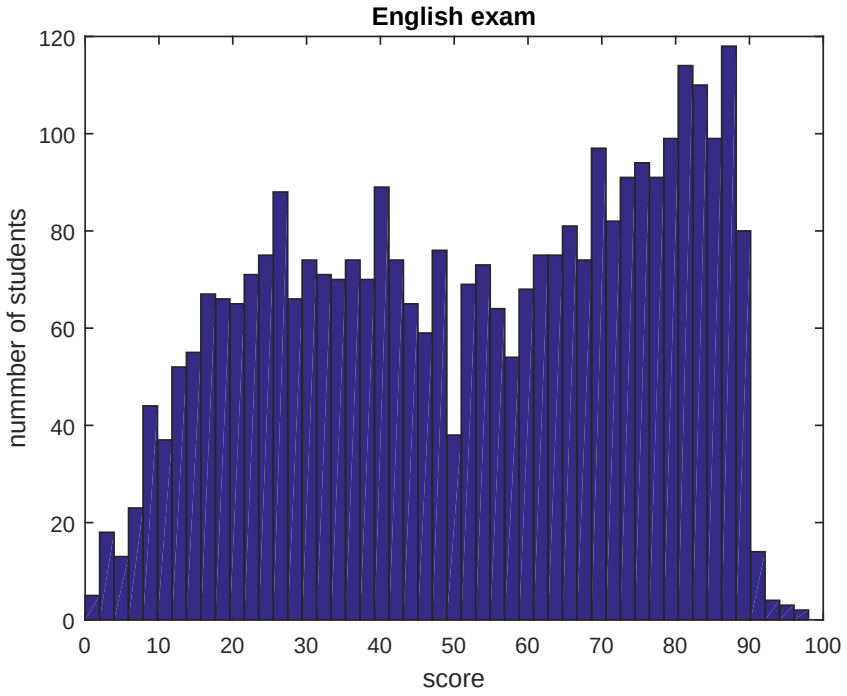

FIGURE 5.3: Distribution of test scores for English exam.

detailed in Appendix A. Table 5.3 shows the viral shedding and symptom scores for subjects who developed clinically relevant symptoms from H3N2, HRV and RSV datasets.

Temporal measurements about symptoms (proxy for fitness) and viral (pathogen) load for each subject were used to derive tolerance scores according to the definition given in (Simms, 2000). In particular, the tolerance score for each subject was calculated by dividing the maximum viral load with the maximum severity of symptoms (Jackson et al., 1958) observed for that subject (Table 5.3). Gene expression measurements were used as an explanatory variables in our ranking task.

Biological rationale behind the task relatedness is that the three infections are viruses that cause similar respiratory symptoms (runny nose, fever, sore throat, cough) and are quantified by the same Jackson score (Jackson et al., 1958), suggesting that some shared genetic mechanisms might be responsible for the disease

Table 5.3: Tolerance scores (R) derived by dividing maximum viral load (V) with maximum severity score (S).

| H3N2 | | | | HRV | | | | RSV | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sub ID | S | V | R | Sub ID | S | V | R | Sub ID | S | V | R |
| FLU05 | 12.00 | 5.45 | 0.45 | HRV06 | 8.00 | 2.72 | 0.34 | RSV01 | 11.00 | 0.00 | 0.00 |
| FLU08 | 10.00 | 4.70 | 0.47 | HRV19 | 2.00 | 0.95 | 0.47 | RSV20 | 6.00 | 0.00 | 0.00 |
| FLU01 | 9.00 | 4.25 | 0.47 | HRV04 | 8.00 | 3.94 | 0.49 | RSV07 | 20.00 | 4.46 | 0.22 |
| FLU07 | 12.00 | 6.25 | 0.52 | HRV15 | 7.00 | 3.45 | 0.49 | RSV02 | 20.00 | 5.10 | 0.26 |
| FLU06 | 7.00 | 5.00 | 0.71 | HRV07 | 7.00 | 4.44 | 0.63 | RSV12 | 4.00 | 2.50 | 0.62 |
| FLU10 | 5.00 | 3.75 | 0.75 | HRV20 | 6.00 | 4.44 | 0.74 | RSV06 | 9.00 | 5.65 | 0.63 |
| FLU12 | 4.00 | 5.00 | 1.25 | HRV16 | 6.00 | 4.69 | 0.78 | RSV14 | 6.00 | 4.54 | 0.76 |
| FLU15 | 2.00 | 4.50 | 2.27 | HRV09 | 3.00 | 2.46 | 0.82 | RSV11 | 5.00 | 3.85 | 0.77 |
| FLU13 | 2.00 | 5.45 | 2.70 | HRV11 | 3.00 | 2.47 | 0.83 | RSV03 | 6.00 | 4.70 | 0.78 |
| | | | | HRV03 | 4.00 | 3.45 | 0.86 | | | | |

manifestations. Consequently, we sought to learn the tolerance scoring functions jointly.

The tolerance scores were used to compile a set of ranked pairs, and the objective was to learn the scoring functions for tolerance to H3N2, HRV and RSV viruses (3 tasks), from high-dimensional gene expression data. Since 12,023 dimensions is very computationally expensive to optimize, we reduced the dimensionality of the data to the 100 most informative genes according to the correlation with the target. Prior to the model fitting, data was normalized using the method proposed in (Cao et al., 2016). The results of learning the scoring functions with different approaches are summarized in the Table 5.4.

Table 5.4: Comparison of accuracy indicators (fraction of correctly ordered pairs) for alternative score learning methods on learning the tolerance to three human respiratory viral infections.

| task | $L_2$ | $L_1$ | $L_*$ | $L_{1,2}$ | $L_* + L_{1,2}$ |
|---|---|---|---|---|---|
| FLU | 0.766 | 0.980 | 0.809 | 0.988 | **0.996** |
| HRV | 0.344 | 0.122 | 0.389 | **0.500** | 0.400 |
| RSV | 0.806 | **0.972** | 0.861 | 0.306 | 0.861 |
| AVG | 0.638 | 0.692 | 0.686 | 0.598 | **0.752** |

The results from the Table 5.4 show that the HRV task is the most difficult one

in the described formulation. Although some alternative approaches achieved better accuracy in two of the tasks, the proposed approach achieved the best generalization trade-off as can be concluded from the highest average (overall) accuracy.

# CHAPTER 6

# RESULTS: HIGH-DIMENSIONAL FRAMEWORK

This chapter is based on the material presented in (Stojkovic and Obradovic, 2017b), and here we present the empirical evaluation of the proposed high-dimensional framework for maximum-margin ranking on temporal data, dubbed Sparse Learning of Disease Severity Score formulation (SLDSS). The high-dimensional framework is evaluated in number of synthetic and two real applications.

## 6.1 Severity score characterization on synthetic data

For the initial assessment of the proposed SLDSS framework, we have generated a synthetic example with properties that motivated the approach. If a large number of variables is measured, many are expected to be irrelevant for the assessment of severity.

We defined the severity score as a linear combination of intensities of the first 10 features after initiating a set of 100. In addition, we set the coefficients to have different magnitudes, as it is expected that contribution of different variables are of various levels (top panel in Fig. 6.1). The remaining ninety features do not affect

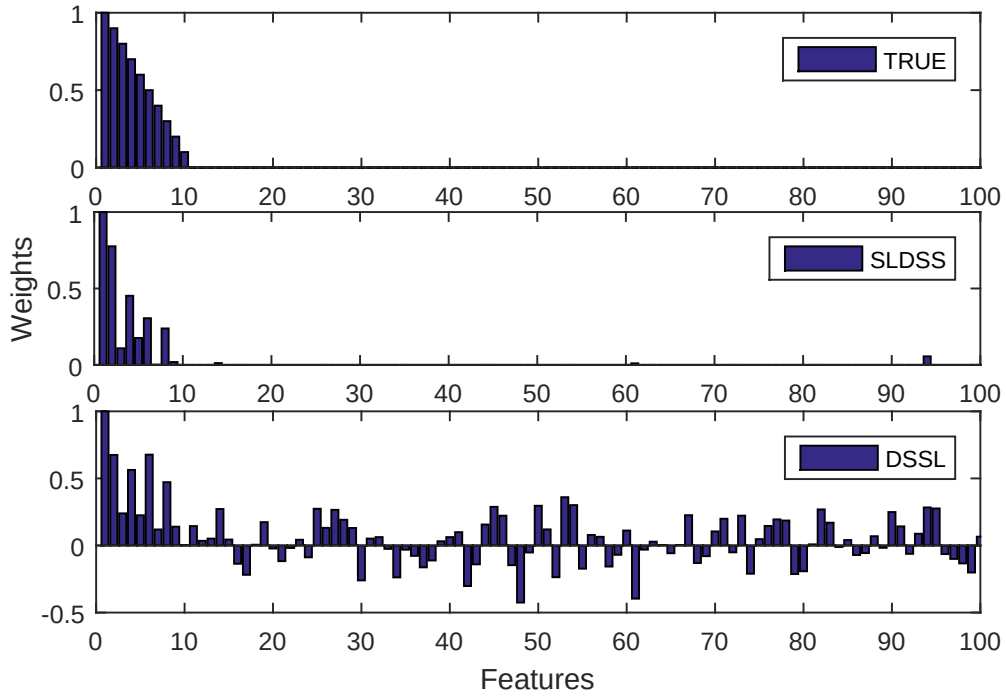


FIGURE 6.1: Comparison of learned weight vectors (normalized) of sparse SLDSS method and dense DSSL method with the ground truth.

severity score at all, they are irrelevant and only introduce uncertainty into the problem. For training purposes, values of all features are randomly sampled from a uniform distribution for 10 fictitious subjects with 10 different measurements each. Severity scores are associated based on a linear function with weights depicted in Fig. 6.1. Comparison labels (pairs) were generated as all possible pairs in which the first element (sample) have substantially higher severity score as compared to the second element. This requirement of substantial gap in severity between pairs serves to mimic the case where a doctor could claim, with high confidence, that one patient is in more severe condition than another. Such generated training data was utilized to fit Sparse LDSS, (dense) DSSL, and DSSL model trained on the exact 10 features that are relevant, which we named Ideal DSSL in Table 6.1.

All models were tested on comparison pairs from an additional 50 test subjects with 10 measurements each. Testing data was generated by the same protocol as

Table 6.1: Performance on synthetic data as measured by correctly ordered pairs - Accuracy, and by aggregated error (magnitude of difference in wrongly ordered pairs) - Hinge loss

| approach | SLDSS | DSSL | IDEAL DSSL |
|---|---|---|---|
| Accuracy | 0.9397 | 0.8373 | 0.9558 |
| Hinge loss | 176.06 | 3110.20 | 180.65 |

explained for training, except the threshold for the required difference of scores was set several times lower, in order to see how learned functions generalize to more subtle differences between the cases.

$$Accuracy = \frac{\#\quad Correctly\quad Ranked}{\#\quad Total\quad Examples} = 1 - \frac{\#\quad Incorrectly\quad Ranked}{\#\quad Total\quad Examples} \tag{6.1}$$

The predictive performance was measured as "Accuracy" eq. 6.1, i.e. the fraction of the total examples that are correctly ordered, meaning that a linear function assigned a higher score to the first component of a pair. The results presented in Table 6.1 show that learning a dense weight vector impairs the predictive accuracy of the model, while learning a sparse vector approaches the ideal accuracy obtained by learning a disease severity score from in advance known relevant features. Fig. 6.1 shows the weights of learned severity functions, and it might be seen that reason for the reduced testing accuracy of the dense DSSL method (bottom panel) is because it assigned nonzero weights to (by design) completely irrelevant features.

### *6.1.1 Feature size analysis*

We have explored how the number of irrelevant features affects the model performance. This time we sampled 100 subjects (with 10 timestep samples each), with $10,000$ features, where only the first 10 contribute to the true score. We varied the number of features from 10 (all features informative), up to $10,000$ in exponentially progressive increments $[10; 30; 100; 300; 1,000; 3,000; 10,000]$. Results presented in

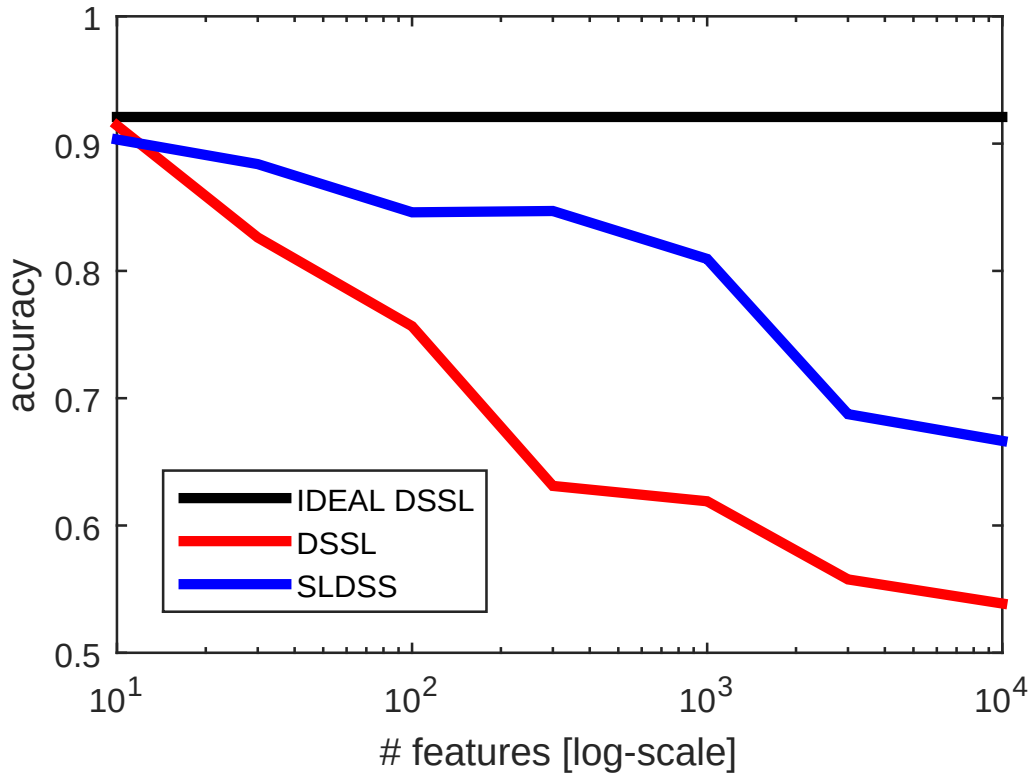


FIGURE 6.2: Influence of the problem dimensionality (number of features) on the accuracy of ranking methods.

Fig. 6.2 suggest that when all available features are informative (10 out of 10) DSSL is slightly better than SLDSS. However, as soon as the irrelevant features are added, the SLDSS approach becomes substantially more accurate than DSSL. As more irrelevant dimensions are added both approaches' performance decreases, but SLDSS at a slower pace.

### *6.1.2 Sample size analysis*

We also investigated how the number of training samples affects the predictive performance of the ranking approaches. We generated another synthetic set of 100 subjects (10 samples each). All samples had 100 features, where the first 10 were relevant for the ground truth score. From such generated examples, we constructed

357, 355 comparison pairs for training. We varied the number of sample pairs, by randomly sampling from 10, up to 300, 000 in exponentially progressive increments [10; 30; 100; 300; 1, 000; 3, 000; 10, 000; 30, 000; 100, 000; 300, 000]. From the results on holdout testing set, presented in Fig. 6.3, it can be seen that accuracy increases with the number of training pairs, and that SLDSS is always more accurate than DSSL. The IDEAL DSSL, which is always trained only on the 10 relevant features, is consistently the most accurate.

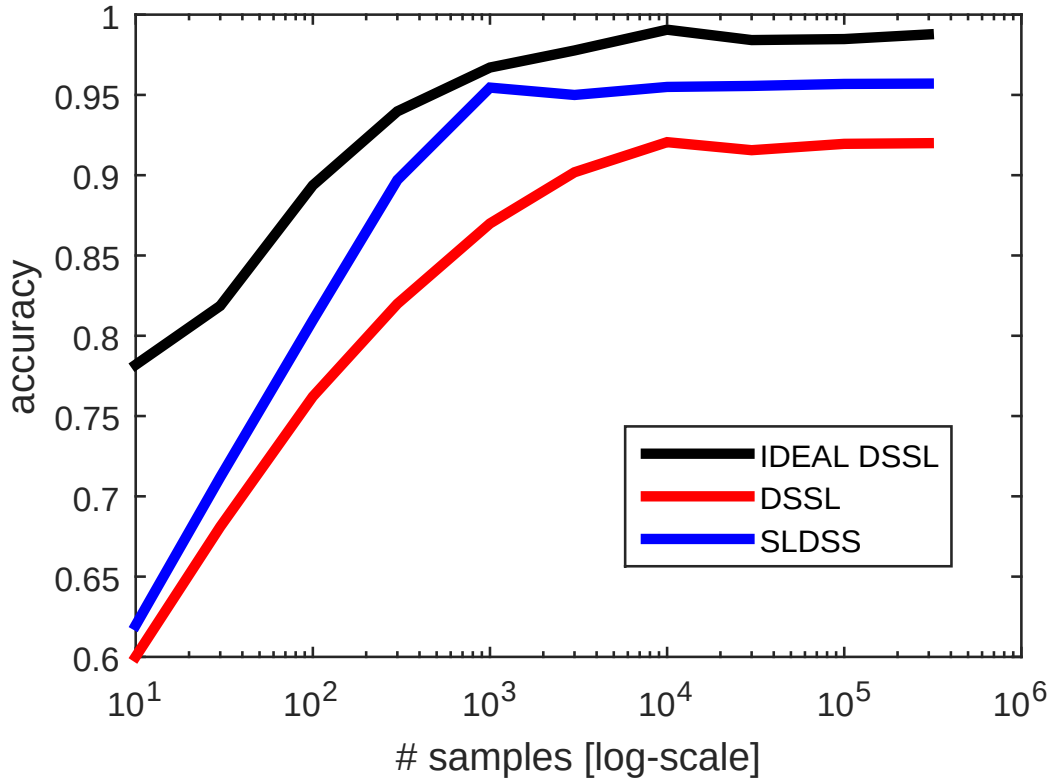


FIGURE 6.3: Influence of the sample size (number of sample pairs) on the accuracy of ranking methods.

## 6.2 Severity score for Influenza A virus

To further assess the proposed approach, we applied it to learning the severity of H3N2 influenza symptoms.

The utilized dataset contains temporally collected gene expression measurements of human subjects infected with H3N2 virus (Zaas et al., 2009). The samples were collected on multiple occasions (approximately every eight hours) during the period of one week after the virus was inoculated in subjects. Concurrently, the severity of their symptoms was tracked (approximately twice a day) and clinically assessed using the modified Jackson score (Jackson et al., 1958).

When measurement time points were not perfectly aligned with severity score estimates, we associated the temporally nearest estimate with the gene expression vector. Having high dimensionality of the measurements ($12,032$ genes), temporally collected samples and associated severity score estimates, this dataset was suited for testing the proposed severity score learning framework. In addition to direct assessments of severity scores, which could be used for regression, the data samples are also accompanied with class labels "symptomatic" and "asymptomatic" (Woods et al., 2013), based on the values of modified Jackson scores.

Our comparison pairs generation process follows the guidelines proposed in (Dyagilev and Saria, 2015a). Ideally, an expert would be presented with example pairs and would assess which one appears more intense (w.r.t. a property of interest), based on visual inspection, clinical report or arbitrary convenient source. The alternative is to use an existing scoring system to generate comparison pairs, and for this application we utilized the Jackson score. We generated a third label type by extracting all possible pairs of samples where the first component is associated with a score that is substantially larger than the second. In our experiments, the "substantial" is defined by setting a threshold to 5 for training and 1 for testing.

On the described dataset consisting of 267 samples (17 subjects with about 16 temporal samples each) we have compared the predictive performance of the following four methods:

Table 6.2: Performance on H3N2 influenza gene expression dataset as measured by the fraction of correctly ordered pairs (accuracy)

| approach | SLDSS | DSSL | LASSO | $L_1$ Logistic Regression |
|---|---|---|---|---|
| Accuracy | 0.8097 | 0.7689 | 0.9490 | 0.7815 |

1. Sparse Learning of Disease Severity Score (SLDSS) from comparison pairs
2. "Dense" Disease Severity Score Learning (DSSL) from comparison pairs
3. LASSO regression on direct values of severity scores
4. $L_1$-regularized Logistic Regression fitted on binary classification labels of symptom severity

All enumerated methods result in a vector of feature weights that can be multiplied with the vector of measured features, and summed up to obtain the estimated value of a scoring function. Except for the DSSL which results in a dense vector of weights, all other approaches typically only have a small number of nonzero weights, while all others are exactly equal to zero.

We compared the mentioned methods in a 10-fold cross-validation procedure (where all samples belonging to one subject are either all in training, or all in testing folds) and the results are shown in Table 6.2.

In conducted experiments, the non-sparse method (DSSL) has the lowest accuracy, which provides evidence that sparse approaches were beneficial. LASSO was the most accurate, due to its direct access to the ground truth values (of the underlying scores), while other methods only had access to partial information. The Logistic Regression only had information if the score was larger than a certain threshold, while the DSSL and SLDSS only knew, for a list of pairs, which element in a given pair had a higher score. This, on the other hand, limits the application of LASSO to cases where scoring function already exists, thus reducing the necessity for learning

it from the data. Among the approaches which learn from indirect information about underlying values of scores (comparison pairs and severity classes) our SLDSS is the most accurate.

### *6.2.1 Robustness of selected features*

We are also interested in prospective use of the SLDSS approach for feature selection, that is for discovering the most relevant variables for the condition. Therefore, we have performed additional analysis regarding the robustness (stability) of the selected features.

Robustness of selected features is a very important aspect of the feature selection algorithms that was relatively neglected up until recently (Saeys et al., 2007). Various fields aim at finding the right subset of variables that would allow reliable prediction, and the more there are candidates to search from, the harder it is to find the right subset. Feature selection methods play a crucial role there, but when the dimensionality of data is much higher than the number of samples, the expectation of consistently finding high-quality solution decreases (Sima and Dougherty, 2006). On the other side, $L_1$ regularized models have far fewer requirement for sample size as compared to rotation invariant models ($L_2$ regularized models, Support Vector Machines, Artificial Neural Networks and DSSL, whose sample complexity grows at least linearly in the number of irrelevant features), as their sample size requirement grows logarithmically in the dimension of (irrelevant) features (Ng, 2004), so they are an attractive tool for such tasks.

Robustness is a metric that quantifies how different training sets affect the affinity of the algorithm towards the particular features and there are different measures proposed (Kalousis et al., 2007). Essentially, any similarity metric, which has higher value when the two compared vectors are more similar, might be used for assessing the feature selection stability. Here we used the common three:

1. Pearson coefficient (eq. 6.2), which measures the correlation between the weight vectors $w$ and $w'$ learned on different data (sub)sets, and tells magnitude stability of the weights. In the case when the weight vector is used as a linear function, it also tells how stable the learned function is.

$$C_P(w, w') = \frac{\sum_i (w_i - \mu_w)(w'_i - \mu_{w'})}{\sqrt{\sum_i (w_i - \mu_w)^2 \sum_i (w'_i - \mu_{w'})^2}} \tag{6.2}$$

2. Spearman rho metric (eq. 6.3), which measures how well the orders (ranks) $r$ and $r'$ of weights' $w$ and $w'$ magnitudes are preserved between different training sets. It is important, for example, in the dense methods where features are selected as some top number of features according to the magnitude of weights.

$$C_S(r, r') = \frac{\sum_i (r_i - \mu_r)(r'_i - \mu_{r'})}{\sqrt{\sum_i (r_i - \mu_r)^2 \sum_i (r'_i - \mu_{r'})^2}} \tag{6.3}$$

3. Jaccard index (eq. 6.4), which measures the overlap between two discrete sets $s$ and $s'$ of nonzero features in $w$ and $w'$, normalized with their union ($|.|$ is cardinality operator). Jaccard index is the most relevant measure (out of the three mentioned) regarding the stability of selected features, as studied frameworks select features in the form of a discrete set of nonzero features.

$$C_J(s, s') = \frac{|s \cap s'|}{|s \cup s'|} = \frac{|s \cap s'|}{|s| + |s'| - |s \cap s'|} \tag{6.4}$$

All four severity score learning methods are assessed for consistency/robustness based on each of the three stability measures (eqs. 6.2-6.4), through a 10-fold cross-validation procedure on H3N2 data. The sparsity level was tuned with free parameters (for sparse methods) such as to produce the average number (over ten folds)

Table 6.3: Stability of selected feature subsets summarized as an average pairwise similarity over ten training folds

| Measure | SLDSS | DSSL | LASSO | $L_1$ Logistic Regression |
|---|---|---|---|---|
| Pearson coefficient | **0.8656** | 0.7402 | 0.7362 | 0.5562 |
| Spearman rank | **0.8163** | 0.7204 | 0.5162 | 0.3988 |
| Jaccard index | **0.6916** | 0.2946 | 0.3595 | 0.2474 |

of nonzero features of about 100 out of $12,032$ possible (SLDSS $97.1 \pm 16.7$; LASSO $99.9 \pm 8.8$; $L_1$LogReg $101.7 \pm 22.7$), with results presented in Appendix B and summarized in Table 6.3. The dense method, DSSL, was compared to others, according to Jaccard index, by taking only the top 100 features according to the largest magnitudes in each of the folds separately. The results show that here proposed SLDSS method is the most stable one according to each of the three measures. This means that it learns the most stable severity score function (according to Pearson correlation), as well as the most stable set of nonzero features (according to Jaccard index). This evidence is suggesting that SLDSS is finding the most reliable signal in the data, out of all the tested approaches. Nevertheless, there are no guarantees that the selected set of features is free of false positives, as previously it was theoretically concluded that LASSO-like approaches select a super-set of the true features (Bühlmann and Van De Geer, 2011).

### *6.2.2 Gene ontology over-representation analysis*

To further check the appropriateness of SLDSS method as a biomarker discovery tool, we performed Gene Ontology Over-Representation Analysis to assess the relevance of a set of features extracted from the influenza dataset. In the robustness analysis section, we found that more than two thirds (0.6916) of the nonzero features are, on average, shared between the different folds of data. In fact, 50 genes were nonzero in all of the folds, so we took that set of genes and submitted it for over-representation analysis in the PANTHER (Mi et al., 2016) online tool.

Table 6.4: Genes selected by the Sparse Disease Severity Score Learning method, listed in alphabetical order

| Gene Symbols | | | | |
|---|---|---|---|---|
| AIM2 | CXCL10 | IFIH1 | MS4A4A | S100A12 |
| ALDH1A1 | EIF2AK2 | IFIT1 | MX1 | SERPING1 |
| ATF3 | EPB41L3 | IFIT2 | MYOF | SIGLEC1 |
| BLVRA | ETV7 | IFIT3 | OAS1 | STAT1 |
| C3AR1 | GBP1 | IL18RAP | OAS2 | TFEC |
| CASP5 | HERC5 | ISG15 | OAS3 | TLR7 |
| CASP7 | IFI35 | LAMP3 | OASL | TNFSF10 |
| CCL2 | IFI44 | LAP3 | RIN2 | TYMP |
| CCL8 | IFI44L | LILRA5 | RSAD2 | XAF1 |
| CDKN1C | IFI6 | MAFB | RTP4 | ZBP1 |

We analyzed the list of 50 selected genes given in Table 6.4, against all the $12,032$ genes in the dataset. Some of the $12,032$ genes were duplicates, and some symbols were not recognized by the database (Annotation Version and Release Date: GO Ontology database, Released 2016-03-25) resulting in the comparison of the 50 selected genes against the reference list of $10,792$ genes using the PANTHER Overrepresentation Test (release 20160321) with Bonferroni correction. Bonferroni correction (Haynes, 2013) is a simple and common method for multiple testing correction of significance value indicators. It is well acknowledged that it might be substantially conservative, especially when multiple tests are not independent. In multiple gene ontology process testing, it might be extremely conservative because descendants of a process are completely dependent on their parents. Nevertheless, even after overly conservative adjustments, a number of processes are found statistically significantly overrepresented with the cutoff value of 0.05 for p-value. Significantly overrepresented GO biological processes (listed in a Table 6.5) are related almost exclusively to immune response and a reaction of the host body to the virus. This is consistent with the fact that the dataset is about the response to viral infection, suggesting that the discovered set of features is indeed relevant for the studied process.

Table 6.5: PANTHER overrepresentation analysis results.
no. - number of associated genes;
exp. - expected number of genes by chance;
fold - number of times enriched

| GO biological process | GOID | no. | exp. | fold | P-value |
|---|---|---|---|---|---|
| defense response to virus | (GO:0051607) | 18 | .62 | 29.21 | 4.00E-18 |
| response to virus | (GO:0009615) | 20 | .96 | 20.75 | 1.01E-17 |
| type I interferon signaling pathway | (GO:0060337) | 14 | .28 | 49.54 | 1.97E-16 |
| cellular response to type I interferon | (GO:0071357) | 14 | .28 | 49.54 | 1.97E-16 |
| response to type I interferon | (GO:0034340) | 14 | .29 | 47.96 | 3.08E-16 |
| immune response | (GO:0006955) | 32 | 5.42 | 5.91 | 2.37E-15 |
| immune system process | (GO:0002376) | 35 | 8.20 | 4.27 | 4.07E-13 |
| innate immune response | (GO:0045087) | 26 | 3.79 | 6.87 | 1.07E-12 |
| defense response | (GO:0006952) | 31 | 6.12 | 5.07 | 1.07E-12 |
| defense response to other organism | (GO:0098542) | 19 | 1.73 | 10.97 | 1.46E-11 |
| immune effector process | (GO:0002252) | 19 | 1.75 | 10.85 | 1.76E-11 |
| cytokine-mediated signaling pathway | (GO:0019221) | 20 | 2.13 | 9.40 | 3.84E-11 |
| cellular response to cytokine stimulus | (GO:0071345) | 21 | 2.72 | 7.72 | 3.01E-10 |
| response to cytokine | (GO:0034097) | 22 | 3.16 | 6.96 | 4.80E-10 |
| response to other organism | (GO:0051707) | 22 | 3.23 | 6.81 | 7.44E-10 |
| response to external biotic stimulus | (GO:0043207) | 22 | 3.23 | 6.81 | 7.44E-10 |
| response to biotic stimulus | (GO:0009607) | 22 | 3.31 | 6.65 | 1.21E-09 |
| negative regulation of viral genome replication | (GO:0045071) | 8 | .19 | 41.11 | 1.83E-07 |
| negative regulation of viral process | (GO:0048525) | 9 | .36 | 24.90 | 7.74E-07 |
| regulation of viral genome replication | (GO:0045069) | 8 | .30 | 26.56 | 5.56E-06 |

Table 6.6: PANTHER overrepresentation analysis results CONTINUED.
no. - number of associated genes;
exp. - expected number of genes by chance;
fold - number of times enriched

| GO biological process | GOID | no. | exp. | fold | P-value |
|---|---|---|---|---|---|
| negative regulation of viral life cycle | (GO:1903901) | 8 | .35 | 23.02 | 1.69E-05 |
| response to stress | (GO:0006950) | 34 | 14.20 | 2.40 | 5.49E-05 |
| negative regulation of multi-organism process | (GO:0043901) | 9 | .60 | 15.06 | 6.01E-05 |
| response to external stimulus | (GO:0009605) | 26 | 8.48 | 3.06 | 1.19E-04 |
| cellular response to interferon-gamma | (GO:0071346) | 8 | .50 | 16.14 | 2.59E-04 |
| regulation of viral process | (GO:0050792) | 9 | .79 | 11.43 | 6.26E-04 |
| response to interferon-gamma | (GO:0034341) | 8 | .57 | 13.93 | 7.95E-04 |
| regulation of symbiosis | (GO:0043903) | 9 | .88 | 10.17 | 1.66E-03 |
| regulation of viral life cycle | (GO:1903900) | 8 | .74 | 10.86 | 5.14E-03 |
| interferon-gamma-mediated signaling pathway | (GO:0060333) | 6 | .31 | 19.33 | 5.39E-03 |
| response to stimulus | (GO:0050896) | 43 | 26.53 | 1.62 | 6.68E-03 |
| regulation of defense response | (GO:0031347) | 14 | 3.03 | 4.61 | 8.01E-03 |
| regulation of cytokine production | (GO:0001817) | 12 | 2.19 | 5.48 | 9.59E-03 |
| cellular response to organic substance | (GO:0071310) | 23 | 8.36 | 2.75 | 1.01E-02 |
| regulation of multi-organism process | (GO:0043900) | 11 | 1.81 | 6.07 | 1.07E-02 |
| response to interferon-alpha | (GO:0035455) | 4 | .09 | 45.44 | 1.55E-02 |
| cellular response to chemical stimulus | (GO:0070887) | 25 | 10.06 | 2.49 | 1.75E-02 |
| cell surface receptor signaling pathway | (GO:0007166) | 23 | 9.04 | 2.54 | 4.07E-02 |
| multi-organism process | (GO:0051704) | 22 | 8.40 | 2.62 | 4.59E-02 |

## 6.3 Tolerance to pathogens in frogs

We applied the SLDSS feature selection approach on the frog tadpole gene expression datasets to identify genes that best explain tolerance behavior, based on labels reported in Table 6.7. Each of the listed pathogen has six samples, that is three tadpoles specimens with two temporally collected samples, which makes a total of 48 samples. Gene expression measurements consists of $8,726$ probes corresponding to genes in frogs.

Briefly, tolerant behavior is deemed as one where an organism preserves its fitness, despite being infected with high level of pathogens. It should be differentiated from the resistant phenotype, which achieves high fitness by actively reducing the levels of pathogen, by acting with his immune system defense mechanisms. Sensitive (or susceptible) phenotype is the case where organism's fitness is deteriorating, that is it cannot withstand pathogen infection, nether through tolerance nor resistance mechanisms. Pairs of ranked examples for learning were compiled by letting samples with the "Tolerant" label be paired with samples with the "Sensitive" label, where "Tolerant" samples were always ranked higher than the "Sensitive" ones. The Tolerant group was represented by A. baumannii and K. pneumonia (12 samples), and sensitive group consists of A. hidrophyla and P. aeruginosa (12 samples). Under this setup, we applied the SLDSS approach on the gene expression dataset with only ENTREZ Maglott et al. (2005) annotated probes and on the complete set of probes, to find the tolerance-related genes. Analysis with only annotated probes is conducted since it allows utilization of prior knowledge, and analysis of a complete set of probes because it provides new insight into previously poorly explored markers.

A set of 35 genes obtained from the ENTREZ annotated subset of the dataset are listed in Table 6.8, along with the annotation details. One of the interesting selected features is trnsferrin, which plays important role in iron metabolism on a

Table 6.7: Phenotypes labeled according to reaction of frog tadpoles to different pathogens.

| Sensitive | Resistant | Tolerant |
|---|---|---|
| P. Aeruginosa | S. pneumoniae | A. baumannii |
| | | K. pneumonia |
| A. hydrophila | LPS - 5000 ug/mL | E. coli F11 |
| | | S. aureus |

cellular level (Hower et al., 2009), which is known to be important factor in infections (Drakesmith and Prentice, 2008).

In the obtained computational models, the higher the magnitude of the coefficient, the larger the effect it has in discriminating two groups, and it can be regarded as gene's higher importance for the tolerance phenotype. It should also be noted that this is not just the top 35 probes obtained by truncating the features with lower coefficients. In fact, all features coefficients, other than for selected 35, are equal to zero. The number of non-zero features is tuned by a penalization parameter, and we have empirically chosen the one that gives a few dozen features, as that is expected number of features with good importance-generalization trade-off. Allowing all features to have nonzero contribution would lead to the problem of overfitting, and the other extreme would assess the correlation of only one biomarker with the target pattern. As tolerance is likely arising from the interaction between many genes, choosing a handful of features (35 in this case) allowed us to investigate complex behavior, while avoiding the pitfall of statistical over-fit by using too many features. In order to test the predictive performance of learned scoring functions, we evaluated the approach on unused samples. The unused samples consist of two groups, Tolerant being the E. coli and S. aureus (12 samples), and Resistant being the S. pneumoniae and LPS (12 samples). Although the originally trained problem was a bit different, Tolerance vs Sensitivity, and not Tolerance vs Resistance, it still makes sense to test in this setup, as Tolerant samples are expected to still have higher tolerance scores

Table 6.8: List of 35 genes with ENTEZ ID selected by the SLDSS approach from the frog data.

| Coefficient | ENTREZ_GENE_ID | GeneSymbol | Description |
|---|---|---|---|
| -0.262989006 | 399287 | 'Prkcl' | 'protein kinase C lambda/iota' |
| 0.097428008 | 446405 | 'MGC83803' | 'MGC83803 protein' |
| 0.096942664 | 446710 | 'MGC83955' | 'MGC83955 protein' |
| 0.091749873 | 444230 | 'MGC80788' | 'MGC80788 protein' |
| -0.088071452 | 378665 | 'sox7-A' | 'SRY protein' |
| -0.060269613 | 443754 | 'MGC81060' | 'MGC81060 protein' |
| 0.058430007 | 443843 | 'MGC83146' | 'MGC83146 protein' |
| 0.046096018 | 399249 | 'Copz1' | 'coatomer protein complex zeta 1' |
| -0.045700082 | 447241 | 'MGC84754' | 'MGC84754 protein' |
| -0.043493382 | 1E+08 | 'rps6ka4' | 'ribosomal protein S6 kinase |
| -0.036352345 | 446777 | 'MGC80410' | 'MGC80410 protein' |
| 0.023231029 | 379523 | 'shsia-3' | 'shisa-3' |
| 0.019199372 | 447309 | 'MGC81740' | 'MGC81740 protein' |
| 0.01854616 | 447271 | 'MGC86386' | 'MGC86386 protein' |
| 0.014342954 | 444095 | 'MGC83624' | 'MGC83624 protein' |
| 0.011724236 | 431901 | 'MGC83731' | 'UPF0566 protein' |
| -0.009496875 | 380498 | 'pdcd2' | 'programmed cell death 2' |
| -0.007811129 | 447201 | 'MGC80305' | 'MGC80305 protein' |
| -0.000383502 | 444105 | 'MGC80424' | 'MGC80424 protein' |

Table 6.9: List of 35 genes with ENTEZ ID selected by the SLDSS approach from the frog data CONTINUED.

| Coefficient | ENTREZ_GENE_ID | GeneSymbol | Description |
|---|---|---|---|
| 9.61E-05 | 443641 | 'cdc5l' | 'cell division cycle 5-like' |
| -9.58E-05 | 378638 | 'hoxa13-A' | 'homeo box A13' |
| 9.44E-05 | 444094 | 'MGC83623' | 'MGC83623 protein' |
| -7.55E-05 | 444120 | 'MGC80493' | 'MGC80493 protein' |
| -6.80E-05 | 779435 | 'b4galt6' | 'galactosyl- -transferase 6' |
| 6.51E-05 | 446532 | 'MGC80279' | 'MGC80279 protein' |
| 4.77E-05 | 1E+08 | 'pi4k2a' | 'Phosphatidylinositol 4-kinase |
| -4.65E-05 | 779259 | 'MGC154458' | 'sideroflexin 1' |
| 4.44E-05 | 379502 | 'MGC64251' | 'transferrin' |
| 4.00E-05 | 733312 | 'des' | 'desmin' |
| -3.71E-05 | 734196 | 'fancd2' | 'Fanconi anemia group D2' |
| -3.49E-05 | 446239 | 'CHML' | 'CHML protein' |
| -3.30E-05 | 414681 | 'cbr4' | 'Carbonyl reductase 4' |
| 3.05E-05 | 379859 | 'odc1' | 'ornithine decarboxylase 1' |
| -1.17E-05 | 779277 | 'rcn1' | 'reticulocalbin 1' |
| -4.93E-06 | 399108 | 'Coro1c' | 'coronin homolog' |

in comparison with the Resistant ones. The model learned on all probes is applied to test samples and estimates of their tolerance scores were obtained. Estimates of the tolerance scores are presented in Figure 6.4. Scores are given on the $y$-axes, and represent dimensionless numeric, while the $x$-axis just gives the sample ID and does not had to have any particular order. Even though it looks like the predicted scores are mixed, statistical tests show that there are significant differences in the

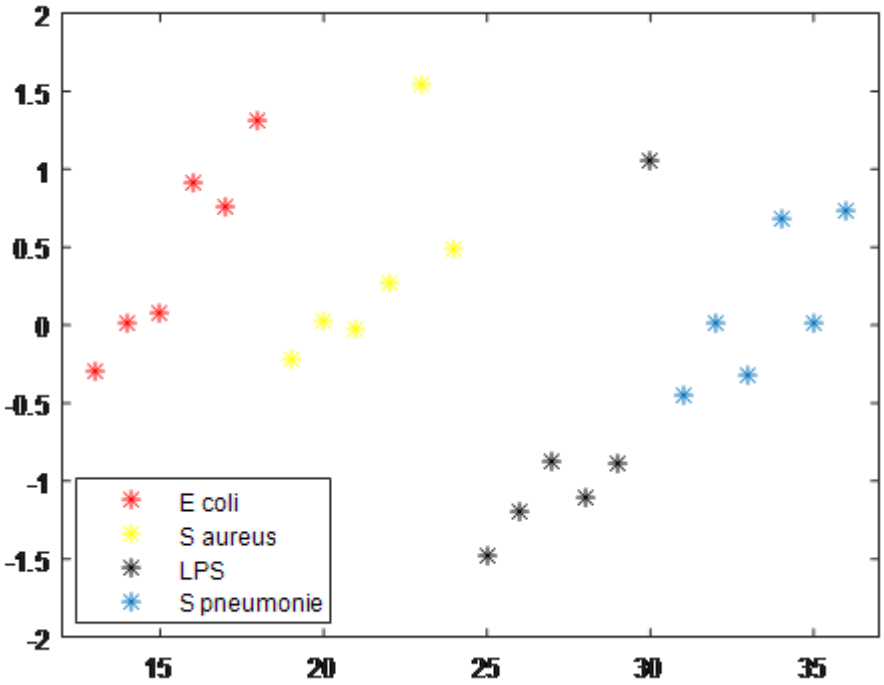


FIGURE 6.4: Predicted values of tolerance scores for testing samples consisting of tolerant (E. coli and S. aureus) and resistant phenotypes (S. pneumoniae and LPS).

mean values of the scores for the two phenotypic groups on the population level. We performed a two-sided (unpaired) t-test and results suggest that the null hypothesis that two groups are no different can be rejected on the significance level of 0.05 (p-value = 0.032). Although not perfectly discriminative, this result shows that the approach learned some patterns in which tolerance samples differs from the other phenotypes (sensitive and resistant).

# CHAPTER 7

## CONCLUSION

Quantifying a properties of interest is important task in many domains, as that is necessary for informing decisions and making appropriate actions. For example, whether to treat the patient with some drug, or whether to hire the candidate for some job position. Unfortunately, it is not an easy task, as often those very properties are latent and difficult to assess. However, even when direct assessment is not an option, many times it is feasible to obtain relative assessment of two examples. Like noticing that some patient is in more severe condition than the other one, or that some patient today appears more healthy than he was yesterday.

There exist methods that can effectively use such "pairwise comparison" information to build predictive models, which can subsequently be used for assessing the property of interest from other observable features. We adopted one such framework named ranking SVM and extended it for use in some special cases. Applications in bio-medical datasets typically have some specific challenges. First, and the major one, is the limited amount of data examples, due to an expensive measuring technology, and/or infrequency of conditions of interest. Such limited number of examples makes both identification of patterns/models and their validation less useful and re-

liable. Repeated samples from the same subject are collected on multiple occasions over time, which breaks IID sample assumption and introduces dependency structure that can be exploited, but needs to be taken into account more appropriately. Also, feature vectors are highdimensional, and typically of much higher cardinality than the number of samples, making models less useful and their learning less efficient.

We proposed a method that jointly learns multiple scoring functions from a set of ranked examples. These efforts are motivated by applications in which there are multiple related tasks, with a limited amount of data for each task. Related tasks commonly share underlying regularities which could be learned more accurately by modeling all tasks together. The multitask approach utilizes composite regularization consisting of the trace norm and row-wise grouped Lasso penalty, to impose structural regularity among the model parameters of different tasks. We also proposed an optimization algorithm, based on the alternate minimization and proximal gradient techniques, for solving such convex multitask ranking based scoring functions learning objective. We also presented an approach to the problem of learning scoring functions in presence of irrelevant or high dimensional measurements. We build on top of existing efforts by simultaneously performing feature selections that are most relevant for the score learning. Both developed frameworks were thoroughly evaluated on synthetic and real examples, in order to assess their characteristics.

Multitask framework empirical evaluations in one synthetic and two real-world datasets suggest the benefits of utilizing the multitask approach for learning related ranking based scoring functions. According to the results, the model with only $L_*$ performs worse than $L_{1,2}$, because sparsity in features seems to be the more dominant pattern in the data relative to the low-rank component. However, utilizing both $L_*$ and $L_{1,2}$ in the same model turned out to be most beneficial for studied applications. We also assessed multiple approaches to learning the severity scores in high-dimensional applications. Our results point to the utility and maybe even neces-

sity of reducing the dimensionality of the problem through sparse learning techniques, with the use of $L_1$ norm regularization. Combination of the advantages of existing solutions turned out to be beneficial for the predictive performance, as measured by accuracy. The robustness of the learned scoring function, on human influenza virus application, as well as features selected by our approach compares very favorably against the alternatives. Conducted gene ontology over-representation analysis supports the relevance of the genes identified by the SLDSS approach. Additional studies are possible to further characterize selected genes and the processes they are involved in, in order to provide further insight into causal relations underlining the influenza infection. For example, "transferrin" gene selected in the frog tolerance to pathogens application, and its role in iron metabolism during the pathogen infection, will be focus of the further studies, with aim to gain some knowledge on how tolerant behavior arises. These are all mounting evidence that proposed approach could be used as a discovery tool for both scoring functions and related informative variables, which could further motivate novel hypotheses.

Also, there are some limitations and drawbacks to the proposed models and algorithms. First the models are linear function of the observed features, and therefore cannot capture well the nonlinear effects that might affect relations between the features and target. The remedy for that might be extension of the proposed frameworks to use the kernel trick to allow for more richer representational ability (Schölkopf et al., 1999). Second, the proposed proximal gradient algorithm with alternating minimization for optimization of the multitask objective proved valuable for applications with low to moderate dimensionality of the feature space. However, as the contemporary applications have ever increasing number of measured variables, more efficient optimization approaches with better scalability are required. One potential way to accelerate the proximal gradient algorithm is to adopt the approach proposed in (Toh and Yun, 2010). The high-dimensional method on the other side

is appropriate for learning scoring functions from high dimensional cases. However, it would probably have problem in applications where the number of cases is also large, since in such applications a quadratic number of comparisons in the number of samples can be a challenge. That difficulty might be alleviated with appropriate sampling of the training samples, similar to techniques proposed for Gaussian Processes (Lawrence et al., 2009).

# BIBLIOGRAPHY


Anderson, R. (2007), *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*, Oxford University Press.

Ando, R. K. and Zhang, T. (2005), "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, 6, 1817–1853.

Argyriou, A., Evgeniou, T., and Pontil, M. (2008), "Convex multi-task feature learning," *Machine Learning*, 73, 243–272.

Ashtawy, H. M. and Mahapatra, N. R. (2015), "Machine-learning scoring functions for identifying native poses of ligands docked to known and novel proteins," *BMC bioinformatics*, 16, S3.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. (2011), "Convex optimization with sparsity-inducing norms," *Optimization for Machine Learning*, 5.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. (2012), "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, 4, 1–106.

Bai, J., Zhou, K., Xue, G., Zha, H., Sun, G., Tseng, B., Zheng, Z., and Chang, Y. (2009), "Multi-task learning for learning to rank in web search," in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1549–1552, ACM.

Beck, A. and Teboulle, M. (2009), "Gradient-based algorithms with applications to signal recovery," *Convex optimization in signal processing and communications*, pp. 42–88.

Bertsekas, D. P. and Tseng, P. (1994), "Partial proximal minimization algorithms for convex pprogramming," *SIAM Journal on Optimization*, 4, 551–572.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. (2003), "Dimensionality reduction via sparse support vector machines," *Journal of Machine Learning Research*, 3, 1229–1243.

Boyd, S. (2011), "Alternating direction method of multipliers," in *Talk at NIPS Workshop on Optimization and Machine Learning*.

Boyd, S. and Vandenberghe, L. (2004), *Convex optimization*, Cambridge university press.

Bühlmann, P. and Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010), "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, 20, 1956–1982.

Cao, X. H., Stojkovic, I., and Obradovic, Z. (2014), "Predicting sepsis severity from limited temporal observations," in *International Conference on Discovery Science*, pp. 37–48, Springer.

Cao, X. H., Stojkovic, I., and Obradovic, Z. (2016), "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BMC bioinformatics*, 17, 1–10.

Chen, J., Tang, L., Liu, J., and Ye, J. (2009), "A convex formulation for learning shared structures from multiple tasks," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 137–144, ACM.

Chen, J., Zhou, J., and Ye, J. (2011), "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 42–50, ACM.

Colburn, W., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., Spilker, B. A., Woodcock, J., et al. (2001), "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Biomarkers Definitions Working Group," *Clinical Pharmacol & Therapeutics*, 69, 89–95.

Ćosić, A., Katić, D., and Stojković, I. (2013), "An Approach to Localization of Mobile Robots Based on Extended Kalman Filter," in *Proceedings of the 57th ETRAN Conference, Society for Electronics, Telecommunications, Computers, Automatic Control and Nuclear Engineering*, pp. 1–4.

De Mol, C., De Vito, E., and Rosasco, L. (2009), "Elastic-net regularization in learning theory," *Journal of Complexity*, 25, 201–230.

Drakesmith, H. and Prentice, A. (2008), "Viral infection and iron metabolism," *Nature Reviews Microbiology*, 6, 541–552.

Dyagilev, K. and Saria, S. (2015a), "Learning (predictive) risk scores in the presence of censoring due to interventions," *Machine Learning*, pp. 1–26.

Dyagilev, K. and Saria, S. (2015b), "Learning severity score for sepsis: a novel approach based on clinical comparisons," in *AMIA Annual Symposium Proceedings*, pp. 1890–1898.

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005), "Learning multiple tasks with kernel methods," *Journal of Machine Learning Research*, 6, 615–637.

Ghalwash, M. F., Cao, X. H., Stojkovic, I., and Obradovic, Z. (2016), "Structured feature selection using coordinate descent optimization," *BMC Bioinformatics*, 17, 1–14.

Ghosh, D. and Chinnaiyan, A. M. (2005), "Classification and selection of biomarkers in genomic data using LASSO," *BioMed Research International*, 2005, 147–154.

Gligorijevic, D., Stojanovic, J., and Obradovic, Z. (2015), "Improving confidence while predicting trends in temporal disease networks," in *SIAM SDM 4th Workshop on Data Mining for Medicine and Healthcare*.

Gligorijevic, D., Stojanovic, J., and Obradovic, Z. (2016), "Uncertainty Propagation in Long-term Structured Regression on Evolving Networks," in *Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*.

Gong, P., Ye, J., and Zhang, C. (2012), "Robust multi-task feature learning," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 895–903, ACM.

Haynes, W. (2013), *Encyclopedia of Systems Biology*, chap. Bonferroni Correction, pp. 154–154, Springer New York, New York, NY.

Hower, V., Mendes, P., Torti, F. M., Laubenbacher, R., Akman, S., Shulaev, V., and Torti, S. V. (2009), "A general map of iron metabolism and tissue-specific subnetworks," *Molecular bioSystems*, 5, 422–443.

Jackson, G. G., Dowling, H. F., Spiesman, I. G., and Boand, A. V. (1958), "Transmission of the common cold to volunteers under controlled conditions: I. The common cold as a clinical entity," *AMA archives of internal medicine*, 101, 267–278.

Jacob, L., Vert, J.-p., and Bach, F. R. (2009), "Clustered multi-task learning: A convex formulation," in *Advances in neural information processing systems*, pp. 745–752.

Jalali, A., Sanghavi, S., Ruan, C., and Ravikumar, P. K. (2010), "A dirty model for multi-task learning," in *Advances in Neural Information Processing Systems*, pp. 964–972.

Joachims, T. (2002), "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, ACM.

Kalousis, A., Prados, J., and Hilario, M. (2007), "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and information systems*, 12, 95–116.

Kumar, S. and Hebert, M. (2004), "Discriminative Fields for Modeling Spatial Dependencies in Natural Images," in *Advances in Neural Information Processing Systems*, pp. 1531–1538.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001), "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *ICML*, pp. 282–289.

Lai, H., Pan, Y., Liu, C., Lin, L., and Wu, J. (2013), "Sparse learning-to-rank via an efficient primal-dual algorithm," *IEEE Transactions on Computers*, 62, 1221–1233.

Lawrence, N. D., Rattray, M., and Titsias, M. K. (2009), "Efficient sampling for Gaussian process inference using control variables," in *Advances in Neural Information Processing Systems*, pp. 1681–1688.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005), "Entrez Gene: gene-centered information at NCBI," *Nucleic acids research*, 33, D54–D58.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2016), "PANTHER version 10: expanded protein families and functions, and analysis tools," *Nucleic acids research*, 44, D336–D342.

Miloradović, B., Popić, S., and Stojković, I. (2013), "Development of Intelligent Behaviour Using Decision Making Methods Based on ANN," in *Proceedings of the 12th INFOTECH-Jahorina*, pp. 1060–1065.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., and Ecob, R. (1988), *School matters: The junior years*, Open Books.

Ng, A. Y. (2004), "Feature selection, L 1 vs. L 2 regularization, and rotational invariance," in *Proceedings of the twenty-first international conference on Machine learning*, p. 78, ACM.

Parikh, N. and Boyd, S. (2014), "Proximal Algorithms." *Foundations and Trends in optimization*, 1, 127–239.

Pavlovski, M., Zhou, F., Stojkovic, I., Kocarev, L., and Obradovic, Z. (2017), "Adaptive Skip-Train Structured Regression for Temporal Networks," in *Proceedings of the ECML/PKDD 17*.

Peng, F. and McCallum, A. (2006), "Information extraction from research papers using conditional random fields," *Information processing & management*, 42, 963–979.

Radosavljevic, V., Vucetic, S., and Obradovic, Z. (2010), "Continuous Conditional Random Fields for Regression in Remote Sensing." in *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010) Lisbon, Portugal*.

Rasmussen, C. E. and Williams, C. K. (2006), *Gaussian processes for machine learning*, vol. 1, MIT press Cambridge.

Rodić, A. and Stojković, I. (2012a), "Integrated Control of Quadrotor Flight Dynamics with Complementary Compensator of System Uncertainties," in *International Micro Air Vehicle Conference and Flight Competition IMAV 2012*, pp. 1–8.

Rodić, A. and Stojković, I. (2013), "Building of Open-Structure Wheel-Based Mobile Robotic Platform," *Interdisciplinary Mechatronics*, pp. 385–421.

Rodić, A., Mester, G., and Stojković, I. (2012), "Navigation and Control of Indoor Mobile Robot in Unknown Environments," in *Proceedings of the 56th ETRAN Conference, Society for Electronics, Telecommunications, Computers, Automatic Control and Nuclear Engineering*, pp. 1–4.

Rodić, A., Mester, G., and Stojković, I. (2013), "Qualitative evaluation of flight controller performances for autonomous quadrotors," in *Intelligent Systems: Models and Applications*, pp. 115–134, Springer.

Rodić, A. D. and Stojković, I. R. (2012b), "Dynamic Inversion Control of quadrotor with complementary Fuzzy logic compensator," in *11th Symposium on Neural Network Applications in Electrical Engineering (NEUREL)*, pp. 53–58, IEEE.

Saeys, Y., Inza, I., and Larrañaga, P. (2007), "A review of feature selection techniques in bioinformatics," *bioinformatics*, 23, 2507–2517.

Santolino, M. and Boucher, J.-P. (2009), "Modelling the disability severity score in motor insurance claims: an application to the Spanish case," *IREA–Working Papers, 2009, IR09/002*.

Sato, K. and Sakakibara, Y. (2005), "RNA secondary structural alignment with conditional random fields," *Bioinformatics*, 21, ii237–ii242.

Scheinberg, K., Ma, S., and Goldfarb, D. (2010), "Sparse inverse covariance selection via alternating linearization methods," in *Advances in neural information processing systems*, pp. 2101–2109.

Schmidt, M., Roux, N. L., and Bach, F. R. (2011), "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Advances in neural information processing systems*, pp. 1458–1466.

Schölkopf, B., Burges, C. J., and Smola, A. J. (1999), *Advances in kernel methods: support vector learning*, MIT press.

Shaobing, C. and Donoho, D. (1994), "Basis pursuit," in *28th Asilomar conf. Signals, Systems Computers*.

Sima, C. and Dougherty, E. R. (2006), "What should be expected from feature selection in small-sample settings," *Bioinformatics*, 22, 2430–2436.

Simms, E. L. (2000), "Defining tolerance as a norm of reaction," *Evolutionary Ecology*, 14, 563–570.

Sokolovska, N., Chevaleyre, Y., Clément, K., and Zucker, J.-D. (2017), "The fused lasso penalty for learning interpretable medical scoring systems," in *Neural Networks (IJCNN), 2017 International Joint Conference on*, pp. 4504–4511, IEEE.

Spasojević, S., Ilić, T. V., Stojković, I., Potkonjak, V., Rodić, A., and Santos-Victor, J. (2017), "Quantitative assessment of the arm/hand Movements in Parkinsons Disease Using a Wireless armband Device," *Frontiers in Neurology*, 8, 1–15.

Stojanovic, J., Jovanovic, M., Gligorijevic, D., and Obradovic, Z. (2015), "Semi-supervised learning for structured regression on partially observed attributed graphs," in *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM 2015) Vancouver, Canada*, SIAM.

Stojanovic, J., Gligorijevic, D., and Obradovic, Z. (2016), "Modeling Customer Engagement from Partial Observations," in *25th ACM International Conference on Information and Knowledge Management (CIKM 2016)*.

Stojković, I. and Katić, D. (2017), "Formation Control of Robotized Aerial Vehicles Based on Consensus-Based Algorithms," *FME Transactions*, 45, 559–564.

Stojkovic, I. and Obradovic, Z. (2017a), "Predicting Sepsis Biomarker Progression under Therapy," in *Proceedings of the 30th IEEE International Symposium on Computer-Based Medical Systems IEEE CBMS-17*, pp. 19–24.

Stojkovic, I. and Obradovic, Z. (2017b), "Sparse Learning of the Disease Severity Score for High-Dimensional Data," *Complexity*, 2017, 1–11.

Stojković, I., Rodić, A., and Stevanović, I. (2012), "Comparison of different flight control techniques for autonomous quadrotors," in *Proceedings of the 56th ETRAN Conference, Society for Electronics, Telecommunications, Computers, Automatic Control and Nuclear Engineering*, pp. 1–4.

Stojkovic, I., Jelisavcic, V., Milutinovic, V., and Obradovic, Z. (2016a), "Distance Based Modeling of Interactions in Structured Regression," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence IJCAI-16*, pp. 2032–2038.

Stojkovic, I., Ghalwash, M. F., Cao, X. H., and Obradovic, Z. (2016b), "Effectiveness of Multiple Blood-Cleansing Interventions in Sepsis, Characterized in Rats," *Scientific Reports*, 6, 1–11.

Stojkovic, I., Jelisavcic, V., Milutinovic, V., and Obradovic, Z. (2017a), "Fast sparse Gaussian Markov Random Fields learning based on Cholesky factorization," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence IJCAI-17*, pp. 2758–2764.

Stojkovic, I., Ghalwash, M., and Obradovic, Z. (2017b), "Ranking based multitask learning of scoring functions," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 1–12.

Sutton, C. and McCallum, A. (2006), "An introduction to conditional random fields for relational learning," *Introduction to statistical relational learning*, pp. 93–128.

Tappen, M. F., Liu, C., Adelson, E. H., and Freeman, W. T. (2007), "Learning gaussian conditional random fields for low-level vision," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR'07)*, pp. 1–8, IEEE.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Toh, K.-C. and Yun, S. (2010), "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, 6, 15.

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. (1996), "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure," *Intensive care medicine*, 22, 707–710.

Vujicic, T., Glass, J., Zhou, F., and Obradovic, Z. (2017), "Gaussian conditional random fields extended for directed graphs," *Machine Learning*, 106, 1271–1288.

Wang, L., Zhu, J., and Zou, H. (2006), "The doubly regularized support vector machine," *Statistica Sinica*, pp. 589–615.

Woods, C. W., McClain, M. T., Chen, M., Zaas, A. K., Nicholson, B. P., Varkey, J., Veldman, T., Kingsmore, S. F., Huang, Y., Lambkin-Williams, R., et al. (2013), "A

host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2," *PLoS One*, 8, e52198.

Wytock, M. and Kolter, Z. (2013), "Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1265–1273.

Yang, S., Shapiro, L., Cunningham, M., Speltz, M., Birgfeld, C., Atmosukarto, I., and Lee, S.-I. (2012), "Skull retrieval for craniosynostosis using sparse logistic regression models," in *Medical Content-Based Retrieval for Clinical Decision Support*, pp. 33–44, Springer.

Zaas, A. K., Chen, M., Varkey, J., Veldman, T., Hero, A. O., Lucas, J., Huang, Y., Turner, R., Gilbert, A., Lambkin-Williams, R., et al. (2009), "Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans," *Cell host & microbe*, 6, 207–217.

Zhou, J., Chen, J., and Ye, J. (2011), "Malsar: Multi-task learning via structural regularization," *Arizona State University*, 21.

Zhou, J., Liu, J., Narayan, V. A., and Ye, J. (2012), "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1095–1103, ACM.

Živanović, M. and Stojković, I. (2013), "Self-Organization in motion of a set of living individuals," *Self-Organiztion: Theories and Methods*, pp. 171–194.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

# Appendix A

## RESPIRATORY VIRAL INFECTION DATA

Used data is obtained from study (Zaas et al., 2009), and is publicly available on Gene Expression Omnibus repository under number GSE17156. For all subjects in each of the three datasets, symptoms were recorded twice a day and quantified by the modified Jackson Score (Jackson et al., 1958). Thereafter, subjects were classified based on the modified Jackson Score values into "symptomatic" and "asymptomatic" groups. In addition, viral load temporal measurements are available for 28 "symptomatic" subjects, given in Table 5.3. Gene expression measurements (for 12,023 genes) were collected temporally, starting at a baseline (24 hours prior to inoculation with virus) and measured at certain time points following the experimental procedure described in detail in (Zaas et al., 2009), making a total of 16, 14 and 21 time-point measurements for H3N2, HRV and RSV datasets, respectively. Subsequent Figures A.1-A.6 shows the viral shedding and symptom scores for subjects who developed clinically relevant symptoms from H3N2, HRV and RSV datasets, and are recreated using the info from supplementary material of (Zaas et al., 2009).

## A.1 Human Influenza Virus - H3N2

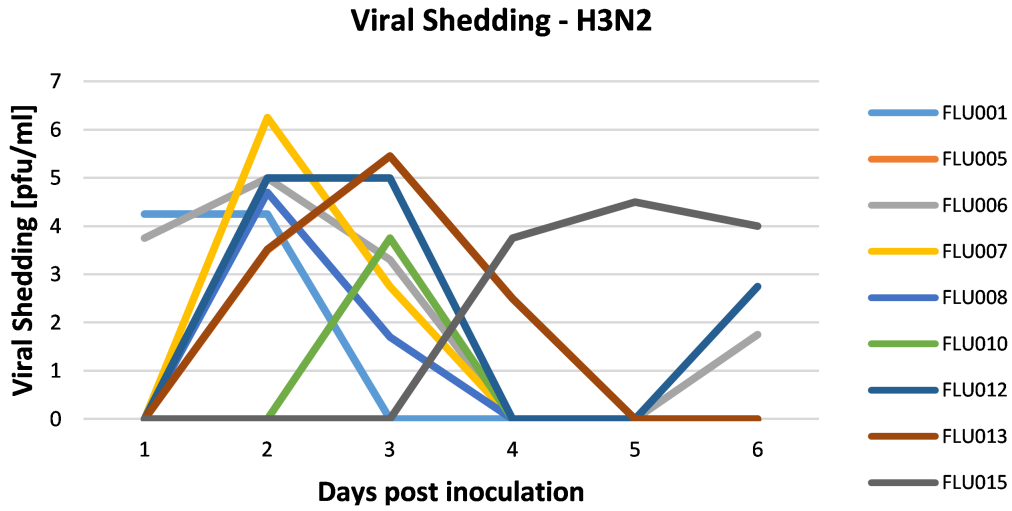


FIGURE A.1: H3N2 patients' viral load over the course of infection.

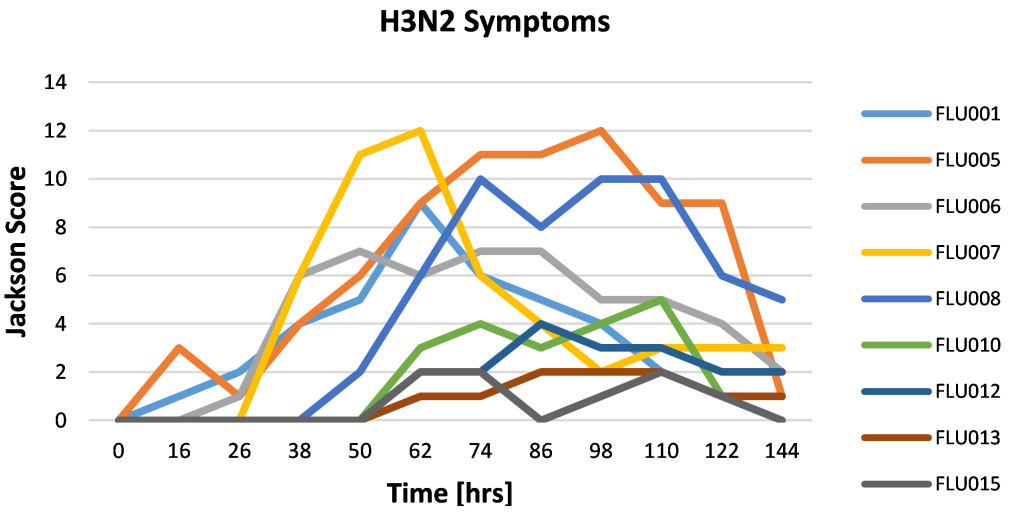


FIGURE A.2: H3N2 patients' symptoms severity over the course of infection.

## A.2 Human Rhino Virus - HRV

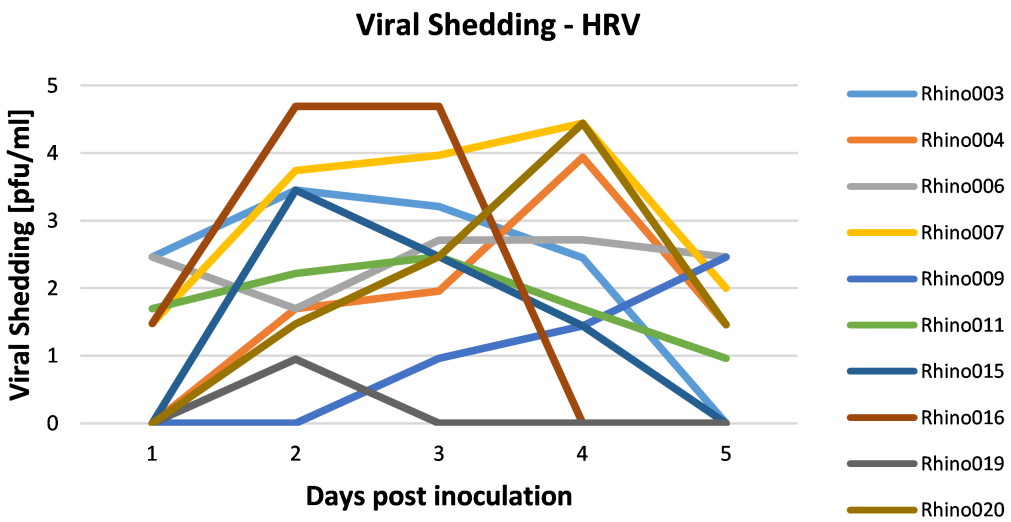


FIGURE A.3: HRV patients' viral load over the course of infection.

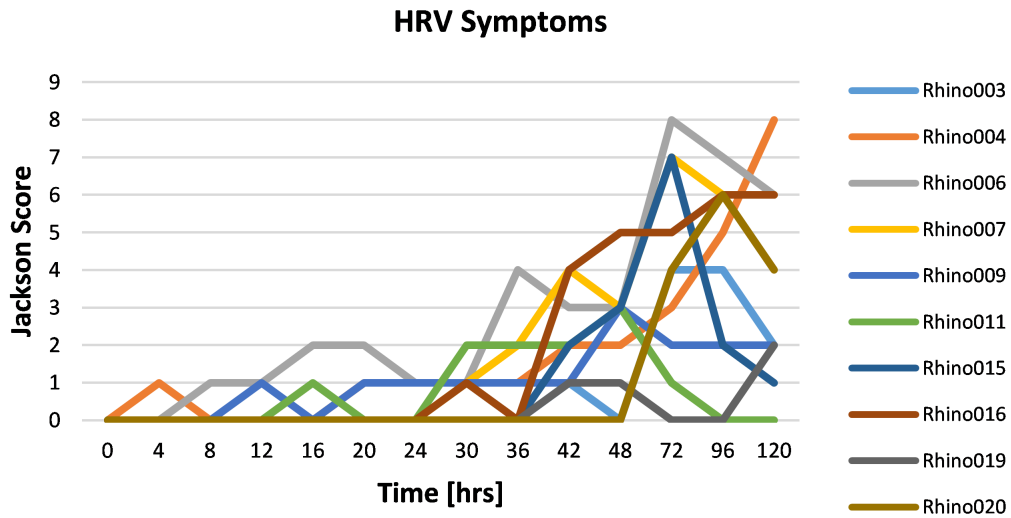


FIGURE A.4: HRV patients' symptoms severity over the course of infection.

## A.3 Respiratory Syncytial Virus - RSV

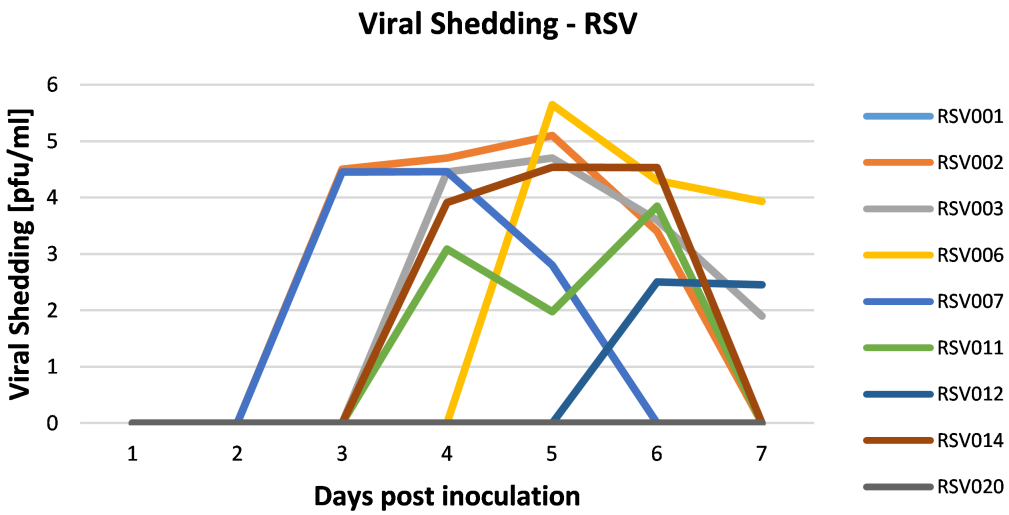


FIGURE A.5: RSV patients' viral load over the course of infection.

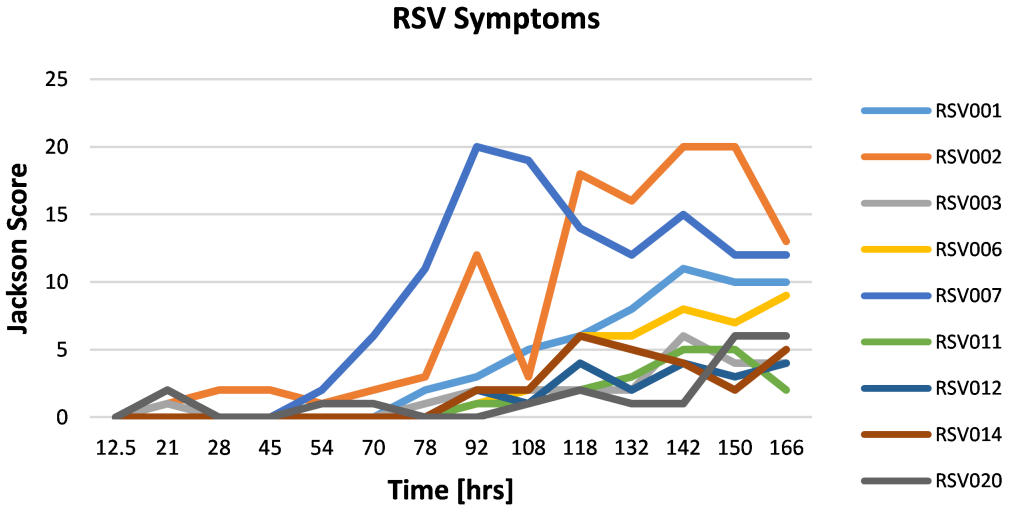


FIGURE A.6: RSV patients' symptoms severity over the course of infection.

# Appendix B

## FEATURE SELECTION STABILITY

Feature selection stability, or in other words robustness, is a metric that quantifies how sensitive are the selection algorithms to different training sets. That is, how likely it is, that they will select similar feature sets, among different training samples' sets. We performed training of four competitive algorithms (SLDSS, DSSL, LASSO and $L_1$ Log Reg) for scoring function learning in a 10-fold crossvalidation procedure, which resulted in 40 sets of selected features.

Robustness of such selected features is characterized in three ways: using Pearson correlation which measures how well are feature magnitudes preserved (Fig B.1); Spearman rank to track how the order of features' magnitudes are preserved (Fig B.2); and Jaccard Index to see how well the sets are overlapping (Fig B.3).

From Figures B.1-B.3, it can be observed that features are most similar within the 10 folds selected by the same algorithm (yellowish squares on the diagonal), and that SLDSS has the brightest square, which suggest largest correlation.

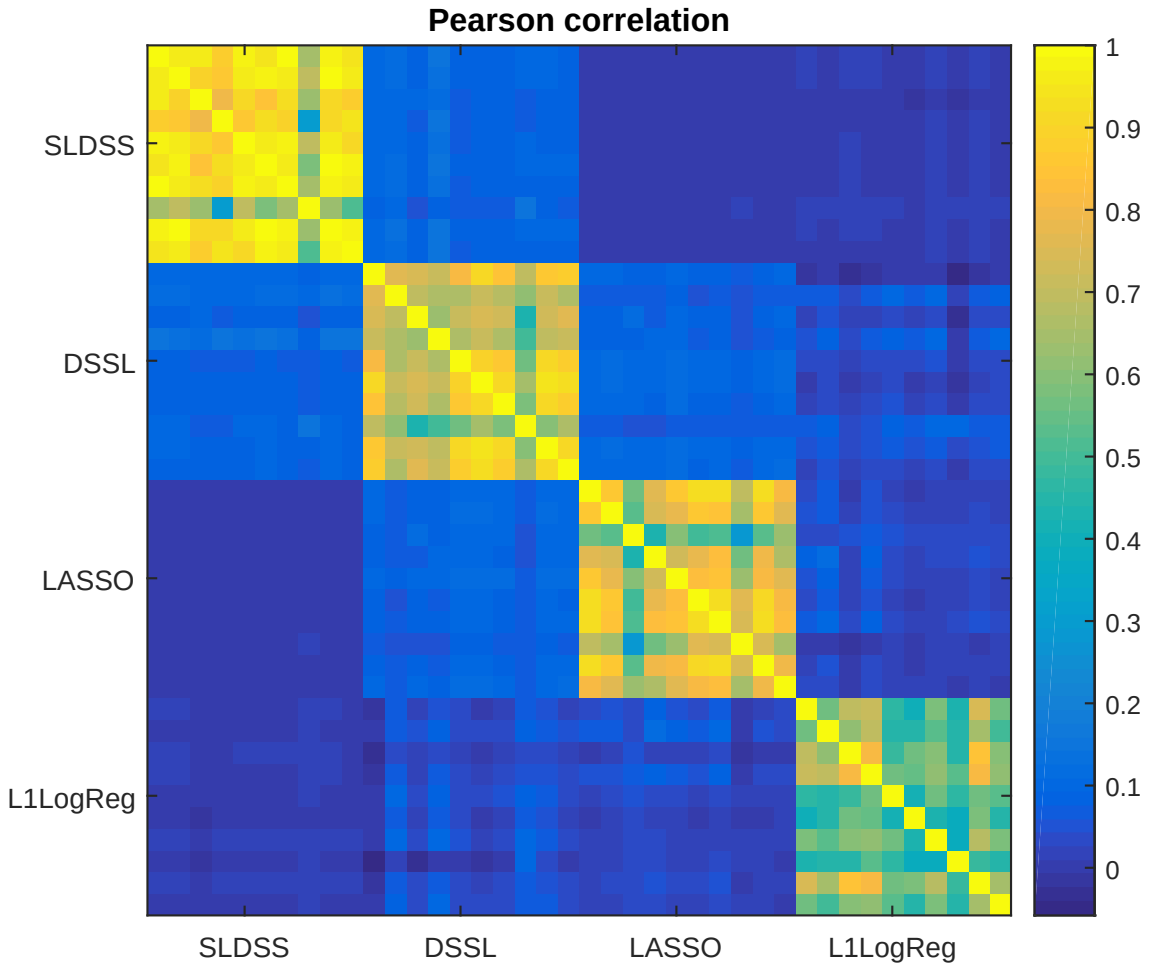


FIGURE B.1: Pearson similarity matrix between weight vectors learned over all 10 folds of data and all four methods. Warmer colors correspond to higher similarity (stability), and cooler tones to lower similarity. SLDSS (upper left square) has the highest similarities among all methods.

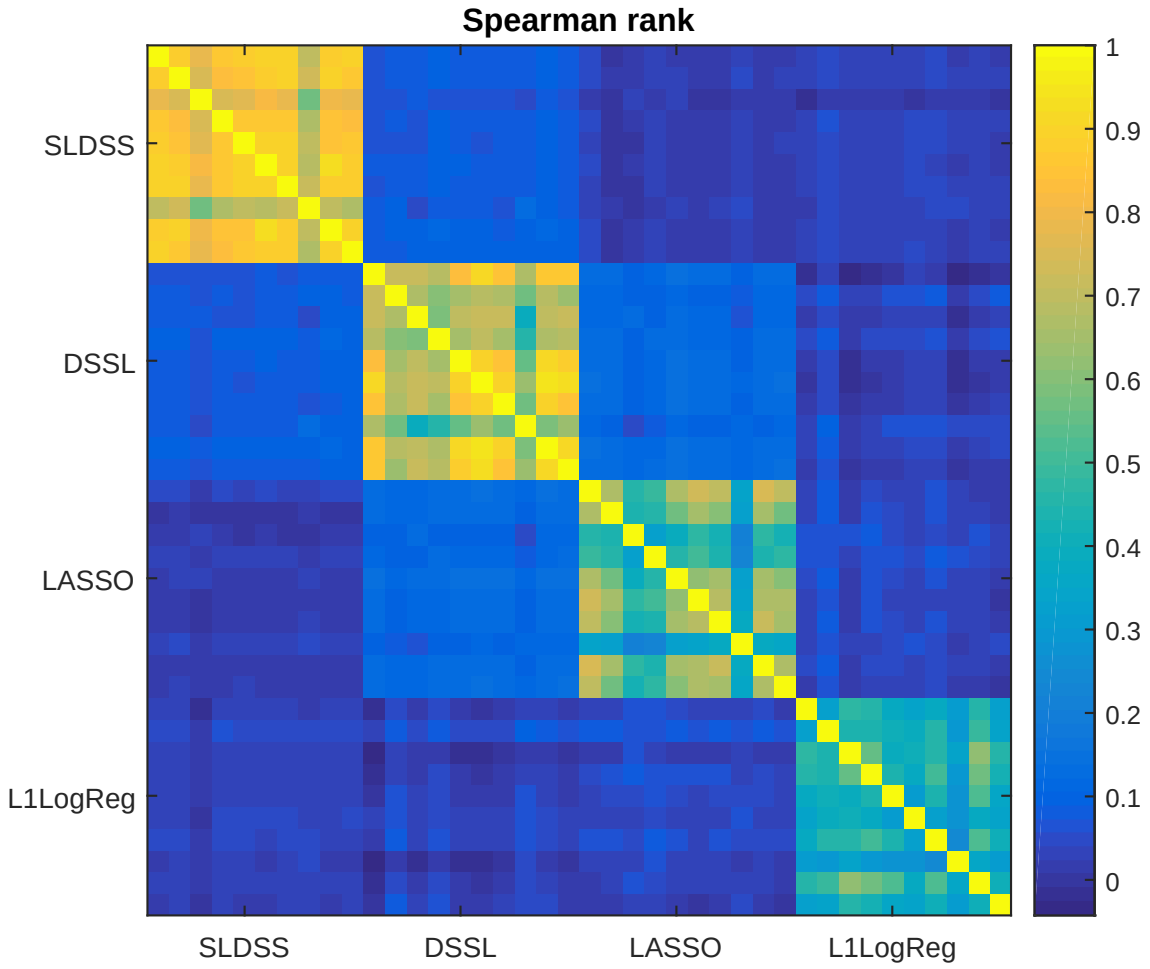


FIGURE B.2: Spearman similarity matrix between weight vectors learned over all 10 folds of data and all four methods. Warmer colors correspond to higher similarity (stability), and cooler tones to lower similarity. SLDSS (upper left square) has the highest similarities among all methods.

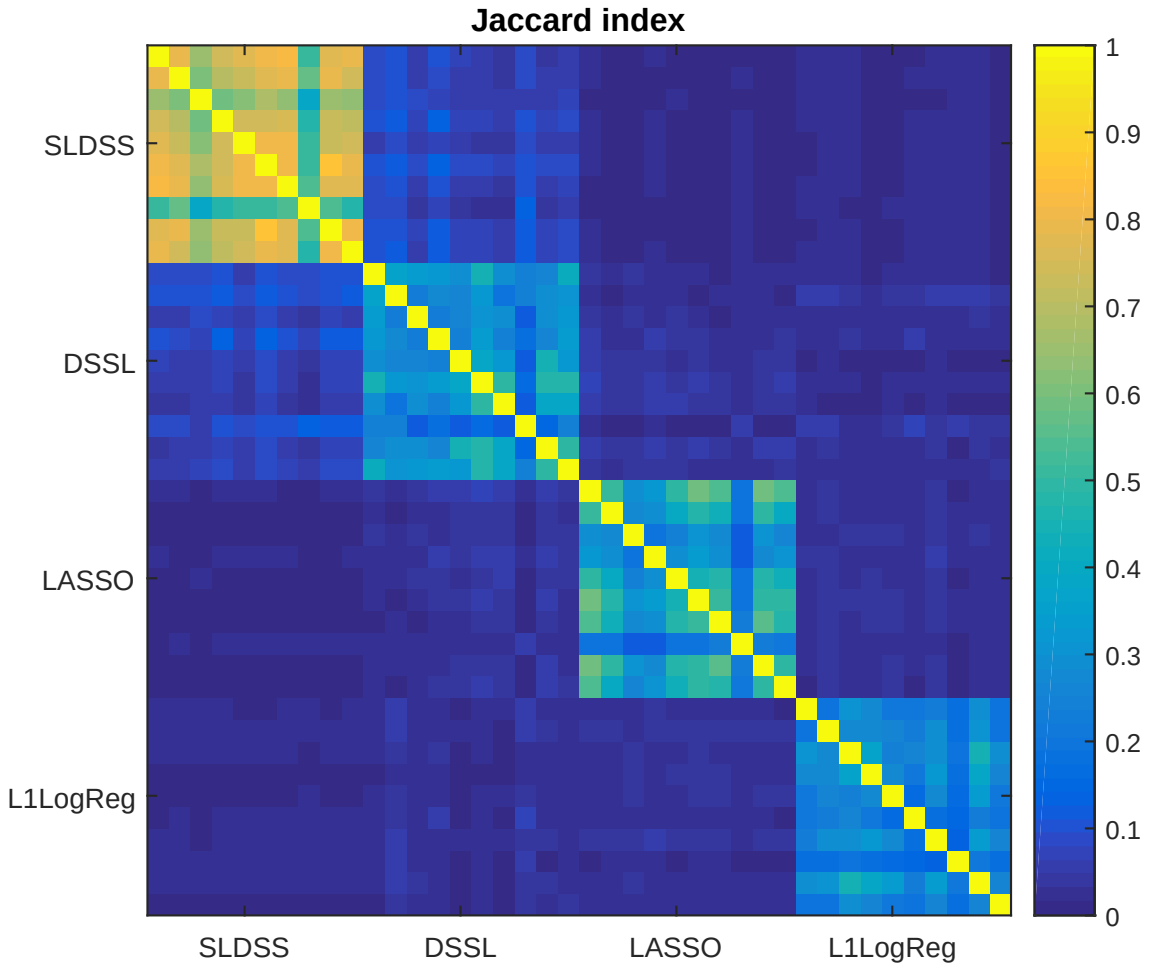


FIGURE B.3: Jaccard similarity matrix between weight vectors learned over all 10 folds of data and all four methods. Warmer colors correspond to higher similarity (stability), and cooler tones to lower similarity. SLDSS (upper left square) has the highest similarities among all methods.

# Biography

Ivan Stojković was born on April 11-th 1987. in Belgrade, Serbia. He finished Mathematical High School for talented pupils in Belgrade. Bachelor and master studies, he completed at the Department for Signals & Systems, School of Electrical Engineering, University of Belgrade, in years 2010. and 2011. respectively. He started doctoral studies in 2012. at Module for Control Systems & Signal Processing at University of Belgrade, and in 2014. at Department for Computer & Information Sciences at Temple University in Philadelphia. He worked as research assistant at Mihailo Pupin Institute in Belgrade, and afterwards at Center for Data Analytics and Biomedical Informatics in Philadelphia. He was involved in three projects funded by DARPA: "Graph-theoretic Research on Algorithms and the Phenomenology of Social networks", "Dialysis-Like Therapeutics" and "Technologies for Host Resilience". He had (co-)authored more than 20 papers in various topics: ranking methods (Stojkovic et al., 2017b; Stojkovic and Obradovic, 2017b), probabilistic graphical models (Stojkovic et al., 2016a, 2017a; Pavlovski et al., 2017), therapy optimization (Stojkovic et al., 2016b; Stojkovic and Obradovic, 2017a), control of aerial (Stojković et al., 2012; Rodić and Stojković, 2012a,b; Rodić et al., 2013) and wheeled mobile robots (Ćosić et al., 2013; Miloradović et al., 2013; Rodić and Stojković, 2013; Rodić et al., 2012), multiagent consensus (Stojković and Katić, 2017; Živanović and Stojković, 2013), and on data mining applications in biomedical domain (Spasojević et al., 2017; Ghalwash et al., 2016; Cao et al., 2016, 2014).

Прилог 1.

# Изјава о ауторству

Потписани-а Иван Стојковић

број уписа 5056/11

**Изјављујем**

да је докторска дисертација под насловом

Примена функционалних норми за регуларизацију рангирања над темпоралним подацима

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

**Потпис докторанда**

У Београду, 01.04.2018.

Иван Стојковић

Прилог 2.

# Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Иван Стојковић

Број уписа 5056/11

Студијски програм Електротехника и Рачунарство

Наслов рада Примена функционалних норми за регуларизацију рангирања над темпоралним подацима

Ментори Зоран Обрадовић и Бранко Ковачевић

Потписани Иван Стојковић

изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

У Београду, 01.04.2018.

**Потпис докторанда**

Иван Стојковић

Прилог 3.

# Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Примена функционалних норми за регуларизацију рангирања над темпоралним подацима

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

**Потпис докторанда**

У Београду, 01.04.2018.

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.