Monitoring and modelling of behavioural changes using smartphone and wearable sensing

Kamronn, Simon Due

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Kamronn, S. D. (2018). Monitoring and modelling of behavioural changes using smartphone and wearable sensing. Kgs. Lyngby: Technical University of Denmark. DTU Compute PHD-2018, Vol.. 489

Monitoring and modelling of behavioural changes using smartphone and wearable sensing

Simon Kamronn

Kongens Lyngby 2018

DTU Compute
Department of Applied Mathematics and Computer Science

Monitoring and modelling of behavioural changes using smartphone and wearable sensing

Simon Kamronn
simon@kamronn.com

Technical University of Denmark
Department of Applied Mathematics and Computer Science

DTU Compute
Richard Petersens Plads
Building 324
2800 Kgs. Lyngby
DENMARK

Tel. +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk
PHD: ISSN 0909-3192

Summary

Increase of sedentary behaviour and obesity has been on the rise for a score of years or more, despite public information campaigns and even the incursion of the latest fad, fitness trackers. The latter were by industry heralded as life–changers that by simple mechanisms would change the behaviour of the wearer to be more active and more healthy. Studies have since shown that they may have an initial positive effect on activity levels and reduced weight, but that it quickly falters and people stop using the trackers altogether. A reoccurring observation seem to be a misunderstanding of what drives human motivation and what it takes to change human behaviour with respect to physical activity. Being reminded of your weight or steps taken throughout the day, is for most people but a mere observation, not an intervention. This misunderstanding, or naïvety, probably stems from conclusions that are drawn from data that are too thin to support them. We propose a paradigm that relies on massive amounts of data, pervasively sampled from smartphones.

We show that smartphone data is able to estimate plausible intervention effects from a randomized controlled trial, and through higher sampling frequency and additional modalities, is able to break up the estimated effects into contextual pieces that can be used to better understand behavioural aspects. We further show that by using a model that adapts to each individual, we can predict a persons total energy expenditure accurately from the same data.

A novel model to recognise human activity semi–supervised and from multiple datasets, is presented. The model combines convolutional neural networks to extract hierarchical features and recurrent neural networks to model temporal dependencies. This is combined with recent developments in domain adaptation where domain separation is penalised through adversarial training of an auxiliary classifier.

Lastly a model is presented that fully unsupervised is able to learn latent states that naturally decompose into static and dynamic representations. The static representations are learnt as a function that maps a high–dimensional observation into a low–dimensional code that is dependent on a structured prior distribution that governs the dynamical system.

Resume

Der er sket en stigning i stillesiddende adfærd og fedme de sidste 20 år, på trods af offentlige kampagner og udbredelsen af den seneste teknologi, fitness trackere. Sidstnævnte blev af industrien udråbt til at have livsstilsændrende egenskaber og kunne med simple metoder gøre brugeren til et mere aktivt og sundere menneske. Studier har vist, at selvom de har en umiddelbar positiv effekt på fysisk aktivitet og reduceret kropsvægt, så svinder effekten hurtigt igen og folk stopper med bruge trackerne. En gentagen observation er misforståelsen af hvad der driver motivation og hvad der kræves for at ændre menneskelig adfærd med hensyn til fysisk aktivitet. At blive mindet om ens vægt eller hvor mange skridt man tager i løbet af dagen, er for de fleste bare en observation, ikke en intervention. Misforståelsen stammer sandsynligvis fra konklusioner der er draget på et grundlag af for tyndt data. Vi foreslår et paradigme der baseres på store mængder data, som er opsamlet ubemærket gennem smartphones.

Vi viser at smartphone data kan estimere plausible interventionseffekter fra et randomiseret kontrolleret forsøg og med højere sampling frekvens og flere modaliteter, kan opdele effekterne i kontekstuelle dele, der kan bruges til bedre at forstå aspekter af menneskelig adfærd. Vi viser yderligere at vi med samme data kan bruge en personligt kalibreret model til at estimere totalt energiforbrug.

En ny model til klassificering af fysisk aktivitet fra delvist annoteret data fra flere kilder er desuden præsenteret. Modellen bruger dybe neurale netværk til udtrække at hierarkiske repræsentationer og modellere den temporale udvikling i tidsserien. Det er kombineret med nylig udvikling indenfor domæne adaptation hvor domæne separering er straffet ved "adversarial" træning af en ekstra klassificerings opgave.

Vi præsenterer til sidst en model der uden brug af annotering kan lære latente underrum der naturlig dekomponerer i statiske og dynamiske repræsentationer. Den statiske repræsentation er lært som en funktion der transformerer en høj–dimensionel observation ned i et lav–dimensionelt rum, der er afhængigt af en struktureret *a priori* sandsynlighedsfordeling, som styrer det dynamiske system.

Preface

This thesis was prepared at the Section for Cognitive Systems, Department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfillment of the requirements for acquiring a Ph.D. degree in engineering.

The thesis consist of a summary report and a collection of three papers: one conference paper and two submitted journal papers. The work was conducted from March 2015 to August 2018

Kongens Lyngby, August 19, 2018

Simon Kamronn

Simon Kamronn

Acknowledgements

I have been privileged to work in the friendly and encouraging environment that the Cognitive Systems department provide, and daily challenged and inspired by all the great minds that inhabit its space.

I thank my supervisor Jakob Eg Larsen for providing a guiding light in especially the many frustrating situations when nothing works, and my co–supervisor, Lars Kai Hansen, for painstakingly explaining probabilistic concepts so that even I may understand, and editing my papers so that others may as well.

I am thankful to all my colleagues, who are much smarter than I, for our stimulating discussions on Bayesian Statistics, Deep Learning, and the many things that fall between. Without you I would not have become so side–tracked and learned so much.

I thank my collaborators for believing in me and our projects, even when the prospects looked bleak.

To the GO-ACTIWE team at the Department of Biomedical Sciences, Copenhagen University, and their leader Bente Stallknecht, I am grateful for all your hard work in conducting the trial and taking the time to explain many aspects of human physiology.

I am thankful for the financial support from Trygfonden, DTU Compute, and Otto Mønsteds Fond.

To my family, especially Terese and Linn, I am grateful for your unwavering support.

List of publications

Included in thesis

Peer reviewed

Mads Rosenkilde, Martin Bæk Blond, Anne Sofie Gram, Jonas Salling Quist, Jonas Winther, **Simon Kamronn**, Desirée Hornbæk Milling, Jakob Eg Larsen, Astrid Pernille Jespersen, and Bente Stallknecht. (2017). *The GO-ACTIWE Randomized Controlled Trial - An Interdisciplinary Study Designed to Investigate the Health Effects of Active Commuting and Leisure Time Physical Activity.* Contemporary Clinical Trials 53 (February): 122–29.

C Marco Fraccaro, **Simon Kamronn**, Ulrich Paquet and Ole Winther. (2017). *A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning.* Advances in Neural Information Processing Systems 30, NIPS 2017.

Manuscripts

A **Simon Kamronn**, Lars Kai Hansen and Jakob Eg Larsen. (2018). *Smartphone pervasive sensing of physical activity of overweight adults in a long-running randomized controlled trial.*

B **Simon Kamronn**, Jonas Salling Quist, Martin Bæk Blond, Anne–Sofie Gram, Kevin Hall, Peter Walter, Mads Rosenkilde, Bente Stallknecht and Jakob Eg Larsen. (2018). *Estimating energy expenditure from smartphones. A free-living long-term comparison with doubly labelled water.*

Not included in thesis

Andreas Trier Poulsen, **Simon Kamronn**, Jacek Dmochowski, Lucas C. Parra and Lars Kai Hansen. (2017).*EEG in the classroom: Synchronised neural recordings during video presentation.* Scientific Reports, 7, 1–9.

Contents

CHAPTER 1 Introduction

1.1 Motivations

"Sitting is the new smoking" and while the associated risks of too little physical activity are widely known (Patel et al., 2010), the obesity pandemic is ever-increasing (Badimon et al., 2017). A recent survey on the Danish health announced that 51% of the population are now overweight (up from 46.8% in 2010) and 16.8% are obese. Not including associated non-communicable diseases such as diabetes and cardiovascular diseases (Ding et al., 2016), obesity is estimated to incur a cost of 12.2 billion Danish kroner annually in healthcare and lost production (Jensen et al., 2018).

Excess weight is due to an imbalance between physical activity and dietary intake (Maher et al., 2013). Accordingly, the survey reports increased unhealthy eating habits (13.3% in 2010 to 15.9% in 2017) and a lack of physical activity. The World Health Organisation (WHO) recommendation of at least 150 minutes of moderate or 75 minutes of vigorous weekly activity are only met by 71.2% of the population and almost 60% are sedentary more than 8 hours a day. Of the ones not meeting the WHO recommendation, 71% wants to be more active. Engaging the inactive and overweight in regular physical activity is an obvious way to reduce overweight and obesity, but how to motivate a sustained increase, is still largely an open question.

Adherence to structured exercise programs generally deteriorates quickly (Seefeldt et al., 2002) and it has been found that individuals participating in a 6-week physical activity program relapsed into pre-program behaviour after 20 weeks (Gomersall et al., 2015). Integrating habitual physical activity is difficult as time usage needs to be allocated from existing habits. Habits that may be stronger or more alluring even after a long period. To sustain long-term changes it may be beneficial to be critical about what kind of activities that are displaced by an exercise program. Substituting passive transport with active transport (e.g. biking to work (Mytton et al., 2016)) will for example displace other activities to a lesser degree (Gomersall et al., 2015).

However, only a small percentage of modern civilisation are able or willing to change mode of transportation so leisure time exercise is still the more important determinant of an active lifestyle. As the forces of old habits and hedonism overcome routines learnt in exercise programs as soon as the program has ended, one solution could be a program that does not end. Internet based (eHealth) interventions have shown modest but significant effects of increased physical activity (Davies et al., 2012) but whether the interventions have long-term effects, is still uncertain (Kohl et al., 2013). Compared to traditional face-to-face interventions, eHealth methods are substantially cheaper and may have large effects on society, when applied to the scale of populations. Smartphone based (mHealth) interventions are potentially more effective than eHealth (Carter et al., 2013) but there is unfortunately no conclusive evidence that smartphone apps are effective in encouraging physical activity as the literature primarily consist of underpowered short-term studies (Stuckey et al., 2017; Direito et al., 2017) despite being a very active research area (Müller et al., 2018a). Future research should focus on large randomized trials with long-term follow-up and include evidence-based components, innovative mHealth apps, and *accurate physical activity monitoring* (Reis et al., 2016; Lewis et al., 2017).

1.2 Measuring causal effects

Observing how a system responds to an external stimuli is a way to learn how a system works. Monitoring people is similarly at the heart of understanding humanity; what characterises certain behaviours and how are they changed. Research in disciplines concerned with physiological or psychological aspects are built around the paradigm of changing a single parameter in the life of a group of people and measuring the response.

If the parameter is independent of the people in the group, it is possible to establish a causal link between the induced change and the measured response. Ensuring conditional independence between the people selected and the changed parameter (the intervention) is difficult as the dependence structure is often complex and impossible to measure. For this reason another group is formed that does not experience the intervention, a control group. If the groups are homogenous it is no longer necessary to ensure conditional independence between the intervention and the people that are intervened. The difference in the measured responses of the two groups will be the average causal effect of the intervention. This paradigm is often referred to as a randomised controlled trial and considered a gold standard in measuring causal effects.

However, a measured effect is only supported within the distribution of personal characteristics that is represented in the group of people, that the effect was measured on. If an intervention effect is only established on males, it is not necessarily transferable to females and it could potentially be dangerous to assume so. Uncertainty of an effect is similarly depended on the number of people in the group and the more varied the personal characteristics are, the more people are needed to establish the effect of an intervention with sufficient certainty. To measure a causal effect with certainty that generalise to everyone therefore requires a lot of people to measure on and preferably for a long time.

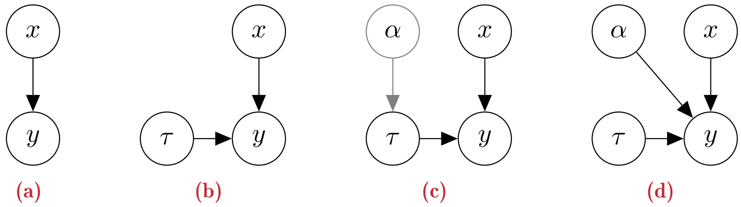


Figure 1.1: Generative models with varying number of parameters and dependency structures. Self–dependency is omitted.

1.3 Statistical models of effects

Probabilistic graphical models (Koller and Friedman, 2009) provide a general framework to define statistical relationships and a way to separate model definition from inference (i.e. parameter estimation). Most models can be expressed as a directed acyclic graph (DAG) which is built from variables with a dependency structure that often reflect the generative process that a priori is believed to have generated the observed data. An example of such a generative process could be the relationship between step counts from a pedometer and bodyweight measurements,

$$y_i = y_{i-1} + \beta_0 - \beta_1 x_i, \tag{1.1}$$

where y is bodyweight, x is step counts, and i denotes a discrete time index. In this model it is assumed that the energy intake (β_0) and energy expenditure per step (β_1) are constant so that weight gain or loss is entirely dependent on how many steps that are taken. A graphical representation of this model is illustrated in Figure 1.1(a). If the data was collected as part of a study inspecting the effects of a new diet, we would expect the bodyweight, y, to differ between the group receiving the diet and the one that did not. A random variable should be included in the model to reflect that

$$y_i = y_{i-1} + \beta_0 - \beta_1 x_i + \boldsymbol{\tau}_{t[i]}. \tag{1.2}$$

Here $\boldsymbol{\tau}_{t[i]}$ models the causal effect of the diet and $t[i]$ indicates whether sample i is from the group receiving the diet or not (Figure 1.1(b)). Additional knowledge can be included in the model to make it fit better to the observed data. It could e.g. be useful to include gender, age, or height, so (1.2) can be expanded to

$$y_i = y_{i-1} + \beta_0 - \beta_1 x_i + \boldsymbol{\tau}_{t[i]} + \boldsymbol{\alpha}_{s[i]}, \tag{1.3}$$

where $\boldsymbol{\alpha}_{s[i]}$ is a random variable with s levels corresponding to the number of possible values it can assume, i.e. two if it represents gender. Conditioning on a variable removes the *confounding* effect that variable has on the response, y, from the other parameters in the model. If height or gender are predictive of energy expenditure and the distribution of these characteristics are imbalanced between the control and intervention group, the variance explained by those variables would implicitly be modelled in the estimated intervention effect, $\boldsymbol{\tau}$, which is the case illustrated in Figure

1.1(c). By including them in the model, that dependency is removed (Figure 1.1(d)). This type of model is known as a hierarchical Bayesian model if all parameters are treated as random variables with prior distributions (Wainwright and Jordan, 2008).

1.4 Pervasive sensing

Pervasive, or ubiquitous, are adjectives used to describe things that are widespread and permeates the surroundings. In western society that is certainly true for the smartphone which is why it is so powerful as a platform to conduct research on (Raento et al., 2009; Althoff et al., 2017). Not only is it possible to collect data from thousands of people worldwide, it is also easier and cheaper to recruit and enroll participants (Chan et al., 2017). The pervasiveness also serves another purpose, however. Establishing whether a measured effect is caused by an intervention typically requires everything else to remain constant. Introducing an instrument to measure the effect can trigger a psychological response affecting the behaviour that is being measured (Sullivan and Lachman, 2016). Unless the instrument is part of the intervention it will be impossible to recover the actual intervention effect. Collecting data on a smartphone thus has the benefit of not affecting the result as long as the behaviour regarding the smartphone remains unchanged, thereby offering improved *ecological validity* (Raento et al., 2009).

1.5 Outline and contributions

This thesis is structured as a summary of the main results presented in the papers in the Appendix with some additional background theory on the used methods, and work that was not included in the papers.

Chapter 2 introduces the randomized controlled trial of which this project has been part of and the technical aspects in monitoring participants in real–time with smartphones.

Chapter 3 introduces the method of measuring physical activity with smartphones and presents the main results of the intervention effects which are then compared to changes in fat–free mass measurements. The second part of the Chapter introduces a novel method to estimate energy expenditure by combining multiple independent data sources.

Chapter 4 introduces a method to identify significant locations using a spatio–temporal clustering algorithm, how to infer the context of these, and their relation to physical activity. The second part introduces a novel semi–supervised method to classify human activity from raw accelerometer signals using deep neural networks.

Chapter 5 introduces the *Kalman variational auto–encoder*, a model that by combining a structured temporal prior distribution and a variational auto–encoder, is able to learn disentangled representations of static and dynamic factors. The model is an efficient simulator that is able to fit high–dimensional complex observations.

Chapter 6 summaries the discoveries, concludes, and highlights some interesting future directions.

CHAPTER 2
The Study

The objective of the study was to investigate the health effects of physical activity during leisure time and active transport in overweight and obese individuals. The primary outcomes were metabolic health, cardiovascular health, and energy balance measured with insulin sensitivity, endogenous thrombin potential, and total body fat, respectively. In a subgroup it was furthermore the aim to investigate the use of smartphones to pervasively monitor physical activity and estimate energy expenditure in different domains of everyday life. It is that subgroup that is the focus of this thesis. The following sections describe the RCT design, technical aspects of pervasive smartphone sensing, and some of the challenges discovered during the study.

2.1 Methods

This study has been carried out by the GO–ACTIWE team of which most are residing at the Biomedical Sciences department, Faculty of Health and Medical Sciences, University of Copenhagen. The group is lead by professor Bente Stallknecht and principal investigator Mads Rosenkilde. Group members include doctoral students Jonas Salling Quist, Martin Bæk Blond, Anne Sofie Gram, and Jonas Winther, and laboratory scientist Desirée Hornbæk Milling.

2.1.1 Participants

Overweight and obese Caucasians were in the period of November 2013 to October 2015 recruited from the greater Copenhagen area. Inclusion criteria were physical inactivity, 20-45 years of age, a body mass index from 25 - 35 kg m^{-2}, non-smoking with a normal plasma glucose ($<$ 6.1 mmol/l) and blood pressure ($<$140/90 mm Hg). All the criteria are listed in Table 2.1. Participants were recruited using internet advertisements (social media, forsøgsperson.dk, etc.) and the local media. Recruits were screened with an online questionnaire and a rigorous medical examination. The study was approved by the ethical committee of The Capital Region of Denmark (H-4-2013-108), registered at the Danish Data Protection Agency and at clinicaltrials.gov

Inclusion criteria	Exclusion criteria
• 20 - 45 years old • Healthy • Physically in-active • Body mass index: 25–35 kg m^{-2} • Caucasian	• Body fat percentage $< \begin{cases} 32\% & \text{for women} \\ 25\% & \text{for men} \end{cases}$ • VO2peak $> \begin{cases} 40\text{ml } O_2\text{kg}^{-1}\text{min}^{-1} & \text{for women} \\ 45\text{ml } O_2\text{kg}^{-1}\text{min}^{-1} & \text{for men} \end{cases}$ • Fasting plasma glucose > 6.1 mmol/l • Blood pressure > 140/90 mm Hg • Abnormal ECG • Type 2 diabetes • Pregnancy or planning of pregnancy • Smoking • Medicine use

Table 2.1: Inclusion and exclusion criteria.

(identifier: NCT01962259 and NCT01973686) and adhered to the principles of the Helsinki declaration. Further information is available in Rosenkilde et al. (2017).

2.1.2 Study design

The study was designed as a 6-month randomised controlled trial, which was described in Section 1.2, with three interventions; an active bicycling commuting group (BIKE) and two leisure time physical activity (LPTA) groups (MOD, VIG) as well as a non-exercise control group (CON). The LPTA groups are defined by the exercise intensities of 50% (MOD) and 70% (VIG) of the peak oxygen uptake (VO_2peak) during exercise. Daily exercise expenditure was prescribed to 320 kcal for women and 420 kcal for men with five sessions per week. Participants were advised to not change their eating habits and were invited to three test periods at baseline, 3 months, and 6 months where each test period consisted of three test days. A two–week free–living assessment period followed each test period in which the total energy expenditure was measured with doubly labelled water (DLW) (Westerterp, 2017) and physical activity by accelerometry (Actigraph GT3x, Actigraph Corp., Pensacola, Florida, USA). Participants were told to use a heart rate monitor (RC3 GPS, Polar Electro Oy, Kempele, Finland) during all exercises and to weigh themselves every day using a wireless internet-connected bodyweight (WiThings Body Composition WiFi scale, WiThings Europe,Issy-les-Moulineaux, France).

Doubly Labelled Water

Doubly labelled water is the most accurate method to measure human energy expenditure in free–living scenarios and often the only viable method. It is, however, rather expensive as the analysis procedure requires highly specialised equipment. The procedure is as follows: a dose of water labelled with ^{18}O and ^{2}H is given to

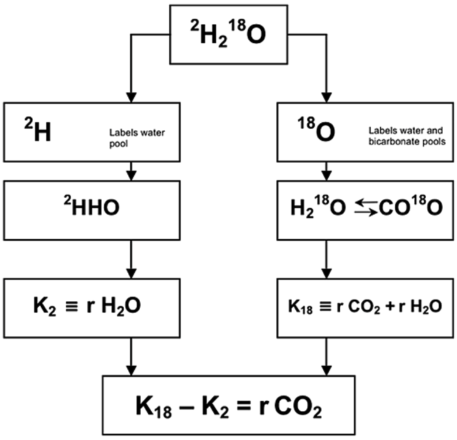

$^2H_2{}^{18}O$
^2H
Labels water pool
^{18}O
Labels water and bicarbonate pools
^2HHO
$H_2{}^{18}O \leftrightarrows CO^{18}O$
$K_2 \equiv r\,H_2O$
$K_{18} \equiv r\,CO_2 + r\,H_2O$
$K_{18} - K_2 = r\,CO_2$

Figure 2.1: Illustration of the process behind the doubly labelled water method to measure energy expenditure. Borrowed from Westerterp (2017).

a participant whom then samples urine every morning the following two weeks. In these samples the enrichment of ^{18}O and ^{2}H are assessed with isotope ratio mass spectrometry to establish the decay rates of the isotopes. The difference of these rates are a function of CO_2 production, as depicted in Figure 2.1, which is directly correlated with energy expenditure.

2.2 Smartphone

Participants received a smartphone (Nexus 5X, LG Electronics, Seoul, South Korea or Galaxy S5, Samsung Electronics Ltd., Suwon, South Korea) to which a custom application had been installed to acquire sensor data. The application was adopted from previous studies (Aharony et al., 2011; Stopczynski et al., 2014) and extended with measuring accelerometry (50 Hz sampling frequency) and activity recognition provided by the internal classifier through the Google Activity Recognition API. The application furthermore collected GPS-based location (6–min sampling interval), step count, and display on/off events. A timeline of when the different data modalities are collected is depicted in Figure 2.2. All data were stored in encrypted SQLite3 databases and uploaded securely to an in–house server when the smartphone had WiFi access. Participants were instructed to use the smartphone as their primary device and carry it as often as possible but otherwise not change any routines. An armband was provided should they want to bring the smartphone when exercising. Assistance was provided in case of technical issues and a replacement device provided in case of hardware failure.

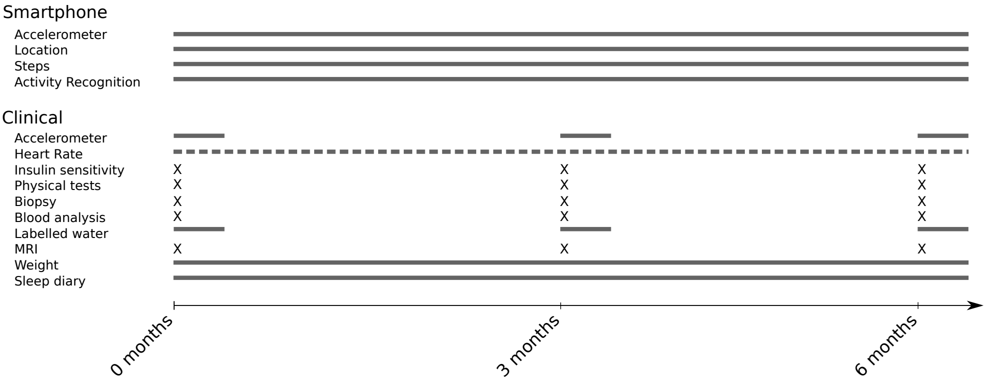


Figure 2.2: Timeline of when the different modalities are collected during the 6 months study.

2.2.1 Data management

Server-side infrastructure was adopted from a previous study in which it had successfully handled smartphone sensing data from 1000 smartphones simultaneously (Stopczynski et al., 2014). Incoming SQLite files were decrypted and all data except for accelerometry was put into a MongoDB database. Accelerometry was found to be too big in size to efficiently handle in a conventional database and therefore stored in a chunked and compressed format designed for time-series called BColz. Around one terabyte of smartphone data was collected in the study. All data were kept exclusively on the servers and computations carried out either through Jupyter Notebooks hosted on the server or with scripts executed through SSH.

2.2.2 Data processing

Data analysis is an interactive discipline where understanding comes from repeated manipulation and visualisation. With large amounts of data it quickly becomes time-consuming to do simple calculations, partly because the data does not fit into the memory of the computer. To manage out-of-memory computations the Python library Dask was used as it is built around a lazy computational graph that can do computations on chunks of data and then aggregate the partial computations. Statistical analyses were done in probabilistic programming libraries Stan (Carpenter et al., 2017) or Edward (Tran et al., 2016) depending which library that fitted the task best. Stan is good for exact inference using sampling methods and Edward is better when flexibility and scalability is needed. Visualisations are created with Altair (Vanderplas and Granger, 2018) and Matplotlib (Hunter, 2007).

2.3 Challenges

It is not without peril that one employs consumer products in research, which was acknowledged from the beginning of the study, and is a the heart of one question this project tries to answer: can smartphones be used to predict energy expenditure? And if so, how accurately. Some margin of relatively big measurement error is tolerated in this exploratory paradigm and while we a priori can try to reason about a possible scale, we must ultimately accept the collected measurements as the current limit.

In research it is not uncommon to experience malfunctions and commit errors that affect the research process in a negative way. Expected errors can be mitigated by incorporating strict routines in e.g. manual data acquisition and be reduced by incorporating systematic checks on data integrity in automatic data acquisition scenarios, such as sampling from smartphone sensors. The latter can, however, present a major workload and be stressful, and therefore a balance should be found in which some errors are accepted.

The cohort receiving a smartphone and doubly labelled water is a subgroup of a larger cohort because the funding for this project was received later. To include as many participants as possible, the enrollment had to start very quickly which meant that there was no time for a pilot study to test the equipment. Furthermore, once enrollment had begun, the study paradigm could not be changed. A pilot study and subsequent efforts to stablise the software, and perhaps adjust the sampling paradigm, would probably have made a significant difference to data integrity.

2.3.1 Missing data

A sampling frequency of 50 Hz paints a very clear picture of when data is missing and in the accelerometry data an observation rate of 30.5% is observed, which is lower than what was expected. Three patterns emerge when inspecting the data; blocks of hours are missing for all participants, blocks of days are missing for some, and all have less data during night time. The latter is not an issue and explainable with participants turning off the device and perhaps some kind of power preserving feature of the smartphones. The missing blocks are, however, still unexplained and the faults may lie in the smartphone, server-side, or both. It is not always that all modalities are missing which may indicate sensor specific issues on the smartphone. Most participants experienced issues with insufficient storage when not connected to WiFi for a prolonged period (days) and some participants had frequently issues with the application not uploading data, even when connected to WiFi. When storage is full, no more data is recorded, which may be the reason behind longer gaps.

Specific issues

Software projects normally use version control tools such as git to document code changes and ensure a local copy is always up–to–date with the latest version. In this project a part–time programmer unfortunately made an update to the Android application in which old source code was used. In this particular code, the timestamp calculation was wrong, which affected all samples acquired while the error persisted, which it did for about a week.

After a power outage at the university hosting the server, the harddrive containing the data was irrecoverably damaged. Fortunately most of the data was backed up in raw SQLite3 files and the database was reconstructed after a few weeks.

2.3.2 Compliance

As the study relied on natural behaviour regarding smartphone use, it is only neglect that was monitored for. Though participants agreed to use the smartphone as their primary, a few stopped using it in the beginning of the study. They where also asked to carry the smartphone during exercises but a cross-reference with the collected heart rate data, show that very few chose to do so.

2.3.3 Software

The software adapted from Stopczynski et al. (2014) was designed in a modular way so that it was easy to add new sensor probes and through a central configuration file, specify which sensors to sample, and how often. Probes could either be sampled with a fixed interval, e.g. location every 5 minutes, or adaptively using a hook to capture broadcasts of new samples, e.g. if another application requests location updates every second, they are also acquired my our application. For the application to request a sensor for a new sample, the smartphone has to be awake, which consumes a lot of power. Accelerometer samples arrive every 20 millisecond, so utilising the existing framework required the smartphone to be active permanently, which it can only be for a few hours. The LG Nexus 5 was one of the first smartphones to include an accelerometer with a first–in–first–out circular buffer in which the most recent 10.000 samples are stored. By enabling that buffer through the Android framework, the application could wake up every 6 minutes and read out the entire buffer of accelerometer samples and go to sleep again, reducing the power consumption drastically. The buffer has a fixed size which also means that old samples will be overwritten as new arrive and failing to read the buffer, results in missing data.

On the server–side it proved very difficult to manage these amounts of data as we had no previous experience doing so and were not equipped with modern infrastructure, e.g. Spark and Hadoop. The final solution was settled on after many failed attempts and is not suited to ensure data integrity, but provides fast, parallel processing.

2.3.4 A rant on smartphones

Modern smartphones are in fact full–fledged computers crammed into pocket-sized containers equipped with huge screens, cameras, fingerprint readers, and various sensors. If this accomplishment is not enough to awe, then consider the perspective of the operating system. Hundreds contribute to the Android Open Source project and it has to support hundreds of different smartphone platforms. This process can be somewhat controlled but then add all the applications written by programmers of questionable competence competing over the users attention and the smartphones resources. In this vast space of possible failures it should seem a wonder that anything works but the answer is simple; error handling. Every second the operating system handles errors from many background processes and running applications. Users may not notice unless an application crashes but if the platform is unstable, background processes may also become so. That is at least an explanation to why this study experienced a high loss of data. Fortunately the smartphones being produced at the time of this writing is already much better than those deployed in the study.

CHAPTER 3
Physical Activity & Energy Expenditure

The many failed attempts to curb the obesity pandemic (Lee et al., 2012) raises the question of whether behavioural aspects of obesity and physical inactivity are understood properly. Most research into physical activity is through underpowered studies with too few participants for too short a time in artificial settings, which persistently have been shown not to generalise to the real world (Jeran et al., 2016). To disentangle the myriad of interacting factors governing physical activity, it is time to move the focus from the individual to *population-scale pervasive health* (Kohl et al., 2012; Hallal et al., 2012; Althoff, 2017). Historically, these factors; behavioural, social, etc., have been difficult to measure and studies have been conducted through self–reported surveys which are biased and only provide sparse, summarising information. To fully capture the spectrum of health–related behaviours, a dense sampling scheme is required, yielding high resolution, multi–modal data from many people over a long time–span. Mobile phones are presently the only realistic option and have the added benefit of providing a powerful platform capable of much more than measuring physical activity (Althoff et al., 2017). With this study we try to move beyond the large body of existing research that uses dedicated monitors (Evenson et al., 2015; Clawson et al., 2015), to investigate the potential of the smartphone as a ubiquitous device to pervasively monitor energy expenditure in free–living, accurately (Maddison et al., 2017). The pervasiveness of smartphones in the population offers an unparalleled reach (Campbell et al., 2008) and as the smartphone is already intimately accepted, it offers an unadulterated view of human behaviour, contrary to introducing a foreign device (Raento et al., 2009). Section 3.1 will summarise the paper in Appendix A, particularly the analysis of intervention effects measured by physical activity derived from the smartphone. Section 3.2 will summarise the paper in Appendix B which investigates the feasibility of predicting total energy expenditure with smartphones.

3.1 Physical activity

A perfect measure of physical activity would be the energy expended by muscles but we are left to deal with what is actually measurable, such as acceleration. Ways to quantify physical activity based on acceleration equals the number of different devices to measure it. Actigraph (Actigraph Corp., Pensacola, Florida, USA) popularised the *activity count* as a measure of physical activity, which originated from the mechanistic internals of the first generation devices, and the company has for the sake of consistency tried to emulate that process in its digital offspring. The precise algorithm is proprietary and while the individual steps have been largely discovered (Peach et al., 2014), important details are still undisclosed. This unfortunately muddles analyses based on activity counts as choices such as the frequency range of a band-pass filter can change results significantly. With the purpose of being able to compare to Actigraph and scientific literature, the Actigraph algorithm was in this study approximated using simultaneously captured data from smartphones and Actigraph.

3.1.1 Quantifying activity in counts

Having a time–series of accelerometer measurements, the objective is to calculate a value that represents physical activity independent of the person, activity, or any other circumstantial factors. The algorithm is simply to band–pass filter the signal to remove noise and then sum the absolute values in bins of one second. Each activity count is thus a summary of the activity in that second. The vector magnitude (i.e. root mean square in a euclidean space) is then calculated to remove the directional component of the triaxial accelerometer. A seemingly easy process to copy but the choice of band–pass filter and its parameters are very important. To approximate the activity count function, from smartphone data to Actigraph counts, two approaches were tried: Bayesian Optimisation and deep neural networks. I'll shortly describe each approach but not go into details as neither method worked well and the results are not used further.

Bayesian Optimisation

Considering the function $f(x)$ we are interested in finding the optimal hyperparameters from some set $\mathcal{X} \in \mathrm{R}^{\mathrm{D}}$ of interest. To do this we can treat $f(x)$ as a probabilistic model with a prior distribution defined by the Gaussian Process (GP, Rasmussen (2004)). Using evaluations of $f(x)$ as data we can get the posterior distribution over functions; the acquisition function $a(x)$. With the acquisition function it is possible to determine which hyperparameters that are expected to yield the best result and therefore should be evaluated next. See Snoek et al. (2012) for further details.

This method was used to find the optimal hyperparameters to the aforementioned simple algorithm to calculate activity counts.

Parameter	Value
Filter type	Butterworth
Lower cut off	0.2 Hz
Higher cut off	5 Hz
Order	6

Table 3.1: Activity count algorithm parameters.

Deep neural networks

Deep neural networks are a form of general purpose function approximators that have proven efficient in fitting models to raw data, i.e. without pre-processing such as filtering or feature extraction such as calculating summary statistics. A second of raw accelerometer data, $\mathbf{X} \in \mathbb{R}^{3\times 50}$, was given as input to the model that tried to predict the corresponding activity count from the Actigraph algorithm. Different neural network architectures were tested in a grid-search. More background theory on neural networks are provided in Section 4.2.2.

Results

Neither method achieved sufficient accuracy to either 1) calculate Actigraph counts from raw Actigraph accelerometer data or 2) to estimate Actigraph counts from raw smartphone accelerometer data. It is a bit surprising that 1) did not work well but 2) was less likely to succeed as the smartphone is carried in various ways whereas the Actigraph is mounted on the body. Since it henceforth is only possible to compare relative values of physical activity from Actigraph and smartphone, the parameters of the activity count algorithm used for the smartphone data are set to capture human motion up to a frequency of 5 Hz (Table 3.1).

3.1.2 Pre–processing

Activity counts are computed in intervals of one second, but that resolution is still too high to efficiently do posterior inference. A suitable resolution was heuristically found to be intervals of 5 minutes in which the activity counts are summed and the amount of missing data is averaged. As the weather possibly have an impact on the physical activity level, hourly precipitation and temperature data at three weather stations in and around Copenhagen, was acquired from the Danish Meteorological Institute and mapped to each sample, using time and location. Temporal information (hour, day, week, month) and information on how many off– and sick–days the participants experienced in each period of the study, were included together with the baseline characteristics: age, gender, BMI, VO_2peak, education, job status, civil status, and the type of smartphone.

Participants with less than 50 days containing data were excluded and samples with 90% missing data were treated as completely missing. Treatment of partially and completely missing data is described further in Appendix A.

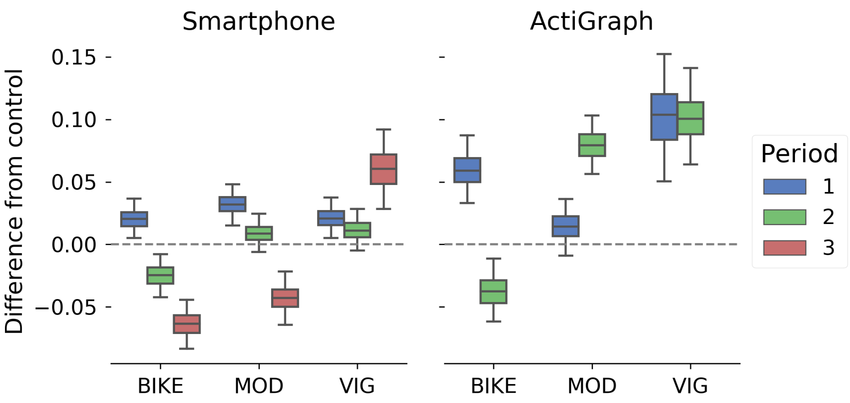


Figure 3.1: Treatment effect of interventions for each intervention in each period contrasted to control. Period 1 is from baseline to 3 months, 2 is from 3 to 6 months and 3 is after 6 months. Whiskers indicate 95% posterior intervals. We see a positive effect in period 1 across the board and then a decline in the second, for the smartphone data. The Actigraph data show a more varied result with e.g. the opposite development in the MOD group.

3.1.3 Model definition

With physical activity represented by activity counts we can use a variant of the model presented in equation (1.3) to estimate the causal intervention effects

$$y_i = \mu + \boldsymbol{\mu}^{pre}_{s[i]} + \boldsymbol{\tau}_{t[i]} + \boldsymbol{\alpha}_{g[i]} + \boldsymbol{\beta} X_i + \epsilon_i, \tag{3.1}$$

where y_i are activity counts, μ is the global intercept, $\boldsymbol{\mu}^{pre}_{s[i]}$ are the subject–specific pre–treatment means, $\boldsymbol{\tau}_{t[i]}$ are the treatment effects, $\boldsymbol{\alpha}_{g[i]}$ are random covariates, $\boldsymbol{\beta}$ are regression coefficients for the linear regressors X, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The $s[i]$ notation denotes which participant sample i belongs to, $t[i]$ denotes treatment, and $g[i]$ denotes group of a particular covariate. In this model we are primarily interested in $\boldsymbol{\tau}$ and the other parts are included to remove confounding effects. The model is used to estimate intervention effects from both smartphone and Actigraph data with the only difference that Actigraph data is only from the three free–living periods as described in Section 2.1.2.

3.1.4 Intervention effects

Estimated intervention effects of each group during each period relative to the control group are depicted in Figure 3.1. In the effects estimated from the smartphone data it is observed that the first period exhibit an overall increase in activity, while the following periods show unchanged or decreased activity levels. Only the vigorous group show an increase throughout the experiment and especially after the experiment is over, in period three. Estimated activity levels based on the Actigraph data are similar except for the moderate group for which the second period is the most active. Differences in activity levels are presumably due to the difference in when the data is from. Smartphone data are acquired continuously and period one therefore includes data from baseline up until period two whereas the Actigraph is only carried for a single week in each period. Missing data and how people use their smartphone may also factor in but are difficult aspects to reason about. Further analysis of the covariates are provided in Appendix C.

Loss of fat–mass

Change in fat–mass from baseline for each group is depicted in Figure 3.2 where the fat–mass is represented as a percentage of total bodyweight. All intervention groups have lost fat–mass during the first period (baseline to 3 months) and maintained their loss in the second period. Apart from the bicycle group in which activity levels decreased in the second period, the loss of fat–mass corresponds well with the measured increase in physical activity. Two sources of bias in the activity estimation may lead to lower effects than what would be expected from the observed loss of fat–mass: 1) the acceleration measured while biking is typically very low and not representative of the physical exertion and 2) the participants tend to leave their smartphone at home while exercising. The latter point holds particularly for the vigorous group.

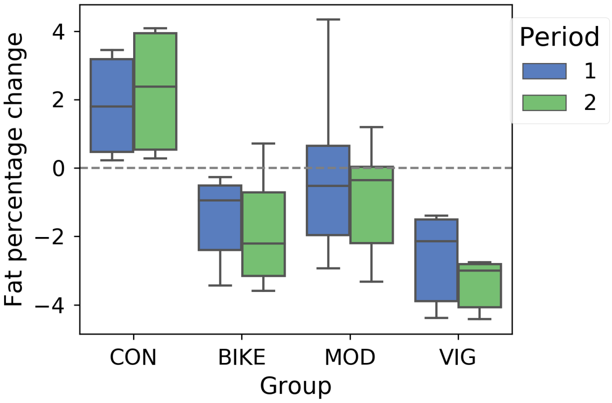


Figure 3.2: Change in fat–mass contrast to baseline. An increase in bodyweight is observed in the control group and a decrease in the rest of the groups though most significantly in VIG. The change from the first to the second test period is not significant in any group.

3.2 Energy expenditure prediction

Energy expenditure (EE) remains difficult to estimate without the use of laborious methods such as doubly labelled water and indirect calorimetry. Reasons include the complexity of human physiology that leads to a natural variation across people. EE is composed of three components; digestion, physical activity, and resting EE, which means that even if physical activity is correctly measured and converted to EE, a large part of the total EE still needs to be accounted for. Predicting just the 'active' part of EE from physical activity measurements is difficult as people move differently, do different kinds of activities, and have varied body compositions. Body weight is for example proportional to energy spent during weight bearing activities (Schoeller and Jefford, 2002) and therefore it would be natural to adjust for body weight in a regression model of EE, however, resting EE is only dependent on fat–free mass (Müller et al., 2018b), so not knowing the ratio of fat to fat–free mass will leave a large part of the variance unexplained. Similarly will not knowing the type of activity lead to unexplained variance as the relationship between EE and physical activity depends on the activity type and intensity (Altini et al., 2015).

Many of these components of variance are specific to the individual and therefore avoidable when doing individual–level model calibration. This is a complicated procedure, however, as it entails exact measurement of EE by calorimetry. In this study the primary goal is to establish how well the smartphone can predict EE, so while we do calibrate to the individual, it is acknowledged that the procedure is not directly applicable to the general population, yet.

3.2.1 Individual–level calibration

We are interested in estimating a function that predicts energy expenditure from physical activity measured by the smartphone, but alas, DLW samples are sparse as they are summarised over a long interval (14 days) and expensive to acquire. This means that a model that is able to adjust to each participant will be over–parameterised and severely overfit to the training data if estimated directly from the DLW measurements. The dataset at hand, however, contains an additional modality that may be used as a mediator between physical activity and energy expenditure: heart rate (HR). A mediator enables the use of additional data by decomposing the model into two parts as depicted in Figure 3.3.

Heart rate prediction

In each exercise session the participants are required to wear a Polar HR monitor and are asked to also carry their smartphone. These data are independent of the EE measurements and may therefore be used to estimate a function that predicts HR from physical activity

$$\mathbf{z}_m^i \sim \mathcal{N}(\beta_1^i + \beta_2^i \mathbf{x}_m^i, \sigma_z^2), \qquad \sigma_z^2 \sim \mathcal{N}(0, 1), \tag{3.2}$$

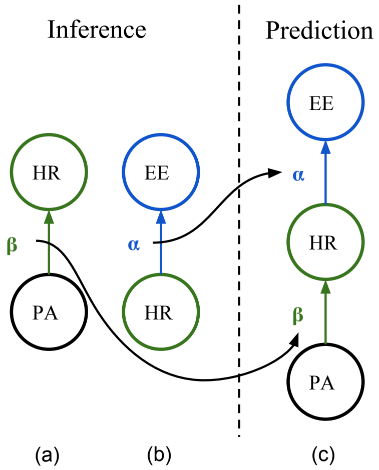


Figure 3.3: Individually calibrated model of energy expenditure prediction. Inference and prediction steps are separated as the data used for each of the models are different.

where $\mathbf{z}$ denotes HR, $\mathbf{x}$ denotes physical activity, and $\boldsymbol{\beta}$ are the regression parameters. Priors on the parameters are given by

$$\beta_1^i \sim \mathcal{N}(\beta_1, 0.1), \qquad \beta_1 \sim \mathcal{N}(0,1) \tag{3.3}$$

$$\beta_2^i \sim \mathcal{N}(\beta_2, 0.1), \qquad \beta_2 \sim \mathcal{N}(1,1), \tag{3.4}$$

which motivates regularisation towards similar coefficients across participants. Coefficients are partially pooled through the prior which has the practical implication that statistical power is borrowed from parameters with low variance by parameters with high variance.

Energy expenditure prediction

Three VO$_2$peak tests are conducted during the study for each participant and in these a linear relationship between HR and EE is estimated. That means we can use these estimated parameters in a model estimating EE

$$\mathbf{y}^i \sim \mathcal{N}\left(\sum_{m=1}^{M}(\alpha_1^i + \alpha_2^i \mathbf{z}_m^i), \sigma_y^2\right), \qquad \sigma_y^2 \sim \mathcal{N}(0,1), \tag{3.5}$$

where $\mathbf{y}^i$ denotes EE for participant i and $\boldsymbol{\alpha}$ are the parameters estimated from the VO$_2$peak test. The sum is from having multiple HR for each EE measurement.

The combined model is able estimate EE from physical activity and have parameters that are calibrated to each participant without overfitting to the data we are trying to predict.

Renormalising parameters

All data are normalised to improve exploration of the sampling algorithm but as the data used for inference of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is not the same, the normalisation is not the same either and the output of (3.2) is not directly applicable in (3.5). $\boldsymbol{\alpha}$ is therefore scaled by

$$\hat{\alpha}_1 = \frac{\alpha_1 + \alpha_2 \text{min}(\mathbf{z}) - \text{min}(\mathbf{y})}{\text{max}(\mathbf{y}) - \text{min}(\mathbf{y})}, \qquad \hat{\alpha}_2 = \alpha_2 \frac{\text{max}(\mathbf{z}) - \text{min}(\mathbf{z})}{\text{max}(\mathbf{y}) - \text{min}(\mathbf{y})}. \tag{3.6}$$

3.2.2 Baseline model

A baseline model is used to evaluate how well a simpler approach without heart rate as a mediator, is able to generalise. The model is a simple regression of physical activity on EE with correction for confounding covariates: gender, age, weight, and fat mass. The model is defined as

$$\mathbf{y}^i \sim \mathcal{N}\left(\beta_1^i + \boldsymbol{\beta}_2^i \mathbf{X}^i, \sigma_y^2\right), \qquad \sigma_y^2 \sim \mathcal{N}(0,1), \tag{3.7}$$

where $\mathbf{y}$ represents EE, $\mathbf{X}$ contains both physical activity and covariates, and $\boldsymbol{\beta}$ are the parameters. The prior on $\boldsymbol{\beta}_2$ is a multivariate normal distribution that is able to model correlation between the covariates. Evaluation in done with leave–one–participant–out cross–validation using mean absolute error (MAE). This is important to test whether the model is able to generalise to unseen people. To test the calibration effect of each covariate, a total of five models are evaluated – from zero to all covariates.

3.2.3 Results

Baseline

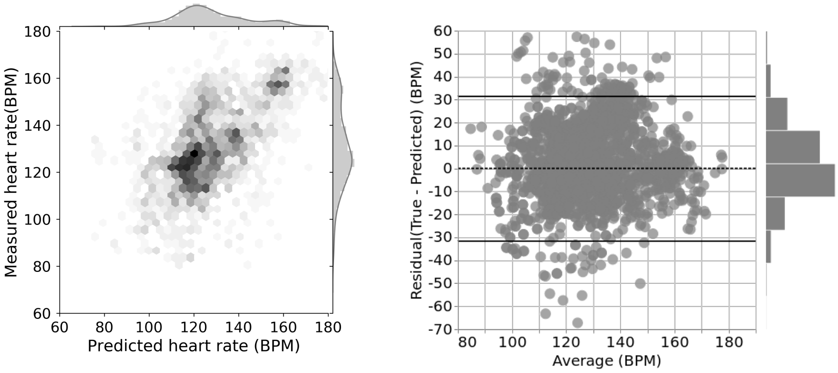
Without proper calibration, the prediction of EE is very poor, as observed in Table 3.2, though comparable to multi–sensor consumer monitors (Chowdhury et al., 2017). The model improves by adjusting for known covariates and is best when all data is included. The best model result correspond to an error of 28.8% relative to the average total energy expenditure.

Covariate	Activity	+Gender	+Age	+Weight	+Fat mass
LOO–CV MAE	1051	878	778	903	729

Table 3.2: Leave–one–subject–out cross–validation (LOO–CV) mean absolute error (MAE) for five baseline models where each column includes one additional covariate. Results are in kcal/day.

Individual calibration

Prediction of the mediator, HR, is unbiased, positive, and strongly correlated to the measured (r=0.86, $P < 0.001$) as illustrated in Figure 3.4. The predicted EE is also positive and strongly correlated to the measured (r=0.84, P $<$ 0.001) with a MAE of 242 kcal/day, though slightly negatively biased (-52 kcal/day, $\approx$ 2.1% of average response) and with a standard deviation of 297 kcal/day of the residuals. The reported MAE correspond to an error of 9.6% relative to the average total energy expenditure. The chart in Figure 3.5 show the predicted against the measured EE where the colors correspond to intervention group and size of the dots correspond to prediction error.



(a) Prediction

(b) Bland-Altman

Figure 3.4: Heart rate regression. In (a) it is shown that heart rate is predicted accurately from smartphone activity counts and in (b) that the residuals are close to normal without bias.

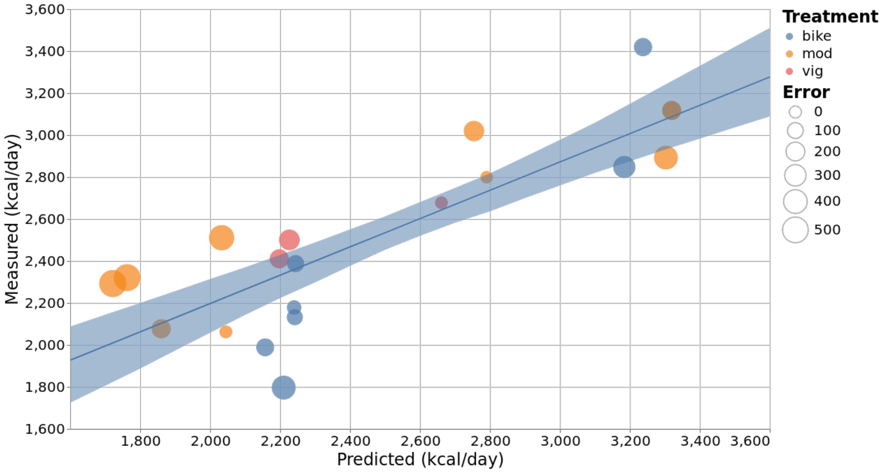


Figure 3.5: Predicted and measured energy expenditure. Each sample corresponds to data over a full DLW period, i.e. multiple days and the size corresponds to the error. The shaded area is the 95% bootstrapped confidence interval.

CHAPTER 4
Domains & Activities

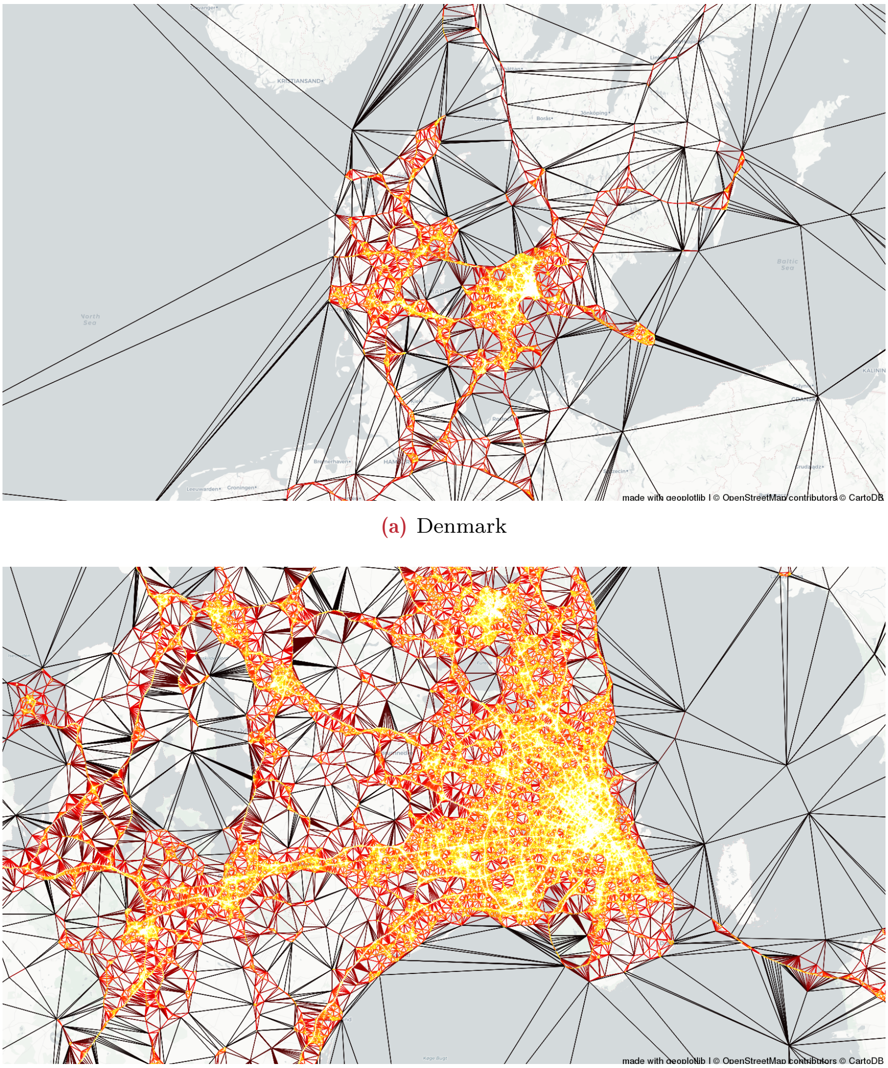
In Section 3.1 it was shown that the smartphone is able to measure physical activity on a scale that is consistent with the loss of fat–free mass and in Section 3.2 it was shown that the smartphone is able to predict total energy expenditure with individual–level calibration. As the smartphone provides additional modalities, particularly location and the derived activity recognition, we can decompose the intervention effect in these contextual directions. Namely, what contextual meaning does a certain location carry and which activities drive an individuals' physical activity. The following sections describe how to infer meaningful locations from GPS data, how the physical activity is distributed among these, and how to predict human activity with a semi–supervised neural network.

4.1 Domains

Domain is an important determinant of understanding behavioural aspects of physical activity. When trying to understand physical activity over a long duration in everyday life settings, important questions include: when and where is the most energy spent and what effect does an intervention have on the distribution of expenditure between domains, e.g. does an increase in one lead to a decrease, a compensation, in another?

4.1.1 Stop locations and paths

Locations measured by a Global Positioning System (GPS) sensor are provided in terms of longitude and latitude and have on their own little meaning. To give locations meaning it is useful to infer spatial clusters which indicate that an individual spend time there. A density–based spatial clustering algorithm (DBSCAN, Ester et al. (1996)) is used to find areas of high sample density in the two dimensional spatial domain. In Figure 4.1 a delaunay tessellations of the location samples illustrate the density of samples in this study. A bright color indicates high sample density and from Figure 4.1(a) it is clear that the participants spent most time in the north–eastern parts of Zealand. When concentrating on this area as in Figure 4.1(b), a


KRISTIANSAND
North Sea
Baltic Sea
made with geoplotlib | © OpenStreetMap contributors © CartoDB
made with geoplotlib | © OpenStreetMap contributors © CartoDB

(a) Denmark

(b) Greater Copenhagen

Figure 4.1: Delaunay tessellations of all location samples in the study to provide an overview of the sample density and range.

more detailed tessellation show the outline of large roads and that Copenhagen is particularly well–visited. GPS samples are in general accurate up to 8 meters outdoors but the uncertainty can go up to 100 meters indoor (Zandbergen and Barbeau, 2011). The scale parameter of the kernel in DBSCAN is therefore set to 150 meters to allow for some leeway. Each location sample also contains temporal information which can be used to refine the accuracy of detected clusters. As we are only interested in clusters in which an individual spends time, a cluster is only defined as such if multiple consecutive location samples fall within the boundaries during a 15 minute window (Cuttone et al., 2014). When labelling the samples in the spatio–temporal domain it is useful to characterise those that fall within a cluster as stops and those that do not, as paths.

4.1.2 Domain inference

Circadian and societal rhythms emerges as clear patterns when visualising the amount of time spent in the individual clusters, decomposed into the hour of the day or the day of the week as in Figure 4.2. With clear enough patterns we can establish *home* and *work*, if the latter is stationary, and the remaining clusters as *leisure*. Time between clusters are naturally defined as *travelling* and times for which no location sample is available, are *unknown*. This procedure is manual but could easily be automated with a bit more background knowledge of the participants.

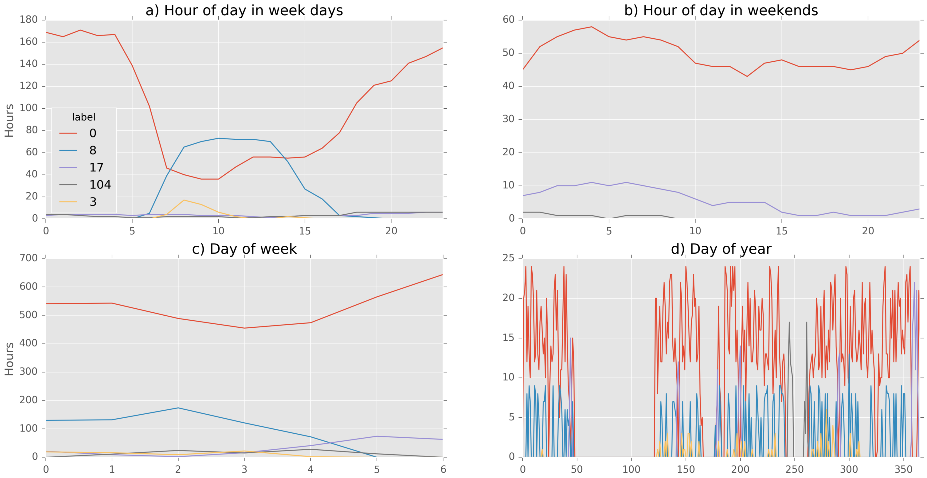


Figure 4.2: Time spent in the top five clusters decomposed into time–units. The y–axes are in hours and the x–axis is mentioned in the subplot titles. This figure illustrates the different ways in which identified clusters can be analysed just by summing over different temporal slices.

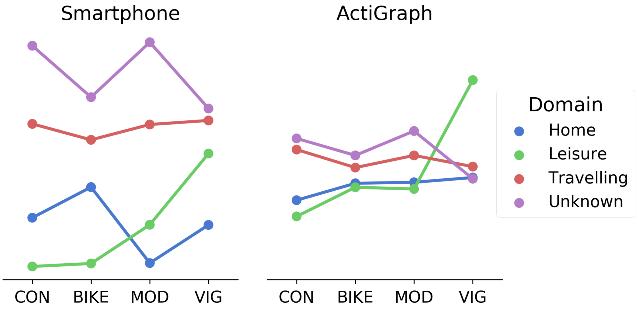


Figure 4.3: Effect of domains. These effects are not contrasted to CON and are therefore not interpretable as causal but can still be interpreted relatively. Most notably are MOD and VIG more active in the *leisure* domain compared to CON and BIKE. In the Actigraph data, the variation is not as great, probably due to a lower sample size.

4.1.3 Effect of domains

As we also show in the paper in Appendix A, is it straightforward to condition the intervention effect τ in equation (3.1) on domains to inspect what effect the domains have on physical activity. The marginal posterior of the intervention effect conditioned on domain is depicted in Figure 4.3. Note that the estimates are not contrasted to the control group (CON) and therefore not interpretable as causal effects but may still be compared internally. The *work* domain is not included as it was impossible to establish the correct cluster for too many of the participants and has therefore been merged into *unknown*. The *unknown* clusters are unfortunately dominating the high–activity samples and it is not clear why, except that few participants are known to have active jobs (e.g. postman). As explained in Section 2.3.1, singular modalities are sometimes missing, and in case of missing location samples, the measured activity ends up in an unknown cluster.

An interesting observation in Figure 4.3 is that MOD and VIG generally have higher activity levels in the *leisure* clusters than at *home* and that this relation is opposite for CON and BIKE. A possible explanation is a compensatory effect of the intervention that moves expenditure from one domain to another.

Effect of time in domains

When integrating out the intervention assignment, we get the effect of domains relative to time, which is depicted in Figure 4.4. Remember that the third period is after the 6 months completion of the study and up to 300 days after the start. While an apparent increase is observed from the first to the second period, this is not observed in the intervention effect in Figure 3.1 which is because this is an average over participants from all interventions as opposed to a contrast to the control group.

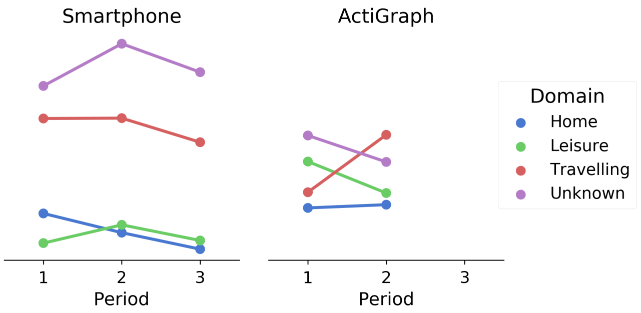


Figure 4.4: Effect of domains over time. Each data point is summed over intervention groups and therefore not causal. A decrease in *home* is observed as well as a general decrease in period 3 when the study has finished.

4.2 Activity Recognition

Human activity recognition is a research field that spans from binary classification (active/sedentary) to detecting daily activities, such as drinking and ironing (Chen et al., 2012; Incel et al., 2013; Reyes Ortiz, 2015; Shoaib et al., 2015). The modality of choice similarly spans from accelerometers to 3–dimensional depth cameras and any combination of what may lie between (Aggarwal and Xia, 2014; Twomey et al., 2016). Applications of activity recognition are many as activity convey information about the context of an individual. The information may for example be used to intelligently assist elderly living at home, understand human behaviour, or more generally put; assist in human-computer interaction.

In Section 4.2.2 a novel method to classify activity from raw accelerometer data will be described. In the initial phase of writing the manuscript describing the model, a nearly identical method was published in Ordóñez and Roggen (2016) after which publishing our work seemed moot. We have later extended the model with domain adaptation and it is that part which is the focus of this section. For a thorough explanation and rigorous experimentation regarding classification of activities with the presented type of model, please consult Ordóñez and Roggen (2016) and Hammerla et al. (2016).

4.2.1 Related methods

Bao and Intille (2004) demonstrated the ability of machine learning methods to classify everyday human activities from an array of worn accelerometers and it was then shown that a single accelerometer is sufficient to recognise basic activities (e.g. standing, walking, running) (Ravi et al., 2005). Work that paved the way for detecting activities with mobile phones (Brezmes et al., 2009; Kwapisz et al., 2011; Wu et al., 2012). Machine learning approaches applied in activity recognition typically relies on features extracted from the signal because they are unable to deal with either the high number of dimensions or noise level in the raw data. Features are hand-crafted and often rely on expert domain knowledge but in signals that does not offer any guidelines to what a good feature might look like, summary statistics are typically used. However, there are no guarantees that relevant information is captured in the features so learning a feature extractor from the data with a deep neural network can improve information retention (Plötz et al., 2011). A deep neural network can in fact discard explicit features entirely by classifying from raw data directly (Zeng et al., 2014) and in real–time when deployed on a smartphone (Lane and Georgiev, 2015). Previous methods applying neural networks primarily focus on classifying a single time–frame of 1–10 seconds of data. Human activity is stationary on a much longer time–scale so the activity in a time–frame is therefore statistically dependent on the neighbouring. On the other hand can a transition from one activity to another span two frames and only by compounding information is that captured. The following section will describe a neural network model that learns to classify activities from raw data with a temporal memory to condition on past information.

4.2.2 Human activity recognition with deep neural networks

Background theory

Deep neural networks refer to the practice of stacking layers of neural networks to expand flexibility and encode structural information. The purpose is to learn the function $y = f(\mathbf{x}; \boldsymbol{\theta})$ that maps input $\mathbf{x}$ to output y by adjusting the parameters $\boldsymbol{\theta}$. To learn the parameters, gradients of $\boldsymbol{\theta}$ with respect to the error signal is *backpropagated* through the network (Goodfellow et al., 2016). Three types of architectures are most often used: dense, convolution, and recurrent.

Dense neural networks are also commonly known as fully–connected (or feed–forward) as the outputs from all nodes in a layer are connected to all nodes in the next layer, as inputs.

Convolution architectures are inspired by the filtering operation in signal processing where a filter is convolved with the signal. Parameters of the filters are learnt to extract meaningful representations and by stacking multiple convolutional layers, a hierarchy of increasingly complex representations are constructed (Krizhevsky et al., 2012). In the domain of human activity, the lower layers may hypothetically learn to recognise upward or downward motion while higher layers may learn to extract perpetual motion of a certain frequency.

Recurrent neural networks (RNN) are a family of architectures designed to model sequences. Instead of just mapping an input to an output, the hidden state of the network is dependent on itself from an earlier time–step, i.e.

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \boldsymbol{\theta}), \tag{4.1}$$

which provide the network with an internal memory. Recurrence occurs between hidden units because the parameters $\boldsymbol{\theta}$ are shared between time–steps. The basic RNN structure can become forgetful after few steps due to *vanishing gradients*. A variant of the RNN cell called *long short-term memory* (LSTM) reduces this problem by embedding a mechanism to control when the internal state is updated, thereby extending its memory span (Hochreiter and Schmidhuber, 1997). The output y is typically modelled with a linear function or dense neural network, i.e. $y = f(\mathbf{h}^{(t)}; \boldsymbol{\phi})$.

Neural networks can be combined in any imaginable way as long as they remain differentiable. It is therefore possible to construct models with multiple inputs, models that contains parallel sub–networks, and even with multiple objective functions, some of which the following model will demonstrate.

Domain adaptation

Deep neural networks require large amounts of data due to the large number of parameters in the models or they will overfit to the training data. In activity recognition it is particularly difficult because of small annotated datasets (Plötz and Guan, 2018) and when generalising to data from previously unseen people (Bao and Intille, 2004). The problem is caused by a shift between the training (*source*) and test (*target*) distributions (Shimodaira, 2000) and a solution is therefore to transform the domains into a common subspace where they have very similar distributions (Ben-David et al.,

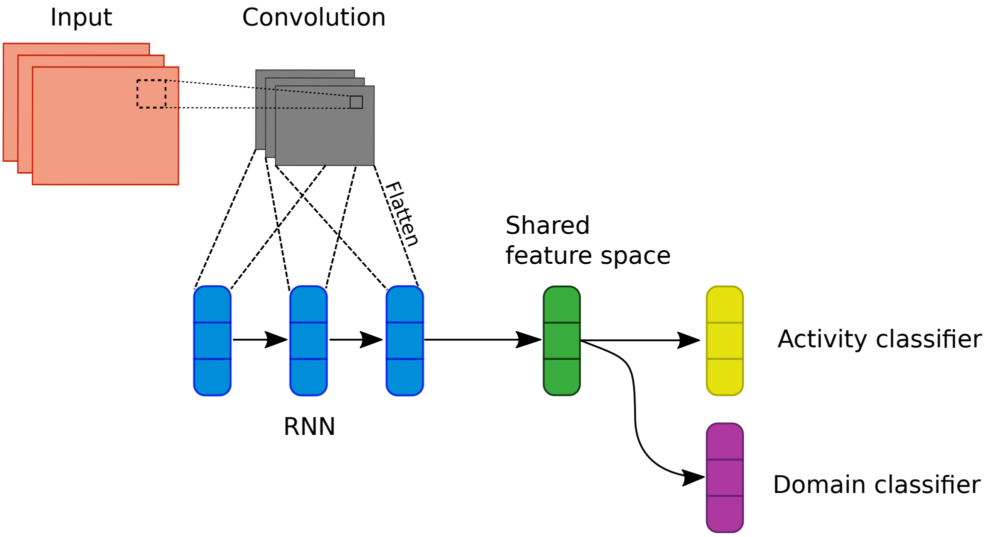


Figure 4.5: Neural network architecture. The input time–series is segmented and provided to the CNN which output is flatten to a single vector for input in the Bi–LSTM RNN. A shared representation is learned in a two–layer dense NN and provided as input in the activity and domain classifiers.

2010). Previous approaches to domain adaptation in activity recognition rely on transferring knowledge stored in a model pre–trained on the source dataset and then potentially fine–tuning the model, if any labels are available, in the target domain (Morales and Roggen, 2016). The rationale being that lower layers of a deep CNN learns to extract representations that are domain–invariant (Donahue et al., 2014). The method works reasonably well but also introduces additional tuning issues, e.g. which layers to re–train in the fine–tuning. An *adversarial training* method can skip this two–stage procedure entirely and simultaneously learn representations that are domain–invariant and predictive of class membership. To be domain–invariant, predictions based on features from the common subspace must not be able to discriminate between source and target domains. In neural network models this is readily achieved by extending the model with an additional *domain classifier* (Ganin et al., 2016). While the model is optimised to reduce the error of the label classifier, the error of the domain classifier is maximized. This encourages the model to learn a shared representation that is domain–invariant but still able to classify activities. The method can be used semi– or unsupervised and is used in this project to strengthen the label classifier with external datasets.

Model definition

Activities are composites of motions that manifests as sequences of increasing and decreasing acceleration when measured with a tri–axial accelerometer. Not very

Dataset	Samples	Activities
iDASH (Wu et al., 2012)	3174	walking, jogging, stairs, step
UCI HAR (Anguita et al., 2013)	5272	walking, upstairs, downstairs, sitting, standing, laying
UCI mHealth (Baños et al., 2012)	4859	walking, jogging, stairs, cycling, running, jump
WIDSM (Kwapisz et al., 2011)	10982	walking, jogging, sitting, standing, upstairs, downstairs

Table 4.1: External datasets with annotated activities used to test the model.

different from images which are composed of two–dimensional gradients of pixel intensity. Similar to how Gabor filters are able to extract local patterns such as edges in images, is it imaginable that a finite set of filters are able to identify acceleration patterns generated by human motion. Convolutional neural networks are used to learn these representations directly from the raw data. As motions are composed of several primitives, increasingly complex representations are learned by stacking multiple layers. The output from the first convolutional layer is therefore used as input to another and in this model two layers are used. A 'maximum pooling' operation is done to the output of each layer to reduce temporal dimensionality. It is done by only keeping the maximal of every two neighbouring values, thereby reducing dimensionality to half.

A sequence of N samples may contain several motions that make up some activity. Partitioning the sequence into K segments and applying the deep convolutional neural network to each segment, the model learns to extract a representation of the set of motions (or lack thereof) within a segment. This is illustrated in Figure 4.5 as going from Input to Convolution. The sequence of K representations, or learned features, are then used as input to a LSTM recurrent neural network that is able to model the temporal dynamics of activities. A second RNN that processes the sequence in the opposite direction is added to increase the model flexibility and account for any variation that may arise due to temporal peculiarities in the data. The RNN thus becomes bi-directional. The vector of hidden units from the last step in the RNN is used as input to a two–layer dense neural network with two heads: a label classifier and a domain classifier. Each classifier is a dense NN composed of two and three layers, respectively.

Data

External datasets are summarised in Table 4.1. The objective is to leverage the combination of multiple datasets to predict four basic activities on the GoActiwe dataset: *still*, *walking*, *running*, *biking*. As the external data are acquired in different ways and with different activities, the labels are bucketed and relabelled to the aforementioned activities as best as possible. Sampling frequency is re–sampled to 50 Hz for all datasets and samples are binned in windows of 5 seconds to represent a single activity sample.

Semi-supervised experiment

Efficacy of domain adaptation is measured by classification accuracy on a dataset with partially observed labels. The scale determines the proportion of labels observed and ranges from unsupervised (0.0) to supervised (1.0) in steps of 0.1. Each mask is sampled before training and therefore constant during training. A single dataset (UCI HAR) was chosen to be the fully observed source and the three others to be target domains. The domain adaptation task is binary, i.e. single source and target, but it is possible to extend this method to multiple domains (Zhao et al., 2017).

Inference

The distribution of activities in the data is not uniform and that may skew the classifier to become better at classifying some activities which will have a detrimental effect on the underrepresented classes. To counter this effect, the data are sampled with a uniform distribution over classes during training.

To train the model a variant of stochastic gradient descent with momentum called Adam is used (Kingma and Ba, 2015). All hyperparameters are fixed for all models.

Results

Classification accuracies on the target domains are listed in Table 4.2 (and illustrated in Figure 4.6) and on the source domain in Table 4.3 (Figure 4.7). The first rows (0.0) are the results for the unsupervised case and as expected a quite poor performance is observed in the target domains whereas a near perfect performance in the source domain. The target domain classification relies in this case on how similar the signals are to the source domain, fx in what position is the monitor carried, how are the activities performed, and so on. The UCI mHealth dataset is apparently very dissimilar, resulting in worse than random performance. Performance increase significantly across all datasets with just 10% available labels, however, the results from classification without domain adaptation listed in Table 4.4 are equally good. The experiment at hand is apparently too easy for the model to gain anything significant from domain adaptation and it would therefore make more sense to test the model on a more demanding problem, fx the experiment detailed in Morales and Roggen (2016).

Dataset Observed	WISDM	iDASH	mHealth
0.0	0.280	0.484	0.060
0.1	0.870	0.890	0.711
0.2	0.961	0.895	0.738
0.3	0.964	0.907	0.769
0.4	0.972	0.913	0.772
0.5	0.969	0.920	0.758
0.6	0.971	0.916	0.759
0.7	0.970	0.920	0.761
0.8	0.969	0.920	0.757
0.9	0.963	0.917	0.735
1.0	0.973	0.919	0.767

Table 4.2: Accuracy of domain adaptation on partially observed target datasets.

Dataset Observed	WISDM	iDASH	mHealth
0.0	0.998	0.995	0.995
0.1	0.972	0.990	0.953
0.2	0.975	0.988	0.948
0.3	0.978	0.985	0.892
0.4	0.927	0.987	0.980
0.5	0.906	0.986	0.956
0.6	0.932	0.972	0.960
0.7	0.871	0.980	0.952
0.8	0.889	0.978	0.967
0.9	0.949	0.974	0.947
1.0	0.912	0.981	0.956

Table 4.3: Accuracy of domain adaptation on fully observed source dataset.

Dataset Observed	WISDM	iDASH	mHealth
0.0	0.0	0.0	0.0
0.1	0.973	0.867	0.757

Table 4.4: Accuracy on partially observed datasets with no domain adaptation.

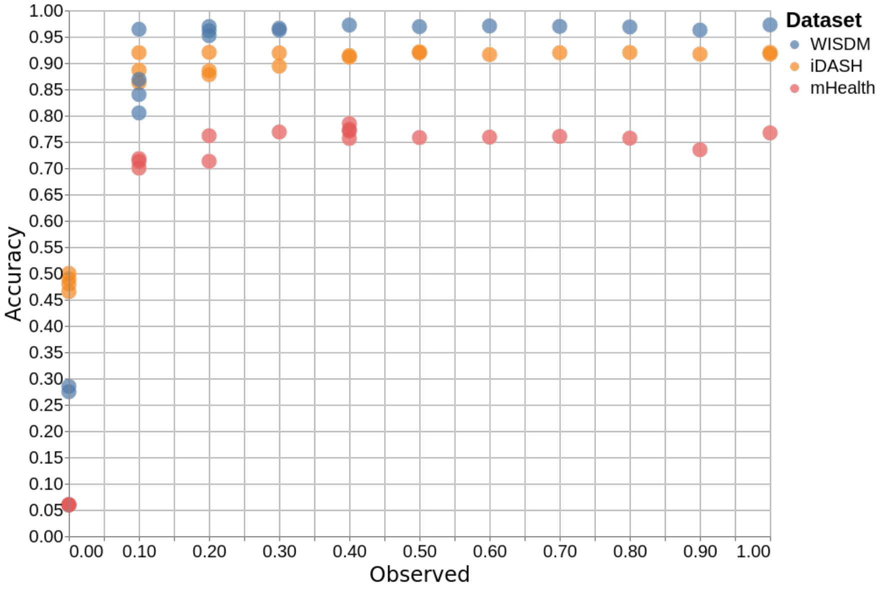


Figure 4.6: Accuracy on target datasets.

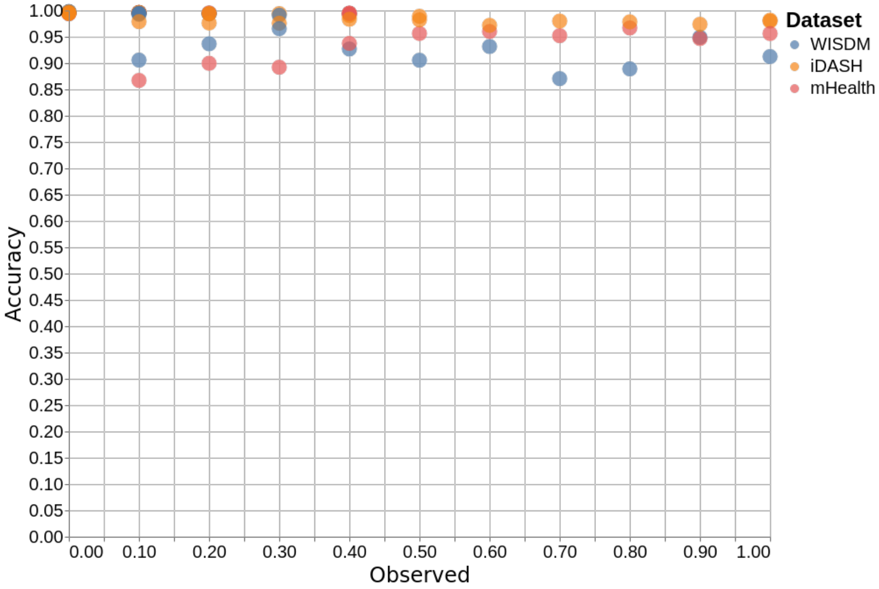


Figure 4.7: Accuracy on source datasets.

CHAPTER 5 Deep latent state space model

If enough information was available, would it be possible to infer the state of a human being perfectly (Franklin, 2017)? Simplicity and predictability in human nature is for example found in the cognitive biases of our minds (Kahneman and Egan, 2011) and recently in our mobility patterns. Individuals have few places they frequently visit (González et al., 2008) and though the moving patterns of an individual evolve, the number of familiar locations is stationary around 25 (Alessandretti et al., 2018), which may be one reason to why human mobility is up to 93% predictable (Song et al., 2010). If our whereabouts are so predictable, other aspects of our nature are likely to be as well. Like how and why we exercise. The aforementioned studies in mobility required data from thousands of people over long periods before the patterns emerged, and that may well hold true for exercise as well.

There is no doubt that a persons level of physical activity (or health) is not an entirely self–contained process, but is constantly affected by external factors, such as weather conditions, and internal factors, such as mood. Modelling that process thus requires not just large amounts of longitudinal data, but varied as well. Making sense of high–dimensional streams of data acquired from many modalities, is not straightforward and typically requires one of two approaches: 1) engineer features that capture most of the relevant information and then model the underlying process with simple statistics, or 2) train a deep neural network that is able to ingest the data raw and model the process in an end–to–end fashion. Neural networks are powerful in learning useful representation from data but can be difficult to train, require large amounts of data, and are hard to interpret. Statistical methods are often quite the opposite but lack the flexibility of learning structure from data and therefore require domain expertise. The following sections will present the model introduced in the paper in Appendix C that combines flexible representation learning with a linear probabilistic model. Key concepts to understanding the model will be summarised but focus is on understanding the model on a conceptual level and how it relates to physical activity.

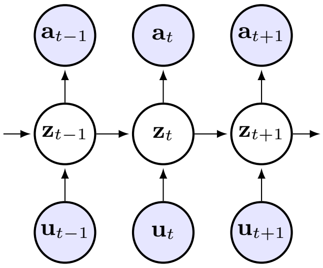


Figure 5.1: Graphical model of the linear Gaussian state space model. Filled circle indicate observed quantities, empty circle indicate latent variables and the arrows indicate dependence.

5.1 Background theory

5.1.1 State space models

It is not unlikely that a persons physical health can be represented by a latent state $\mathbf{z}_t$ to some point in time, t, and that this state lives in a relatively low–dimensional space. As humans evolve rather slowly, in the absence of death, it is also not unlikely that the current state is dependent on the previous, as in a first–order Markov process, but with potential external influence from $\mathbf{u}_t$. Linear Gaussian state space models (LGSSM) are widely used to model this type of process in the case of normally distributed variables

$$p_{\gamma_t}(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{u}_t) = \mathcal{N}(\mathbf{A}_t\mathbf{z}_{t-1} + \mathbf{B}_t\mathbf{u}_t, \mathbf{Q}). \tag{5.1}$$

The state $\mathbf{z}_t$ is then used as a prior to model the observations $\mathbf{a}_t$,

$$p_{\gamma_t}(\mathbf{a}_t|\mathbf{z}_t) = \mathcal{N}(\mathbf{C}_t\mathbf{z}_t, \mathbf{R}), \tag{5.2}$$

where $\gamma_t = [\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t]$ are the transition, control, and emission matrices. $\mathbf{Q}$ and $\mathbf{R}$ are covariance matrices of the process and measurement noise. All distributions are assumed Gaussian and all dependencies linear, which ensures that posterior inference is analytically tractable. This model is often referred to as the Kalman filter (Kalman and Others, 1960) and is depicted graphically in Figure 5.1. A thorough explanation is available in Durbin and Koopman (2012) and Fraccaro (2018).

5.1.2 Variational inference

Bayesian probabilistic models are typically not tractable and therefore require an approximative inference procedure to estimate unknown parameters. Variational inference (VI) is a method that is able to approximate high–dimensional posteriors by factorisation and recasting the problem as an optimisation (Bishop, 2006; Zhang et al., 2017). Lets assume that we have the joint distribution $p(\mathbf{a}, \mathbf{x})$ and we want to find

$$p(\mathbf{a}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{a})p(\mathbf{a})}{p(\mathbf{x})} \tag{5.3}$$

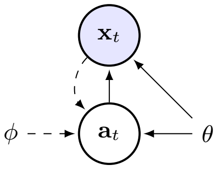


Figure 5.2: Graphical model of the variational auto–encoder. Solid lines denote the generative model $p_\theta(\mathbf{x}|\mathbf{a})p_\theta(\mathbf{a})$ and dashed lines denote the variational approximation $q_\phi(\mathbf{a}|\mathbf{x})$.

where $\mathbf{a}$ are latent variables that encode the hidden structure in the data, $\mathbf{x}$, and we are interested in learning the posterior $p(\mathbf{a}|\mathbf{x})$, i.e. the conditional distribution of the hidden structure given the data. The normalisation term, $p(\mathbf{x}) = \int_a p(\mathbf{x}, \mathbf{a})$, is not tractable to compute, so instead VI approximates the posterior with a simpler distribution, $q_\phi(\mathbf{a})$, which is from some family of parametric distributions, parameterised by the variational parameters ϕ. These parameters are optimised by reducing the Kullback–Leibler (KL) divergence between $p(\mathbf{a}|\mathbf{x})$ and $q_\phi(\mathbf{a})$ which can be computed in closed form when using Gaussian distributions. The divergence can be formulated as

$$\mathbb{KL}(q_\phi(\mathbf{a})||p(\mathbf{a}|\mathbf{x}))) = -\mathbb{E}_{q_\phi(\mathbf{a})}\left[\log \frac{p(\mathbf{a}|\mathbf{x})}{q_\phi(\mathbf{a})}\right] \tag{5.4}$$

$$= -\mathbb{E}_{q_\phi(\mathbf{a})}\left[\log \frac{p(\mathbf{a},\mathbf{x})}{q_\phi(\mathbf{a})} - \log p(\mathbf{x})\right] \tag{5.5}$$

$$= -\underbrace{\mathbb{E}_{q_\phi(\mathbf{a})}\left[\log \frac{p(\mathbf{a},\mathbf{x})}{q_\phi(\mathbf{a})}\right]}_{\text{ELBO}} + \log p(\mathbf{x}). \tag{5.6}$$

Notice that the log marginal likelihood, or log normaliser, does not depend on q and can therefore be removed from the expectation in the second line. In the last line we have two terms: the negative evidence lower bound (ELBO) and the log marginal probability of $\mathbf{x}$. As the KL divergence is non–negative, the ELBO represents a lower bound on $\log p(\mathbf{x})$ (hence the name), which is constant for all $q_\phi(\mathbf{a})$. Therefore will maximising the ELBO always minimize the KL divergence. Because q typically is much simpler than p, the posterior approximation will rarely capture the complexity of the true posterior though. VI is computationally less intensive than exact inference methods such as MCMC sampling and can be scaled to large datasets with stochastic VI (Hoffman et al., 2013; Ranganath et al., 2013). In latent variable models, some stochastic VI methods are still too computationally intensive for very large datasets, however, as the local latent variables needs to be optimized for each data point. Amortized variational inference replaces the local variables with global parameters by learning a function that maps from a data point to the distribution over the latent variables.

5.1.3 Variational auto–encoder

The variational auto–encoder is a deep neural network structure used for amortized inference in deep latent variable models (Kingma and Welling, 2014; Rezende et al., 2014). Assume we have a generative random process in which a sample in drawn from the prior distribution $\mathbf{a} \sim p_\theta(\mathbf{a})$, which typically is a centered isotropic multivariate Gaussian, and $\mathbf{x}$ is generated from the likelihood

$$p_\theta(\mathbf{x}|\mathbf{a}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{a}), \boldsymbol{\sigma}(\mathbf{a})) \tag{5.7}$$

which often is a Gaussian distribution in case of continuous data but other distributions can be used as well. The likelihood is known as the *probabilistic decoder* or *generative network* and is dependent on the latent variable $\mathbf{a}$ through the non–linear functions $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ which are neural networks parameterised by θ. The posterior distribution $p(\mathbf{a}|\mathbf{x})$ is approximated with the amortized variational distribution

$$q_\phi(\mathbf{a}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\sigma}(\mathbf{x})) \tag{5.8}$$

which is dependent on $\mathbf{x}$ through $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ that are parameterised by ϕ. This is known as the *probabilistic encoder* or *inference network*. The inference network thus learns a probabilistic encoding into a potentially low–dimensional space, that contains as much relevant information as possible. It is this encoding–decoding structure, which is also seen in Figure 5.2, that is the inspiration for the name. Parameters of the neural networks are trained simultaneously by maximizing the evidence lower bound (ELBO) with stochastic gradient descent and by plugging the terms into the ELBO from equation (5.6) we see that it decomposes into two terms

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_{q_\phi(\mathbf{a}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{a},\mathbf{x})}{q_\phi(\mathbf{a}|\mathbf{x})}\right] && (5.9)\\ &= \mathbb{E}_{q_\phi(\mathbf{a}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{a}) - \log \frac{q_\phi(\mathbf{a}|\mathbf{x})}{p_\theta(\mathbf{a})}\right] && (5.10)\\ &= \mathbb{E}_{q_\phi(\mathbf{a}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{a})\right] - \mathbb{KL}(q_\phi(\mathbf{a}|\mathbf{x})||p_\theta(\mathbf{a})). && (5.11) \end{aligned}$$

Here the first part is a reconstruction term that motivates the encoder to learn a code with enough information that the decoder can reconstruct the input. The second part is the KL divergence between the posterior and the prior which regularises the posterior towards the prior. Random variables are generally not differentiable so to obtain stochastic gradients of the parameters, a differentiable unbiased estimator of the ELBO can be obtained by using the *reparameterisation trick* (Kingma and Welling, 2014; Rezende et al., 2014).

5.2 Kalman variational auto–encoders

Lets step away from monitoring human activity and consider a simpler system for a moment. Imagine a ball flying through the air. Humans are able to instantly recognise the ball, its direction, and predict the path it is going to take. This process requires

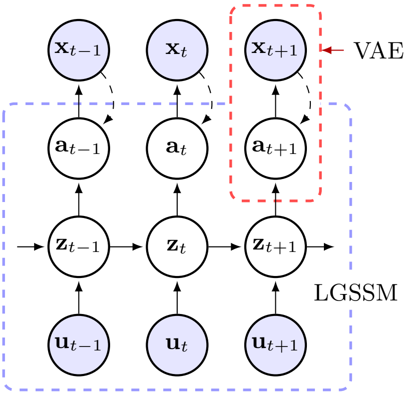


Figure 5.3: Graphical model of Kalman variational auto–encoder. Solid lines denotes the generative network and dashed lines denotes the inference network.

the ability to recognise the ball and an understanding of the system in which the ball operates, i.e. Newtonian dynamics (Ungerleider and Haxby, 1994). Two independent cognitive processes. *Kalman variational auto–encoders* (KVAE) are able to disentangle these processes into two latent representations, i.e. the object recognition and the temporal dynamics. Given a video of a moving object, each high–dimensional frame, $\mathbf{x}_t$, is mapped to a low–dimensional code, $\mathbf{a}_t$, with the probabilistic encoder $q_\phi(\mathbf{a}|\mathbf{x})$. Since we know that we want to model the dynamics of the ball, we can use that prior knowledge to inform the model structure. To model the dynamics, all we need to know about the ball in each frame, is its position. The high–dimensional observation can therefore be compressed into a two–dimensional latent space, without loss of information. As illustrated in Figure 5.3, the code $\mathbf{a}_t$ can now act as a pseudo–observation in the LGSSM and with consecutive encodings of the balls position, the LGSSM is able to model the underlying dynamical system in $\mathbf{z}_t$, that governs the balls trajectory. Simulation of the future and missing data imputation can be done cheaply in the latent space $\mathbf{z}$ by computing

$$p_\theta(\mathbf{z}_{t+1}|\mathbf{x}_{1:t}, \mathbf{u}_{1:t+1}), \tag{5.12}$$

from which an output can be sampled from the generative model

$$p(\mathbf{x}_{t+1}, \mathbf{a}_{t+1}|\mathbf{z}_{t+1}) = p_\theta(\mathbf{x}_{t+1}|\mathbf{a}_{t+1})p_\gamma(\mathbf{a}_{t+1}|\mathbf{z}_{t+1}). \tag{5.13}$$

All parameters in the model can be learned simultaneously with stochastic gradient descent by maximisation of the ELBO (Fraccaro et al., 2017).

5.2.1 Modelling non–linear dynamics

Linear state space models are not able to handle non–linearities like the ball hitting a wall, but we are able to handle these while keeping local linearity by modelling the time–varying dynamics of the LGSSM, γ_t, as a function of the previous encodings, $\mathbf{a}_{0:t-1}$. We introduce the *dynamics parameter network*

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t(\mathbf{a}_{0:t-1}) \tag{5.14}$$

which is implemented as a recurrent neural network with a softmax output so $\boldsymbol{\alpha}_t$ sum to one. Instead of learning a single set of dynamics, we learn K different dynamics $\{\mathbf{A}^{(k)}, \mathbf{B}^{(k)}, \mathbf{C}^{(k)}\}$ which are then modulated by $\boldsymbol{\alpha}_t$ in each time–step, e.g.

$$\mathbf{A}_t = \sum_{k=1}^{K} \alpha_t^{(k)}(\mathbf{a}_{0:t-1})\mathbf{A}^{(k)}. \tag{5.15}$$

The dynamics in each time–step are thus a weighted sum of the K globally learned dynamics. If the model observes that the ball hits a wall to time $t-1$, it is able to select the proper dynamics so the ball bounces off the wall to time t.

5.3 Human dynamics and internal states

Abstracting from the example with the ball, we can imagine many dynamical systems in which the underlying latent state lives on a manifold of much smaller dimensionality than the observations. Take for example the massive amount of data produced by cameras and LIDARs in self–driving cars. A single camera may produce a million–dimensional observation (pixels) many times a second while the state of the vehicle may be represented in less than ten dimensions. Or the multi–modal and unstructured data in electronic health records. Consolidating high–dimensional X–ray scans with low–dimensional, but high–frequency, ECG and unstructured text from the doctors notes, each with irregular sampling intervals, quickly becomes a major project. Each signal require domain expertise feature engineering and annotation for most applications. The Kalman variational auto–encoder is trained unsupervised and therefore only requires large amounts of time–series data. The probabilistic encoder learns to extract low–dimensional representations from complex and high–dimensional observations and the structured time–varying prior distribution, the LGSSM, learns a hidden state that evolves over time, governed by dynamics learned from the data. If trained on healthcare data, this trajectory will describe a persons health in as many aspects as the data spans. If only ECG data from stroke patients is observed, then that is what the model learns to encode. Once the model has observed sufficient amounts of data, however, it can be initialised with a small amount of data from a previously unseen person and then simulate the future. The trajectory of this simulation will tell if there is an increased risk of stroke.

In a similar scenario, imagine that we are able to acquire continuous streams of physical activity data from the smartphones of a large part of the population. We could perhaps learn a 'physical state vector' that is able to describe the internal conditions of a person. If additional information is available, e.g. activity type, social context, weight, or mood, then that can be used to ask counterfactual questions and simulate potential outcomes. A reduction in activity levels may be traced back to its source and we may learn what motivates an individual to be more active. A key assumption is that humans are similar enough that the model can identify a latent space in which data from similar persons with similar behaviour, or whatever drives the generative process, is mapped to the same location in this space.

CHAPTER 6
Conclusions & Perspectives

Central to this dissertation has been the question of how well smartphones can be used to measure physical activity and energy expenditure, pervasively. To this question we demonstrate that not only do they provide a capable platform, the platform is capable of much more.

The question is not answered in full, however, as the complexity of human physiology is too vast for the smartphone to provide the solution on its own. At least with the present state of tools at disposal. It is, for example, not possible to accurately predict energy expenditure with a simple linear model, which is the most prevalent method, due to the wide variation between individuals. The reported error of 28.8% (Section 3.2.3) is a lot when considering it is of the total energy expenditure, i.e. including resting energy expenditure which is the largest component. It should be noted that these results are specific to the relatively small cohort in this study and are therefore not necessarily generalisable, which is the crux of all studies. By going through heart rate as a mediator, a personalised model was obtained, although at the expense of using professional equipment. It is compelling to divide the inference procedure into two parts around the mediator, however, as data for the first part is much easier to acquire compared to the second. That enables continued life–long learning in which the parameters are updated whenever new data is available. The inference is so simple that it is even possible to compute on the smartphone. This type of personalisation is also radically different than normalising for background information such as age and gender, as the model can be more flexible. The only requirement is that the output is a regression on heart rate, so if a persons mobility patterns change or a new type of activity is engaged, a model that continues to learn, can adapt to these changes, pervasively.

In Chapter 3 we show two things. First, with data from the entire study, the estimated intervention effects correspond well with the measured loss of fat–mass and we can use additional modalities to decompose this effect into e.g. domains of everyday life. Second, the exact same data can predict total energy expenditure when used in combination with measurements from a VO_2peak test. The two–stage structure reduces the model complexity from trying to predict energy expenditure from acceleration measurements, which is difficult to generalise between people and activities, to predicting heart rate, a much easier problem. Heart rate also depends heavily on physiological fitness which makes a generic model difficult to create, but wearables provide measurements of sufficient quality to create a personal model. Substituting physical activity with energy expenditure in the first result, we effectively show that smartphones can estimate energy expenditure in everyday life and that we can decompose it in as many directions as we can measure. Domains, activity, social context, nearby people, mood, soundscape, etc. This extrapolation is limited in two aspects, however. We measured total energy expenditure and it is therefore not obvious how to isolate the energy spent on some activity or in a particular time–span, and the measured expenditure was an average over two weeks, further complicating this issue. In cases with zero missing data it should be easier to do this mapping and in most cases the exact energy spent in a particular exercise is not as important as the relative expenditure.

Activity recognition was not used in this project to analyse behaviour around activity levels as the amount of missing data for this modality in particular made inference too uncertain. Another group published the same model for human activity recognition as we proposed (Ordóñez and Roggen, 2016), and we therefore decided not to publish our work, but we did manage to define a model that achieved state–of–the–art and is general for many time–series applications. Our later addition of domain adaptation is important as the available datasets in general are small but the variation between activities and people, big. The presented experiment unfortunately only showed a small gain in the unsupervised case as the problem was too simple for the model without adaptation, even with just 10% observed labels. Due to time–constraints, a new experiment was not possible.

Analysis of *how* and *where* energy is expended was from the onset of this study one of the primary objectives. Especially *how*, including the type of activity, and the temporal change herein due to an intervention, provide insight into behavioural mechanisms that may elucidate *what* to do if a general increase in energy expenditure is desired. But the *where* may in turn be the key to unlock *when* to provide a gentle nudge so that the neural autopilot (system one; Kahneman and Egan (2011)) is temporally disabled and the concious self (system two) can take an active decision to reflect. In a metaphorical sense, a behavioural 'GPS' to provide just–in–time adaptive interventions (Lathia et al., 2014; Nahum-Shani et al., 2015; Pejovic et al., 2015). The cause–effect relationships needed to create this guide may be discovered in a principled way with Structural Equation Models (Pearl, 2009) or implicitly modelled in the parameters of a joint model. Unless interventions are tested systematically, a large amount of data is required to yield results with sufficient certainty.

The Kalman variational auto–encoder is an attempt to define a model that is geared to handle the massive amounts of data that we can gather today, but that are often difficult to annotate and process. Big data often require big models as you either have unlabelled or unstructured data or with 'small' models risk finding spurious correlations and treat them as casual relations. Instead of placing all our bets on deep neural networks, we believe in taking advantage of their strengths, such as learning complex data distributions, and combine them with structured models that provide interpretability, data efficiency, and a principled way to incorporate prior knowledge. Models do not become less powerful by removing flexibility, quite contrary do they become more powerful when providing the correct structure. In Appendix C we show that the model is able to learn a sensible encoding of a complex observation and learn the parameters of the underlying dynamical system, but it is still uncertain how difficult applications to real–world problems will be. Depending on the complexity of the system that is being modelled, a potentially large amount of data is required.

6.1 Future perspectives

- The model for human activity recognition in Section 4.2 relies on segmenting the observed time–series into windows and then treat them as individual samples. Recently an alternative training procedure in which windows of varying length are randomly sampled from the input, was proposed (Guan and Plötz, 2017). This approach avoids the issue of selecting an appropriate fixed window length and the theoretical bias that it may introduce, but as long as the window length is sufficiently long to cover any activity, it is not obvious that random lengths are advantageous. Automatic segmentation of the input signal can, however, be readily be achieved by borrowing methods from the single–shot learning literature in image segmentation and classification and would be applicable in both training and prediction (Liu et al., 2016; Chambon et al., 2018).

- Domain adaptation is useful in combining datasets but the same method, adversarial training, can be used to learn representations that are e.g. biometrically agnostic, which is important in the context of privacy. The flexible structure of the model in Section 4.2 enables any number of objective functions to be applied, one of which may encourage the model to learn a representation that is not able to discriminate between individuals. Another might encourage the model to learn a representation that does not discriminate gender or race, which is important in *fairness* (Barocas and Selbst, 2016; Xie et al., 2017).

- The current model formulation of KVAE includes a recurrent neural network, the dynamics parameter network (Section 5.2.1), that in each step selects the most probable out of a set of K parameters, given previous observations. The RNN is a black box but we could instead imagine a Hidden Markov Model (HMM) that selects the dynamics of the LGSSM, thereby providing complete transparency and the ability to analyse the transition dynamics of the HMM (Ghahramani and Hinton, 2000). Berardi et al. (2018) show that with temporally dense longitudinal data, the dynamics of the HMM can give a clearer picture of behavioural differences between interventions. Instead of just analysing whether physical activity increases after an intervention is started, the transition dynamics can provide insights into behaviour by looking at the probability of jumping between states.

- Most exciting is the potential that will be revealed by truly large scale studies and the application of advanced models as the one presented in Chapter 5. Current studies that qualify as large have not gathered multi–modal data and base their analysis on rather simple methods (Althoff et al., 2017).

6.2 Closing remarks

We set out to investigate how well smartphones would fare as tools in a clinical trial when used to measure energy expenditure and other modalities available on the platform. While we experienced significant hurdles along the way, we show that the smartphone is able to acquire a battery of useful information that is relevant in many disciplines. The two most important findings are the smartphones ability to track physical activity and predict energy expenditure, pervasively. Adoption of smartphones in large scale studies is increasing which means that tools to deploy experiments and collect data, should become increasingly available and reliable. We identified some troubling aspects, that other research has identified too, with using too small cohorts to make general claims from discoveries. And even if a claim is made within boundaries supported by the data, readers and popular media will too often disregard these bounding assumptions altogether. Based on these findings we suggest that pervasive, large scale experiments should become the default setting to do research, when possible. We acknowledge the difficulty of modelling large amounts of complex observations and propose a model that is suitable for this task.

APPENDIX A
Smartphone pervasive sensing of physical activity of overweight adults in a long-running randomized controlled trial

Simon Kamronn, Lars Kai Hansen and Jakob Eg Larsen. (2018). *Smartphone pervasive sensing of physical activity of overweight adults in a long-running randomized controlled trial.* Submitted JMIR mHealth

Original Paper

Smartphone pervasive sensing of physical activity of overweight adults in a long-running randomized controlled trial

Simon Kamronn, Lars Kai Hansen, Jakob Eg Larsen
Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

Corresponding Author:
Simon Kamronn
Cognitive Systems
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby
Denmark
Phone: +45 40929792
Email: simon@kamronn.com

Abstract

Background: Clinical trials are expensive why it should be a priority to acquire as much data as possible during the trial. The burden on participants and staff is often the limiting factor on the amount of data feasibly acquired, which is why trials may benefit from incorporating the readily available small sensor-packed ubiquitous device; the smartphone.
Objective: The aim of this study was to assess whether a smartphone can assist or replace existing practices in evaluating a physical activity intervention study in overweight sedentary adults.
Methods: We introduce the smartphone as an additional sensing device in a physical activity intervention study that investigates the effects of active commuting and leisure-time exercise on a range of biological measures.
Results: We find the smartphone is able to measure multiple modalities ubiquitously over a long duration and a hierarchical Bayesian analysis reveal estimates that are well in line with an independent analysis of the biological measures.
Conclusions: The smartphone has in this study shown that it, while not being without limitations, is able to augment current research methodologies and add value in historically infeasible ways. We can now ask questions that factorizes temporally on a minute-scale resolution and conceptually over the domains of everyday life.
Trial Registration: Clinicaltrials.gov NCT01962259

Keywords: mobile health, rct, smartphone, physical activity

Introduction

With more than half of the western population being overweight and nearly a quarter obese [1], the obesity pandemic and the associated non-communicable diseases [2] has become one of modern times biggest challenges and is only getting worse. Behavioural aspects of managing body weight by being physically active are largely not understood [3], and sometimes even counter-intuitive [4], why more research using the latest available technology is needed [5].

Before trying to manage body weight we must first monitor and understand the underlying dynamics. In this study we ask what effect active commuting (bicycle) and leisure-time activities have on physical activity in everyday life, when not exercising. Will the extra bouts of exercise be absorbed and decrease activity in other domains, maintaining status quo, or will they inspire and motivate more activity?

Conducting a randomized controlled trial (RCT) is typically expensive but also necessary if we want to understand systems in which causality is important. It is expensive primarily due to the excessive time spent on compliance and acquiring data (clinical sampling) and even though the sampling is time consuming, it is often only possible to shed light on the big picture questions. Pervasive monitoring can take the analysis to more detailed levels and explore novel directions [6]. Instead of viewing a three months period as one sample, activity tracking using smartphones can break that period into samples of one minute or less. Quantifying energy expenditure in terms of domains, places, and type of activity is simply not feasible in a traditional long-running study but the sensors of smartphones allow us to query these different directions and may provide insights otherwise missed.

ActiGraph (ActiGraph GT3x, ActiGraph Corp, Florida, USA), the de-facto standard of monitoring physical activity, has established its usefulness by being widely adopted and validated, but at least three aspects are in its disfavor. It uses an obsolete and unknown algorithm based on how mechanical activity trackers worked once and by being static is not able to adapt to different activities, e.g. walking or running, or different groups of people, e.g. children, adults, and handicapped, which would enable much more accurate estimates [7]. Secondly, conducting a large scale, long-term study quickly becomes cumbersome as the price of acquisition and the personal burden on participants of carrying and maintaining an extra device can be limiting factors in studies. Lastly, and this is for many monitoring methods: the psychological effects of being in a study, and in particular being reminded continually from a device on the hip, can have significant impact on the behaviour that is being studied [5]. Simply by carrying a pedometer, it was found in a systematic review [8], people increase their physical activity by 26.9%.

Given the dominance of ActiGraph and the many previously established health outcomes based on the device [9,10], [11] investigated to what extent Android smartphone based activity sensing correlated with ActiGraph in both laboratory and free-living conditions. They found that physical activity estimates from the smartphones were highly correlated with estimates from ActiGraph and with one another, and that it mattered little in what position the smartphone was carried. Using a smartphone has several advantages, beyond avoiding to carry an additional device. The vast majority already owns a smartphone which makes it a cheap and easy way to deploy scalable, real time, and truly pervasive studies, if it reliably can estimate energy expenditure and activity. It further has the ability to provide feedback to participants [4,12,13], which may be an important component in fighting obesity after tracking and understanding it, hereby closing the loop and hopefully guiding us towards more healthy choices.

We apply the smartphone as a monitoring device in a physical activity intervention study spanning six months with the primary aim of investigating the effects of active commuting and leisure-time activities on physically inactive obese participants. Will we be able to measure the effects of interventions based on smartphone sensing data and will they be in accordance with the clinical results?

Methods

Participants

Caucasian, overweight, and physically inactive healthy adults were recruited in the Copenhagen area from November 2013 to October 2015. Inclusion criteria included age of 20 to 45, body mass index (BMI) from 25 to 35 kg/m^2, and a self-reported physically inactive lifestyle. The full list of of criteria are available in [14].

Study design

The study design is described in full elsewhere [14] but most importantly it was designed as a randomized controlled trial to test the effectiveness of three interventions on physical activity and one control group. They were prescribed a daily exercise energy expenditure of 320 kcal for women and 420 kcal for men. The four groups are:
BIKE: bicycle commuting to work
MOD: leisure time moderate activity (50% of VO$_2$max)
VIG: leisure time vigorous activity (70% of VO$_2$max)
CON: The control group was asked to maintain the same lifestyle as prior to the study.

Clinical sampling

Participants were asked to attend three different test days at baseline, 3 months, and 6 months where a range of physiological tests were conducted and biological samples acquired. Immediately after the test-days a free-living assessment period (7-14 days) followed in which they were monitored on physical activity (ActiGraph GT3x, 7 days), sleep, diet, and daily energy expenditure through doubly labelled water [15]. They were furthermore required to wear a Polar heart rate monitor (RC3 GPS, Polar Electro Oy, Kempele, Finland) every time they exercised and weigh themselves every day of the study using a wireless internet-connected bodyweight (WiThings Body Composition WiFi scale, WiThings Europe,Issy-les-Moulineaux, France).

Pervasive sampling

A mobile sensing application developed for a previous smartphone sensing study [16] was adopted and configured for use in the present study. The application was configured to acquire acceleration (50Hz sampling frequency), GPS-based location (6-minute sampling interval), recognized activity through the Google Activity Recognition API, step count, and screen on/off status.

Implementation details

To accommodate the high sampling rate of the accelerometer without reducing the battery-life of the smartphone, samples are 'batched', i.e., queued in hardware before read and saved to the storage. At the time of the study only two smartphones supported batching, LG Nexus 5 and Samsung S5. The former was chosen as it was successfully used previously [16] but during the enrollment period it was unfortunately decommissioned and Samsung S5 was used hereafter. A total of 19 LG and 11 Samsung smartphones are used in the cohort that completed the experiment.

Missing data

In an experiment with limited control some level of missing data must be expected. In the present study we also experienced significant missing data. The observed patterns of missingness are not always "random" or explainable. For example, during night-time the smartphones may reduce the sampling of the accelerometer, such loss of data may increase variance but not introduce a bias, as the effects can be modeled. In addition data loss occurred during the day with no apparent pattern. In some cases all the modalities are lost, indicating a general failure to record, send, or receive the data, and at other times it is a single modality indicating a failure on the smartphone. If such losses occur without dependence of the outcome variables it can again lead to increased variance but should not introduce a bias.

Pre-processing

Activity level calculation

The primary outcome is the level of physical activity but since that is not a well-defined measure, especially using smartphone sensors, *activity counts* are used to make the conclusions comparable with ActiGraph results. Activity counts are a measure of physical activity within a time span, or more precisely, a motion summary count calculated from the area under the accelerometer curve [7]. In practice this means that the signal is bandpass filtered with a range that preserves as much human motion related acceleration as possible while discarding noise, which is around 0.25 - 5Hz. The absolute values are then summed in bins of 1 second. This leaves activity counts for each dimension but can without loss of information be reduced to the *vector magnitude count* (VMC) which is the euclidean distance spanned by the count vector.

Domain inference

Spatial location data is complex and noisy so to reduce complexity we use a simple behavioral cluster representation [17]. We identify 4 different *domains* for each participant, namely: Home, Leisure, Travelling, and Unknown. Home is manually identified by having the most samples associated and typically active during the evening/night. Travelling is identified as being en route between two locations with a sufficiently high average speed. Leisure is identified as samples that are not Home, Work or Travelling. We also wanted to identify when participants were at work but it turned out to be impossible for more than half of the participants due to either too much missing data, unemployment, or a non-stationary work such as postman. For this reason the locations identified as work has been merged into the Unknown category.

Aggregation

The accelerometer samples at 50Hz and therefore gathers a very large amount of data in six months. To enable statistical analysis all modalities are summarized in 5 minute bins. Continuous data such as the activity count is summed and with categorical data the most frequent category within a bin is used.

Statistical Analysis

A long running experiment conducted in the real world in which the participants are left to their own devices will naturally include many confounding factors which are mostly unobservable. Due to the randomized treatment assignment we can theoretically ignore all covariates but may be able to detect a weaker signal by including them [18]. Including all covariates and interactions will provide the most flexible model [19]but also a model that is often over-parameterized and non-identifiable why covariates should be chosen carefully [20]. Especially when the number of participants is low or covariates are correlated with treatment assignment. In model-based analysis of randomized controlled trials the response variable can be approximated with a hierarchical Bayesian model

$$y_i = \mu + \mu_{s[i]}^{pre} + \tau_{t[i]} + \alpha_{g[i]} + \beta X_i + \epsilon_i,$$

where y_i is the response, μ is the intercept, μ_s^{pre} is the subject-specific pre-treatment mean, τ is the treatment assignment, α are random covariates, β are regression coefficients, X are the linear regressors, and $\epsilon_i \sim N(0, \sigma^2)$. The $g[i]$ notation specifies which group, of a particular covariate, that sample *i* is from. $s[i]$ denotes participants and $\tau[i]$ denotes treatment. μ_s^{pre} and the covariates are sum-to-zero contrasted which ensures the model is identifiable and the posterior distribution of τ is thus a measure of the treatment effect. This is a hierarchical Bayesian model when proper prior distributions are placed on the parameters [21] and otherwise known as a linear mixed model. The prior both regularizes parameter estimates and shares statistical strength between sub-groups in the data through the concept of partial pooling [22].

Varying effects

It may be overreaching to assume the treatment effects are constant over the duration of the experiment, and certainly after the experiment is finished, so the treatment effects have been allowed to vary with the natural periods of the study. The intervention is divided into 4 periods; pre-treatment, 0 to 3 months, 3 to 6 months, and after 6 months. The pre-treatment period is already modelled in the pre-treatment mean μ_s^{pre} and thus not modelled in the treatment effect. It is further assumed that the effect varies with domain (Home, Leisure, Travelling, Unknown). We thus have a treatment effect partitioned into 4 groups, 3 periods, and 4 domains.

Controlling covariates

To control for some of the variability caused by observed baseline variables they are included in the model in the term $\alpha_{g[i]}$, also called varying intercepts. Seasonal, weekly, and daily variation are modelled by including the month, weekday, and hour of the day as covariates, i.e. for month a 12-dimensional varying intercept. Each of these have a normally distributed zero-centered prior with standard deviation of 1. We further control for gender (male, female), relationship status (single, cohabiting), education (no college, college, grad school), and employment status (employed, student). In the analysis of the smartphone data the type of smartphone is also added as varying intercept. These have a normally distributed zero-centered prior with standard deviation of 0.01.

Linear regressors

In the linear term β we include age, body mass index (BMI), maximal oxygen uptake (VO_2max), temperature and rainfall in the vicinity and within the hour of a sample, and the number of sick/off-days they have from the study in a given period. All the coefficients have a normally distributed zero-centered prior with standard deviation of 0.005.

Missing data

Because the activity level data is originally on the scale of a second and then summed over a fixed bin size *N*, if any samples in a bin is missing we get that

$$y_i = \sum_{n=1}^{N_{obs}} y_{in}^{obs} + \sum_{n=1}^{N_{mis}} y_{in}^{mis}.$$

One way of dealing with the missing samples is treating them as latent variables, or model parameters, but if the bin size is small enough or data is missing at random, we can assume equal observations within the bin, i.e. $y_i = N y_{in}^{obs}$ for any $n \in [1, \ldots, N_{obs}]$, and thus assume

$$y_i^{obs} = \frac{N^{obs}}{N} y_i,$$

which scales the regression so we calculate an approximately correct likelihood. To relax this assumption the scaling is itself being scaled with a parameter $\lambda_{s[i]}$ for each participant

$$y_i^{obs} = \lambda_{s[i]} \frac{N^{obs}}{N} y_i,$$

which enables the model to learn if a particular participant has a behaviour that warrants an even lower scaling.

Completely missing data

Often we have no data within the summarized bin and so have to treat the sample as missing to ensure a proper posterior distribution. In these cases we treat the missing sample as a latent variable so that $y_i^{mis} \sim N(0, \sigma_{mis}^2)$. If the amount of observed samples in a bin is less than 10% it is treated as unobserved as well.

Non-compliance

Not carrying the smartphone, especially while exercising, is a situation of non-compliance and unfortunately often observed in the data. We are able to estimate the compliance at few intervals due to multiple data sources but often we are not. As we observe many instances of non-compliance during the prescribed exercises, which we know due to Polar HR data, all periods containing exercises have been removed from the data and the samples are treated as unobserved, following standard practice [22].

Inference

The model is implemented in the probabilistic programming framework Edward [23] because it enables fast inference even when scaling to millions of samples. Relying on variational inference may be less accurate than sampling methods such as Markov Chain Monte Carlo (MCMC). To test this MCMC inference using Stan [24] was carried out for a subset of the data, which showed very similar results to Edward for the same data. To select model architecture, i.e. which covariates to include and how to manage missing data, the model with the highest log-likelihood was selected.

Results

Of the 47 recruits receiving a smartphone, 30 participants completed the study. The distribution of participants among the interventions are: 6 CON, 7 BIKE, 12 MOD, and 5 VIG. The mean age of participants was 36 years (range: 22–45), with 19 females (63%). Most were cohabiting (n=22, 73%), in employment (n=27, 90%), had no college degree (n=17, 57%), and only 2 had gone to grad school (7%). After summarization we have a total of 2,311,151 samples, or 8024 days. For our outcome variable we have an observation rate of 30.5% and the other modalities are adaptively sampled, i.e. without a fixed sampling rate. The total number of days for each participant in the study is capped at 300 from the first day, including baseline, and the average number of days in the study is 275 with a standard deviation of 27.

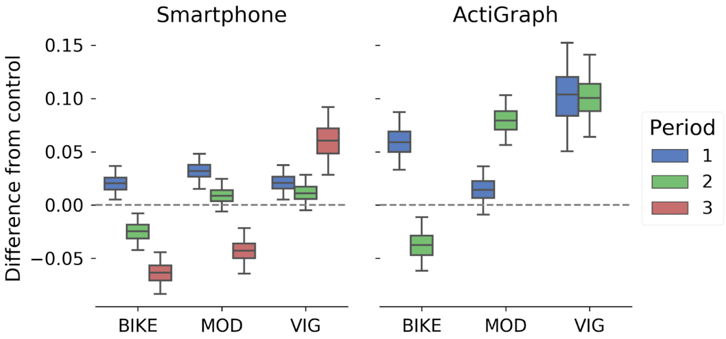


Figure 1: Treatment effect

Posterior estimate of each treatment relative to CON in period 1 (baseline to 3 months), 2 (3–6 months), and 3 (after 6 months) of the study. The whiskers indicate 95% interval. For smartphone data we see a positive effect for all of the interventions in the first period and then a decline in the second. For the ActiGraph data we see a more varied result with positive effects in the second period as well.

Effects of interventions

Treatment effects measured by non-exercise activity of each intervention contrasted to the control group for each of the two periods in the study measured by the two devices are shown in Figure 1. As a participant-specific pre-treatment intercept is also modelled, the effects are in contrast to the estimated baselines as well. In the effects measured by the smartphone we see a pattern in which it appears that the participants initially are motivated to be more active but after 3 months the activity declines, not unlike similar studies [4]. However, even though the effects for MOD and VIG in the second period are not significantly different from CON, they are neither significantly different from the effects in period one, so while we do observe a relative decrease in activity with respect to CON, we can not be sure of an absolute decrease of each group. BIKE is the group with the most significant drop in activity with both period two and three being less active than CON and we see the same for the third period of MOD. The third period for VIG, however, is the most active of them all, a result currently with no obvious explanation as it is an uncontrolled period without data from any other modality.

In the effects measured by ActiGraph the effects for the first period at 3 months is positive but becomes negative in the second period for BIKE, following the trend observed from the smartphone. MOD and VIG, however, becomes more positive and stays the same, respectively, which is almost the opposite of what we see in the smartphone data.

A comparison to loss of fat

In [25] the fat mass of the same participants as in the present study was measured with dual-energy X-ray absorptiometry. Here a significant loss of fat mass was found in all three groups compared to CON in the first period (baseline to 3 months), and a sustained fat mass in the second period (from 3 to 6 months), for BIKE and MOD but further decreased for VIG. If we compare the loss of fat mass to the physical activity effect (Figure 1) measured by the smartphone, those results are highly correlated and make sense if we assume that increase in non-exercise activity leads to loss of fat mass. If we were to include activity measured during prescribed training exercises, the measured effect would be even greater and probably better correlated to loss of fat mass. This suggests that smartphones are able to predict the loss of fat mass.

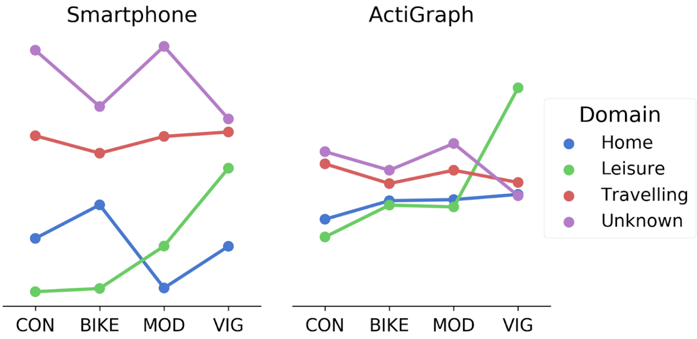


Figure 2: Effect of domains between groups

Treatment effect summed over the periods. The posterior estimates are not contrasted to CON (as in Figure 1) and therefore not directly interpretable as a treatment effect why no y-axis is shown but they can be used for comparison between domains and interventions. For the smartphone we see e.g. MOD and VIG are more active during Leisure, and that Travelling and Unknown generally elicit higher activity.

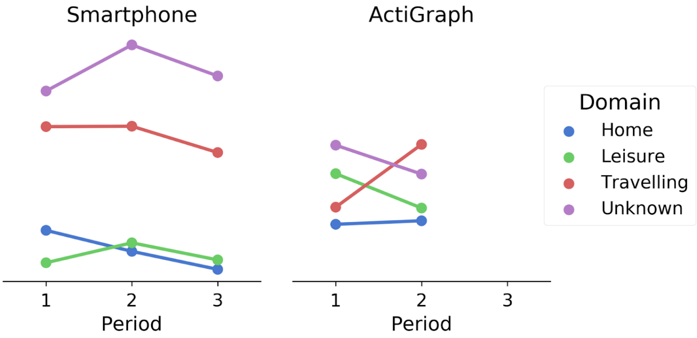


Figure 3: Effect of domains over time

Treatment effect summed over the interventions. In the smartphone data we see a general increase of activity in the second period followed by a decrease after the end of the study in the third period. The ActiGraph data shows a change in which domain the activity is high between the periods.

Effects of domain and time

In Figure 2 we see significant differences in the amount of physical activity spent in the different domains even though the samples are distributed relatively evenly with about a third in the Unknown domain and an equal share in the three others. The Unknown domain is unfortunately also the most active across all interventions. Secondary is Travelling which suggests the participants regularly walk between destinations, e.g. to the grocery store, as motorized transportation typically elicit a low accelerometer response. The groups CON and BIKE are more active at Home than away whereas MOD and VIG are inactive at Home and VIG is very active in Leisure. The exercises are excluded from the data and can therefore not explain why VIG is so active during Leisure but some of the explanation might be that it has less Unknown samples than e.g. MOD. In the ActiGraph data the effect of domains are less pronounced, perhaps because VIG in Leisure is so dominating, but apart from that the trend follows the smartphone data rather well. We see a bit more activity at Home in MOD and VIG which might be due to participants not carrying their phones.

If Figure 3 the effect of domains is shown for each period where the third period is after the 6 months when the study is completed but participants have continued collecting data on their smartphones for a time up to 300 days post commencement. It appears the total level of physical activity increases from the first to the second period which naively contradicts the decrease of treatment effect (Figure 1). The explanation is simply that while the treatment effect is in contrast to CON, the domain effect is a sum over all the groups and can therefore not be used to evaluate the effect of interventions. It is, however, interesting, and perhaps expected, that the level of activity decreases significantly in all domains after the study is completed. The results from the ActiGraph (Figure 3) show a quite different pattern in which it seems the Travelling domain has absorbed activity from both Leisure and Unknown from period 1 to 2.

Why do the results differ between devices?

Data from the ActiGraph is acquired at three times; at baseline, 3 months, and 6 months, and for a maximum of 7 days each time. This leads to less data and therefore less stable inference even though the posterior estimates are significant. In a free-living study there will be unobservable confounding factors and one might be the psychological effect of carrying an accelerometer on the hip. The ubiquitous nature of the smartphone leads in this case to more reliable estimates.

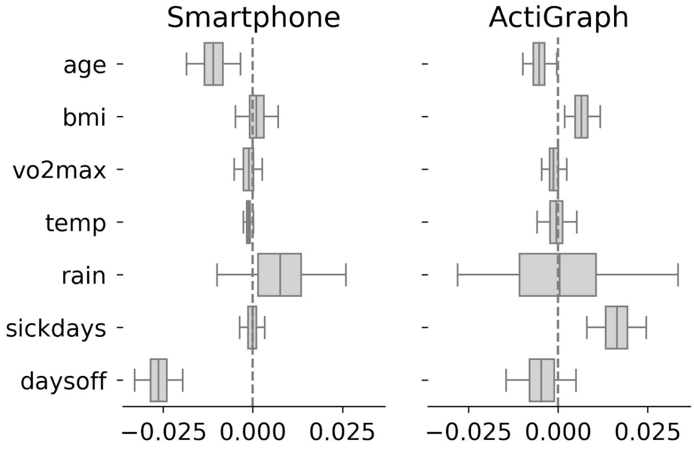


Figure 4: Linear regressors

Coefficients β of the linear regressors with whiskers indicating 95% interval. A positive value indicates a positive correlation with, and therefore increase in, physical activity. Age is slightly negatively correlated, BMI, VO_2max, temperature, rain and sick days have no effect and the number of days off from the study is also negatively correlated.

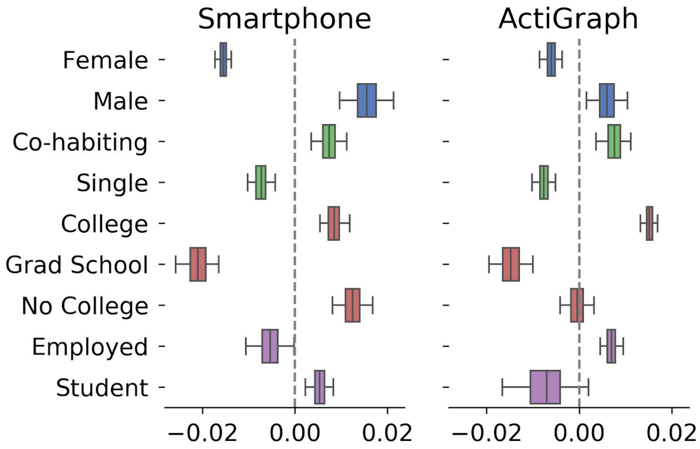


Figure 5: Covariates

The varying intercepts α with whiskers indicating 95% interval. All groups (e.g. gender or education) are sum-to-zero contrasted. Being a cohabiting male without a graduate school degree you have a higher probability of being physically active. It is uncertain if being employed or a student is associated with higher levels of activity.

Characteristics

External factors and personal characteristics may influence a measured response to a degree in which the effect of interest is either lost in noise or incorrectly estimated due to e.g. confounders. For this reason all relevant information of the participants and their surroundings are included in the model either as a linear predictor or a varying intercept.

Linear predictors

In Figure 4 estimates of the regression coefficients are illustrated in a horizontal boxplot with whiskers indicating 95% intervals. Most of the coefficients are either very small or insignificant but age is interestingly a predictor of less physical activity, which is a common finding [26,27]. The number of days off the study is the strongest predictor of less activity indicating that for whatever reason participation was hindered also limited their non-exercise activity. In the ActiGraph data VO$_2$max is a small but positive contributor and we see a moderately strong contribution from the number of sick days a participant has had during a given period in the study. It is unsure if sick days leads to more exercise, which would be unexpected, or the activity during the days of wearing the ActiGraph happen to correlate with the number of sick days.

Varying intercepts

We observe large significant effects of most covariates when modelling the smartphone data with the one exception that it is uncertain whether having a job or being a student is important (Figure 5). Notably it seems men are more active than women, which has been previously observed [26], highly educated people are less active than lower educated, and participants with a partner are more active, perhaps due to motivational support. Covariate estimates from the ActiGraph data are quite similar with the exception of having a reversed relationship between employed and student participants but as the ratio between these groups are very skewed, this result should not be interpreted strongly.

Time-dependent intercepts show the expected behaviour of high activity during the day (Figure 6), more activity in the summer months (Figure 8), and less activity in the weekends compared to the weekdays (Figure 7).

Discussion

Principal findings

This is the first long-term randomized controlled trial to investigate the effects of active commuting and leisure-time exercise on physical activity in sedentary obese and overweight people using smartphones and ActiGraph. We find the smartphone is a resourceful platform for research but also that is has both strengths and limitations that needs to be better quantified.

Strengths

As a strength the smartphone may lower the bar to including multidimensional, longitudinal data streams in any study regardless of participant technological proficiency and furthermore have the potential to not only provide easier enrollment but also recruitment through platforms such as the ResearchKit (Apple, Cupertino, CA, USA), increasing reach and decreasing cost of marketing [28]. Selection bias may be lessened by using smartphones as the primary sensor since the vast majority owns a smartphone already and time-demanding activities such as spending multiple days on physiological examinations may no longer be needed, reducing the burden on participants. Reducing examinations can in turn allow researchers to spend more time on activities that improve retention. Clinicians will initially, and rightly, argue that high-variance data as we acquire from most smartphone sensors, are no substitution for low-variance momentary clinical measurements and depending on the endpoint that may be correct, but the sheer volume and variety of data made available through smartphones not only enables low-variance estimates, they allow researchers to explore new directions.

Limitations

In terms of limitations, a too short time-span was available to tailor the existing mobile sensing application before enrolling the first participant which also meant that insufficient time was available to conduct a pilot study to sufficiently test the application. In hindsight it would have been prudent to improve sampling consistency and server infrastructure to reduce the amount of missing data. Newer developments in this area with ResearchKit and ResearchStack [29] will undoubtedly improve much upon this issue and general adaptation. Having access to raw data is always preferred but sampling multidimensional streams at a high frequency introduces a multitude of potential issues with battery-life, storage, data-transfer, etc., that can be mitigated by edge-computing; carefully choosing some pre-processing to take place on the smartphone.

Previous research find strong correlations between activity estimates from smartphones and ActiGraph [11] why the observed discrepancies in the present study are likely due to the difference in the time-span covered by the observations and non-observable confounders. The relatively small sample size and psychological effects of carrying a physical activity tracker could also cause enough bias to throw off the analysis [5]. An explanation that is supported by the agreement of linear regressor and covariate estimates.

Conclusions

Pervasive sensing technologies show great potential in clinical research as either a substitution, enhancement, or addition to existing methodology. We find that the smartphone provides reliable estimates in correspondence with independently measured loss of fat mass. In the results we see a decrease in physical activity with time but we also see a shift in the distribution of activity over the domains with activity moving from Home to Leisure and that VIG in general has a much greater tendency to be active in Leisure than at Home. By combining multi-modal data it is possible to decompose measured effects into time, domain, and many other directions which may contain previously hidden insights to research questions.

Acknowledgements

The authors would like to thank the team behind GO-ACTIWE; Bente Stallknecht, Mads Rosenkilde, Jonas Salling Quist, Martin Bæk Petersen, and Anne Sofie Gram, for conducting the randomized controlled trial and collecting ActiGraph data. We would also like to thank TrygFonden for funding the project.

Conflict of Interest

None declared

Abbreviations

BMI: body mass index
RCT: randomized controlled trial

References

1. Badimon L, Bugiardini R, Cenko E, Cubedo J, Dorobantu M, Duncker DJ, Estruch R, Milicic D, Tousoulis D, Vasiljevic Z, Vilahur G, de Wit C, Koller A. Position paper of the European Society of Cardiology-working group of coronary pathophysiology and microcirculation: obesity and heart disease. Eur Heart J 2017 Jul 1;38(25):1951–1958. PMID:28873951

2. Hallal PC, Andersen LB, Bull FC, Guthold R, Haskell W, Ekelund U, Lancet Physical Activity Series Working Group. Global physical activity levels: surveillance progress, pitfalls, and prospects. Lancet Elsevier; 2012 Jul 21;380(9838):247–257. PMID:22818937

3. Finkelstein EA, Haaland BA, Bilger M, Sahasranaman A, Sloan RA, Nang EEK, Evenson KR. Effectiveness of activity trackers with and without incentives to increase physical activity (TRIPPA): a randomised controlled trial. Lancet Diabetes Endocrinol Elsevier; 2016 Dec;4(12):983–995. PMID:27717766

4. Jakicic JM, Davis KK, Rogers RJ, King WC, Marcus MD, Helsel D, Rickman AD, Wahed AS, Belle SH. Effect of Wearable Technology Combined With a Lifestyle Intervention on Long-term Weight Loss: The IDEA Randomized Clinical Trial. JAMA jamanetwork.com; 2016 Sep 20;316(11):1161–1171. PMID:27654602

5. Sullivan AN, Lachman ME. Behavior Change with Fitness Technology in Sedentary Adults: A Review of the Evidence for Increasing Physical Activity. Front Public Health frontiersin.org; 2016;4:289. PMID:28123997

6. Maddison CJ, Lawson D, Tucker G, Heess N, Norouzi M, Mnih A, Doucet A, Teh YW. Filtering Variational Objectives [Internet]. arXiv [csLG]. 2017. Available from: http://arxiv.org/abs/1705.09279

7. Peach D, Van Hoomissen J, Callender HL. Exploring the ActiLife® filtration algorithm: converting raw acceleration data to counts. Physiol Meas IOP Publishing; 2014 Nov 12;35(12):2359.

8. Bravata DM, Smith-Spangler C, Sundaram V, Gienger AL, Lin N, Lewis R, Stave CD, Olkin I, Sirard JR. Using pedometers to increase physical activity and improve health: a systematic review. JAMA

jamanetwork.com; 2007 Nov 21;298(19):2296–2304. PMID:18029834

9. Strath SJ, Kaminsky LA, Ainsworth BE, Ekelund U, Freedson PS, Gary RA, Richardson CR, Smith DT, Swartz AM, American Heart Association Physical Activity Committee of the Council on Lifestyle and Cardiometabolic Health and Cardiovascular, Exercise, Cardiac Rehabilitation and Prevention Committee of the Council on Clinical Cardiology, and Council. Guide to the assessment of physical activity: Clinical and research applications: a scientific statement from the American Heart Association. Circulation 2013 Nov 12;128(20):2259–2279. PMID:24126387

10. Buman MP, Hekler EB, Haskell WL, Pruitt L, Conway TL, Cain KL, Sallis JF, Saelens BE, Frank LD, King AC. Objective light-intensity physical activity associations with rated health in older adults. Am J Epidemiol 2010 Nov 15;172(10):1155–1165. PMID:20843864

11. Hekler EB, Buman MP, Grieco L, Rosenberger M, Winter SJ, Haskell W, King AC. Validation of Physical Activity Tracking via Android Smartphones Compared to ActiGraph Accelerometer: Laboratory-Based and Free-Living Validation Studies. JMIR mHealth and uHealth 2015;3:e36.

12. Buman MP, Epstein DR, Gutierrez M, Herb C, Hollingshead K, Huberty JL, Hekler EB, Vega-López S, Ohri-Vachaspati P, Hekler AC, Baldwin CM. BeWell24: development and process evaluation of a smartphone “app” to improve sleep, sedentary, and active behaviors in US Veterans with increased metabolic risk. Behav Med Pract Policy Res Springer US; 2015 Nov 9;1–11.

13. Harries T, Eslambolchilar P, Rettie R, Stride C, Walton S, van Woerden HC. Effectiveness of a smartphone app in increasing physical activity amongst male adults: a randomised controlled trial. BMC Public Health 2016 Sep 2;16:925. PMID:27590255

14. Rosenkilde M, Petersen MB, Gram AS, Quist JS, Winther J, Kamronn SD, Milling DH, Larsen JE, Jespersen AP, Stallknecht B. The GO-ACTIWE randomized controlled trial - An interdisciplinary study designed to investigate the health effects of active commuting and leisure time physical activity. Contemp Clin Trials 2017 Feb;53:122–129. PMID:28007633

15. Westerterp KR. Doubly labelled water assessment of energy expenditure: principle, practice, and promise. Eur J Appl Physiol 2017 Jul;117(7):1277–1285. PMID:28508113

16. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, Lehmann S. Measuring large-scale social networks with high resolution. PLoS One 2014 Jan;9(4):e95978. PMID:24770359

17. Cuttone A, Lehmann S, Larsen JE. Inferring human mobility from sparse low accuracy mobile sensing data. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp ’14 Adjunct 2014;995–1004.

18. Gelman A. Experimental reasoning in social science [Internet]. 2010. Available from: http://www.stat.columbia.edu/~gelman/research/published/yalecausal2.pdf

19. Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang [Internet] 2013 Apr;68(3). PMID:24403724

20. Bates D, Kliegl R, Vasishth S, Baayen H. Parsimonious mixed models. arXiv preprint arXiv:1506 04967 [Internet] 2015; Available from: https://arxiv.org/abs/1506.04967

21. Feller A, Gelman A. Hierarchical Models for Causal Effects. Emerging Trends in the Social and

Behavioral Sciences John Wiley & Sons, Inc.; 2015.

22. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis, Third Edition. CRC Press; 2013. ISBN:9781439840955

23. Tran D, Kucukelbir A, Dieng AB, Rudolph M, Liang D, Blei DM. Edward: A library for probabilistic modeling, inference, and criticism [Internet]. arXiv [statCO]. 2016. Available from: http://arxiv.org/abs/1610.09787

24. Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. Stan: A Probabilistic Programming Language. Journal of Statistical Software, Articles uvm.edu; 2017;76(1):1–32.

25. Quist JS, Rosenkilde M, Petersen MB, Gram AS, Sjödin A, Stallknecht B. Effects of active commuting and leisure-time exercise on fat loss in women and men with overweight and obesity: A randomized controlled trial. Int J Obes [Internet] 2017 Oct 10; PMID:28993707

26. Althoff T, Sosič R, Hicks JL, King AC, Delp SL, Leskovec J. Large-scale physical activity data reveal worldwide activity inequality. Nature 2017 Jul 20;547(7663):336–339. PMID:28693034

27. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, White T, van Hees VT, Trenell MI, Owen CG, Preece SJ, Gillions R, Sheard S, Peakman T, Brage S, Wareham NJ. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. PLoS One 2017 Feb 1;12(2):e0169649. PMID:28146576

28. Chan Y-FY, Wang P, Rogers L, Tignor N, Zweig M, Hershman SG, Genes N, Scott ER, Krock E, Badgeley M, Edgar R, Violante S, Wright R, Powell CA, Dudley JT, Schadt EE. The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. Nat Biotechnol [Internet] Nature Research; 2017 Mar 13 [cited 2017 Mar 13]; [doi: 10.1038/nbt.3826]

29. Estrin D, Carroll M, Lakin N. ResearchStack [Internet]. researchstack.org. [cited 2018 Apr 10]. Available from: https://www.webcitation.org/6yZVKkfwM

Time-dependent varying intercepts

Figure 6: Hourly variation
The effect of hour of the day on the activity level. The effects between devices are quite similar and follow a trend that is agreeable with previous results [27].

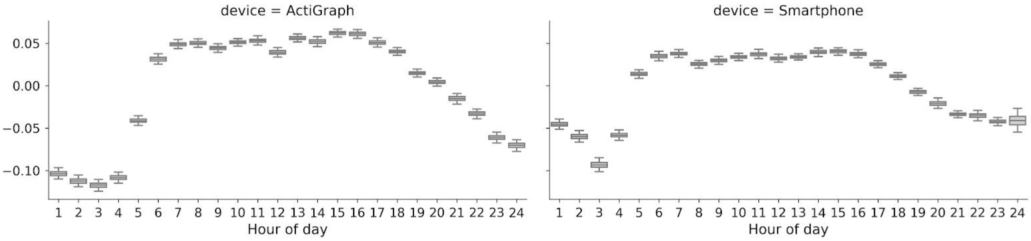


Figure 7: Daily variation
The effect of day of the week on the activity level. Somewhat similar variation between devices where mostdays are not significantly different from each other and the weekend seems to be for resting.

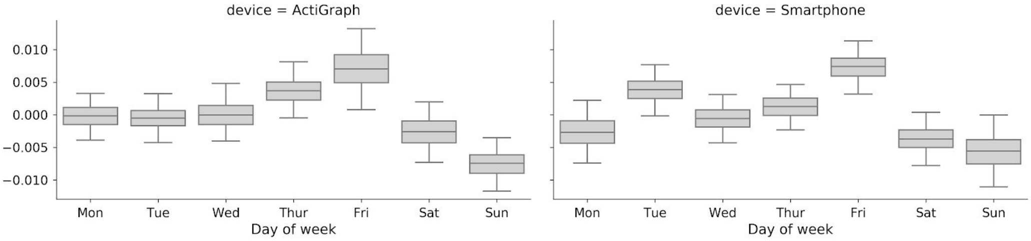
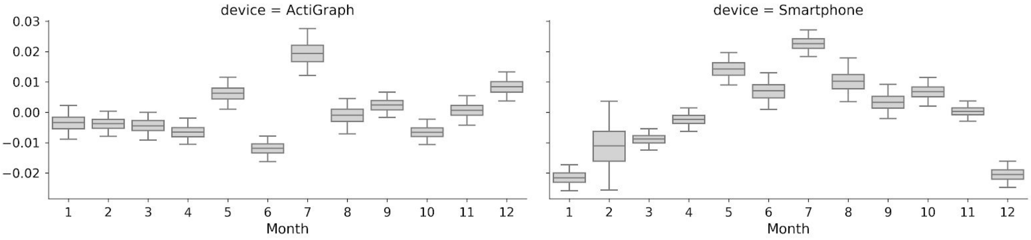


Figure 8: Monthly variation
The effect of month of the year on the activity level. We see little correspondence between the devices and note that the variations observed in the smartphone data are closer to what would be expected with better weather yielding higher activity.



APPENDIX B

Estimating energy expenditure from smartphones. A free-living long-term comparison with doubly labelled water

Simon Kamronn, Jonas Salling Quist, Martin Bæk Blond, Anne–Sofie Gram, Kevin Hall, Peter Walter, Mads Rosenkilde, Bente Stallknecht and Jakob Eg Larsen. (2018). *Estimating energy expenditure from smartphones. A free-living long-term comparison with doubly labelled water.* Submitted to Scientific Reports

Estimating energy expenditure from smartphones. A free-living long-term comparison with doubly labelled water

Abstract

Doubly labelled water (DLW) is the gold standard method to measure free-living energy expenditure (EE), but it is expensive and laborious to use. With the advent of wearable sensor technologies capable of measuring mechanical and physiological signals, it is possible to estimate EE in a pervasive way that in addition provides higher temporal resolution. The objective of this study was to look beyond the dedicated devices; to ascertain how well a smartphone with integrated sensors is capable of estimating EE. In a 6 month randomised controlled trial of sedentary obese individuals, EE was measured with DLW, and physical activity was determined using the accelerometer sensor in a smartphone. A simple linear regression model to predict EE from smartphone data was found inaccurate due to physiological inter-person variation. Taking advantage of multiple sources of data, a two-stage regression model with a common mediator that is calibrated to each individual, can predict EE with strong correlation ($r=0.84$, $P < 0.001$). Accurate EE estimates from smartphones have the potential to provide novel insights into behavioural aspects of physical activity, but how to generalize regression models of EE to populations remains an open question.

Introduction

Prediction of physical activity energy expenditure (PAEE) from pervasive technology such as embedded inertial sensors is an active research area with great commercial, societal, and scientific value (Jeran, Steinbrecher, and Pischon 2016). With nearly a quarter of the western population being obese, the associated non-communicable diseases such as diabetes and stroke threaten to overload the healthcare systems (Badimon et al. 2017). Exercise and a hypocaloric diet are the preferred remedies for obesity (Maher et al. 2013), with exercise being more effective in reducing visceral adiposity (Verheggen et al. 2016).

Physical activity level is a product of genetic, cultural, cognitive, and many other factors. Objective measures of physical activity can e.g. be used as a predictor of changes in mental state (Faurholt-Jepsen et al. 2014; Chapman et al. 2017) or to influence people's behaviour into more healthy lifestyle patterns when used in combination with behaviour-change techniques (Chin et al. 2016; Direito et al. 2017). Wearable activity trackers can motivate people to increase physical activity, but there is no evidence that the increase is maintained over time (Sullivan and Lachman 2016; Finkelstein et al. 2016). Individuals who purchase activity trackers are already motivated to develop healthy behaviours. However, increasing the physical activity level in the “unmotivated” group has great societal implications and smartphones might prove to be an important entry point, as most people already possess one.

For decades inertial sensors, such as accelerometers, have been used to approximate PAEE (Montoye et al. 1983; Bouten et al. 1994). They have been validated when continuously worn on a known position on the body, e.g. the hip, leg, wrist etc., but with varying degrees of success (Plasqui and Westerterp 2007; Sardinha and Júdice 2017; Jeran, Steinbrecher, and Pischon 2016; de Graauw et al. 2010; Ferguson et al. 2015). Due to different levels of cardiorespiratory fitness and body weight of individuals, it is not possible to accurately model EE directly from physical activity alone (Bonomi and Westerterp 2012), as it requires calibration to the individual or at least the activity performed (Crouter, Bassett, and Freedson 2009; Bassett, Rowlands, and Trost 2012; Altini et al. 2015; Bonomi et al. 2009).

A significant increase in prediction accuracy is observed when including characteristics such as sex, body weight, and age in the model (Sardinha and Júdice 2017; Ingraham, Ferris, and Remy 2017), but these measures are crude and as the heterogeneity of the sample population increases, model performance decreases. Combinations of accelerometers and other modalities such as heart rate (HR) monitors (Brage et al. 2015; Murakami et al. 2016; Chowdhury et al. 2017; Beltrame et al. 2017; Ingraham, Ferris, and Remy 2017) will generally increase prediction accuracy, but are also considerably more obtrusive than a simple accelerometer and are therefore as yet infeasible in large-scale studies.

The objective of this study was to assess how well the smartphone can predict EE over a longer period (14 days). No restrictions were imposed upon the participants, and therefore unobserved non-wear time and variability from how the smartphone is carried and used must be expected. Smartphones have previously been compared to research-grade accelerometers in free-living conditions and found comparable (Hekler et al. 2015; Maddison et al. 2017; Bort-Roig et al. 2014), but only during a short period (≤ 7 days) and only evaluated on PAEE in laboratory settings (Maddison et al. 2017). To our knowledge, this is the first study to evaluate the level of accuracy of the smartphone as a pervasive monitor of EE in everyday life using DLW as ground truth.

Methods

Participants

In the period from November 2013 to October 2015, Caucasians with overweight or class 1 obesity were recruited from the greater Copenhagen area. They were required to be physically inactive, 20-45 years of age, have a body mass index from 25 - 35 kg/m^2, non-smoking, and with normal plasma glucose (< 6.1 mmol/l) and blood pressure (<140/90 mm Hg). They were recruited using advertisements in local media and on the Internet, including social media. Further information is available elsewhere (Rosenkilde et al. 2017).

Characteristics of included participants

Gender	Count	Weight (kg)	Fat %	BMI	Age
Female	7	84.2 (7.5)	43.7 (2.9)	29.1 (2.3)	37.3 (7.1)
Male	4	94.6 (7.7)	31.8 (2.3)	30.4 (1.6)	37.0 (7.2)

Table 1. Baseline characteristics. BMI: body mass index, kg ▪ m^{-2}. Data is shown as mean (standard deviation).

Study design

The study was a 6-month randomised controlled trial with three identical test periods at baseline, 3 months, and 6 months. Participants were randomised to one of three exercise interventions (BIKE, MOD, VIG) or a control group. Prescribed daily exercise energy expenditure was set to 320 kcal for women and 420 kcal for men, five days a week. The BIKE group was tasked with commuting to and from school/work by bicycle, MOD was asked to exercise at moderate intensity (50% of peak oxygen uptake (VO$_2$peak)-reserve) and VIG at vigorous intensity (70% of VO$_2$peak-reserve).

Data Collection

Doubly labelled water

Participants were administered a dose of DLW at each test period, with which their total EE was estimated for the following 14 days. Urine samples were taken every day over the period and in each sample the enrichment of ^{18}O and ^{2}H were measured using isotope ratio mass spectrometry to establish the turnover rates (Westerterp 2017). The difference between these decay rates is a function of carbon dioxide production, which in turn is a function of EE. Measurement error is correlated with the biological half-life of the isotopes, which varies with age and physical activity level. Based on the half-life, an observation interval of 2 weeks is recommended for adults.

Accelerometer

Participants were provided with a smartphone (Nexus 5X, LG electronics, Seoul, South Korea or Galaxy S5, Samsung Electronics Ltd., Suwon, South Korea) for the duration of the study, which they used as their primary phone. A custom-made mobile sensing application enabled ubiquitous data acquisition. The application was from a previous study (Stopczynski et al. 2014; Aharony et al. 2011) extended with measure of 3-axis acceleration sampled with a frequency of 50 Hz.

Heart rate

During the exercise sessions that were prescribed throughout the interventions, participants wore a Polar heart rate monitor (RC3 GPS, Polar Electro Oy, Kempele, Finland) that sampled HR with a frequency of 1 Hz. HR monitors were individually adjusted after 6 weeks and 3 months for changes in maximal HR, VO2peak and body weight.

Model

Total EE consists of three components; the basal metabolic rate, physical activity, and digestion (Westerterp 2009). If we assume an approximately constant EE due to digestion, a linear relationship between physical activity and EE is feasible. However, physical activity varies a lot from a sedentary person to an athlete, and therefore it is physiologically motivated to model this relationship on an individual basis. Even with people of comparable body composition, activity types and patterns have significant impact on EE. Due to a long observational window in which EE is measured using DLW, and because each DLW sample is expensive, we have a maximum of three observations per participant. With a low number of samples, the model is restricted in its number of parameters to avoid overfitting and it is thus not immediately possible to model EE with participant-level parameters. To overcome this limitation, we expand the model to use HR as a mediator between physical activity and EE, and participant-level parameters are estimated from independent data.

Individual calibration

We can approximate EE from HR with the linear model

$$\mathbf{y}^i = \mathrm{N}\left(\sum_{m=1}^{M}(\alpha_1^i + \alpha_2^i \mathbf{z}_m^i), \sigma_y^2\right), \qquad \sigma_y^2 \sim \mathrm{N}(0,1), \tag{1}$$

where N(· , ·) is the normal distribution, **y** represents EE, **z** represents HR, *i* indicates participant, and the sum is due to multiple HR samples for each EE. To estimate α, each individual was subjected to a VO_2peak test in which HR was measured using a HR monitor (RC3 GPS, Polar Electro Oy, Kempele, Finland) and EE through indirect calorimetry. We can also make the approximation of estimating HR from physical activity

$$\mathbf{z}_m^i \sim \mathrm{N}(\beta_1^i + \beta_2^i \mathbf{x}_m^i, \sigma_z^2), \qquad \sigma_z^2 \sim \mathrm{N}(0,1)_, \tag{2}$$

in which the physical activity is represented by **x** and is derived from the accelerometer and *β* denotes the parameters. Priors on the parameters are

$$\beta_1^i \sim \mathrm{N}(\beta_1, 0.1), \qquad \beta_1 \sim \mathrm{N}(0,1)$$

$$\beta_2^i \sim \mathrm{N}(\beta_2, 0.1), \qquad \beta_2 \sim \mathrm{N}(1,1)$$

where the relatively narrow variance ensures regularization across participants. We rarely have HR and smartphone accelerometer data simultaneously, as participants rarely carried their smartphones during exercise, but due to high sampling frequency, even short durations are sufficient to estimate this relationship. If physical activity data is partly unavailable it is modelled through

$$\mathbf{x} \sim \mathrm{N}(\boldsymbol{\theta}, \sigma_x^2),$$

where $\boldsymbol{\theta}$ are observed covariates and predictors (Kamronn, Hansen, and Larsen 2018). Figure 1 depicts the resulting probabilistic graphical model for each subject, with *N* EE observations each with *M* physical activity samples.

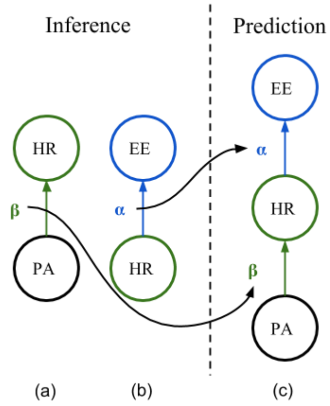


Figure 1. Illustration of generative model during inference and prediction. **a)** Prediction model of HR using Polar and smartphone data acquired during free-living exercises. **b)** Prediction model of EE using data acquired during VO2peak test. **c)** Combination of model parameters to create the full generative model from physical activity to EE. Data used for prediction in this model is from free-living during DLW periods.

Baseline calibration

To compare with existing methodology, a linear regression model is fitted to the physical activity and EE data directly, without using independent HR data as a mediator. The model definition is

$$\mathbf{y}^i \sim \mathrm{N}(\beta_1 + \beta_2 \mathbf{X}^i, \sigma_y^2), \qquad \sigma_y^2 \sim \mathrm{N}(0,1)_,$$

where **X** contains both physical activity and baseline characteristics; gender, age, weight, and fat mass. The prior on $\boldsymbol{\beta}_2$ is a multivariate normal distribution to model possible correlation between the covariates. The model is evaluated with mean absolute error (MAE) and leave-one-out cross-validation (LOO-CV), where participants in turn are left out when estimating parameters and then used for prediction. To inspect the added predictive value of each characteristic, five models are estimated, where each model adds another characteristic, i.e. the first has none and the last includes all.

Model fit

Parameters are estimated using Stan (Carpenter et al. 2017) with the No-U-Turn sampler and all Gelman-Rubin $\hat{R}$ statistics are ≤ 1.01, indicating convergence of sampling chains.

Data preprocessing

The full dataset collected during this study spans 6 months, and two subsets of the full dataset are considered here. One used for model inference and one used for EE prediction.

Inference

Data are extracted from periods defined by exercise sessions from the HR monitor, which was only worn during prescribed exercise sessions. A session is omitted if it has insufficient physical activity data (< 20 minutes), if the total physical activity level is too low, or if the correlation between HR and physical activity is low (r < 0.7). The latter two criteria exclude non-compliant sessions during which the smartphone was not carried during the whole session. Individual samples are excluded if the HR is too low (< 80 beats per minute), as the initial HR in the beginning of a training session is not representative of the true labour. By basing the model entirely on high-intensity data, we risk modelling local noise, and therefore resting HR measurements are included to represent zero physical activity. Both activity and HR data are summed to bins of 5 minutes and normalised to lie between 0 and 1.

Prediction

When predicting EE, the data subset is not defined by exercise sessions but by the 14-day period in which total EE is measured using DLW. The physical activity data is summed to bins of 1 hour, and if a bin is partly missing physical activity data, it is imputed from a hierarchical Bayesian model that was fitted from the full data set (Kamronn, Hansen, and Larsen 2018). Days with too much missing data (> 80%) are not included, and to ensure consistency across predictions, only the 5 days with least missing data are included for each period. If a DLW period does not contain 5 days with enough data, it is excluded.

Data Availability

The datasets generated during and/or analysed during the current study are not publicly available due to privacy regulation but are available from the corresponding author on reasonable request.

Results

A total of 47 participants were recruited for the smartphone study of whom 30 completed the 6-month intervention. Eight participants were excluded due to extensive missing smartphone data while exercising, and 11 participants were excluded due to missing smartphone data during the DLW periods. A two-sided t-test report no significant difference of baseline characteristics between the groups of included (Table 1) and excluded participants.

Baseline calibration

If the model is not calibrated to each individual, the correlation is very poor (r=0.05, P=0.83). The results of the baseline calibrated models are reported in Table 2. It is observed that including body weight decreases performance, while the best model includes all baseline characteristics.

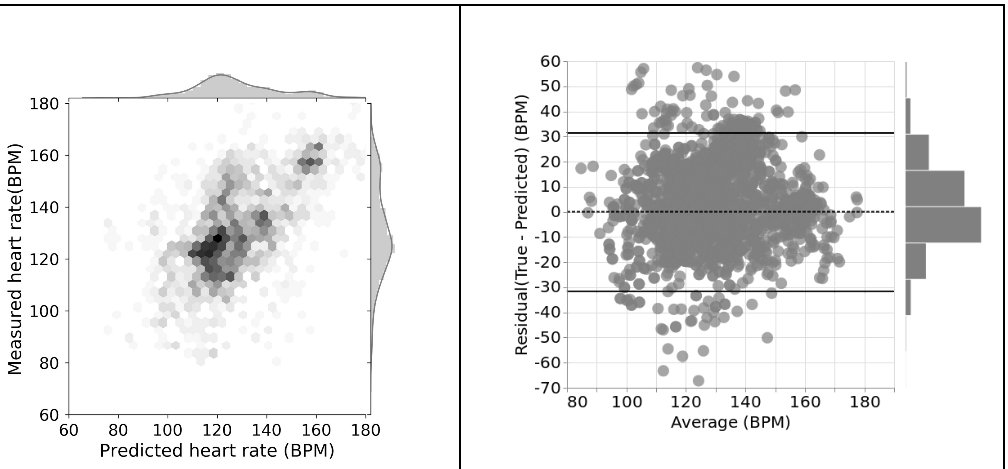
Characteristic	Activity	+Gender	+Age	+Weight	+Fat mass
LOO-CV MAE (kcal・day^{-1})	1051	878	778	903	729

Table 2. Leave-one-out cross-validation (LOO-CV) mean absolute errors (MAE) for the baseline calibrated model. Each column adds another baseline characteristic as predictor to the model.

Individual calibration

The two-stage model structure makes it possible to investigate the intermediate step of estimating HR from smartphone activity as specified in equation (2) and illustrated in Figure 1a. In Figure 2a, the predicted HR is visualised against the measured HR, and while the prediction is positive and strongly correlated (r=0.86, P < 0.001), we observe difficulties in modelling low HR, so only measurements above 80 beats per minute are included. The prediction is unbiased, as depicted in the Bland-Altman plot in Figure 2b, and the residuals are close to normally distributed.

Figure 3 depicts the comparison of the predicted EE against the measured EE for the individually calibrated model (depicted in Figure 1c). The correlation is strongly positive (r=0.84, P < 0.001) with a mean absolute error of 242 kcal・day^{-1}. The prediction is slightly negatively biased (-52 kcal・day^{-1} , ≈ 2.1% of average response) with a standard deviation of 297 kcal・day^{-1} of the residuals. The estimated model parameters are provided in Table 3 to provide a transparent overview of the contribution to the predicted EE from the different parts of the model. By comparing model parameters it easy to see the large differences in e.g. bias, which would not be possible to estimate with a simpler model.



(a) Hexagonal binned heatmap of predicted versus measured heart rate. The colour-scale of the joint plot (centre) illustrates how many samples fall within a particular bin (the cells) with darker colours indicating higher density. The marginal plots (top and right) show histograms overlaid with kernel density estimations of the marginal sample distributions of the predicted and measured heart rates, respectively.

(b) Bland-Altman plot showing a negligible bias and that most samples fall within 2 standard deviations.

Figure 2. Heart rate prediction

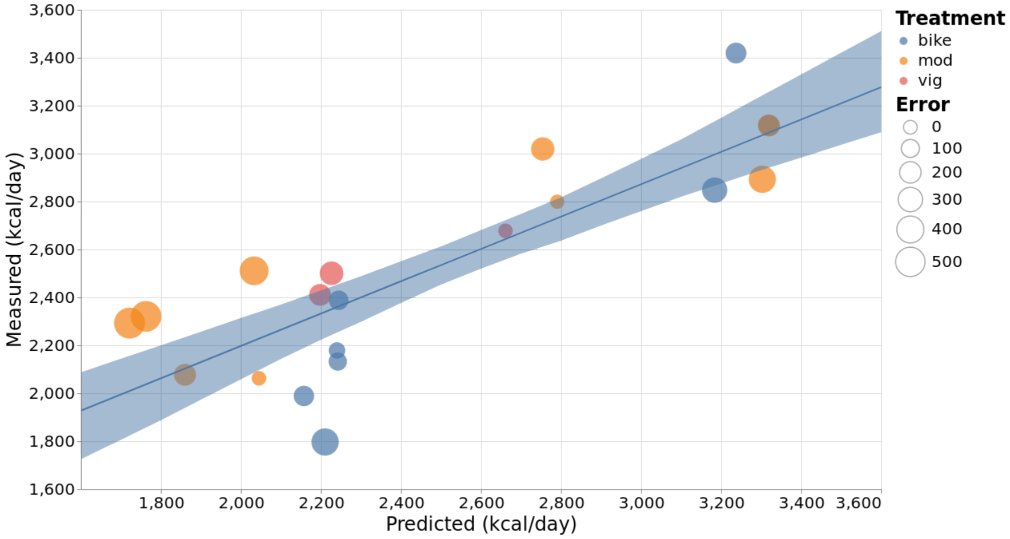


Figure 3. Predicted and measured energy expenditure. The colours represent intervention group and the size of the mark represents error of the estimate. The shaded area depicts the 95% confidence interval of the regression calculated by bootstrapping.

Participant ID	β_1	β_2	α_1	α_2
1	0.131552	0.420998	0.477086	0.491608
2	0.247892	0.603714	0.100037	0.47053
3	0.144971	0.350678	0.862894	0.764199
4	0.273536	0.540913	0.627859	0.693642
5	0.258299	0.433329	-0.073656	0.378969
6	0.411951	0.253482	0.087761	0.555243
7	0.507432	0.330656	0.022673	0.522429
8	0.082066	0.386173	0.120447	0.414975
9	0.459349	0.218257	0.732309	0.681283
10	0.133753	0.3504	0.573855	0.601198
11	0.280383	0.622213	-0.06886	0.2325

Table 3. Estimated model parameters ($\beta_{1,}\beta_{2,}\alpha_{1,}\alpha_2$) for 11 participants (ID 1-11)

Discussion

The principal finding of this study is that smartphones can be used to estimate EE in free-living individuals with limited constraints on how the smartphone is used or carried. The proposed model is robust to the various sources of noise introduced by the smartphone, but it requires calibration to the individual. The model is capable of adapting to individuals by combining data from two domains. Using data acquired from a one-time VO$_2$peak test conducted with accurate equipment, the model compensates for individual-level physiology and body composition. Using data acquired from wearable devices under free-living conditions, it models variation in inter-person movement patterns and variation between naturally occurring activities. All these sources of variation in EE and physical activity must be accounted for.

Our finding that a single uncalibrated model is too simple to estimate EE across all individuals reaffirms previous studies (Sardinha and Júdice 2017; Jacobi et al. 2007). Body weight is proportional to PAEE in weight-bearing activities (Schoeller and Jefford 2002), which means that an obese individual spends more energy being physically active than a lean individual (Prentice et al. 1986). Body weight is therefore often used to calibrate regression models of EE. However, the amount of fat-free mass explains more than 80% of the variance in resting EE (Müller et al. 2018), a major component of the total EE, which is why normalising by body weight is not appropriate when applied to individuals of varying body composition. Tissue, organs, and genetics are also important determinants in resting EE, adding to the complexity (Müller et al. 2018). The results in Table 2 show that normalizing by gender and age improves the estimate but including body weight actually worsens it. With the error measure being calculated on a leave-one-subject-out basis, we can conclude that including body weight leads to overfitting. By including fat mass, the prediction improves significantly, which is due to the model's ability to adjust for both fat and fat-free mass. Using accurate data from the laboratory (VO$_2$peak test) when estimating EE from HR, the prediction is three times better (242 vs. 729 kcal ・ day^{-1}). The estimated model parameters in Table 3 stress the need for calibration. A large variation is observed in e.g. the bias term in predicting EE from HR (α_1=[-0.073656; 0.862894]), suggesting a large difference in resting EE and similarly for the slope (α_2=[0.2325; 0.764199]), suggesting variation in the amount of energy expended at a certain HR. It should be stressed that the reported results include resting EE, which is often the largest component of total EE, and therefore not directly modelled by the data acquired by the smartphone. This distinction is important when evaluating the predictive performance.

Models that do well in a laboratory setting generally do much worse in a free-living setting due to the nature of the activities performed in each. A linear model estimated on data from a small set of activities on a treadmill is not able to generalise to the set of activities performed in active daily living and will substantially underestimate the PAEE (Nilsson et al. 2008; de Graauw et al. 2010). In this study, that limitation is partially circumvented by using data obtained under free-living conditions when estimating HR from physical activity with small (Polar chest strap), or even pervasive (Shcherbina et al. 2017), equipment.

The high prediction accuracy found in many studies of PAEE that do not calibrate models to individuals probably overfit severely due to small homogenic cohorts in simplified scenarios (Jeran, Steinbrecher, and Pischon 2016). Human physiology is too complex to model with simple means like linear models on high variance signals, or even with complex models (Crouter, Bassett, and Freedson 2009) if the data foundation is not adequate, and current research indicates that no study to date is adequately sized (Sardinha and Júdice 2017). To create this foundation, a ubiquitous device is needed to capture the true behavioural aspects of physical activity and it needs to be available in all corners of the high dimensional space that encompasses human variability. The smartphone is currently the only device that fits the bill.

Because the smartphone is able to measure modalities besides acceleration, such as location through GPS, local proximity through WiFi and Bluetooth (Stopczynski et al. 2014), and mood through the voice (Hargreaves, Starkweather, and Blacker 1965), the smartphone has much greater potential than stand-alone devices. From location, the geospatial movement patterns can be used to understand human behaviour in a visual manner by illustrating the spatiotemporal trajectories (Maddison et al. 2017) or by inferring the context of different locations such as home and work (Kamronn, Hansen, and Larsen 2018). This information can be leveraged to increase and generalise prediction accuracy and to decompose predicted PAEE into a more nuanced picture. Being able to model the behavioural mechanisms of PAEE and being aware of the environment, it becomes possible to take a decision-theoretical approach to just-in-time interventions that adapts to the context of an individual.

Limitations

The DLW method is considered the gold standard for measuring EE under daily living conditions (Westerterp 2017). Samples obtained in this way are, however, rather coarse compared to alternative methods that sample with intervals on the second or minute scale, so even though these devices are capable of higher resolution, we can only make claims to the validity of a long sample interval.

Furthermore, smartphones are unreliable in terms of sample consistency, partly due to adaptive sampling schemes that optimise when to sample to reduce power usage and partly due to an unreliable software platform. As the smartphone is not mounted on the participant, they may not always carry it on their body, resulting in non-compliant missing data. These issues result in lower physical activity estimates and we are only able to model away some of the missingness. It is acknowledged that the reported results are accompanied by uncertainty due to the high level of missingness and the low number of participants.

Due to both the physiological and behavioural variation between individuals, the model requires calibration to the individual, which is a major limitation in applicability outside clinical settings.

Future work

Calibration of model parameters to individual participants is only feasible in small to moderately sized studies, while the method of monitoring through smartphones has population-sized potential (Stopczynski et al. 2014; Althoff et al. 2017). More flexible models using context such as activity-type and location can be used to increase the precision (Altini et al. 2015), but as the relationship between physical activity level and EE varies across not only types of activities, but more importantly, across individuals, a more ambitious approach is warranted. The relationship between human physiology and behaviour measured by the smartphone and EE needs to be explored for large heterogeneous groups to build a model that is able to generalize across populations.

Conclusion

Total EE can be predicted from smartphone usage with the right approach and the right equipment. The proposed model can adapt to individuals using a combination of high-accuracy clinical and high-variation free-living data. It is shown that an uncalibrated model, and arguably a complex model without the necessary information available, is not able to accurately predict EE. Developing a model that is able to generalise to populations is a monumental task but with increasing adoption of ubiquitous sensors, increasingly feasible.

Acknowledgements

This project was funded in part by TrygFonden.

References

Aharony, Nadav, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. "Social fMRI: Investigating and Shaping Social Mechanisms in the Real World." *Pervasive and Mobile Computing* 7 (6): 643–59.

Althoff, Tim, Rok Sosič, Jennifer L. Hicks, Abby C. King, Scott L. Delp, and Jure Leskovec. 2017. "Large-Scale Physical Activity Data Reveal Worldwide Activity Inequality." *Nature* 547 (7663): 336–39.

Altini, Marco, Pierluigi Casale, Julien F. Penders, and Oliver Amft. 2015. "Personalization of Energy Expenditure Estimation in Free Living Using Topic Models." *IEEE Journal of Biomedical and Health Informatics* 19 (5): 1577–86.

Badimon, Lina, Raffaele Bugiardini, Edina Cenko, Judit Cubedo, Maria Dorobantu, Dirk J. Duncker, Ramón Estruch, et al. 2017. "Position Paper of the European Society of Cardiology-Working Group of Coronary Pathophysiology and Microcirculation: Obesity and Heart Disease." *European Heart Journal* 38 (25): 1951–58.

Bassett, David R., Alex Rowlands, and Stewart G. Trost. 2012. "Calibration and Validation of Wearable Monitors." *Medicine and Science in Sports and Exercise* 44 (7): 32–38.

Beltrame, T., R. Amelard, A. Wong, and R. L. Hughson. 2017. "Prediction of Oxygen Uptake Dynamics by Machine Learning Analysis of Wearable Sensors during Activities of Daily Living." *Scientific Reports* 7 (April): 45738.

Bonomi, A. G., G. Plasqui, A. H. C. Goris, and K. R. Westerterp. 2009. "Improving Assessment of Daily Energy Expenditure by Identifying Types of Physical Activity with a Single Accelerometer." *Journal of Applied Physiology* 107 (3). Am Physiological Soc: 655–61.

Bonomi, A. G., and K. R. Westerterp. 2012. "Advances in Physical Activity Monitoring and Lifestyle Interventions in Obesity: A Review." *International Journal of Obesity* 36 (2). nature.com: 167–77.

Bort-Roig, Judit, Nicholas D. Gilson, Anna Puig-Ribera, Ruth S. Contreras, and Stewart G. Trost. 2014. "Measuring and Influencing Physical Activity with Smartphone Technology: A Systematic Review." *Sports Medicine* 44 (5): 671–86.

Bouten, C. V., K. R. Westerterp, M. Verduin, and J. D. Janssen. 1994. "Assessment of Energy Expenditure for Physical Activity Using a Triaxial Accelerometer." *Medicine and Science in Sports and Exercise* 26 (12). alexandria.tue.nl: 1516–23.

Brage, Søren, Kate Westgate, Paul W. Franks, Oliver Stegle, Antony Wright, Ulf Ekelund, and Nicholas J. Wareham. 2015. "Estimation of Free-Living Energy Expenditure by Heart Rate and Movement Sensing: A Doubly-Labelled Water Study." *PloS One* 10 (9): e0137206.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1). uvm.edu: 1–32.

Chapman, Justin J., James A. Roberts, Vinh T. Nguyen, and Michael Breakspear. 2017. "Quantification of Free-Living Activity Patterns Using Accelerometry in Adults with Mental Illness." *Scientific Reports* 7 (March). nature.com: 43174.

Chin, Sang Ouk, Changwon Keum, Junghoon Woo, Jehwan Park, Hyung Jin Choi, Jeong-Taek Woo, and Sang Youl Rhee. 2016. "Successful Weight Reduction and Maintenance by Using a Smartphone Application in Those with Overweight and Obesity." *Scientific Reports* 6 (November): 34563.

Chowdhury, Enhad A., Max J. Western, Thomas E. Nightingale, Oliver J. Peacock, and Dylan Thompson. 2017. "Assessment of Laboratory and Daily Energy Expenditure Estimates from Consumer Multi-Sensor Physical Activity Monitors." *PloS One* 12 (2). journals.plos.org: e0171720.

Crouter, S., D. Bassett, and P. Freedson. 2009. "An Artificial Neural Network to Estimate Physical Activity Energy Expenditure and Identify Physical Activity Type from an Accelerometer." *Journal of Applied*. Am Physiological Soc. http://www.physiology.org/doi/abs/10.1152/japplphysiol.00465.2009.

Direito, Artur, Eliana Carraça, Jonathan Rawstorn, Robyn Whittaker, and Ralph Maddison. 2017. “mHealth Technologies to Influence Physical Activity and Sedentary Behaviors: Behavior Change Techniques, Systematic Review and Meta-Analysis of Randomized Controlled Trials.” *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine* 51 (2): 226–39.

Faurholt-Jepsen, Maria, Mads Frost, Maj Vinberg, Ellen Margrethe Christensen, Jakob E. Bardram, and Lars Vedel Kessing. 2014. “Smartphone Data as Objective Measures of Bipolar Disorder Symptoms.” *Psychiatry Research* 217 (1-2): 124–27.

Ferguson, Ty, Alex V. Rowlands, Tim Olds, and Carol Maher. 2015. “The Validity of Consumer-Level, Activity Monitors in Healthy Adults Worn in Free-Living Conditions: A Cross-Sectional Study.” *The International Journal of Behavioral Nutrition and Physical Activity* 12 (March). ijbnpa.biomedcentral.com: 42.

Finkelstein, Eric A., Benjamin A. Haaland, Marcel Bilger, Aarti Sahasranaman, Robert A. Sloan, Ei Ei Khaing Nang, and Kelly R. Evenson. 2016. “Effectiveness of Activity Trackers with and without Incentives to Increase Physical Activity (TRIPPA): A Randomised Controlled Trial.” *The Lancet. Diabetes & Endocrinology* 4 (12). Elsevier: 983–95.

Graauw, Suzanne M. de, Janke F. de Groot, Marco van Brussel, Marjolein F. Streur, and Tim Takken. 2010. “Review of Prediction Models to Estimate Activity-Related Energy Expenditure in Children and Adolescents.” *International Journal of Pediatrics* 2010 (June). downloads.hindawi.com: 489304.

Hargreaves, W. A., J. A. Starkweather, and K. H. Blacker. 1965. “VOICE QUALITY IN DEPRESSION.” *Journal of Abnormal Psychology* 70 (3): 218–20.

Hekler, Eric B., Matthew P. Buman, Lauren Grieco, Mary Rosenberger, Sandra J. Winter, William Haskell, and Abby C. King. 2015. “Validation of Physical Activity Tracking via Android Smartphones Compared to ActiGraph Accelerometer: Laboratory-Based and Free-Living Validation Studies.” *JMIR mHealth and uHealth* 3: e36.

Ingraham, K. A., D. P. Ferris, and C. David Remy. 2017. “Using Wearable Physiological Sensors to Predict Energy Expenditure.” In *2017 International Conference on Rehabilitation Robotics (ICORR)*, 340–45.

Jacobi, David, Anne-Elisabeth Perrin, Natacha Grosman, Marie-France Doré, Sylvie Normand, Jean-Michel Oppert, and Chantal Simon. 2007. “Physical Activity-Related Energy Expenditure With the RT3 and TriTrac Accelerometers in Overweight Adults*.” *Obesity* 15 (4). Wiley Online Library: 950–56.

Jeran, S., A. Steinbrecher, and T. Pischon. 2016. “Prediction of Activity-Related Energy Expenditure Using Accelerometer-Derived Physical Activity under Free-Living Conditions: A Systematic Review.” *International Journal of Obesity* 40 (8). nature.com: 1187–97.

Kamronn, Simon, Lars Kai Hansen, and Jakob Eg Larsen. 2018. “Smartphone Pervasive Sensing of Physical Activity of Overweight Adults in a Long-Running Randomized Controlled Trial.” *Journal of Medical Internet Research*. https://doi.org/10.2196/preprints.10745.

Maddison, Ralph, Luke Gemming, Javier Monedero, Linda Bolger, Sarahjane Belton, Johann Issartel, Samantha Marsh, et al. 2017. “Quantifying Human Movement Using the Movn Smartphone App: Validation and Field Study.” *JMIR mHealth and uHealth* 5 (8): e122.

Maher, Carol A., Emily Mire, Deirdre M. Harrington, Amanda E. Staiano, and Peter T. Katzmarzyk. 2013. “The Independent and Combined Associations of Physical Activity and Sedentary Behavior with Obesity in Adults: NHANES 2003-06: Adult Activity Patterns and Obesity Risk.” *Obesity* 21 (12). Wiley Online Library: E730–37.

Montoye, H. J., R. Washburn, S. Servais, A. Ertl, J. G. Webster, and F. J. Nagle. 1983. “Estimation of Energy Expenditure by a Portable Accelerometer.” *Medicine and Science in Sports and Exercise* 15 (5). europepmc.org: 403–7.

Müller, Manfred J., Corinna Geisler, Mark Hübers, Maryam Pourhassan, Wiebke Braun, and Anja Bosy-Westphal. 2018. “Normalizing Resting Energy Expenditure across the Life Course in Humans: Challenges and Hopes.” *European Journal of Clinical Nutrition* 72 (5): 628–37.

Murakami, Haruka, Ryoko Kawakami, Satoshi Nakae, Yoshio Nakata, Kazuko Ishikawa-Takata, Shigeho Tanaka, and Motohiko Miyachi. 2016. “Accuracy of Wearable Devices for Estimating Total Energy Expenditure: Comparison With Metabolic Chamber and Doubly Labeled Water Method.” *JAMA Internal*

Medicine, March. https://doi.org/10.1001/jamainternmed.2016.0152.

Nilsson, A., S. Brage, C. Riddoch, S. A. Anderssen, L. B. Sardinha, N. Wedderkopp, L. B. Andersen, and U. Ekelund. 2008. “Comparison of Equations for Predicting Energy Expenditure from Accelerometer Counts in Children.” *Scandinavian Journal of Medicine & Science in Sports* 18 (5). Wiley Online Library: 643–50.

Plasqui, Guy, and Klaas R. Westerterp. 2007. “Physical Activity Assessment With Accelerometers: An Evaluation Against Doubly Labeled Water**.” *Obesity* 15 (10): 2371–79.

Prentice, A. M., A. E. Black, W. A. Coward, H. L. Davies, G. R. Goldberg, P. R. Murgatroyd, J. Ashford, M. Sawyer, and R. G. Whitehead. 1986. “High Levels of Energy Expenditure in Obese Women.” *British Medical Journal* 292 (6526). bmj.com: 983–87.

Rosenkilde, Mads, Martin Bæk Petersen, Anne Sofie Gram, Jonas Salling Quist, Jonas Winther, Simon Due Kamronn, Desirée Hornbæk Milling, Jakob Eg Larsen, Astrid Pernille Jespersen, and Bente Stallknecht. 2017. “The GO-ACTIWE Randomized Controlled Trial - An Interdisciplinary Study Designed to Investigate the Health Effects of Active Commuting and Leisure Time Physical Activity.” *Contemporary Clinical Trials* 53 (February): 122–29.

Sardinha, L. B., and P. B. Júdice. 2017. “Usefulness of Motion Sensors to Estimate Energy Expenditure in Children and Adults: A Narrative Review of Studies Using DLW.” *European Journal of Clinical Nutrition* 71 (3): 331–39.

Schoeller, D. A., and G. Jefford. 2002. “Determinants of the Energy Costs of Light Activities: Inferences for Interpreting Doubly Labeled Water Data.” *International Journal of Obesity and Related Metabolic Disorders: Journal of the International Association for the Study of Obesity* 26 (1). nature.com: 97–101.

Shcherbina, Anna, C. Mikael Mattsson, Daryl Waggott, Heidi Salisbury, Jeffrey W. Christle, Trevor Hastie, Matthew T. Wheeler, and Euan A. Ashley. 2017. “Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort.” *Journal of Personalized Medicine* 7 (2). https://doi.org/10.3390/jpm7020003.

Stopczynski, Arkadiusz, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. 2014. “Measuring Large-Scale Social Networks with High Resolution.” *PloS One* 9 (4): e95978.

Sullivan, Alycia N., and Margie E. Lachman. 2016. “Behavior Change with Fitness Technology in Sedentary Adults: A Review of the Evidence for Increasing Physical Activity.” *Frontiers in Public Health* 4. frontiersin.org: 289.

Verheggen, R. J. H. M., M. F. H. Maessen, D. J. Green, A. R. M. M. Hermus, M. T. E. Hopman, and D. H. T. Thijssen. 2016. “A Systematic Review and Meta-Analysis on the Effects of Exercise Training versus Hypocaloric Diet: Distinct Effects on Body Weight and Visceral Adipose Tissue: Effects of Exercise versus Diet on Visceral Fat.” *Obesity Reviews: An Official Journal of the International Association for the Study of Obesity* 17 (8): 664–90.

Westerterp, Klaas R. 2009. “Assessment of Physical Activity: A Critical Appraisal.” *European Journal of Applied Physiology* 105 (6). Springer: 823–28.

———. 2017. “Doubly Labelled Water Assessment of Energy Expenditure: Principle, Practice, and Promise.” *European Journal of Applied Physiology* 117 (7): 1277–85.

APPENDIX C
A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning

Marco Fraccaro*, **Simon Kamronn***, Ulrich Paquet and Ole Winther. (2017). *A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning.* Advances in Neural Information Processing Systems 30, NIPS 2017.

**Equal contribution*

A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning

Marco Fraccaro[†*] **Simon Kamronn** [†*] **Ulrich Paquet**[‡] **Ole Winther**[†]
[†] Technical University of Denmark
[‡] DeepMind

Abstract

This paper takes a step towards temporal reasoning in a dynamically changing video, not in the pixel space that constitutes its frames, but in a latent space that describes the non-linear dynamics of the objects in its world. We introduce the Kalman variational auto-encoder, a framework for unsupervised learning of sequential data that disentangles two latent representations: an object's representation, coming from a recognition model, and a latent state describing its dynamics. As a result, the evolution of the world can be imagined and missing data imputed, both without the need to generate high dimensional frames at each time step. The model is trained end-to-end on videos of a variety of simulated physical systems, and outperforms competing methods in generative and missing data imputation tasks.

1 Introduction

From the earliest stages of childhood, humans learn to represent high-dimensional sensory input to make temporal predictions. From the visual image of a moving tennis ball, we can imagine its trajectory, and prepare ourselves in advance to catch it. Although the act of recognising the tennis ball is seemingly independent of our intuition of Newtonian dynamics [29], very little of this assumption has yet been captured in the end-to-end models that presently mark the path towards artificial general intelligence. Instead of basing inference on any abstract grasp of dynamics that is learned from experience, current successes are autoregressive: to imagine the tennis ball's trajectory, one forward-generates a frame-by-frame rendering of the full sensory input [4, 6, 21, 22, 27, 28].

To disentangle two latent representations, an object's, and that of its dynamics, this paper introduces *Kalman variational auto-encoders (KVAEs)*, a model that separates an intuition of dynamics from an object recognition network (section 3). At each time step t, a variational auto-encoder [16, 23] compresses high-dimensional visual stimuli $\mathbf{x}_t$ into latent encodings $\mathbf{a}_t$. The temporal dynamics in the learned $\mathbf{a}_t$-manifold are modelled with a linear Gaussian state space model that is adapted to handle complex dynamics (despite the linear relations among its states $\mathbf{z}_t$). The parameters of the state space model are adapted at each time step, and non-linearly depend on past $\mathbf{a}_t$'s via a recurrent neural network. Exact posterior inference for the linear Gaussian state space model can be performed with the Kalman filtering and smoothing algorithms, and is used for imputing missing data, for instance when we imagine the trajectory of a bouncing ball after observing it in initial and final video frames (section 4). The separation between recognition and dynamics model allows for missing data imputation to be done via a combination of the latent states $\mathbf{z}_t$ of the model and its encodings $\mathbf{a}_t$ only, without having to forward-sample high-dimensional images $\mathbf{x}_t$ in an autoregressive way. KVAEs are tested on videos of a variety of simulated physical systems in section 5: from raw visual stimuli, it "end-to-end" learns the interplay between the recognition and dynamics components. As KVAEs can do smoothing, they outperform an array of methods in generative and missing data imputation tasks (section 5).

[*] Equal contribution.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Background

Linear Gaussian state space models. Linear Gaussian state space models (LGSSMs) are widely used to model sequences of vectors $\mathbf{a} = \mathbf{a}_{1:T} = [\mathbf{a}_1, .., \mathbf{a}_T]$. LGSSMs model temporal correlations through a first-order Markov process on latent states $\mathbf{z} = [\mathbf{z}_1, .., \mathbf{z}_T]$, which are potentially further controlled with external inputs $\mathbf{u} = [\mathbf{u}_1, .., \mathbf{u}_T]$, through the Gaussian distributions

$$p_{\gamma_t}(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{u}_t) = \mathcal{N}(\mathbf{z}_t; \mathbf{A}_t\mathbf{z}_{t-1} + \mathbf{B}_t\mathbf{u}_t, \mathbf{Q}) , \qquad p_{\gamma_t}(\mathbf{a}_t|\mathbf{z}_t) = \mathcal{N}(\mathbf{a}_t; \mathbf{C}_t\mathbf{z}_t, \mathbf{R}) . \tag{1}$$

Matrices $\gamma_t = [\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t]$ are the state transition, control and emission matrices at time t. $\mathbf{Q}$ and $\mathbf{R}$ are the covariance matrices of the process and measurement noise respectively. With a starting state $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{z}_1; \mathbf{0}, \mathbf{\Sigma})$, the joint probability distribution of the LGSSM is given by

$$p_\gamma(\mathbf{a}, \mathbf{z}|\mathbf{u}) = p_\gamma(\mathbf{a}|\mathbf{z})\, p_\gamma(\mathbf{z}|\mathbf{u}) = \textstyle\prod_{t=1}^T p_{\gamma_t}(\mathbf{a}_t|\mathbf{z}_t) \cdot p(\mathbf{z}_1) \prod_{t=2}^T p_{\gamma_t}(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{u}_t) , \tag{2}$$

where $\gamma = [\gamma_1, .., \gamma_T]$. LGSSMs have very appealing properties that we wish to exploit: the filtered and smoothed posteriors $p(\mathbf{z}_t|\mathbf{a}_{1:t}, \mathbf{u}_{1:t})$ and $p(\mathbf{z}_t|\mathbf{a}, \mathbf{u})$ can be computed exactly with the classical Kalman filter and smoother algorithms, and provide a natural way to handle missing data.

Variational auto-encoders. A variational auto-encoder (VAE) [16, 23] defines a deep generative model $p_\theta(\mathbf{x}_t, \mathbf{a}_t) = p_\theta(\mathbf{x}_t|\mathbf{a}_t)p(\mathbf{a}_t)$ for data $\mathbf{x}_t$ by introducing a latent encoding $\mathbf{a}_t$. Given a likelihood $p_\theta(\mathbf{x}_t|\mathbf{a}_t)$ and a typically Gaussian prior $p(\mathbf{a}_t)$, the posterior $p_\theta(\mathbf{a}_t|\mathbf{x}_t)$ represents a stochastic map from $\mathbf{x}_t$ to $\mathbf{a}_t$'s manifold. As this posterior is commonly analytically intractable, VAEs approximate it with a variational distribution $q_\phi(\mathbf{a}_t|\mathbf{x}_t)$ that is parameterized by ϕ. The approximation q_ϕ is commonly called the recognition, encoding, or inference network.

3 Kalman Variational Auto-Encoders

The useful information that describes the movement and interplay of objects in a video typically lies in a manifold that has a smaller dimension than the number of pixels in each frame. In a video of a ball bouncing in a box, like Atari's game Pong, one could define a one-to-one mapping from each of the high-dimensional frames $\mathbf{x} = [\mathbf{x}_1, .., \mathbf{x}_T]$ into a two-dimensional latent space that represents the position of the ball on the screen. If the position was known for consecutive time steps, for a set of videos, we could learn the temporal dynamics that govern the environment. From a few new positions one might then infer where the ball will be on the screen in the future, and then imagine the environment with the ball in that position.

The *Kalman variational auto-encoder* (KVAE) is based on the notion described above. To disentangle recognition and spatial representation, a sensory input $\mathbf{x}_t$ is mapped to $\mathbf{a}_t$ (VAE), a variable on a low-dimensional manifold that encodes an object's position and other visual properties. In turn, $\mathbf{a}_t$ is used as a pseudo-observation for the dynamics model (LGSSM). $\mathbf{x}_t$ represents a frame of a video[2] $\mathbf{x} = [\mathbf{x}_1, .., \mathbf{x}_T]$ of length T. Each frame is encoded into a point $\mathbf{a}_t$ on a low-dimensional manifold, so that the KVAE contains T separate VAEs that share the same decoder $p_\theta(\mathbf{x}_t|\mathbf{a}_t)$ and encoder $q_\phi(\mathbf{a}_t|\mathbf{x}_t)$, and depend on each other through a time-dependent prior over $\mathbf{a} = [\mathbf{a}_1, .., \mathbf{a}_T]$. This is illustrated in figure 1.

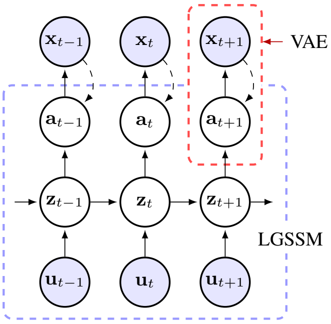


Figure 1: A KVAE is formed by stacking a LGSSM (dashed blue), and a VAE (dashed red). Shaded nodes denote observed variables. Solid arrows represent the generative model (with parameters θ) while dashed arrows represent the VAE inference network (with parameters ϕ).

3.1 Generative model

We assume that $\mathbf{a}$ acts as a latent representation of the whole video, so that the generative model of a sequence factorizes as $p_\theta(\mathbf{x}|\mathbf{a}) = \prod_{t=1}^T p_\theta(\mathbf{x}_t|\mathbf{a}_t)$. In this paper $p_\theta(\mathbf{x}_t|\mathbf{a}_t)$ is a deep neural network parameterized by θ,

[2] While our main focus in this paper are videos, the same ideas could be applied more in general to any sequence of high dimensional data.

that emits either a factorized Gaussian or Bernoulli probability vector depending on the data type of $\mathbf{x}_t$. We model $\mathbf{a}$ with a LGSSM, and following (2), its prior distribution is

$$p_\gamma(\mathbf{a}|\mathbf{u}) = \int p_\gamma(\mathbf{a}|\mathbf{z})\, p_\gamma(\mathbf{z}|\mathbf{u})\, \mathrm{d}\mathbf{z}\ , \tag{3}$$

so that the joint density for the KVAE factorizes as $p(\mathbf{x}, \mathbf{a}, \mathbf{z}|\mathbf{u}) = p_\theta(\mathbf{x}|\mathbf{a})\, p_\gamma(\mathbf{a}|\mathbf{z})\, p_\gamma(\mathbf{z}|\mathbf{u})$. A LGSSM forms a convenient back-bone to a model, as the filtered and smoothed distributions $p_\gamma(\mathbf{z}_t|\mathbf{a}_{1:t}, \mathbf{u}_{1:t})$ and $p_\gamma(\mathbf{z}_t|\mathbf{a}, \mathbf{u})$ can be obtained exactly. Temporal reasoning can be done in the latent space of $\mathbf{z}_t$'s and via the latent encodings $\mathbf{a}$, and we can do long-term predictions without having to auto-regressively generate high-dimensional images $\mathbf{x}_t$. Given a few frames, and hence their encodings, one could "remain in latent space" and use the smoothed distributions to impute missing frames. Another advantage of using $\mathbf{a}$ to separate the dynamics model from $\mathbf{x}$ can be seen by considering the emission matrix $\mathbf{C}_t$. Inference in the LGSSM requires matrix inverses, and using it as a model for the prior dynamics of $\mathbf{a}_t$ allows the size of $\mathbf{C}_t$ to remain small, and not scale with the number of pixels in $\mathbf{x}_t$. While the LGSSM's process and measurement noise in (1) are typically formulated with full covariance matrices [24], we will consider them as isotropic in a KVAE, as $\mathbf{a}_t$ act as a prior in a generative model that includes these extra degrees of freedom.

What happens when a ball bounces against a wall, and the dynamics on $\mathbf{a}_t$ are not linear any more? Can we still retain a LGSSM backbone? We will incorporate nonlinearities into the LGSSM by regulating γ_t from *outside* the exact forward-backward inference chain. We revisit this central idea at length in section 3.3.

3.2 Learning and inference for the KVAE

We learn θ and γ from a set of example sequences $\{\mathbf{x}^{(n)}\}$ by maximizing the sum of their respective log likelihoods $\mathcal{L} = \sum_n \log p_{\theta\gamma}(\mathbf{x}^{(n)}|\mathbf{u}^{(n)})$ as a function of θ and γ. For simplicity in the exposition we restrict our discussion below to one sequence, and omit the sequence index n. The log likelihood or evidence is an intractable average over all plausible settings of $\mathbf{a}$ and $\mathbf{z}$, and exists as the denominator in Bayes' theorem when inferring the posterior $p(\mathbf{a}, \mathbf{z}|\mathbf{x}, \mathbf{u})$. A more tractable approach to both learning and inference is to introduce a variational distribution $q(\mathbf{a}, \mathbf{z}|\mathbf{x}, \mathbf{u})$ that approximates the posterior. The evidence lower bound (ELBO) $\mathcal{F}$ is

$$\log p(\mathbf{x}|\mathbf{u}) = \log \int p(\mathbf{x}, \mathbf{a}, \mathbf{z}|\mathbf{u}) \geq \mathbb{E}_{q(\mathbf{a},\mathbf{z}|\mathbf{x},\mathbf{u})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{a}) p_\gamma(\mathbf{a}|\mathbf{z}) p_\gamma(\mathbf{z}|\mathbf{u})}{q(\mathbf{a}, \mathbf{z}|\mathbf{x}, \mathbf{u})}\right] = \mathcal{F}(\theta, \gamma, \phi)\ , \tag{4}$$

and a sum of $\mathcal{F}$'s is maximized instead of a sum of log likelihoods. The variational distribution q depends on ϕ, but for the bound to be tight we should specify q to be equal to the posterior distribution that only depends on θ and γ. Towards this aim we structure q so that it incorporates the exact conditional posterior $p_\gamma(\mathbf{z}|\mathbf{a}, \mathbf{u})$, that we obtain with Kalman smoothing, as a factor of γ:

$$q(\mathbf{a}, \mathbf{z}|\mathbf{x}, \mathbf{u}) = q_\phi(\mathbf{a}|\mathbf{x})\, p_\gamma(\mathbf{z}|\mathbf{a}, \mathbf{u}) = \textstyle\prod_{t=1}^{T} q_\phi(\mathbf{a}_t|\mathbf{x}_t)\, p_\gamma(\mathbf{z}|\mathbf{a}, \mathbf{u})\ . \tag{5}$$

The benefit of the LGSSM backbone is now apparent. We use a "recognition model" to encode each $\mathbf{x}_t$ using a non-linear function, after which exact smoothing is possible. In this paper $q_\phi(\mathbf{a}_t|\mathbf{x}_t)$ is a deep neural network that maps $\mathbf{x}_t$ to the mean and the diagonal covariance of a Gaussian distribution. As explained in section 4, this factorization allows us to deal with missing data in a principled way. Using (5), the ELBO in (4) becomes

$$\mathcal{F}(\theta, \gamma, \phi) = \mathbb{E}_{q_\phi(\mathbf{a}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{a})}{q_\phi(\mathbf{a}|\mathbf{x})} + \mathbb{E}_{p_\gamma(\mathbf{z}|\mathbf{a},\mathbf{u})}\left[\log \frac{p_\gamma(\mathbf{a}|\mathbf{z}) p_\gamma(\mathbf{z}|\mathbf{u})}{p_\gamma(\mathbf{z}|\mathbf{a}, \mathbf{u})}\right]\right]\ . \tag{6}$$

The lower bound in (6) can be estimated using Monte Carlo integration with samples $\{\widetilde{\mathbf{a}}^{(i)}, \widetilde{\mathbf{z}}^{(i)}\}_{i=1}^{I}$ drawn from q,

$$\hat{\mathcal{F}}(\theta, \gamma, \phi) = \frac{1}{I} \sum_i \log p_\theta(\mathbf{x}|\widetilde{\mathbf{a}}^{(i)}) + \log p_\gamma(\widetilde{\mathbf{a}}^{(i)}, \widetilde{\mathbf{z}}^{(i)}|\mathbf{u}) - \log q_\phi(\widetilde{\mathbf{a}}^{(i)}|\mathbf{x}) - \log p_\gamma(\widetilde{\mathbf{z}}^{(i)}|\widetilde{\mathbf{a}}^{(i)}, \mathbf{u})\ . \tag{7}$$

Note that the ratio $p_\gamma(\widetilde{\mathbf{a}}^{(i)}, \widetilde{\mathbf{z}}^{(i)}|\mathbf{u})/p_\gamma(\widetilde{\mathbf{z}}^{(i)}|\widetilde{\mathbf{a}}^{(i)}, \mathbf{u})$ in (7) gives $p_\gamma(\widetilde{\mathbf{a}}^{(i)}|\mathbf{u})$, but the formulation with $\{\widetilde{\mathbf{z}}^{(i)}\}$ allows stochastic gradients on γ to also be computed. A sample from q can be obtained by first sampling $\widetilde{\mathbf{a}} \sim q_\phi(\mathbf{a}|\mathbf{x})$, and using $\widetilde{\mathbf{a}}$ as an observation for the LGSSM. The posterior $p_\gamma(\mathbf{z}|\widetilde{\mathbf{a}}, \mathbf{u})$ can be tractably obtained with a Kalman smoother, and a sample $\widetilde{\mathbf{z}} \sim p_\gamma(\mathbf{z}|\widetilde{\mathbf{a}}, \mathbf{u})$ obtained from it. Parameter learning is done by *jointly* updating θ, ϕ, and γ by maximising the ELBO on $\mathcal{L}$, which decomposes as a sum of ELBOs in (6), using stochastic gradient ascent and a single sample to approximate the intractable expectations.

3.3 Dynamics parameter network

The LGSSM provides a tractable way to structure $p_\gamma(\mathbf{z}|\mathbf{a},\mathbf{u})$ into the variational approximation in (5). However, even in the simple case of a ball bouncing against a wall, the dynamics on $\mathbf{a}_t$ are not linear anymore. We can deal with these situations while preserving the linear dependency between consecutive states in the LGSSM, by non-linearly changing the parameters γ_t of the model over time as a function of the latent encodings up to time $t-1$ (so that we can still define a generative model). Smoothing is still possible as the state transition matrix $\mathbf{A}_t$ and others in γ_t do not have to be constant in order to obtain the exact posterior $p_\gamma(\mathbf{z}_t|\mathbf{a},\mathbf{u})$.

Recall that γ_t describes how the latent state $\mathbf{z}_{t-1}$ changes from time $t-1$ to time t. In the more general setting, the changes in dynamics at time t may depend on the history of the system, encoded in $\mathbf{a}_{1:t-1}$ and possibly a starting code $\mathbf{a}_0$ that can be learned from data. If, for instance, we see the ball colliding with a wall at time $t-1$, then we know that it will bounce at time t and change direction. We then let γ_t be a learnable function of $\mathbf{a}_{0:t-1}$, so that the prior in (2) becomes

$$p_\gamma(\mathbf{a},\mathbf{z}|\mathbf{u}) = \textstyle\prod_{t=1}^{T} p_{\gamma_t(\mathbf{a}_{0:t-1})}(\mathbf{a}_t|\mathbf{z}_t) \cdot p(\mathbf{z}_1) \prod_{t=2}^{T} p_{\gamma_t(\mathbf{a}_{0:t-1})}(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{u}_t) \ . \tag{8}$$

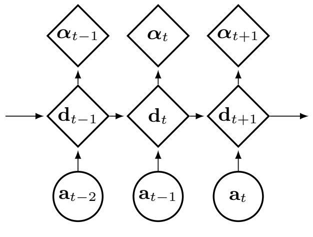


Figure 2: Dynamics parameter network for the KVAE.

During inference, after all the frames are encoded in $\mathbf{a}$, the dynamics parameter network returns $\gamma = \gamma(\mathbf{a})$, the parameters of the LGSSM at all time steps. We can now use the Kalman smoothing algorithm to find the exact conditional posterior over $\mathbf{z}$, that will be used when computing the gradients of the ELBO.

In our experiments the dependence of γ_t on $\mathbf{a}_{0:t-1}$ is modulated by a *dynamics parameter network* $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t(\mathbf{a}_{0:t-1})$, that is implemented with a recurrent neural network with LSTM cells that takes at each time step the encoded state as input and recurses $\mathbf{d}_t =$ *LSTM*$(\mathbf{a}_{t-1},\mathbf{d}_{t-1})$ and $\boldsymbol{\alpha}_t = \mathrm{softmax}(\mathbf{d}_t)$, as illustrated in figure 2. The output of the dynamics parameter network is weights that sum to one, $\sum_{k=1}^{K} \alpha_t^{(k)}(\mathbf{a}_{0:t-1}) = 1$. These weights choose and interpolate between K different operating modes:

$$\mathbf{A}_t = \sum_{k=1}^{K} \alpha_t^{(k)}(\mathbf{a}_{0:t-1})\mathbf{A}^{(k)}, \quad \mathbf{B}_t = \sum_{k=1}^{K} \alpha_t^{(k)}(\mathbf{a}_{0:t-1})\mathbf{B}^{(k)}, \quad \mathbf{C}_t = \sum_{k=1}^{K} \alpha_t^{(k)}(\mathbf{a}_{0:t-1})\mathbf{C}^{(k)} \ . \tag{9}$$

We globally learn K basic state transition, control and emission matrices $\mathbf{A}^{(k)}$, $\mathbf{B}^{(k)}$ and $\mathbf{C}^{(k)}$, and interpolate them based on information from the VAE encodings. The weighted sum can be interpreted as a soft mixture of K different LGSSMs whose time-invariant matrices are combined using the time-varying weights $\boldsymbol{\alpha}_t$. In practice, each of the K sets $\{\mathbf{A}^{(k)},\mathbf{B}^{(k)},\mathbf{C}^{(k)}\}$ models different dynamics, that will dominate when the corresponding $\alpha_t^{(k)}$ is high. The dynamics parameter network resembles the locally-linear transitions of [14, 31]; see section 6 for an in depth discussion on the differences.

4 Missing data imputation

Let $\mathbf{x}_{\mathsf{obs}}$ be an observed subset of frames in a video sequence, for instance depicting the initial movement and final positions of a ball in a scene. From its start and end, can we imagine how the ball reaches its final position? Autoregressive models like recurrent neural networks can only forward-generate $\mathbf{x}_t$ frame by frame, and cannot make use of the information coming from the final frames in the sequence. To impute the unobserved frames $\mathbf{x}_{\mathsf{un}}$ in the middle of the sequence, we need to do inference, not prediction.

The KVAE exploits the smoothing abilities of its LGSSM to use both the information from the past and the future when imputing missing data. In general, if $\mathbf{x} = \{\mathbf{x}_{\mathsf{obs}},\mathbf{x}_{\mathsf{un}}\}$, the unobserved frames in $\mathbf{x}_{\mathsf{un}}$ could also appear at non-contiguous time steps, e.g. missing at random. Data can be imputed by sampling from the joint density $p(\mathbf{a}_{\mathsf{un}},\mathbf{a}_{\mathsf{obs}},\mathbf{z}|\mathbf{x}_{\mathsf{obs}},\mathbf{u})$, and then generating $\mathbf{x}_{\mathsf{un}}$ from $\mathbf{a}_{\mathsf{un}}$. We factorize this distribution as

$$p(\mathbf{a}_{\mathsf{un}},\mathbf{a}_{\mathsf{obs}},\mathbf{z}|\mathbf{x}_{\mathsf{obs}},\mathbf{u}) = p_\gamma(\mathbf{a}_{\mathsf{un}}|\mathbf{z})\, p_\gamma(\mathbf{z}|\mathbf{a}_{\mathsf{obs}},\mathbf{u})\, p(\mathbf{a}_{\mathsf{obs}}|\mathbf{x}_{\mathsf{obs}}) \ , \tag{10}$$

and we sample from it with ancestral sampling starting from $\mathbf{x}_{\mathsf{obs}}$. Reading (10) from right to left, a sample from $p(\mathbf{a}_{\mathsf{obs}}|\mathbf{x}_{\mathsf{obs}})$ can be approximated with the variational distribution $q_\phi(\mathbf{a}_{\mathsf{obs}}|\mathbf{x}_{\mathsf{obs}})$. Then, if γ is fully known, $p_\gamma(\mathbf{z}|\mathbf{a}_{\mathsf{obs}}, \mathbf{u})$ is computed with an extension to the Kalman smoothing algorithm to sequences with missing data, after which samples from $p_\gamma(\mathbf{a}_{\mathsf{un}}|\mathbf{z})$ could be readily drawn.

However, when doing missing data imputation the parameters γ of the LGSSM are not known at all time steps. In the KVAE, each γ_t depends on all the previous encoded states, including $\mathbf{a}_{\mathsf{un}}$, and these need to be estimated before γ can be computed. In this paper we recursively estimate γ in the following way. Assume that $\mathbf{x}_{1:t-1}$ is known, but not $\mathbf{x}_t$. We sample $\mathbf{a}_{1:t-1}$ from $q_\phi(\mathbf{a}_{1:t-1}|\mathbf{x}_{1:t-1})$ using the VAE, and use it to compute $\gamma_{1:t}$. The computation of γ_{t+1} depends on $\mathbf{a}_t$, which is missing, and an estimate $\widehat{\mathbf{a}}_t$ will be used. Such an estimate can be arrived at in two steps. The filtered posterior distribution $p_\gamma(\mathbf{z}_{t-1}|\mathbf{a}_{1:t-1}, \mathbf{u}_{1:t-1})$ can be computed as it depends only on $\gamma_{1:t-1}$, and from it, we sample

$$\widehat{\mathbf{z}}_t \sim p_\gamma(\mathbf{z}_t|\mathbf{a}_{1:t-1}, \mathbf{u}_{1:t}) = \int p_{\gamma_t}(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{u}_t)\, p_\gamma(\mathbf{z}_{t-1}|\mathbf{a}_{1:t-1}, \mathbf{u}_{1:t-1})\, \mathrm{d}\mathbf{z}_{t-1} \tag{11}$$

and sample $\widehat{\mathbf{a}}_t$ from the predictive distribution of $\mathbf{a}_t$,

$$\widehat{\mathbf{a}}_t \sim p_\gamma(\mathbf{a}_t|\mathbf{a}_{1:t-1}, \mathbf{u}_{1:t}) = \int p_{\gamma_t}(\mathbf{a}_t|\mathbf{z}_t)\, p_\gamma(\mathbf{z}_t|\mathbf{a}_{1:t-1}, \mathbf{u}_{1:t})\, \mathrm{d}\mathbf{z}_t \approx p_{\gamma_t}(\mathbf{a}_t|\widehat{\mathbf{z}}_t)\,. \tag{12}$$

The parameters of the LGSSM at time $t+1$ are then estimated as $\gamma_{t+1}([\mathbf{a}_{0:t-1}, \widehat{\mathbf{a}}_t])$. The same procedure is repeated at the next time step if $\mathbf{x}_{t+1}$ is missing, otherwise $\mathbf{a}_{t+1}$ is drawn from the VAE. After the forward pass through the sequence, where we estimate γ and compute the filtered posterior for $\mathbf{z}$, the Kalman smoother's backwards pass computes the smoothed posterior. While the smoothed posterior distribution is not exact, as it relies on the estimate of γ obtained during the forward pass, it improves data imputation by using information coming from the whole sequence; see section 5 for an experimental illustration.

5 Experiments

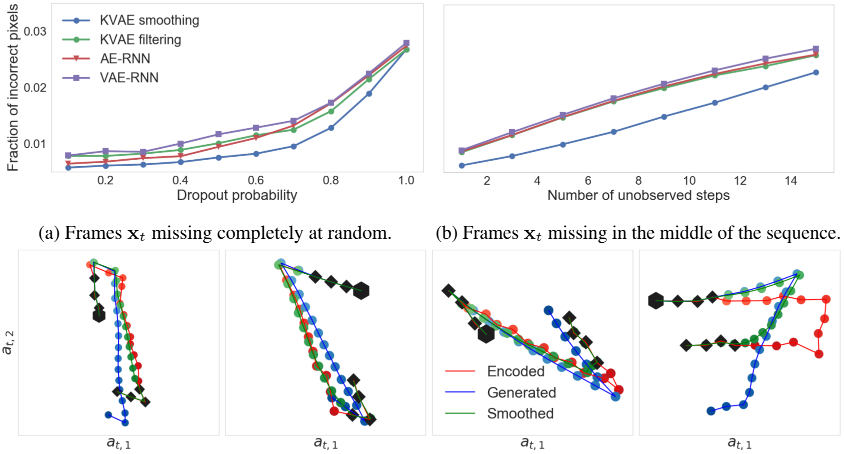
We motivated the KVAE with an example of a bouncing ball, and use it here to demonstrate the model's ability to separately learn a recognition and dynamics model from video, and use it to impute missing data. To draw a comparison with deep variational Bayes filters (DVBFs) [14], we apply the KVAE to [14]'s pendulum example. We further apply the model to a number of environments with different properties to demonstrate its generalizability. All models are trained end-to-end with stochastic gradient descent. Using the control input $\mathbf{u}_t$ in (1) we can inform the model of known quantities such as external forces, as will be done in the pendulum experiment. In all the other experiments, we omit such information and train the models fully unsupervised from the videos only. Further implementation details can be found in the supplementary material (appendix A) and in the Tensorflow [1] code released at github.com/simonkamronn/kvae.

5.1 Bouncing ball

We simulate 5000 sequences of 20 time steps each of a ball moving in a two-dimensional box, where each video frame is a 32x32 binary image. A video sequence is visualised as a single image in figure 4d, with the ball's darkening color reflecting the incremental frame index. In this set-up the initial position and velocity are randomly sampled. No forces are applied to the ball, except for the fully elastic collisions with the walls. The minimum number of latent dimensions that the KVAE requires to model the ball's dynamics are $\mathbf{a}_t \in \mathbb{R}^2$ and $\mathbf{z}_t \in \mathbb{R}^4$, as at the very least the ball's position in the box's 2d plane has to be encoded in $\mathbf{a}_t$, and $\mathbf{z}_t$ has to encode the ball's position and velocity. The model's flexibility increases with more latent dimensions, but we choose these settings for the sake of interpretable visualisations. The dynamics parameter network uses $K = 3$ to interpolate three modes, a constant velocity, and two non-linear interactions with the horizontal and vertical walls.

We compare the generation and imputation performance of the KVAE with two recurrent neural network (RNN) models that are based on the same auto-encoding (AE) architecture as the KVAE and are modifications of methods from the literature to be better suited to the bouncing ball experiments.[3]

[3] We also experimented with the SRNN model from [7] as it can do smoothing. However, the model is probably too complex for the task in hand, and we could not make it learn good dynamics.



(a) Frames $\mathbf{x}_t$ missing completely at random.

(b) Frames $\mathbf{x}_t$ missing in the middle of the sequence.

(c) Comparison of encoded (ground truth), generated and smoothed trajectories of a KVAE in the latent space $\mathbf{a}$. The black squares illustrate observed samples and the hexagons indicate the initial state. Notice that the $\mathbf{a}_t$'s lie on a manifold that can be rotated and stretched to align with the frames of the video.

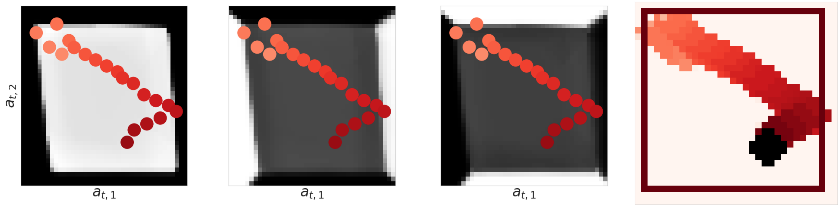
Figure 3: Missing data imputation results.

In the *AE-RNN*, inspired by the architecture from [27], a pretrained convolutional auto-encoder, identical to the one used for the KVAE, feeds the encodings to an LSTM network [11]. During training the LSTM predicts the next encoding in the sequence and during generation we use the previous output as input to the current step. For data imputation the LSTM either receives the previous output or, if available, the encoding of the observed frame (similarly to filtering in the KVAE). The *VAE-RNN* is identical to the AE-RNN except that uses a VAE instead of an AE, similarly to the model from [5].

Figure 3a shows how well missing frames are imputed in terms of the average fraction of incorrectly guessed pixels. In it, the first 4 frames are observed (to initialize the models) after which the next 16 frames are dropped at random with varying probabilities. We then impute the missing frames by doing filtering and smoothing with the KVAE. We see in figure 3a that it is beneficial to utilize information from the whole sequence (even the future observed frames), and a KVAE with smoothing outperforms all competing methods. Notice that dropout probability 1 corresponds to pure generation from the models. **Figure 3b** repeats this experiment, but makes it more challenging by removing an increasing number of *consecutive* frames from the middle of the sequence ($T = 20$). In this case the ability to encode information coming from the future into the posterior distribution is highly beneficial, and smoothing imputes frames much better than the other methods. **Figure 3c** graphically illustrates figure 3b. We plot three trajectories over $\mathbf{a}_t$-encodings. The *generated* trajectories were obtained after initializing the KVAE model with 4 initial frames, while the *smoothed* trajectories also incorporated encodings from the last 4 frames of the sequence. The *encoded* trajectories were obtained with no missing data, and are therefore considered as ground truth. In the first three plots in figure 3c, we see that the backwards recursion of the Kalman smoother corrects the trajectory obtained with generation in the forward pass. However, in the fourth plot, the poor trajectory that is obtained during the forward generation step, makes smoothing unable to follow the ground truth.

The smoothing capabilities of KVAEs make it also possible to train it with up to 40% of missing data with minor losses in performance (appendix C in the supplementary material). Links to videos of the imputation results and long-term generation from the models can be found in appendix B and at sites.google.com/view/kvae.

Understanding the dynamics parameter network. In our experiments the dynamics parameter network $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t(\mathbf{a}_{0:t-1})$ is an LSTM network, but we could also parameterize it with any differentiable function of $\mathbf{a}_{0:t-1}$ (see appendix D in the supplementary material for a comparison of various



(a) $k = 1$ (b) $k = 2$ (c) $k = 3$ (d) Reconstruction of $\mathbf{x}$

Figure 4: A visualisation of the dynamics parameter network $\alpha_t^{(k)}(\mathbf{a}_{t-1})$ for $K = 3$, as a function of $\mathbf{a}_{t-1}$. The three $\alpha_t^{(k)}$'s sum to one at every point in the encoded space. The greyscale backgrounds in **a)** to **c)** correspond to the intensity of the weights $\alpha_t^{(k)}$, with white indicating a weight of one in the dynamics parameter network's output. Overlaid on them is the full latent encoding $\mathbf{a}$. **d)** shows the reconstructed frames of the video as one image.

architectures). When using a multi-layer perceptron (MLP) that depends on the previous encoding as mixture network, i.e. $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t(\mathbf{a}_{t-1})$, figure 4 illustrates how the network chooses the mixture of learned dynamics. We see that the model has correctly learned to choose a transition that maintains a constant velocity in the center ($k = 1$), reverses the horizontal velocity when in proximity of the left and right wall ($k = 2$), the reverses the vertical velocity when close to the top and bottom ($k = 3$).

5.2 Pendulum experiment

Model	Test ELBO
KVAE (CNN)	810.08
KVAE (MLP)	807.02
DVBF	798.56
DMM	784.70

Table 1: Pendulum experiment.

We test the KVAE on the experiment of a dynamic torque-controlled pendulum used in [14]. Training, validation and test set are formed by 500 sequences of 15 frames of 16x16 pixels. We use a KVAE with $\mathbf{a}_t \in \mathbb{R}^2$, $\mathbf{z}_t \in \mathbb{R}^3$ and $K = 2$, and try two different encoder-decoder architectures for the VAE, one using a MLP and one using a convolutional neural network (CNN). We compare the performaces of the KVAE to DVBFs [14] and deep Markov models[4] (DMM) [17], non-linear SSMs parameterized by deep neural networks whose intractable posterior distribution is approximated with an inference network. In table 1 we see that the KVAE outperforms both models in terms of ELBO on a test set, showing that for the task in hand it is preferable to use a model with simpler dynamics but exact posterior inference.

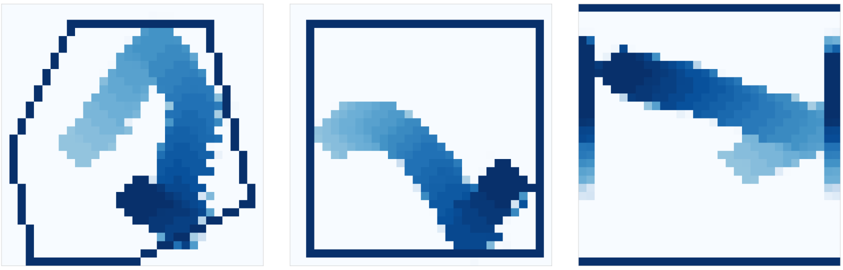
5.3 Other environments

To test how well the KVAE adapts to different environments, we trained it end-to-end on videos of (i) a ball bouncing between walls that form an irregular polygon, (ii) a ball bouncing in a box and subject to gravity, (iii) a Pong-like environment where the paddles follow the vertical position of the ball to make it stay in the frame at all times. Figure 5 shows that the KVAE learns the dynamics of all three environments, and generates realistic-looking trajectories. We repeat the imputation experiments of figures 3a and 3b for these environments in the supplementary material (appendix E), where we see that KVAEs outperform alternative models.

6 Related work

Recent progress in unsupervised learning of high dimensional sequences is found in a plethora of both deterministic and probabilistic generative models. The VAE framework is a common work-horse in the stable of probabilistic inference methods, and it is extended to the temporal setting by [5, 7, 14, 17]. In particular, deep neural networks can parameterize the transition and emission distributions of different variants of deep state-space models [7, 14, 17]. In these extensions, inference networks

[4] Deep Markov models were previously referred to as deep Kalman filters.



(a) Irregular polygon. (b) Box with gravity. (c) Pong-like environment.

Figure 5: Generations from the KVAE trained on different environments. The videos are shown as single images, with color intensity representing the incremental sequence index t. In the simulation that resembles Atari's Pong game, the movement of the two paddles (left and right) is also visible.

define a variational approximation to the intractable posterior distribution of the latent states at each time step. For the tasks in section 5, it is preferable to use the KVAE's simpler temporal model with an exact (conditional) posterior distribution than a highly non-linear model where the posterior needs to be approximated. A different combination of VAEs and probabilistic graphical models has been explored in [13], which defines a general class of models where inference is performed with message passing algorithms that use deep neural networks to map the observations to conjugate graphical model potentials.

In classical non-linear extensions of the LGSSM like the extended Kalman filter and in the locally-linear dynamics of [14, 31], the transition matrices at time t have a non-linear dependence on $\mathbf{z}_{t-1}$. The KVAE's approach is different: by introducing the latent encodings $\mathbf{a}_t$ and making γ_t depend on $\mathbf{a}_{1:t-1}$, the *linear* dependency between consecutive states of $\mathbf{z}$ is preserved, so that the exact smoothed posterior can be computed given $\mathbf{a}$, and used to perform missing data imputation. LGSSM with dynamic parameterization have been used for large-scale demand forecasting in [25]. [18] introduces recurrent switching linear dynamical systems, that combine deep learning techniques and switching Kalman filters [20] to model low-dimensional time series. [9] introduces a *discriminative* approach to estimate the low-dimensional state of a LGSSM from input images. The resulting model is reminiscent of a KVAE with no decoding step, and is therefore not suited for unsupervised learning and video generation. Recent work in the non-sequential setting has focused on disentangling basic visual concepts in an image [10].

Great strides have been made in the reinforcement learning community to model how environments evolve in response to action [4, 21, 22, 28, 30]. In similar spirit to this paper, [30] extracts a latent representation from a PCA representation of the frames where controls can be applied. [4] introduces action-conditional dynamics parameterized with LSTMs and, as for the KVAE, a computationally efficient procedure to make long term predictions without generating high dimensional images at each time step. As autoregressive models, [27] develops a sequence to sequence model of video representations that uses LSTMs to define both the encoder and the decoder. [6] develops an action-conditioned video prediction model of the motion of a robot arm using convolutional LSTMs that models the change in pixel values between two consecutive frames.

While the focus in this work is to define a generative model for *high dimensional* videos of simple physical systems, several recent works have combined physical models of the world with deep learning to learn the dynamics of objects in more complex but *low-dimensional* environments [2, 3, 8, 32].

7 Conclusion

The KVAE, a model for unsupervised learning of high-dimensional videos, was introduced in this paper. It disentangles an object's latent representation $\mathbf{a}_t$ from a latent state $\mathbf{z}_t$ that describes its dynamics, and can be learned end-to-end from raw video. Because the exact (conditional) smoothed posterior distribution over the states of the LGSSM can be computed, one generally sees a marked improvement in inference and missing data imputation over methods that don't have this property.

A desirable property of disentangling the two latent representations is that temporal reasoning, and possibly planning, could be done in the latent space. As a proof of concept, we have been deliberate in focussing our exposition to videos of static worlds that contain a few moving objects, and leave extensions of the model to real world videos or sequences coming from an agent exploring its environment to future work.

Acknowledgements

We would like to thank Lars Kai Hansen for helpful discussions on the model design. Marco Fraccaro is supported by Microsoft Research through its PhD Scholarship Programme. We thank NVIDIA Corporation for the donation of TITAN X GPUs.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] P. W. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *NIPS*, 2016.

[3] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum. A compositional object-based approach to learning physical dynamics. In *ICLR*, 2017.

[4] S. Chiappa, S. Racanière, D. Wierstra, and S. Mohamed. Recurrent environment simulators. In *ICLR*, 2017.

[5] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *NIPS*, 2015.

[6] C. Finn, I. J. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *NIPS*, 2016.

[7] M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther. Sequential neural models with stochastic layers. In *NIPS*, 2016.

[8] K. Fragkiadaki, P. Agrawal, S. Levine, and J. Malik. Learning visual predictive models of physics for playing billiards. In *ICLR*, 2016.

[9] T. Haarnoja, A. Ajay, S. Levine, and P. Abbeel. Backprop KF: learning discriminative deterministic state estimators. In *NIPS*, 2016.

[10] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2017.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997.

[12] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[13] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams. Composing graphical models with neural networks for structured representations and fast inference. In *NIPS*, 2016.

[14] M. Karl, M. Soelch, J. Bayer, and P. van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. In *ICLR*, 2017.

[15] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[16] D. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

[17] R. Krishnan, U. Shalit, and D. Sontag. Structured inference networks for nonlinear state space models. In *AAAI*, 2017.

[18] S. Linderman, M. Johnson, A. Miller, R. Adams, D. Blei, and L. Paninski. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. In *AISTATS*, 2017.

[19] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017.

[20] K. P. Murphy. Switching Kalman filters. Technical report, 1998.

[21] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh. Action-conditional video prediction using deep networks in atari games. In *NIPS*, 2015.

[22] V. Patraucean, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. *arXiv:1511.06309*, 2015.

[23] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

[24] S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–45, 1999.

[25] M. W. Seeger, D. Salinas, and V. Flunkert. Bayesian intermittent demand forecasting for large inventories. In *NIPS*, 2016.

[26] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[27] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using LSTMs. In *ICML*, 2015.

[28] W. Sun, A. Venkatraman, B. Boots, and J. A. Bagnell. Learning to filter with predictive state inference machines. In *ICML*, 2016.

[29] L. G. Ungerleider and L. G. Haxby. “What” and “where” in the human brain. *Curr. Opin. Neurobiol.*, 4:157–165, 1994.

[30] N. Wahlström, T. B. Schön, and M. P. Deisenroth. From pixels to torques: Policy learning with deep dynamical models. *arXiv:1502.02251*, 2015.

[31] M. Watter, J. Springenberg, J. Boedecker, and M. Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *NIPS*, 2015.

[32] J. Wu, I. Yildirim, J. J. Lim, W. T. Freeman, and J. B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *NIPS*, 2015.

A Experimental details

We will describe here some of the most important experimental details. The rest of the details can be found in the code at github.com/simonkamronn/kvae.

Data generation. All the videos were generated using the physics engine Pymunk. We generated 5000 videos for training and 1000 for testing.

Encoder/Decoder architecture for the KVAE. As we only use image-based observations, the encoder is fixed to a three layer convolutional neural network with 32 units in each layer, kernel-size of 3x3, stride of 2, and ReLU activations. The decoder is an equally sized network using the Sub-Pixel[26] procedure for deconvolution. In the pendulum experiment however we also test MLPs.

Optimization. As optimizer we use ADAM [15] with an initial learning rate of 0.007 and an exponential decay scheme with a rate of 0.85 every 20 epochs. Training one epoch takes 55 seconds on an NVIDIA Titan X and the model converges in roughly 80 epochs.

Training tricks for end-to-end learning. The biggest challenge of this optimization problem is how to avoid poor local minima, for example where all the focus is given to the reconstruction term, at the expense of the prior dynamics given by the LGSSM. To achieve a quick convergence in all the experiments we found it helpful to

- downweight the reconstruction term from of VAEs during training, that is scaled by 0.3. By doing this, we can in fact help the model to focus on learning the temporal dynamics.
- learn for the first few epochs only the the VAE parameters θ and ϕ and the globally learned matrices $\mathbf{A}^{(k)}$, $\mathbf{B}^{(k)}$ and $\mathbf{C}^{(k)}$, but not the parameters of the dynamics parameter network $\boldsymbol{\alpha}_t(\mathbf{a}_{0:t-1})$. After this phase, all parameters are learned jointly. This allows the model to first learn good VAE embeddings and the scale of the prior, and then learn how to utilize the K different dynamics.

Choice of hyperparameters for the LGSSM. In most of the experiments we used $\mathbf{a}_t \in \mathbb{R}^2$, $\mathbf{z}_t \in \mathbb{R}^4$ and $K = 3$. In the *gravity* experiments we used however $\mathbf{z}_t \in \mathbb{R}^5$ as the model has no controls applied to it and needs to be able to learn a bias term due to the presence of the external force of gravity. The *polygon* experiments uses $K = 7$ as it needs to learn more complex dynamics. In general, we did not find difficult to tune the parameters of the KVAE, as the model can learn to prune unused components (if flexible enough).

B Videos

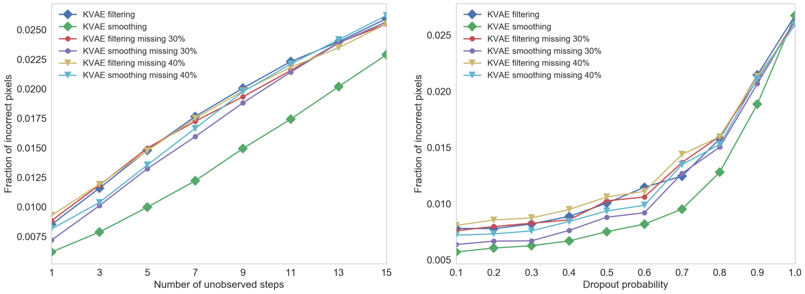
Videos are generated from all models by initializing with 4 frames and then sampling. The *filtering* and *smoothing* versions are allowed to observe part of the sequence depending on the masking scheme. All the *filtering* and *smoothing* videos are generated from sequences applied with a random mask with a masking probability of 80% (as in figure 3a) except for the videos with the suffix *consecutive* in which only the first and last 4 frames are observed (as in figure 3b). Only the KVAE models have *smoothing* videos. For the bouncing ball experiment (named *box* in the attached folder), we also show the videos from a model trained with 40% missing data.

In most videos the black ball is the ground truth, and the red is the one generated from the model, except for the ones marked *long_generation* in which the true sequence is not shown.

Videos are available from Google Drive and the website sites.google.com/view/kvae.

C Training with missing data.

The smoothed posterior described in section 4 can also be used to train the KVAE with missing data. In this case, we only need to modify the ELBO by masking the contribution of the missing data points



(a)

(b)

Figure 6: Training with missing data

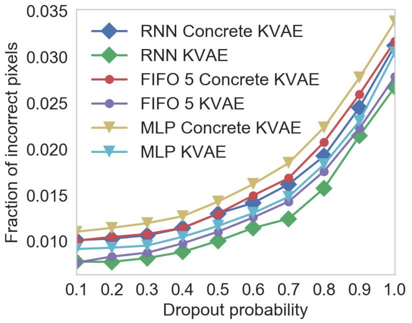


Figure 7: Comparison of modelling choices wrt. the α-network

in the joint probability distribution and variational approximation:

$$p(\mathbf{x}, \mathbf{a}, \mathbf{z}, \mathbf{u}) = p(\mathbf{z}_1) \prod_{t=2}^{T} p_{\gamma_t}(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{u}_t) \prod_{t=1}^{T} p_{\gamma_t}(\mathbf{a}_t|\mathbf{z}_t)^{\mathcal{I}_t} \prod_{t=1}^{T} p_\theta(\mathbf{x}_t|\mathbf{a}_t)^{\mathcal{I}_t}$$

$$q_\phi(\mathbf{a}|\mathbf{x}) = \prod_{t=1}^{T} q_\phi(\mathbf{a}_t|\mathbf{x}_t)^{\mathcal{I}_t} \ ,$$

where $\mathcal{I}_t$ is 0 if the data point is missing, 1 otherwise. Figure 6 illustrates a slight degradation in performance when training with respectively 30% and 40% missing data but, remarkably, the accuracy is still better when using smoothing in these conditions than with filtering with all training data available.

D Dynamics parameter network architecture

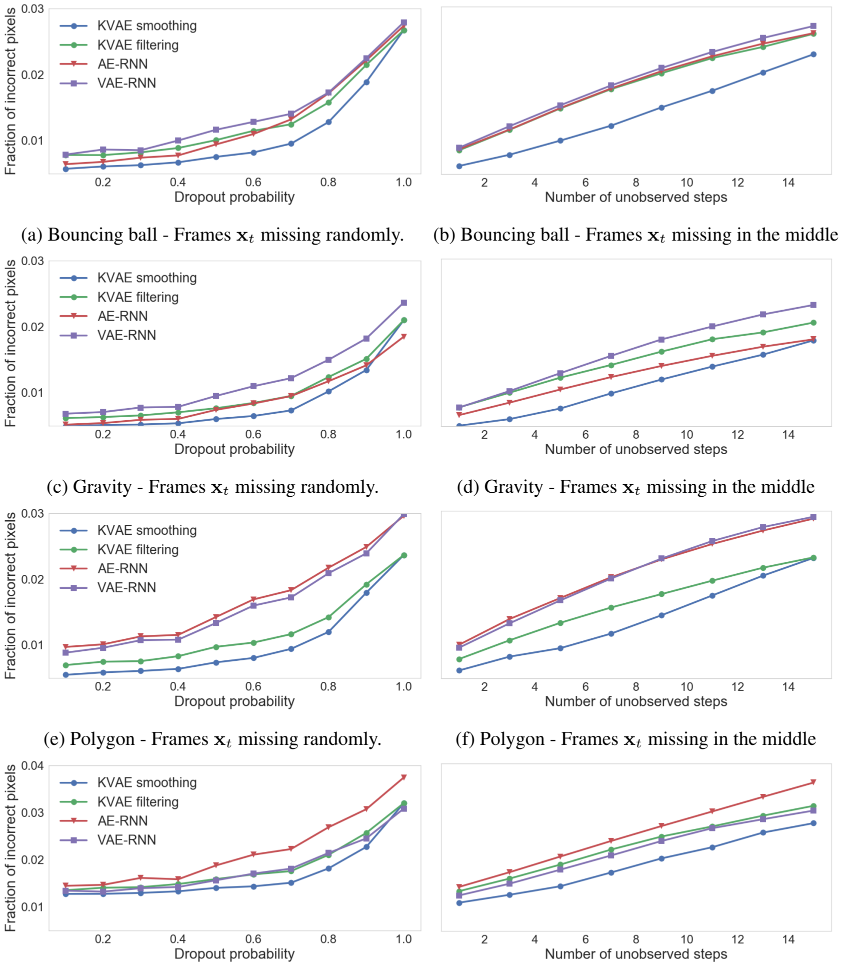
As the α-network governs the non-linear dynamics, it has a significant impact on the modelling capabilities. Here we list the architectural choices considered:

- **MLP** with two hidden layers.
- **Recurrent Neural Networks** with LSTM units.
- **’First in, first out memory’ (FIFO) MLP** with access to 5 time steps.

In all cases, we can also model α as an (approximate) discrete random variable using the the Concrete distribution [19, 12]. In this case we can recover an approximation to the switching Kalman filter[20].

In figure 7 the different choices are tested against each other on the bouncing ball data. In this case all the alternative choices result in poorer performances than the LSTM chosen for all the other experiments. We believe that LSTMs are able to better model the discretization errors coming from the collisions and the 32x32 rendering of the trajectories computed by the physics engine.

E Imputation in all environments



(a) Bouncing ball - Frames $\mathbf{x}_t$ missing randomly.

(b) Bouncing ball - Frames $\mathbf{x}_t$ missing in the middle

(c) Gravity - Frames $\mathbf{x}_t$ missing randomly.

(d) Gravity - Frames $\mathbf{x}_t$ missing in the middle

(e) Polygon - Frames $\mathbf{x}_t$ missing randomly.

(f) Polygon - Frames $\mathbf{x}_t$ missing in the middle

(g) Pong - Frames $\mathbf{x}_t$ missing randomly.

(h) Pong - Frames $\mathbf{x}_t$ missing in the middle

Figure 8: Imputation results for all environments

Bibliography

Aggarwal, J. K. and Xia, L. (2014). Human activity recognition from 3D data: A review. *Pattern Recognit. Lett.*, 48:70–80.

Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive Mob. Comput.*, 7(6):643–659.

Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., and Baronchelli, A. (2018). Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, 2(7):485–491.

Althoff, T. (2017). Population-Scale Pervasive Health. *IEEE Pervasive Comput.*, 16(4):75–79.

Althoff, T., Sosič, R., Hicks, J. L., King, A. C., Delp, S. L., and Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature*, 547(7663):336–339.

Altini, M., Casale, P., Penders, J. F., and Amft, O. (2015). Personalization of Energy Expenditure Estimation in Free Living Using Topic Models. *IEEE J Biomed Health Inform*, 19(5):1577–1586.

Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J. L. (2013). A Public Domain Dataset for Human Activity Recognition Using Smartphones. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (April):24–26.

Badimon, L., Bugiardini, R., Cenko, E., Cubedo, J., Dorobantu, M., Duncker, D. J., Estruch, R., Milicic, D., Tousoulis, D., Vasiljevic, Z., Vilahur, G., de Wit, C., and Koller, A. (2017). Position paper of the European Society of Cardiology-working group of coronary pathophysiology and microcirculation: obesity and heart disease. *Eur. Heart J.*, 38(25):1951–1958.

Baños, O., Damas, M., Pomares, H., Rojas, I., Tóth, M. A., and Amft, O. (2012). A Benchmark Dataset to Evaluate Sensor Displacement in Activity Recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 1026–1035, New York, NY, USA. ACM.

Bao, L. and Intille, S. S. (2004). Activity Recognition from User-Annotated Acceleration Data. *Pervasive Computing*, pages 1–17.

Barocas, S. and Selbst, A. D. (2016). Big Data's Disparate Impact.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., and others (2010). A theory of learning from different domains. *Mach. Learn.*

Berardi, V., Carretero-González, R., Bellettiere, J., Adams, M. A., Hughes, S., and Hovell, M. (2018). A Markov Approach for Increasing Precision in the Assessment of Data-Intensive Behavioral Interventions. *J. Biomed. Inform.*

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer, 4 edition.

Brezmes, T., Gorricho, J.-L., and Cotrina, J. (2009). Activity Recognition from Accelerometer Data on a Mobile Phone. In *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, pages 796–799. Springer Berlin Heidelberg.

Campbell, A. T., Eisenman, S. B., Lane, N. D., Miluzzo, E., Peterson, R. A., Lu, H., Zheng, X., Musolesi, M., Fodor, K., and Ahn, G. (2008). The Rise of People-Centric Sensing. *IEEE Internet Comput.*, 12(4):12–21.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles*, 76(1):1–32.

Carter, M. C., Burley, V. J., Nykjaer, C., and Cade, J. E. (2013). Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial. *J. Med. Internet Res.*, 15(4):e32.

Chambon, S., Thorey, V., Arnal, P. J., Mignot, E., and Gramfort, A. (2018). A deep learning architecture to detect events in EEG signals during sleep.

Chan, Y.-F. Y., Wang, P., Rogers, L., Tignor, N., Zweig, M., Hershman, S. G., Genes, N., Scott, E. R., Krock, E., Badgeley, M., Edgar, R., Violante, S., Wright, R., Powell, C. A., Dudley, J. T., and Schadt, E. E. (2017). The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. *Nat. Biotechnol.*

Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., and Yu, Z. (2012). Sensor-based activity recognition. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.*, 42(6):790–808.

Chowdhury, E. A., Western, M. J., Nightingale, T. E., Peacock, O. J., and Thompson, D. (2017). Assessment of laboratory and daily energy expenditure estimates from consumer multi-sensor physical activity monitors. *PLoS One*, 12(2):e0171720.

Clawson, J., Pater, J. A., Miller, A. D., Mynatt, E. D., and Mamykina, L. (2015). No Longer Wearing: Investigating the Abandonment of Personal Health-tracking Technologies on Craigslist. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 647–658, New York, NY, USA. ACM.

Cuttone, A., Lehmann, S., and Larsen, J. E. (2014). Inferring human mobility from sparse low accuracy mobile sensing data. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct*, pages 995–1004.

Davies, C. A., Spence, J. C., Vandelanotte, C., Caperchione, C. M., and Mummery, W. K. (2012). Meta-analysis of internet-delivered interventions to increase physical activity levels. *Int. J. Behav. Nutr. Phys. Act.*, 9:52.

Ding, D., Lawson, K. D., Kolbe-Alexander, T. L., Finkelstein, E. A., Katzmarzyk, P. T., van Mechelen, W., Pratt, M., and Lancet Physical Activity Series 2 Executive Committee (2016). The economic burden of physical inactivity: a global analysis of major non-communicable diseases. *Lancet*, 388(10051):1311–1324.

Direito, A., Carraça, E., Rawstorn, J., Whittaker, R., and Maddison, R. (2017). mHealth Technologies to Influence Physical Activity and Sedentary Behaviors: Behavior Change Techniques, Systematic Review and Meta-Analysis of Randomized Controlled Trials. *Ann. Behav. Med.*, 51(2):226–239.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *International Conference on Machine Learning*, pages 647–655. jmlr.org.

Durbin, J. and Koopman, S. J. (2012). Time series analysis by state space methods.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., and Others (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231. aaai.org.

Evenson, K. R., Goto, M. M., and Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int. J. Behav. Nutr. Phys. Act.*, 12:159.

Fraccaro, M. (2018). *Deep Latent Variable Models for Sequential Data.* PhD thesis, Technical University of Denmark.

Fraccaro, M., Kamronn, S., Paquet, U., and Winther, O. (2017). A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning. In *Advances in Neural Information Processing Systems 30.*

Franklin, R. L. (2017). *Freewill and Determinism: A Study of Rival Conceptions of Man.* Taylor & Francis.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., and others (2016). Domain-adversarial training of neural networks. *The Journal of Machine.*

Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural Comput.*, 12(4):831–864.

Gomersall, S., Maher, C., English, C., Rowlands, A., and Olds, T. (2015). Time regained: when people stop a physical activity program, how does their time use change? A randomised controlled trial. *PLoS One*, 10(5):e0126665.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Guan, Y. and Plötz, T. (2017). Ensembles of Deep LSTM Learners for Activity Recognition Using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2):11:1–11:28.

Hallal, P. C., Andersen, L. B., Bull, F. C., Guthold, R., Haskell, W., Ekelund, U., and Lancet Physical Activity Series Working Group (2012). Global physical activity levels: surveillance progress, pitfalls, and prospects. *Lancet*, 380(9838):247–257.

Hammerla, N. Y., Halloran, S., and Ploetz, T. (2016). Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic Variational Inference. *J. Mach. Learn. Res.*, 14(1):1303–1347.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, 9(3):90–95.

Incel, O. D., Kose, M., and Ersoy, C. (2013). A Review and Taxonomy of Activity Recognition on Mobile Phones. *Bionanoscience*, 3:145–171.

Jensen, H. A. R., Davidsen, M., Ekholm, O., and Christensen, A. I. (2018). *Danskernes sundhed: den nationale sundhedsprofil 2017*. Sundhedsstyrelsen.

Jeran, S., Steinbrecher, A., and Pischon, T. (2016). Prediction of activity-related energy expenditure using accelerometer-derived physical activity under free-living conditions: a systematic review. *Int. J. Obes.*, 40(8):1187–1197.

Kahneman, D. and Egan, P. (2011). Thinking, fast and slow.

Kalman, R. E. and Others (1960). A new approach to linear filtering and prediction problems. *Int. J. Eng. Trans. A*, 82(1):35–45.

Kingma, D. and Welling, M. (2014). Auto-encoding variational Bayes. In *ICLR*.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*. arxiv.org.

Kohl, 3rd, H. W., Craig, C. L., Lambert, E. V., Inoue, S., Alkandari, J. R., Leetongin, G., Kahlmeier, S., and Lancet Physical Activity Series Working Group (2012). The pandemic of physical inactivity: global action for public health. *Lancet*, 380(9838):294–305.

Kohl, L. F. M., Crutzen, R., and de Vries, N. K. (2013). Online prevention aimed at lifestyle behaviors: a systematic review of reviews. *J. Med. Internet Res.*, 15(7):e146.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques.* MIT press.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*

Kwapisz, J., Weiss, G., and Moore, S. (2011). Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations . . .*, 12(2):74–82.

Lane, N. D. and Georgiev, P. (2015). Can Deep Learning Revolutionize Mobile Sensing? In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, HotMobile '15, pages 117–122, New York, NY, USA. ACM.

Lathia, N., Pejovic, V., Rachuri, K., Mascolo, C., Musolesi, M., and Rentfrow, P. J. (2014). Smartphones for Large-scale Behaviour Change Interventions. *IEEE Pervasive Comput.*, pages 1–11.

Lee, I.-M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., Katzmarzyk, P. T., and Lancet Physical Activity Series Working Group (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *Lancet*, 380(9838):219–229.

Lewis, B. A., Napolitano, M. A., Buman, M. P., Williams, D. M., and Nigg, C. R. (2017). Future directions in physical activity intervention research: expanding our focus to sedentary behaviors, technology, and dissemination. *J. Behav. Med.*, 40(1):112–126.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*, pages 21–37. Springer International Publishing.

Maddison, R., Gemming, L., Monedero, J., Bolger, L., Belton, S., Issartel, J., Marsh, S., Direito, A., Solenhill, M., Zhao, J., Exeter, D. J., Vathsangam, H., and Rawstorn, J. C. (2017). Quantifying Human Movement Using the Movn Smartphone App: Validation and Field Study. *JMIR Mhealth Uhealth*, 5(8):e122.

Maher, C. A., Mire, E., Harrington, D. M., Staiano, A. E., and Katzmarzyk, P. T. (2013). The independent and combined associations of physical activity and sedentary behavior with obesity in adults: NHANES 2003-06: Adult Activity Patterns and Obesity Risk. *Obesity*, 21(12):E730–E737.

Morales, F. J. O. and Roggen, D. (2016). Deep Convolutional Feature Transfer Across Mobile Activity Recognition Domains, Sensor Modalities and Locations. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, ISWC '16, pages 92–99, New York, NY, USA. ACM.

Müller, A. M., Maher, C. A., Vandelanotte, C., Hingle, M., Middelweerd, A., Lopez, M. L., DeSmet, A., Short, C. E., Nathan, N., Hutchesson, M. J., Poppe, L., Woods, C. B., Williams, S. L., and Wark, P. A. (2018a). Physical Activity, Sedentary Behavior, and Diet-Related eHealth and mHealth Research: Bibliometric Analysis. *J. Med. Internet Res.*, 20(4):e122.

Müller, M. J., Geisler, C., Hübers, M., Pourhassan, M., Braun, W., and Bosy-Westphal, A. (2018b). Normalizing resting energy expenditure across the life course in humans: challenges and hopes. *Eur. J. Clin. Nutr.*, 72(5):628–637.

Mytton, O. T., Panter, J., and Ogilvie, D. (2016). Longitudinal associations of active commuting with body mass index. *Prev. Med.*, 90:1–7.

Nahum-Shani, I., Hekler, E. B., and Spruijt-Metz, D. (2015). Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychol.*, 34S:1209–1219.

Ordóñez, F. J. and Roggen, D. (2016). Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors*, 16(1).

Patel, A. V., Bernstein, L., Deka, A., Feigelson, H. S., Campbell, P. T., Gapstur, S. M., Colditz, G. A., and Thun, M. J. (2010). Leisure time spent sitting in relation to total mortality in a prospective cohort of US adults. *Am. J. Epidemiol.*, 172(4):419–429.

Peach, D., Van Hoomissen, J., and Callender, H. L. (2014). Exploring the ActiLife® filtration algorithm: converting raw acceleration data to counts. *Physiol. Meas.*, 35(12):2359.

Pearl, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.*, 3(0):96–146.

Pejovic, V., Mehrotra, A., and Musolesi, M. (2015). Anticipatory Mobile Digital Health: Towards Personalised Proactive Therapies and Prevention Strategies. *arXiv [cs.CY]*.

Plötz, T. and Guan, Y. (2018). Deep Learning for Human Activity Recognition in Mobile Computing. *Computer*, 51(5):50–59.

Plötz, T., Hammerla, N. Y., and Olivier, P. (2011). Feature learning for activity recognition in ubiquitous computing. *Proceeding IJCAI'11 Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, Volume 2:1729–1734.

Raento, M., Oulasvirta, A., and Eagle, N. (2009). Smartphones: An Emerging Tool for Social Scientists. *Sociol. Methods Res.*, 37(3):426–454.

Ranganath, R., Gerrish, S., and Blei, D. M. (2013). Black Box Variational Inference.

Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning*, Lecture Notes in Computer Science, pages 63–71. Springer Berlin Heidelberg.

Ravi, N., Dandekar, N., Mysore, P., and Littman, M. M. L. (2005). Activity recognition from accelerometer data. *Proceedings of the national ...*, pages 1541–1546.

Reis, R. S., Salvo, D., Ogilvie, D., Lambert, E. V., Goenka, S., Brownson, R. C., and Lancet Physical Activity Series 2 Executive Committee (2016). Scaling up physical activity interventions worldwide: stepping up to larger and smarter approaches to get people moving. *Lancet*, 388(10051):1337–1348.

Reyes Ortiz, J. L. (2015). *Smartphone-Based Human Activity Recognition.*

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *ICML*.

Rosenkilde, M., Petersen, M. B., Gram, A. S., Quist, J. S., Winther, J., Kamronn, S. D., Milling, D. H., Larsen, J. E., Jespersen, A. P., and Stallknecht, B. (2017). The GO-ACTIWE randomized controlled trial - An interdisciplinary study designed to investigate the health effects of active commuting and leisure time physical activity. *Contemp. Clin. Trials*, 53:122–129.

Schoeller, D. A. and Jefford, G. (2002). Determinants of the energy costs of light activities: inferences for interpreting doubly labeled water data. *Int. J. Obes. Relat. Metab. Disord.*, 26(1):97–101.

Seefeldt, V., Malina, R. M., and Clark, M. A. (2002). Factors affecting levels of physical activity in adults. *Sports Med.*, 32(3):143–168.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference*, 90(2):227–244.

Shoaib, M., Bosch, S., Incel, O., Scholten, H., and Havinga, P. (2015). A Survey of Online Activity Recognition Using Mobile Phones. *Sensors*, 15:2059–2085.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2951–2959. Curran Associates, Inc.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., and Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PLoS One*, 9(4):e95978.

Stuckey, M. I., Carter, S. W., and Knight, E. (2017). The role of smartphones in encouraging physical activity in adults. *Int. J. Gen. Med.*, 10:293–303.

Sullivan, A. N. and Lachman, M. E. (2016). Behavior Change with Fitness Technology in Sedentary Adults: A Review of the Evidence for Increasing Physical Activity. *Front Public Health*, 4:289.

Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism.

Twomey, N., Diethe, T., Kull, M., Song, H., Camplani, M., Hannuna, S., Fafoutis, X., Zhu, N., Woznowski, P., Flach, P., and Craddock, I. (2016). The SPHERE Challenge: Activity Recognition with Multimodal Sensor Data.

Ungerleider, L. G. and Haxby, L. G. (1994). “What” and “where” in the human brain. *Curr. Opin. Neurobiol.*, 4:157–165.

Vanderplas, J. and Granger, B. (2018). Altair: Declarative Visualization in Python. Accessed: 2018-8-8.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Westerterp, K. R. (2017). Doubly labelled water assessment of energy expenditure: principle, practice, and promise. *Eur. J. Appl. Physiol.*, 117(7):1277–1285.

Wu, W., Dasgupta, S., Ramirez, E. E., Peterson, C., and Norman, G. J. (2012). Classification accuracies of physical activities using smartphone motion sensors. *J. Med. Internet Res.*, 14(5):e130.

Xie, Q., Dai, Z., Du, Y., Hovy, E., and Neubig, G. (2017). Controllable Invariance through Adversarial Feature Learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 585–596. Curran Associates, Inc.

Zandbergen, P. A. and Barbeau, S. J. (2011). Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones. *J. Navig.*, 64(3):381–399.

Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., and Zhang, J. (2014). Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors. *MobiCASE*.

Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. (2017). Advances in Variational Inference. *arXiv [cs.LG]*.

Zhao, H., Zhang, S., Wu, G., Costeira, J. P., Moura, J. M. F., and Gordon, G. J. (2017). Multiple Source Domain Adaptation with Adversarial Training of Neural Networks.