Bayesian Transfer Learning for personalised well-being forecasting from scarce, sporadic observations

Eirini Christinaki

School of Computer Science and Electronic Engineering

University of Essex

A thesis submitted for the degree of

Doctor of Philosophy

October 2020

Dedicated to those who love me unconditionally and are always there for me.
Thank you so much!

Acknowledgements

I would like to express my gratitude and appreciation to my supervisor Prof. Luca Citi for his warm support, advice and encouragement. I am also fortunate to have Prof. Riccardo Poli as my co-supervisor and I am grateful for his timely and invaluable comments. I deeply appreciate how both of you have been continuously guided me all the way through this journey. You have always been so friendly and I really enjoyed every moment we spent together. This work would not have been possible without your tremendous support.

I would also like to thank my senior colleagues from University of Essex Dr. Xinyang Li and Dr. Tasos Papastylianou. It was a great experience and pleasure working with you on the NEVERMIND project. Regarding my PhD, I would like to express my sincere acknowledgement in the constructive suggestions and support of Tasos. I am also sincerely grateful to my supervisory panel members Prof. Francisco Sepulveda and Dr. Ana Matran-Fernandez for providing me with insightful feedback.

I would like to pay my special regards to my fellow and lab mates in Brain-Computer Interfaces and Neural Engineering (BCI-NE) Laboratory at University of Essex for being part of this adventure (in one way or another). I would also like to acknowledge the NEVERMIND consortium partners for providing the relevant infrastructure and the NEVERMIND dataset, which I have extensively used in my experiments. I further wish to thank our NEVERMIND partner *Università deglistudi di Torino* and especially Sara Carletto and Luca Ostacoli for providing me with the correlation matrix of the Davidson's Emotional Styles Questionnaire.

In addition, I deeply thank my friends in Greece, UK and all around the world for always being there for me.

Lastly, I would like to give the deepest thank you to my family for loving me and supporting me unconditionally, no matter what!

Funding body

This research was supported by the EU-funded project with the acronym NEVERMIND[1] (H2020 PHC-2015-689691).

[1]NEurobehavioural predictiVE and peRsonalised Modelling of depressIve symptoms duriNg primary somatic Diseases with ICT-enabled self-management procedures (online at http://www.nevermindproject.eu/).

Abstract

The research presented in this dissertation has been conducted within the context of the NEVERMIND[2] project. The main objective of this PhD was to explore and propose novel approaches for addressing the challenges associated with creating personalised models and making predictions in real world health-related applications when training is performed incrementally on scarce sporadic biomedical data. A particular challenge was being able to provide reliable personalised predictions in the early stage of data collection when insufficient data are available for training. The solution proposed in this dissertation is centred on Bayesian Transfer Learning techniques that allowed me to make informed predictions even in such challenging conditions by leveraging information coming from other patients.

Firstly, I proposed a non-parametric transfer learning approach, which allowed me to make more accurate predictions about a specific patient by combining models trained on other "donor" patients in proportion to how well these models fit the specific patient's past observations. Secondly, I developed a parametric transfer learning approach, which incorporated a modified prior that accounts for the knowledge available from all other "donor" patients. Finally, I proposed modified versions of the previous two approaches, where I controlled how much information is borrowed for transfer based on the similarity in emotional profiles between the patient under test and each "donor" patient. The results show that the proposed transfer learning methods not only naturally dealt with the uneven, sporadic data in the dataset but also performed very well even in the hardest forecasting scenarios, such as the case where only seven days of data are available, and the system is required to forecast for the next seven days. In general these approaches produced better-suited models for participants with very few sporadic training samples and performed significantly better than a number of competing models.

[2]NEurobehavioural predictiVE and peRsonalised Modelling of depressIve symptoms duriNg primary somatic Diseases with ICT-enabled self-management procedures (online at http://www.nevermindproject.eu/).

List of Publications

C1. Eirini Christinaki, Riccardo Poli, and Luca Citi. "Bayesian Transfer Learning for the Prediction of Self-reported Well-being Scores". *In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2018, pp. 41–44
— (relevant work presented in chapters 3 & 5).

C2. Eirini Christinaki et al. "Parametric transfer learning based on the fisher divergence for well-being prediction". *In: Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019*. Institute of Electrical and Electronics Engineers Inc., Oct. 2019, pp. 288–295
— (relevant work presented in chapters 4 & 5).

J1. Eirini Christinaki et al. "Well-being Forecasting using a Parametric Transfer-Learning method based on the Fisher Divergence and Hamiltonian Monte Carlo". *In: EAI Endorsed Transactions on Bioengineering and Bioinformatics* 1.1 (Oct. 2020)
— (relevant work presented in chapters 4 & 5).

Contents

List of Figures

List of Tables

Glossary and Abbreviations

AD . . . . . *Alzheimer's disease*

AVEC . . . . . *Audio-Visual Emotion recognition Challenge*

BCI . . . . . *Brain-Computer Interface*

BMA . . . . . *Bayesian Model Averaging*

DTSVM . . . *Domain Transfer Support Vector Machine*

EM . . . . . *Expectation-Maximization*

ESQ . . . . . *Emotional Styles Questionnaire*

FD . . . . . *Fisher Divergence*

GP . . . . . *Gaussian Process*

GPR . . . . . *Gaussian Process Regression*

HMC . . . . . *Hamiltonian Monte Carlo*

ICU . . . . . *Intensive Care Unit*

LDS . . . . . *Linear Dynamic System*

LL . . . . . *Log Likelihood*

LSTM . . . . . *long short-term memory*

MAP . . . . . *Maximum A Posteriori*

MCI . . . . . *Mild Cognitive Impairment*

MCI-C . . . . *Mild Cognitive Impairment converters*

MCI-NC . . . *Mild Cognitive Impairment nonconverters*

MCMC . . . . *Markov Chain Monte Carlo*

MH . . . . . *Metropolis-Hastings*

MKL . . . . . *Multiple Kernel Learning*

ML . . . . . *Machine Learning*

MTL . . . . . *Multi-Task Learning*

MTMKL . . . *Multi-task, Multi-Kernel learning*

NC . . . . . *Normal Control*

NUTS . . . . . *No-U-Turn Sampler*

RCT . . . . . *Randomised Controlled Trial*

RMSE . . . . . *Root Mean Squared Error*

SD . . . . . *Standard deviation*

SVM . . . . . *Support Vector Machine*

TL . . . . . *Transfer Learning*

WHO . . . . . *World Health Organization*

1 Introduction

This chapter introduces the research topic and the problems investigated in this thesis. It also outlines the motivation and the aim of this research, as well as the research questions and the hypothesis explored in the study. It further presents the main contributions alongside with a brief overview of the chapters ahead.

Contents

In the "era of big data", the curse of small datasets in *machine learning* (ML) still exists in medical applications where experimental datasets can be limited in size due to high cost or complexity of patient data collection, thus making in some cases traditional ML algorithms impractical for predictive modelling. Such a case is the modelling of human affect which is rather challenging for traditional ML algorithms.

1.1 Problem statement

Modelling overall well-being is a difficult task since well-being is a complex internal state consisting of several related dimensions. It is composed of multiple mental and physical factors that are usually measured with self-reported surveys. For depression, self-reported scores are used because studies have confirmed relationships between self-reported affect and clinical ratings of depression [1, 2].

Research on constructing models with the aim of predicting future mood states of individuals has shown that there is individual variation in how someone's mood is affected by different factors and what puts one person in a good or depressed mood state. Such individual differences exist even in terms of how people's mood is affected by the weather [3, 4]. Hence, there is need to move toward personalised approaches since personalised models can take advantage of attributes specific to the individual. In addition, forecasting people's depressed mood, based on self-reported historical moods, behavioural profiles and medical records, collected by their smart-phone has shown that the long-term historical information of a user improves the accuracy of forecasting depressed mood. This fact further stresses the need for accumulating enough data from each participant, since effective prediction in predictive modelling, requires reliable and systematic historical data [5].

Training fully personalised models by using all the available training data for each person would theoretically produce the most accurate model for that person because this model would account for inter-individual variability. However, personalisation that requires tuning/learning a model targeted for the individual user poses some challenges. Ideally, predictions should be provided from day one. This means that initially, when a patient has just been enrolled in a study or has just started using such a system, the model will be expected to make meaningful predictions for that individual despite the fact that no, or very limited patient-specific data will be available. While it is possible to to train/update a model incrementally on the data available at a given time, it is difficult to give reliable predictions when insufficient data are available. Consequently, such a system will only be able to make useful predictions after an adequate amount of data has been obtained which in practice means that a user might not benefit from such a model, until after several weeks or months of use have taken place. Moreover, when training data is scarce it is difficult to adequately capture the complexity of affective states.

Datasets are also very likely to be sparse and contain sporadic observations, both because of the nature of the data acquired and because users may have the option

to refuse or postpone interacting with the system (e.g., answering questionnaires) or wearing sensors, thereby exacerbating the problem. Challenges faced when training a model on such sparse/limited datasets with traditional ML algorithms include over-fitting, difficulties in handling outliers, and inappropriate assumptions of equivalence between training and test data distributions, a concept known as dataset shift [6]. Moreover, models trained on datasets that have a highly unbalanced representation due to sparsity, run the risk of being meaningless or unreliable in practice. Even in the absence of over-fitting, in theory, there could be a large number of specific models that could fit a sparse dataset equally faithfully, but only a small number of those models would correspond to clinical reality. In order to account for the uncertainty surrounding data sparsity, the models need to be of sufficient generality but without risking creating models being of no practical value. Similarly, for any given model, sparsity complicates assessing the effective fit to the data, since there are not enough sample points to help meaningfully differentiate between more specific and more generalised but meaningless models.

Based on the above, it is therefore essential to overcome these limitations by finding new ways to build personalised well-being prediction models that can account for individual differences in the absence of sufficient personal data of suitable nature for training.

1.2 Motivation and aim

NEVERMIND[1] is an EU-funded[2] research project, which aims to provide effective smart-phone-based self-management tools alongside clinical support, to help individuals at risk of developing depressive symptoms as a consequence of a primary

[1]NEurobehavioural predictiVE and peRsonalised Modelling of depressIve symptoms duriNg primary somatic Diseases with ICT-enabled self-management procedures (online at http://www.nevermindproject.eu/).

[2]NEVERMIND has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 689691.

disease (e.g., cancer, myocardial infarction, amputation, nephropathy). In this project, sparse multimodal biomedical and subjective data, including a collection of physiological data, body movement, speech, and the recurrence of social interactions, are collected via a smart-phone and a lightweight sensorized shirt. The data from individual users are collected over time and become available in a sequential manner yet, the predictions are expected to be in real-time and from the day each patient is enrolled, and thus the model needs to be trained/updated incrementally on the data available at a given time. This means that initially, when the patient has just been enrolled in the study, the model will be expected to make predictions for that individual using only the small dataset available for that person. On top of that, this dataset is also very likely to be sparse, since patients have the option to refuse or postpone answering questionnaires or wearing their sensorised T-shirts.

The research presented in this dissertation has been conducted within the context of the NEVERMIND project. The main objective of the research undertaken in this PhD is to explore and propose novel approaches with the aim of addressing the challenges on creating subject-specific models and making predictions in real world health-related applications when training is performed incrementally on scarce sporadic biomedical data. The focus of this work is on providing reliable personalised predictions in the early stage of data collection when insufficient personal data of suitable nature are available for training. In this dissertation, I tackle the challenges and overcome limitations that traditional ML algorithms have, by devising Bayesian Transfer Learning techniques that allow informed predictions to be made by leveraging information coming from other patients in the study. Specifically, when we have a task in a domain of interest, but we do not have sufficient training data, as in the case studied in this PhD, transfer learning can be used to improve learning from this domain by transferring information from a related domain. This information transfer will be particularly beneficial in that case because by appropriately sharing knowledge between personalised models, there can be an opportunity to enhance performance (i.e., to improve the quality of subject-specific models) even with scarce sporadic observations available for each

subject. In other words, more accurate models could be built for each subject while controlling for the risk of over fitting because the limited amount of data available for each subject will be compensated by the presence of data from subjects with similar characteristics. This can be achieved with Bayesian methods which use probabilities for quantifying the uncertainty in inferences based on statistical data analysis. The use of Bayesian methods arise from the need to quantify and reduce uncertainty which inevitable occurs from possible errors in model structure and inputs. Although uncertainties in data, model inputs and model structure can not be quantified directly, they can be represented using probability distributions and this stimulates the use of Bayesian methods. In our case, with these methods we can include the uncertainty considering the scarce, sporadic observations and at the same time we can incorporate any prior information into the estimates.

Formally, I address the following research questions:

Q.1 How does transfer learning affect the forecasting accuracy of personalised models in real-world applications, when person-specific data is scarce?

Q.2 How many days worth of person-specific data are required when training such a personalised model, before the contribution of a transfer learning component to the model becomes negligible?

Q.3 Does transferring from patients' groups with similar characteristics to the patient in question have any further benefit to the transfer learning process, over the more general transfer learning scenario based on transferring from the entire-population?

With these research questions in mind, I formulate the following hypotheses for formal testing:

H.1 *In the presence of limited person-specific data available for training (e.g. at the early stage of data collection), a prediction model leveraging both patient-specific*

data (i.e. the "target" domain), and knowledge gained from other patients (i.e. a different but related "source" domain), will perform significantly better than a model which relies on the target domain alone for training.

H.2 *The benefit of transfer learning on predictive performance will be greater when the amount of person-specific data available for training is small; as person-specific data availability increases, the relative contribution of the transfer learning component to the overall accuracy will diminish, until it becomes negligible.*

H.3 *A transfer learning model mostly incorporating highly relevant information from a source domain (i.e. biasing population sampling towards participants having similar characteristics to the patient being modelled) will perform significantly better than a model which utilizes the source domain in a general manner (i.e. sampling the population with equal probability over all other participants)*

1.3 Claims and contributions

In this dissertation I devised Bayesian transfer learning techniques to address the challenges on creating subject-specific models and making predictions in real world health-related applications when training is performed incrementally on scarce sporadic biomedical data. The proposed approaches were developed within the scope of the EU–funded Horizon 2020 NEVERMIND project. In this thesis, the effectiveness of these techniques has been demonstrated in the context of personalised prediction of self-reported well-being scores, using data from the NEVERMIND project [7, 8]. Finally, the proposed methods have been already and are still used by the project in the context of the self-management tool. In summary, the main claims and contributions are the following:

Chapter 3:

(i) A non-parametric transfer learning approach based on a Markov Chain Monte Carlo sampler and Bayesian Model Averaging, which allows to make more accurate predictions about a specific patient by combining models trained on other "donor" patients in proportion to how well these models fit the specific patient's past observations (published as [9]).

Chapter 4:

(ii) A parametric transfer learning approach based on the Fisher divergence, which expresses external information coming from "donor" patients as a prior probability distribution used within a Hamiltonian Monte Carlo framework. (published as [10, 11]).

Chapter 5:

(iii) Modified versions of the previous two approaches presented in Chapters 3 & 4, which controls how much information is borrowed for transfer based on the similarity in emotional profiles between the patient under test and each "donor" patient.

1.4 Thesis outline

The content of this thesis report is organised in the following manner:

Chapter 2 describes the relevant background and provides an overview of related work.

Chapter 3 presents a non-parametric Bayesian transfer learning method based on a Markov Chain Monte Carlo sampler and Bayesian Model Averaging.

Chapter 4 presents a parametric Bayesian inference method making use of transfer learning in the context of a Hamiltonian Monte Carlo sampling, which allows

a population prior to be directly represented in the sampling process through the use of the Fisher divergence.

Chapter 5 presents a similarity-based transfer learning approach mostly incorporating highly relevant information from the source domain by biasing population sampling towards participants having similar characteristics, based on their emotional profiles, to the patient being modelled.

Chapter 6 concludes the thesis by summarising the main contributions of this thesis, discussing the limitations of the work and looking at promising avenues for future research.

2

Background and Related Work

This chapter gives the readers the relevant background and material on Bayesian method and transfer learning techniques and provides an overview of related work.

Contents

2.1 Markov Chain Monte Carlo inference

Bayesian methods can be used in many fields, from ecology [12] to bioinformatics [13] conservation biology [14], drug discovery, epidemiology and biostatistics [15], just to mention a few. They have a number of advantages over other statistical modelling and data analysis techniques. In Bayesian methods, the models can easily accommodate unobserved variables, the use of prior probability distributions allows to incorporate prior information and the posterior probability can be used as easily interpretable alternative to p value. As a result, there are a large number of articles, books and courses that present the foundations and key theoretical concepts of Bayesian methods and approaches. Well structured descriptions of the

rationale together with a brief overview of Bayesian data analysis can be found in [16] and [17] or see [18] for a comprehensive review.

Markov Chain Monte Carlo (MCMC) are techniques for estimating *posterior distributions* in Bayesian inference or more general for obtaining information about distributions. MCMC is a very general and powerful framework which allows sampling from a large class of distributions and it constitutes the most popular method for sampling from high-dimensional distributions. A particular strength of this method is that it can be used when we cannot directly draw samples and works well even for complicated distributions in high-dimensional spaces. It can be used to obtain posterior parameter estimates when these are difficult to express in closed form. However, MCMC presents the drawback that it is often very slow, especially for high-dimensional models [19].

MCMC generates samples from the posterior distribution by constructing a Markov chain that has as its equilibrium distribution the target posterior distribution. According to Bayes' theorem, the posterior probability of the model parameters θ given the data D is:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}. \tag{2.1}$$

The posterior is a probability distribution representing what we think about the parameters after having seen the data. It can be found by combining the *prior* $p(\theta)$, which is what we think about the parameters θ before we have seen any data D, with the likelihood $p(D|\theta)$, i.e., how we think our data is distributed. In the denominator, we have the normalizing constant $p(D)$, which is called the evidence and can be regarded as a normalizing constant to ensure that $p(\theta|D)$ is a proper density and can be computed by integrating over all possible parameter values:

$$p(D) = \int_\theta p(D, \theta)d\theta = \int_\theta p(D|\theta)p(\theta)d\theta. \tag{2.2}$$

In several cases, this integral can be high-dimensional and thus difficult to compute. Moreover, the MCMC technique uses only θ to find how good the fit is and $p(D)$

does not rely on θ. Therefore, we can ignore $p(D)$ because we are just trying to get the target distribution, which needs to be proportional to the posterior distribution (2.3) and this normalization factor makes no difference to that. That is:

$$p(\theta|D) \propto p(D|\theta) \times p(\theta). \tag{2.3}$$

2.1.1 Metropolis Hastings algorithm

Metropolis-Hastings (MH) is a specific implementation of MCMC, which works well in high dimensional spaces. This technique requires a simple distribution q called the *proposal distribution* to help draw samples from an intractable posterior distribution. In MH, at each step, we propose to move from the current θ to a new proposed θ' that is a random sample drawn from $q(\theta'|\theta)$. The proposal distribution is used to randomly walk in the distribution space and to accept or reject jumps to new positions based on how likely the sample is. The likelihood of each new sample θ' is decided by an acceptance function which must be proportional to the posterior we want to sample from. It is common to choose the acceptance function being a probability density function that expresses this proportionality. In the case of symmetric proposals, where $q(\theta'|\theta) = q(\theta|\theta')$, the acceptance probability α is given by:

$$\alpha = \min\left(1, \frac{p(\theta')}{p(\theta)}\right) \tag{2.4}$$

This means that if θ' is more likely than the current θ, then we always accept the move. If it is less probable than the current θ, we might reject the move or still accept it and move there anyway depending on the relative probabilities. So instead of greedily moving to only high density regions, we occasionally allow visiting low-density regions.

In order to be able to use an asymmetric proposal distribution, where $q(\theta'|\theta) \neq q(\theta|\theta')$, the Metropolis-Hastings algorithm includes an additional correction factor in the acceptance probability:

$$\alpha = min\left(1, \frac{p(\theta')}{p(\theta)} \times c\right) \quad \text{where} \quad c = \frac{q(\theta|\theta')}{q(\theta'|\theta)}. \tag{2.5}$$

This correction is needed to compensate for the fact that the proposal distribution itself (rather than the target distribution) might favour certain states [20].

2.1.2 Hamiltonian MCMC

Hamiltonian Monte Carlo (HMC) is a gradient-based MCMC sampler that uses the derivatives of the density function being sampled to generate efficient transitions spanning the posterior. It avoids random walk behaviour by simulating a physical system governed by Hamiltonian dynamics. It uses an approximate Hamiltonian dynamics simulation based on numerical integration which is then corrected by performing a Metropolis acceptance step. In other words, it uses Hamiltonian dynamics to make proposals as part of an MCMC method. To do so, HMC introduces auxiliary momentum variables p and draws from a joint density $p(\mathrm{p}, \theta) = p(\mathrm{p}|\theta)p(\theta)$. The auxiliary momentum variables p are introduced to create an auxiliary probability distribution which admits the target distribution as a marginal. The joint density $p(\mathrm{p}, \theta)$ defines a Hamiltonian $H(\mathrm{p}, \theta) = -\log p(\mathrm{p}, \theta) = -\log p(\mathrm{p}|\theta) - \log p(\theta)$ where we can interpret the first term as a "kinetic energy" $T(\mathrm{p}|\theta) = -\log p(\mathrm{p}, \theta)$ and the second term as a "potential energy" $V(\theta) = -\log p(\theta)$ [21].

With HMC, the basic idea is to think of the parameters as a particle in a multi-dimensional space and create auxiliary variables which represent the "momentum" of this particle. The algorithm works as follows:

- Starting from the current value of the parameters θ, a transition to a new state is generated in two stages:
    1. a value for the momentum p is drawn independently of the current parameter values
    2. the joint system (θ, p) made up of the current parameter values θ and new momentum p is evolved via Hamilton's equations obtaining (p', θ').
- Finally, we apply a Metropolis acceptance step, where the probability of keeping the proposal (p', θ') generated by transitioning from (p, θ) is:

$$\alpha = \min\left(1, \frac{p(\mathrm{p}', \theta')q(\mathrm{p}, \theta|\mathrm{p}', \theta')}{p(\mathrm{p}, \theta)q(\mathrm{p}', \theta'|\mathrm{p}\theta)}\right) = \min\left(1, \exp\left(H(\mathrm{p}, \theta) - H(\mathrm{p}', \theta')\right)\right). \quad (2.6)$$

In summary, the HMC approach involves alternating between a series of leapfrog updates and a re-sampling of the momentum variables from their marginal distribution. The decision whether to update the new state or keep the existing state is taken by applying the Metropolis acceptance step.

2.2 Transfer Learning

In machine learning, we have a model defined up to some parameters, and learning is the execution of a computer program to optimise the parameters of the model using the training data or the past experience [22]. The model may be predictive to make forecasts in the future or descriptive to gain knowledge from data or both. For machine learning tasks, it is essential to use a training set of data to discover potential predictive relationships, while a test and validation set is further used for evaluating whether the discovered relationships hold more generally that just for the training data. In other words, machine learning focuses on designing algorithms that can learn from data and make predictions based on properties learned from them.

Traditional ML algorithms work under the common assumption that the training and test data are drawn from the same feature space and have the same distribution [23]. These methods make predictions on future data using mathematical models that are trained on previously collected (labelled or unlabelled) data, which are similar in nature to future data. However, in many real-world applications these assumption does not hold. There are cases where the feature space or the distribution of the test data changes, and, so, the prediction models cannot be used and must be rebuilt after having collected enough new training data, which is very expensive and sometimes not practically possible. This issues is particularly relevant to our research since the data gathered in the NEVERMIND project have the problems of being sparse and non-stationary.

Transfer Learning (TL) methods are a recent class of techniques, which enable one to work around the strict requirement that the test and training data should necessarily conform to the same probability distribution [24]. These methods can use data from unrelated or partially related tasks [25], and allow the domains, tasks, and distributions used in training and testing to be different up to a certain point, thereby solving the problem of otherwise having to build a completely new model from scratch [26]. They rely on the basic assumption that the source and target domains, while not necessarily of the same underlying distribution, may still be related in other ways, i.e., via an explicit or implicit relationship between the feature space of the two domains. The goal of TL is to improve learning in the target domain by leveraging previously acquired knowledge gained in the source domain.

There are three common measures by which transfer learning might improve learning [27]: (1) the initial performance, (2) the learning time and (3) the final performance. In more detail, as shown in Fig. 2.1, with transfer learning we can achieve a better initial performance in the target task (a higher start), less time spent to fully learn the target task (a higher slope) or a higher final performance level in the target task (a higher asymptote). However, the effectiveness of a transfer learning method depends on the source task and how it is related to the target task. If the relationship is strong and the transfer method can take advantage of it, the performance in the target task can significantly improve through transfer. On the other hand, if the source task is not sufficiently related to the target task or if the relationship is not well leveraged by the transfer method, with many approaches the performance may fail to improve or even worsen [28]. If the source domain/task data yields a reduced performance of learning in the target domain/task then negative transfer has occurred [29]. Avoiding transfer that actually decrease performance is a very important issue. Ideally, a transfer learning method must produce positive transfer between appropriately related tasks while avoiding negative transfer when the tasks are not a good match. Following the description and notation in [23], a domain $D = \{\mathcal{X}, P(X)\}$ is defined by two parts, a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, ..., x_n\} \in \mathcal{X}$. The example given

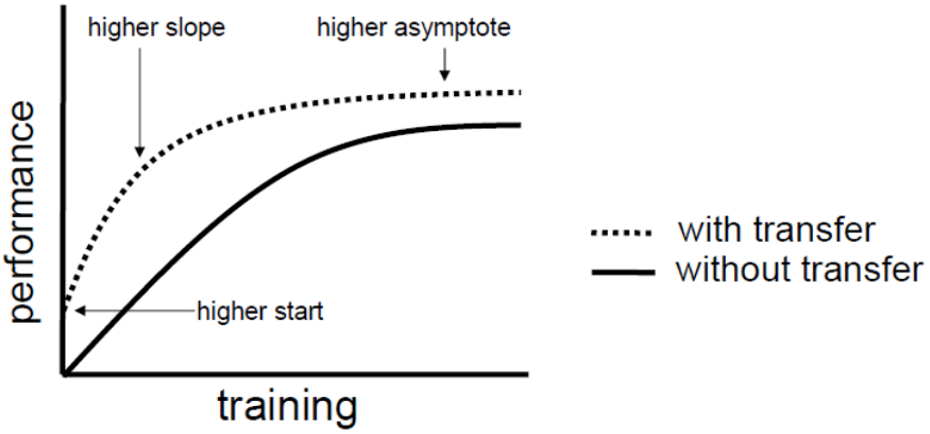


Figure 2.1: How transfer might improve learning (image taken from [27])

by Pan et.al [23] is a document classification learning task where each term is taken as a binary feature. Then, $\mathcal{X}$ is the space of all possible term vectors, x_i is the i-th term vector corresponding to some documents, n is number of term vectors in X and X is a particular learning sample.

Also, a task $T = \{\mathcal{Y}, f(\cdot)\}$ is defined by two components, a label space $\mathcal{Y}$ and an objective predictive function $f(\cdot)$ which is not observable but can be learned from the training data which consist of pairs $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in \mathcal{Y}$. The function $f(\cdot)$ can be used to predict the corresponding label, $f(x_i)$, of a new instance x_i. From a probabilistic point of view $f(x_i)$ can be written as $P(y_i|x_i)$ thus, the task can be also denoted as $T = \{\mathcal{Y}, P(Y|X)\}$.

With this notation, transfer learning is defined as follow:

Definition 1: (*Transfer Learning*) Given a source domain D_S and learning task T_S, a target domain D_T and learning task T_T, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$.

In the above definition, a domain is a pair $D = \{\mathcal{X}, P(X)\}$. Thus, if two domains D_S and D_T are different, i.e., $D_S \neq D_T$, this implies that either the feature

spaces between the domains are different ($\mathcal{X}_\mathcal{S} \neq \mathcal{X}_\mathcal{T}$) or the marginal probability distributions between domain data are different ($P(X_S) \neq P(X_T)$ where $X_{S_i} \in \mathcal{X}_\mathcal{S}$ and $X_{T_i} \in \mathcal{X}_\mathcal{T}$) or both. Similarly, a task is defined as a pair $T = \{\mathcal{Y}, P(Y|X)\}$ thus, if the learning tasks T_S and T_T are different, this implies that either the label spaces between the domains are different ($\mathcal{Y}_\mathcal{S} \neq \mathcal{Y}_\mathcal{T}$) or the conditional probability distributions between the domains are different ($P(Y_S|X_S) \neq P(Y_T|X_T)$, where $Y_{S_i} \in \mathcal{Y}_\mathcal{S}$ and $Y_{T_i} \in \mathcal{Y}_\mathcal{T}$) or both. When $D_S = D_T$ and $T_S = T_T$, the learning problem becomes a traditional ML problem.

Based on the availability of labelled data and the different situations in relation to the source and target domains and tasks, transfer learning approaches can be categorized as *inductive*, *transductive* or *unsupervised*. Inductive learning refers to learning techniques which try to learn the objective predictive function $f_T(\cdot)$. Transductive learning techniques try to learn the relationship between instances. Unsupervised transfer learning focuses on solving tasks such as clustering and dimensionality reduction.

As Pan et.al [23] explain in their survey for transfer learning, in inductive transfer learning the target task T_T is different from the source task T_S whereas the source domain D_S and target domain D_T can be the same or different. In cases where a lot of labelled data are available in the source domain D_S, inductive transfer is similar to *multi-task learning* (MTL). When there are no labelled data available in the source domain D_S, inductive transfer is similar to self-taught learning. In both cases, a few labelled data in the target domain D_T are required. On the other hand, in transductive learning, the source task T_S and the target task T_T are the same, while the source domain D_S and the target domain D_T are different. This means that either the feature spaces between the domains are different ($\mathcal{X}_\mathcal{S} \neq \mathcal{X}_\mathcal{T}$) or the marginal probability distributions of the input data are different ($P(X_S) \neq P(X_T)$) which is a case related to domain adaptation. As for the data, in both these cases, a lot of labelled data are available in the source domain D_S but there are no labelled data available in the target domain D_T. Finally, in unsupervised transfer learning,

the target task T_T is different from the source task T_S but related to it. In this case, there are no labelled data available in either source domain D_S or target domain D_T.

Transfer learning techniques can be further grouped into *homogeneous* and *heterogeneous* [30]. In the case that the two domains D_S and D_T are different ($\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$) but the feature spaces are the same ($\mathcal{X}_\mathcal{S} = \mathcal{X}_\mathcal{T}$), the transfer learning is defined as homogeneous, whereas when the feature spaces between the domains are different ($\mathcal{X}_\mathcal{S} \neq \mathcal{X}_\mathcal{T}$), the transfer learning is defined as heterogeneous.

Transfer learning techniques can be also categorised into classification, regression and clustering problems [23, 30] or transfer learning for reinforcement learning [31]. Lastly, based on what knowledge can be transferred across domains and tasks, the existing transfer learning approaches fall into four categories [23, 26, 32]:

(1) ***Instance transfer*** where some samples/instances from the source domains are reused for learning in the target domain by re-weighting them. These techniques work well when the feature spaces are the same ($X_S = X_T$).

(2) ***Feature-representation transfer*** where a good feature representation of the data is used to reduce the differences between the source feature space $\mathcal{X}_\mathcal{S}$ and target feature spaces $\mathcal{X}_\mathcal{T}$. These methods try either to make the target and the source distributions look similar, or they try to find an abstracted representation for domain-specific features.

(3) ***Parameter transfer*** where the main assumption is that the source task T_S and the target task T_T share some parameters or prior distributions. The knowledge can be transferred across tasks by discovering the shared parameters or learning the prior distributions shared between the source and the target datasets.

(4) ***Relational-knowledge transfer*** where the main assumption is that in cases where the data are not independent and identically distributed (i.i.d.) but represented by multiple relationships, the knowledge to be transferred is

a similarity relationship among the data in the source domain D_S and the target domain D_T.

Alternative categorisations can been also found in studies based on applications in different research areas. For instance, in the transfer-based activity recognition [32], authors categorise the existing approaches by sensor modality, by differences between source and target environments, by data availability, and by type of information that is transferred. As another example, in Brain-Computer Interface (BCI) classification problems [33] authors organise the existing transfer learning methods into three different settings based on the transfer strategies: (a) feature representation-transfer learning, (b) instance-transfer learning and (c) classifier-transfer learning.

For a more general overview of the history, taxonomy, and state of the art in transfer learning methods for classification, regression, and clustering problems, see [23, 26, 30–33].

2.2.1 Bayesian transfer learning approaches

Bayesian transfer learning methods have been successfully applied in making predictions or classification tasks. These techniques refer to methods that are related to statistical inference and rely on Bayes theorem.

Hierarchical Bayesian approaches have been widely considered for transfer learning because they can effectively combine data from multiple sources [25, 34]. For instance, it has been shown that when tasks are hierarchically related, a hierarchical Bayesian transfer framework can significantly improve learning speed [35]. This framework can be adapted to solve sequential decision problems where the training examples, unlike the standard supervised setting, are not independently identically distributed.

In non-transfer learning settings, Gaussian Process Regression (GPR) has been recently successfully applied to personalised time-series modelling [36] where the Gaussian Process (GP) inference was expressed in a Bayesian framework for the

optimisation of this specific domain. Likewise, in traditional ML, the standard *naïve Bayes* is among the most popular and effective classification algorithms [37]. In a relatively recent work, a hierarchical extension of the classic *naïve Bayes classifiers* has been proposed as an alternative and efficient method for MTL [25]. In that work, the focus was on making prediction from multiple related datasets via transfer learning. The main idea was to partition the dataset into a number of clusters, such that the data for all tasks in a cluster had the same distribution. In particular, the *naïve Bayes classifier* was extended to a multi-task setting by training one classifier for each cluster and then, all classifiers were combined using a Dirichlet process. The resulting model was tested on real data, in a multi-task classification problem. To evaluate the clustered model, its predictive performance was assessed in a transfer learning setting, predicting labels for a user[1] with sparse data, having observed all the labelled data for the remaining users. Results suggested that the *clustered Naïve Bayes* model, which uses a Dirichlet process prior to coupling the parameters of several models applied to separate tasks, improved the performance in situations where the model was presented with multiple, related tasks. The clustered model could use data from related users to provide better prediction even with very few examples. This approach could immediately be applicable to any collection of tasks the data of which are modelled by the same parametrised family of distributions, whether those models were generative or discriminative.

A rather similar approach has been proposed in [34] where a hierarchical non-parametric Bayesian model was developed, although this time the model learns from single training examples. This model transfers the acquired knowledge from previously learned categories to a novel category, in the form of a prior over category means and variances. However, in the case where only a single example from a new category is provided, estimating the variance and similarity metric for categorising an object is very difficult. Thus, the proposed model initially discovers how to group categories into meaningful super-categories that express

[1] the terms user, participant and subject will be used interchangeably throughout this thesis

different priors for new classes and then, when given a single example of a novel category, the model can efficiently infer the super-category that the novel category belongs to. Consequently, the model estimates not only the new category's mean but also an appropriate similarity metric based on parameters inherited from the super-category. This method was tested on image datasets where, according to the authors, the model learned useful representations of novel categories based on just a single training example and performed significantly better than simpler hierarchical Bayesian approaches.

2.3 Well-being modelling and prediction

Affect is a collective term for describing feeling states like emotions and moods. Emotions and moods are mainly distinguished by their duration and by whether they are directed at a specific cause [38]. In contrast to emotion, *mood* is defined as a transient, low-intensity, non-specific and subtle affective state that often has no clear cause. The mood states are short-term and transient feelings, whereas affective traits are stable, long-term individual differences in the tendency to experience a certain mood state [39]. The affective states provide valuable information about personal traits, sociability and well-being.

Well-being can be described as the experience of happiness, health and prosperity. It can be understood as how people feel and how they function, both on a personal and a social level and how they evaluate their lives as a whole. Studies have investigated the role of mood-related factors in judgements of general well-being and results have shown that people use their momentary affective states in making judgements of how happy and satisfied they were with their lives in general [40]. Similarly, there is strong evidence that the global judgements of affect depends on how people feel at the time they provide their affective state [41]. Important contributors to subjective well-being, like happiness and life satisfaction, are associated with depression. A recent cross-sectional study in Korea [42] has shown that life satisfaction and happiness

were significantly associated with a lower risk of depression. In that study, university students with depression showed a lower level of life satisfaction and happiness than the non-depressed students. Results also indicated that individuals with higher levels of life satisfaction and happiness had a decreased probability of having depression. In another longitudinal study [43], self-reported life satisfaction strongly predicted subsequent depressive symptoms in a 15-year follow-up of healthy adults. Results also showed that life satisfaction is strongly associated with concurrent depressive symptoms and the level of life satisfaction can help to detect a group of people from the general population with low subjective well-being and a high risk of having, or developing, depressive symptoms.

According to the World Health Organization (WHO), depression is estimated to affect more than 264 million people of all ages, in all communities across the world [44]. It is a leading cause of disability worldwide, the third leading contributor to the global burden of disease for females and a major contributor to the overall global burden of disease. In the United States of America (U.S.) alone, it is estimated that 17.3 million adults had at least one major depressive episode in their lifetime (7.1% of all U.S. adults [45]). The major depressive disorder, also known as clinical depression, which is a mood disorder characterized by persistent feelings of sadness, low self-esteem, and loss of interest [46]. It affects both men and women (although major depression is more prevalent in women), across all ages and ethnicities worldwide [47]. In general, depression is a common mental disorder and is characterized by sadness, loss of interest or pleasure, feelings of guilt or low self-esteem, disturbed sleep or appetite, feelings of tiredness, poor concentration and even medically unexplained symptoms. It often also comes with symptoms of anxiety [48], it is associated with increased risk of mortality [49, 50] and there is evidence for the strong bi-directional association between physical illness and depression [51].

Research has shown that having a physical illness is a strong risk factor for depression and especially people with severe (primary somatic) disease have a higher risk of depression [52]. At the same time, depression is a risk factor for

developing or exacerbating existing physical illness and is also related to the onset or worsening/improvement of a chronic medical illness [53, 54]. The co-occurrence can make the symptoms worse and recovery more difficult in both cases. In order to understand the complexities of this association and the best ways to treat each, both the depression and the medical illness need to be considered. Based on the comorbidity and the fact that the onset of depressive symptoms can be an aggravating factor of psychosomatic diseases, identifying early warning signs or predicting the severity and onset of these symptoms is of great importance and can play a key role for effective interventions to mitigate or even prevent negative consequences. This was the reason for the NEVERMIND project too.

2.3.1 Monitoring systems in health research

In the last decade, researchers have developed monitoring systems that incorporate wearable sensors. The key benefits of these systems, in addition to enabling ubiquitous service provision, are their low-cost, small size, lightness and low-power consumption. They mainly are devices for monitoring blood pressure, blood glucose levels, cardiac activity, respiratory activity [55] and electrodermal activity [56]. The applications of such systems related to medicine and healthcare include wearable sensors in the form of, e.g., gloves [57] or sensorized t-shirts [58] that have been developed for the characterization of depressive states in bipolar patients [59] and the physiological monitoring in affective computing studies [60].

Monitoring systems for automatic depression assessment based on visual cues is another rapidly growing research domain [61]. These systems employ visual cues for automatic diagnosis and/or severity assessment of depression. Significant progress has also been made on the automatic detection and prediction of depression through the analysis of speech acoustics [62], vocal prosody [63], head pose and movement analysis [64], facial expression and gesture analysis [65] as well as combinations of these [65–67].

The rapid growth in the use of smart-phones has also played an important role in the integration of these devices in health research. Nowadays, researchers have begun to explore the use of healthcare apps utilizing the built-in sensors that are available in smart phones. Current research includes apps for clinical assessment, care screening and symptom monitoring. For example, apps are available for monitoring dementia patients [68], detect falls for the elderly [69] and testing cholesterol [70]. The sensing data collected from smart-phones have been also used to extract features related to depressive mood [71] or to recognize depressive and manic states and detect state changes of patients suffering from bipolar disorder [72].

In recent years, there has been an increase in the number of mental health applications, including those related to depression, available to the public. However, many of the proposed systems and their purported benefits are often not properly backed up by evidence obtained from appropriate scientific research or clinical studies [73]. The effectiveness of these applications has not yet been established and research has shown that although they have the potential to improve treatment accessibility and reduce symptoms [74], the majority of them lack of proof about their efficacy [75]. Therefore, finding a tool supported by robust evidence to manage the depression has become a challenge [76].

Recently, transfer learning has been introduced into medical predictive modelling. For example, Cheng *et al.* [77] developed a domain transfer learning method to predict mild cognitive impairment (MCI) conversion using a multimodal dataset (MRI, FDG-PET, and cerebrospinal fluid data) from normal control (NC) subjects and patients with MCI and Alzheimer's disease (AD). In the proposed framework, MCI converters (MCI-C) and MCI nonconverters (MCI-NC) patients were recognized by using AD and NC subjects as auxiliary domain. The authors implemented a Domain Transfer Support Vector Machine (DTSVM) which was used to classify MCI-C and MCI-NC patients (i.e., target data) with the help of AD and NC patients as the auxiliary data. This approach showed a great performance improvement,

compared to traditional Support Vector Machine (SVM) classification algorithm, with an overall MCI-C and MCI-NC classification accuracy of 79.4%.

2.3.2 Machine learning approaches for well-being prediction

Monitoring and predicting mood states through non-invasive means has attracted a significant research interest over the past decade. This has mainly arose from evidence which suggest that well-being is at least as sensitive a predictor for the risk of adverse outcomes as formal clinical assessment of depression based on established diagnostic criteria [78–83]. While some of these approaches involve laboratory-based techniques such as electroencephalography (e.g. [84]), or elaborate, bespoke equipment (e.g. the "Smart Mirrors" project [85]), smart-phones and wearable devices due to their ubiquity and convenience have now become the predominant research focus for the non-invasive collection of information and signals for the purpose of predicting mood states (e.g. [5, 71, 86–91]). For example is Suhara *et al.* [5] a deep-learning-based approach was used to forecast people's depression mood, using self-reported historical moods, behavioural types and medical records, collected by a smart-phone. The method used long short-term memory (LSTM) neural network models to predict the mood in the following day given two weeks of daily mood reports. Results showed that the method can forecast the severely depressed mood of a user based on self-reported histories with high accuracy, while the history of the previous two weeks was sufficient to forecast future severe depression. In addition, the analysis indicated that the present mental state of a user depends not only on the mood in the previous day but also on the mood in other days. Still, the results showed that long-term historical information on the mood state of a user improves the accuracy over a systems that relies only on information from the previous day, which stresses the need for accumulating enough data from each participant. These findings are also inline with outcomes from the *MoodScope* study [86], where a mood inference classifier managed to statistically infer a user's daily mood average with an accuracy of 93% after a

two-month personalised training period. In that study, smart-phone usage logs and self-reported mood data were collected over two months from a small size, fairly homogeneous, population. For the initial stages of data collection, when the training data for a new user are not enough, authors proposed the use of an all-user model created from an aggregate of all of the users' data to predict the mood.

Regarding transfer learning, related work includes a Multi-task, Multi-Kernel learning approach applied to the problem of predicting students' well-being using survey, mobility, smart-phone and physiology data obtained over a period of 30 days [92]. In that work, the proposed method combines data from multiple modalities and shares information among multiple related tasks. More specifically, the classifier combines the kernels for each modality using a set of kernel weights for each task and these weights are then regularised globally, allowing information about the weights to be shared across tasks. According to preliminary results, this method could successfully classify five dimensions of well-being (happiness, health, alertness, energy and stress) within a single model. Furthermore, this approach provided performance improvements compared to both traditional SVM classifiers and Multiple Kernel Learning (MKL).

In [93], researchers treated well-being prediction as an MTL problem. The method uses MTL to predict future well-being of college students by treating the classification of different well-being states as related tasks. In that work, a *Multi-task, Multi-Kernel learning* (MTMKL) method, a Hierarchical Bayesian model and a Deep Neural Network, which are three formulations of MTL, were compared. In MTMKL method, information across tasks was shared through kernel weights on feature types. In the Hierarchical Bayesian model, tasks shared a common Dirichlet prior in order to constrain the tasks to be similar. Lastly, in the Deep Neural Network, several hidden layers were shared between tasks, while the final layers were task specific. The experimental results have shown that accounting for individual differences through MTL, dramatically improved the well-being prediction performance. This

improvement was based on the fact that MLT allows to have a model specifically trained for each user which also benefits from the data of other users.

MTL has been also employed in [88] to train personalised ML models which were customized to the needs of each individual, but still leveraged data from across the population. The goal of this work was to predict students well-being (mood, stress and health). The models were again (as in [93]) a MTL deep neural network that share several hidden layers but have final layers unique to each task, a MTMKL that feeds information a cross tasks through kernel weights on feature types and finally a Hierarchical Bayesian model in which tasks share a common Dirichlet Process prior. In binary classification, dramatic improvements to mood prediction performance were observed when MTL was used to personalise ML models by multi-tasking over clusters of similar people.

Transfer learning techniques that take into account population heterogeneity have been also proposed in domains involving sequential data modelling. Recently, an online transfer learning technique for hidden Markov models with Gaussian mixture models [94] was proposed for addressing the problem of inferring a sequence of hidden states associated with a sequence of observations produced by an individual within a population. This approach learns different transition and emission models (it estimates the parameters of the transition and emission distributions) for each individual in the training population. These models are then treated as basis models to speed up the online learning process for new individuals. In this way, individuals in the population are used to make predictions about similar individuals by identifying those individuals who closely resemble each other. The approach appears to outperform online EM and online Variational Bayes when tested in real-world applications, which included activity recognition, sleep classification and prediction of packet flow direction in telecommunication networks.

Transfer learning methodologies have also been employed to improve model accuracy in the presence of scarce data. In [91], objective data (measurements deriving from

smart-phone sensors) and subjective data (self-reported questionnaires) were used to model stress behaviour of healthy employees. The labelled data for subjects were scarce. To cope with that, information obtained from other subjects in the study was transferred. The proposed approaches were based on defining a distance among models and using similar models to improve predictions. Given a set of previously learned models along with their respective data, the method either transferred instances from another, close model (sampling based approach) or simply used close models from other subjects (ensemble approach). Results showed that transferring from a few, similar, subjects was better than using more subjects which are not close to the target model. Furthermore, transferring using other models (ensemble approach) was better than transferring instances.

The literature previously outlined indicates that transfer learning could be considered for cases with limited training data, multiple or partially related datasets (one for each participant) or even sparse datasets, to overcome issues related to personalised predictive modelling. Moreover, the literature reports different ML algorithms and transfer learning techniques, such as Bayesian or hierarchical methods, that could be explored to attempt to construct the effective and efficient algorithms for personalised prediction in the presence of sparse and scarce data. However, the majority of studies in this field are conducted in a healthy population and focus on mood detection and classification, and only very few focus on the more challenging problem of long-term forecasting. Studies commonly employ neural network methods long-term forecasting. E.g. Spathis *et al.* [89] used smartphones to acquire a sequence of self-reported mood states over three weeks, by asking users to select a point into a two-dimensional grid the dimensions of which represent "valence" and "arousal". They then trained a multi-task encoder-decoder recurrent neural network to produce a sequence of valence/arousal forecasts (expressed as points on the same grid) for up to 7 days. Their model performed well, though the authors noted performance was less reliable in participants with high mood variability. Similarly, Yu *et al.* [90] used data from the SNAPSHOT study [95], which comprised detailed data from 251 college students, including data from surveys,

mobile phones, wearables and weather information. These were used to define mood, health, and stress scores, on which they compared a series of MTL approaches including regularized linear models and several varieties of neural networks, in next-day, and up to 7-day forecasts. Their findings showed good performance for next-day scenarios; however, even after selecting the best-performing algorithm, there was a significant reduction in accuracy in 7-day forecasts.

One limitation of neural-network based approaches like the above mentioned ones, is that any transfer learning is typically applied to provide an initial estimate of the network's parameters. This is typically then either used as an initialization point for subsequent fine-tuning, or the topmost layers are "frozen", meaning that they are excluded from subsequent training [96]. While this approach can achieve a significant initial speed-up in terms of learning, it is less robust, in that it does not allow for any uncertainty present in the transfer domain to be propagated to the prediction. Expressing the transfer learning component as a prior probability in the context of Bayesian inference methods [18] could potentially allow us to make use of this information. However, this is not necessarily a straightforward thing to do: when dealing with complicated distributions defined in high-dimensional spaces, obtaining posterior parameter estimates expressed in closed form is typically not feasible, as the integrals involved in the inference process tend to be computationally intractable.

3

Non-parametric Transfer Learning based on Bayesian Model Averaging

This chapter introduces a Bayesian Transfer Learning modelling method based on a Markov Chain Monte Carlo sampler and Bayesian Model Averaging (BMA) for dealing with the challenge of building user-specific predictive models able to make predictions of self-reported well-being scores when scarce sporadic observations used for training the model.[1]

Contents

3.1 Introduction

Within the NEVERMIND project, the proposed method for modelling participants and predict their self-reported well-being scores is based on a *Linear Dynamic*

[1] Published as: Eirini Christinaki, Riccardo Poli, and Luca Citi. "Bayesian Transfer Learning for the Prediction of Self-reported Well-being Scores". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2018, pp. 41–44

System (LDS) [97, 98]. The method assumes that the well-being of the user is represented by a state vector, and that its dynamics can be captured by an LDS of the following form:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1) + \mathbf{B}\mathbf{u}(t) + \boldsymbol{\varepsilon}_x(t) \tag{3.1a}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \boldsymbol{\mu}_y + \boldsymbol{\varepsilon}_y(t) \tag{3.1b}$$

where $\mathbf{x}(t) \in \mathbb{R}^{n_\mathbf{x}}$ is the latent state for the model, reflecting the user's underlying state of well-being; $\mathbf{y}(t) \in \mathbb{R}^{n_\mathbf{y}}$ is a vector of observations corresponding to the measurements collected from the user (including biomedical signal features and self-reported well-being scores); $\mathbf{u}(t) \in \mathbb{R}^{n_\mathbf{u}}$ is the input vector (representing external interventions, or influences from the external environment, e.g., weather or day of the week); and $\boldsymbol{\mu}_y$ is the baseline value of the observation vector. In addition, $\boldsymbol{\varepsilon}_x$ and $\boldsymbol{\varepsilon}_y$ represent noise (i.e. uncertainty) over the state and observation vectors, and are assumed to be distributed as $\boldsymbol{\varepsilon}_x(t) \sim \mathcal{N}(0, \mathbf{S}_x)$ and $\boldsymbol{\varepsilon}_y(t) \sim \mathcal{N}(0, \mathbf{S}_y)$ respectively. Finally, the $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are the transition, input and observation matrices of the LDS model, while the parameters of this model hereinafter will be collectively referred to as $\boldsymbol{\theta}$. In the initial work done within the NEVERMIND project [97], model fitting, i.e. the identification of such matrices, was accomplished by using the *Expectation Maximization* (EM) method.

3.1.1 Model inputs

In this work, we use the "NEVERMIND pilot study dataset" which includes data collected from 45 participants enrolled in the pilot study of the NEVERMIND project and the "NEVERMIND clinical trial dataset" which includes data collected from the first 47 participants enrolled in the recently initiated clinical trial of the same project. Both datasets consist of participants aged 18 or older, who have received a diagnosis of a severe somatic disease, including myocardial infarction, breast cancer, prostate cancer, kidney failure and lower limb amputation. Each participant takes part in the study for a period of around 12 weeks from the time of their enrolment (which is independent per participant). The data are collected

in Pisa, Turin and Lisbon [8]. Kidney disease patients were recruited from the *Cisanello University Hospital, University of Pisa, Italy.* Breast and prostate cancer patients were recruited from the following centres within the *Piedmont Oncological Network, at San Luigi Gonzaga University Hospital, Turin, Italy* and *Breast Unit-Oncology Department and Urology Department at Città della Salute e della Scienza University Hospital, Turin, Italy.* Myocardial infarction patients were recruited from the *Cardiology Department at the Santa Maria Hospital, Lisbon.* Patients with lower limb amputations were recruited at the *Rehabilitation Department at the Santa Maria Hospital, Lisbon.* Appropriate informed consent is obtained from the patients in writing. The studies further received ethical approval by the European Commission as a prerequisite of funding approval for the project. Ethical approval was also sought in each of the site where the intervention is implemented (Pisa, Turin and Lisbon) by the local Research Ethics Committees:

- Pisa-Italy: Comitato Etico di Area Vasta Nord Ovest (Comitato Etico Spermentazione Farmaco – CESF)
- Turin-Italy: Turin Ethical Committee of Città della Salute e della Scienza University Hospital and Ethical Committee of San Luigi Gonzaga University Hospital, Orbassano
- Lisbon-Portugal: Ethics Committee of the Medical Academic Centre of the University of Lisbon.

These datasets consist of subjective data in the form of questionnaires, as well as other multimodal data, collected over time from individual subjects via a smartphone and a specialised lightweight sensorised T-shirt. The full datasets include a collection of physiological signals, accelerometer data, and voice recordings; however, for the purposes of this work, we will only consider the three self-reported well-being scales that the user is prompted to provide on a daily basis. The resulting daily scores from each scale will be fed into the LDS model as the observation vector

$\mathbf{y}(t)$. Each scale's numerical input is obtained from the participant via a sliding scale, which takes values from 1.0 to 6.8 (at 0.2 increments), where lower values represent better outcomes. The three scales correspond to the following questions:

- *"How are you feeling today?"* — the *Feel* score: a measure of the participant's subjective assessment of their morning / waking mood;

- *"How was your sleep?"* — the *Sleep* score: a measure of the participant's subjective assessment of sleep quality for the night before; and

- *"How was your day?"* — the *Day* score: a measure of the participant's subjective assessment of the quality of (potentially stressful) events over the course of the day.

Each question is prompted daily and participants may refuse to provide an answer, contributing to the scarce, sporadic nature of the dataset. According to the clinical protocol used in the randomised controlled trial (RCT) [8], human review may be triggered if no significant interaction has occurred for a certain time interval. Participants for whom there were no available data (e.g. patients who had already been enrolled in NEVERMIND, but had not yet started using the system) or have answered less than 10% of the time on average or those that their total data length was less than two weeks, were excluded from the analysis carried out here.

3.1.2 Model parameters

The LDS model (3.1) can describe the current state as an auto-regression of arbitrary order simply by extending the state to include its most recent values, e.g. by writing $\mathbf{x}(t) = [\xi(t-2), \xi(t-1), \xi(t)]^{\mathsf{T}}$ where $\xi(t)$ is the original latent state and $\mathbf{x}(t)$ is the extended one. In this work, we consider a unit-root third-order autoregressive model with a single state capturing all three observations, which can be represented

by the LDS model (3.1) with:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a_0^* & a_2 & a_1 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix}, \ \mathbf{S}_x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & s_1 \end{bmatrix},$$
$$\mathbf{C} = \begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_7 & c_8 & c_9 \\ c_{10} & c_{11} & c_{12} \end{bmatrix}, \ \mathbf{S}_y = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}.$$

The diagonal of the $\mathbf{S}_y$ matrix has been chosen empirically by estimating the variance of the error made by the subjects when using the slider to provide answers to the questionnaires, also accounting for the fact that the scales were quantized. In addition, the observation vector $\mathbf{y}(t)$ includes only the three self-reported well-being scales while the baseline value of the observation vector ($\boldsymbol{\mu}_y$), was set as the mean state. Finally, the a_0^* represent parameters computed via the constraint $a_0^* = 1 - a_1 - a_2$ to ensure the model has a unit root.

While in [97] estimates of the unknown model matrices were obtained using EM, i.e. maximising the likelihood (marginalised over the latent state), here we parametrise them through $\boldsymbol{\theta} = [a_1, a_2, b, c_1, \ldots, c_{12}, s_1]^\intercal$ and follow a Bayesian approach to obtain their posterior probabilities and perform transfer learning, as explained in the following sections.

3.2 Method

The proposed TL technique is a non-parametric method based on an MCMC sampler and BMA for sharing knowledge between personalised models with a focus on providing reliable predictions for a given patient even when scarce and sporadic observations or inconsistent and irregular data is available for that patient. Non-parametric models assume that the data distribution are defined in terms of an infinite-dimensional parameter space. This approach initially requires that we sample from the posterior distribution of the parameters, i.e., our beliefs about the parameters after having seen the data for a given patient. This can be achieved by

using an MCMC sampler, which constructs a Markov chain of samples (i.e. parameter sets), having as their equilibrium distribution the target posterior distribution. The MCMC sampler can use the data from each patient to create his/her own chains while the TL will occur because the model will make predictions about a specific patient under test by combining models trained on other patients ("donor patients") according to how well they fit the patient's under test past observations. As we have seen in Sec. 2.2, this TL method could fall under the category of instance transfer where some samples/instances from the source domains are reused for learning in the target domain by re-weighting them.

3.2.1 Markov Chain Monte Carlo sampler

Bayesian inference offers an alternative to maximum likelihood and allows us to determine the posterior probability of the model parameters given the data. MCMC methods can be used to obtain posterior parameter estimates when these are difficult to express in closed form and works well even for complicated distributions in high-dimensional spaces. Although MCMC tends to be more computationally intensive method than other methods like e.g. variational inference, it provides guarantees of producing (asymptotically) exact samples from the target density [99]. MCMC constructs a Markov-chain having as its equilibrium distribution the target posterior distribution. To sample from the posterior distribution of the parameters (our beliefs about the parameters after having seen the data), in this work, we use the affine invariant ensemble sampler for MCMC (emcee) proposed in [100]. emcee has been chosen as an easy to use, well tested, pure Python module, where the underlying algorithm also has an affine invariance property that allows it to perform equally well under all linear transformations, and therefore be insensitive to covariances among parameters. It is also an ensemble method which relies on multiple walkers (the members of the ensemble) sampling in parallel. For any given walker in the ensemble, their next position is proposed by choosing another walker from the ensemble at random and choose a new position that is a random linear combination of the positions of both walkers. This proposed move is called

"stretch move" since, the proposal is generated by stretching along the straight line connecting the two walkers [100, 101]. However, as we will discuss in Chapter 4 (see Sec. 4.2), emcee method also has disadvantages like facing issues in high-dimensional parameter spaces or with models whose parameters are highly correlated.

According to Bayes Theorem, given a vector of observations $\mathbf{y}$, and a vector of parameters $\boldsymbol{\theta}$, the posterior probability $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ is related to the likelihood term $p(\mathbf{y} \,|\, \boldsymbol{\theta})$ and the prior term $p(\boldsymbol{\theta})$ via:

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \tag{3.2}$$

Therefore, given a way to compute the product $p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$, the ensemble sampler generates random vectors $\boldsymbol{\theta}$ distributed according to $p(\boldsymbol{\theta}|\mathbf{y})$. The likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is in our case the marginal likelihood of the LDS model in Sec. 3.1 marginalised over the latent state $\mathbf{x}$:

$$p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \int_{\mathbf{x}} p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta})\; p(\mathbf{x} \,|\, \boldsymbol{\theta})\; \mathrm{d}\mathbf{x}. \tag{3.3}$$

This likelihood term can be readily obtained from the LDS model using a Kalman filter applied to the participant's data (see [97]). Additionally, we specify a prior probability distribution $p(\boldsymbol{\theta})$ to inform and constrain our model. Specifically, for the model parameters described in Sec. 3.1.2, we place a Gaussian prior over the c_i coefficients and an inverse gamma prior over the non-zero diagonal element s_1 of S_x. We adopt diffuse priors because they express vague or general information so they are dominated by the likelihood function and have minimal effect, relative to the data, on the final inference.

3.2.2 Bayesian Model Averaging

When making a prediction, we want to take into consideration information coming from both the patient under test, as well as more general information available from other patients under a transfer learning framework. Formally, we want to obtain the posterior predictive distribution $p(\widetilde{\mathbf{y}} \,|\, \mathbf{y}, \mathbf{Y}_N)$ for a given patient (without loss of generality we consider the one with index $N{+}1$), where $\widetilde{\mathbf{y}}$ is the desired prediction, $\mathbf{y}$

represents the patient's existing observations, and $\mathbf{Y}_N = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ corresponds to the information coming from all other N "donor" participants' observations. In theory, this expression can be obtained by marginalising $\boldsymbol{\theta}$ out as follows:

$$p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \mathbf{Y}_N) = \int_{\boldsymbol{\theta}} p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \boldsymbol{\theta})\; p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{Y}_N)\; \mathrm{d}\boldsymbol{\theta}. \tag{3.4}$$

Unfortunately, in practice this integral is generally intractable. Alternatively, assuming conditional independence with respect to $\boldsymbol{\theta}$ across data coming from different participants, we can expand equation (3.4) as:

$$\begin{aligned} p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \mathbf{Y}_N) &= \int p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \boldsymbol{\theta})\, \frac{p(\mathbf{y} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \mathbf{Y}_N)}{\int p(\mathbf{y} \,|\, \boldsymbol{\theta}')\, p(\boldsymbol{\theta}' | \mathbf{Y}_N)\, \mathrm{d}\boldsymbol{\theta}'}\, \mathrm{d}\boldsymbol{\theta} \approx \\ &\approx \sum_{k=1}^{K} p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \boldsymbol{\theta}_k)\, \frac{p(\mathbf{y} \,|\, \boldsymbol{\theta}_k)}{\sum_{j=1}^{K} p(\mathbf{y} \,|\, \boldsymbol{\theta}_j)}. \end{aligned} \tag{3.5}$$

with each of the K samples in $\{\boldsymbol{\theta}_k\}_{k=1}^{K}$ distributed according to $p(\boldsymbol{\theta} \,|\, \mathbf{Y}_N)$, as approximated by running the MCMC sampler (described in Sec. 3.2.1) on each of the N participants and then pooling together the resulting samples. Under this scheme, assuming each run creates S samples, we obtain the K vectors of parameters used in equation (3.5) via uniform random sampling from the mixed sample pool of $N \times S$ model parameters. The probabilities $p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \boldsymbol{\theta}_k)$ and $p(\mathbf{y} \,|\, \boldsymbol{\theta}_k)$ are then obtained by using the Kalman filter as described earlier.

The fractional term in the summation shown in equation (3.5) represents the probability that, out of the K models considered, the given model $\boldsymbol{\theta}_k$ generated the observed data $\mathbf{y}$. Therefore, using equation (3.5) to estimate equation (3.4) corresponds to performing BMA [102] over the K candidate models.

Calling $\mu_k(t)$ and $\sigma_k^2(t)$ the mean and variance of the future self-reported well being $\tilde{\mathbf{y}}(t)$ as predicted by the k-th model $\boldsymbol{\theta}_k$ through the Kalman filter, the mean and variance of the Bayesian model-averaged $\tilde{\mathbf{y}}(t)$ are obtained as follows:

$$\mu(t) = \frac{\sum_{k=1}^{K} \mu_k(t)\, p(\mathbf{y} | \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} p(\mathbf{y} | \boldsymbol{\theta}_k)}, \tag{3.6a}$$

$$\sigma^2(t) = \frac{\sum_{k=1}^{K} \left[\sigma_k^2(t) + (\mu_k(t) - \mu(t))^2\right]\, p(\mathbf{y} | \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} p(\mathbf{y} | \boldsymbol{\theta}_k)} \tag{3.6b}$$

The pipeline of our proposed method can be seen in Fig. 3.1.

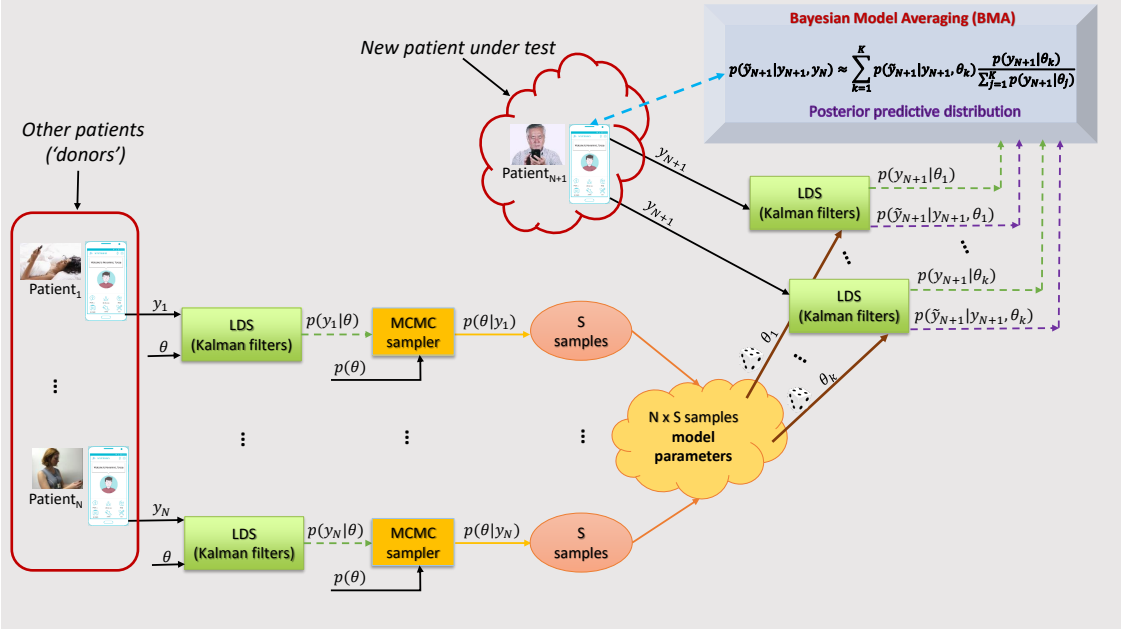


Figure 3.1: Pipeline of the proposed non-parametric TL method.

3.3 Results

In this section we first present the initial evaluation of this approach on a subset of the "NEVERMIND pilot study dataset". The proposed method is evaluated for $K = 1000$ candidate models per participant since here we have a non-parametric method which requires a lot of data-points. We further compare this transfer learning approach with the previously non-transfer method used by the NEVERMIND project and we present the results from this comparison. Finally, the proposed approach is validated on the NEVERMIND clinical trial and the MIMIC II dataset [103] and it is also compared against four standard baselines.

3.3.1 Evaluation on NEVERMIND pilot study dataset

To evaluate the transfer learning model, denoted here as M_{BMA}, we assess its predictive performance and compare it against the non-transfer model that uses the EM method (trained by maximum likelihood) denoted as M_{EM}. Both models

are evaluated using real-world data collected within the NEVERMIND project during the pilot study (details in Sec. 3.1.1).

For the MCMC method, we train a separate ensemble for each of the N participants. Each ensemble comprised 130 walkers that samples our 16-dimensional parameter space for 4,500 iterations, of which the first 1,500 are considered as a "burn-in" period during which all samples are discarded. The fraction of steps accepted for each walker was around 0.37, which is within the suggested range 20%-50% [100].

The corner plot in Fig. 3.2 shows the one- and two-dimensional projections of the samples obtained by MCMC using the actual self-reported well-being data $\mathbf{y}$ from one of our participants. These can be interpreted as sampled estimates of the marginal (diagonal plots):

$$p(\boldsymbol{\theta}_i|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta}_1 \ldots \mathrm{d}\boldsymbol{\theta}_{i-1} \mathrm{d}\boldsymbol{\theta}_{i+1} \ldots \mathrm{d}\boldsymbol{\theta}_n \tag{3.7}$$

and joint (off-diagonal plots) posterior distributions:

$$p(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta}_1 ... \mathrm{d}\boldsymbol{\theta}_{i-1} \mathrm{d}\boldsymbol{\theta}_{i+1} ... \mathrm{d}\boldsymbol{\theta}_{j-1} \mathrm{d}\boldsymbol{\theta}_{j+1} ... \mathrm{d}\boldsymbol{\theta}_n. \tag{3.8}$$

In this section, we present the evaluation results from 17 randomly selected test patients. Initially, for each patient under test, the models are trained with 7 days (1 week) of historic data $\mathbf{y}$ from the same patient to predict the observations $\tilde{\mathbf{y}}$ of the next 7 days (test week). The predictions are then compared with the actual observations for the test week. For comparing and quantifying the prediction ability of each model we use metrics including *accuracy* and *Root Mean Squared Error* (*RMSE*). The *accuracy* is computed considering a prediction as correct if the predicted value is less than 0.5 away from the actual observation. In addition, the actual forecasting error for the predictions of each individual, is assessed with the more traditional *RMSE* calculated as $RMSE = \left[\frac{1}{n}\sum_t (r_t - \mu_t)^2\right]^{\frac{1}{2}}$, where t here corresponds to time-points within the forecasting period, r_t is the corresponding actual value of a well-being score observed at time t and used as a target value for validation, n is the number of targets present when missing values are excluded,

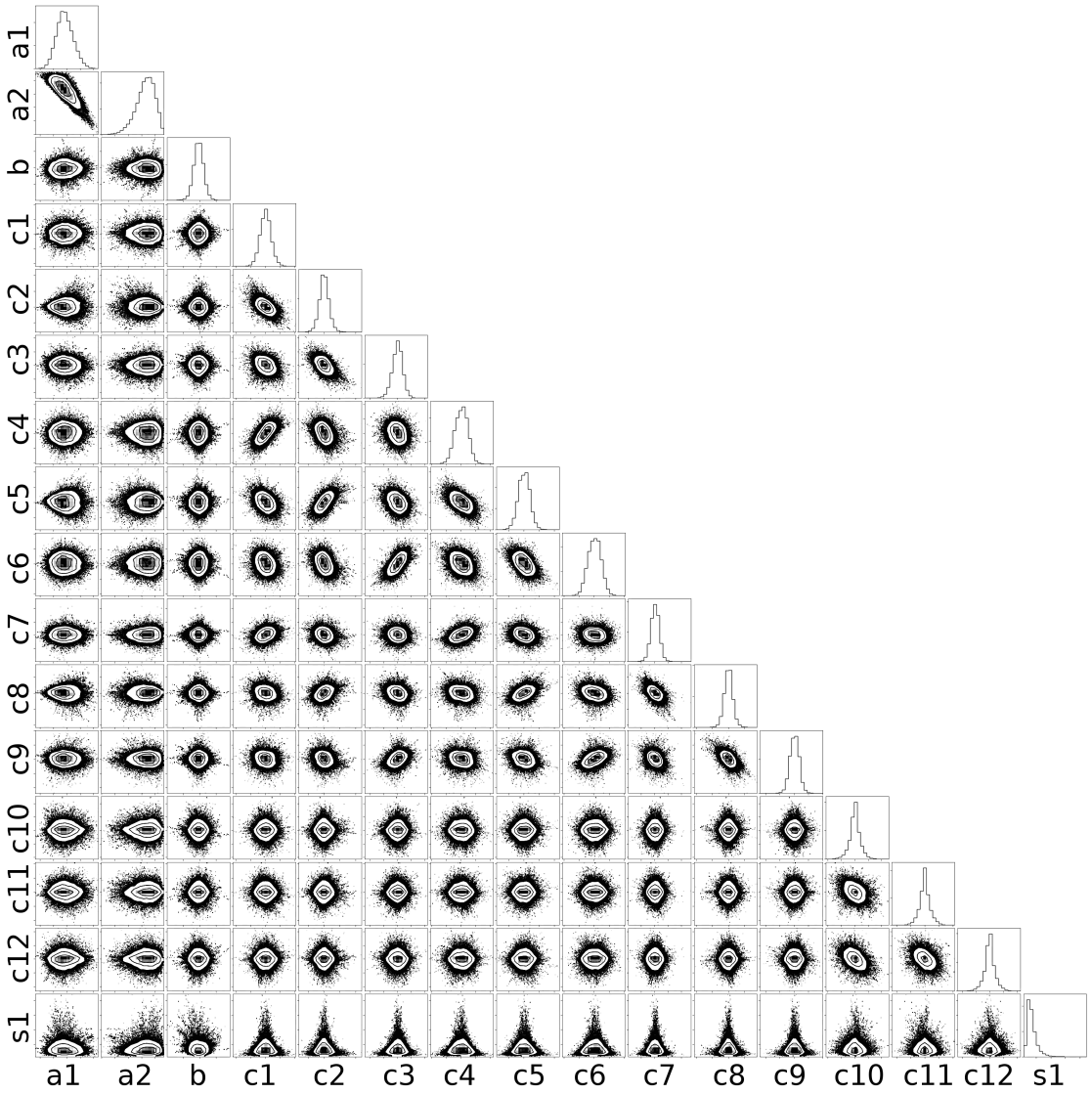


Figure 3.2: Corner plot of the 16 parameters of the model. The histograms along the diagonal presents the marginalized distribution for each parameter independently. The other panels show the marginalized two dimensional distributions (the covariance between parameters).

and where $r_t - \mu_t \stackrel{\text{def}}{=} 0$ when r_t is missing. Note that *RMSE* ignores how accurate our estimates of the prediction variance are.

An indicative example of the predicted mean and the variance as learned by our model along with the sporadic self-reported well-being scores from one of our participants can be seen in Fig. 3.3. The figure illustrates the mean and standard deviation, respectively, of the state predicted by the model according to equations (3.6a) and (3.6b). As can be seen from this figure, our method adequately

deals with the uneven, sporadic data representation in the dataset and performs well in a difficult forecasting scenario that requires to train on previous seven days and predict next seven days. However, this is a very initial result.

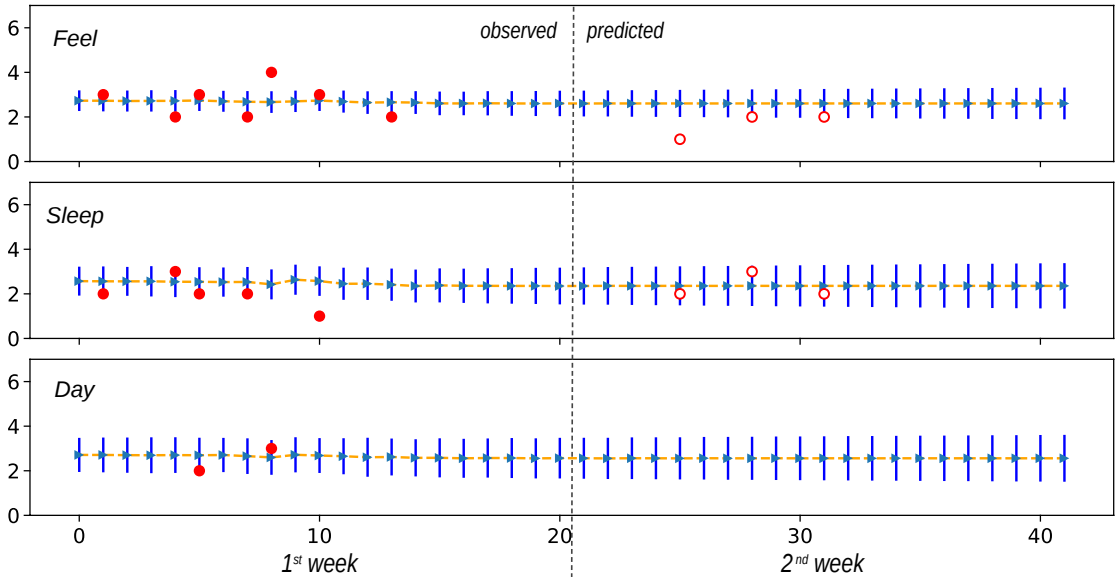


Figure 3.3: Example of self-reported well-being score modelling and prediction (this is a very initial result). The model was trained with one week of data (the leftmost 21 points at 8-hour resolution) and tested on the following week. The solid red circles represent the reported scores that were used by the model, while the empty ones in the second week are only reported for reference. The blue triangles and the associated whiskers represent the mean and standard deviation.

For the same forecasting scenario, we compare the performance of our transfer learning approach (M_{BMA}) against the competing no-transfer learning model (M_{EM}). The results presented in Fig. 3.4 show that the transfer learning model yields the lowest *RMSE* in most cases (for 11 patients out of 17). The overall *RMSE* for this scenario was 0.62 for our method and 0.67 for the no-transfer M_{EM} model. The *accuracy* measured from the average predicted results was 62.74% for the transfer learning approach and 60.55% for the no-transfer M_{EM} method.

To investigate how the predictive performance changes as more data become available, we trained the models for each participant with 14 days (2 weeks), 21 days (3 weeks) and 28 days (4 weeks) of historic data to predict the following 7 days (test week). The average *RMSE* and *accuracy* for the experiments performed

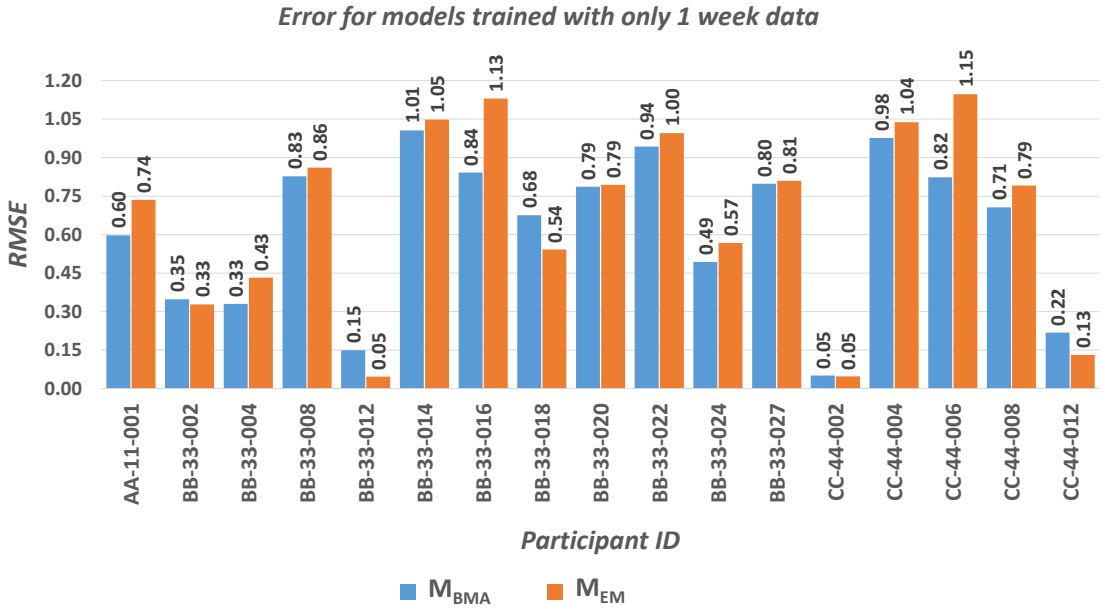


Figure 3.4: RMSE per participant for the transfer and non-transfer prediction models trained with only 1 week data and forecast next week.

Table 3.1: Comparison of prediction results for the transfer (M_{BMA}) and no-transfer learning (M_{EM}) models.

	RMSE		***Accuracy [%]***		
Data length	M_{BMA}	M_{EM}	M_{BMA}	M_{EM}	***participants***
1 week	**0.62**	0.67	**62.74**	60.55	17
2 weeks	**0.61**	0.74	**63.93**	56.62	16
3 weeks	**0.66**	0.77	**52.44**	49.81	11
4 weeks	**0.55**	0.91	**61.45**	34.05	13
Average	**0.61**	0.77	**60.14**	50.26	

Bold values show the best evaluation scores (lowest error, highest accuracy) among the predictive models. Note that there is a different number of "valid" participants per data length because the analysis includes only participants who have data within this period.

with varying training data length are presented in Table 3.1. These results are based only on "valid" participants (i.e. participants who have data within this period) and given the fact that patients are allowed to refuse to answer some or all questions on any particular day, the number of observations present within that time-interval may well differ between patients. The evaluation metrics for the predictive models show that the best scores were obtained using the transfer learning method in all four scenarios.

3.3.2 Validation on NEVERMIND clinical trial data

The previously presented model has been extended by increasing the number of parameters to be estimated. However, the focus again is on performing inferences about the model parameters $\theta \equiv \{A, B, C, S_x, \mu_y\}$, where this time instead of setting μ_y, the baseline value of the observation vector, as the mean state, we parametrize it through θ in order to be able to set this baseline based on information we borrow from other patients:

$$\mu_y = \begin{bmatrix} \mu_{y_1} \\ \mu_{y_2} \\ \mu_{y_3} \end{bmatrix}$$

This configuration has been chosen in order to produce comparable results with the validation of our approach on an external dataset presented in Sec. 3.3.3. After setting a Gaussian prior over the μ_y coefficients, we re-compute an estimate of the posterior probability from which the affine invariant ensemble sampler for MCMC will draw samples. Once we obtain the chain of the samples, we compute the marginalized posterior probability distribution of each parameter and the joint posterior between the parameters. Now, for the MCMC simulation, we specify 150 samplers to sample our 19-dim parameter space, 1500 steps length of "burn-in" period and 4.500 iterations. The average fraction of steps accepted for each walker for this example is 0.3 (suggested acceptance rate 20%-70%).

Since the data from the pilot study and the newly collected data from participants enrolled in the clinical trial of the NEVERMIND project were both collected in a similar way, we follow the same criteria as before for the data cleaning. Thus, we remove from this analysis subjects that have answered less than 10% of the time on average or those that their total data length is less than two weeks. In total, these criteria remove 14 patients out of 47 available at that time thus, the remaining dataset consists of a total of 33 patients.

The predictive performance of the current TL approach, namely the *BMA* approach (M_{BMA} model), is evaluated this time in relation to the following "baseline" models: *a)* A *patient-average prediction* model (M_{A}), *b)* A *population-average* model (M_{P}),

c) A *last-datapoint* model (M_{L}), and *d)* An *ordinary least squares regression* model (M_{R}). Given that the new data arrive incrementally, it is necessary to rebuild each patient's chains on a daily basis in order to keep the models up-to-date, and ensure that all observations available for that participant are being used. It is important to note that, for any given time-interval, given the fact that patients are allowed to refuse to answer some or all questions on any particular day, the number of observations present within that time-interval may well differ between patients.

We further define the following notation. Let:

- L_{tr} be the length of the "training period", i.e., the number of weeks used for training (regardless of the number of actual observations that happen to be contained within), chosen from the set $\{1, 2, ..., 10\}$,

- L_{fc} be the length of the "forecasting period", i.e., the number of future data-points (three per day) to be forecasted, chosen from the set $\{1, 3, 7\}$,

- μ_t and σ_t^2 be the mean and variance of the forecasted prediction at the time-point t,

- r_t be the corresponding *actual* value of a well-being score observed at time t, (which may be missing if no answer was provided), used as a target value for validation.

In these experiments, we evaluate the forecasting scenarios that result from all possible combinations of training and forecasting period-length pairs $\{L_{\text{tr}},\ L_{\text{fc}}\}$. However, for brevity, we only show here a representative subset from these results (more results will be presented in the following chapters), namely: *a)* train for L_{tr}=1 week and predict for L_{fc}=7, and *b)* train for $L_{\text{tr}} \in \{1, 2, 3, 4\}$ and predict for L_{fc}=7. Please, note that the number of participants for which it possible to obtain predictions, depends on the choice of L_{tr} and L_{fc}.

Since our goal is to obtain predictions with an associated measure of uncertainty (i.e., how much we can trust the prediction itself), the quality of our predictive algorithms must be assessed using a measure of accuracy that takes both the mean prediction accuracy and the estimated uncertainty into account.

One such measure is the *Log Likelihood* (LL), which, for a predictive model M_i, is given by

$$LL = \sum_t \log p(r_t \mid \mu_t, \sigma_t^2) = \sum_t \log \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{(r_t-\mu_t)^2}{2\sigma_t^2}} \tag{3.9}$$

where t corresponds to time-points within the forecasting period for which an actual observations is available. The higher the LL measure, the better the probabilistic predictions are. Note also that, while the output of M_{BMA} is not strictly speaking i.i.d Gaussian, for the purposes of obtaining an LL measure, we represent them as i.i.d Gaussians of their respective mean and variance.

While we believe LL is a better measure to account for both mean prediction accuracy and accuracy in the uncertainty around the prediction, we further calculate the actual forecasting error for the predictions of each individual, using again the *RMSE* as we did during the evaluation presented in Sec. 3.3.1.

Furthermore, for any pair of competing models, we calculate a "winning percentage", as a measure of predictive superiority for one model over another. This is computed as:

$$\begin{aligned} \mathsf{wins} &= \mathsf{games} - \mathsf{ties} - \mathsf{losses}, \\ \mathsf{winning\,percentage} &= [\mathsf{wins} + (\mathsf{ties}/2)] \;/\; \mathsf{games}, \end{aligned} \tag{3.10}$$

where games represents the total number of participants for whom it was possible to obtain predictions given a specific $\{L_{\text{tr}}, L_{\text{fc}}\}$ pair, and wins corresponds to the subset of those participants, for whom the model in question performs *better* than its counterpart, with respect to a particular performance measure (i.e. either LL or *RMSE*). Furthermore, we used the exact Wilcoxon Signed-Rank test-statistic [104] to make pairwise comparisons for these methods, effectively investigating the extent

to which the winning percentages represent a genuine and statistically significant improvement, per pair of competing models.

The results presented in Fig. 3.5 show the winning percent based on LL for the M_{BMA} when training with only 1 weeks of past data, and predicting 7 days ahead. As can be seen, the M_{BMA} model scores more wins compared to all of its competitors, when both the accuracy and the uncertainty of the predictions is taken into account (i.e. when using LL as the performance measure). In addition, it scores significantly better (at the 5% level, using a one-tailed hypothesis) compared to three out of the four competitors, but shows no significant difference compared to its M_{P} competitor (the *population-average* model). These results confirm our first hypothesis (H.1), that in the presence of limited person-specific data available for training (i.e. at the early stage of data collection), the transfer learning based prediction model performs significantly better than models which rely only on the limited patient-specific data (i.e. the 'target' domain).

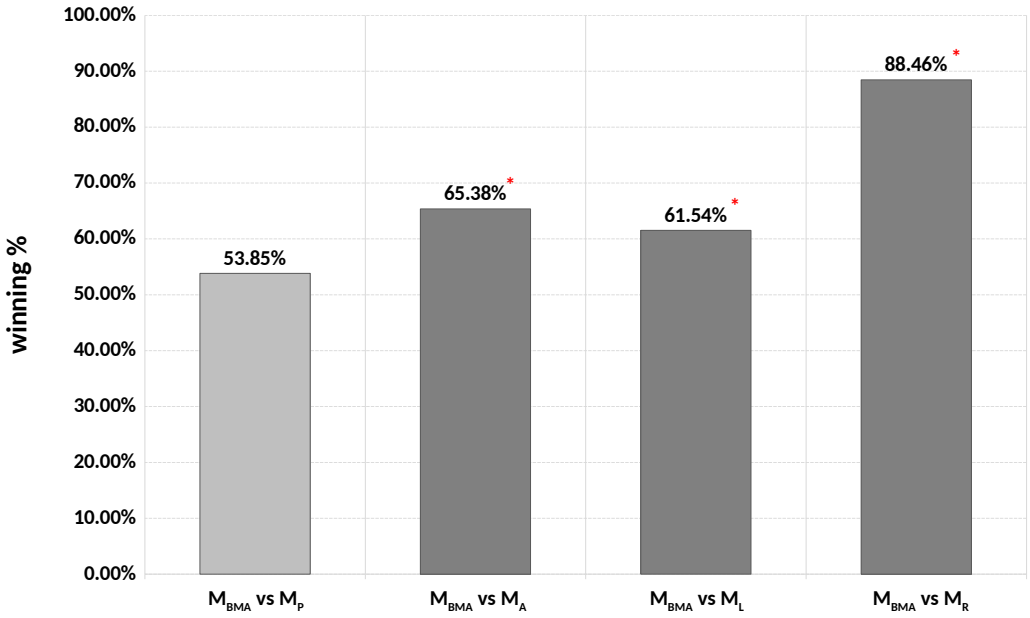


Figure 3.5: Comparison of the winning results based on LL for the M_{BMA} against the four competing models when training with 1 week of past data and predicting 7 days ahead. The * indicates statistical significance.

We further present results for the experiments performed with varying training

Table 3.2: Effect of training period length on performance

LL				
	1-week	*2-weeks*	*3-weeks*	*4-weeks*
M_{BMA}	**-13.08**	**-14.12**	**-7.24**	**-8.45**
M_{P}	-15.75	-16.11	-12.18	-12.53
M_{A}	-48.27	-21.50	-10.95	-8.85
M_{L}	-66.03	-29.36	-13.09	-21.14
M_{R}	-14328.78	-1016.50	-41.16	-257.75
participants	26	25	19	18
RMSE				
	1-week	*2-weeks*	*3-weeks*	*4-weeks*
M_{BMA}	**0.805**	**0.861**	0.608	**0.535**
M_{P}	0.961	0.912	0.856	0.827
M_{A}	0.883	0.862	**0.549**	0.634
M_{L}	0.979	0.912	0.834	0.829
M_{R}	1.838	1.008	1.142	1.044
participants	26	25	19	18

Bold values show the best validation scores (highest likelihood, lowest error) among the baselines and the transfer learning predictive model. Note that there is a different number of "valid" participants per data length because the analysis includes only participants who have data within this period.

data length in order to investigate how the predictive performance changes as more data become available. This analysis has been conducted for training periods $L_{\text{tr}} \in \{1, 2, 3, 4\}$ and forecasting period $L_{\text{fc}} = 7$. Table 3.2 shows the results obtained with respect both to the *LL* and the *RMSE* evaluation measure. The top half of the table shows the mean *LL* values while the bottom half presents the mean *RMSE* values, both obtained over all patients for whom data was available for the corresponding $\{L_{\text{tr}}, L_{\text{fc}}\}$ pair. The values closer to zero represent better performance. These results show that the better the probabilistic predictions were obtained using the transfer learning method in all four scenarios. However, based on the actual error for the predictions (*RMSE*), when we train with 3 weeks of past data, the *patient-average prediction* model (M_{A}) is superior to all other competitors.

3.3.3 Validation on MIMIC II dataset

In this section, we study the accuracy of the proposed method using the openly available MIMIC-II database [105]. The data used for this validation is a subset of

the data used on the PhysioNet/CinC Challenge 2012. The focus of this challenge was to develop methods for patient-specific prediction of in-hospital mortality. The full data used for the challenge consisted of records from 12,000 Intensive Care Unit (ICU) stays. According to [106] *"each record contained general descriptors recorded at the time of admission to the ICU (age, gender, weight, height, and type of ICU) and up to 37 time-series measurements (for example, the diastolic/mean/systolic arterial blood pressure and lab tests) that may be observed (never, once, or more than once) during the first 48 hours after admission. For each time series measurement, the associated time stamp indicating the time elapsed since admission, was also recorded. Two subsets, A and B, each one made of 4,000 of the 12,000 records, were available to the participants"*. In this work, we decided to use the first 72 entries from the set-a with ids 132618 to 132705 used for test and ids 132539 to 132617 used for training. These data can be found online at https://physionet.org/physiobank/database/challenge/2012/. Finally, from our analysis we removed patients that have less than 5 measurements in total or the total data length was less than 3.5 hours. These two criteria removed 1 patient out of 36 from the test set and thus, the remaining dataset consists of a total of 71 patients.

Once more, in order to perform a similar validation to the one previously presented in Sec. 3.3.2 where the NEVERMID dataset was used, we decided out of the 37 measurements included in this dataset, to select *Heart Rate*, *Temperature* and *Urine* to be used as our three observations. These variables were selected as they were among the ones more often observed. Regarding the *Urine*, we take the log due to heavy-tailed distribution [106]. Since these observations have very different ranges, the data pre-processing further includes rescaling them such that their ranges extend from 0 to 1. We have also empirically chosen the non-zero diagonal elements of S_y by estimating the error made by the person taking the measurement in the ICU also accounting the nature of the different observations. Thus, in this case the state variance was set as $S_y = \text{diag}(.008, .013, .014, .001)$. The rest of the approach regarding the MCMC simulation and the BMA is the same as previously and the focus again is on performing inferences about the model

parameters $\theta \equiv \{A, B, C, S_x, \mu_y\}$ and make predictions based on scarce, sporadic observations. The predictive performance of the current TL approach is evaluated again in relation to the same 'baseline' models.

An indicative example of the predicted mean and the variance as learned by our model along with the sparse *ICU observations* from one of the patients can be seen in Fig. 3.6. The figure illustrates the mean and standard deviation, respectively, of the state predicted by the model according to equations (3.6a) and (3.6b). The time interval between two consecutive points is 10 minutes (10 minutes resolution). As can be seen from this figure, our method again adequately deals with the uneven sparse data representation in the dataset and performs well in a difficult forecasting scenario that requires to train on previous three hours and predict next three hours. In particular, we can notice that when there is an interval without observations, the confidence of the prediction made by the model is low while after seeing an observation, the prediction variance shrinks since the model is more confident for the prediction and the mean changes towards the observed direction.

The results presented in Fig. 3.7 show the winning percent based on *LL* for the M_{BMA} when training with only 3 hours of past data, and predicting as far as 3 hours ahead. As can be seen, the M_{BMA} model scores more wins compared to all of its competitors, when both the accuracy and the uncertainty of the predictions is taken into account (i.e. when using LL as the performance measure). In addition, it scores significantly better (at the 5% level, using a one-tailed hypothesis) compared to three out of the four competitors, but shows no significant difference compared to its M_{P} competitor (the *population-average* model). These results are in-line with the results presented in Sec. 3.3.2 and further confirm our first hypothesis (H.1), that in the presence of limited person-specific data available for training (i.e. at the early stage of data collection), the transfer learning based prediction model performs significantly better than models which rely only on the limited patient- specific data (i.e. the 'target' domain).

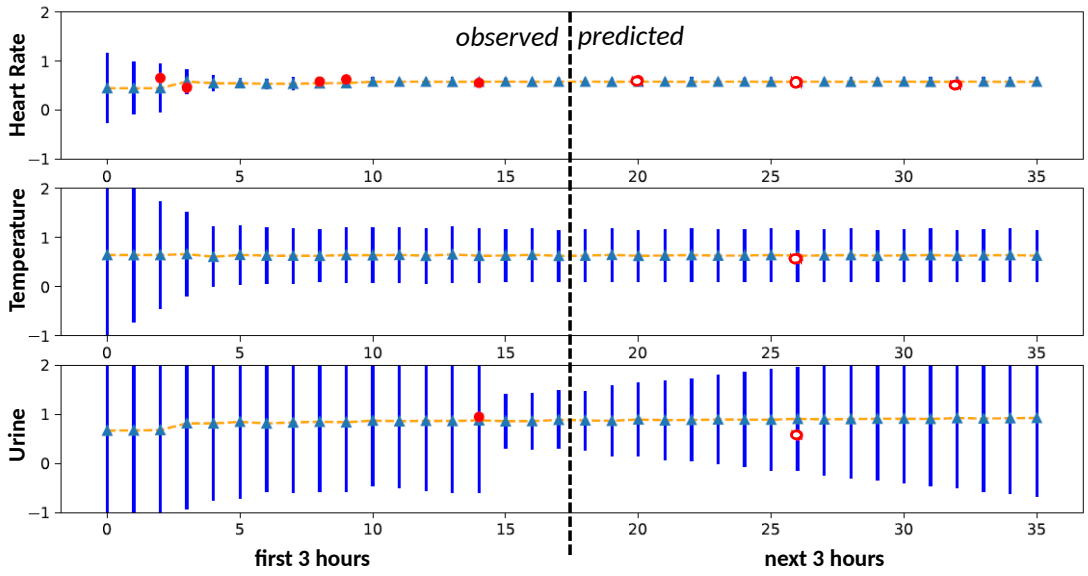


Figure 3.6: Example of *ICU observations'* modelling and prediction. The model was trained with three hours of data (the leftmost 18 points at 10-minutes resolution) and tested on the three hours. The solid red circles represent the reported scores that were used by the model, while the empty ones in the predict part on the right are only reported for reference. The blue triangles and the associated bars represent the mean and standard deviation.

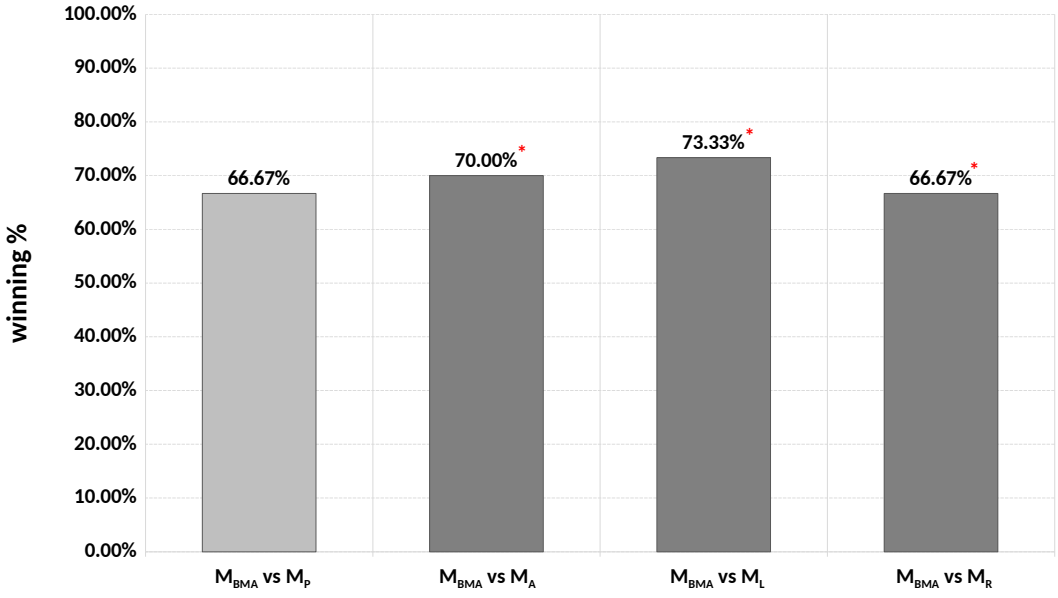


Figure 3.7: Comparison of the winning results based on LL for the M_{BMA} against the four competing models when training with 3 hours of past data and predicting 3 hours ahead. The * indicates statistical significance.

Finally, we present again results for the experiments performed with varying training data length in order to investigate how the predictive performance changes as more data become available. Following the previous notation, the $\{L_{\text{tr}}, L_{\text{fc}}\}$ pair now refers to hours instead of weeks. The analysis was conducted for training periods $L_{\text{tr}} \in \{3, 6, 9, 12\}$ and forecasting periods $L_{\text{fc}} = 3$. Table 3.3 shows the results obtained with respect both to the *LL* and the *RMSE* evaluation measure. The top half of the table shows the mean *LL* values while the bottom half presents the mean *RMSE* values, both obtained over all patients for whom data was available for the corresponding $\{L_{\text{tr}}, L_{\text{fc}}\}$ pair. These results show that the better probabilistic predictions were obtained using the transfer learning method in all four scenarios. However, based on the actual error for the predictions (*RMSE*), when we train with 3 hours of past data, the *patient-average prediction* model (M_{A}) is superior to all other competitors. This could probably be explained due to the nature of the data used and thus, the naive baseline of the average of the patient's observations over the past 3 hours could potentially constitute a good predictor for the next 3 hours.

3.4 Discussion

In this chapter, we proposed a Bayesian TL framework based on an MCMC sampler and a BMA approach that deals with the challenge of building user-specific predictive models able to make predictions in the presence of scarce, sporadic observations. According to the experimental results, the TL model shows an improvement over previous work, which relied on Maximum Likelihood parameter estimation using a standard Expectation Maximisation approach [97]. It also performs better than a population-based model and achieves a significant improvement over training separate models for each participant by using solely their examples. Its overall performance shows the advantage of delivering better results for participants with very few training samples. This method adequately deals with the inconsistent and irregular data representation in the dataset and produces a better suited model for

Table 3.3: Effect of training period length on performance

LL				
	3-hours	*6-hours*	9-hours	12-hours
M_{BMA}	**3.38**	**6.42**	**7.29**	**8.16**
M_{P}	-1.89	1.45	2.20	2.72
M_{A}	-60.21	-13.98	-71.65	-17.07
M_{L}	-165.27	-72.07	-45.44	-51.10
M_{R}	-5895.64	-3449.16	-1913.66	-2399.82
participants	32	34	35	35
RMSE				
	3-hours	*6-hours*	9-hours	12-hours
M_{BMA}	0.145	**0.117**	**0.093**	**0.077**
M_{P}	0.167	0.156	0.153	0.149
M_{A}	**0.136**	0.129	0.119	0.097
M_{L}	0.158	0.146	0.138	0.137
M_{R}	0.186	0.185	0.168	0.119
participants	32	34	35	35

Bold values show the best validation scores (highest likelihood, lowest error) among the baselines and the transfer learning predictive model. Note that there is a different number of "valid" participants per data length because the analysis includes only participants who have measurements within this period.

participants with very few training samples. In addition, our proposal of obtaining probabilistic predictions, expressed as having a mean and a variance, is beneficial since we are producing a prediction and an associated measure of uncertainty over each prediction that allows a measure of how much we can trust the prediction itself.

The previously presented framework is also flexible and can be used in different applications by making the appropriate configurations i.e. adding or removing parameters. For example, in the NEVERMIND we know that we have three questions all of them taking values from 1-6 thus, the baseline value of the observation vector can be initially empirically set to the mean state while in the ICU dataset, the three observations take different values which empirically can not be justified. In such case, our model gives us the flexibility to parametrize the observation vector in order to get this baseline value.

However, such an approach does not exploit the full potential of the MCMC

sampler with respect to estimating the integral in equation (3.4). The proposed BMA approach makes use of MCMC, simply in order to explore and generate samples from $p(\boldsymbol{\theta} \mid \mathbf{Y}_N)$; these samples are then adjusted using $p(\mathbf{y} \mid \boldsymbol{\theta}_k)$ as shown in equation (3.5), to obtain the probability $p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{Y}_N)$ required to approximate equation (3.4). To make fuller use of the potential of the MCMC sampler for the estimation of the integral in equation (3.4), we formulated a new approach, which allows MCMC to explore and sample from $p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{Y}_N)$ directly. This improved framework will be described in the next chapter.

4

Parametric Transfer Learning based on the Fisher divergence

This chapter introduces a personalised Bayesian inference method making use of Transfer Learning in the context of Hamiltonian Monte Carlo sampling, which allows a population prior to be directly represented in the sampling process through the use of the Fisher divergence.[1,2]

Contents

[1] As published in: Eirini Christinaki et al. "Parametric transfer learning based on the fisher divergence for well-being prediction". In: *Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019*. Institute of Electrical and Electronics Engineers Inc., Oct. 2019, pp. 288–295

[2] As published in: Eirini Christinaki et al. "Well-being Forecasting using a Parametric Transfer-Learning method based on the Fisher Divergence and Hamiltonian Monte Carlo". In: *EAI Endorsed Transactions on Bioengineering and Bioinformatics* 1.1 (Oct. 2020)

4.1 Introduction

In the early stages of a patient's enrolment, when data are limited and sporadic, rather than facing the risk of over-fitting with an inappropriately personalised approach, we propose a model that initially relies on a more generalised prediction, borne from prior knowledge and as more personal data become available, the model will slowly starts *transitioning* to more personalised predictions, in an organic, data-driven manner. The prior knowledge required for the initial generalised predictions will be obtained from other participants in the study, and will be used to inform the prediction of a specific patient through an appropriate transfer learning mechanism.

The technique presented in this chapter is a parametric transfer-learning approach based on the *Fisher divergence* (FD), which expresses external information coming from other patients as a prior probability distribution used within a Hamiltonian Monte Carlo framework. Parametric models assume that the data distribution is defined in terms of a finite set of parameters. This approach allows us to create patient-specific models and make informed predictions about a specific individual, even in the early stages of data collection, when data are sporadic, limited, and acquired slowly over time, by leveraging general information available from other patients in the form of priors. This technique allows for a seamless transition from generalised to highly personalised models, as data become gradually available. The effectiveness of that method will be demonstrated in the context of personalised prediction of self-reported well-being scores, using data from the NEVERMIND project [8].

4.1.1 Model inputs

For the purposes of this work, we will use data collected from 112 participants (66 Male, 46 Female) during the NEVERMIND clinical trial [8] for the period between end of December 2017 and end of March 2019. The dataset includes patients who have a diagnosis of myocardial infarction (19 Male, 1 Female), breast

cancer (39 Female), prostate cancer (32 Male), kidney failure (8 Male, 3 Female) and lower limb amputation (7 Male, 3 Female).

In this study, we will again only consider the three self-reported well-being scales, mentioned earlier in Sec. 3.1.1. This time though, each scale's numerical input takes values from 1.0 to 6.0* (at 0.2 increments), where lower values represent better outcomes. Each question is still prompted daily and participants may refuse to provide an answer, contributing to the scarce, sporadic nature of the dataset. Participants for whom there were no available data (e.g. patients who had already been enrolled in NEVERMIND, but had not yet started using the system), or whose total data recording-length is less than two weeks, are excluded from the analysis carried out here.

A comparison of the data distribution of self-report "Day", "Feel" and "Sleep" scores to a normal distribution with mean and standard deviation corresponding to the data can be seen in Fig. 4.1.

4.1.2 Model parameters

As described earlier in Sec. 3.1.2, the latent state of the LDS model (3.1) at any time t can be extended to describe an auto-regression of arbitrary order, simply by extending the state-vector to include its most recent values, e.g. by writing:

$$\mathbf{x}(t) = [\xi(t), \xi(t-1), \xi(t-2), ...]^{\mathsf{T}} \tag{4.1}$$

where $\xi(t)$ is the original, "base" latent state, and $\mathbf{x}(t)$ is the extended one. In this work, we use a simplified version of the LDS model described by (3.1), which ignores the influences from the external environment; in other words, the inputs $\mathbf{u}(t)$ are absent and, therefore, the matrix $\mathbf{B}$ is unused since we are not considering external inputs at this stage. Furthermore, the observation vector $\mathbf{y}(t)$ is limited to

*The upper limit of the scale was 6.8 during the pilot study and the early stages of the clinical trial but this was later capped at 6.0, following interface and user design considerations. This does not affect our data, since in the current dataset approximately only 0.15% of the observations are > 6.

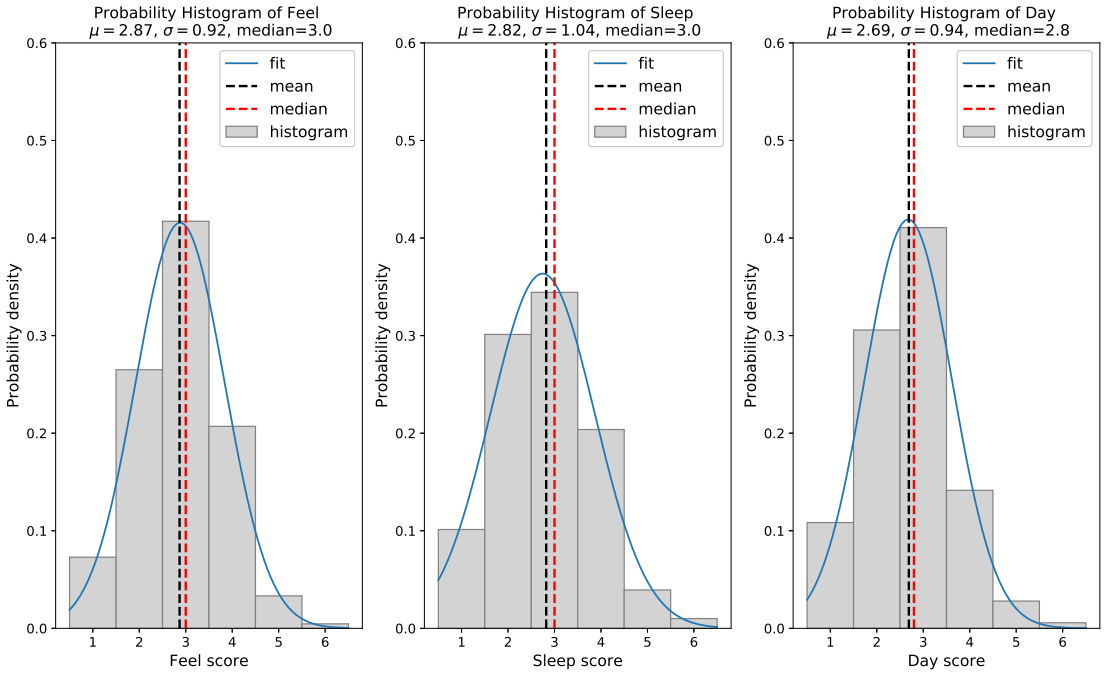


Figure 4.1: Empirical distribution of self-report well-being scales.

reflect only self-reported well-being scales. Given the above, we consider a unit-root, third-order autoregressive model with a single state capturing all three observations, which can be represented by the LDS model in (3.1) with:

$$\mathbf{A} = \begin{bmatrix} a_0^* & a_1 & a_2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix},$$

$$\mathbf{S}_x = \begin{bmatrix} s_x & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{S}_y = \begin{bmatrix} s_y & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_y \end{bmatrix},$$

$\mathbf{B}$=[], $\boldsymbol{\mu}_y$=$[\mu_y, \mu_y, \mu_y]^\intercal$ and $\mathbf{x}(0) = [\xi_1, \xi_1, \xi_1]^\intercal$. The a_0^* represents parameters computed via the constraint $a_0^* = 1 - a_1 - a_2$ to ensure the model has a unit root. The transition matrix $\mathbf{A}$ is totally equivalent to the one presented in Sec. 3.1.2 since if we re-order it, we will still get the same result, since it corresponds to a re-ordering of the state vector. Finally, the estimates of the unknown model matrices are parametrised again through $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = [a_1, a_2, c_{11} \ldots c_{33}, s_x, s_y, \mu_y, \xi_1]$.

4.2 Method

In the previous method [9] presented in Chapter 3, we used the affine invariant ensemble sampler for MCMC (emcee) proposed in [100]. However, there are cases where the affine-invariant ensemble sampler may not perform well or shows unusual and undesirable properties. In particular, when the target density is a multi-modal landscape, the walkers can become stuck in different modes [107] or in lower dimensional subspaces. Furthermore, in high dimensions, the chains can show insufficient convergence and slow mixing, or appear to have converged when they have not [101]. For these reasons, in this work, we decided to utilise an HMC sampler [108], written in the Stan language [109], and more specifically its adaptive extension, the *No-U-Turn Sampler* (NUTS) [110].

4.2.1 Hamiltonian Monte Carlo sampler

The HMC approach exploits Hamiltonian dynamics in order to propose future states in the Markov chain. Effectively, the system simulates the movement of particles over a surface, such that the overall energy of the system is conserved, and can be expressed as the sum of two energy components a "kinetic energy", and a "potential energy" component. The kinetic energy component is generated via a pre-determined probability distribution, and thus plays the part of the proposal component in MCMC, whereas the potential energy component maps directly to the underlying probability distribution we are trying to sample from. Standard HMC algorithms generally depend on, and are sensitive to an appropriate choice of hyperparameters, namely the step-size and number of steps to use during exploration of the domain. The NUTS variant modifies the proposal component of the base algorithm slightly, in that it evolves the initial system both forwards and backwards in time to form a balanced binary tree. The system then stops automatically when the algorithm detects that the sampler has started retracing earlier steps (i.e. making a "U-turn"), thus eliminating the need to define a pre-determined

number of steps. At the same time, the step-size parameter is adapted on the fly, completely eliminating the need to hand-tune HMC.

The HMC algorithm is advantageous, in that it organically makes use of gradient information, enabling it to move faster toward regions of high probability and explore the parameter space more efficiently compared to standard random walks. Consequently, with this sampler we obtain faster convergence in high-dimensional target distributions, while the resulting Markov chain is less correlated. In addition, like in emcee, multiple chains can be allowed to run in parallel. Finally, the use of HMC allows for straightforward scaling up of models to even higher dimensionality and complexity, which could be required in future work.

As previously mentioned in Sec. 3.2.1, according to Bayes Theorem, given a vector of observations $\mathbf{y}$, and a vector of parameters $\boldsymbol{\theta}$, the posterior probability $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ is related to the likelihood term $p(\mathbf{y} \,|\, \boldsymbol{\theta})$ and the prior term $p(\boldsymbol{\theta})$ via equation (3.2). Given a way to compute the product $p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, the HMC sampler allows us to generate K random vectors $\boldsymbol{\theta}_k$, distributed according to $p(\boldsymbol{\theta}|\mathbf{y})$. We can then use this fact to estimate a posterior expectation $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}\Big[h(\boldsymbol{\theta})\Big] = \int h(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta}$ with respect to an arbitrary function $h(\boldsymbol{\theta})$, as the sample average $\frac{1}{K}\sum_{k=1}^{K} h(\boldsymbol{\theta}_k)$, evaluated at the posterior samples $\boldsymbol{\theta}_k$ [18, Sec. 10.1].

Both our previously described in Sec. 3.2.2 TL approach based on BMA [9] and the one described in this section, also presented in [10], are essentially ML techniques for estimating the intractable integral in equation (3.4); however, they do so in a substantially different way.

The approach delineated in this section is a parametric TL method based on the *Fisher divergence*, which can be used to fit a sample of data points to given probabilistic models defined up to a normalisation constant [111–113]. In this approach, the HMC sampler uses the data from each participant directly, to create chains of parameter vectors reflecting the posterior probability distribution of their personalised models. However, in doing so, the MCMC process itself makes use

of a TL component internally, in the sense that the generated chains are obtained based on a modified prior, where this prior accounts for the knowledge available from all other participants, in a manner reminiscent of empirical Bayes approaches. Stated formally, we estimate equation (3.4) as

$$p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \mathbf{Y}_N) \approx \frac{1}{K} \sum_{k=1}^{K} p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \boldsymbol{\theta}_k), \tag{4.2}$$

where the samples $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{Y}_N)$ are obtained via MCMC using the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ from equation (3.3) and a prior $p(\boldsymbol{\theta} \,|\, \mathbf{Y}_N)$ obtained as a mixture of the posterior distributions of the N previous participants:

$$p(\boldsymbol{\theta} \,|\, \mathbf{Y}_N) = \frac{1}{N} \sum_{n}^{N} p(\boldsymbol{\theta} \,|\, \mathbf{y}_n), \tag{4.3}$$

that we approximate in parametric form as:

$$p(\boldsymbol{\theta} \,|\, \mathbf{Y}_N) \approx q_{\boldsymbol{\beta}}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{4.4}$$

where $q_{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is a function governed by a vector of hyperparameters $\boldsymbol{\beta}$. By design, this is not a standard variational inference, it is just a way (a choice) to approximate the likelihood in order to modify a weak prior into a prior that accounts for the information from the other "donor" patients. Note also that while $p(\boldsymbol{\theta} \,|\, \mathbf{Y}_N)$ in equation (4.3) is, from a theoretical point of view, the same as in the BMA approach, the fact that we now consider an approximation of it in parametric form allows us to explore it fully using the HMC sampler, rather than being constrained to using only the $N \times S$ samples previously obtained from the N other participants.

In this work, the specific $q_{\boldsymbol{\beta}}(\boldsymbol{\theta})$ used is an exponentiated quadratic w.r.t. a non-linear mapping of $\boldsymbol{\theta}$:

$$q_{\boldsymbol{\beta}}(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\, g(\boldsymbol{\theta})^{\intercal} \mathbf{Q}_{\boldsymbol{\beta}}\, g(\boldsymbol{\theta}) - \mathbf{v}_{\boldsymbol{\beta}}^{\intercal} g(\boldsymbol{\theta}) \right), \tag{4.5}$$

where $\boldsymbol{\beta} = \mathrm{vec}([\mathbf{Q}_{\boldsymbol{\beta}}, \mathbf{v}_{\boldsymbol{\beta}}])$ and g is a vector function such that its i-th element is $\log(\theta_i)$ if θ_i is a parameter representing a variance (e.g. s_x) and θ_i otherwise. The quadratic parameter $\mathbf{Q}_{\boldsymbol{\beta}}$ is chosen in the set of positive semi-definite matrices so that $q_{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is bounded and $q_{\boldsymbol{\beta}}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ is a proper prior. The hyperparameters $\boldsymbol{\beta}$ leading to the best approximation equation (4.4) can then be found by minimising the Fisher divergence from $q_{\boldsymbol{\beta}}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta} \,|\, \mathbf{Y}_N)$.

4.2.2 Fisher Divergence

The *Fisher divergence* from a distribution $q(\mathbf{x})$ to a distribution $p(\mathbf{x})$, denoted $D_F(p\|q)$, is defined as:

$$D_F(p\|q) = \int_{\mathbf{x}} p(\mathbf{x}) \left\| \frac{\nabla_{\mathbf{x}} p(\mathbf{x})}{p(\mathbf{x})} - \frac{\nabla_{\mathbf{x}} q(\mathbf{x})}{q(\mathbf{x})} \right\|^2 \mathrm{d}\mathbf{x} = \tag{4.6}$$
$$= \int_{\mathbf{x}} p(\mathbf{x}) \left\| \nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x}) \right\|^2 \mathrm{d}\mathbf{x}$$

Much like its better known counterpart — the *Kullback-Leibler divergence* (also known as *relative entropy*) — the Fisher divergence can be understood as an asymmetric measure of distance between a target distribution p, and an approximating distribution q serving as a model for p. In the same manner that the Kullback-Leibler divergence is tightly linked to the concept of entropy (in that it corresponds to the entropy difference between p and q) the definition of the Fisher divergence is similarly tightly linked to the concept of the *Fisher information*, defined[3] by $J(p) = \int p(\mathbf{x}) \left\| \nabla_{\mathbf{x}} \log p(\mathbf{x}) \right\|^2 \mathrm{d}\mathbf{x}$.

A practical disadvantage of the Kullback-Leibler divergence is that, for the result to be meaningful, it requires that both the target, and approximating function be expressed as appropriately *normalised* probability density functions. However, when one only has unnormalised quantities to work with, the computation of an appropriate normalisation constant, whose proper evaluation requires integration over the entire domain of the function, tends to be intractable in the context of high-dimensional problems. By contrast, the Fisher divergence obviates the need for computing such a normalisation constant, since the fractional nature of the calculation with respect to both the target and the approximating function, means that a normalization constant would cancel out from either of those two terms anyway, and therefore lack of appropriate normalization does not affect the final result. This makes the Fisher divergence an advantageous measure of distance to use when dealing with high-dimensional, unnormalised probability

[3] As also noted in [111, 113], while the Fisher information can be defined with respect to *any* parameter, this particular formulation is specifically defined with respect to a hypothetical location parameter.

density functions; this is indeed the case in our TL approach, since both our $q_\beta(\boldsymbol{\theta})$ model, and any distribution represented by the output of an MCMC sampler, will necessarily represent unnormalised quantities.

4.2.2.1 Minimising the Fisher divergence of the mixture distribution

Initially we derive a relationship between the Fisher divergence of a mixture distribution and those of its mixture components. Given a mixture distribution $p(\mathbf{x}) = \sum_i w_i p_i(\mathbf{x})$ with $\sum_i w_i = 1$ and $w_i \geq 0$, the Fisher divergence between p and an approximating distribution q_β can be computed as:

$$\begin{aligned} D_F(p\|q_\beta) &= \int_{\mathbf{x}} p(\mathbf{x}) \left\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log q_\beta(\mathbf{x})\right\|^2 d\mathbf{x} = \\ &= \int_{\mathbf{x}} p(\mathbf{x}) \left\| \frac{\nabla_{\mathbf{x}} p(\mathbf{x})}{p(\mathbf{x})} - \nabla_{\mathbf{x}} \log q_\beta(\mathbf{x}) \right\|^2 d\mathbf{x} = \\ &= \int_{\mathbf{x}} \left\{ \frac{\|\nabla_{\mathbf{x}} p(\mathbf{x})\|^2}{p(\mathbf{x})} - 2\, \nabla_{\mathbf{x}} p(\mathbf{x})^{\mathsf{T}} \nabla_{\mathbf{x}} \log q_\beta(\mathbf{x}) \right. \\ &\qquad \left. + p(\mathbf{x}) \|\nabla_{\mathbf{x}} \log q_\beta(\mathbf{x})\|^2 \right\} d\mathbf{x} \end{aligned} \tag{4.7}$$

Since we have based our choice of the divergence on the original paper by [111] (Equation 2), we follow the same direction therefore the selection of the direction $D_F(p\|q_\beta)$ instead of $D_F(q_\beta\|p)$. Now, if we rearrange the equation (4.7) by adding and subtracting a convenient term, then breaking up the mixture distribution and regrouping, we will obtain:

$$\begin{aligned} D_F(p\|q_\beta) = \int_{\mathbf{x}} \Bigg\{ & \frac{\|\nabla_{\mathbf{x}} p(\mathbf{x})\|^2}{p(\mathbf{x})} - \sum_i w_i \frac{\|\nabla_{\mathbf{x}} p_i(\mathbf{x})\|^2}{p_i(\mathbf{x})} \\ & + \sum_i w_i \left[\frac{\|\nabla_{\mathbf{x}} p_i(\mathbf{x})\|^2}{p_i(\mathbf{x})} - 2\, \nabla_{\mathbf{x}} p_i(\mathbf{x})^{\mathsf{T}} \nabla_{\mathbf{x}} \log q_\beta(\mathbf{x}) \right. \\ & \left. + p_i(\mathbf{x}) \|\nabla_{\mathbf{x}} \log q_\beta(\mathbf{x})\|^2 \right] \Bigg\} d\mathbf{x} = \\ = J(p) - \sum_i w_i\, J(p_i) + \sum_i & w_i\, D_F(p_i\|q_\beta), \end{aligned} \tag{4.8}$$

where $J(p) = \int_{\mathbf{x}} p(\mathbf{x}) \|\nabla_{\mathbf{x}} \log p(\mathbf{x})\|^2 d\mathbf{x}$ is the Fisher information [113] of p while $J(p_i)$ is that of p_i. This is especially useful when looking for the best approximation q_β to a mixture distribution p. Interestingly, since $J(p)$ and $J(p_i)$ do not depend on q_β, the best approximation to the mixture distribution can be computed by

minimising a weighted sum of the Fisher divergence between the approximant and the mixture components:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} D_F(p\|q_{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\beta}} \sum_i w_i \, D_F(p_i\|q_{\boldsymbol{\beta}}). \tag{4.9}$$

Regarding our parametric TL approach, as we showed above, for a mixture distribution, the Fisher divergence of the mixture is simply the weighted sum of the individual divergences from $q_{\boldsymbol{\beta}}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ to the mixture components. From a collection of samples distributed according to $p(\boldsymbol{\theta} \,|\, \mathbf{y}_n)$, obtained by running MCMC separately for each one of the N "prior" patients, we can derive the Fisher divergence for each mixture component as follows:

$$\begin{aligned} F_n(\boldsymbol{\beta}) &= D_F\Big(p(\boldsymbol{\theta}|\mathbf{y}_n) \,\Big\|\, q_{\boldsymbol{\beta}}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})\Big) = \\ &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y}_n) \left\| \nabla_{\boldsymbol{\theta}} \left[\log \frac{p(\mathbf{y}_n|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{y}_n)} \right] \right. \\ &\qquad \left. - \nabla_{\boldsymbol{\theta}} \Big[\log \Big(q_{\boldsymbol{\beta}}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \Big) \Big] \right\|^2 \mathrm{d}\boldsymbol{\theta} \approx \\ &\approx \frac{1}{K} \sum_{k=1}^{K} \Big\| \nabla_{\boldsymbol{\theta}} [\log p(\mathbf{y}_n|\boldsymbol{\theta}_k)] - \nabla_{\boldsymbol{\theta}} [\log q_{\boldsymbol{\beta}}(\boldsymbol{\theta}_k)] \Big\|^2 \end{aligned} \tag{4.10}$$

with $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}|\mathbf{y}_n)$. In this case, $w = 1/n$ because we borrow equally from all n "donor" patients and therefore, it becomes possible to obtain an optimal value $\boldsymbol{\beta}^*$ by solving the following constrained optimisation problem:

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\,:\,\mathbf{Q}_{\boldsymbol{\beta}} \in S_+} \frac{1}{N} \sum_{n=1}^{N} F_n(\boldsymbol{\beta}); \tag{4.11}$$

where S_+ is the set of symmetric positive semi-definite matrices. Given our choice of $q_{\boldsymbol{\beta}}$, the problem in equation (4.11) is an instance of a cone quadratic program, which we solve efficiently using the `cvxopt` library [114].

The prior $q_{\boldsymbol{\beta}^*}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ so obtained is then used alongside the likelihood provided by the LDS model in the context of MCMC, to produce the K samples required for equation (4.2), thus giving rise to our final prediction as the average of K individual prediction components.

The mean and variance of the overall prediction can be obtained at each future time-point by pooling the means and variances of the individual prediction components (i.e. μ_k and σ_k^2 respectively) as follows:

$$\mu(t) = \frac{1}{K}\sum_{k=1}^{K}\mu_k(t), \tag{4.12a}$$

$$\sigma^2(t) = \frac{1}{K}\sum_{k=1}^{K}\left\{\sigma_k^2(t) + \left[\mu_k(t) - \mu(t)\right]^2\right\}. \tag{4.12b}$$

The pipeline the proposed method can be seen in Fig. 4.2.

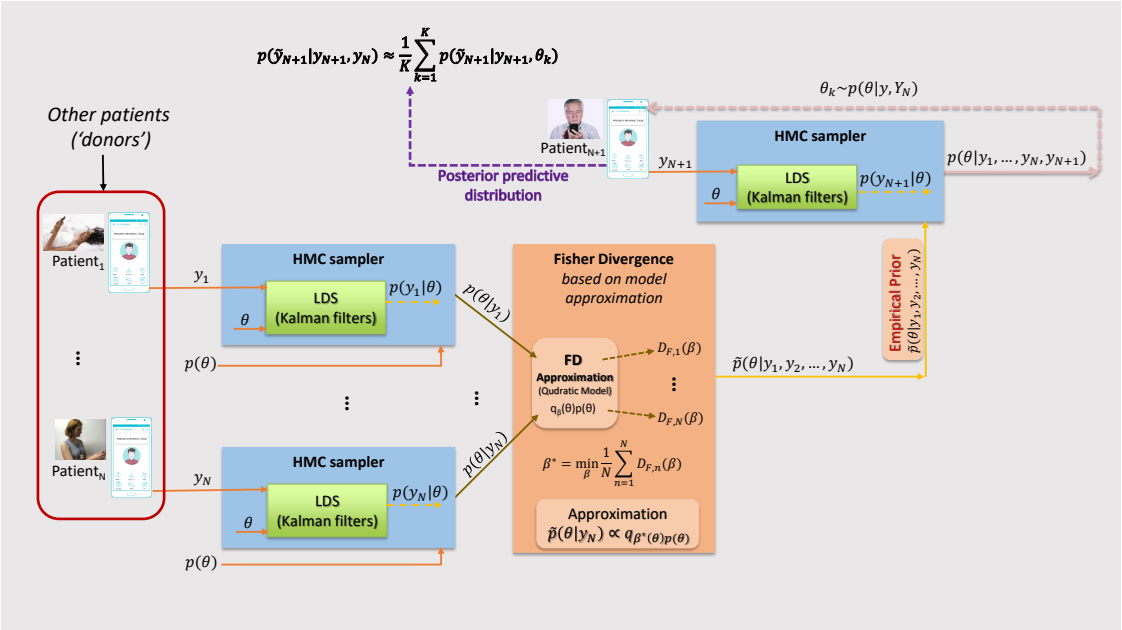


Figure 4.2: Pipeline of the proposed parametric TL method.

4.3 Model instantiation for the NEVERMIND data

This section describes the specific values and implementation of the model defined above, as used in the experiments, as well as the approach used to validate the method on the "NEVERMIND dataset". The predictive performance of the current TL approach, namely the *Fisher-divergence minimization* approach (referred to as the M_{FD} model henceforth), is evaluated in relation to a number of competing

models. In the first instance, we compare this model against the BMA approach (model M_{BMA}) presented in Chapter 3 and used in [9]. To ensure a fairer comparison between M_{FD} and M_{BMA}, the chains for M_{BMA} were created using HMC rather than emcee as previously done in [9]. In addition, we compare M_{FD} against a Maximum A Posteriori (MAP) model (M_{MAP}) and the 4 "baseline" models M_{A}, M_{P}, M_{L}, and M_{R} previously presented in Sec. 3.3.2. The M_{MAP} model was obtained by running Stan in 'optimization' mode instead of "sampling" mode, which uses the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization algorithm under the hood [109]; this directly provides a single "best" estimate for the parameter vector $\boldsymbol{\theta}$, corresponding to the MAP estimate of the posterior distribution, as computed by Stan. This time, since we have a parametric method which does not require as much data as our previous non-parametric method, the proposed approach is evaluated for $K = 1000$ candidate models in total as opposed to $K = 1000$ candidate models per "donor" participant used in our previous method in Sec. 3.3 which is computationally very expensive.

Regarding the priors on model parameters, we have chosen for the "donor" model priors to use weakly informative priors expressing vague or general information, while the "recipient" priors are a combination of these "donor" model priors and the transfer. This has the effect that model selection is primarily driven by the likelihood function, such that in the presence of adequate data, the specific choice of prior has a minimal effect on the final inference, relative to the data. Specifically, with regard to the parameters described in Sec. 4.1.2, we place a diffuse Gaussian prior over elements $a_1, a_2 \sim \mathcal{N}(0, 0.5^2)$ of the transition matrix $\mathbf{A}$, and over the coefficients $c_i \sim \mathcal{N}(0, 1)$ of the observation matrix $\mathbf{C}$. A diffuse Gaussian prior is also placed over the initial state vector $\mathbf{x}(0)$ where $\xi_1 \sim \mathcal{N}(1, 2)$; this distribution was centred away from zero to break the symmetry of the problem and reduce the occurrence of multiple equivalent modes. We further place an inverse gamma prior over the non-zero diagonal element of the state-noise matrix $\mathbf{S}_x$, as $s_x \sim \Gamma^{-1}(\alpha, \beta)$ with shape parameter α=2 and scale parameter $\beta \approx 0.06$. Small values of α lead to wide distributions and in particular α=2 corresponds to a prior with infinite

variance, thus allowing the inference mechanism to explore values of s_x as large as needed. The mode of the inverse gamma distribution is given by $s_x^* = \beta(\alpha+1)^{-1}$. We also set the baseline value of the observation vector to the fixed value μ_y=3, and the parameter s_y to the fixed value of 0.04 (i.e. 0.2^2); the latter was chosen empirically, by estimating the variance of the error made by the subjects when they provide answers to the questionnaires, given the fact that they use a slider in order to do so, and that this results in the scales being quantised at a resolution of 0.2.

The HMC sampler was set to compute Markov chains using 8 walkers working in parallel, such that each sample corresponds to a vector $\boldsymbol{\theta}$ consisting of the scalar parameters described in Sec. 4.1.2 (i.e. a_1, s_x, etc). Each walker was set to create 275 samples, where the first 150 obtained samples were discarded as "burn-in", leaving 125 representative samples per chain. The individual chains generated from each walker were then combined into a single larger chain, having a total of 1000 samples. Please note that the HMC for "donors" and "recievers" is identical (i.e. number of chains, etc) with the only exception of the q_beta additional prior.

We further monitor the convergence of the chains, by computing the *potential scale reduction factor on split chains*, typically referred to more concisely as the 'split-$\hat{R}$' measure [18, Sec. 11.4]. The split-$\hat{R}$ provides a measure of convergence and mixing quality of the chains in an MCMC simulation, which can be used to gain insight into the rate and degree of convergence, as well as in terms of detecting non-stationarity, allowing for better evaluation of the underlying algorithms. We also obtain the log-posterior density (denoted by the '`lp__`' variable in Stan) and summary-statistics for each model parameter, including means, standard deviations (SD) and various quantiles computed from the draws. The summary also reports the Monte Carlo Standard Errors (SE_{mean}), and the effective sample sizes (n_{eff}). The Monte Carlo Standard Error is the uncertainty about a statistic in the sample due to sampling error; the smaller the standard error, the closer the mean estimate of the posterior draws of the parameter is expected to be to the true value. The effective sample size, n_{eff}, measures the amount by which autocorrelation in samples increases uncertainty

(standard errors) relative to an independent sample; if the samples are independent, the effective sample-size equals the actual sample-size. It is particularly important in terms of gauging the reliability of the split-$\hat{R}$ measure, as a small n_{eff} can lead to unreliable values for $\hat{R}$. Table 4.1 presents an indicative example of a summary for the parameters of interest, as estimated from a collection of samples corresponding to one of the participants in the study. The results show that all values for the split-$\hat{R}$ are approximately 1.0 (above 0.9 and below 1.1) and n_{eff} is well above the minimum recommended value of 100 effective samples per chain [115], indicating that chains had mixed well and the model had successfully converged.

Table 4.1: Summary of results using stan for the parameters of interest estimated by the samples for a single participant

Parameter	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a_{01}	0.0863	0.0029	0.0871	-0.0770	0.0278	0.0856	0.1426	0.2599	882.9309	1.0010
a_{02}	0.0032	0.0061	0.1497	-0.3080	-0.0944	0.0076	0.1004	0.2788	604.7551	0.9982
a_{03}	0.1771	0.0144	0.3475	-0.5423	-0.0506	0.1996	0.4243	0.8121	585.1530	1.0006
a_{04}	0.0006	0.0027	0.0855	-0.1759	-0.0538	0.0027	0.0603	0.1637	1019.4277	1.0004
a_{05}	0.1319	0.0054	0.1451	-0.1602	0.0338	0.1338	0.2388	0.3981	717.8639	1.0029
a_{11}	0.0536	0.0041	0.1454	-0.2280	-0.0340	0.0480	0.1435	0.3494	1257.8595	1.0007
a_{12}	-0.0532	0.0079	0.2459	-0.5414	-0.2140	-0.0448	0.1095	0.4344	959.2048	1.0040
a_{13}	0.5468	0.0124	0.3718	-0.1838	0.2800	0.5573	0.7887	1.2924	892.5546	1.0028
a_{14}	0.5410	0.0056	0.1582	0.2074	0.4403	0.5527	0.6531	0.8261	802.2508	1.0018
a_{15}	-0.0049	0.0079	0.2330	-0.4417	-0.1745	-0.0086	0.1596	0.4523	861.1409	1.0042
a_{21}	0.0718	0.0046	0.1517	-0.2087	-0.0301	0.0621	0.1736	0.3867	1081.4613	1.0019
a_{22}	-0.2019	0.0107	0.2570	-0.6839	-0.3797	-0.2204	-0.0287	0.3285	573.6596	1.0013
a_{23}	0.3631	0.0119	0.3585	-0.3241	0.1276	0.3621	0.6134	1.0868	908.1137	0.9985
a_{24}	-0.1932	0.0049	0.1449	-0.4598	-0.2937	-0.1999	-0.0907	0.0972	887.2103	0.9997
a_{25}	-0.1025	0.0065	0.1899	-0.4588	-0.2386	-0.1038	0.0286	0.2722	847.3869	1.0032
$S_{x_{00}}$	0.2353	0.0032	0.0917	0.0819	0.1731	0.2233	0.2862	0.4436	811.2320	1.0041
$S_{x_{11}}$	1.3516	0.0138	0.4259	0.7136	1.0479	1.2784	1.5717	2.3702	949.6742	0.9984
$S_{x_{22}}$	0.5590	0.0118	0.3038	0.0943	0.3412	0.5125	0.7272	1.3054	665.3888	1.0022
ξ_1	1.0116	0.0381	1.0741	-1.5606	0.4478	1.1017	1.7275	2.8530	796.3929	1.0065
ξ_2	1.2758	0.0478	1.5653	-2.1297	0.2680	1.4053	2.3020	4.2206	1073.7744	0.9990
ξ_3	0.7028	0.0703	1.8407	-3.2101	-0.5245	0.7903	1.9692	4.0550	685.7082	1.0085
$lp\_\_$	-262.1961	0.1446	3.0299	-268.7097	-264.1561	-261.8284	-259.9745	-257.0868	438.9218	1.0004

Note: Rows correspond to model parameters, and columns to the various summary metrics. **mean** denotes the posterior mean, **se_mean** denotes the Monte Carlo standard error, and **sd** denotes the posterior standard deviation. The numbers 2.5%, 25%, 50%, 75%, and 97.5% denote quantiles. **n_eff** denotes the effective sample size, and **Rhat** denotes the split-$\hat{R}$ statistic.

4.3.1 Model output

The output of the model at each timepoint is a 3-dimensional probability distribution expressing a probabilistic prediction for the values of the three questions involved, i.e. the *Feel*, *Sleep*, and *Day* scores. For a different value of the length-of-training hyperparameter L_{tr}, a different model output is obtained over both the training and forecasting period. For visualisation purposes, we graph the individual questions independently as three separate graphs, each reporting score as a function of time t.

Fig. 4.3 shows a typical example of the means and variances of the model's probabilistic outputs per timepoint as learned by our model, along with the sporadic self-reported well-being scores for one of our participants, with L_{tr}=1 and L_{fc}=7. In the figure, the mean prediction is represented as a dashed line, and the uncertainty around the prediction is indicated by a shaded $\pm\sigma$ area around the mean. As expected, most (but not all) reported scores fall within the shaded region, even in the forecast phase, where, however, as expected σ grows progressively bigger due to the absence of inputs to the LDS. The *LL* (see equation (3.9)) and *RMSE* measures (see Sec. 3.3.2) corresponding to the model outputs over the timepoints in Fig. 4.3 were -4.41 and 0.5447, respectively.

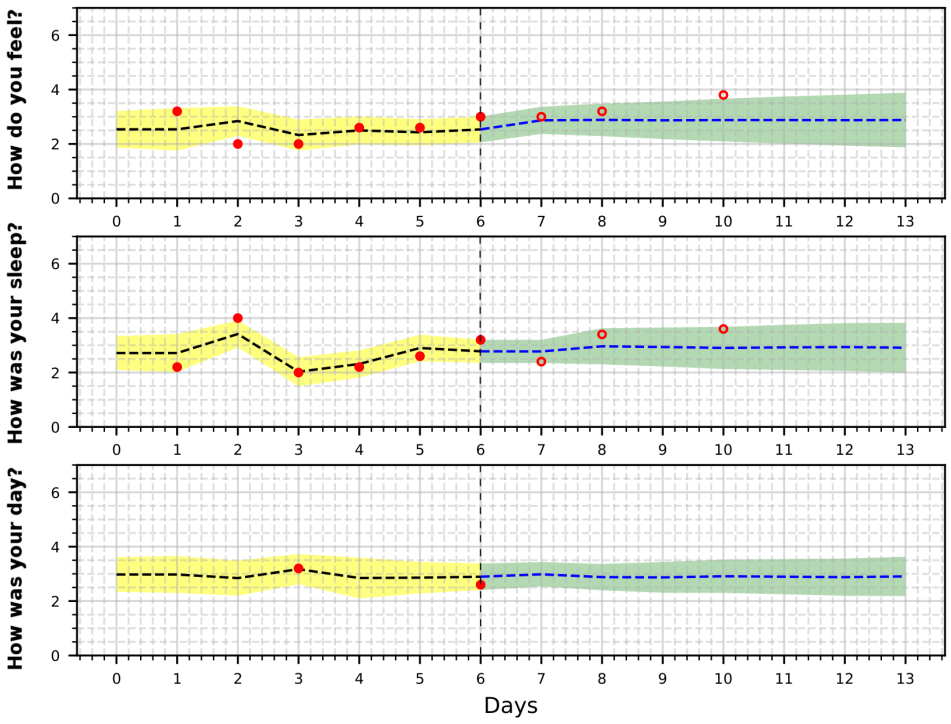


Figure 4.3: Example of self-reported well-being score modelling and prediction. The dashed vertical lines mark the last time point available to the model for training and visually separate past observations from future predictions. The solid red circles mark the reported scores that were used by the model, while the empty ones are reported to visually assess the prediction accuracy. The dashed black and blue lines and the associated yellow and green shadows represent the mean and standard deviation, respectively, of the distribution of the model outputs.

4.4 Results

In this chapter, we will initially present the results from the comparison of the performance of the proposed method against the competing models and the evaluation of our method when different number of previous days are used for training as well as when we forecast a different number of future days.

4.4.1 Comparison against competing models

We compared the performance of the M_{FD} method against the competing models outlined in Sec. 4.3 in two ways: *a)* by analysing the distribution of performance differences directly; *b)* by analysing "winning percentages" as per equation (3.10). Both analyses were performed using *LL* and *RMSE* measures, separately. Furthermore, we used again the exact Wilcoxon Signed-Rank test-statistic [104] to make pairwise comparisons for the methods.

A representative example of the first type of analysis can be seen in Fig. 4.4 for the *LL* differences and Fig. 4.5 for the *RMSE* differences between models. The results were obtained for L_{tr}=3 and L_{fc}=7, which was the most conservative choice for comparing M_{FD} and M_{BMA} (more on this in Sec. 4.4.2). The box and whisker graph plots show the median, interquartile range, and extreme cases of *LL* differences between the transfer learning model M_{FD} and its six competitors. It is clear that, for both performance measures, M_{FD} performs better across the board. Also, among all other competitors, M_{BMA} is the one with the least spread in terms of pairwise differences over all patients. Similar results were obtained with other values of L_{tr} and L_{fc}.

Figs. 4.6 and 4.7 look at the same predictions (obtained for L_{tr}=3 and L_{fc}=7) using the second type of analysis. These results show that the M_{FD} model scores significantly more wins (at the 5% level, using a one-tailed hypothesis) compared to all of its competitors, when both the accuracy and the uncertainty of the predictions is taken into account (i.e. when using *LL* as the performance measure). When only the *RMSE* is used, M_{FD} scores significantly better than four out of the six

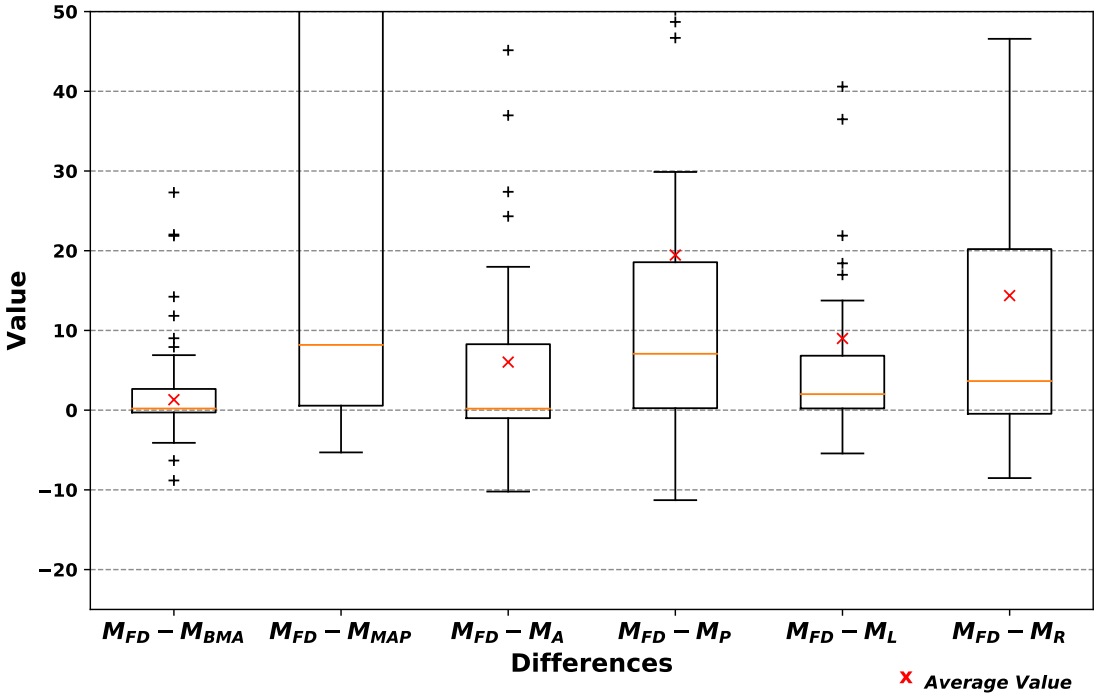


Figure 4.4: Box and whisker graph plots showing median, interquartile range, and extreme cases of *LL* differences when training with 3 weeks of past data and predicting 7 days ahead. Values *above* 0 represent cases where M_{FD} is better than its competitors.

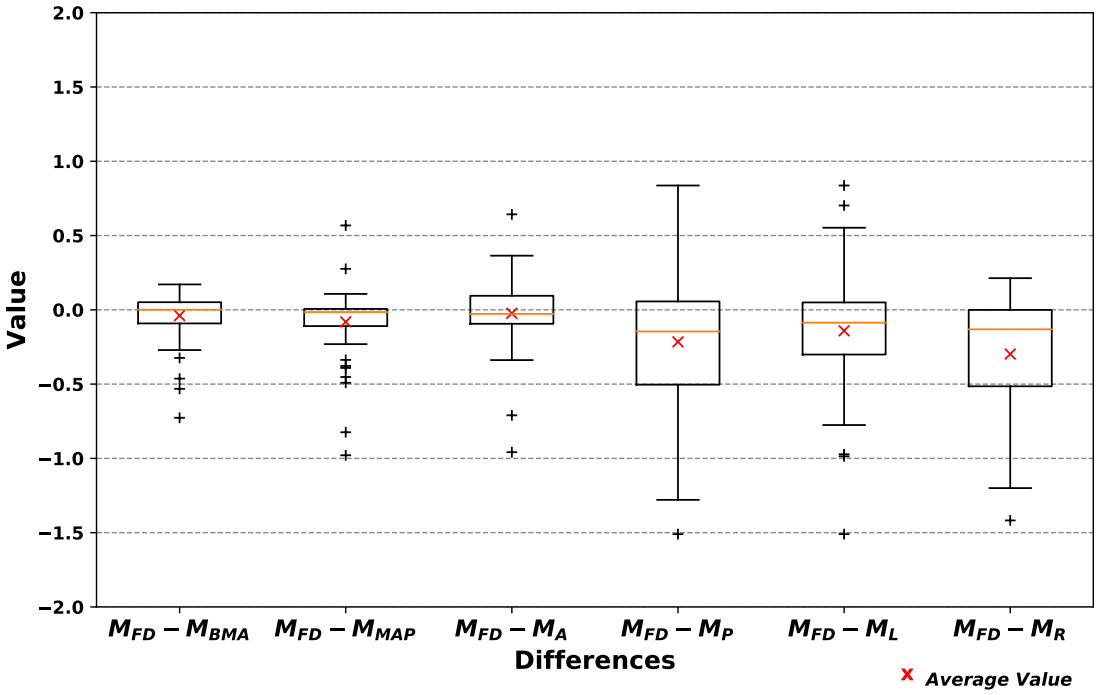


Figure 4.5: Box and whisker graph plots showing median, interquartile range, and extreme cases of *RMSE* differences when training with 3 weeks of past data and predicting 7 days ahead. Values *below* 0 represent cases where M_{FD} is better than its competitors.

competitors, but shows no significant difference compared to its M_{BMA} and M_{A} competitors. Although the winning percentages does not seem to represent a genuine and statistically significant improvement for these two pair of competing models, it still is on par which is something that we could expect for longer periods of training specifically for the non-transfer methods.

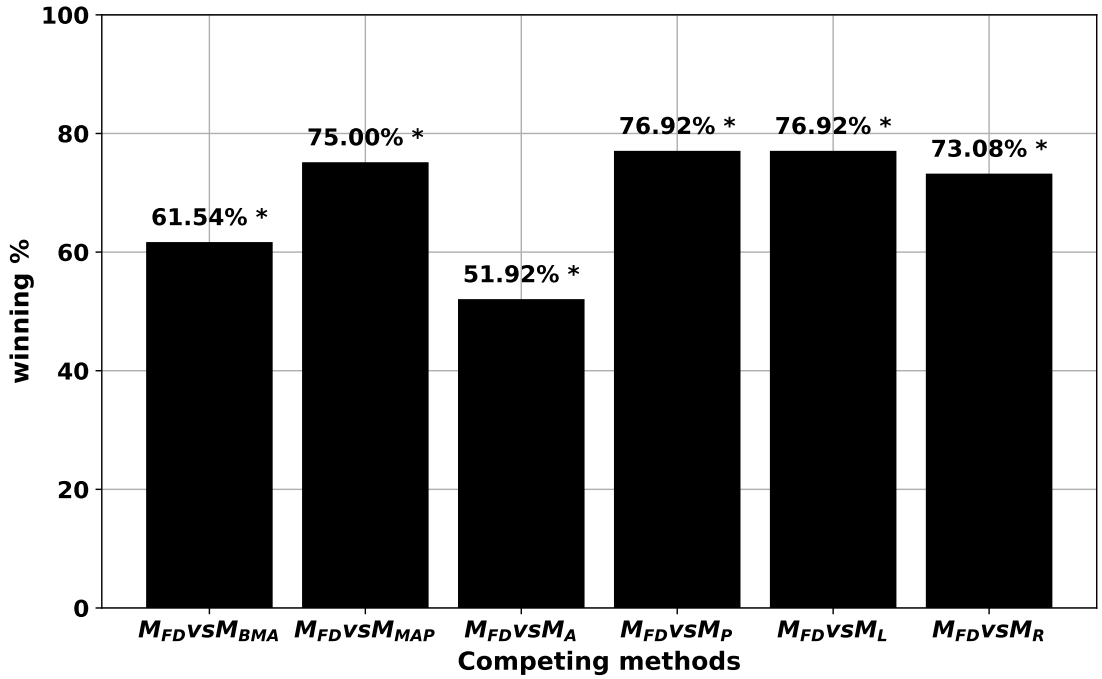


Figure 4.6: Comparison of the winning results based on LL for the M_{FD} against all competing models when training with 3 weeks of past data and predicting 7 days ahead. The * indicates statistical significance.

For shorter training period, when L_{tr}=1 and L_{fc}=7, the analysis of the 'winning percentages' as presented in Figs. 4.8 and 4.9 shows that the M_{FD} model scores significantly more wins (at the 5% level, using a one-tailed hypothesis) compared to all of its competitors, when both the accuracy and the uncertainty of the predictions is taken into account (i.e. when using LL as the performance measure). When only the $RMSE$ is used, M_{FD} scores significantly better than five out of the six competitors, but shows no significant difference compared to its M_{BMA} competitor. Overall, the results are in-line with our expectation for the transfer learning methods where we anticipate that for fewer weeks of training data when only a limited amount

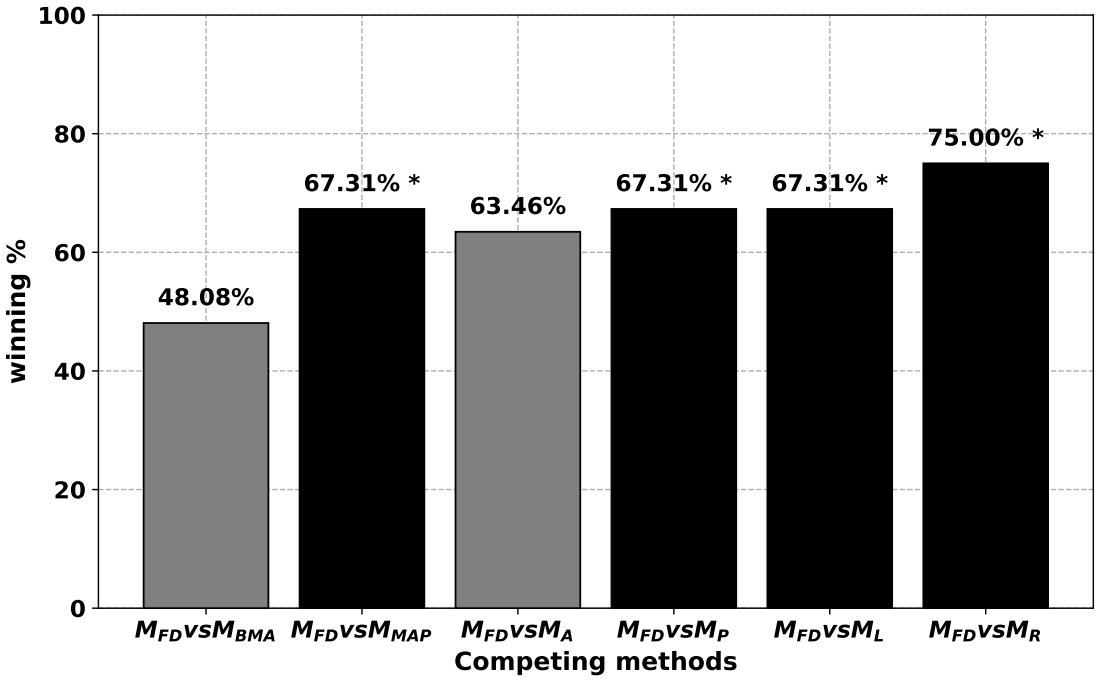


Figure 4.7: Comparison of the winning results based on *RMSE* for the M_{FD} against all competing models when training with 3 weeks of past data and predicting 7 days ahead. The * indicates statistical significance.

of patient-specific data are available for training, transfer learning approaches will be significantly better than non-transfer learning methods and virtually identical, if not still significant, when more weeks of patient-specific training data are available. The effect of the training and testing period length on the performance of the transfer learning methods will be further explored in the next section.

4.4.2 Effect of training / testing period length on performance

One would generally expect that increasing training-period length would improve performance over all models and that predictions further away from the last training time-point would diminish in accuracy. An analysis related to our second hypothesis (H.2) was therefore conducted to confirm this, for training periods $L_{\text{tr}} \in \{1, 3, 7\}$ weeks and forecasting periods $L_{\text{fc}} \in \{1, 3, 7\}$ days. Table 4.2 shows the results obtained with respect to the LL evaluation measure[4]. The top half of the table shows the median LL

[4] The corresponding table for *RMSE* is not shown as it was very similar.

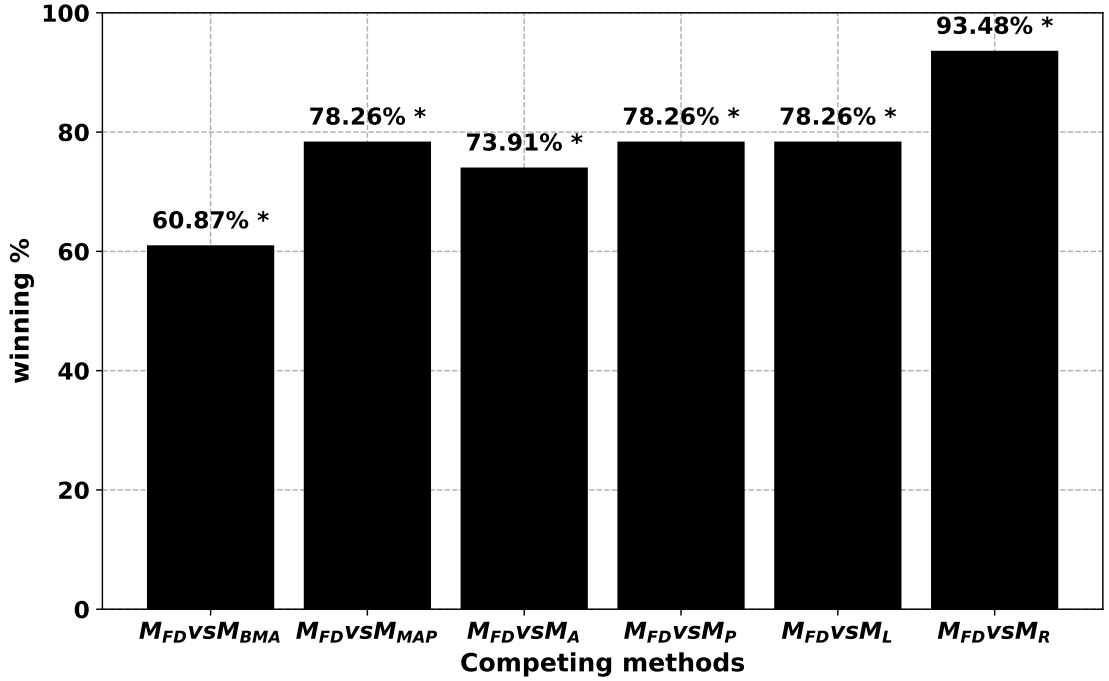


Figure 4.8: Comparison of the winning results based on LL for the M_{FD} against all competing models when training with 1 weeks of past data and predicting 7 days ahead. The * indicates statistical significance.

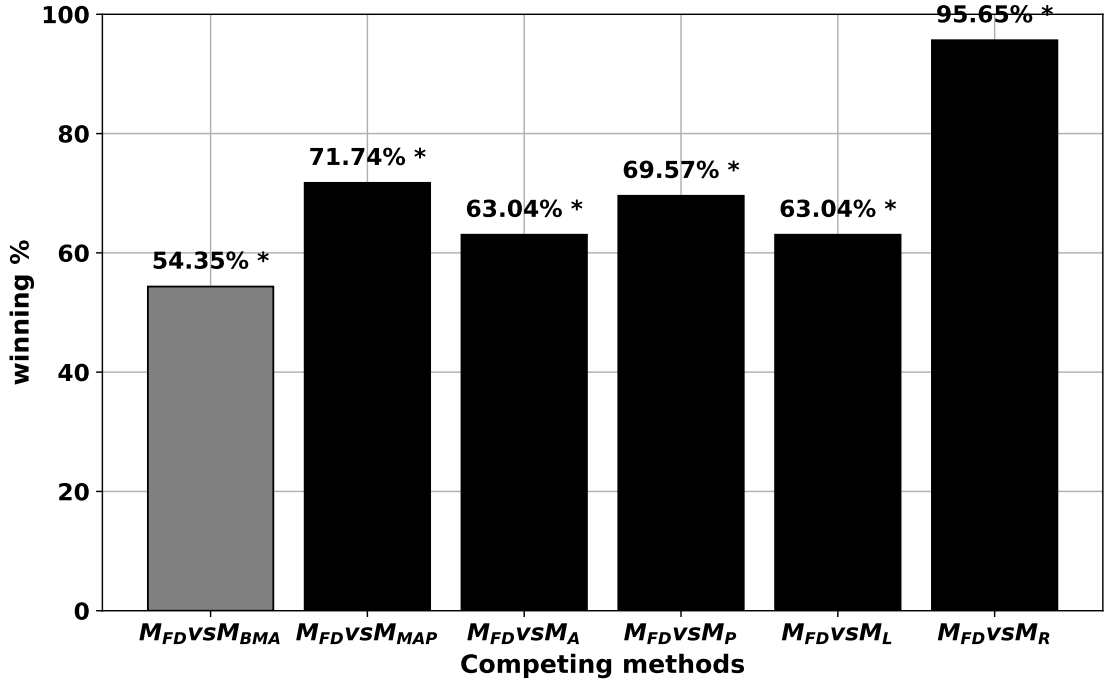


Figure 4.9: Comparison of the winning results based on $RMSE$ for the M_{FD} against all competing models when training with 1 weeks of past data and predicting 7 days ahead. The * indicates statistical significance.

Table 4.2: Medians of LL (top) and of LL-differences (bottom) for various durations of training and forecast periods.

Model Type	$L_{\text{tr}}=1$			$L_{\text{tr}}=3$			$L_{\text{tr}}=7$		
	L_{fc}=1	L_{fc}=3	L_{fc}=7	L_{fc}=1	L_{fc}=3	L_{fc}=7	L_{fc}=1	L_{fc}=3	L_{fc}=7
M_{FD}	-0.15	-2.12	-5.07	-0.76	-2.40	-5.54	-0.72	-1.09	-3.20
M_{BMA}	-1.23	-2.63	-4.87	-2.07	-3.59	-6.66	-1.23	-2.15	-3.61
M_{MAP}	0.49	-8.25	-26.60	-1.44	-5.80	-17.57	-1.41	-3.10	-6.32
M_{A}	-0.98	-3.67	-8.40	-1.80	-2.60	-11.03	-1.11	-1.40	-3.45
M_{P}	-5.97	-8.02	-14.80	-7.61	-7.93	-13.85	-2.14	-4.45	-12.80
M_{L}	-2.07	-4.14	-10.90	-1.95	-4.11	-10.03	-1.56	-3.37	-6.59
M_{R}	-1.33	-7.87	-33.61	-1.74	-4.13	-14.73	-1.14	-2.08	-3.76
Models Compared									
M_{FD} vs M_{BMA}	0.88	0.60	0.30	0.71	0.25	0.20	0.58	0.57	0.69
M_{FD} vs M_{MAP}	-0.75	4.28	22.96	0.95	1.78	8.18	0.17	0.16	0.40
M_{FD} vs M_{A}	0.54	0.64	2.27	0.58	0.05	0.19	0.50	0.23	-0.12
M_{FD} vs M_{P}	5.35	4.24	8.75	5.77	4.95	7.07	1.07	2.87	4.49
M_{FD} vs M_{L}	1.66	1.83	4.70	0.89	0.50	2.01	0.73	1.56	1.32
M_{FD} vs M_{R}	1.73	6.95	26.70	0.29	0.24	3.65	0.29	0.09	0.09
participants	**24**	**37**	**46**	**24**	**40**	**52**	**27**	**40**	**48**

Note that there is a different number of "valid" participants per data length because the analysis includes only participants who have data within this period.

values, obtained over all patients for whom data was available for the corresponding $\{L_{\text{tr}}, L_{\text{fc}}\}$ pair; values closer to zero represent better performance. The bottom half shows pairwise median differences with respect to M_{FD} versus competing models. Note that because of the nature of these intervals, each $\{L_{\text{tr}}, L_{\text{fc}}, \}$ pair will consist of a different number of "valid" participants (i.e. participants who have data within this period), and therefore it is important to note that the above medians are calculated over different sets of patients. The last row shows the number of such valid participants per $\{L_{\text{tr}}, L_{\text{fc}}\}$ pair. Also note that, pairwise median differences are generally not equivalent to pairwise differences between medians.

As shown in the table, some of the methods under comparison initially struggle to make acceptable 7-day predictions, when only one week or three weeks of data is available to them (e.g. M_{MAP}, M_{P} and M_{R}); it is not until 7 weeks of training that

most models can predict one or three days ahead with reasonable accuracy. By contrast, looking at M_{FD}, we see that, not only is it able to make predictions from week one, but it can even make 7-day predictions with remarkable stability for any number of training weeks. Similarly as shown in the bottom half of Table 4.2, we can see that M_{FD} is superior to most methods in most conditions (with the median LL difference being negative only in 2 out of 54 comparisons). In fact, focusing on the M_{FD} vs M_{BMA} row, we see that the new method is also superior to its predecessor, under all scenarios considered; the choice L_{tr}=3 and L_{fc}=7 corresponding to the graphs shown earlier in Secs. 4.3.1 and 4.4.1 showing a relatively narrow interquartile range between the two, was simply selected on the basis that it represented the worst-case scenario for M_{FD} in relation to M_{BMA}.

4.5 Discussion

In this chapter, we proposed a parametric transfer learning approach based on the Fisher divergence in the context of HMC sampling and Bayesian inference that deals with the challenge of building user-specific predictive models able to make predictions in the presence of scarce, sporadic observations. This approach makes it possible to create patient-specific models and make useful predictions of self-reported well-being scores, even when the data available for initial training are sporadic and limited, such that training is performed incrementally as more data become slowly available over time. The approach allows us to make informed predictions even in the early stages of data collection, by leveraging external information coming from other patients, in the form of a prior used within a MCMC process.

We demonstrated this approach on data obtained by the NEVERMIND clinical trial, and measured its performance against previous work (e.g. the BMA method introduced in Sec. 3.2 and in [9]), and a number of baseline approaches. Our results show that this approach yields a significant improvement over its competitors, and is particularly useful in difficult training/forecasting scenarios, e.g. when one

requires a distant, patient-specific forecast, with only a limited initial amount of patient-specific data available for training. These results confirm again our first hypothesis (H.1), that in the presence of limited person-specific data available for training (i.e. at the early stage of data collection), the transfer learning based prediction model performs significantly better than models which rely only on the limited patient- specific data (i.e. the 'target' domain). Our results are also supportive of the hypothesis H.2 that the effect of transfer learning on predictive performance will be maximal when the amount of person-specific data available for training is minimal, while as person-specific data availability increases, the relative contribution of the transfer learning component to the overall accuracy will diminish. This is clearly demonstrated when the models are compared on the most difficult forecasting scenario that requires the models to predict 7 days ahead. However, our second hypothesis (H.2) will be further tested in our next chapter.

Finally, as shown by the performance of transfer learning, the background patient population acting as the source domain for the transfer learning component is informative. However, a limitation of our approach is that we made no attempt to quantify, or investigate ways in which this background knowledge could be made more informative. Thus, in the next chapter the focus will be on investigating whether applying preprocessing strategies that promote individuals in the population that are known to be similar in some way to the person being modelled, enhance transfer learning. At the same time, regarding the LDS model, we will explore a three-state representation which will allow allow for bidirectional interaction between the three individual states and more flexibility and expressive power to the model.

5

Transfer Learning modulated by similarity

This chapter introduces a similarity-based transfer learning approach mostly incorporating highly relevant information from the source domain by biasing population sampling towards participants having similar characteristics, based on their emotional profiles, to the patient being modelled.

Contents

5.1 Introduction

We previously discussed two TL approaches, a non-parametric transfer learning approach based on an MCMC sampler and BMA [9] (see Chapter 3), as well as a parametric transfer learning approach based on an HMC sampler and minimizing the Fisher Divergence [10, 11] (see Chapter 4). These models make predictions about a specific patient by leveraging general information available from other patients. However, these methods are likely to encounter difficulties when transfer is not mutually beneficial: for instance when participants are sufficiently dissimilar or

their affective mood changes over time. This is something that a prediction method should take into account when borrowing information from other patients.

In this section, we will present an approach that extends and complements the work of previous chapters by modulating how much information is transferred from the source population, based on a similarity between patients. This method not only accounts for individual differences but also benefits from the data of similar patients by performing a "soft" clustering (where a given patient can belong to many clusters and also, belonging to a cluster is not binary but gradual) to identify groups of similar individuals in order to better modulate the transfer. In the methods mentioned above, we transferred information under the assumption of uniformly distributed patients. Here we use self-reported personality information (which can be consider quite stable over time) in order to assign probabilities of a patient being used for transfer based on the similarity with the target patient. In other words, the main idea is to use the output of a personality questionnaire to assess the similarity between a new patient and each one of the other patients he/she will borrow information from and transfer less or more based on this similarity. The similarity can be transformed into a probability which can be used to weight the information or the models we transfer.

5.1.1 Model inputs

For the purposes of this work, we used the dataset collected from 182 participants enrolled in the NEVERMIND randomised controlled trial [8] in the period between December 2017 and October 2019. As mentioned in previous chapters, the data were collected in Pisa, Turin and Lisbon however, in this chapter the data are only from Turin and Pisa, since the personality information is missing for the other site. The experiments where approved by local ethical committees and all participants had signed an informed consent form. Here, we again only consider the three self-reported well-being scales that the patient is prompted to provide on a daily basis. Patients may refuse to provide an answer, contributing to the scarce, sporadic nature of the dataset. All participants were adults aged 18 or

older, and had received a diagnosis of a severe somatic disease including myocardial infarction, breast cancer, prostate cancer, kidney failure or lower limb amputation. Participants that had no data (i.e., patients recently enrolled within the period being considered but had not started using the system yet) were excluded from the analysis carried out here, leaving 141 patients.

We also used data coming from a pre-enrolment personality questionnaire [116] that all our patients in Turin answered. This was the 60-items (10 per dimension) Davidson's Emotional Styles Questionnaire (ESQ) [116], an earlier version of the Davidson's ESQ presented in [117], which aims to capture the dimensions of Emotional Style. The questionnaire returns a score in 6 dimensions (*Outlook*, *Resilience*, *Social Intuition*, *Self-Awareness*, *Sensitivity to Context* and *Attention*). Davidson's ESQ also stands on its own as an integrative measure of healthy emotionality and can be used as an assessment tool to assess where each patient falls with regard to each dimension. Here we use the output scores of ESQ in order to find the similarity among patients, which then use together with our previously proposed methods to modulate how mush information we will transfer.

5.1.2 Model parameters

In NEVERMIND, we propose to model patients and predict their self-reported well-being scores using an LDS model (3.1) [97]. This LDS model can describe the current state as an auto-regression of arbitrary order simply by extending the state to include its most recent values, e.g. equation (4.1). In Chapters 3 and 4, the latent state $\mathbf{x}$ was one-dimensional and the output $\mathbf{y}$ was a three-dimensional vector, reflecting the three measurements used: the subjective morning mood, the subjective sleep quality and the subjective end-of-day mood. The latent state $\mathbf{x}$ was assumed to correspond to a state that represents in some manner the degree of the patient's underlying mental well-being. Thus, the above model effectively assumed that a patient's well-being depends only on the degree of their previous well-being, plus any explicit interventions that have occurred, but not on the patient's underlying state of sleep quality or mood, as these are considered purely manifestations of the

patient's own degree of well-being, rather than factors contributing to it. While this may be a reasonable approximation, clinical experience tells us that in the case of underlying sleep and mood states, such a uni-directional model with respect to depression is an over-simplistic representation of the relationship between them, and that the relationship is more likely to be bi-directional [118]. Therefore, for the purpose of the NEVERMIND project, a decision was made to adapt the model slightly, so as to reflect this bi-directionality between the "mental well-being" latent state, and the underlying "sleep" and "mood" latent states, as well as their role in terms of producing the subjective measurements which we then obtained. Here, the output measurements vector $\mathbf{y}$ remains the same as before but the new latent state $\mathbf{x}$ is extended from the one-dimensional state (of arbitrary order) described in equation (4.1), to a three-dimensional one. In the extended model each dimension directly corresponds to collected observations of interest, such as the perceived "waking/morning mood", "quality of sleep" and "end-of-day stress" as subjectively measured by the "feel", "sleep" and "day" questionnaire. Within the NEVERMIND project, the modified LDS model with the multi-state extension further includes a "depression" latent state corresponding to the collected observations from the PHQ-9 score and an optional additional general "well-being" latent state which does not assume any direct correspondence to any of the collected measurements. However, for the purpose of this thesis, we adopted the model with one state for each of the three observations only. The three-dimensional extension of the LDS model, representing the three latent states can be described mathematically as follows:

$$\mathbf{x}(t) \equiv [\xi_0(t), \xi_0(t-1), \ldots, \xi_1(t), \xi_1(t-1), \ldots, \xi_2(t), \xi_2(t-1), \ldots]^{\mathrm{T}} \tag{5.1}$$

where the sizes of the parameter matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ from equations (3.1a) and (3.1b) are adapted appropriately.

The expressive power of the above model is increased compared to the previous one in the following ways:

- The patient's overall mental well-being, sleep quality, and mood throughout the day will depend on both the nature of these individual states from the day before, as well as any (bi-directional) interactions between them (as defined by state transition matrix $\mathbf{A}$).
- The effectiveness of an intervention (as determined by the input matrix $\mathbf{B}$) can now be estimated with respect to its separate effects over all the above states.
- The subjective answers given by patients with respect to sleep and mood are interpreted to be a direct manifestation of their respective underlying states, as modulated by their mutual interaction with the other states, including that corresponding to overall mental wellness.

At the point of updating the latent states through observations (i.e. the update step describe in equation (3.1b)), the parameter matrix $\mathbf{C}$ (governing how observations adjust prior knowledge of the latent states) is constrained in such a manner that observations relating to sleep, waking mood, and day quality only contribute to a single state each. In this way, we ensure that the three added latent states to the model have biological relevance, since they are made to reflect the underlying latent states that give rise to their corresponding observations rather than some other weighted combination of such biological states.

Here, we consider a unit-root, second-order autoregressive model, which can be represented by the LDS model in (3.1) with:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{I}_3 & \mathbf{0}_{3x3} \end{bmatrix}, \mathbf{A}_{11} = \begin{bmatrix} a_0^* & a_{01} & a_{02} \\ a_1^* & a_{11} & a_{12} \\ a_2^* & a_{21} & a_{22} \end{bmatrix}, \mathbf{A}_{12} = \begin{bmatrix} a_{03} & a_{04} & a_{05} \\ a_{13} & a_{14} & a_{15} \\ a_{23} & a_{24} & a_{25} \end{bmatrix},$$

where variables $a_{j,k}$ represent adjustable parameters, and a_i^* represent parameters computed via the constraint $a_i^* = 1 - \sum_n a_{i,n}$ where $n = 6$, to ensure the model has a unit root,

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0}_{3\times3} \end{bmatrix},$$

where the effective identity matrix ensures a one-to-one relationship between the specific latent states and their questionnaires at time t,

$$\mathbf{S}_x = \begin{bmatrix} \mathbf{s}_x & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \end{bmatrix},$$

where $\mathbf{s}_x$=diag$(s_{x_{00}}, s_{x_{11}}, s_{x_{22}})$. Finally $\mathbf{S}_0$=diag$(s_{00}, s_{11}, ..., s_{55})$, where the variance s_{00}=s_{11}=...=s_{55}=0.04, $\mathbf{S}_y$=$[0.04, 0.04, 0.04]^\intercal$, $\mathbf{x}(0) = [\xi_1, \xi_2, \xi_3, 0, 0, 0]^\intercal$ and $\boldsymbol{\mu}_y$=$[3, 3, 3]^\intercal$. The estimates of the unknown model matrices are again parametrised through $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = [a_{01}, \ldots, a_{05}, a_{11}, \ldots, a_{15}, a_{21}, \ldots, a_{25}, s_{x_{00}}, s_{x_{11}}, s_{x_{22}}, \xi_1, \xi_3, \xi_3]^\intercal$.

5.2 Method

As we previoulsy described in Sec. 5.1.1, the 60-items (10 per dimension) Davidson's emotional style questionnaire returns a score in 6 dimensions called *Outlook*, *Resilience*, *Social Intuition*, *Self-Awareness*, *Sensitivity to Context* and *Attention*.

Let $Q \equiv \{q_1, q_2, ..., q_N\}$ denote the set of N patients. For each patient $q_i \in Q$, the returned scores $\boldsymbol{s}^{(i)}$=$[s_1^{(i)}, s_2^{(i)}, \cdots s_6^{(i)}]^T$ are used in order to find a distance between the vector of scores $\boldsymbol{s}^{(i)}$ of, i.e., a "donor" patient from the study and the vector scores $\boldsymbol{s}^{(j)}$ coming from another patient $q_j \in \{Q \setminus q_i\}$, i.e., a new patient under test, recently enrolled in the study. According to [117], the *Outlook* and *Resilience* dimensions are highly correlated. Thus, for measuring the similarity between patients, we use the Mahalanobis distance rather than Euclidean, since it takes into account these correlations in the data. In general, the Mahalanobis distance is an effective multivariate distance metric that measures the distance between a point and the centre of a multivariate normal distribution. This distance measure is also scale-invariant, i.e. does not dependent on the scale of measurements. However, Mahalanobis distance requires the knowledge of the covariance matrix which, in our case, is the covariance matrix of the different dimensions in the general population. The matrix Δ of the mutual Mahalanobis distance between two pair of score vectors

$\boldsymbol{s}^{(i)}$ and $\boldsymbol{s}^{(j)}$ is the dissimilarity measure and is defined as:

$$\Delta_{ij} = \sqrt{(\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(j)})^T \Sigma^{-1} (\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(j)})}. \tag{5.2}$$

The 6×6 covariance matrix Σ is computed by using the column vector $\boldsymbol{\sigma}$ and a correlation matrix R coming from the general population[1] as follow:

$$\Sigma = (\boldsymbol{\sigma}\boldsymbol{\sigma}^T) \odot R, \tag{5.3}$$

where $\odot$ denotes the Hadamard product. Then, we convert the dissimilarity into similarity by relating the distance to probability through the chi-square (χ^2) distribution. It turns out that the squared Mahalanobis distance with covariance Σ between two multivariate normal samples with the same covariance Σ, follows a chi-square distribution with ν degrees of freedom (in our case, based on our sample estimates, $\nu = 6$) as shown below:

$$\begin{aligned} \Delta_{ij}^2 &= (\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(j)})^T \Sigma^{-1} (\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(j)}) \\ &= (\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(j)})^T \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(j)}) \\ &= (\Sigma^{-\frac{1}{2}} (\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(j)}))^T (\Sigma^{-\frac{1}{2}} (\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(j)})) \end{aligned} \tag{5.4}$$

where Σ^{-1} is the inverse of the covariance matrix Σ and $\Sigma^{-\frac{1}{2}}$ is the inverse square root of the covariance matrix ("Mahalanobis whitening"). If we denote the whitened vector D as $\Sigma^{-\frac{1}{2}}(\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(j)})$ then $D^T D = \|D\|^2 = \sum_{k=1}^{\nu} D_k^2 \sim \chi_\nu^2$.

In order to define a similarity measure that can be used to cluster patients and assign weights based on the closeness of each patient under test to other "donor" patients, we transform the Δ_{ij}^2 values into probabilities using the χ^2 cumulative probability distribution. For a random variable $X \sim \chi_\nu^2$ with ν degrees of freedom and evaluated at x, the cumulative distribution function F denoting the probability $p(X \leq x)$ when $x \geq 0$ is:

$$\begin{aligned} F_{\chi_\nu^2}(x) = p(X \leq x) &= \int_0^x \frac{t^{\frac{\nu-2}{2}} e^{\frac{-t}{2}}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} \, dt \\ &= 1 - \frac{\Gamma(\frac{\nu}{2}, \frac{x}{2})}{\Gamma(\frac{\nu}{2})} = \frac{\gamma(\frac{\nu}{2}, \frac{x}{2})}{\Gamma(\frac{\nu}{2})} = P\left(\frac{\nu}{2}, \frac{x}{2}\right) \end{aligned} \tag{5.5}$$

[1] We use the older yes/no version of the Davidson's ESQ and an unpublished correlation matrix has been provided to us by Sara Carletto and Luca Ostacoli (Department of Clinical and Biological Sciences, Università deglistudi di Torino, Italy), our partners in the NEVERMIND project. For the newer version of Davidson's ESQ, the matrix of correlations can be found in [117].

where $\Gamma(\nu/2)$ denotes the gamma function, $\Gamma(\nu/2, x/2)$ is the upper incomplete gamma function, $\gamma(\cdot)$ is the lower incomplete gamma function and $P(\nu/2, x/2)$ is the regularized gamma function.

For a random variable $\Delta^2 \sim \chi^2_\nu$, the $F_{\chi^2_\nu}(\Delta^2)$ acts as a dissimilarity measure taking values in the range [0,1] and indicates the probability $p(\chi^2_\nu \leq \Delta^2_{ij})$. Suppose χ^2_ν is a chi-square distribution with ν degrees of freedom (in our case $\nu = 6$) and Δ^2_{ij} is the squared Mahalanobis distance between personality scores $\boldsymbol{s}^{(i)}$ and $\boldsymbol{s}^{(j)}$ of the pair of patients (q_i, q_j). Then the probability that these two patients are less distant (more similar) than any two random patients (thus more likely to be originated from the same population or belonging in the same cluster) is defined as:

$$p(\chi^2_\nu > \Delta^2_{ij}) = 1 - p(\chi^2_\nu \leq \Delta^2_{ij}) = 1 - F_{\chi^2_\nu}(\Delta^2_{ij}) \tag{5.6}$$

Therefore, in the current approach, we can define, for any pair of patients $q_i \in Q$ and $q_j \in \{Q \backslash q_i\}$, a similarity measure $w_{ij} = 1 - F_{\chi^2_\nu}(\Delta^2_{ij})$, which can then be used within the previous TL frameworks in order to modulate how much information we borrow from each "donor" participant. For simplicity, when we consider predictions for a new patient q_{N+1} (as denoted in Chapters 3 and 4) and we transfer from "donor" patients q_n coming from the set of N other patients, we denote the similarity measure as:

$$w_n = 1 - F_{\chi^2_\nu}(\Delta^2_{n,N+1}). \tag{5.7}$$

In our non-parametric TL approach based on BMA, the MCMC sampler creates S samples for each of the N patients. This time, the K vectors of parameters used in equation (3.5) page 35 are drawn by randomly choosing each time a "donor" patient q_n from the set of N other patients, with a probability

$$p(q_n) = \frac{w_n}{\sum_{l=1}^{N} w_l}, \tag{5.8}$$

and then uniformly randomly sampling a set of parameters from this patient's S samples. In other words, we use the clusters previously formed from the personality traits to place a prior on the random selection of model parameters so that models

from subjects with similar characteristics will have a higher probability of being selected. In this way, we allow the TL to be more or less influenced by the similarity between the two patients and control to which extent each of the model parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$, which represent information coming from the other patients, will contribute.

Fig. 5.1 shows a graphical overview of the non-parametric transfer learning approach based on an MCMC sample and BMA, modulated by similarity.

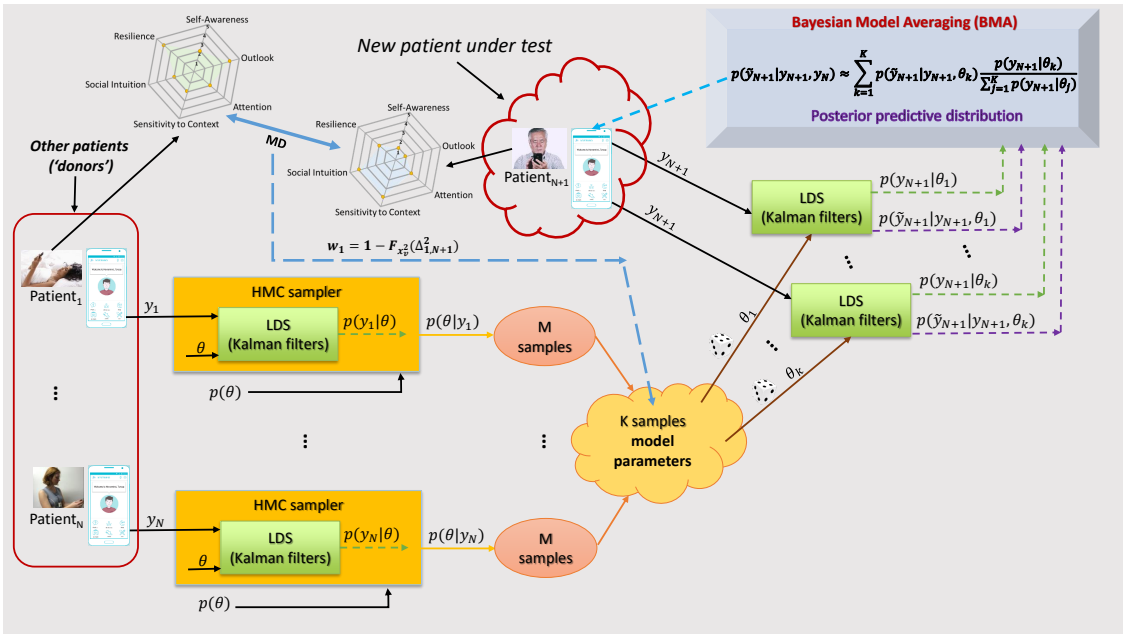


Figure 5.1: Pipeline of the proposed non-parametric TL method modulated by similarity.

Regarding the parametric TL approach based on minimizing the Fisher divergence, as we showed in Sec. 4.2.2, the best approximation to the mixture distribution can be computed by minimising a weighted sum of the individual divergences from $q_\beta(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ to the mixture components. However, this time we are not assuming a uniform mixture and therefore the mixture components do not have the same weights as in equation (4.11) thus, we now obtain an optimal value $\boldsymbol{\beta}^*$ by solving the following constrained optimisation problem:

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \sum_{n=1}^{N} w_n F_n(\boldsymbol{\beta}), \tag{5.9}$$

where the weight w_n corresponds to the similarity measure between a "donor" patient q_n and the patient under test. This means that we use the similarity to weight the

different divergences produced in this process (when we try to find the optimal parameter for the model) and this introduces a transfer learning component that it is biased towards more similar patients in the cohort to the patient of interest.

Fig. 5.2 shows a graphical overview of the version of the parametric transfer learning approach based on the Fisher divergence modulated by similarity.

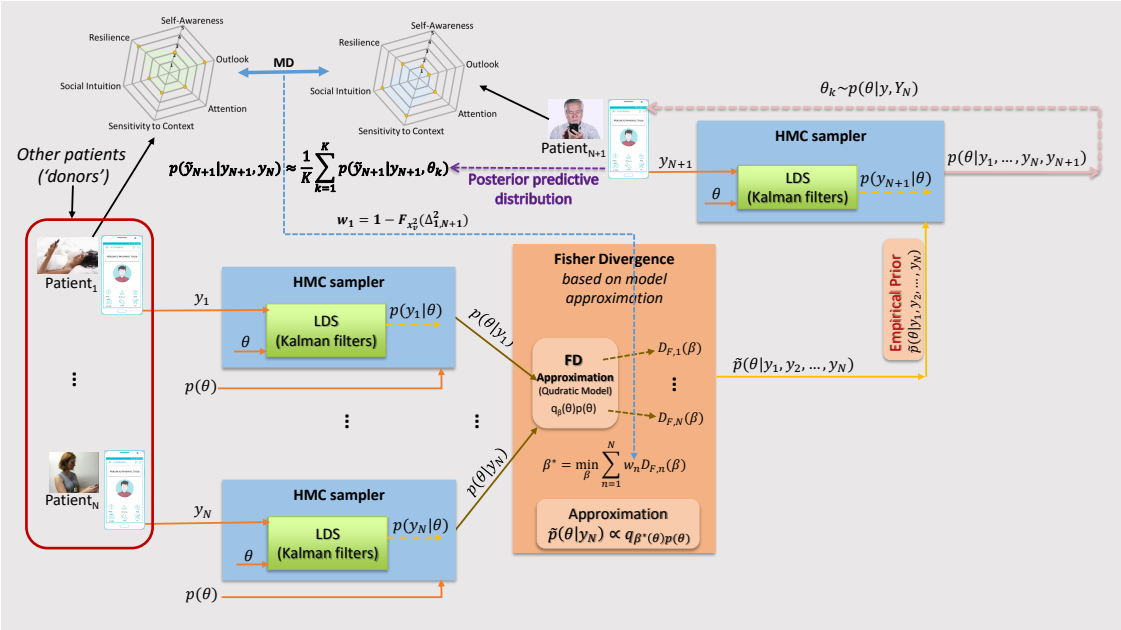


Figure 5.2: Pipeline of the proposed parametric TL method modulated by similarity.

5.3 Results

The predictive performance of the current TL method, that based on the approach *modulated by similarity* as opposed to the *uniform-sampling* approach presented in Chapters 3 and 4, was evaluated on the NEVERMIND dataset (see Sec. 5.1.1). To assess whether TL provides significant performance benefits and whether clustering patients by similarity improves the predictions, we compared the proposed TL methods against a number of competing models. These models are the maximum a posteriori (MAP) set of parameters denoted as M_{MAP} and presented in Sec. 4.3, a non-transfer learning Bayesian model (M_{No}), which utilizes parameters from the

S samples created for each patient using the HMC sampler and the four baseline models M_{A}, M_{P}, M_{L}, and M_{R} previously presented in Sec. 3.3.2. As per the current method, the version of the non-parametric TL approach based on BMA modulated by similarity, is denoted here as M_{BMA_m}, while the version of the parametric TL approach based on the FD modulated by similarity is referred as M_{FD_m} henceforth. The corresponding uniform-sampling versions will be referred as M_{BMA_u} and M_{FD_u}, respectively. We further consider a modified form of M_{FD}, which borrows information only from patients with at least 3 weeks of valid data. This selection made based on finding from previous studies reviewed in Sec. 2. The uniform sampling approach of this model is denoted as $M_{\text{FD}^3_u}$, while the corresponding approach modulated by similarity is referred as $M_{\text{FD}^3_m}$. Finally, the chains for all models are created using the HMC sampler while the selection of the priors for the model parameters remains the same as in Sec. 4.3. For the output of the models see Sec. 4.3.1.

5.3.1 Trade-off in computational cost and accuracy

In Chapter 3 we evaluated the proposed non-parametric TL method for $K = 1000$ candidate models per participant since the non-parametric methods work better when we have many data-points. However, this high number of models make this approach computationally very expensive. Therefore, for the parametric TL method presented in Chapter 4, we made a decision to set $K = 1000$ models in total.

In this section, we want to evaluate and compare non-parametric with parametric TL approaches and so we need to examine the trade-off in computational cost and accuracy for these two classes of methods. In order to find the best trade-off between the number of candidate models, K, and accuracy, we evaluate the non-parametric M_{BMA_u} and M_{BMA_m} approaches using different K values. Based on these results, we will then select the number of candidate models for the rest of our experiments.

The analysis was conducted for three equally spaced snapshots over the training periods, namely $L_{\text{tr}} \in \{1, 6, 11\}$ weeks and forecasting periods $L_{\text{fc}} \in \{1, 3, 7\}$ days. Assuming N is the number of "donor" patients and $K = 1000$ is the number

of candidate models, then the performance of the non-parametric approach was evaluated using $K_{min}(M_{\text{BMA}_u}) = K$ candidate models for transfer, sampled uniformly at random with replacement and $K_{max}(M_{\text{BMA}_u}) = K \times N$ models sampled uniformly at random without replacement. For the modulated by similarity version, we set $K_{min}(M_{\text{BMA}_m}) = \sum_{n=1}^{N} Kp(q_n)$ models sampled at random with replacement, where the probability for each "donor" patient to be selected is determined by its relative weight according to equation (5.8). In addition, we have $K_{max}(M_{\text{BMA}_m}) = \sum_{n=1}^{N} Kw_n$ weighted random samples without replacement, which is a fraction of the $K_{max}(M_{\text{BMA}_u})$ models. In this latest approach, the "donor" patients are selected based on the probability according to equation (5.7).

The performance of models was evaluated using the *RMSE* while statistical significance of differences between the models was tested using the Wilcoxon Signed-Rank test. We used this test for pairwise comparisons in order to evaluate the null hypothesis that the M_{BMA_u} using $K_{min}(M_{\text{BMA}_u})$ models and M_{BMA_u} using $K_{max}(M_{\text{BMA}_u})$ models will have equal distribution of prediction errors as well as that the comparison of these predictive models will not show a significant difference. We followed the same analysis for the M_{BMA_m} but this time using $K_{min}(M_{\text{BMA}_m})$ and $K_{max}(M_{\text{BMA}_m})$ donor sampling schemes.

Table 5.1 shows the average *RMSE* and the medians of *RMSE*-differences obtained over all patients for whom data were available for the corresponding $(L_{\text{tr}}, L_{\text{fc}})$ pair. In most cases, both our approaches show a lower error when the number of candidate models is high. However, significant differences were found only for the uniform-sampling approach with training period L_{tr}=11 and forecasting periods $L_{\text{fc}} \in \{3, 7\}$. This represents the case where the relative contribution of the transfer learning component to the overall accuracy reduces or even becomes negligible. For the approach modulated by similarity there is no statistical evidence to prefer $K_{max}(M_{\text{BMA}_m})$ over $K_{min}(M_{\text{BMA}_m})$ models. Therefore, for the rest of our experiments we will consider $K = 1000$ candidate models in total since both methods work

Table 5.1: Average *RMSE* and medians of *RMSE*-differences for various durations of training and forecast periods, as well as the number of "valid" participants within each period.

		M_{BMA_u}			M_{BMA_m}			
		average RMSE (K_{min})	*average RMSE* (K_{max})	*median of differences*	*average RMSE* (K_{min})	*average RMSE* (K_{max})	*median of differences*	*sample size*
Train 1	**Predict 1**	0.7001	0.7152	-0.0002	0.6928	0.7045	-0.0033	55
	Predict 3	0.8012	0.8052	0.0051	0.8049	0.8033	0.0010	82
	Predict 7	0.8181	0.8150	0.0045	0.8211	0.8250	-0.0019	99
Train 6	**Predict 1**	0.4230	0.3831	0.0103	0.4447	0.4341	0.0019	34
	Predict 3	0.5744	0.5629	-0.0020	0.5916	0.5774	-0.0016	55
	Predict 7	0.5713	0.5671	-0.0022	0.5848	0.5788	0.0028	65
Train 11	**Predict 1**	0.5004	0.4558	-0.0001	0.4294	0.4667	-0.0093	28
	Predict 3	0.5300	0.4933	**0.0085**	0.4898	0.4911	-0.0002	36
	Predict 7	0.5322	0.5032	**0.0103**	0.5063	0.5011	0.0013	42

Bold values indicate statistical significance ($p < .05$). Also note that there is a different number of "valid" participants per data length because the analysis includes only participants who have data within this period.

reasonable well with this K and it is a less costly alternative compared with $K_{max}(M_{\text{BMA}_u})$ and $K_{max}(M_{\text{BMA}_m})$.

5.3.2 Comparison against uniform-sampling TL models

In order to examine whether transferring from patients' groups with similar characteristics has any further benefit to the learning process, we compared the models modulated by similarity (M_{BMA_m} and M_{FD_m}) against the model M_{BMA_u} and M_{FD_u}. In this comparison, we also included the models which borrow information only from patients with at least 3 weeks of valid data ($M_{\text{FD}^3_m}$ and $M_{\text{FD}^3_u}$).

Our models were trained for all combinations of training data length and forecasting periods. Here we show a representative subset from these results that includes the combinations of training periods $L_{\text{tr}} \in \{1, 2, 3\}$ and and predict for all L_{fc}. This subset represents the most difficult scenarios where the amount of available historical training data for each individual is less than a month worth. Again, note that the number of participants for which it is possible to obtain predictions depends

on the choice of L_{tr} and L_{fc}. The statistical significance of differences between the models were again tested using the Wilcoxon Signed-Rank test (a one-tailed hypothesis, at the 5% level).

Table 5.2 shows the results obtained with respect to the LL evaluation measure calculated over different sets of patients. The last row of the table shows the number of such valid participants per $(L_{\text{tr}}, L_{\text{fc}})$ pair. The top half shows the average LL values for each pair of models, the models giving a better probabilistic prediction being the ones having the higher (less negative) value. The bottom half shows pairwise median differences (the median of the differences) between the uniform-sampling version and the version modulated by similarity of the models. In addition, the bottom half of the table, compares the models borrowing from "donors" with at least 3 weeks of valid data with the corresponding unrestricted version of the same models.

As seen in the table, the M_{BMA_m} performs better than its uniform-sampling version (M_{BMA_u}) and makes, on average, better probabilistic predictions in all $(L_{\text{tr}}, L_{\text{fc}})$ pairs except for L_{tr}=3 where the uniform-sampling versions starts to become better. However, even though M_{BMA_u} scores better than M_{BMA_m}, the differences between the models were not significant. On the other hand, when training with only 1 weeks of past data (L_{tr}=1) and forecasting for $L_{\text{fc}}\in\{1, 3, 7\}$ days, the M_{FD_m} and $M_{\text{FD}^3_m}$ models score significantly better compared to their uniform-sampling versions. Finally, for the training periods $L_{\text{tr}}\in\{1, 2\}$ and forecasting periods $L_{\text{fc}}\in\{1, 3, 7\}$, the models that borrow information only from patients with at least 3 weeks of valid data ($M_{\text{FD}^3_u}$ and $M_{\text{FD}^3_m}$) perform better than their counterpart version without this restriction (M_{FD_u} and M_{FD_m}). Notably, when the training is performed with the minimum training period (one week), these differences are significant. In addition, the model modulated by similarity that borrows information only from patients with at least 3 weeks of valid data ($M_{\text{FD}^3_m}$) performs slightly better that its unrestricted version (M_{FD_m}) when $L_{\text{tr}} = 3$ and $L_{\text{fc}} = 7$.

Table 5.2: Performance comparison of different pair of models based on average LL (top) and the median of LL-differences (bottom) for various durations of training and forecast periods.

Model Type	$L_{\text{tr}} = 1$			$L_{\text{tr}} = 2$			$L_{\text{tr}} = 3$		
	L_{fc}=1	L_{fc}=3	L_{fc}=7	L_{fc}=1	L_{fc}=3	L_{fc}=7	L_{fc}=1	L_{fc}=3	L_{fc}=7
M_{BMA_m}	-1.821	-4.538	-8.614	-2.038	-3.910	-7.097	-1.710	-3.755	-7.749
M_{BMA_u}	-2.086	-4.818	-9.123	-2.156	-4.217	-7.340	-1.640	-3.546	-7.456
M_{FD_m}	-2.175	-4.770	-8.976	-2.204	-4.209	-7.306	-2.137	-3.844	-7.468
M_{FD_u}	-2.271	-4.875	-9.174	-2.182	-4.176	-7.290	-2.058	-3.827	-7.480
$M_{\text{FD}^3_m}$	-2.014	-4.551	-8.781	-2.088	-4.111	-7.224	-2.196	-3.865	-7.462
$M_{\text{FD}^3_u}$	-2.106	-4.651	-8.894	-2.091	-4.096	-7.220	-2.086	-3.860	-7.530
Models Compared									
M_{BMA_u} vs M_{BMA_m}	-0.065	-0.024	-0.004	0.044	-0.114	-0.124	0.025	0.005	-0.029
M_{FD_u} vs M_{FD_m}	**-0.031**	**-0.026**	**-0.061**	0.008	-0.010	-0.023	0.006	-0.003	0.000
$M_{\text{FD}^3_u}$ vs $M_{\text{FD}^3_m}$	**-0.008**	**-0.057**	**-0.037**	-0.023	-0.012	0.020	-0.000	0.002	-0.011
M_{FD_u} vs $M_{\text{FD}^3_u}$	**-0.048**	**-0.051**	**-0.058**	0.001	0.002	0.003	0.031	0.053	0.076
M_{FD_m} vs $M_{\text{FD}^3_m}$	**-0.060**	**-0.077**	**-0.046**	-0.004	-0.009	0.003	0.028	0.052	0.056
participants	*55*	*82*	*99*	*54*	*75*	*92*	*41*	*74*	*94*

Bold values indicate statistical significance ($p < .05$). Also note that there is a different number of "valid" participants per data length because the analysis includes only participants who have data within this period.

The training period L_{tr}=1 represents the early stage of data collection when limited person-specific data are available which is the period where the benefit of transfer learning on predictive performance will be greater. In reference to our research question RQ.3 (H.3), the results offer support to the idea that *a transfer learning model biasing population sampling towards participants having similar characteristics to the patient being modelled will perform significantly better than a model sampling the population with equal probability over all other participants.* This is particularly clear for the M_{FD} model, which incorporates a modified prior that accounts for the knowledge available from all other participants, where it is beneficial to transfer from patients' groups with similar characteristics to the patient in question rather than transferring equally from the entire-population.

5.3.3 Comparison against competing models

In this section, we compare the performance of the transfer learning approaches modulated by similarity against the competing models listed in Sec. 5.3. To do this, we analyse "winning percentages" defined in equation (3.10). The analyses was performed for the early stage of data collection which includes training periods $L_{\text{tr}} \in \{1, 2, 3\}$ and predicting one week ahead ($L_{\text{fc}} = 7$). We used again the exact Wilcoxon Signed-Rank test-statistic [104] to make pairwise comparisons between methods but this time, we have multiple comparisons which means that some results may have p-values less than the 5% significant level, even if the null hypotheses is really true. Thus, we adjusted the false discovery rate by using the Benjamini-Hochberg procedure [119].

Table 5.3: Winning results based on LL for each transfer learning method against the competing models for L_{fc}=7.

		M_{No}	M_{MAP}	M_{A}	M_{P}	M_{L}	M_{R}	**part.**
L_{tr}=1	$M_{\text{BMA}m}$	**60.61%**	**86.87%**	**67.68%**	**60.61%**	**64.65%**	**85.57%**b	
	$M_{\text{FD}m}$	**62.63%**	**88.89%**	**57.58%**	55.56%	**60.61%**	**82.47%**	99
	$M_{\text{FD}^3_m}$	**66.67%**	**88.89%**	**56.57%**	57.58%	**64.65%**	**83.51%**	
L_{tr}=2	$M_{\text{BMA}m}$	**58.70%**	**82.61%**	**55.43%**	**58.70%**	**65.22%**	**74.73%**	
	$M_{\text{FD}m}$	**61.96%**	**82.61%**	59.78%	57.61%	**58.70%**	**75.82%**	92
	$M_{\text{FD}^3_m}$	**56.52%**	**82.61%**	56.52%	57.61%	**63.04%**	**75.82%**	
L_{tr}=3	$M_{\text{BMA}m}$	53.19%	**76.60%**	44.68%	**68.09%**	**70.21%**	**66.67%**	
	$M_{\text{FD}m}$	55.32%	**78.72%**	48.94%	**62.77%**	**62.77%**	**67.74%**	94
	$M_{\text{FD}^3_m}$	47.87%	**80.85%**	45.74%	**64.89%**	**63.83%**	**69.89%**	

Bold values indicate statistical significance using Benjamini-Hochberg procedure ($p < .05$). Also note that there is a different number of "valid" participants [part.] per data length because the analysis includes only participants who have data within this period.

Table 5.3 shows the winning percentage of each TL approach and training period against the competing models, under the most challenging scenario of training the models with limited past observations and predicting for a full week ahead. The winning percentage values are based on the LL performance measure for this analysis.

So, they take into account both the accuracy *and* the uncertainty of the predictions. Results indicate that overall the modulated by similarity transfer learning methods score more "wins" than their competitors in 50 out of 54 comparisons (92.6%). The winning percentages range from 44.68% to 89.90% with all of the proposed transfer learning models scoring significantly more wins (at the 5% level, using a one-tailed hypothesis) compared to the Maximum A Posteriori model (M_{MAP}), the *ordinary least squares regression* model (M_{R}) and the *last-datapoint* model (M_{L}). M_{MAP} and M_{R} are less accurate models compared to our approach and they struggle to make acceptable 7-day predictions, at least in this initial stage where the data available for training are sporadic and limited. M_{L}, which does not incorporate training since we simply choose the last observed value as the prediction for the sequence of next-week's observations, is not informative because it does not reflect any changes in the mood. As for M_{P}, our methods score better more than it 50% of the time, in all three scenarios, with most of the results being statistically significant. This method appears to be also suboptimal since the population-based estimate is too generalized and it does not take into account, for example, the variability in behavioural patterns of the patient. These results corroborate our previous finding presented in Sec. 4.4.2 (Table 4.2). Regarding the *patient-average prediction* model (M_{A}) when we train with 3 weeks of past data, the accuracy increases and it turns out that the average over a long period of time is actually a good predictor. We have previously observed similar results in Sec. 3.3.2 (Table 3.2) where based on the actual error for the predictions (*RMSE*), when we trained with 3 weeks of past data on the ICU dataset, the model was superior to all its competitors. However, the transfer learning models modulated by similarity score significantly more wins than the M_{A} when minimal data are available for training and they are on par when more weeks of data are available for training. Finally, the non-transfer learning Bayesian model (M_{No}) follows closely our transfer learning methods. However, under the most challenging scenario of training the models with only one or two weeks of past observations and predicting the following week, the transfer learning approaches modulated by similarity perform significantly better. It is worth mentioning that

the version modulated by similarity of the non-parametric TL approach based on BMA (M_{BMA_m}) scores significantly more wins compared to all six competing models when only one week or two weeks of data are available for training. This result also confirms the superiority of the modulated by similarity TL approach and emphasizes that during the TL process, it is beneficial to transfer samples coming from similar patients in the cohort more than patients that are dissimilar.

Moreover, the above results support our previous findings presented in Sec. 3.3.2 (Fig. 3.5), Sec. 3.3.3 (Figs. 3.7) and Sec. 4.4.1 (Figs. 4.8) where the transfer learning approaches were significantly better than non-transfer learning methods when only a limited amount of patient-specific data were available for training. In reference to our research question RQ.1 (H.1), the results further support the idea that *in the presence of limited person-specific data available for training* (e.g., at the early stage of data collection), *a prediction model leveraging both patient-specific data* (i.e., the 'target' domain) *and knowledge gained from other patients* (i.e., a different but related 'source' domain) *will perform significantly better than a model which relies on the target domain alone for training.*

5.3.4 Effect of training data availability on performance

In this section, we investigate the effects of data availability on the model performance by training the models with different amount of historical data. Naturally the expectation is that when using less data, the prediction error will increase but the extent of this is unknown. In relation to second hypothesis (H.2), we further believe that *the benefit of transfer learning on predictive performance will be greater when the amount of person-specific data available for training is small and that as person-specific data availability increases, the relative contribution of the transfer learning component to the overall accuracy will diminish, until it becomes negligible.*

The performance of the proposed TL methods modulated by similarity against reference estimation approaches was assessed for a different number of observations used for training to predict one week (L_{fc}=7) of well-being scores. The analysis

was conducted for training periods $L_{\text{tr}} \in \{1, 2, ..., 12\}$ (regardless of the number of actual observations that happen to be contained within each period).

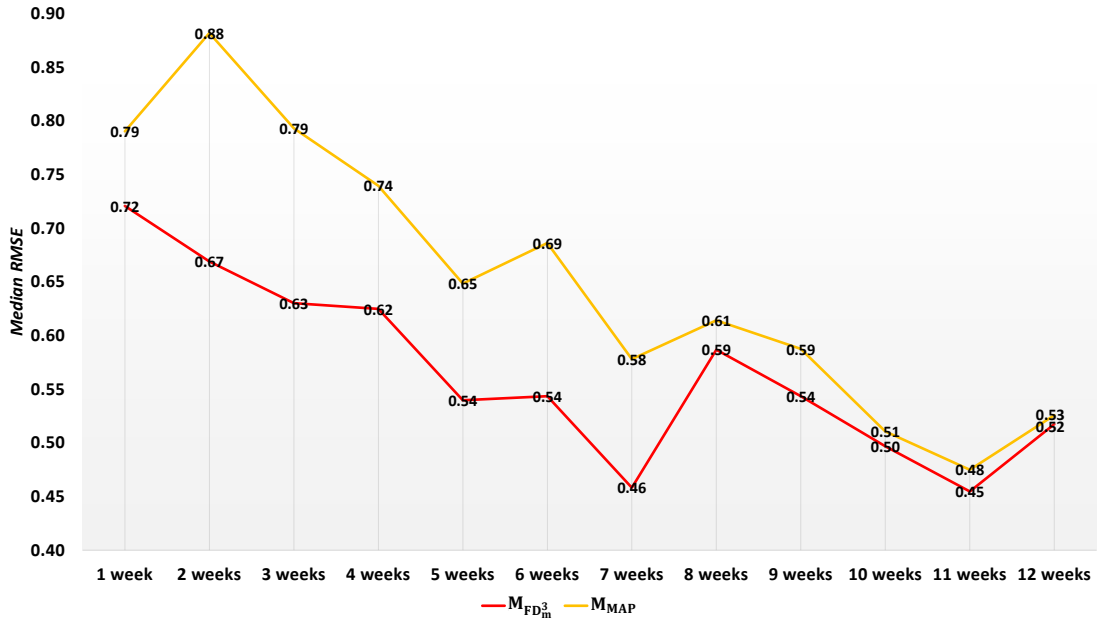
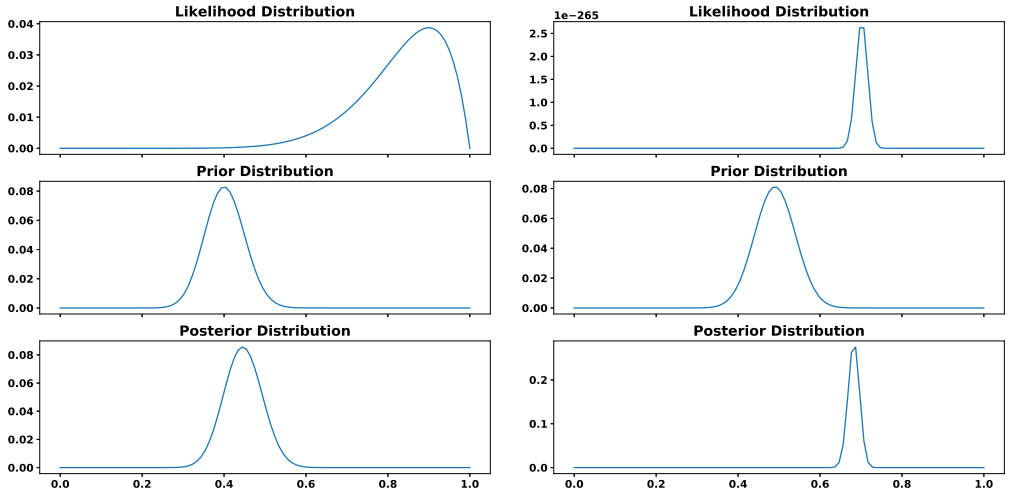


Figure 5.3: The median RMSE values for a transfer learning and a standard estimation approach, obtained over all patients for whom data was available for the corresponding training periods when predicting next week's scores.

A representative example of this analysis can be seen in Fig. 5.3, which shows the median forecasting error for the predictions of each individual obtained from one of our modulated by similarity TL methods ($M_{\text{FD}_m^3}$) and the standard estimation approach (M_{MAP}). The other transfer learning models perform very similarly and, so, have been omitted here. The figure displays the changes in *median RMSE* over time. Based on these results, we observe that M_{MAP} initially struggles to make acceptable 7-day predictions and it is the less accurate method. However, predictions improve with the amount of data available for training. At the same time, we observe that for shorter training periods, our transfer learning method outperforms M_{MAP}, whereas it is on par for longer training periods. Initially, when we have less data, the prior dominates the posterior distribution. Our approach is able to use models all around the posterior while the MAP estimation uses only the maximum point, which makes our approach more robust. However, as the data availability increase,

the likelihood term increases until it dominates over the prior and the posterior distribution becomes almost identical to the likelihood (Fig. 5.4). This makes our method to gradually and automatically become virtually identical to M_{MAP}. Finally, it is important to note that the *median RMSE* obtained with our algorithm training with only one-week worth of data is matched by the competing state of the art algorithm only when more than four-weeks worth of data are used for training. This result demonstrates the effect of transfer learning in bootstrapping the learning process and allowing us to achieve a better accuracy much earlier.

The above results further support our previous findings presented in Sec. 4.4.2 (Table 4.2) and Sec. 5.3.3 (Table 5.3) where the transfer learning methods were superior to other approaches and our method was not only able too make predictions from week one, but showed also a remarkable stability for any number of training weeks.



(a) With less data, the prior dominates the posterior distribution

(b) With more data, the likelihood dominates the posterior distribution

Figure 5.4: Posterior distribution dominated by prior or the data.

5.4 Discussion

In this chapter, we presented a modified version of the methods presented in Chapters 3 and 4, which controls how much information we borrow for the transfer

learning process according to a measure of similarity in emotional profiles between the patients under test and "donor" patients. More specifically, we use a measure of distance between the personalities of the patients in order to weight the samples that come from similar "donor" patients in the cohort more than the samples coming from dissimilar "donor" patients. In this way, we modulate the transfer learning slightly to effectively get more information from the patients in the cohort that are more suitable for the transfer. Similarly, in our parametric transfer learning approach, we use the similarity to weight the different divergences produced in this process when we try to find the optimal parameter for the model and this introduces a transfer learning component that is biased towards more similar patients in the cohort with the patient of interest.

Our analysis suggests that applying preprocessing strategies that promote individuals in the population which are known to be similar, in some way, to the person being modelled, enhances the transfer learning. We demonstrated that a "soft" clustering which splits the population into subsets of patients that exhibit similar relationships between their personality traits improves the accuracy of our algorithms. The proposed approaches modulated by similarity outperform the uniform-sampling version of these methods and in the early stages of data collection, when the data available for initial training are sporadic and limited, they also predict better than a number of competing models.

Our results highlight the value of transfer learning and verify the effectiveness of the proposed method. They further corroborate our previous findings in relation to our first and second hypothesis (H.1 and H.2) and are also supportive of the third hypothesis (H.3). Overall, the experiments suggest that the proposed approach constitutes a good transfer learning strategy.

6

Conclusions and future research

This chapter summarizes the main contributions of this thesis, discusses the limitations of the work and looks at promising avenues for future research.

Contents

6.1 Summary

The use of smart-phones and wearable sensors for quantifying and providing personalised predictions of well-being states is an actively growing field, with potential applications in the prevention and self-management of depression and other disorders. Personalisation refers to the ability to learn a model, which is specifically tuned to its intended user. However, a major obstacle in this endeavour so far has been that traditional forms of machine learning (which typically require the availability of large datasets of uniformly sampled data) are generally not applicable to solve this problem. There are two main reasons for this: firstly the kind of data provided by patients through smart-phones and wearable sensors tend to be sporadic and/or intermittently available; secondly, realistically, for such a personalised system to be useful, users need predictions virtually from day one, whereas in a typical situation the data available to the system for personalisation

will initially be very limited, and acquired incrementally over time. As demonstrated in previous studies and in this thesis, it could take *weeks* before a sufficient amount of data has accumulated that guarantees reliable predictions.

6.2 Contributions

The present research focused on providing reliable personalised predictions in the early stage of data collection when insufficient personal data of suitable nature are available for training. In this thesis, I proposed novel approaches with the aim of addressing the challenges on creating subject-specific models and making predictions in real world health-related applications when training is performed incrementally on scarce sporadic biomedical data.

To tackle the challenges and overcome the limitations of traditional ML algorithms, Bayesian transfer learning techniques were proposed that allow to make informed predictions by leveraging information coming from other patients in our study. These algorithms are probabilistic and allow not only to make predictions about future trends (a sequence of future moods) but also to provide an associated measure of uncertainty around the predictions allowing to know how much we can trust the predictions themselves. The approach I follow further allows for a seamless transition from generalised to highly personalised models, as data become gradually available.

The proposed methods are the following:

(i) a non-parametric transfer learning approach based on Bayesian Model Averaging, which allows to make predictions about a specific patient with limited sparse training data by combining models trained on other "donor" patients according to how well these models fit the specific patient's past observations and performing BMA on the candidate model.

(ii) a parametric transfer-learning approach based on the Fisher divergence, which expresses external information coming from "donor" patients as a prior probability distribution used within a Hamiltonian Monte Carlo framework.

(iii) a modified version of the methods in (1) and (2), which is a similarity-based transfer learning approach, that controls how much information is borrowed for transfer according to a measure of similarity in emotional profiles between the patients under test and "donor" patients.

These approaches make it possible to create patient-specific models and make useful predictions of self-reported well-being scores, even when the data available for initial training are sporadic and limited, such that training is performed incrementally as more data become slowly available over time. The application of transfer learning allows each patient to have a model tailored for them but still benefit from other patients data. In this way, someone can start making predictions that are neither too general, nor unreliably specific, until there is sufficient data to converge to a small representative pool of models that are highly personalised to the patient.

We demonstrated the proposed approaches on data obtained by the NEVERMIND pilot and clinical trial, and measured their performance against each other and also against a number of competing models. We analysed the effectiveness of the proposed transfer learning approaches for predicting personalised well-being scores utilizing individual's data alongside with external information coming from other patients. The effectiveness of these methods was examined by conducting a range of experiments on the real world datasets collected from individual patients. Our results showed that these approaches, in most cases, yield a significant improvement over their competitors, and are particularly useful in difficult training/forecasting scenarios, e.g., when one requires a distant, patient-specific forecast, with only a limited initial amount of patient-specific data available for training.

This study further demonstrated the importance of transferring more information from patients in a cohort who closely resemble each other than from patients

that are dissimilar. In these approaches, transfer learning was enhanced by using personality trait information in order to assign weights to each "donor" patient in the population based on the similarity in personality traits to the target patient. In this way, well-being predictions for the target patient are based mostly on the observed well-being scores from the similar patients and this speeds up the learning process and the accuracy of the algorithms.

Regarding the research questions:

- ⋆ The results presented in Chapters 3, 4, 5 support the hypothesis that:
  in the presence of limited person-specific data available for training (e.g. at the early stage of data collection), a prediction model leveraging both patient-specific data (i.e. the "target" domain), and knowledge gained from other patients (i.e. a different but related "source" domain), will perform significantly better than a model which relies on the target domain alone for training.

- ⋆ The results presented in Chapters 4, 5 support the hypothesis that:
  the benefit of transfer learning on predictive performance will be greater when the amount of person-specific data available for training is small; as person-specific data availability increases, the relative contribution of the transfer learning component to the overall accuracy will diminish, until it becomes negligible.

- ⋆ The results presented in Chapters 5 support the hypothesis that:
  a transfer learning model mostly incorporating highly relevant information from a source domain (i.e. biasing population sampling towards participants having similar characteristics to the patient being modelled) will perform significantly better than a model which utilizes the source domain in a general manner (i.e. sampling the population with equal probability over all other participants)

6.3 Limitations and future work

One limitation of this study is that the Linear Dynamical System model used was limited to training and prediction on observations that reflected questionnaire responses only. Furthermore, the LDS model was initially limited to a single "general" latent state reflecting well-being in a general sense. This was later extended to three "observation-specific" states for the reason that a single latent-state model did not seem to be expressive enough to capture the underlying complexity of depressive states. The three-state model provided a richer representation of the underlying biological states by allowing linking latent states to observations directly and by allowing individual states to interact with each other. However, in this work, I have not explored the differences between the single state and three state model in much detail or the particular manner in which modelling observations with specific latent states improves prediction compared to when using such observations to simply update the prior probability of more general states. Future work will therefore focus on establishing the best combination of "modelled-as-state" vs "update-only" observations, as well as considering richer observation vectors, e.g., by also including information coming from physiological signals or the addition of PHQ9 questionnaire information, which could potentially strengthen the model's predictive abilities further, as well as the clinical utility. In this case, it will be interesting to investigate the effect that such extensions might have on transfer learning, as well as whether this is worth the increased complexity of the model.

In this study, I found that the background patient population acting as the source domain for the transfer learning component was informative as shown by the performance of transfer learning. Therefore, I applied a preprocessing strategy that promote individuals in the population that are known to be similar in personality to the person being modelled and this enhanced transfer learning. However, a limitation is that I made no attempt to quantify or investigate ways in which other background knowledge like the primary disease, could be exploited which could potentially provide even better results.

According to very recent studies, high fluctuations in well-being label were associated with larger error in the predictive models [90] while mood variability, personality traits and day of the week played an important role in model performance [89]. This is something that ought to be further examined for the proposed transfer learning algorithms through a post-hoc analysis since based on the results transfer learning appeared to be more robust to this situation.

Finally, I recognise that this method has wider applicability to other domains, such as finance, recommender systems, training initiatives, etc, and generally any scenario where limited or sporadic data arrive in a sequential manner, and a seamless transition from generalised to personalised models is required. Therefore, future work will also need to focus on verifying the performance and generality of this approach, both on the complete NEVERMIND dataset (including data coming from both physiological and questionnaire-based sources), as well as other known external datasets (such as the MIMIC-III critical care database [120] used also in the beginning of this thesis). During the verification, the analysis should further include all of the methods described in the Background chapter.

References

[1] Jacqueline JMH Strik, Adriaan Honig, Richel Lousberg, and Johan Denollet. "Sensitivity and specificity of observer and self-report questionnaires in major and minor depression following myocardial infarction". In: *Psychosomatics* 42.5 (2001), pp. 423–428.

[2] John A. Rush, Thomas J. Carmody, Hisham M. Ibrahim, Madhukar H. Trivedi, Melanie M. Biggs, Kathy Shores-Wilson, M. Lynn Crismon, Marcia G. Toprac, and T. Michael Kashner. "Comparison of Self-Report and Clinician Ratings on Two Inventories of Depressive Symptomatology". In: *Psychiatric Services* 57.6 (2006), pp. 829–837.

[3] Jaap J. A. Denissen, Ligaya Butalid, Lars Penke, and Marcel A. G. van Aken. "The effects of weather on daily mood: A multilevel approach." In: *Emotion* 8.5 (2008), pp. 662–667.

[4] Theo A. Klimstra, Tom Frijns, Loes Keijsers, Jaap J. A. Denissen, Quinten A. W. Raaijmakers, Marcel A. G. van Aken, Hans M. Koot, Pol A. C. van Lier, and Wim H. J. Meeus. "Come rain or come shine: Individual differences in how weather affects mood." In: *Emotion* 11.6 (2011), pp. 1495–1499.

[5] Yoshihiko Suhara, Yinzhan Xu, and Alex 'Sandy' Pentland. "DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks". In: *Proceedings of the 26th International Conference on World Wide Web - WWW '17*. New York, New York, USA: ACM Press, 2017, pp. 715–724.

[6] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. "A unifying view on dataset shift in classification". In: *Pattern Recognition* 45.1 (Jan. 2012), pp. 521–530.

[7] *NEurobehavioural predictiVE and peRsonalised Modelling of depressIve symptoms duriNg primary somatic Diseases with ICT-enabled self-management procedures.* Online at http://www.nevermindproject.eu/. 2018.

[8] Vladimir Carli, Danuta Wasserman, Gergö Hadlaczky, Nuhamin Gebrewold Petros, Sara Carletto, Luca Citi, Sergio Dinis, Claudio Gentili, Sergio Gonzalez-Martinez, Aldo De Leonibus, et al. "A protocol for a multicentre, parallel-group, pragmatic randomised controlled trial to evaluate the NEVERMIND system in preventing and treating depression in patients with severe somatic conditions". In: *BMC psychiatry* 20.1 (2020), pp. 1–10.

[9] Eirini Christinaki, Riccardo Poli, and Luca Citi. "Bayesian Transfer Learning for the Prediction of Self-reported Well-being Scores". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2018, pp. 41–44.

[10] Eirini Christinaki, Tasos Papastylianou, Riccardo Poli, and Luca Citi. "Parametric transfer learning based on the fisher divergence for well-being prediction". In: *Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019*. Institute of Electrical and Electronics Engineers Inc., Oct. 2019, pp. 288–295.

[11] Eirini Christinaki, Tasos Papastylianou, Sara Carletto, Sergio Gonzalez-Martinez, Luca Ostacoli, Manuel Ottaviano, Riccardo Poli, and Luca Citi. "Well-being Forecasting using a Parametric Transfer-Learning method based on the Fisher Divergence and Hamiltonian Monte Carlo". In: *EAI Endorsed Transactions on Bioengineering and Bioinformatics* 1.1 (Oct. 2020).

[12] Michael A. McCarthy. *Bayesian Methods for Ecology*. Cambridge University Press, 2007.

[13] Darren J. Wilkinson. "Bayesian methods in bioinformatics and computational systems biology". In: *Briefings in Bioinformatics* 8.2 (Dec. 2006), pp. 109–116.

[14] Paul R. Wade. "Bayesian Methods in Conservation Biology". In: *Conservation Biology* 14.5 (Oct. 2000), pp. 1308–1316.

[15] Yangxin Huang, Dacheng Liu, and Hulin Wu. "Hierarchical Bayesian Methods for Estimation of Parameters in a Longitudinal HIV Dynamic System". In: *Biometrics* 62.2 (June 2006), pp. 413–423.

[16] John K. Kruschke, Herman Aguinis, and Harry Joo. "The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences". In: *Organizational Research Methods* 15.4 (Oct. 2012), pp. 722–752.

[17] Peter Congdon. *Bayesian statistical modelling*. 2nd Editio. John Wiley & Sons, 2006, p. 596.

[18] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[19] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006, p. 738.

[20] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, 2012, p. 1067.

[21] Stan Development Team and their assignees. *14.1 Hamiltonian Monte Carlo | Stan Reference Manual*. URL: https://mc-stan.org/docs/2%7B%5C_%7D19/reference-manual/hamiltonian-monte-carlo.html (visited on 06/04/2019).

[22] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[23] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (Oct. 2010), pp. 1345–1359.

[24] Ricardo Sousa, Luis M Silva, Luis A Alexandre, and Jorge Santos. "Transfer Learning: Current Status, Trends and Challenges". In: *20th Portuguese Conference on Pattern Recognition, RecPad*. 2014, pp. 57–58.

[25] Daniel M. Roy and Leslie Pack Kaelbling. "Efficient Bayesian Task-Level Transfer Learning". In: *IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence* (2007), pp. 2599–2604.

[26] Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. "Transfer learning using computational intelligence: A survey". In: *Knowledge-Based Systems* 80 (2015), pp. 14–23.

[27] Lisa Torrey, Jude Shavlik, Trevor Walker, and Richard MacLin. "Transfer learning via advice taking". In: *Studies in Computational Intelligence* 262 (2010), pp. 147–170.

[28] Lisa Torrey and Jude Shavlik. "Transfer learning". In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 2009, pp. 242–264.

[29] Liang Ge, Jing Gao, Hung Ngo, Kang Li, and Aidong Zhang. "On handling negative transfer and imbalanced distributions in multiple source transfer learning". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 7.4 (Aug. 2014), pp. 254–271.

[30] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. "A survey of transfer learning". In: *Journal of Big Data* 3.1 (Dec. 2016), p. 9.

[31] Matthew E Taylor and Peter Stone. "Transfer Learning for Reinforcement Learning Domains: A Survey". In: *Journal of Machine Learning Research* 10 (2009), pp. 1633–1685.

[32] Diane Cook, Kyle D Feuz, and Narayanan C Krishnan. "Transfer learning for activity recognition: A survey". In: *Knowledge and Information Systems* 36.3 (Sept. 2013), pp. 537–556.

[33] Peitao Wang, Jun Lu, Bin Zhang, and Zeng Tang. "A review on transfer learning for brain-computer interface classification". In: *2015 5th International Conference on Information Science and Technology, ICIST 2015*. IEEE, Apr. 2015, pp. 315–322.

[34] Ruslan R. Salakhutdinov, Joshua B. Tenenbaum, and Antonio Torralba. "One-Shot Learning with a Hierarchical Nonparametric Bayesian Model". In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. 2012, pp. 195–206.

[35] Aaron Wilson, Alan Fern, and Prasad Tadepalli. "Transfer Learning in Sequential Decision Problems: A Hierarchical Bayesian Approach". In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. Vol. 27. 2012, pp. 217–227.

[36] Glen Wright Colopy, Stephen J. Roberts, and David A. Clifton. "Bayesian Optimization of Personalized Models for Patient Vital-Sign Monitoring". In: *IEEE Journal of Biomedical and Health Informatics* 22.2 (Mar. 2018), pp. 301–310.

[37] Harry Zhang. "The optimality of naive Bayes". In: *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference-FLAIRS2004*. 2004.

[38] Karen Niven. "Affect". In: *Encyclopedia of Behavioral Medicine*. Ed. by Marc D. Gellman and J. Rick Turner. New York, NY: Springer New York, 2013, pp. 49–50.

[39] Maria Kleinstäuber. "Mood". In: *Encyclopedia of Behavioral Medicine*. Ed. by Marc D Gellman and J Rick Turner. New York, NY: Springer New York, 2013, pp. 1259–1261.

[40] Norbert Schwarz and Gerald L. Clore. "Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states". In: *Journal of Personality and Social Psychology* 45.3 (Sept. 1983), pp. 513–523.

[41] Annette Brose, Ulman Lindenberger, and Florian Schmiedek. "Affective states contribute to trait reports of affective well-being". In: *Emotion* 13.5 (Oct. 2013), pp. 940–948.

[42] Eun Hyun Seo, Seung Gon Kim, Sang Hoon Kim, Jung Ho Kim, Jung Hyun Park, and Hyung Jun Yoon. "Life satisfaction and happiness associated with depressive symptoms among university students: A cross-sectional study in Korea". In: *Annals of General Psychiatry* 17.1 (Dec. 2018), p. 52.

[43] Heli Koivumaa-Honkanen, Jaakko Kaprio, Risto Honkanen, Heimo Viinamäki, and Markku Koskenvuo. "Life satisfaction and depression in a 15-year follow-up of healthy adults". In: *Social Psychiatry and Psychiatric Epidemiology* 39.12 (Dec. 2004), pp. 994–999.

[44] Spencer L. James et al. "Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159 (Nov. 2018), pp. 1789–1858.

[45] *NIMH » Major Depression*. URL: https://www.nimh.nih.gov/health/statistics/major-depression.shtml (visited on 04/30/2020).

[46] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. Pub, American Psychiatric, 2013.

[47] Alize J. Ferrari, Fiona J. Charlson, Rosana E. Norman, Scott B. Patten, Greg Freedman, Christopher J.L. Murray, Theo Vos, and Harvey A. Whiteford. "Burden of Depressive Disorders by Country, Sex, Age, and Year: Findings from the Global Burden of Disease Study 2010". In: *PLoS Medicine* 10.11 (Nov. 2013). Ed. by Phillipa J. Hay, e1001547.

[48] P. J. de Jong, B. E. Sportel, E. De Hullu, and M. H. Nauta. "Co-occurrence of social anxiety and depression symptoms in adolescence: differential links with implicit and explicit self-esteem?" In: *Psychological Medicine* 42.3 (2012), pp. 475–484.

[49] Stephen E. Gilman, Ewa Sucha, Mila Kingsbury, Nicholas J. Horton, Jane M. Murphy, and Ian Colman. "Depression and mortality in a longitudinal study: 1952-2011". In: *CMAJ* 189.42 (Oct. 2017), E1304–E1310.

[50] Maria Chiu, Simone Vigod, Farah Rahman, Andrew S. Wilton, Michael Lebenbaum, and Paul Kurdyak. "Mortality risk associated with psychological distress and major depression: A population-based cohort study". In: *Journal of Affective Disorders* 234 (July 2018), pp. 117–123.

[51] David M Clarke and Kay C Currie. "Depression, anxiety and their relationship with chronic diseases: a review of the epidemiology, risk and treatment evidence". In: *Medical Journal of Australia* 190.7 (2009), S54.

[52] Hee-Ju Kang, Seon-Young Kim, Kyung-Yeol Bae, Sung-Wan Kim, Il-Seon Shin, Jin-Sang Yoon, and Jae-Min Kim. "Comorbidity of depression with physical disorders: research and clinical implications". In: *Chonnam medical journal* 51.1 (Apr. 2015), pp. 8–18.

[53] Wayne Katon, Elizabeth H B Lin, and Kurt Kroenke. "The association of depression and anxiety with medical symptom burden in patients with chronic medical illness". In: *General Hospital Psychiatry* 29.2 (2007), pp. 147–155.

[54] Wayne J. Katon. "Epidemiology and treatment of depression in patients with chronic medical illness". In: *Dialogues in Clinical Neuroscience* 13.1 (2011), pp. 7–24.

[55] Tuba Yilmaz, Robert Foster, and Yang Hao. "Detecting Vital Signs with Wearable Wireless Sensors". In: *Sensors* 10.12 (Dec. 2010), pp. 10837–10862.

[56] Alberto Greco, Antonio Lanata, Luca Citi, Nicola Vanello, Gaetano Valenza, and Enzo Pasquale Scilingo. "Skin Admittance Measurement for Emotion Recognition: A Study over Frequency Sweep". In: *Electronics* 5.3 (Aug. 2016), p. 46.

[57] Rosalind W. Picard and Jocelyn Scheirer. "The galvactivator: A glove that senses and communicates skin conductivity". In: *Proceedings 9th Int. Conf. on HCI.* 2001.

[58] Gaetano Valenza, Claudio Gentili, Antonio Lanatà, and Enzo Pasquale Scilingo. "Mood recognition in bipolar patients through the PSYCHE platform: Preliminary evaluations and perspectives". In: *Artificial Intelligence in Medicine* 57 (2013), pp. 49–58.

[59] Gaetano Valenza, Luca Citi, Claudio Gentili, Antonio Lanata, Enzo Pasquale Scilingo, and Riccardo Barbieri. "Characterization of Depressive States in Bipolar Patients Using Wearable Textile Technology and Instantaneous Heart Rate Variability Assessment". In: *IEEE Journal of Biomedical and Health Informatics* 19.1 (Jan. 2015), pp. 263–274.

[60] Antonio Lanatà, Gaetano Valenza, and Enzo Pasquale Scilingo. "A novel EDA glove based on textile-integrated electrodes for affective computing". In: *Medical & biological engineering & computing* 50.11 (Nov. 2012), pp. 1163–1172.

[61] Anastasia Pampouchidou, Panagiotis Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis. "Automatic Assessment of Depression Based on Visual Cues: A Systematic Review". In: *IEEE Transactions on Affective Computing* 10.4 (2017), pp. 445–470.

[62] Margaret Lech, Lu-Shih Low, and Kuan Ee Ooi. "Detection and Prediction of Clinical Depression". In: *Mental Health Informatics*. Springer Berlin Heidelberg, 2014, pp. 185–199.

[63] Ying Yang, Catherine Fairbairn, and Jeffrey F. Cohn. "Detecting depression severity from vocal prosody". In: *IEEE Transactions on Affective Computing* 4.2 (2013), pp. 142–150.

[64] Sharifa Alghowinem, Rol Goecke, Michael Wagner, Gordon Parkerx, and Michael Breakspear. "Head Pose and Movement Analysis as an Indicator of Depression". In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, Sept. 2013, pp. 283–288.

[65] Loic Kessous, Ginevra Castellano, and George Caridakis. "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis". In: *Journal on Multimodal User Interfaces* 3.1 (Mar. 2010), pp. 33–48.

[66] Markus Kächele, Michael Glodek, Dimitrij Zharkov, Sascha Meudt, and Friedhelm Schwenker. "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression". In: *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM*. 2014, pp. 671–678.

[67] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information". In: *Image and Vision Computing* 31.2 (2013), pp. 137–152.

[68] Frank Sposaro, Justin Danielson, and Gary Tyson. "iWander: An Android application for dementia patients". In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, Aug. 2010, pp. 3875–3878.

[69] George Vavoulas, Matthew Pediaditis, Emmanouil G Spanakis, and Manolis Tsiknakis. "The MobiFall dataset: An initial evaluation of fall detection algorithms using smartphones". In: *13th IEEE International Conference on BioInformatics and BioEngineering*. IEEE, Nov. 2013, pp. 1–4.

[70] Vlad Oncescu, Matthew Mancusob, and David Erickson. "Cholesterol testing on a smartphone". In: *Lab on a Chip* 14.4 (2014), pp. 759–763.

[71] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study". In: *Journal of medical Internet research* 17.7 (2015), e175.

[72] Agnes Grunerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Ohler, Gerhard Troster, Oscar Mayora, Christian Haring, and Paul Lukowicz. "Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients". In: *IEEE Journal of Biomedical and Health Informatics* 19.1 (Jan. 2015), pp. 140–148.

[73] Anna Huguet, Sanjay Rao, Patrick J. McGrath, Lori Wozney, Mike Wheaton, Jill Conrod, and Sharlene Rozario. *A systematic review of cognitive behavioral therapy and behavioral activation apps for depression*. May 2016.

[74] Patricia A. Arean, Kevin A. Hallgren, Joshua T. Jordan, Adam Gazzaley, David C. Atkins, Patrick J. Heagerty, and Joaquin A. Anguera. "The use and effectiveness of mobile apps for depression: Results from a fully remote clinical trial". In: *Journal of Medical Internet Research* 18.12 (Dec. 2016), e330.

[75] Tara Donker, Petrie Katherine, Judy Proudfoot, Janine Clarke, Mary-Rose Birch, and Helen Christensen. "Smartphones for Smarter Delivery of Mental Health Programs A Systematic Review Donker Journal of Medical Internet Research". In: *J Med Internet Res* 15.2013 (Nov. 2013), pp. 1–19.

[76] Nelson Shen, Michael-Jane Levitan, Andrew Johnson, Jacqueline Lorene Bender, Michelle Hamilton-Page, Alejandro Alex R Jadad, and David Wiljer. "Finding a depression app: a review and content analysis of the depression app marketplace". In: *JMIR mHealth and uHealth* 3.1 (2015), e16.

[77] Bo Cheng, Mingxia Liu, Daoqiang Zhang, Brent C. Munsell, and Dinggang Shen. "Domain Transfer Learning for MCI Conversion Prediction". In: *IEEE Transactions on Biomedical Engineering* 62.7 (July 2015), pp. 1805–1817.

[78] Fredric O Finkelstein and Susan H Finkelstein. "Depression in chronic dialysis patients: assessment and treatment". In: *Nephrology Dialysis Transplantation* 15.12 (2000), pp. 1911–1913.

[79] David E Bush, Roy C Ziegelstein, Matthew Tayback, Daniel Richter, Sandra Stevens, Howard Zahalsky, and James A Fauerbach. "Even minimal symptoms of depression increase mortality risk after acute myocardial infarction". In: *The American journal of cardiology* 88.4 (2001), pp. 337–341.

[80] Mary A. Whooley, Peter de Jonge, Eric Vittinghoff, Christian Otte, Rudolf Moos, Robert M. Carney, Sadia Ali, Sunaina Dowray, Beeya Na, Mitchell D. Feldman, Nelson B. Schiller, and Warren S. Browner. "Depressive symptoms, health behaviors, and risk of cardiovascular events in patients with coronary heart disease". In: *Jama* 300.20 (2008), pp. 2379–2388.

[81] Janine Giese-Davis, Kate Collie, Kate M S Rancourt, Eric Neri, Helena C Kraemer, and David Spiegel. "Decrease in depression symptoms is associated with longer survival in patients with metastatic breast cancer: a secondary analysis". In: *Journal of clinical oncology* 29.4 (2011), p. 413.

[82] Marij Zuidersma, Henk Jan Conradi, Joost P van Melle, Johan Ormel, and Peter de Jonge. "Self-reported depressive symptoms, diagnosed clinical depression and cardiac morbidity and mortality after myocardial infarction". In: *International Journal of Cardiology* 167.6 (2013), pp. 2775–2780.

[83] Annelieke M Roest, Anne Heideveld, Elisabeth J Martens, Peter de Jonge, and Johan Denollet. "Symptom dimensions of anxiety following myocardial infarction: Associations with depressive symptoms and prognosis." In: *Health Psychology* 33.12 (2014), p. 1468.

[84] Omid G Sani, Yuxiao Yang, Morgan B Lee, Heather E Dawes, Edward F Chang, and Maryam M Shanechi. "Mood variations decoded from multi-site intracranial human brain activity". In: *Nature Biotechnology* 36.10 (Nov. 2018), pp. 954–961.

[85] Pedro Henriquez, Bogdan J. Matuszewski, Yasmina Andreu, Luca Bastiani, Sara Colantonio, Giuseppe Coppini, Mario D'Acunto, Riccardo Favilla, Danila Germanese, Daniela Giorgi, Paolo Marraccini, Massimo Martinelli, Maria-Aurora Morales, Maria Antonietta Pascali, Marco Righi, Ovidio Salvetti, Marcus Larsson, Tomas Stromberg, Lise Randeberg, Asgeir Bjorgan, Giorgos Giannakakis, Matthew Pediaditis, Franco Chiarugi, Eirini Christinaki, Kostas Marias, and Manolis Tsiknakis. "Mirror Mirror on the Wall... An Unobtrusive Intelligent Multisensory Mirror for Well-Being Status Self-Assessment and Visualization". In: *IEEE Transactions on Multimedia* 19.7 (July 2017), pp. 1467–1481.

[86] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. "MoodScope: building a mood sensor from smartphone usage patterns". In: *Proceeding of the 11th annual international conference on Mobile systems, applications, and services - MobiSys '13*. New York, New York, USA: ACM Press, 2013, pp. 389–402.

[87] Natasha Jaques, Sara Taylor, Ehi Nosakhare, Akane Sano, Elizabeth B. Klerman, and Rosalind Picard. "Importance of Sleep Data in Predicting Next-Day Stress, Happiness, and Health in College Students". In: *Journal of Sleep and Sleep Disorders Research* 40.suppl_1 (2017), A294–A295.

[88] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. "Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health". In: *IEEE Transactions on Affective Computing* 11.2 (2020), pp. 200–213.

[89] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. "Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, July 2019, pp. 2886–2894.

[90] Han Yu, Elizabeth B. Klerman, Rosalind W. Picard, and Akane Sano. "Personalized Wellbeing Prediction using Behavioral, Physiological and Weather Data". In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, May 2019, pp. 1–4.

[91] Alban Maxhuni, Pablo Hernandez-Leal, L. Enrique Sucar, Venet Osmani, Eduardo F. Morales, and Oscar Mayora. "Stress modelling and prediction in presence of scarce data". In: *Journal of Biomedical Informatics* 63 (2016), pp. 344–356.

[92] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. "Multi-task, multi-kernel learning for estimating individual wellbeing". In: *NIPS Workshop on Multimodal Machine Learning*. 2015.

[93] Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. "Multi-task Learning for Predicting Health, Stress, and Happiness". In: *NIPS Machine Learning for Health Care Workshop*. 2016.

[94] Priyank Jaini, Zhitang Chen, Pablo Carbajal, Edith Law, Laura Middleton, Kayla Regan, Mike Schaekermann, George Trimponias, James Tung, and Pascal Poupart. "Online Bayesian Transfer Learning for Sequential Data Modeling". In: *International Conference on Learning Representations (ICLR)*. 2017.

[95] Akane Sano, Sara Taylor, Andrew W McHill, Andrew Jk Phillips, Laura K Barger, Elizabeth Klerman, and Rosalind Picard. "Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study." In: *Journal of medical Internet research* 20.6 (June 2018), e210.

[96] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems*. 2014, pp. 3320–3328.

[97] Xinyang Li, Riccardo Poli, Gaetano Valenza, Enzo Pasquale Scilingo, and Luca Citi. "Self-reported Well-being Score Modelling and Prediction : Proof-of-Concept of an Approach based on Linear Dynamic Systems". In: *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'17)*. 2017, pp. 2205–2208.

[98] Zoubin Ghahramani and Geoffrey E. Hinton. *Parameter Estimation for Linear Dynamical Systems*. Tech. rep. University of Totronto, Dept. of Computer Science, Technical Report CRG-TR-92-2, 1996.

[99] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877.

[100] Jonathan Goodman and Jonathan Weare. "Ensemble Samplers With Affine Invariance". In: *Communications in Applied Mathematics and Computational Science* 5.1 (2010), pp. 65–80.

[101] David Huijser, Jesse Goodman, and Brendon J. Brewer. *Properties of the Affine Invariant Ensemble Sampler in high dimensions*. Sept. 2017. arXiv: arXiv:1509.02230v2.

[102] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. "Bayesian Model Averaging: A Tutorial". In: *Statistical Science* 14.4 (1999), pp. 382–401.

[103] Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. "Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database". In: *Critical Care Medicine*. Vol. 39. 5. Lippincott Williams and Wilkins, 2011, pp. 952–960.

[104] Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric statistical methods*. Vol. 751. John Wiley & Sons, 2013.

[105] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and Harry Eugene Stanley. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." In: *Circulation* 101.23 (2000), e215–e220.

[106] Luca Citi and Riccardo Barbieri. "PhysioNet 2012 Challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm". In: *2012 Computing in Cardiology*. IEEE, 2012, pp. 257–260.

[107] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. "emcee : The MCMC Hammer". In: *Publications of the Astronomical Society of the Pacific* 125.925 (Mar. 2013), pp. 306–312.

[108] Radford M Neal. "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*. Ed. by Galin Jones, Steve Brooks, Andrew Gelman, and Xiao-Li Meng. 2011. Chap. 5, pp. 113–162.

[109] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. "Stan: A probabilistic programming language". In: *Journal of statistical software* 76.1 (2017).

[110] Matthew D. Hoffman and Andrew Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1593–1623.

[111] Aapo Hyvärinen. "Estimation of Non-Normalized Statistical Models by Score Matching". In: *Journal of Machine Learning Research* 6 (2005), pp. 695–709.

[112] Aapo Hyvärinen. "Some extensions of score matching". In: *Computational statistics & data analysis* 51.5 (2007), pp. 2499–2512.

[113] Siwei Lyu. "Interpretation and generalization of score matching". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2009, pp. 359–366.

[114] Martin Andersen, Joachim Dahl, Zhang Liu, and Lieven Vandenberghe. "Interior-point methods for large-scale cone programming". In: *Optimization for machine learning*. The MIT Press, 2011. Chap. 3, pp. 55–83.

[115] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Burkner. "Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC". In: *Bayesian Analysis* (2020).

[116] Richard J. Davidson and Sharon Begley. *The emotional life of your brain : how its unique patterns affect the way you think, feel, and live–and how you can change them*. Hudson Street Press, 2012, p. 279.

[117] Pelin Kesebir, Agata Gasiorowska, Robin Goldman, Matthew J. Hirshberg, and Richard J. Davidson. "Emotional Style Questionnaire: A Multidimensional Measure of Healthy Emotionality". In: *Psychological Assessment* 31.10 (Oct. 2019), pp. 1234–1246.

[118] Pasquale K. Alvaro, Rachel M. Roberts, and Jodie K. Harris. "A Systematic Review Assessing Bidirectionality between Sleep Disturbances, Anxiety, and Depression". In: *Sleep* 36.7 (July 2013), pp. 1059–1068.

[119] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (Jan. 1995), pp. 289–300.

[120] Alistair E W Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3.1 (2016).