Technical Note PR-TN-2016/–

Issued: September/2016

Measurement methods for body fat (%) assessment from 3D Kinect scans

Andra Alexa

Philips Research Europe

Company Confidential until December/2021

Author's address: Andra Alexa
andra.alexa@philips.com

Title: Measurement methods for body fat (%) assessment from 3D Kinect scans

Author: Andra Alexa

Supervisors: Sebastiaan den Boer, Makiko Sadakata

Reviewer(s): Jason Farquhar

Technical Note: PR-TN-2016/–

Additional numbers:

Subcategory: Graduation Project

Project: 2010-0376 Healthy Selfie

Customer:

Keywords: 3D imaging, body fat(%), Kinect

Abstract: Body composition is a better health indicator than just scale weighting. Previous research internally conducted at Philips showed that a model based on full 3D body representation can accurately measure body fat(%) (RMSE = 2.22%) on pregnant women. The current study investigates the plausibility of performing body fat assessment from single 3D depth-maps, thus an incomplete 3D representation of human body. A Kinect v2 device was used for data acquisition and two predictive models based on hand crafted features extracted from the point cloud data have been developed. A Lasso regression lead to a 3 parameter fat prediction model with $adj - R^2 =$ 0.72 and RMSE = 8.02%. A multivariate linear regression with a stepwise elimination routine resulted into a predictive model with adj-R^2 = 0.60, and RMSE = 9.85%.

Conclusions: Our predictive model provides a feasible approach for fat mass measurement using Kinect single depth maps. However, an RMSE of around 8% may fall on the edge of the acceptable error range for a market product. To implement our approach into a smart gadget, the predictive power has to be enhanced by improving both the features extracted and predictive models. Nevertheless, the method deserves great merit for further research as the prospect of a home body composition monitoring device addresses a great audience concerned about physical well-being.

Management summary

Body composition depicts the amount of water, bone, muscle and fat mass in the human body. It is considered a better health indicator than just weight measurement as any imbalance between these elements are associated with health risks. Often the term is used to distinguish between fat mass and fat-free mass. Body fat is also the prevalent focus of the study. In the context of technological development, novel, convenient, and cost efficient measurement methods are necessary to ease tracking of fat mass variations and satisfy a demanding market. In this way, people can easily track their body composition variations in their own home.

Recent research internally conducted at Philips (see PR-TN-2015/00174) successfully showed that 3D optical imaging can be used to develop accurate measurement techniques for fat mass assessment in pregnant women. The fat prediction model was based on perimeter values around thigh, waist, arm and neck. Although very promising, the approach holds two main drawbacks: (1) it requires a second person and a lot of room to walk around and scan the body and (2) demands laborious manual processing of data analysis as the anatomical locations were manually annotated.

The objective of this study is to address these shortcomings by proposing new data acquisition and analysis processes. As previous research performed fat prediction on full 3D models acquired with Structure Sensor (an iPad mounted 3D scanning device), we aim at conducting body fat assessment on single depth-images rather than a full 3D body scan. For the purpose of current study, a different device has been used, namely Kinect v2. Two main benefits derived from the switch. The data was returned in the format we need, as a single depth map returning the distance in mm from the camera plane which solved the issue of an additional person necessity. In addition, a toolbox with a dedicated Kinect module performs a skeletal tracking, thus the manual annotation problem was solved.

To address these issues, 38 participants, all Philips interns and employees, have been measured. Firstly, BIA data (body fat percentage) has been collected using a commercially available scale. Thereafter, 5 body scans have been acquired, with the participants facing 5 different (predetermined) angles towards the Kinect v2 plane. The scans and the corresponding fat labels were processed and used to build a fat prediction model. Two predictive models based on hand crafted features extracted from the point cloud data have been developed. A Lasso regression lead to a 3 parameter fat prediction model with $adj - R^2 = 0.72$ and RMSE = 8.02%. A multivariate linear regression with a stepwise elimination routine resulted into a predictive model with $adj - R^2 = 0.60$, and RMSE = 9.85%.

Contents

1 Introduction

1.1 Background

1.1.1 Body Composition

Body composition is a term that generally describes the amount of water, bone, muscle and fat mass in the human body (Blakemore et al., 2001). Body composition is a better health indicator than just scale weighting as any imbalance between these components such as dehydration or excessive fat is associated with health risk. Although not technically correct, often the term refers to the 2C model (two-component model to discern between fat mass and fat-free mass) (Astorino et al., 2012). Body fat is also the prevalent focus of the study.

Obesity is an old, but still very relevant topic. The associated risks have been documented over the last decades and includes, for example, that *'in the age group 45-50 the death rate increases roughly 1% for each pound of excess'* (Newburgh, 1942; Sloan et al., 1962). Excessive adipose tissue is related to health risks such as hypertension, cardiovascular diseases, stroke, metabolic syndrome or diabetes (Janssen et al., 2004; Sohlstroem et al., 1993). This comes at high health care costs, therefore keeping fat retention under control has benefits both in terms of care spending as well as long-term quality of life.

1.1.2 Measurement methods

To prevent excessive weight gain it is necessary for people to firstly be aware of their body fat retention. Several measurement methods have been developed. A more detailed description is given in TN2012/00016 and PR-TN-2015/00174 [1]. For an extensive overview of all assessment methods please refer to Wells and Fewtrell (2006).

Hydrostatic weighing is considered one of the most accurate methods for assessing body composition (Warner et al., 1986). Estimation of body fat is derived from body volume using the formula shown in (Brožek et al., 1963). Body density is calculated by measuring water displacement (Sloan et al., 1967) of fully immersed participants into a tank of water at 35 °C using the principle of Archimedes. Although highly accurate, it is costly, which make it suitable for experimental settings, but not for daily use.

Magnetic Resonance Imaging is a more accurate method than hydrostatic weighing with a relative standard deviation of 1.5% when estimating the body adipose tissue (Sohlstroem et al., 1993). However, accuracy of the method is shadowed by high maintenance costs of MRI which is also an unsuitable method for home user.

Caliperometry involves hypodermic fat measurement at specific anatomical locations by pinching the skin. The pressure applied by a spring-loaded caliper releases water from the tissue and measures the thickness of the skinfold. Several formulas are available to derive body fat from skinfold measurements. The method allows for easy and high accuracy fat assessment (Cross et al., 2011; Taggart et al., 1967), which makes the caliper a widely used device for research and clinical purposes. However, the necessity of a specialized person and restriction to only subcutaneous fat are its main disadvantages.

Bioelectrical Impedance Analysis (BIA) measures the body resistance of a weak electrical current (800A, f=50 kHz) through the body tissues. The electrical impedance of tissue is directly related to the body water. The method works on the principle that fat-free mass is a better conductor than fat mass. A BIA device is basically a weight scale with integrated electrodes which measure the head-to-foot water resistance of the electrical current. BIA derived percent body fat is highly correlated with BMI (Ranasinghe et al., 2013), making it suitable for clinical practice and home use. However, it's accuracy is affected by level of hydration and humidity.

Dual-energy X-ray absorptiometry (DXA) scan is considered the most accurate fat test available which can assess body composition for each square inch by using X-ray scans. However, the method is not recommended for

[1] Philips internal technical notes

frequent use as it's costly, a full-body scan session takes about 10 minutes and involves usage of X-ray radiation, which poses way more health risks than fat itself.

All these techniques have both advantages and disadvantages depending on purpose of use and equipment accessibility. However, most of them require trained personnel and expensive equipment. For the home user, an accurate, quick, cost effective and convenient method is needed. A novel approach is gaining ground, namely 3D imaging methods.

1.1.3 3D imaging

3D imaging is the general term for graphics that use three-dimensional (Cartesian) description of geometrical data to build 3D digital models. 3D data can be of different types: (1) obtained from overlapping stereo image pairs (mostly used in movies industry and it's basically an 3D illusion from 2D images) or (2) obtained directly in 3D format using 3D depth cameras and is usually represented as a point cloud.

Several 3D imaging technologies that use depth cameras are available on the market. Few examples are Google Tango [2], Structure Sensor [3] Microsoft Kinect Sensor used for Xbox or Intel Perceptual which is similar to Microsoft Kinect but designed for a less Microsoft-centered environment. This technology is currently used for state of the art development in gaming (Zhang, 2012), computer vision and industrial design (Izadi et al., 2011) but it could also serve home health monitoring (at the moment, home health fat monitoring devices using 3D depth cameras are not available on the market).

Astorino et al. (2012) was the first study to investigate body composition measured with DXA scan from 3D scans acquired with Textile/Clothing Technology Corp. ([TC]) 3D body scanner (Cary, NC, USA). Although still not as reliable as traditional methods, due to its relatively low costs and portability 3D imaging is a promising technique for home use.

1.1.3.1 Motivation

Given that body volume is correlated with fat mass and 3D scanners can create an accurate 3D (volumetric) representation of the body, fat related features can be extracted from 3D scans. Not only volumetric information, but also body proportions can be depicted. Significant correlations between abdominal fat and waist/hip ratio have been shown (Sohlstroem et al., 1993). Several studies proposed features hiding in 3D body scans such as curvature (Laws et al., 2006) or full body analysis using depth maps (Suau Cuadros et al., 2013; Zhang et al., 2014).

1.1.3.2 3D scanners

Out of the available depth cameras, only two are of interest for the given project: Structure Sensor which was used in an underlying study, and Kinect which is used in current approach.

Structure Sensor (SS) is an iPad mounted 3D scanner (Occipital, USA) that uses structured light for acquiring depth information. Skanect [4] is a dedicated software used for data acquisition. The operating mode is simple: A person has to walk with the scanner around an object to capture the geometry of its surface. An LCD projector spans the physical space with Infrared (IR) light patterns which distort when hitting the object. A depth camera with 640x480 resolution catches the pattern distortion and computes the distance from camera plane to each pixel in the field of view. A color camera can also be used for colorization. The depth values represent the 3^{rd} dimension of the digital model. The volumetric depiction of the object is created by calibrating a high number of depth maps acquired from a 360 °view points. The acquisition frame rate is 30 frames/s.

Kinect One also referred to as Kinect v2. sensor (Microsoft, USA) is probably the most well-known depth camera on the market. The sensor contains an RGB camera for color and a depth camera consisting of IR projector and IR camera. In addition, a multi-array microphone can give information about 3D motion capture (Zhang, 2012). To produce a depth image, each pixel of the ToF camera encodes the distance to the analogous point in space. The IR projector spans the physical environment and the IR camera detects the reflected IR component. The distance of each pixel in space is derived by measuring the phase difference between the radiated and reflected IR waves (Hansard

[2] For details check: https://developers.google.com/project-tango/
[3] For details check: http://structure.io/
[4] Official web page: http://skanect.occipital.com/

et al., 2012). The depth camera has a resolution of 512 x 424 pixels and a frame rate of 30 frames/s. A more through technical description is given in (Pagliari and Pinto, 2015).

1.1.3.3 Terminology

In 3D imaging, the notions of depth map (DM), point cloud (PCL) and coordinate system are widely used, thus some relevant clarification follows.

A *depth map* (DM) is a 3D representation of the environment from a single view point. It captures information relating to distances of surface objects from a predefined plane to capture the geometry of surfaces. Its just like a normal picture, which adds up a 3^{rd} dimension which represents the depth distance of object in space.

A *point cloud* (PCL) is also a 3D representation of the external surface of the object in three-dimensional coordinate system. The point cloud is the full set of points a 3D scanner has measured. In the case of Kinect, the point cloud data is derived from a single depth map, while in the case of SS, the point cloud of an object is derived from multiple depth maps.

A *collection of points* is a subset of points extracted from the original PCL.

We will also refer to a *full 3D model* of an object as the 3D surface obtained from the point cloud by unifying the points in a 3x3 neighborhood via triangulation.

A point cloud can be represented in *3D pixel coordinates* or in *3D world coordinates*. Each pixel has an x and y position within an image. The pixel indices are integers which range from 1 to the length of camera resolution. Therefore, the 3D pixel coordinates of an image contains the (x, y) pair of pixel indices and the corresponding depth value z. Although this representation is already providing information about the depth values within the image, the actual position in m (or any other unit measure) relative to the camera, thus 3D world coordinates, may better serve measurement purposes. The transformation from 3D pixel coordinates to 3D world coordinates is dependent on the field of view (FOV) and resolution of the camera.

1.2 Objective

The objective of this study is to develop an automated routine for body fat measurement from single depth maps. The method should provide the incentives of an efficient and cost effective method for the home user to track fat retention and help preventing the associated health risks.

1.2.1 Healthy Selfie. Problem description

The premises of the current project are partly based on a previous study internally conducted at Philips, which assessed the feasibility of 3D imaging as a technique to estimate maternal fat mass in a sample of pregnant women. The findings are detailed in PR-TN-2015/00174.

Here was proven that depth cameras can be used to perform a better fat prediction on pregnant women than BIA. 55 subjects were scanned using the SS and the volumetric representation of the body was depicted as a PCL. A four-parameter predictive model was built based on circumference values around thigh, waist, arm and neck derived from the full 3D body model. The model had an average $RMSE = 2.22\%$ $(Adj - R^2 = 0.89, F(4, 17) = 34.02, p < 0,001)$ which compares favorably with BIA model ($RMSE = 3.80\%, Adj - R^2 = 0.68, F(5, 16) = 6.67, p < 0.001$).

Although very promising, the method holds few shortcomings which makes it unfeasible for an end product: (1) an additional person and a lot of space is required to acquire the scan, which is not a customer-friendly approach, and (2) data analysis requires the use of manually annotated anatomical locations (thus analysis needs to be automatized).

1.2.2 Hypothesis and research question

In this study we aim to further improve the findings of the previous study by proposing a more convenient method for home use. To produce an accurate prediction, instead of a full 3D body model, separate depth images, thus single-view-point point clouds, may be sufficient. The main advantage of such an approach is that the participant can take just a self-scan. In addition, this would reduce data dimensionality and save expensive computation time. However, such a drastic viewpoints reduction comes at the cost of less accurate human body representation.

Thus the current study revolves around the following research question: Can single depth maps be used to build an automatic routine for body fat assessment?

1.2.3 Solution approach

To answer the research question a new experiment with a different approach was performed. A Kinect camera was used instead of Structure Sensor as it better serves the purposes of the study: the camera is static, thus no rotation around the scanned object is necessary (this allows the user to self-capture a full body scan) and returns only single depth maps of the environment. In addition, the Kinect has a skeleton tracking function which solves the problem of manual anatomical annotation.

Participants will be scanned using a Kinect depth camera and fat mass data will be acquired using a commercially available BIA scale. After data acquisition and data preprocessing, a set of hand crafted features will be extracted from the point cloud, mainly anthropometric measurements. The features will be paired with the corresponding fat values and a predictive model will be trained using supervised techniques.

2 Methods

2.1 Participants

38 Philips interns and employees were recruited internally. No special inclusion criteria such as gender, weight or nationality were set. Subjects were instructed about the purpose and procedure of the study prior to participation, which was exclusively voluntary. Part of the instruction was also to wear casual clothing, preferably as tight as possible. Bulky clothing was removed as well as all accessories and badges. Testing consisted of a single measurement session that last for about 20 minutes. No monetary compensation was provided.

2.2 Study design

The study was conducted in one of the lab facilities at Philips Reasearch in High Tech Campus, Eindhoven. All test sessions were performed in the same conditions. The set-up consisted of a Kinect device situated on a table 125 cm height. A 100 x 100 cm wooden base situated 170 cm away from the Kinect camera. The base had some lines, 60 cm apart from each other, drawn to guide participants' foot placement. The Kinect device was connected to a laptop (from which data acquisition commands were executed). A scheme is given in Figure 2.1. A BIA scale was placed on a solid, wooden surface. The experiment was conducted in artificial light condition, on moderate intensity.

Data acquisition consisted of 3 parts. Firstly, demographic information such as age, gender, nationality and height were acquired. Thereafter, participants were instructed to sit barefooted on a BIA scale while keeping the scale leads in their hands. The scale returned information regarding weight, total body water, muscle mass and fat mass (per total and per body parts including hands, legs and torso). Finally, 5 Kinect scans were acquired in different poses.

For the Kinect scans, all participants were instructed to sit on the base with limbs spread apart. The legs had to follow the guiding lines drawn on the base. If necessary the hair had to be tight. The scans were collected from 5 predetermined body angles: 0°, 45°, 90°, 75°and 180 °relative to the camera, as we wanted to acquire as much 3D information as possible. Per each pose, 300 repeated measurements have been acquired with the intention to later perform signal-to-noise averaging. With an acquisition rate of 30 frames/sec this resulted into 10 seconds of continuous scanning where participants had to stand motionless in front of the camera. In the final model only 0°(frontal) scans were used.

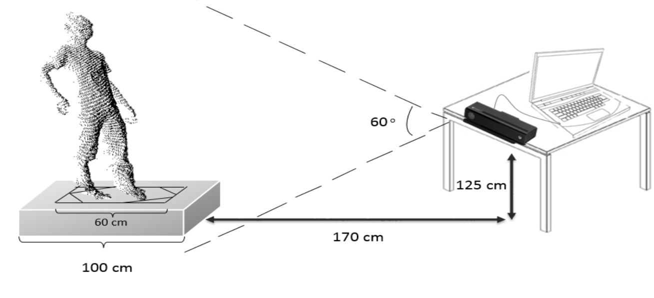


Figure 2.1: 3D scanning set-up scheme.

2.3 Hardware and Software

Hardware used for the study design consisted of a stand-alone, commercially available Kinect One (Microsoft, USA) device with 512x424 pixels time-of-flight camera. A four-lead Ironman BC-558 (Tanita, USA) BIA scale was used for acquiring fat mass data.

Kinect One requires at least an Windows 8 operating system and Matlab 2016a. Image Acquisition Toolbox and Kinect SDK v2. were used for data collection. Analysis was performed in Matlab 2015b.

2.4 Data analysis

The problem we had to solve is a prediction problem when the ground truth (fat %) is known. The conceptual scheme of a prediction with supervised learning adapted to our approach is given in Figure 2.2. These steps were followed in our analysis methodology and are further discussed.

2.4.1 Data description

The raw data acquired from 3D scanning consisted of:

- a depth map
- depth metadata

The depth map is a 512x424 matrix containing the pixels' depth values (in mm) of the objects in physical space spanned by IR projector, thus the human body and the surrounding environment.

The depth metadata contains information about the corresponding depth map such as body joints coordinates and a bit map indicating body position. The skeleton tracking function is able to trace human movement and returns the skeleton joints position in 3D pixel coordinates. The bit map is a 424x542 matrix with values of 1 for pixels that belong to the person, and 255 for pixels belonging to the surrounding environment.

2.4.2 Preprocessing

Before building the predictive model, some preliminary processing was necessary in order to have a good foundation for feature extraction.

Firstly, the PCL of the full scene was extracted from the depth map as a (512x424)x3 matrix where every (x,y,z) pair represents a pixel (x,y) and the corresponding depth value (z). Thereafter, the PCL that describes the person is extracted by matching the bit map coordinates with the full-scene coordinates. Data smoothing and removal of isolated points was performed by applying a median filter on the bit map matrix where each z-value takes the median value of the 3-by-3 neighborhood. At this stage, a selection of the body PCL in pixel coordinates is obtained, but 3D world coordinates is a more informative metric.

Finally, a coordinate transformation from 3D pixel coordinates to 3D world coordinates was performed as shown in Kowsari and Alassaf (2016). The transforming formulas are given in the Appendix A.1. The width and height range had to be scaled upon the corresponding field of view (FOV) value of the depth camera: 60°vertical and 70°horizontal. The values have been gathered from Pagliari and Pinto (2015).

2.4.3 Feature extraction

By having obtained a clean PCL of the human body in 3D world coordinates, we proceed towards feature extraction. A set of 557 hand-crafted features has been computed from the PCL for each participant. A list of the full feature

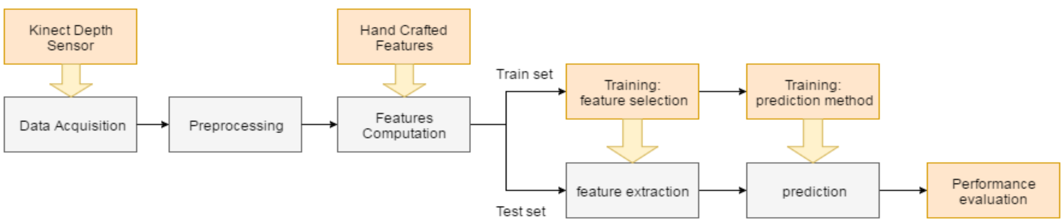


Figure 2.2: Conceptual scheme of the supervised prediction model

set is given in Appendix, Table A.1. The features consist of body measurements, ratios, curvature and combinations between these at specific anatomical locations: thigh, lower hip, hip, waist, upper waist, chest area and arms. In defining the anatomical locations, we made use of joint coordinates. Kinect v2 can track up to 25 joints as shown in Appendix, Figure A.4. As waist measurements are commonly used in assessing fat mass, we will often refer to area *j1j2* as the lumbar area, i.e. the vertical line between joint 1(lower hip) and joint 2 (upper waist).

Features extracted were:

- *height*: by computing the difference between min and max value of the y-coordinate from the body PCL in 3D world coordinates.
- *anthropometric measurements*: Euclidean Distance of the fitted polynomial for selected slices (see Appendix, Figure A.1) as well as width[1] of body area at specific anatomical locations (thigh, waist, torso, arm). The degree of the fitted polynomial varies, as to capture a better surface description: on thighs and arms 2^{nd} degree poly-fit has been performed while on torso and belly area, a 4^{th} degree polynomial would give a more accurate description of the surface. A visualization is given in Appendix, Figure A.2.
- *body ratios*: different combinations of measurement ratios which may be correlated to body fat (including waist-to-hip ratio).
- *bone length*: length of relevant bones (femur, arm) used mainly for scaling
- *body curvature*: which has been shown to relate to fat percentage and is the amount by which a line or surface deviates from being a flat plane. Curvature is the reciprocal of the radius of the fitted circle. An example is given in Appendix, Figure A.3.

2.4.4 Predictive model

The predictive model was built based on features extracted from human body PCL which were paired for each participant with the corresponding BIA fat measurements (%) which was used as gold standard.

Prediction models are just as good as the features provided. Given the large feature space relative to the limited number of observations (38 participants), feature selection is necessary, i.e. determine a smaller set that exhibits the strongest result. There is an empirical rule that, in order to avoid over-fitting (tunning the model too much on training set and having a poor performance on test set) there should be an estimator for every 10 measurements.

According to Guyon and Elisseeff (2003) there are 3 main categories of feature selection algorithms: filters, wrappers and embedded methods. In our approach, we train two multivariate regression models: (1) a Lasso reduction technique as it is one of the most common and well-documented embedded method for feature selection and (2) a stepwise regression as it is the most popular wrapper method in traditional statistics.

Lasso regression aims to minimize not only the ordinary least squares (OLS) but also penalizes model complexity by minimizing the absolute value of the regression coefficients (L1-norm) in which many of them become 0. A detailed explanation follows in Appendix A.2.

Stepwise elimination routine is a reduction method for linear regression that begins with no terms in the model and adds/removes regressors according to their statistical significance, i.e. an entrance tolerance for coefficients with p-values lower than 0.05 and exit tolerance for coefficients with p-value higher than 0.10. A new model is trained on each feature subset via a greedy search algorithm until no more improvement is possible.

In our final prediction approach, we trained two different regression models with the two reduction techniques presented above. Model performance is evaluated with respect to RMSE (Root Mean Squared Error) and $adj - R^2$ (coefficient of determination adjusted for model complexity) values. For a realistic estimate, a 60/40 partition into train/test set has been applied on entire dataset. This ratio has been chosen as we want to have a robust model and we need more test data for verification via bootstrapping. Model performance (RMSE and $adj - R^2$) of chosen models are reported on both test set, but evaluation is done only on the test set. Robustness has been verified by comparing it with 1000 models based on bootstrap data from test set.

[1] By width we account for the length of the surface area exposed to the field of view of the Kinect, at a given anatomical location. It's basically the same as ED, but ignoring the 3^{rd} dimension

3 Results

3.1 Demographics

38 participants (24 males and 14 females) aged between 21 and 57 (mean 27.26±6.7 years) were tested. No participants were excluded from the analysis. Average height was 174±8.06 cm and average total body fat(%) was 20.4±8.89% with major gender differences: 27.85±7.48% for women and 16.3±6.67% for men.

3.2 Data quality

The quality of selected human PCL is strongly dependent on the skeleton tracking function and correctness of joint coordinates. For 13 out of 38 participants, the fitted skeleton did not properly localized upper and lower limbs' joints which lead to an incomplete selection of human PCL. This means that some features could not be used, as the selection routine would displace the noisy and uninformative predictors. However, precise height measurement from Kinect scans is bounded by accuracy of skeleton fitting.

3.3 Height measurement

Extraction of human PCL in 3D world coordinates was used to measure the height of a person by computing the difference between the maximum and minimum y coordinate. The overall average correlation value between the tape-measured and PCL-extracted height for all participants is 0.74 ($p \leq 0.001$). Average measurement error is -1.12 cm (see box plot in Figure 3.1). Removing the 13 measurements with inaccurate limbs' joint annotation, leads to a correlation of 0.96 ($p \leq 0.001$). This means that on a clean scan, a height measurement with 96% accuracy can be performed.

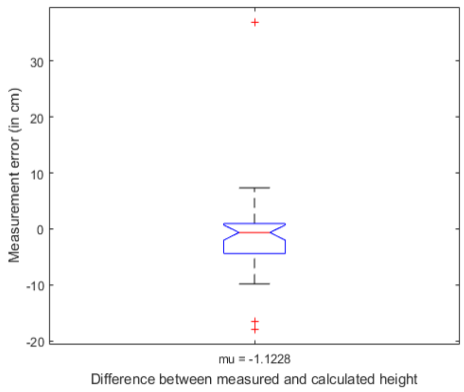


Figure 3.1: Box plot describing the distributional characteristics of PCL-extracted height compared to actual height of participants. Average measurement error is -1.12 cm with a slight tendency towards smaller values. A few outliers stand out.

3.4 Fat prediction model

The two fat prediction models were trained on 60% of the data, and validated on the remaining 40%. The performances of the selected models in regard to $adj - R^2$ and RMSE values on both train and test set are summarized in Table 3.3.

3.4.1 Lasso model

Lasso regression is a reduction routine where feature selection is part of the model construction process (embedded method). 100 Lasso models have been trained on 60% of the data (train set) with 10-folds cross validation (see Appendix, Figure A.5). The optimal λ value (shrinkage parameter within one standard error) is automatically chosen by the model. The final selected model was the one with best cross-validated $adj-R^2$ value ($adj-R^2$ = 0.85, RMSE = 8.86%). The features space was reduced to only 3 predictors which are shown in Table 3.1. On the remaining 40% of the data (test set), the model had $adj - R^2 = 0.72$ and RMSE = 8.02%. Robustness of the model was checked via comparison with a set of 1000 models based on bootstrap sampling on test set (see Appendix, Figure A.6; explanations follow in figure description).

Feature name	Description
EDHchestV	ED of the vertical chest section adjusted for participant height
EDboneLRatchestVj1j2	ratio between the ED vertical chest section and j1j2 distance
widthboneLRatj1j1j2	ratio between waist width and j1j2 distance

Table 3.1: Features selected by Lasso fit in the final model. The feature names are code names used in the model. ED stands for Euclidean distance and j1j2 is the distance between joints j1(lower hip) and j2(upper waist) that define the lumbar area (see Appendix, Figure A.4)

3.4.2 Stepwise elimination routine

Stepwise elimination is a wrapper method for feature selection that adds or deletes regressors at each step in regard to the p-enter criterion. A bidirectional stepwise model selection has been trained on 60% of the dataset and validated on the remaining 40% ($adj - R^2 = 0.83$, RMSE=3.73% on training set). The features were selected based on their statistical significance ($p - value \leq 0.5$) and final model was reduced to 4 predictors shown in Table 3.2. On the test set, the predictive model has $adj - R^2 = 0.60$, and RMSE = 9.85%. Robustness has been examined by comparing it with 1000 models based on bootstrapped test set data (see Appendix, Figure A.6).

Feature name	Description
boneLj1j2	lumbar area length
EDRatchestVwaist	ratio between the ED of chest vertical section and ED of waist
EDRatchestVaR	ratio between ED of chest vertical section and ED of arm
EDwidthRathipj2	ratio between ED of hip area and width of upper waist

Table 3.2: Features selected by Stepwise fit in the final model. The feature names are code names used in the model. The features selected were those with $p - value \leq 0.5$. ED stands for Euclidean distance and j1j2 is the distance between joints j1(lower hip) and j2(upper waist).

	$adj - R^2$ train set	RMSE train set	$adj - R^2$ test set	RMSE test set
Lasso regression	0.85	8.86%	0.72	8.02%
Stepwise regression	0.83	3.73%	0.60	9.85%

Table 3.3: Summary of Lasso and Stepwise regression results, reported on both train set and test set.

4 Discussion

The aim of this project was to predict body fat (%) on single depth maps acquired from Kinect. To our knowledge, no study so far made such an attempt, thus our method extends the pool of possible applications of Kinect devices and similar scanners. Compared to full 3D model, fat prediction from single depth maps is a more convenient method for home use. Our method addressed the limitations of previous study, namely the requirement of an additional person for data acquisition and automatic joint annotation for data analysis. However, our approach comes at the cost of information loss such as volumetric representation and full 3D body patterns.

Therefore, there is no surprise that our models have higher prediction error (RMSE = 8.02% for Lasso model and RMSE = 9.85% for Stepwise regression) compared to the full 3D model (RMSE = 2.22%), mainly due to less informative features. RMSE, as an absolute measure of fit, shows the standard deviation of the random component in the model. Lasso regression has a standard error of 8.02% (RMSE value) which means that an 8.02% error rate in fat prediction (%) is expected, explaining 72% of the variability. The stepwise model had a slightly higher RMSE value (9.85%), which means that a higher error rate is expected with a lower explained variability ($adj - R^2 = 0.60$ on test set).

Lasso regression that we chose for the final predictive model has a better performance than the Stepwise model: higher $adj - R^2$ values (see Table 3.3) which means that it explains more variance and a lower RMSE value, therefore a lower predictive error. Also, the Stepwise regression model has a low RMSE on test set (3.73%), but a higher value on test set(9.85%), which means that the model may overfit on train data. RMSE values of Lasso regression on test and train set are close which proves that the model does not overfit.

To better asses our model performance within existing techniques, accuracy of other measurements methods are reported as follows: MRI scans have 1% accuracy, underwater weighing has 1.5% and DEXA scans have 2% accuracy. However, these are used only in clinical practice, expensive and time-consuming. For home use, a smart scale has around 6% accuracy, and BMI around 9%. Therefore, our Lasso model outperforms BMI and can approach the performance of smart scales. At this stage, the accuracy of the fat prediction method we propose has been limited by other factors which are going to be further discussed, but with great potential for improvement.

4.1 Limitations

First of all, measurement errors can be related to the devices we used. Kinect scanner is prone to measurement flaws as is limited by its hardware specifications. Camera resolution, for example, restricts the accuracy of geometric surface of scanned objects. Substantial improvement has been achieved during transition from Kinect v1 to Kinect v2 and the technology is being further advanced. Another important aspect is sensitivity to environmental lightning. Complete black surfaces (as clothing for instance) can absorb the IR light and therefore, some areas of a surface cannot be detected. Thus, avoidance of dark black wear is highly recommended. In addition, BIA data can also be prone to errors due mainly to body water variations.

Thereafter, software related boundaries can interfere. Our method is dependent on a Kinect dedicated skeleton tracking function implemented in Matlab. Most of the times, a good skeleton fit can be tracked, but this require some motion for calibration. However, as in the case of 13 participants in our study, sometimes joint annotation faces slight oversights. In addition, Kinect v2 is restricted to a Microsoft environment, although similar scanners are available.

Finally, participants-related limitations can also be addressed. Subjects were tested wearing casual clothing with bulky items removed. This prevented us from having a precise representation of body surface exposed to the scanner. For very clean scans, participants should wear tight clothes and even a cap to prevent artifacts due to coarse hair. In addition, the number of participants we manage to scan restricted the number of predictors we could use in our models. An increased sample size will also allow us to add more features without running the risk of overfitting, but this require more observations to be processed.

4.2 Future research

Further research should focus on enhancing the predictive power. In the current approach we used two linear methods as we assumed there is a linear dependence between measured fat (%) and extracted features. However, if the underlying data is non-linear, a non-linear prediction can lead to better results as for instance an SVM with Recursive Feature Selection (RFS). Is a suitable embedded feature selection method which we did not include in the analysis.

Predictions from different classifiers could also be collected into an ensemble that would boost overall performance. Also, feature engineering is an essential step in identifying informative fat predictors (or even physical markers) in 3D scans. Such methods come at the cost of a more thoroughly expertise requirement, an increased number of observations and sometimes interpretability loss.

Future research should also focus on optimal trade-off between user convenience and prediction accuracy. This revolves around the question: "Is 8% fat prediction accuracy good enough?". Awareness over accepted error rate in an end product allows for a more insightful assessment of model limitations. Such a fat prediction method can be offered either as a service, i.e. data would be sent to the cloud, thereafter results would be sent back to the user, or as a product that would perform scan processing and fat prediction on the devices' CPUs. In this case, a computationally efficient method is necessary.

Our predictive model provides a feasible approach for fat mass measurement using Kinect single depth maps. However, an RMSE of around 8% may fall on the edge of the acceptable error range for a market product. To implement our approach into a smart gadget, the predictive power has to be enhanced by improving both the features extracted and predictive models. Nevertheless, the method deserves great merit for further research as the prospect of a home body composition monitoring device addresses a great audience concerned about physical well-being.

References

Todd Astorino, Julien Baker, Steve Brock, Lance Dalleck, Eric Goulet, Robert Gotshall, Alexander Hutchison, M Knight-Maloney, Len Kravitz, James Laskin, et al. Three-dimensional body scanning as a novel technique for body composition assessment: A preliminary investigation. *Journal of Exercise Physiologyonline*, 15(1), 2012.

Colin Blakemore, Sheila Jennett, and Sheila Jennett. *The Oxford companion to the body*, volume 7. Oxford University Press New York, 2001.

Josef Brožek, Francisco Grande, Joseph T Anderson, and Ancel Keys. Densitometric analysis of body composition: revision of some quantitative assumptions. *Annals of the New York Academy of Sciences*, 110(1):113–140, 1963.

P Cross, A Shim, A Verhulst, A Trost, R Schmidt, R Lynde, and P Hauer. Assessing various body composition measurements as an appropriate tool for estimating body fat in division i female collegiate athletes. *Journal of Science and Medicine in Sport*, 14:e93, 2011.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012.

Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011.

Ian Janssen, Peter T Katzmarzyk, and Robert Ross. Waist circumference and not body mass index explains obesity-related health risk. *The American journal of clinical nutrition*, 79(3):379–384, 2004.

Kamran Kowsari and Manal H Alassaf. Weighted unsupervised learning for 3d object detection. *arXiv preprint arXiv:1602.05920*, 2016.

Joseph Laws, Nathaniel Bauernfeind, and Yang Cai. Feature hiding in 3d human body scans. *Information Visualization*, 5(4):271–278, 2006.

LH Newburgh. Obesity. *Archives of Internal Medicine*, 70(6):1033–1096, 1942.

Diana Pagliari and Livio Pinto. Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors. *Sensors*, 15(11):27569–27589, 2015.

Chathuranga Ranasinghe, Prasanna Gamage, Prasad Katulanda, Nalinda Andraweera, Sithira Thilakarathne, and Praveen Tharanga. Relationship between body mass index (bmi) and body fat percentage, estimated by bioelectrical impedance, in a group of sri lankan adults: a cross sectional study. *BMC Public Health*, 13(1):1, 2013.

A We Sloan et al. Estimation of body fat in young men. *Journal of Applied Physiology*, 23:311–315, 1967.

AW Sloan, JJ Burt, and CS Blyth. Estimation of body fat in young women. *Journal of Applied Physiology*, 17(6): 967–970, 1962.

Annica Sohlstroem, Lars-Olof Wahlund, and Elisabet Forsum. Adipose tissue distribution as assessed by magnetic resonance imaging and total body fat by magnetic resonance imaging, underwater weighing, and body-water dilution in healthy women. *The American journal of clinical nutrition*, 58(6):830–838, 1993.

Xavier Suau Cuadros et al. Human body analysis using depth data. 2013.

Nan R Taggart, Ruth M Holliday, WZ Billewicz, FE Hytten, and AM Thomson. Changes in skinfolds during pregnancy. *British Journal of Nutrition*, 21(02):439–451, 1967.

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

JG Warner, Rachel Yeater, Lauralee Sherwood, and K Weber. A hydrostatic weighing method using total lung capacity and a small tank. *British journal of sports medicine*, 20(1):17–21, 1986.

JCK Wells and MS Fewtrell. Measuring body composition. *Archives of disease in childhood*, 91(7):612–617, 2006.

Qing Zhang, Bo Fu, Mao Ye, and Ruigang Yang. Quality dynamic human body modeling using a single low-cost depth camera. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–683. IEEE, 2014.

Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.

A Appendix

A.1 Coordinate transformation

A coordinate transformation from 3D pixel coordinates to 3D world coordinates has been performed as shown in Kowsari and Alassaf (2016) by following Equations A.1, A.2, A.3, A.4 and A.5 where:

- fov_x, fov_y are fields of view of x and y range, respectively (in radians)
- $scale_x$ and $scale_y$ are scaling factors
- W_x,W_y and W_z are the 3D world coordinates
- P_x,P_y and P_z are the 3D pixel coordinates
- R_x and R_y is the resolution of x and y range, respectively

$$scale_x = 2 * tan(\frac{fov_x}{2}) \tag{A.1}$$

$$scale_y = 2 * tan(\frac{fov_y}{2}) \tag{A.2}$$

$$W_x = p_z * scale_x * (\frac{P_x}{R_y} - 0.5) \tag{A.3}$$

$$W_y = P_z * scale_y * (\frac{P_y}{R_x} - 0.5) \tag{A.4}$$

$$W_z = P_z \tag{A.5}$$

The fov_x and fov_y for Kinect v2 are 60°and 70°, respectively (Pagliari and Pinto, 2015). The transformation in radians has been performed by following equation: $Rad = \frac{Degrees * \pi}{180}$.

The origin of the pixel coordinates system is in the top-left corner, while the origin of the 3D world coordinates system is in the middle point of the picture. The translation is carried over during the transformation.

A.2 Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) is a regularization technique used for performing linear regression that penalizes model complexity via a regularization parameter λ (Tibshirani, 1996). λ is a penalty term that constraints the size of the estimated coefficients (aims at minimizing the absolute value of the regression coefficients - L1 norm). It is a model selection and dimensionality reduction technique. For a nonnegative parameter λ, Lasso regression solves the following problem:

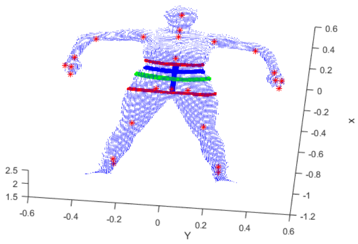
$$min_{\beta_0,\beta}(\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j1}^{p}(\mid \beta_j \mid) \tag{A.6}$$

where:

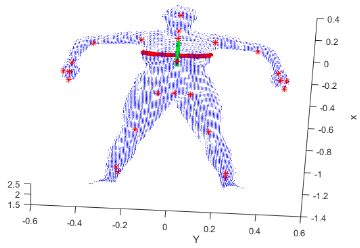
- n : the number of observations
- y_i: the response at observation i
- x_i: the data (feature matrix), a vector of p values at observation i
- λ: shrinkage parameter
- β_0,β: scalar and coefficients of regressors

The bigger the λ value is, the bigger the number of zero components of β is, so the penalization is stronger.

A.3 Features extracted

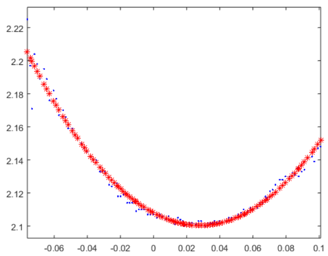


(a) Lumbar area

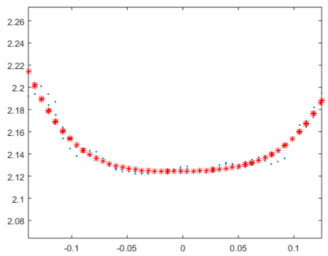


(b) Torso area

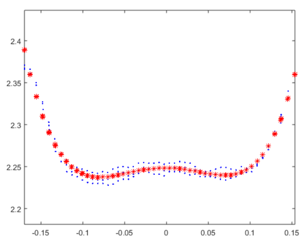
Figure A.1: Slices (sections) at (a) lumbar area: upper waist, waist, hip, lower hip, upper waist-symphysis line and (b) torso area: horizontal and vertical slice. Red circles represent tracked joints.



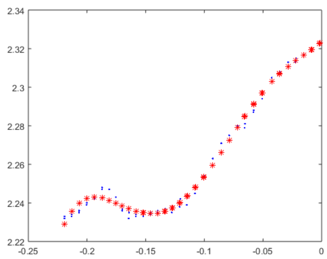
(a) Thigh section: 2^{nd} degree polynomial fitted



(b) Upper waist section: 4^{nd} degree polynomial fitted

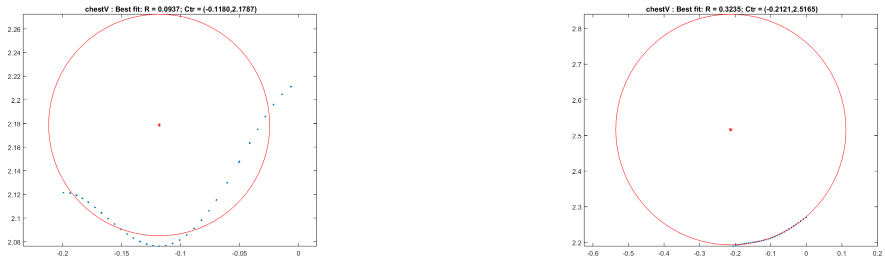


(c) Torso vertical section: 4^{nd} degree polynomial fitted



(d) Torso horizontal section: 4^{nd} degree polynomial fitted

Figure A.2: Fitted polynomials on selected slices used for feature computation, namely euclidean distances at given anatomical location of body region exposed to the Kinect.



(a) (b)

Figure A.3: Circle fit on sagital section over torso area scaled to the magnitude of the fitted circle. The fitted circle over a flat chest (b) is bigger, thus has a higher value of the radius, therefore lower curvature ($curvature = \frac{1}{radius}$).

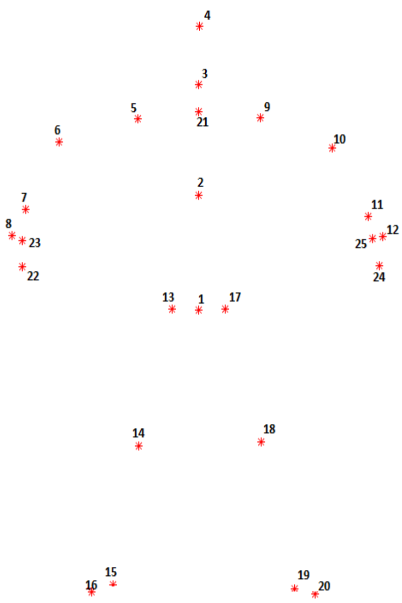


Figure A.4: Skeleton joints numbering as tracked by the Kinect v2. Often we will refer to area j1j2 as to the lumbar area situated between joint 1 and joint 2.

Feature name	Field (code) name	Description
h	height	height extracted from pcl
ED	j1	ED of J1 slice (4^{nd} degree poly-fit)
	j2	ED of J2 slice (4^{nd} degree poly-fit)
	waist	ED of waist slice (4^{nd} degree poly-fit)
	hip	ED of hip slice (4^{nd} degree poly-fit)
	j1j2	ED of waist-symphysis slice (4^{nd} degree poly-fit)
	chestV	ED of chest vertical slice (4^{nd} degree poly-fit)
	chestH	ED of chest horizontal slice (4^{nd} degree poly-fit)
	tL	ED of thigh left slice (2^{nd} degree poly-fit)
	tR	ED of thigh right slice (2^{nd} degree poly-fit)
	aL	ED of arm left slice (2^{nd} degree poly-fit)
	aR	ED of arm right slice (2^{nd} degree poly-fit)
width	j1	width of slice at J1
	j2	width of slice at J2
	waist	width of slice at waist line
	hip	width of slice at hip
	j1j2	width of slice at J1-J2 (longitudinal) line
	chestV	vertical width (i.e. length) of chest line
	chestH	horizontal width (i.e. length) of chest line
	tL	width of thigh left line at middle bone location
	tR	width of thigh right line at middle bone location
	aL	width of arm left line at middle bone location
	aR	w idth of arm right line at middle bone location
boneL	j1j2	length (in m) of lumbar area extracted from pcl
	chest	length (in m) of chest area
	tL	length of left femurs extracted from pcl
	tR	length of right femur
	aL	length of left arm bone
	aR	length of right arm bone
EDH	same as ED	ED fields divided by subject height ⇒ 11 features
widthH	same as width	width fields divided by subject height ⇒ 11 features
EDRat	combination of ED	all ED fields ratios ⇒ 11*11 = 121 features
widthRat	combination of width	all width fields ratio ⇒ 11*11 = 121 features
EDwidthRat	combination of ED & width	ED & width fields ratio ⇒ 11*11 = 121 features
EDboneLRat	combination of ED & bone	ED & boneL fields ratio ⇒ 11*6 = 66 features
widthboneLRat	combination of width & bone	width & boneL fields ratio ⇒ 11*6 = 66 features
curveR	j1	radius of circle fitted on j1 slice points
	j2	radius of circle fitted on j2 slice points
	waist	radius of circle fitted on waist slice points
	hip	radius of circle fitted on hip slice points
	j1j2	radius of circle fitted on waist-symphysis slice points
	chestV	radius of circle fitted on vertical chest line
	chestH	radius of circle fitted on horizontal chest line
	tL	radius of circle fitted on thigh left points
	tR	radius of circle fitted on thigh right points
	aL	radius of circle fitted on arm left points
	aR	radius of circle fitted on arm right points

Table A.1: Full feature set extracted from the PCL of a single participant. *ED* stands for Euclidean Distance, *width* is same as ED ignoring the 3^{rd} dimension, *boneL* is bone length, and *curveR* is the radius of the fitted curve. *j1, j2* represent joints of the skeleton as shown in Figure A.4

.

A.4 Fat prediction

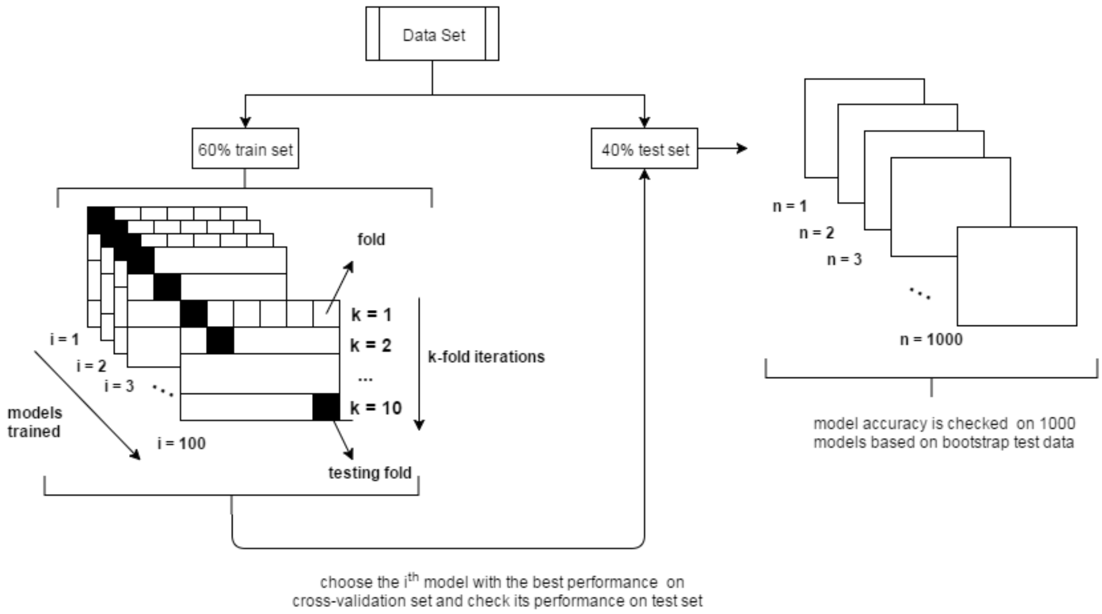
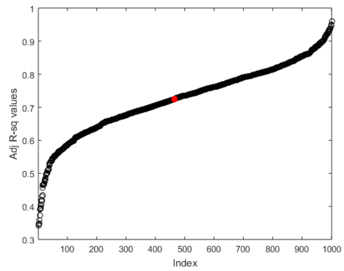
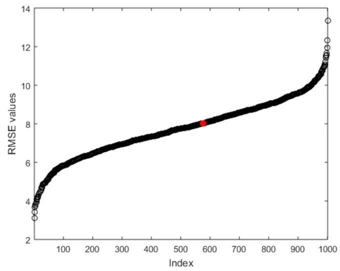


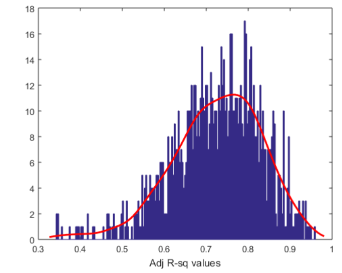
Figure A.5: Model selection performed on Lasso regression. 100 Lasso models have been trained on 60% of the data (train set) with 10-folds cross-validation. The model selected was the one with the best cross-validates performance (highest $adj-R^2$). The model has thereafter been tested on the remaining 40% of the data. Robustness of the model performance has been verified by comparing the results on test set with 1000 bootstrapped models on test data.



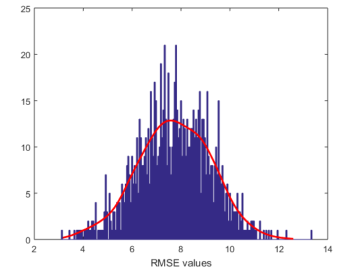
(a) Bootstrapped Adj R-sq values



(b) Bootstrapped RMSE values

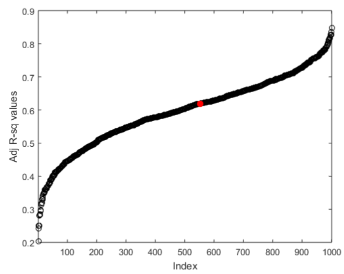


(c) Bootstrapped Adj-Rsq distribution

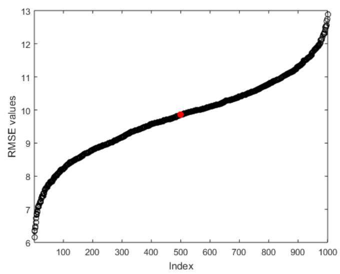


(d) Bootstrapped RMSE distribution

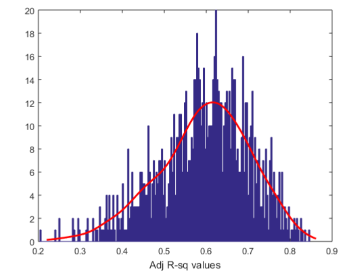
Figure A.6: Bootstrapped Lasso models created on 1000 random sampling (with repetition) from the test set. Red circles represent the $adj-R^2$ and RMSE values of original model. Black circles represent $adj-R^2$ and RMSE values of bootstrapped models. The figures show that better or worse performances than those already reported are possible, but the accuracy of our model is placed roughly in the peak of the normal distribution.



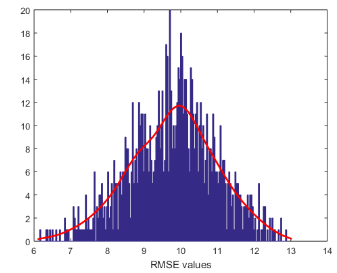
(a) Bootstrapped Adj R-sq values



(b) Bootstrapped RMSE values



(c) Bootstrapped Adj-Rsq distribution



(d) Bootstrapped RMSE distribution

Figure A.7: Bootstrapped Stepwise models created on 1000 random sampling(with repetition) from the full dataset. The RMSE and $adj-R^2$ values on the 1000 bootstrap samples we tested our model on, can be better or worse than the reported performance of the model, but follow a normal distribution. The accuracy of our model is placed roughly in the peak of the distribution.