Body composition prediction by locally weighted and Bayesian networks modeling

Simiao Tian

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture
Agriculture
Alimentation
Food
Biologie
Biology
Environnement
Environment
Santé
Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

Spécialité : Statistiques

présentée et soutenue publiquement par

Simiao TIAN

le 29 Novembre 2013

Body composition prediction by locally weighted and Bayesian networks modeling

Prédiction de la composition corporelle par modélisation locale et les réseaux bayésiens

Directeur de thèse : **Jean-Baptiste DENIS**
Co-encadrement de la thèse : **Béatrice MORIO / Laurence MIOCHE**

Jury

M. François HUSSON, Professeur, Laboratoire de math-appliqués, Agrocampus Ouest	Rapporteur
M.Régis HANKARD, Professeur, U1069 Nutrition Croissance Cancer, Inserm	Rapporteur
M. Léopold FEZEU, Maître de conférence, UMR U557, Inserm	Examinateur
Mme. Christelle GUILLET, Maître de conférence, UFR Médecine, Université d'Auvergne	Examinatrice
M. Stéphane ROBIN, Directeur de recherche, UMR 518, AgroParisTech	Examinateur
Mme. Laurence MIOCHE, Chargée de recherche, UMR1019 Nutrition Humaine, INRA	Examinatrice
Mme. Béatrice MORIO, Directrice de recherche, UMR1019 Nutrition Humaine, INRA	Examinatrice
M. Jean-Baptiste DENIS, Directeur de recherche, UR MIA-Jouy, INRA	Examinateur

AgroParisTech
UR MIA-Jouy / UMR 1019 Nutrition Humaine
INRA, Domaine de Vilvert 78352 Jouy-en-josas / 58 rue Montalembert 63009 Clermont-Ferrand

Summary

The assessment of human body composition is important for evaluating health and nutritional status. Among health issues, overweight and obesity are worldwide problems. Increased fat mass, especially in the trunk location, has been associated with an increased risk of metabolic diseases, such as type 2 diabetes and cardiovascular disease. The lean body mass, especially appendicular muscle mass, is also directly related to health and particularly with the mortality rate. Also, aging is associated with substantial changes in body composition. Reduction in body lean or body fat-free mass occurs during aging (Kyle *et al.*, 2001) together with an increase of body fat related to accumulation of adipose tissues, particularly in abdominal region (Kuk *et al.*, 2009); therefore assessing these changes in segmental body composition may be important because the study will lead to a pre-diagnosis for the prevention of morbidity and mortality risk. Accurate measurements of body composition can be obtained from different methods, such as underwater weighing and dual-energy X-ray absorptiometry (DXA). However, their applications are not always convenient, because they require fixed equipment and they are also time consuming and expensive. As a result, they are not convenient for use as a part of routine clinical examinations or population studies. The potential uses of statistical methods for body composition assessment have been highlighted (Snijder *et al.*, 2006), and several attempts to predict body composition, particularly body fat percentage (BF%), have been made (Gallagher *et al.*, 2000a; Jackson *et al.*, 2002; Mioche *et al.*, 2011b).

The first aim in this thesis was to develop a multivariate model for predicting simultaneously body, trunk and appendicular fat and lean masses from easily measured anthropometric covariables. We proposed a linear solution published in the *British Journal of Nutrition*. There are two main advantages in our proposed multivariate approach. The first consists in using very simple covariables, such as body weight and height, because these measurements are easy and not expensive. The usefulness of waist circumference is also investigated and combined with age, height and weight as predictor variables. The second advantage is that the multivariate approach enables to take into account the correlation structure between the responses into account, which is useful for a number of inference tasks, *e.g.* to give simultaneous confidence regions for all the responses together. Then the prediction accuracy of the multivariate approach is justified by comparing with that of the available univariate models that predict body fat percentage (BF%). With a good accuracy, the multivariate outcomes might then be used in studies necessitating the assessment of metabolic risk factors in large populations.

The second aim in this thesis was to study age-related changes in segmental body compositions, associated with anthropometric covariables. Two Bayesian modeling methods are proposed for the exploration of age-related changes. The main advantage of these methods is to propose a surrogate for a longitudinal analysis from the cross-sectional datasets. Moreover, the Bayesian modeling enables to provide a prediction distribution, rather than a simple estimate, this is more relevant for exploring the uncertainty or accuracy problems. Also we can incorporate the previous findings in the prior distribution, by combining it with the datasets, we could obtain more suitable conclusions.

The previous predictions were based on models supposing any correlation structure within the variables, the third aim in this thesis was to propose a parsimonious sub-model of the multivariable model described by a Gaussian Bayesian network (GBN), more precisely Crossed Gaussian Bayesian Networks (CGBN). One of the advantages of the Bayesian networks formulation is to allow non-statistician, typically expert of one domain, to enter into their mechanism

through the easy understanding directed acyclic graphs (DAGs) presentations. The hope is to obtain a better multivariable prediction than with a plain linear regression model. The idea is applied using structured DAGs when the set of nodes is the product of two series of items. This novel statistical method is applied to the prediction of segmental body compositions adding anthropometric covariables, such as height and weight. The results show that CGBNs globally perform better than the saturated model according to Standard Error of Prediction. In addition, the reduction of the parametric dimension, with respect to the saturated model is striking, especially for the variance parameters. We demonstrated that, at least for GBNs modelling, it was possible to introduce a known structure on the set of variables of interest, and that can lead to very effective results to obtain an interpretable predictive formula.

Résumé

La composition corporelle est importante pour évaluer l'état de santé et le statut nutritionnel d'individus. Le surpoids et l'obésité deviennent des problèmes de santé à l'échelle mondiale. L'accroissement de la masse grasse, notamment celle du tronc, a été associée à une augmentation du risque de maladies métaboliques, telles que le diabète de type 2 et les maladies cardiovasculaires. La masse musculaire, en particulier appendiculaire, est également un indice de santé, et est liée au taux de mortalité. En outre, le vieillissement s'accompagne de changements importants dans la composition corporelle. La masse maigre diminue (Kyle *et al.*, 2001) et la masse grasse augmente, liée à une accumulation de tissus adipeux, en particulier dans la région abdominale (Kuk *et al.*, 2009). Il est donc important d'étudier ces changements en fonction de l'âge pour tenter d'établir un pré-diagnostic et aider à la prévention de la morbidité et de mortalité. La composition corporelle se mesure par différentes méthodes, telles que le pesage sous l'eau ou l'absorption bi-photonique à rayons X (DXA). Cependant, ces méthodes de mesure ne sont pas adaptées pour des populations de taille très grande, car elles nécessitent un équipement fixe, demandent des manipulations longues et sont coûteuses. En revanche, le potentiel de méthodes de prédiction statistique a été mis en évidence pour estimer la composition corporelle (Snijder *et al.*, 2006), et plusieurs modèles ont été proposés pour prédire la composition corporelle, notamment le pourcentage de la masse grasse (BF%) (Gallagher *et al.*, 2000a; Jackson *et al.*, 2002; Mioche *et al.*, 2011b).

Le premier objectif de cette thèse est de développer un modèle multivarié à partir de covariables anthropométriques pour prédire simultanément les masses grasse et maigre de différents segments du corps. Pour cela, nous avons proposé une régression linéaire multivariable publiée dans le *British Journal of Nutrition*. Notre proposition multivariée présente deux avantages principaux. Le premier avantage consiste à utiliser les covariables très simples que sont l'âge, le poids et la taille dont la mesure est facile et peu coûteuse. L'utilité d'ajouter comme covariable le tour de taille a été évaluée. Le deuxième avantage est que l'approche multivariée prend en compte la structure de corrélation entre les variables, ce qui est utile pour certaines études d'inférence où on s'intéresse à des fonctions des variables prédites. La qualité de la précision multivariée a été évaluée par comparaison avec celle des modèles univariés déjà publiés. Nous avons montré que la prédiction multivariée est bonne et que notre approche peut donc être utilisée pour des études de risques métaboliques en grandes populations.

Le second objectif de cette thèse est d'étudier l'évolution de la composition corporelle au cours du vieillissement, en tenant compte des covariables anthropométriques. Deux modélisations bayésiennes ont été retenues et développées. Un des avantages principaux de nos propositions est, grâce à une modélisation, de réaliser une analyse longitudinale à partir de données transversales. En outre, la modélisation bayésienne permet de fournir une distribution prédictive, et non pas une simple valeur prédite, ce qui permet d'explorer l'incertitude de la prédiction. Également, des résultats antérieurs ou publiés peuvent être incorporés dans la distribution priore, ce qui conduit à des conclusions plus précises.

Les prédictions précédentes sont fondées sur des modèles où la structure de corrélation entre les variables est laissée libre, le troisième objectif de notre travail a été d'imposer une structure de corrélation particulière adaptée au problème. L'avantage est l'utilisation d'un sous-modèle parcimonieux du modèle multivarié précédent. Cette structure est décrite au moyen d'un réseau bayésien gaussien (GBN). Le principe est d'obtenir une prédiction multivariée plus robuste car basée sur l'estimation d'un plus petit nombre de paramètres. Cette idée est

mise en œuvre en utilisant des graphes orientés acycliques (DAG) structurés lorsque l'ensemble des nœuds est le produit cartésien de deux ensembles. Nous avons appelé, réseaux bayésiens gaussiens croisés (CGBN), les réseaux bayésiens obtenus. Cette nouvelle méthode statistique a été appliquée à la prédiction de la composition corporelle déjà réalisée. Les résultats montrent que les CGBNs donnent une prédiction plus précise que le modèle multivarié saturé selon l'erreur standard de prédiction. En outre, par rapport au modèle saturé, la dimension paramétrique est diminuée de manière remarquable, en particulier pour les paramètres associés à la variance. Nous avons donc montré qu'il était possible d'introduire une structure connue sur l'ensemble des variables d'intérêt, aboutissant à des résultats efficaces et proposant une formule prédictive facile d'interprétation. La modélisation par les GBNs est avantageuse pour les non-statisticiens, notamment les expert d'un domaine, parce que leurs connaissances s'y incorporent facilement et qu'elle facilite la compréhension à travers les présentations par DAG.

Acknowledgements

This thesis is the fruit of three years of research that would not have been carried out without support.

First and foremost, I would like to express my deep gratitude to my three advisors Dr. J-B. Denis, Dr. L. Mioche and Dr. B. Morio for their time, effort, motivation, enthusiasm, and immense knowledge. Their trust and patience during the thesis were really appreciated. I owe a special thanks to Dr. J-B. Denis who has always taken the time to help me, when I had some questions.

I am gratefull to all the members of my Ph.D dissertation committee : Pr. Y. Boirie, Dr. J-M. Chardigny, Dr. M. Duclos, Dr. H. Monod, Dr. E. Kuhn, Dr. C. Bidot, Dr. S. Robin, whose help and feedback allowed me to clarify my ideas and greatly improve this work.

My sincere thanks also goes to my colleagues from Applied Mathematics and Informatics (MIA) Unit at INRA-jouy and Humain Nutrition Unit (UNH) at INRA-Clermont for creating a propitious work environment.

I would like to acknowledge the financial support of the department MIA-Jouy and UNH-Clermont.

Last but not the least, my particular gratitude is dedicated to my parents Ping Tian and Hua Li, and my brother Silei, none of this would have been possible without their immense encouragement. I would like to acknowledge the sacrifices made by my parents and thank them for their unwavering support. This dissertation is dedicated to my parents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Body composition is an important indicator for evaluating healthy and nutritional status. It is a consequence of biological and non-biological factors such as genetic[1], processes of aging, lifestyle and socio-economic level (Sobal and Stunkard, 1989; Winkleby *et al.*, 1992; Fezeu *et al.*, 2006; Yen and Moss, 1999; McLaren, 2007).

The assessment of body composition is essential in health study. Indeed, body composition study allows not only to better understand the pathophysiology of many diseases, but also to monitor disease following-up and to help guide treatment. Nutritional status is a result from the interaction of body composition, energy balance and body functionality. Also body composition is the best long-term indicator of nutritional status (Bedogni *et al.*, 2006). It is of interest to clinicians and researchers because of its association with body functionality. Some body composition indices were potentially useful to diagnose and monitor the course of certain kind of mal-nutrition. VanItallie *et al.* (1990) proposed a fat-free mass (FFM) index (FFMI = FFM/height^2) and a body fat mass (BFM) index (BFMI = BFM/height^2). They showed that these new indices can be expected to provide more meaningful information about nutritional status.

By measuring body composition, a person's health status can be more accurately assessed and the effects of both dietary and physical activity programs better directed. Besides that, body composition assessment can be widely used in many applications such as (Heyward and Stolarczyk, 1996) :

- Identify individual health risk associated with segmental body composition;
- Monitor changes associated with specific diseases that alter body composition;
- Assess the effectiveness of nutrition programs and exercise interventions;
- Estimate ideal body weight and formulate dietary recommendations and exercise prescriptions;
- Investigate the relationship between body composition and increased morbidity and mortality, and between body composition and decreased function in the elderly;
- Monitor growth, development, maturation and age-related changes in body composition;
- Formulate interventions to prevent chronic diseases later in life.

[1]Gender is a very important genetic factor, and it is easy to get.

1.1 Background

1.1.1 Body composition research and measurement methods

Body Mass Index (BMI) is generally considered the good way to assess overweight and obesity (Committee *et al.*, 1995; Keys *et al.*, 1972), and it is calculated as weight in kilograms divided by height in squared metres (kg/m^2). The most commonly used classification of adult normal weight, overweight and obesity according to BMI is :

Table 1.1: Classification of adult normal weight, overweight and obesity according to BMI.

BMI	Classification
<18.5	Underweight
18.5-24.9	Normal weight
25.0-29.9	Overweight
30.0-34.9	Obese class I (Moderately obese)
35-39.9	Obese class II (Severely obese)

In adults, BMI levels above 25 are associated with a high risk of morbidity and mortality, with BMI levels of 30 and greater indicating obesity (Chumlea *et al.*, 2000). Besides, BMI has been studied extensively for its potential in predicting risk of premature death, disease and disability, therefore BMI over 25 is considered as a risk factor for several non-communicable diseases such as : cardiovascular diseases (mainly heart disease and stroke), diabetes, cancer and etc. It is found that BMI correlates with cardiovascular risk factors, and it can be considered as a surrogate measure of cardiovascular risk factor (Freedman *et al.*, 2001; Ice *et al.*, 2009). Above 25, BMI is associated strongly and positively with mortality attributed to diabetes (Lancet, 2009). Higher BMI value is associated with a significant increase in the risk of cancer (Calle *et al.*, 2003). For instance, among postmenopausal women in the UK, 5% of all cancers (about 6000 annually) are attributable to being overweight or obese (Reeves *et al.*, 2007). Furthermore, the authors concluded that the risk of death from cardiovascular disease, cancer or other diseases increased with heavy weight, regardless of age or gender. The risk of type 2 diabetes has also been linked to BMI, with research demonstrating that the relative risk increases for every additional unit of BMI over 22 (Colditz *et al.*, 1995).

Proxy for body fatness, a significant advantage of BMI is the availability of extensive national reference data and its established relationships with levels of body fatness, morbidity, and mortality in adults (Committee *et al.*, 1995). However, BMI has some limitations, such as that it does not take into account age and gender factors. In adults, even if they have the same BMI, women are more likely to have more body fat than men. In addition, individual who possesses a great amount of muscle mass may be classified as overweight or obese, when in reality they are healthy. Despite of these limitations, BMI is still a popular tool because of its simplicity and cost-efficiency. Body composition measurements enable to overcome these limitations.

Body composition research began during the 1940s and a variety of methods has been introduced to quantify body composition (Behnke, 1942). Most body composition methods are based upon the model in which the body consists of two distinct compartments, fat and fat-free (Brožek *et al.*, 1963). During the derivation of the two compartment model, four compartment model has been proposed (Keys and BroŽek, 1953), and these four groups are respectively water, protein, bone mineral, and fat. The two and four compartment models served as the basis upon which all body composition methods were developed. The available measurement methods range from simple to complex with all methods having limitations and some degree

of measurement error. The clinical significance of the body compartment to be measured must be determined before a measurement method is selected, as the more advanced techniques are less accessible and more costly.

The assessment of body composition involves the use of multicompartment models that are not readily available in clinical practice and epidemiological research. Indirect methods, *i.e.*, methods making use of predictive algorithms, are often used in these settings. Anthropometry is the single most universally applicable and inexpensive body composition method and is of great importance because of its association with health status. Even though not yet a gold-standard technique, dual-energy x-ray absorptiometry (DXA) holds significant promise for the assessment body composition in clinical practice and epidemiological studies.

Anthropometric measurements play an important role in clinical practice. Besides BMI, waist circumference is considered as as powerful predictor of type 2 diabetes. Waist circumference greater than 102 cm for men and 88 cm for women lead to higher risk of type 2 diabetes (Wei *et al.*, 1997). Waist-to-hip ratio (WHR) and waist-to-height ratio are used as surrogates for body fat centralization (Gallagher *et al.*, 1996; Pouliot *et al.*, 1994). WHR higher than 0.9 for men and 0.8 for women have been associated with cardiovascular risk factors (Ko *et al.*, 1997; Hsieh and Yoshinaga, 1995). However, recent studies show that waist circumference has better potential than WHR for assessing health risks (Dobbelsteyn *et al.*, 2001; Chan *et al.*, 2003; de Koning *et al.*, 2007), even though there is often no significant difference between waist circumference and WHR in the accuracy of risk factor prediction. Therefore, the use of waist circumference is widely recommended in prevention and management of risk factors. Despite of safety, cost-effectiveness, convenience for the patient and ease of use, one main limitation of this anthropometric approach is the reduced ability to differentiate levels of fatness and leanness among individuals.

Bioelectrical impedance analysis (BIA) is a commonly used method for measuring body composition (Kyle *et al.*, 2004). BIA determines the electrical impedance, or opposition to the flow of an electric current through body tissues which can then be used to calculate an estimate of total body water (TBW). TBW can be used to estimate fat-free body mass and, by difference with body weight, body fat (Kyle *et al.*, 2004). The impedance index is proportional to the volume of total water and is a predictor variable in regression equations to predict body composition. BIA method has become popular owing to its ease of use, portability of the equipment and its relatively low cost. BIA is useful in describing mean body composition for individuals, however large errors for an individual limit its clinical application, especially among obese people. Moreover, since BIA method has been seldom applied to overweight or obese population, the available BIA prediction equations are not necessarily applicable to overweight or obese children or adults.

Dual energy X-ray absorptiometry (DXA) is the most popular method for quantifying fat, lean, and bone tissues and DXA technique is accepted as a noninvasive measurement method that can be applied in humans of all ages. The two low-energy levels used in DXA and their differential attenuation through the body allow the discrimination of total body adipose and soft tissue, in addition to bone mineral content and bone mineral density. A typical whole-body scan takes approximately 2 mins and exposes the subject to <5 mrem of radiation. The repeatability is also very high for all reported total body measures. The precision is about 1% (standard deviation) for percent fat and 2% (coefficient of variation) for total fat and lean mass measures (Lohman and Chen, 2005; Leonard *et al.*, 2009). Moreover, with the ease of use, availability, and safety of DXA, there is much interest in using the technology for studies of catabolic diseases, obesity, and bone density. Reference populations have been scanned and

defined by sex, ethnicity, and age. The largest study for body composition in the United States was the National Health and Nutrition Survey (NHANES) that scanned 22000 participants from 8–85 years old since 1999.

The advantages of DXA include good accuracy and reproducibility, and provides for the assessment of regional body composition and nutritional status in disease states and growth disorders. DXA is in broad clinical use worldwide in a variety of settings from radiology departments to exercise/ physiology labs. Also it is one of the few methods with a large amount of reference population data, *e.g.*, the NHANES, the Supplémentation en Vitamines et Minéraux Antioxydants (SU.VI.MAX) study (Hercberg *et al.*, 2004). Nevertheless, DXA technique has some limitations, such as that the scanning bed has an upper weight limit and the whole-body field-of-view can not accommodate very large persons; DXA yield a small amount of radiation. Additionally, some studies indicate that DXA may not be as reliable in extreme populations, including the obese. Despite these limitations, DXA is a widely used and convenient method for measuring body composition, owing to its ease of use and availability, also it is currently included in the ongoing large survey.

These "direct" body composition measurement methods have their own advantages and disadvantages, but one common limitation of them is that they are impractical in large studies. Time-consumming and expensive for some measurements, these sophisticated methods require the fixed equipement in a clinical setting. Thus, statistical methods for predicting body composition have been developed, and are more applicable for determining the body composition of large studies in a nonlaboratory setting. Generally, statistical prediction of body composition relies on predictor variables, such as easily acquired anthropometric variables. The rationale for the use of such variables is based on the high multiple correlations and low standard errors of prediction found between variables and the criterion body composition determined by more complex techniques. The statistical prediction can not only provide a current estimation of individual body composition based on observed anthropometric information, but also it enables to estimate an evolution of body composition associated with age-related changes in anthropometry, typically body weight. Indeed, the statistical prediction has significance in the follow-up studies, because this kind of studies is often not pratical in clinical setting. In subsection 1.1.2, we will give a brief introduction of statistical prediction in the general framework.

1.1.2 Statistical prediction

Prediction is the processus by which based on available dataset a model is created or choosen to try to best predict a value or a probability of an outcome. The variables to be predicted by a model are the dependent variables and the variables used in the model to achieve the prediction are predictor variables. The precision of a predictive model refers to its performance within the sample from which it is derived, whereas the accuracy is a measure of the performance of a predictive model when it is applied to an independent sample (Sun and Chumlea, 2005). There are several factors affecting the precision and accuracy of a predictive model, such as the precision of the measured values of the predictor and dependent variables, the statistical relations among the predictor variables and between the predictor variables and the dependent variable, the statistical methods used to formulate the model, and the size and nature of the sample.

The root mean square error (RMSE) is a measure of the precision of a predictive model (Hyndman and Koehler, 2006). The RMSE value can be standardized for the mean value of the dependent variable. This standardized value is useful in comparing the predictive models with different dependent variable and different units (Hyndman and Koehler, 2006). The coefficient of determination, denoted by R^2, is the proportion of the total variance in the dependent variable that is explained by the predictor variables in the model. The larger the R^2 value, the better the model fits the data (Nagelkerke, 1991).

The generalization ability of a predictive model is related to its ability to predict on independent datasets. Its study can be conducted by a validation process. More precisely, a validation process consists of comparing the model results to observations in some way. There is internal validation, when models are checked against data used to develop them, and external validation when models are tested against external independent datasets. External validation is the necessary, by contrary internally validated models can fit the dataset well but perform poorly on novel dataset (called overfitting) (Justice *et al.*, 1999; Steyerberg *et al.*, 2001; Bleeker *et al.*, 2003). Another use of validation process consists of comparing prediction accuracy of different models and of making model selections.

To assess the prediction accuracy, a set of "Standard Error of Prediction k" (SEPk) and "Relative Error of Prediction k" (REPk) are used, they read :

$$SEPk = \left[\frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|^k\right]^{\frac{1}{k}} \tag{1.1}$$

$$REPk = \left[\frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - \widehat{y_i}}{y_i}|^k\right]^{\frac{1}{k}} \tag{1.2}$$

where y_i is the observed value and $\hat{y}_i$ is the corresponding predicted value. For instance, SEP2 is the usual prediction error. Compared to SEP2, the criterion SEP1 is preferable in prediction application, because it is less sensitive to some outlier subjects. For instance, we build a statistical model from a training dataset, it happens that there is only one subject whose prediction strays enormously from its observed value, therefore this important difference leads to a great SEP2. Nevertheless as a matter of fact, the accuracy of the prediction by this model is not so unsuitable according to the SEP2, because by excluding the subject concerned (*e.g.*, outlier somehow), we could obtain a relevant SEP2 which is used to justify the quality of the prediction. Thus, to reduce the effect of outliers without eliminating it, it is usual to consider SEP1, because it is more stable. It is worth noting that in general between the models, if one

model yields a greater SEP2 value compared to others, it also yields a greater SEP1 value.

With respect to REPk criteria, they have the advantage of being scale independent, and so are frequently used to compare prediction performance across different datasets (Hyndman and Koehler, 2006).

In the prediction framework, we attempt to select a predictive model with a minimum SEP (or REP) value, however there is a lower boundary that the SEP value can not exceed. Indeed, let us assume that there exists a mathematical relationship between the predicted variable y and a set of covariables $\mathbf{X}$ as that $y = f(\mathbf{X}) + \varepsilon$ where $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$. A set of possible models $\mathcal{F}$ from a training dataset $\mathcal{D} = \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \ldots, (\mathbf{X}_n, y_n)\}$, *i.e.*, a linear model, a Bayesian linear model, a linear discriminant model, etc. The aim is to find a suitable model $\hat{f} \in \mathcal{F}$ which yields a minimal prediction error. For a subject i, the expected prediction error is (Hastie *et al.*, 2009, p. 223) :

$$\begin{aligned} E\Big\{(y_i - \hat{y}_i)^2 | \mathbf{X} = \mathbf{X}_i\Big\} &= E\Big\{(y_i - \hat{f} + \hat{f} - \hat{y}_i)^2 \Big| \mathbf{X} = \mathbf{X}_i\Big\} \\ &= \sigma_\varepsilon^2 + \text{Bias}^2(\hat{f}(\mathbf{X}_i)) + \text{Var}(\hat{f}(\mathbf{X}_i)) \\ &= \text{intrinsic variance} + \text{Bias}^2 + \text{Variance} \end{aligned} \tag{1.3}$$

The first term is the variance of the target around its true mean $f(\mathbf{X}_i)$, and cannot be avoided no matter how well we estimate $f(\mathbf{X}_i)$, unless imposing $\sigma_\varepsilon^2 = 0$ which is not very realistic. The second term is the squared bias, the amount by which the average of our estimate differs from the true mean; the last term is the variance; the expected squared deviation of $\hat{f}(\mathbf{X}_i))$ around its mean. Typically the more complex we make the model $\hat{f}$, the lower the (squared) bias but the higher the variance.

In practice, we attempt to build a statistical model which could yield a SEP value close to the intrinsic variance. Contrary to easy computation of SEP, the determination of intrinsic variance is difficult, because we don't know the real joint distribution of $(\mathbf{X}, y)$. However, a intuitive way is to take a subset of k "closest" subjects of the predicted subject in a reference dataset, then to calculate the mean $\widehat{\mu}_{y_0}$ and the variance $\widehat{\sigma}^2_{y_0}$ of this subset, finally we can obtain an empirical distribution of $(\mathbf{X}_0, y_0)$ based on $\widehat{\mu}_{y_0}$ and $\widehat{\sigma}^2_{y_0}$.

Now we will give a formula to calculate intrinsic variance of a random variable y :

$$\text{intrinsic.var}(y) = E(|y - \mu_y|) \tag{1.4}$$

The use of "| |" in equation (1.4) is to have consistency with the formula SEP1. This equation allows us to determine the intrinsic variance of each subject in a dataset and to approximate a lower boundary of SEP1. If there exists a relevent number[2] of repetitions for a given value of $\mathbf{X}$ in the dataset, it is sufficient to apply the equation (1.4) to find the intrinsic variance. As it is not the case, we will create the fictive repetitions for each predicted subject according to the principle of "k nearest neighbors" :

1. First of all, a mathematical distance in the covariable space is proposed to define the neighborhood of the predicted subject i. All subjects in this neighborhood are considered to have similar characteristics with the predicted subject i, and those neighbor subjects will form a subset, denoted by s_i. This subset represents the fictive repetitions.

[2]Generally, a good estimation results from a sufficient sample size.

2. From s_i, we calculate the intrinsic variance of the variable y for the corresponding subject i.

3. Following this process, we will obtain an intrinsic variance for each predicted subject using equation (1.4).

Also, it is possible to quantify an average intrinsic variance of y in a training dataset with n observations, denoted by $\overline{\text{intrinsic.var}}$:

$$\overline{\text{intrinsic.var}}(y) = \frac{1}{n}\sum_{i=1}^{n} \text{intrinsic.var}(y_i)$$

The value of $\overline{\text{intrinsic.var}}$ enables to give an indication of the lower boundary at which we hope the prediction accuracy (*i.e.*, SEP) of the models can reach.

1.2 Thesis aims

The research on the prediction of body composition by statistical models has evolved since 1950's (Brožek and Keys, 1951; Jackson *et al.*, 1984). Most of statistical models attempt to predict body fat or percentage of body fat, since it is an important component in body composition research. Moreover, many predictive models using anthropometry as predictor variables are based on the concept of the two compartment model (body mass is divided into fat mass and fat-free mass), because the two compartment model involves an assumption of constant densities of fat mass and fat-free mass but this is not true in all individuals. As a result, prediction models are population-specific).

Appropriate statistical models must not only enable to yield an accurate body composition prediction, but also they should have a good generalization on independent datasets. Building model process involves data assessment, selection of predictor variables, choice of model types. When an appropriate statistical model is build, it will be applied to predict body composition in different contexts, such as anthropometric variation or aging. Therefore, this thesis will be organized by following three main aims :

Statistical multivariate prediction for segmental body compositions. The potential uses of statistical methods for body composition assessment have been highlighted (Snijder *et al.*, 2006), and several attempts to predict body composition, particularly body fat percentage (BF%), using linear models with simple predictor variables have been proposed (Levitt *et al.*, 2007; Gallagher *et al.*, 2000a; Jackson *et al.*, 2002; Mioche *et al.*, 2011a,b). Therefore, the first aim of the present study is to develop sex-specific multivariate models for estimating some segmental compartments of metabolic importance (*i.e.*, lean body mass, appendicular muscle mass and trunk fat) from age and easily accessible anthropometric variables, such as height and weight. The usefulness of including waist circumference is also investigated and combined with age, height and weight as predictor variables.

Age-related changes in body composition. Aging is associated with substantial changes in body composition. Reduction in body lean or body fat-free mass occurs during aging (Kyle *et al.*, 2001) together with an increase of body fat related to accumulation of adipose tissue, particularly in abdominal region (Kuk *et al.*, 2009). The loss of muscle mass, known as sarcopenia, may have a negative impact on physical activity, which could lead to a higher prevalence of metabolic abnormalities, as well as autonomous loss. Moreover, these changes in body composition lead to an increased prevalence of chronic metabolic diseases, particularly type 2 diabetes and cardiovascular diseases (Ezzati *et al.*, 2002). Therefore, the second aim of the present study is to study age-related changes in segmental body compositions, associating with anthropometric variables. The Bayesian and Frequentist modelings are respectively investigated for age-related change study. The main advantage of these methods is to allow conducting a longitudinal analysis from the cross-sectional datasets.

Bayesian methods for a parsimonious prediction. In the previous two studies, we assume that there is no correlation among several segmental body compositions. Now we will turn to more elaborated studies where we use the prior knowledges to define the dependences between some of segmental body compositions. The third aim of the present study is to propose a parsimonious sub-model of the multivariable model, which is used in the first aim, described by a Gaussian Bayesian network in order to obtain a better multivariable prediction than with a plain linear regression model. There are two key ideas : the first is to reduce the parametric dimension by using a directed acyclic graphs (DAGs)

presentation; the second is that a DAG can be build on a crossed structure between the variables to be predicted. This novel statistical method is applied to the prediction of segmental body compositions adding simple easy acquired covariables.

Chapter 2 will provide insight into different statistical methodologies for body composition prediction. The first section briefly discusses the data splitting, followed by a general description of locally weighted approaches. A brief literature reviews of Support Vector Machines Regression and Bayesian network are presented respectively in sections 3 and 4. Also in the end of section 4, a comprehensive development of a novel statistical method, crossed linear gaussian Bayesian network, will be given.

Chapter 3 will focus on the applications of proposed statistical methods. The first section briefly introduces the available datasets in the present study. The second section will concentrate on multivariate modeling for body composition prediction, and different statistical proposals will be described and compared for model selection. Also a published paper related to body composition prediction will be presented. The third section will discuss the applications for age-related change in body compositions. A general Bayesian and Frequentist modeling will be proposed to assess this issue.

The concluding chapter will discuss some encountered difficulties and expected results, make an enlightenment of the main contributions and suggest several recommendations for extending the current research in further works.

Chapter 2

Statistical tools

The statistical prediction for body composition is a good alternative of direct measurements in large studies. The accurate prediction model can be considered as a prognostic tool in the clinical setting when the directment measurement is lacking. This chapter provides insight on our proposed statistical models for body composition prediction, including an interesting discussion about data splitting in the data rich situation (section 2.1). As we are in a statistical prediction situation, to evaluate the model performance, it is recommended to split the full dataset into different parts, and each part will have its own usefulness. In the present study for body composition prediction, there are two main applied statistical approaches : the locally weighted approach and the Bayesian network approach. Locally weighted approach is a form of instance-based algorithm and its prediction is done by local functions which are using only a subset of the data. This approach involves the distance function (subsection 2.2.1) and the weighting function (subsection 2.2.2) that help select a subset of the data with strong similarity for a predicted subject. The locally weighted concept is used respectively in a linear model, Support Vector Machine (SVM) model and Bayesian linear model framework, but only SVM will be briefly introduced (section 2.3). Section 2.4 will focus on the Bayesian networks approach, particularly including the contribution of a novel modeling denoted *Crossed Gaussian Bayesian networks* for the multivariate prediction.

Otherwise, all application studies will be discussed in the next chapter, but it is worth mentioning that we are in data rich situation. For instance, one of main datasets at our disposal is an extracted NHANES dataset with more 3000 subjects[1]. Therefore, our proposed statistical modelings will be applied in data rich situation. For ease of understanding, it is also necessary to mension in advance that the predicted variables are the segmental body compositions, such as body fat, body lean, trunk fat and appendicular lean masses. In addition, for all application studies, we conduct a seperate statistical analysis on men and women.

2.1 Data splitting

The generalization performance of a statistical prediction model is closely related to its prediction capability on independent dataset. Assessment of this performance is extremely important in practice, since it enables to advise the model selection, and gives us a measure of the quality of the ultimately chosen model.

Data splitting is an important step in prediction model development process. A history of data splitting can be found in Stone (1974). The dangers of using the same data to both fit and select the model have been known for many years. Briefly, when using the same dataset

[1]All details about available datasets will be provided in section 3.1.

for model fit and selection, it will underestimate the true error of prediction and overestimate model generalization giving the preference to the more complex models. To avoid the over-optimism induced by using the same dataset, data splitting is a simple technique for dealing with it that was practically easy to manipulate (Faraway, 1995).

Normally dataset is divided into training, test and validation subsets to ensure good generalization ability of the model. Nevertheless, it may occur that data splitting introduces potential bias and variance into model development processus, for example, a data splitting method could allocate extreme observations into the training set, therefore, the test and validation sets contain fewer patterns compared to the training set. Consequently, the generalization ability of the model may be compromised and the trained model cannot be adequately validated. Therefore, the way the data is split has a significant impact not only on model selection, but also on the performance of the final chosen model.

To get a idea about usefulness of the stratification of data, we performed a small simulation study of prediction with a predictor variable X and a predicted variable Y. Also a third variable Z, either continuous or discrete, is used for splitting the dataset into two parts (one for training, another for validation), either randomly or using stratification. In the continuous case of Z, we assume that $(X, Y, Z) \sim \mathcal{N}(\boldsymbol{\mu}_{XYZ}, \boldsymbol{\Sigma}_{XYZ})$ with

$$\boldsymbol{\mu}_{XYZ} = (0, 0, 0) \quad \text{and} \quad \boldsymbol{\Sigma}_{XYZ} = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix};$$

in the discrete case of Z, we assume that $Z \in \{1, 2\}, Pr(Z = 1) = 0.5$ and $(X, Y)|Z \sim \mathcal{N}(\boldsymbol{\mu}_{XY|Z}, \boldsymbol{\Sigma}_{XY|Z})$ with

$$\begin{cases} \boldsymbol{\mu}_{XY|Z} = (10, -5) & \text{for} \quad Z = 1 \\ \boldsymbol{\mu}_{XY|Z} = (20, 5) & \text{for} \quad Z = 2 \end{cases}$$

and

$$\boldsymbol{\Sigma}_{XY|Z} = \boldsymbol{\Sigma}_{XY} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

The size of the dataset ranges from 8, 16, 32, 64 to 128, and for each size of the dataset, the simulation is repeated 1000 times. For each simulation within a given sample size, the random splitting is conducted by a random split-up of equal size in the whole dataset, while the stratified splitting is conducted by a split-up of equal size in the reordering dataset according to Z value (by defaut from minimum to maximum value), then a simple linear model is build on the training subset by considering X as the predictor variable and Y the predicted variable, finally the build model is used to predict Y in the validation subset with a Standard Error of Prediction (SEP) calculated. Table (2.1) presents some descriptive statistics of the 1000 SEP values for each size of the dataset. The results show that when size of the dataset is higher than 30, there is no difference of SEP between random splitting and stratified splitting.

If we are in a data-rich situation, the best approach for both problems is to randomly divide the dataset into three parts: a training set, a validation set, and a test set. The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model (Hastie *et al.*, 2009, p. 222). In the present study, we primarily started carrying on a stratified splitting on a covariate, which was calculated by two other covariates. It turned out that we overestimated generalization quality (quantified by SEP2 and REP2 criteria, *cf.* section 1.1.2) of the chosen prediction model, because under this circumstances, test or validation dataset were correlated with the training one; therefore the three subsets were closely similar.

Table 2.1: Comparison of SEP value between a simple random splitting and a stratified splitting processus.

Continuous variable for data splitting						
n	Splitting type	Min.	1st Qu.	Mean	3rd Qu.	Max
8	Random	0.15	0.77	1.15	1.39	6.07
	Stratified	0.17	0.75	1.14	1.37	7.38
16	Random	0.30	0.76	0.97	1.15	3.35
	Stratified	0.33	0.77	0.96	1.12	4.15
32	Random	0.47	0.78	0.90	1.02	1.38
	Stratified	0.44	0.79	0.91	1.02	1.47
64	Random	0.55	0.82	0.89	0.96	1.24
	Stratified	0.55	0.8	0.88	0.96	1.29
128	Random	0.66	0.82	0.87	0.92	1.12
	Stratified	0.68	0.82	0.87	0.92	1.14
Discrete variable for data splitting						
8	Random	0.13	0.82	1.31	1.50	18.63
	Stratified	0.15	0.84	1.16	1.42	2.81
16	Random	0.30	0.87	1.09	1.28	2.30
	Stratified	0.26	0.86	1.07	1.26	2.17
32	Random	0.48	0.90	1.04	1.18	1.76
	Stratified	0.46	0.90	1.04	1.16	2.15
64	Random	0.69	0.93	1.02	1.10	1.49
	Stratified	0.60	0.93	1.02	1.10	1.46
128	Random	0.72	0.94	1.00	1.06	1.33
	Stratified	0.73	0.94	1.00	1.06	1.32

2.2 Locally weighted approaches

For the body composition prediction, the first statistical approach used in the present study is the locally weighted approach. Locally weighted approach is a form of instance-based algorithm and its prediction is done by local functions which are using only a subset of the data. The basic idea behind locally weighted approach is that instead of building a global model for the whole function space, for each predicted subject a local model is created based on neighboring training subjects in a reference dataset. For this purpose each training subject becomes a weighting factor which expresses the influence of the data point for the prediction. In general, training subjects which are in the close neighborhood to the subject to be predicted are receiving a higher weight than those which are far away. For instance, *Nearest neighbor* local model, one of typical locally weighted approaches, simply chooses the closest point and use its output value. The selection of neighboring training subjects and their associated weighting are controled by the distance function and the weighting function; thus we will describe some useful distance functions, following by a weighting function applied in our locally weighted approach. In subsection 2.2.3, we will give the algorithm description related to our locally weighted approach.

2.2.1 Distance functions

Locally weighted approach is critically dependent on the retained distance function. There are many different ways to define a distance function, and we will briefly describe the functions which were used for our work.

The most common used distance is the unweighted Euclidean distance. Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, unweighted Euclidean distance function is written :

$$\mathcal{D}_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} = \sqrt{(\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y})} \tag{2.1}$$

This Euclidean distance can be easily generalized to a Minkowski distance with norm power to the k, and the latter is called also unweighted L_k norm :

$$\mathcal{D}_M(\mathbf{x}, \mathbf{y}) = \Big(\sum_{i=1}^{p}|x_i - y_i|^k\Big)^{\frac{1}{k}} \tag{2.2}$$

In the present study, two another unweighted distances were used, and they are respectively :

$$\mathcal{D}_{max}(\mathbf{x}, \mathbf{y}) = \max_{i=1}^{p}|x_i - y_i| \tag{2.3}$$

$$\mathcal{D}_{min}(\mathbf{x}, \mathbf{y}) = \min_{i=1}^{p}|x_i - y_i| \tag{2.4}$$

which can be viewed as limit case of equation (2.2) when $k \to \infty$ or $k \to 0$. The distance functions (2.1) - (2.4) are unweighted, and they depend on the scale of each component in $\mathbf{x}$ (or $\mathbf{y}$). For example, in our study, we used a four-dimension vector, composed by age, height, weight and waist circumference, to assess individual inter-similarity. The units of this vector are respectively year, centimeter, kilogram and centimeter, consequently it is difficult to assume equal unit among these scales. One solution is to add a coefficient associated to each covariable. In fact, this coefficient will combine covariate importance and unit together. Following this way, previous unweighted distance function could be rewritten, *i.e.*, $\mathcal{D}_M$:

$$\mathcal{D}'_M(\mathbf{x}, \mathbf{y}) = \Big(\sum_{i=1}^{p}(\rho_c|x_i - y_i|)^k\Big)^{\frac{1}{k}} \tag{2.5}$$

Use of covariate coefficient ρ_c could fix scaling problem, but the choice of coefficient value is not easy to make. We proposed to use some weighted distance functions, such as Mahalanobis distance, to take into account covariate scaling problem.

The Mahalanobis distance is based on correlations and variances between variables. Its main application is to introduce a score based on the covariate characteristics to determine the degree of dissimilarity. One of the main applications of Mahalanobis distance is to introduce a scale based on all characteristics to measure the degree of abnormality. It measures the distances in multidimensional spaces taking into account the correlations between variables or characteristics. In the multivariate framework, it is superior to unweighted Euclidean distance, because it takes into account the distribution of the points (correlation) (Srinivasaraghavan and Allada, 2006). Mathematically, Mahalanobis distance is written, given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$:

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T\Sigma^{-1}(\mathbf{x} - \mathbf{y})} \tag{2.6}$$

where Σ is covariance matrix which is supposed to be invertible.

In comparison with other Euclidean distances (*e.g.*, distance function (2.1)), some advantages of Mahalanobis distance are :

1. It takes into account not only the average value but also the variance and the covariance of the variables measured.

2. It accounts for ranges of acceptability (variance) between variables.

3. It compensates for interactions (covariance) between variables.

4. It is dimensionless.

2.2.2 Weighting function

The distance functions allow to measure the relevance of subjects to the predicted subject. Nearby subjects have high relevance, therefore they will contribute more weighting in the prediction processus. The requirements on a weighting function are straightforward (Fedorov *et al.*, 1993). The maximum value of the weighting function should be at zero distance, and the function should decline smoothly as the distance increases. In general, the smoother the weighting function, the smoother the estimated function. The weighting function should always be non-negative, since a negative value would lead to the training process increasing training error in order to decrease the training criterion. A detailed review of weighting functions can be found in Atkeson *et al.* (1997).

In the present study, we specified a continuous and decreasing weighting function, and it follows the principle such that : for a given predicted subject, the closer between a candidate subject and the predicted subject, the higher is the weighting, as a result, the more contribute this candidate suject to prediction precedure. If the weighting is zero, the candidate subject is considered as not useful at all in the prediction. More precisely, we proposed a uniform-normal curve, and the weighting function is written as follows :

$$w(d) = kexp\Big(- \frac{max(0, (d - d_0))^2}{2\sigma^2} \Big) \tag{2.7}$$

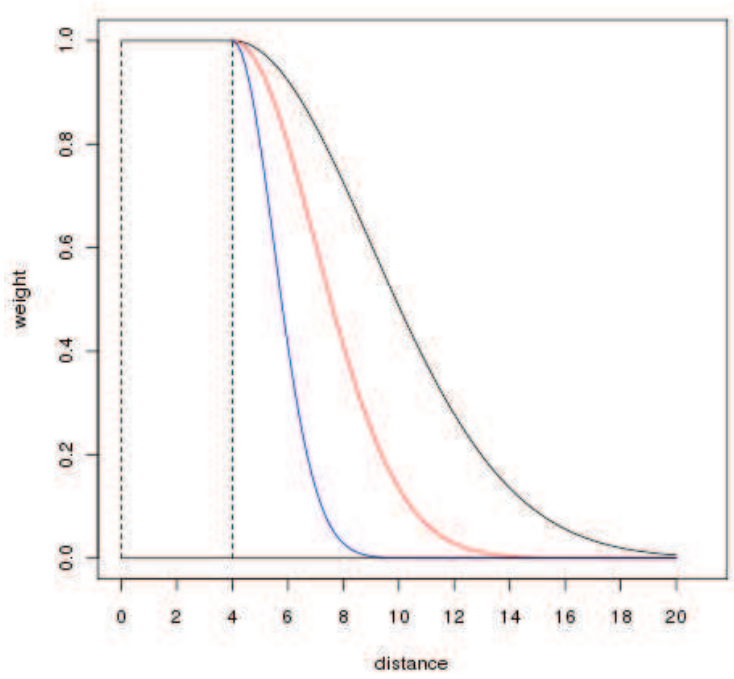
where k is a constant parameter that sets the weighting at distance less than d_0, d_0 is the threshold at which the weighting function decreases, and σ is a parameter that controls the decreasing rate. Figure 2.1 shows the form of this function. From equation (2.7), we can make some remarks :

- when the distance is lower than d_0, the weighting is equal to k;
- when the distance is higher than d_0, increasing σ value leads to increase the weighting;
- For a given predicted subject, by adjusting the value of d_0, we can acquire a subset with desired number of candidate subjects whose weighting is equal to 1.
- The weights are equal up to a constant multiplier.
- With respect to σ, it seems difficult to give it a precise meaning, thus we prefer to use d_0. d_0 can be considered as a distance which controls a proportion of candidates (*i.e.*, $\alpha = \frac{5}{100}$) under the curve of weighting function. One of the advantages of this presentation is that the proportion is expressed as a distance.

2.2.3 Algorithm description

The locally weighted approach consists in, for a given predicted subject, gathering a subset of "close" candidate subjects from training dataset, and also taking into account information provided by relatively "distant" candidates. The meaning of closeness is based on covariate distance, and will be calculated by the distance function introduced in the previous section.

Figure 2.1: The weighting value calculated from equation (2.7) with $k = 1$ and $d_0 = 4$. The distance is on the x-axis, and the weighting on the y-axis. The blue line represents $\sigma = 1.5$, the red $\sigma = 3$ and the black $\sigma = 5$.



The locally weighted approach is composed of the distance and weighting function. More precisely, the distance function allows to assess similarity among subjects, and the weighting function is used to transform the distance value to the contribution (weighting value) of candidate subjects in the prediction procedure. The algorithm of the locally weighted approach can

be described as follows :

Algorithm 1: Locally weighted algorithm

1 Assume that we have a training dataset containing n subjects (or observations) and a set of covariates $\mathcal{C}$
2 Define a distance function $\mathcal{D}$ on $\mathcal{C}$
3 Define a weighting function $w(d)$ based on the distance value such that, $\mathcal{F}_w : R^* \to [0, 1]$ and is never increasing
4 **for** *a given predicted subject* i **do**
5 **for** $1 \leq j \neq i \leq n$ *# candidate sujects used to predict subject* i **do**
6 Calculate the distance $d_{i,j}$ between candidate j and predicted subject i
7 Apply function w_d on $d_{i,j}$ for transformation of a distance $d_{i,j}$ to a weighting $w_{i,j}$
8 **end**
9 **if** $\sum_j w_{i,j} > W_L$ **then**
10 Build a prediction model $\mathcal{M}_i$ by integrating $w_{i,j}, 1 \leq j \leq n$
11 Apply $\mathcal{M}_i$ to make a prediction $\widehat{y}_i$ on predicted subject i
12 **end**
13 **else**
14 Predicted subject i is considered as unpredictable
15 **end**
16 **end**
17 Study prediction accuracy based on criterion

In the step **10** within algorithm 1, we can use plenty of nonparamtric or parametric models, more precisely in the present study, we apply the locally weighted concept in the framework of the linear regression, Support Vector Machine (SVM) regression and bayesian linear models. The weighted linear regression is well developed in the literature (Wang and Tsaur, 2000). It works by incorporating extra nonnegative weights, associated with each data point, into the fitting criterion. The size of the weight indicates the precision of the information contained in the associated observation. By optimizing the weighted fitting criterion, the parameters are estimated, and this allows to determine the contribution of each observation to the final parameter estimates. It is important to note that the weight for each observation is given relative to the weights of the other observations; so different sets of absolute weights can have identical effects. As a locally SVM regression model is used for the body composition prediction, a brief introduction about SVM will be given in the following section 2.3.

2.3 Support Vector Machine modeling

In the previous section, the locally weighted approach algorithm was given. This approach can be incoporated with different parametric models. For a predicted subject, the distance function and the weighting function are used to select a subset of similar subjects in the reference datasets, then a statistical model is applied using this subset for the prediction. In the present study, Support Vector Machine regression model is chosen for predicting the body composition, because it is well known method and shown a good usefulness in various applications.

The Support Vector Machine (SVM) are a set of supervised learning techniques to solve problems of discrimination and regression. They are a generalization of linear classification and this technique has been developped during 1990s with a theorical contribution of Vapnik (1998) about statistical learning theory: Vapnik-Chervonenkis theory. Initially, the SVMs are used as

a binary classification tool, and the principles were further extended to regression framework by Vapnik and co-workers. The SVM were quickly adopted for their ability to work with large data, the small number of hyper-parameters, the fact that they are well founded theoretically, and gave good results in practice.

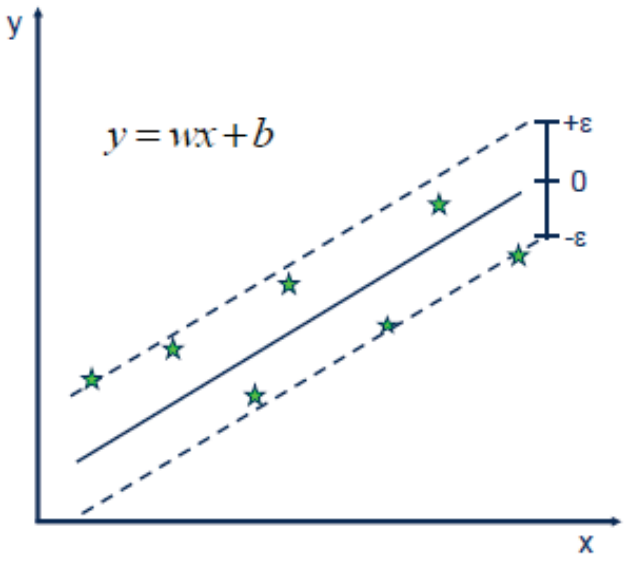
2.3.1 Overview

Support Vector Machine (SVM) is based on the statistical learning theory initially developed by Vapnik in the late 70s and later developed to a more complex concept of structural risk minimization (SRM) (Vapnik, 2000, section 4.1). In many applications, SVM has been shown to provide higher performance than traditional learning machines and has been introduced as powerful tools for solving classification problems (Burges, 1998; Lin and Wang, 2002).

The theory of SVM is formulated on the structural risk minimization (SRM) principle which minimizes an upper bound on the generalization error (Vapnik, 2000). The SVM theory starts from simple ideas on linear separable classes, then progresses into studying the case of linear non-separable classes. The separation of classes using linear separation functions is extended to the nonlinear case. In the classification problem, the SVM first maps the input points into high-dimensional feature space by using the dot product functions, called kernels, then constructs a linear separating hyperplane that maximizes the margin between different classes. The wildly used kernels for SVM are polynomials, splines, radial basis functions, and multilayer perceptrons with one hidden layer (Burges, 1998; Gunn, 1998; Hofmann *et al.*, 2008). For classification problems, the parameters which are related to these kernel functions are chosen so as to minimize an upper bound on the Vapnik-Chervonenkis (VC) dimension of the SVM. Only a subset of input points determines the SVM classifier and these points are called support vectors (SVs).

A version of a SVM for regression has been proposed by Vapnik *et al.* (1997). This method is called support vector machine regression (SVMR). The goal of SVMR is to identify a function $f(\mathbf{x})$ that for all training points $\mathbf{x}$ has a maximum deviation ε from the predicted variable values y and has a maximum margin. More precisely, using the training points, SVMR generates a model representing a tube with radius ε fitted to the data (Figure 2.2). Generally, SVMR has the following advantages :

Figure 2.2: A simple example of SVMR to predict weight on height. x is individual height, and y is individual weight. The observations are shown by the green star.



- SVMR has got a strong generalizability capability that stems from penalizing model complexity (Bergeron *et al.*, 2005);
- SVMR returns the globally extremal solution, subject to minimizing both empirical risk and model complexity, a property referred to as structural risk (Bergeron *et al.*, 2005, p. 974);
- SVMR can lead to a global model which is capable of dealing efficiently with high dimensional input vectors (Thissen *et al.*, 2004, p. 169);
- SVMR can yield a global solution which is often unique, in addition, due to the use of a constrained optimisation problem, the data enter the model in an inner product-which means that, numerically, the dimension of the data is irrelevant;
- SVMR can incorporate a kernel that enables nonlinear regression in an efficient way (Thissen *et al.*, 2004, p. 172).

For more realistic application, a soft margin SVMR was introduced with slack variables. In the standard SVMR framework, the parameter ε is used to control the trade-off between the training error and the generalisation error. However as it is often difficult to select an appropriate ε in the usual SVM framework, Schölkopf *et al.* (2000) introduced a new parameter ν to control the number of training error and support vectors (SVs), denoted as ν-SVM. In our studies of segmental body composition prediction, we used mainly ν-SVM regression (ν-SVMR) and an extension of ν-SVMR integrated with weighting value; therefore we feel necessaire to provide some mathmatical formulations. In section 2.3.2, we will discuss the ν-SVMR and then fuzzy SVMR in section 2.3.3. For more mathematical details related to SVMR, several in-depth overview could be consulted in Gunn (1998); Mangasarian and Musicant (2000); Gao *et al.* (2003); Smola and Schölkopf (2004).

2.3.2 ν-SVMR

Suppose we are given training data $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\} \in \mathcal{X} \times R$, where $\mathcal{X}$ denotes the space of the input patterns (*e.g.*, $\mathcal{X} = R^d$). These might be, for instance, body weight for a group of students with their corresponding age and body height. In standard ε-SVM regression, the goal is to find a function $f(\mathbf{x})$ that has at most ε deviation from the actually obtained variable y_i for all the training data, and at the same time is as flat as possible. In other words, we do not care about errors as long as they are less than ε, but will not accept any deviation larger than this. The function $f(\mathbf{x})$ takes the form :

$$f(\mathbf{x}) =< \mathbf{w}, \mathbf{x} > +b \quad \text{with } \mathbf{w} \in \mathcal{X}, b \in R \tag{2.8}$$

where $< \cdot, \cdot >$ denotes the dot product in $\mathcal{X}$. Flatness in equation (2.8) means to seek a small $\mathbf{w}$. One way to ensure this is to minimize the norm, *i.e.*, $\| \mathbf{w} \|^2 =< \mathbf{w}, \mathbf{w} >$. This problem can be writen as a convex optimization problem :

$$\min \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{2.9}$$

$$\text{with the constraints} \quad \begin{cases} y_i - < \mathbf{w}, \mathbf{x}_i > -b & \leq & \varepsilon \\ < \mathbf{w}, \mathbf{x}_i > +b - y_i & \leq & \varepsilon \\ \varepsilon & \geq & 0 \end{cases} \tag{2.10}$$

The above conditions can be easily extended for the soft margin SVM regression with introduction of the slack variable ξ and C (Smola and Schölkopf, 2004). As it is sometimes not easy to select an appropriate ε in the usual SVM framework, Schölkopf *et al.* (2000) proposed

another formulation in which the parameter C is replaced by a parameter $\nu \in [0,1]$. They demonstrated that ν is an upper bound on the fraction of errors (training error) and a lower bound on the fraction of support vectors (number of SVs). One advantage of ν-SVMR is that it automatically computes ε (Schölkopf *et al.*, 2000, p. 1214), whereas it is often difficult to select an appropriate ϵ in the usual SVM framework, Schölkopf *et al.* (2000) introduced a new parameter ν to control the number of training error and support vectors (SVs), hence comes the name ν-SVM. The ν-SVM can be used for both classification and regression, as discussed in detail in several reviews, by Ivanciuc (2007). The convex optimization problem is written (Schölkopf *et al.* (2000); Chang and Lin (2002)) :

$$\min \qquad \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\nu\varepsilon + \frac{1}{n}\sum_i^n(\xi_i + \xi_i^*)\right) \tag{2.11}$$

$$\text{with the constraints} \quad \begin{cases} y_i - < \mathbf{w}, \mathbf{x}_i > -b & \leq & \varepsilon + \xi_i^* \\ < \mathbf{w}, \mathbf{x}_i > +b - y_i & \leq & \varepsilon + \xi_i \\ \xi_i, \xi_i^* & \geq & 0 \\ \varepsilon & \geq & 0 \end{cases} \tag{2.12}$$

The ε determines the limits of the approximation tube, and the constant $C > 0$ determines the trade-off between the flatness of f and the deviations larger than ε to be tolareted (Vapnik, 2000, chap. 6, p. 188; Smola and Schölkopf, 2004; Ivanciuc, 2007). In the case of the ε-insensitive loss function (Vapnik, 2000, section 6.1), the deviations are described by:

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \tag{2.13}$$

On the basis of the previous loss function (2.13), the associated primal objective function is represented by the Lagrange function by introducing multipliers $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$ and β :

$$\begin{aligned} L_P(\mathbf{w}, b, \alpha_i, \alpha_i^*, \beta, \eta_i, \eta_i^*, \varepsilon, \xi, \xi^*) &= \frac{1}{2}\|\mathbf{w}\|^2 + C\nu\varepsilon + \frac{C}{n}\sum_{i=1}^n(\xi_i + \xi_i^*) - \beta\varepsilon - \sum_{i=1}^n(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ &- \sum_{i=1}^n \alpha_i(\varepsilon + \xi_i + y_i - < \mathbf{w}, \mathbf{x}_i > -b) \\ &- \sum_{i=1}^n \alpha_i^*(\varepsilon + \xi_i^* + < \mathbf{w}, \mathbf{x}_i > +b - y_i) \end{aligned} \tag{2.14}$$

The Karush–Kuhn–Tucker conditions (Boyd and Vandenberghe, 2004) for the primal problem are as follows:

$$\frac{\partial L_P}{\mathbf{w}} = \mathbf{w} - \sum_i^n(\alpha_i^* - \alpha_i)\mathbf{x}_i = 0 \tag{2.15}$$

$$\frac{\partial L_P}{\varepsilon} = C\nu - \sum_i^n(\alpha_i^* + \alpha_i) - \beta = 0 \tag{2.16}$$

$$\frac{\partial L_P}{b} = \sum_i^n(\alpha_i - \alpha_i^*) = 0 \tag{2.17}$$

$$\frac{\partial L_P}{\xi_i} = \frac{C}{n} - \alpha_i - \eta_i = 0 \tag{2.18}$$

$$\frac{\partial L_P}{\xi_i^*} = \frac{C}{n} - \alpha_i^* - \eta_i^* = 0 \tag{2.19}$$

The dual optimization problem is obtained by substituing (2.15) - (2.19) into (2.14) :

$$\begin{aligned} \max \quad L_D(\alpha,\alpha^*) &= -\frac{1}{2}\sum_i^n\sum_j^n(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)<\mathbf{x}_i,\mathbf{x}_j>-\varepsilon\sum_i^n(\alpha_i+\alpha_i^*) \\ &\quad +\sum_i^n(\alpha_i^*-\alpha_i)y_i \end{aligned} \tag{2.20}$$

$$\text{with the constraints} \quad \begin{cases} \sum_i^n(\alpha_i-\alpha_i^*) &= 0 \\ \alpha_i\in[0,\frac{C}{n}] & \\ \alpha_i^*\in[0,\frac{C}{n}] & \\ \sum_i^n(\alpha_i+\alpha_i^*) &\leq C\nu \end{cases}$$

The Lagrange function of the dual problem (2.20) can be solved through the technique of "quadratic programming". Specifically it first determine the values of α_i et α_i^*. With α_i and α_i^* found, parameters $\mathbf{w}$ can be written (Gunn, 1998, p. 31, eq (5.8); Ivanciuc, 2007, p. 343, eq (101)) :

$$\widehat{\mathbf{w}}=\sum_i^n(\alpha_i^*-\alpha_i)\mathbf{x}_i \tag{2.21}$$

Details for computation of b and ε can be found in Schölkopf *et al.* (2000); Smola and Schölkopf (2004). Finally the regression estimate for a new observation $\mathbf{x}$ is expressed as follows :

$$f(\mathbf{x}) = \sum_i^n(\alpha_i^*-\alpha_i)<\mathbf{x}_i,\mathbf{x}>+b \tag{2.22}$$

or

$$f(\mathbf{x}) = \sum_i^n(\alpha_i^*-\alpha_i)K(\mathbf{x}_i,\mathbf{x})+b \quad \text{, SVMR with a kernel function } K(\cdot) \tag{2.23}$$

2.3.3 Fuzzy SVMR

There are more and more applications using the SVMR techniques. However, in many applications, each input data point can a have different contribution to the learning of the regression function. To take into account these contributions associated to data points in the SVMR model, Lin and Wang (2002) developed Fuzzy Support Vector Machine (FSVM) on the theory of SVM. In FSVM, each sample is given a fuzzy membership which denotes the attitude of the corresponding point toward one class. The membership represents how important is the sample to the decision surface. The bigger the fuzzy membership, the corresponding point is treated more important; thus, different input points can make different contributions to the learning of decision surface.

Here we will give a short description of FSVR formulation. Suppose we are given a set of dataset $\{(\mathbf{x}_i,y_i,s_i)\}_{i=1}^n$ with s_i a fuzzy membership. The primal problem is written:

$$\min \quad \frac{1}{2}\|\mathbf{w}\|^2+C\sum_i^n s_i(\xi_i+\xi_i^*) \tag{2.24}$$

$$\text{with the constraints} \quad \begin{cases} y_i-<\mathbf{w},\mathbf{x}_i>-b &\leq \varepsilon+\xi_i \\ <\mathbf{w},\mathbf{x}_i>+b-y_i &\leq \varepsilon+\xi_i^* \\ \xi_i,\xi_i^* &\leq 0 \end{cases}$$

Analogously, by conducting a Lagrange function and solving the dual problem with a standard quadratic programming technique, we can obtain $\mathbf{w}$ and b in the FSVR (Lin and Wang, 2002; Yang and Na, 2008). Generally speaking, Fuzzy SVM technique could enhance prediction accuracy and reduce outlier effect and noise.

2.4 Bayesian networks

For the body composition prediction, the second approach used in the present study is the Bayesian network modeling. In this section, we will first give an overview of Bayesian networks, including structure learning. In subsection 2.4.3, we will propose a novel Bayesian network modeling and give its concrete application in body composition prediction.

Bayesian networks are a formalism for probabilistic reasoning and they have been introduced by Kim and Pearl (1987); Lauritzen and Spiegelhalter (1988); Jensen (1996). Bayesian network theory can be thought of a fusion of incidence diagrams and Bayes' theorem. A Bayesian network, or belief network, shows conditional probability (and eventually causality relationships) between variables. The probability of an event occurring given that another event has already occurred is called a conditional probability. The probabilistic model is qualitatively described by a directed acyclic graph, or DAG. The structure of a DAG is defined by two sets : the set of nodes (vertices) and the set of arcs (directed edges). The nodes are represented as circles containing the variable name. The connections between the nodes are called arcs. The edges represent causality, relevance or direct dependencies between variables and are drawn by arrows between nodes. In particular, an edge from node X_i to node X_j represents a statistical dependence between the corresponding variables. Thus, the arrow indicates that a value taken by variable X_j depends on the value taken by variable X_i. Node X_i is then referred to as a parent of X_j and, similarly, X_j is referred to as the child of X_i.

Advantages of Bayesian networks :

- Bayesian networks visually represent all the relationships between the variables in the system with connecting arcs.
- It is easy to recognize the dependence and independence between various nodes.
- Bayesian networks can maps scenarios where it is not feasible/practical to measure all variables due to system constraints (costs, not enough sensors, etc.)
- Bayesian networks readily facilitate use of prior knowledge.
- By using probabilistic and casual semantics, it's ideal to use a Bayesian network for representing prior knowledge and data.
- Over-fitting of data can be avoided when using Bayesian networks and Bayesian statistical methods.

However Bayesian networks are critized by some following limitations :

- The quality of the results of the Bayesian networks depends on the quality of the prior beliefs or model. A variable is only a part of a Bayesian network if you believe that the system depends on it.

- Computational difficulty of exploring a previously unknown network. To calculate the probability of existence of any arc of the network, all arcs must be calculated. While the resulting ability to describe the network can be performed in linear time, this process of network discovery is an NP-hard task which might either be too costly to perform, or impossible given the number and combination of variables.
- Calculations and probabilities using Baye's rule and marginalization can become complex and care must be taken to calculate them properly.

2.4.1 Preliminary

In probability theory and statistics, given two random variables X and Y, the conditional probability distribution of X given Y is the probability distribution of X when Y is known to have a particular value. For discrete random variables, the conditional probalility distribution of X given Y can be written :

$$\Pr(X|Y) = \frac{\Pr(X,Y)}{\Pr(Y)} = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)} \tag{2.25}$$

If X and Y are continuous variables, then their probability density function is known as the conditional density function, and it reads :

$$f_X(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_Y(y|x)f_X(x)}{f_Y(y)} \tag{2.26}$$

where $\Pr(X,Y)$ (and $f_{X,Y}(x,y)$) are the joint distribution (and density) of two random variables X and Y. From the joint distribution (or density), We can marginalize across some of the variables by adding up across all possible values of those variables. For example, given $f_{X,Y}(x,y)$ we can get the marginal probability density $f_X(x)$ by :

$$f_X(x) = \int f_{X,Y}(x,y)\mathrm{d}y \tag{2.27}$$

By combining equation (2.26) and (2.27), we can obtain the Bayes's formula :

$$f_X(x|y) = \frac{f_{X,Y}(x,y)}{\int f_{X,Y}(x,y)\mathrm{d}x} = \frac{f_Y(y|x)f_X(x)}{\int f_Y(y|x)f_X(x)\mathrm{d}x} \tag{2.28}$$

This formula is fundamental to many modern statistical techniques. Sometimes, the random variables X and Y might not be marginally independent. However, they can become independent conditioning by a third random variable Z. In the standard notation of probability theory, X and Y are conditionally independent given Z if and only if :

$$f(x \mid y, z) = f(x \mid z) \tag{2.29}$$

or equivalently,

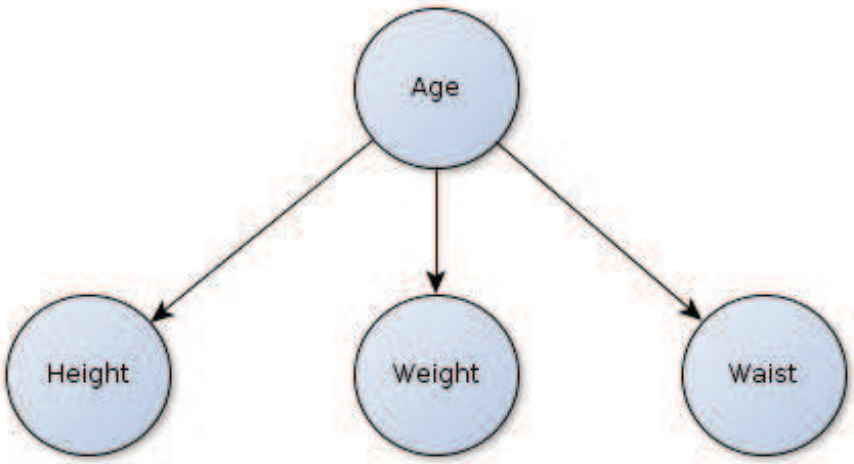
$$f(y \mid x, z) = f(y \mid z) \tag{2.30}$$

the conditional independence is often denoted as $(X \perp\!\!\!\perp Y) \mid Z$, which means that variable X is conditionally independent of variable Y given variable Z, under f (or under probability distribution Pr in the discrete case). More precisely, that is, knowledge of Y's value doesn't affect your belief in the value of X, given a value of Z. The notion of conditional independence is central to Bayesian networks and many other models dealing with probabilistic relationships. To summarize, this short introduction on probability theory, Bayes' rule and conditional independence are fundamental to the theory of bayesian networks. These ideas help a better

understanding of context.

The conditional independence properties can be expressed by a DAG. They enable an effective representation and computation of the joint probability distribution over a set of random variables. The DAG structure of the BN represents the qualitative component of the BN. However, even though the arrows represent direct dependence connection between the variables, the reasoning process can operate on BNs by propagating information in any direction. For instance, a qualitative analysis of the Bayesian network in Figure 2.3 indicates that the three anthropometric variables, Height, Weight and Waist, are not (unconditionally) independent, for instance, the information on Height might implicitly be taken into account for Weight from the edges Height → Age and Age → Weight; nevertheless they are conditionally independent when Age is known, because Age directly influences weight and additional information of height will be not useful.

Figure 2.3: Bayesian network example for a qualitative analysis.



Here is a standard definition of BNs. Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be a non-empty and finite set of n random variables, let S be a DAG whose vertex set V is in bijection with $\mathcal{X}$. A Bayesian network ($\mathcal{BN}$) is defined by a pair (S, θ) where

- S structure of Bayesian network;
- $\theta = \{\Pr(X_i|Pa(X_i)\}_{X_i \in \mathcal{X}}$, with $Pa(x_i)$ denotes the parents of node X_i.

Pearl (1988) has shown that Bayesian networks expressed a joint probability. The joint probability of several variables can be calculated from the product of individual probablities of the nodes. This demonstration leads to a following theorem :

Theorem 1. *(Jensen and Nielsen, 2007, pp. 38-39)*
Let BN be a Bayesian network over $\mathcal{X} = \{X_1, ..., Xn\}$. *Then BN specifies a unique joint probability distribution* $\Pr(\mathcal{X})$ *given by the product of all conditional probability tables specified in BN :*

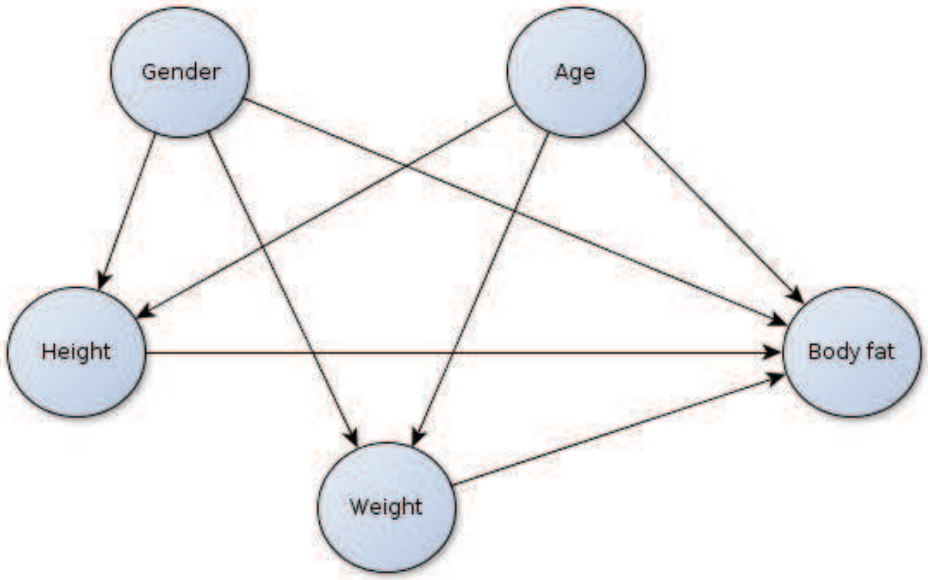
$$\Pr(\mathcal{X}) = \prod_{i=1}^{n} \Pr(X_i|Pa(X_i))$$

where $Pa(X_i)$ *are the parents of* X_i *in BN, and* $\Pr(\mathcal{X})$ *reflects the properties of BN.*

Let us present an example of Bayesian networks. Figure 2.4 emphasizes a typical relationship of variables in our study. The interpretation of the network is the following. One individual

is collected in a dataset during a routine examination. Gender and Age are independent, and height and weight are affected by individual gender and age, and this is shown by the arrows which go from Gender and Age to Height and Weight, respectively (the arrows imply a direct dependency between two variables). Height and Weight are not marginally independent, but conditionally independent given Age and Gender. The most interesting variable is body fat and it is influenced by four other variables.

Figure 2.4: Another simple example of a Bayesian network.



Learning Bayesian networks consists of finding the network that best fits the dataset for a certain scoring function. This problem is not straightforward: Cooper (1990) showed that the inference of a general Bayesian networks (BNs) is a NP-hard problem. Nevertheless heuristic search methods have been proposed to learn BNs structure from datasets. Briefly, there are two main classes of algorithms for learning BNs structure : constraint-based and score-based algorithm. Before introducing wildly used constraint-based and score-based algorithms, it is necessary to give some conditional dependence tests and score functions, respectively used in two classes of algorithms.

For the constraint-based approaches, the algorithms are mainly based on tests of conditional independence between a pair of variables. We now describe several frequently used tests, and we will borrow Tsamardinos *et al.* (2006)'s terminology. Let S_{ijk}^{abc} be the number of times in the data $\mathbf{D}$ where $X_i = a, X_j = b$ and $X_k = c$. Following the same manner, we define S_{ik}^{ac}, S_{jk}^{bc} and S_k^c. The G^2 statistic is defined as (Spirtes *et al.*, 2000; Tsamardinos *et al.*, 2006, p. 42; Neapolitan, 2004, p. 593)

$$G^2 = 2 \sum_{abc} S_{ijk}^{abc} \quad \log \frac{S_{ijk}^{abc} S_k^c}{S_{ik}^{ac} S_{jk}^{bc}} \tag{2.31}$$

The G^2 statistic is asymptotically distributed as a χ^2 with df degrees of freedom where :

$$df = (|D(X_i)| - 1)(|D(X_j)| - 1) \prod_{X_l \in \mathbf{X}_k} |D(X_l)| \tag{2.32}$$

where $D(X)$ is the domain (number of distinct values) of variable X. Moreover, the G^2 statistic is equal to Mutual Information (MI) by divising $2n$, thus MI, an information-theoretic distance measure, defined as

$$\mathbf{MI}(X_i, X_j|X_k) = \frac{G^2}{2n} \tag{2.33}$$

A detailed discussion on MI test is given in Bacciu *et al.* (2013). They argued that for categorical data the mutual information is the most effective test statistic for conditional independence, and give a rationale for its use. Also, there is the classical Pearson's $\mathcal{X}^2$ test for contingency tables (Wasserman, 2004, pp. 189-190) :

$$\mathcal{X}^2(X_i, X_j|X_k) = \sum_{abc} \frac{S_{ijk}^{abc} - T_{ijk}^{abc}}{T_{ijk}^{abc}} \quad \text{, where } T_{ijk}^{abc} = \frac{S_{ik}^{ac} S_{jk}^{bc}}{S_k^c} \tag{2.34}$$

Pearson's $\mathcal{X}^2$ test is implemented to learn BNs structure in Tsamardinos *et al.* (2006). The key rationale is that Pearson's $\mathcal{X}^2$ is a statistical test, and is asymptotically correct for a general discrete multinomial distribution. Moreover Pearson's $\mathcal{X}^2$ test is relatively easy to compute.

For the other series of algorithms, scoring criteria consist of two parts one that rewards a better match of the data to the structure and one that rewards a simpler structure. It is common to classify scoring functions into two main categories: information-theoretic and Bayesian scores. In general, for efficiency purposes, these scores need to decompose over the network structure. The decomposability property allows for efficient learning algorithms based on local search methods. More precisely, let $\mathbf{D}$ a dataset, a scoring function $Score(\mathcal{BN}, \mathbf{D})$ for learning a Bayesian network structure is called decomposable, if it can be expressed as a sum of local scores

$$Score(\mathcal{BN}, \mathbf{D}) = \sum_{i=1} Score(X_i, Pa(X_i), \mathbf{D}) \tag{2.35}$$

Among information-theoretic score functions, the general idea is to search the trade-off between fitting to the data and the complexity of the model. Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) are two popular scores for learning Bayesian network structures, in particular they are decomposable. AIC is written :

$$\mathbf{AIC}(\mathcal{BN}, \mathbf{D}) = -\log \Pr(\mathbf{D}|\hat{\Theta}, \mathcal{BN}) + d \tag{2.36}$$

where $\hat{\Theta}$ are the maximum likelihood parameters for $\mathcal{BN}$, d is the number of free parameters in the model. By multiplying d by $\frac{1}{2}\log n$, we obtain the BIC :

$$\mathbf{BIC}(\mathcal{BN}, \mathbf{D}) = -\log \Pr(\mathbf{D}|\hat{\Theta}, \mathcal{BN}) + \frac{d}{2}\log n \quad \text{, where } n \text{ is the sample size} \tag{2.37}$$

Recently, Cruz-Ramírez *et al.* (2006) evaluated the performance of the BIC and Minimum Description Length (Rissanen, 1978) principle as model selection metrics. Following their notation, the Minimum Description Length (MDL) is defined as :

$$\mathbf{MDL}(\mathcal{BN}, \mathbf{D}) = -\log \Pr(\mathbf{D}|\hat{\Theta}, \mathcal{BN}) + \frac{d}{2}\log n + C_k \tag{2.38}$$

where k is the number of variables, $C_k = \sum_i^k \big(1 + \#Pa(X_i)\big)\log k$ and $\#Pa(X_i)$ is the number of parent of X_i. Hansen and Yu (2001) contributed an in-depth review of MDL in both practical and theoretical aspects for model selection.

AIC, BIC and MDL belong to information-theorectic scoring functions, and they are widely used in many statistical contexts. Meanwhile, Bayesian scoring functions have been studied

and developed. The principle consists in giving a prior probability distribution on the possible networks, and computing the posterior probability distribution conditioned to the available data $\mathbf{D}$, $\Pr(\mathcal{BN}|\mathbf{D})$. The best network is the one that maximizes the posterior probability.

One of the first Bayesian scoring functions, called K2, was proposed by Cooper and Herskovits (1992, p. 321, eq (12)). It relies on several assumptions (multinomiality, lack of missing values, *etc*). Heckerman *et al.* (1995) generalized Cooper and Herskovits (1992)'s equation and established it in a sound theoretical framework. This generalized scoring criteria is called Bayesian Dirichlet criterion (BDe), and it is written (Heckerman *et al.*, 1995, p. 296, eq (10); Heckerman *et al.*, 1998, eq 35) :

$$\mathbf{BDe}(\mathcal{BN}, \mathbf{D}) = \Pr(\mathcal{BN}) \prod_{i}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \tag{2.39}$$

where Γ is the Gamma function, it satisfies $\Gamma(n) = (n-1)!$ where n is integer. An important property of BDe criteria is that its function is decomposable and can be written in terms of the local nodes of the graph, as equation (2.35). More theorical details about Bayesian Dirichlet scores can be found in de Campos and Ji (2010).

2.4.2 Learning Bayesian networks structure

Structure learning for Bayesian Networks is a difficult problem. The large search space and tendency for learned models to overfit make structure learning complicated. However, several heuristic and statistical based algorithms have been developed for structure learning. The first one is an approach based on constraints, denoted as constraint-based algorithms, which poses the learning process as a constraint satisfaction problem, and then constructs a network structure by testing conditional independence (CI) relations. The second one is an approach based on scores, denoted as score-based algorithms, which view learning as a model selection problem; by defining a scoring function which assesses the fitness of each model, it searches for a high-scoring network structure. For the organization of this part, we will start to introduce the contraint-based algorithms, then followed by score-based algorithms.

The contraint-based algorithms (Pearl, 1988; Spirtes *et al.*, 2000) establish a set of conditional independence statements holding for the data, and use this set to build a network with d-separation properties corresponding to the determined conditional independence properties. Constraint-based algorithms generally have smaller likelihood scores than score-based methods; however, constraint-based methods is more efficient and create structures more accurately representing the conditional independencies of the original dataset.

Most of the constraint-based algorithms are generally based on *Inductive Causation* (IC) algorithm (Verma and Pearl, 1991), which consists of three main steps :

Step 1 Connect nodes $X - Y$ if and only when no set of variables S_{XY} (excluding X, Y) can be found with $(X \perp\!\!\!\perp Y)|S_{XY}$, *i.e.* X, Y are independent given all variables in S_{XY}.

Step 2 For each substructure $X - Z - Y$ (X and Y nonadjacent), orientate the edges to $X \rightarrow Z \leftarrow Y$ (a so-called v-structure), if $Z \notin S_{XY}$.

Step 3 Orientate as many of undirected edges as possible subject to the condition that neither a new v-structure nor a directed cycle should be created.

The most basic algorithm is the SGS (Spirtes-Glymour-Scheines) algorithm (Spirtes *et al.*, 2000), statistically consistent, but very computationally inefficient : the SGS algorithm requires a number of d-separation tests that increases exponentially with the number of variables

n. Therefore, some variations of SGS algorithm have been proposed. One best known of these variations is PC algorithm (Spirtes *et al.*, 2000, pp. 84-89). It works exactly like the SGS algorithm, except for the edge removal step, where it tries to condition on as few variables as possible (as above), and only conditions on adjacent variables. The PC algorithm has the same assumptions as the SGS algorithm, and the same consistency properties, but generally runs much faster, and does many fewer statistical tests.

Recently, another kind of algorithm based on Markov Blanket property has drawn a lot of attention. The Markov Blanket of a variable of X_i, MB(X_i), is the minimal set for which $(X_j \perp\!\!\!\perp X_i|\text{MB}(X_i))$, for all $X_j \in \mathcal{X} - X_i - MB(X_i)$ where $\mathcal{X}$. The Markov Blanket of a variable X_I is the minimal and unique[2] set of variables which can completely shield variable X_i from all other variables. All other variables are probabilistically independent of the variable X_i conditioned on the Markov Blanket of variable X_i.

There are several Markov Blanket learning methods such as : Koller-Sahami (KS) algorithm (Koller and Sahami, 1996) , Grow-Shrink (GS) algorithm (Margaritis and Thrun, 1999), Incremental association Markov Blanket (IAMB) algorithm (Tsamardinos *et al.*, 2003b), Max-Min Parents and Children (MMPC) (Tsamardinos *et al.*, 2003a) and Max-Min Markov Blanket (MMMB) algorithm (Tsamardinos *et al.*, 2003a).

Koller-Sahami (KS) algorithm is the first algorithm to employ Markov Blanket concept for feature selection. Koller and Sahami (1996) provided a theoretical justification for optimal feature selection based on using cross-entropy to minimize the amount of predictive information lost during feature elimination. Although KS algorithm is theoretically sound, there is no theoretical guarantee for KS algorithm to find optimal Markov Blanket sets (Tsamardinos *et al.*, 2003b, p. 378). To low the computation, the KS algorithm requires two parameters : (1) the number of variables to retain, and (2) the maximum number of variables the algorithm is allowed to condition on. These two limits are helpful to reduce the search complexity greatly, but with a loss of of correctness. KS algorithm uses a backward heuristic removing strategy which is claimed unsound and the output of which is an approximate Markov blanket of the target attribute.

The Grow-Shrink (GS) Markov blanket algorithm has been proposed by Margaritis and Thrun (1999). The GS algorithm is claimed to be the first sound algorithm for Markov blanket discovery and will output the correct Markov blanket of target attribute under certain assumptions. The key idea is to first learn the Markov blanket of each node in the network, denoted as *grow phase*, then remove each false member of Markov blanket, denoted as *shrink phase*. More precisely, by using a static forward heuristic search strategy in *grow phase* and false positive judgment stragegy in *shrink phase*, we will get the unique Markov blanket of target variable. However, it is indicated that in many cases, especially in the case of the small size dataset, GS algorithm is not faithful and it couldn't discover the correct Markov blanket subset (Tsamardinos *et al.*, 2003b). The GS algorithm is described as follows, and more discussions related to

[2]If the data set is faithful to the $\mathcal{BN}$, then the MB of each variable is unique (Frey *et al.*, 2003).

GS algorithm can be found in Margaritis (2003) :

Algorithm 2: Grow-Shrink Algorithm

1 $S \leftarrow \varnothing$
2 /* Grow Phase */
3 While $\exists Y \in V - \{X\}$ such that $(X not \perp\!\!\!\perp Y|S)$ do
4 $MB(X) \leftarrow MB(X) \cup \{Y\}$
5 /* Shrink Phase */
6 While $\exists Y \in S$ such that $(X \perp\!\!\!\perp Y|S - \{Y\})$ do
7 $S \leftarrow S - \{Y\}$
8 return S)

Similar to GS algorithm with two search procedures (forward and backward), Tsamardinos *et al.* (2003b) have proposed the Incremental Association Markov blanket algorithm (IAMB) for classification problems in microarray research. Tsamardinos *et al.* (2003b) pointed that GS algorithm is theoretically sound but uses a static and potentially inefficient heuristic in *grow phase*. Therefore, IAMB algorithm enhances GS by using a dynamic heuristic. The "dynamic heuristic" means that IAMB reorder the set of variables each iteration a new variable enters the blanket in the growing phase, which may get more accurate blankets under some conditions. Comparing to GS algorithm, IAMB might achieve a better performance with fewer false positives admitted during the forward phase (Yaramakala and Margaritis, 2005). In spite of the improvement, IAMB algorithm, like GS, is still not data efficient because it requires a very large number of samples to perform well[3]. Three variants of IAMB algorithm have also been developed to reduce the size of CI test. InterIAMBnPC algorithm (1) interleaves the *grow phase* of IAMB with *shrink phase* attempting to keep the size of MB as small as possible; (2) it substitutes the *shrink phase* as implemented in IAMB with the PC algorithm instead. Two other IAMB variants are interIAMB and IAMBnPC, and they only either interleave the first two phases or rely on PC for the *shrink phase*, respectively (Tsamardinos *et al.*, 2003b). From experimental study, these three IAMB variants achieve better performance on data efficiency than IAMB.

To take account of data inefficient problem in GS, IAMB algorithm and IAMB variants, Max-Min Markov Blanket (MMMB) algorithm (Tsamardinos *et al.*, 2003a), HITON-MB (Aliferis *et al.*, 2003) and Parents and Children based Markov Boundary (PCMB) algorithm (Peña *et al.*, 2007) were proposed [4]. All these three algorithms follow a divide-and-conquer method that breaks down the problem of identifying Markov Blanket of a target variable X_i into two subproblems: First, identifying parents and children of a target variable X_i (denoted as $PC(X_i)$) and, second, identifying the spouses of X_i. Meanwhile, they have the same two assumptions as IAMB (*i.e.*, faithfulness and correct independence test) and take into account the graph topology to improve data efficiency. However, results from MMMB and HITON-MB are not always correct since some descendants of X_i other than its children will enter $PC(X_i)$ during the first step of identifying parents and children of X_i (Peña *et al.*, 2007). Nevertheless, PCMB can be proved overcoming this problem, more importantly, it is the first trial proved sound theoretically. The full algorithm specification and theorectic demonstration can be referred in Peña *et al.* (2007).

On the other hand, the score-based algorithms involve searching over possible Bayesian Net-

[3]IAMB may require a large amount of learning data to identify the MB of a variable, because in practice, its CI tests may condition on the whole MB of a variable (Fu and Desmarais, 2008).
[4]data efficient because the number of instances required to identify MB(T) does not depend on the size of MB(T) but on the topology of G.

work structures in an attempt to maximize a scoring function. Scoring functions are generally a variation on likelihood penalized to discourage overly complex network structures. Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978), and Bayesian Dirichlet criterion (BDe) (Heckerman *et al.*, 1995) are three common penalized likelihood metrics used in score-based algorithms. Although model likelihood is maximized, the search problem grows exponentially with the size of the dataset. Due to the large size of the problem space, search algorithms are generally coupled with heuristics that limit the size of the problem (Heckerman *et al.*, 1995; Cooper and Herskovits, 1992). The most common search method applied in scored-based approach is greedy hill climbing search over all DAG structures (Bouckaert, 1995), which performs local operations on the network that lead to the best change in score until no more positive changes can be made. Due to its greediness, hill climbing is computationally expensive and cannot be run on high-dimensional networks. A modified greedy search can be augmented with methods for escaping local, sub-optimal maxima. For instance, random restarts, simulated annealing, or incorporation of a TABU list are often added to a search procedure (Chickering *et al.*, 1995; Bouckaert, 1995; Glover, 1989). The basic hill climbing search algorithm is given as :

Algorithm 3: Hill-Climbing Algorithm

1 Choose an initial network structure $\mathcal{BN}$ over $\mathcal{X}$
2 Compute the score of $\mathcal{BN}$, denoted as $Score(\mathcal{BN}, \mathbf{D}) = Score(\mathcal{BN})$
3 Set maxscore = $Score(\mathcal{BN})$
4 Repeat the following steps as long as maxscore increase :
   (a) for every possible arc addition, deletion or reversal not resulting in a cyclic network:
      (i) compute the score of the modified network $\mathcal{BN}^*$, $Score(\mathcal{BN}^*)$
      (ii) if $Score(\mathcal{BN}^*) > Score(\mathcal{BN})$, set $\mathcal{BN} = \mathcal{BN}^*$ and $Score(\mathcal{BN}) = Score(\mathcal{BN}^*)$
   (b) update maxscore with the new value of $Score(\mathcal{BN})$
5 Return the directed acyclic graph $\mathcal{BN}$.

The size of the search space of greedy search (i.e., the number of possible screened DAGs) is superexponential to the number of variables. Moreover it is reported that searching DAG-space is slightly redundant, since some local moves result in a graph that is in the same equivalence class as its predecessor in search. Several methods have emerged to improve the efficiency of greedy search. One of the most relevant algorithms in this class is the Greedy Equivalent Search (GES) algorithm (Chickering, 2003). Greedy Equivalence Search performs greedy search in the the space of equivalence classes and represents an equivalence class by a partially directed acyclic graph (PDAG). GES theoretically finds the most probable network in the sample limit if the distribution of the data is faithful. Greedy Equivalent Search searches in the space of equivalence classes (PDAGs), however, it has the attractive property that it is guaranteed to identify in the sample limit the most probable a posteriori Bayesian network provided that the data distribution is faithful. In practice, GES is only locally optimal to overcome local maxima, a modified version has been proposed by (Nielsen *et al.*, 2002). More precisely, Nielsen *et al.* (2002) introduced randomness into GES that provides a trade-off between greediness and randomness. This approach improves results but does not solve the problem of local maxima holistically. Other heuristic search methods have also been developed, *i.e.*, genetic algorithms (Larrañaga *et al.*, 1996), simulated annealing (Chickering *et al.*, 1995), tabu search (Acid and de Campos, 2003), ant colony optimization (De Campos *et al.*, 2002), etc. A well documented and comprehensive review on this topic is provided in Russell and Norvig (2009).

Another way to improve efficiency of the search uses constraints placed on the search. The most widely used is the K2 algorithm (Cooper and Herskovits, 1992). It combines the K2 metric[5] with a greedy hill-climbing search and requires an ordering of the total variables. The K2 algorithm starts by assuming that a node lacks parents, after which, in every step, it adds incrementally the parent whose addition most increases the probability of the resulting structure given the data. The K2 algorithm stops adding parents to the nodes when the addition of a single parent cannot increase the probability. However the K2 algorithm is sensitive to the ordering. The same idea using a quite different approach is taken in the Optimal Reinsertion (OR) algorithm (Moore and Wong, 2003). Given a starting structure, OR algorithm picks a target node and all arcs into and out of the node are severed. Then subject to some constraints, OR algorithm finds the optimal set of in-arcs and out-arcs with which to reinsert it. This procedure continues, running through repeated cycles in which all nodes take turns at being the target, until no step changes the DAG structure. Optimal Reinsertion makes use of specialized data-structures (Moore and Schneider, 2002; Moore and Wong, 2003, p. 553) to make tractable the evaluation of search operators.

Both the constraint-based and score-based algorithms have their advantages. Score-based algorithms typically works better with less data than the constraint-based algorithms and with probability distributions that admit dense graphs[6]. They also allow probability distributions over models to be easily represented and have better mechanisms for dealing with missing data. However, contraint-based algorithms work well with sparse graphs, are generally quick and have good ways of finding hidden common causes and selection bias. The constraint-based algorithms are inaccurate in dense networks and few data because the CI tests become unreliable in such cases. The score-based algorithms are more accurate, but they do not scale up to high-dimensional problems due to a super-exponential growth of the search space. As both approaches have proper advantages, hybrid algorithms have been developed to combine the good points of both. Typically, a hybrid algorithm starts with a constraint-based algorithm to find the skeleton of the network and then employs a score-based algorithm to identify the best set of edge orientations. Among these hybrid algorithms, we will only mention several of them, such as the first hybrid algorithm the CB algorithm (Singh and Valtorta, 1993), the BENEDICT algorithm (Acid and de Campos, 2001), Cooperative Coevolution Genetic Algorithm (CCGA) (Wong *et al.*, 2004), the Max-Min Hill-Climbing (MMHC) algorithm (Tsamardinos *et al.*, 2006), the Hybrid HPC (H2PC) (Gasse *et al.*, 2012). For other structure learning techniques, a very comprehensive discussion is given in Daly *et al.* (2011).

Besides, it is worth pointing out one of available free softwares for learning Bayesian network structure, bnlean. bnlearn[7] written by Scutari (2010) is an R package for learning the graphical structure of Bayesian networks. The package implements a range of learners including constraint-based, search-and-score and hybrid learners. Each learner has a range of appropriate arguments and options, and the learner's implementations allow them to work with, at least, moderately complicated networks. In addition to the learning algorithms the library also includes representations of a number of 'well known' networks with the ability to generate datasets for these networks. It is possible to export these generated datasets in a simple format, and additional datasets can be imported using the same simple format.

[5]K2 metric was generalized to the Bayesian Dirichlet metric expressed in equation (2.39) (Heckerman *et al.*, 1995)

[6]In mathematics, a dense graph is a graph in which the number of edges is close to the maximal number of edges. The opposite, a graph with only a few edges, is a sparse graph. The distinction between sparse and dense graphs is rather vague, and depends on the context.

[7]The software and documentation is available for download from http://www.bnlearn.com/.

2.4.3 Crossed Linear Gaussian Bayesian Networks (Paper submitted to *Journal de la Société Française de Statistique*)

In this part, we will describe a novel Bayesian networks modeling, a parsimonious sub-model described by a linear Gaussian Bayesian network. The main aim is to improve the multivariate predictive accuracy of the classic linear regression model. The theoretical description will first be given, then we will apply the proposed approach to the body composition prediction from easily measurable covariables.

A standard statistical approach to multivariate prediction is the multivariate multiple regression model (Anderson, 2003, Chapter 8). Let us suppose that we have n samples with p variables to predict with the help of q covariables. The model reads

$$\begin{array}{rcl} \mathbf{Y} & = & \mathbf{X}\Theta + \mathbf{E} \\ V\left(vec\left(\mathbf{E}\right)\right) & = & \mathbf{I}_n \otimes \Sigma \end{array} \tag{2.40}$$

where $\mathbf{Y}$ is the variable matrix $[n \times p]$, $\mathbf{X}$ is the covariable matrix augmented with a $\mathbf{1}$ vector $[n \times (q+1)]$, Θ is the expectation parameter matrix $[(q+1) \times p]$, $\mathbf{E}$ is the error matrix $[n \times p]$ and Σ is the covariance matrix $[p \times p]$; n being the number of observations, p the number of variables and q the number of covariables. The number of parameters is $p(q+1)$ for the expectation and $p(p+1)/2$ for the covariance matrix. When p and q are large, n must be large as well to obtain estimates with desirable statistical properties. Of course, variances and covariances are more demanding in terms of sample size: it is well known that much easier to estimate the location than either scale or shape of a distribution.

Many proposals have been made in the literature to offer more sophisticated and convenient statistical tools for multivariate regression problems. Some examples are the undirected graphical models used in Whittaker (1990), the multivariate analysis of variance (MANOVA) and seemingly unrelated regression (SUR) models in Timm (2002), the multivariate generalised linear models in Fahrmeir and Tutz (1994), and more recently the graphical lasso in Friedman *et al.* (2007).

The idea we develop in this paper is to use linear prediction in a more parsimonious framework using linear Gaussian Bayesian networks [GBN]. We explore two possible approaches: (i) a general GBN and (ii) a crossed GBN to take advantage of known structures on the set of variables. After introducing some general principles and results, a real-world application to the prediction of Human Body Composition from easily measurable covariables is proposed.

Linear Gaussian Bayesian Networks

A GBN is a Bayesian network (Neapolitan, 2004, sections 4.1.3 and 7.2.3 ; Korb and Nicholson, 2011, section8.2) where every variable (or node) follows a Normal distribution. For each node, conditionally to its ascendants, the variance is constant and the expectation depends only on the direct parents through an affine transformation of the parent values. As a consequence, the joint probability distribution of the set of variables is multinormal with a free expectation and a constrained covariance matrix. In addition the acyclicity constraint of Bayesian networks induces a partial ordering on the nodes, and their relationships can be represented with a directed acyclic graph [DAG] (Pearl, 1988; Pearl, 2009; Leray, 2006, chapter 1; Koller and Friedman, 2009). More precisely, it exists at least one topological order on the node set, say $([1], [2], ..., [p])$ such that the distributions can be defined by the following p equations:

$$Y_{[i]} \mid Y_{[1]}, ..., Y_{[i-1]} \sim N\left(\mu_{[i]} + \sum_{u=1}^{i-1} \rho_{[i],[u]} Y_{[u]}, \sigma^2_{[i]}\right) \text{ for } i = 1, ..., p \tag{2.41}$$

where the summation term vanishes if node $Y_{[i]}$ has no parent. When the $p(p-1)/2$ regression coefficients $\rho_{[i],[u]}$ are all unknown and unconstrained, the GBN is saturated and there is no

restriction on the form of the covariance matrix of the implied multinormal distribution. In that case, the model has $p(p+3)/2$ free parameters. If we denote the number of parents of the ith node with $p(i)$, there are p free parameters for the μs, p for the σs and $m = \sum_{i=1}^{p} p(i)$ for the ρs. It is easy to see that the μs and σs are respectively associated to the location and scale parameters of the variables, so we can assume that all variables have marginal zero expectations and unity variances without altering the intrinsic properties of the model. As a result, the maximum number of parameters is $m = p(p-1)/2$, corresponding to the conveniently modified ρs and related to the $p(p-1)/2$ correlation parameters of the multinormal distributions.

Just to give a small example, let us consider a GBN with three marginally centred and normalised nodes with conditional distributions:

$$\begin{aligned} Y_1 &\sim N(0,1), \\ Y_2 \mid Y_1 &\sim N\left(\rho_{12} Y_1, \left(1-\rho_{12}^2\right)\right), \\ Y_3 \mid Y_1, Y_2 &\sim N\left(\rho_{23} Y_2, \left(1-\rho_{23}^2\right)\right); \end{aligned} \tag{2.42}$$

and the following joint distribution:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{12}\rho_{23} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{12}\rho_{23} & \rho_{23} & 1 \end{pmatrix}\right).$$

Compared to an unconstrained distribution on Y_1, Y_2 and Y_3, there is one less free parameter (ρ_{13}), thus inducing the following constraint on the correlation matrix:

$$Cor(Y_1, Y_3) = Cor(Y_1, Y_2).Cor(Y_2, Y_3)$$

For any GBN, the number of free parameters in the correlation matrix is simply given by the number of arcs in the associated DAG, which is equal to m. It is important to note that this way to impose constraints on the correlation matrix is quite efficient and intuitive. However, expressing the induced constraints is not always as straightforward as in this small example, even though the rules to get the correlation coefficients from the regression coefficients are themselves simple.

To define the DAG associated with a GBN, it is convenient to use a $p \times p$ adjacency matrix, say A. Each row and each column of A is associated with one of the nodes in the DAG, and when Y_i is a parent of $Y_{i'}$, then $A_{i,i'}$ is equal to one, and zero otherwise. Since it is equivalent to the DAG, the adjacency matrix shares all its properties; for instance, the number of arcs in the DAG is given by the sum of all the elements of A. Another interesting property is that A^u provides the number of paths of length u joining any ordered pair of nodes built with successive arcs of the DAG (Bang-Jensen and Gutin, 2009). In case of GBNs, we can also define the regression coefficient matrix, R, which has the same dimension and the same zeros as the adjacency matrix, but the regression coefficients instead of ones. It is of use in finding the joint distribution of the nodes as explained below.

Indeed the computation of the joint distribution of the set of nodes is not an easy task even assuming a multinormal distribution. Three algorithms to compute its expectation vector and covariance matrix are reported in the following; the first two are based on the topological order. The first one is related to the algorithm illustrated in Korb and Nicholson (2011, section 2.4.1) for discrete BNs. It is also of interest to look at proposals made by Neapolitan (2004, section 4.1.3).

1. Recursive construction.

   (a) Start with the marginal distribution of the first node in the topological ordering, which by definition has no parent.

(b) From the second node until the last node, using the joint distribution of the previous nodes and the conditional distribution to them of the considered node, compute the marginal expectation, the marginal variance and the covariances. This is quite straightforward since the current node is defined as a linear combination of the previous nodes plus an independent normal variable, and the parameter values are given by the conditional definition.

2. Affine transformation of a vector of independent centred and normalised normal variables denoted by E, that is, the identification of the vector M and matrix G such that $Y = M + G \cdot E$.

   (a) Express the first node as $M_1 + G_{1,1}E_1$.

   (b) From the second node until the last node, the distribution of the ith node can be expressed as $M_i + \sum_{u=1}^{i} G_{i,u}E_u$.

   It is important to note that the matrix G is lower triangular and that all its diagonal components can be imposed to be strictly positive.

3. Use of the matrix R defined above.

   (a) Compute R.

   (b) Compute the matrix

$$R^* = \mathbf{I}_p + \sum_{u=1}^{p-1} R^u. \tag{2.43}$$

   Only the powers of R for which u is less or equal to the maximum path length in the DAG are needed, because R^u is a zero matrix for larger values of u. An upper bound for the maximum path length is $p - 1$. There are algorithms to compute them for a specific DAG (Bang-Jensen and Gutin, 2009; Sedgewick, 2011) which can be of interest when the number of nodes is large; otherwise, a numerical test can be performed at each new power to check whether all the elements of R^u are equal to zero. Note that R^* is upper triangular when it rows and columns are ordered following the same topological order of the nodes, and that all its diagonal components are equal to one.

   (c) Compute the expectation vector. Let μ be the vector of the constant terms of the conditional expectations of all nodes. Then

$$E[\mathbf{Y}] = R^{*'} \cdot \mu. \tag{2.44}$$

   (d) Compute the covariance matrix. Let $\mathrm{diag}(\sigma)$ be a diagonal matrix whose components are the conditional standard deviations of the nodes. Then

$$V[\mathbf{Y}] = R^{*'} \cdot \mathrm{diag}(\sigma)^2 \cdot R^*. \tag{2.45}$$

Crossed Gaussian Bayesian Networks

In some situations, the set of variables has a crossed structure, that is the variables can indexed by a couple of indexes, all couples being present. In the following we will denominate these indexes: series of items. The most widely-known case is dynamic BNs, in which the same set of variables is observed at different successive times, but other situations are possible as shown in the example below (§2.4.3). In order to obtain a parsimonious model, requiring only a small number of parameters, it can be profitable to use a crossed structure. To do so, we propose to use crossed DAGs, which are the product of one DAG on the first series of items by

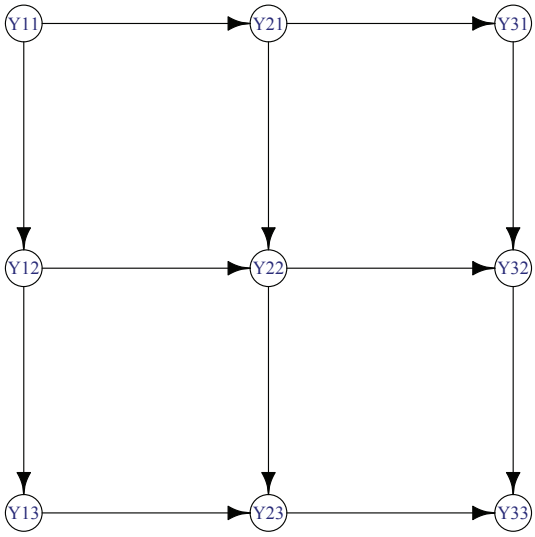


Figure 2.5: Crossed DAG obtained by crossing the serial DAG defined in (2.42) by itself.

another DAG on the second series of items. In fact a crossed DAG is the Cartesian product of DAGs associated to each series of items (Bang-Jensen and Gutin, 2009). More formally, let us denote each variable with a pair of indices associated to the two series of items: $Y_{(i_1,i_2)}$ with $i_1 = 1, ..., p_1$, $i_2 = 1, ..., p_2$ and $p = p_1 p_2$; also be $A^{(1)}$ (and $A^{(2)}$) the adjacency matrices associated onto the p_1 (and p_2) items. The adjacency matrix of the resulting crossed DAG is given by the simple rule:

$$\begin{aligned} A_{(i_1,i_2),(j_1,j_2)} &= 1 \text{ when } \begin{cases} i_1 = j_1 \quad \text{and} \quad A^{(2)}_{i_2,j_2} = 1 \\ \quad \text{or} \\ i_2 = j_2 \quad \text{and} \quad A^{(1)}_{i_1,j_1} = 1 \end{cases} \\ &= 0 \text{ otherwise.} \end{aligned} \tag{2.46}$$

That is, for each set of nodes having a common item of series 1, the DAG for series 2 is applied; and conversely for each set of nodes having a common item of series 2, the DAG for series 1 is applied. Equation (2.46) is equivalent to the matrix formula

$$A = \mathbf{I}_{p_1} \otimes A^{(2)} + A^{(1)} \otimes \mathbf{I}_{p_2}. \tag{2.47}$$

Crossing the DAG defined by (2.42) with itself produces the crossed DAG of Figure 2.5. This DAG can be used to propose a GBN, and obviously the number of parameters is reduced from a maximum of 36 to 12. In addition to those, following from the crossed structure, more constraints can be imposed on the remaining regression coefficients. For instance, some equalities can be imposed, like those implied by:

$$R = \mathbf{I}_{p_1} \otimes R^{(2)} + R^{(1)} \otimes \mathbf{I}_{p_2} \tag{2.48}$$

where $R^{(1)}$ ($R^{(2)}$) is some regression matrix associated to the DAG of the first (second) series of items. If m_1 (m_2) is the parametric dimension of the first (second) DAG over the items, the parametric dimension of the crossed DAG is reduced from $p_2 m_1 + p_1 m_2$ to $m_1 + m_2$, which can be a drastic drop. Intermediate proposals can be made; some examples are given in Table 2.2.

Table 2.2: Different restrictions on the regression coefficients of a crossed DAG. p_1 and p_2 are the node numbers of the elementary DAGs generating the crossed DAG, and m_1 and m_2 are their respective parametric dimensions.

type	constraints	parametric dimension
F.F	no one	$p_2m_1 + p_1m_2$
C.F	identical for each item of series 1	$p_2m_1 + m_2$
F.C	identical for each item of series 2	$m_1 + p_1m_2$
C.C	identical for both series	$m_1 + m_2$

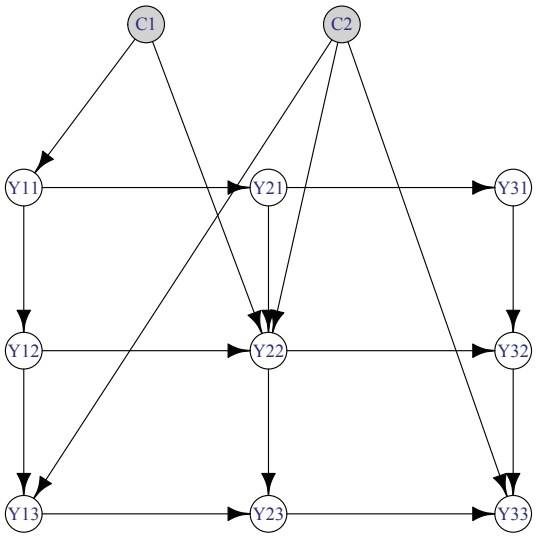


Figure 2.6: Crossed DAG from Figure 2.5 completed with two covariables ($C1$ and $C2$). The covariables intervene only on some of the variables for parsimony.

Introducing Covariables

When we introduced GBNs in the previous sections, we focused only on the variables to predict. We will now incorporate covariables as well to match the regression model described in Equation (2.40), defining the probability distribution of $\mathbf{Y}$ conditional to the covariables $\mathbf{X}$ as a crossed GBN. Starting from the simple example in Figure 2.5, the result is shown in Figure 2.6. The presence of the conditioning covariables alters the properties of GBNs introduced in §2.4.3. The expectation cannot be further supposed to be null since it depends on the covariables' values; in addition, the correlation matrix loses the simplicity of Equation (2.45).

As an example of the increased complexity introduced by the covariables, consider a toy example of one covariable intervening in two nodes of a 2×2 crossed DAG. Suppose that the joint distribution between variables and covariables can be described with a centred and

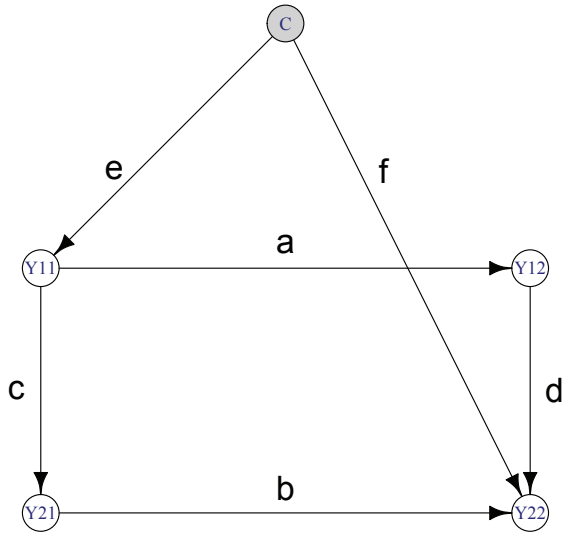


Figure 2.7: Toy 2×2 crossed DAG with one covariable (**C**). Regression coefficients of the centred normalised distribution are indicated on each arc of the DAG.

normalised GBN as proposed in Figure 2.7, that is:

$$\begin{array}{rcc} C & \sim & N(0,1) \\ Y_{1,1} \mid C & \sim & N(eC, 1-e^2) \\ Y_{1,2} \mid Y_{1,1} & \sim & N(aY_{1,1}, 1-a^2) \\ Y_{2,1} \mid Y_{1,1} & \sim & N(cY_{1,1}, 1-c^2) \\ Y_{2,2} \mid C, Y_{1,2}, Y_{2,1} & \sim & N\left(fC + dY_{1,2} + bY_{2,1}, \sigma^2_{2,2}\right) \end{array} \tag{2.49}$$

where

$$\sigma^2_{2,2} = \; 1 - (f^2 + d^2 + b^2 + 2(efad + efcb + adcb))\,.$$

In the equations above, the main difficulty lies in defining the conditional variances to achieve all the marginal variances to be ones. The structure of the covariance (here correlation) matrix is more evident giving only the upper part:

$$V \begin{bmatrix} C \\ Y_{1,1} \\ Y_{2,1} \\ Y_{1,2} \\ Y_{2,2} \end{bmatrix} = \begin{pmatrix} 1 & e & ce & ae & f + ade + bce \\ - & 1 & c & a & ef + ad + bc \\ - & - & 1 & ac & cef + acd + b \\ - & - & - & 1 & aef + d + abc \\ - & - & - & - & 1 \end{pmatrix}. \tag{2.50}$$

It is trivial to show that every correlation is the sum of the products of the regression coefficients following every possible path linking the considered pair of nodes, as in the third proposal to compute the joint distribution of the nodes.

The induced regression formulae can be computed from the equations above as follows:

$$E\left[\begin{pmatrix} Y_{1,1} \\ Y_{2,1} \\ Y_{1,2} \\ Y_{2,2} \end{pmatrix} \mid C\right] = \begin{pmatrix} e \\ ce \\ ae \\ f + ade + bce \end{pmatrix} \cdot C, \tag{2.51}$$

$$V\left[\begin{pmatrix} Y_{1,1} \\ Y_{2,1} \\ Y_{1,2} \\ Y_{2,2} \end{pmatrix} \mid C\right] = \begin{pmatrix} 1-e^2 & c(1-e^2) & a(1-e^2) & (ad+bc)(1-e^2) \\ - & 1-c^2e^2 & ac(1-e^2) & acd(1-e^2)+b(1-c^2e^2) \\ - & - & 1-a^2e^2 & abc(1-e^2)+d(1-a^2e^2) \\ - & - & - & 1-(f+ade+bce)^2 \end{pmatrix} \tag{2.52}$$

Some of these expressions can be interpreted in terms of paths over the DAG from Figure 2.7, but other are more complicated and have no obvious graphical interpretation. The presence of two or more covariables (Figure 2.6) means that the algebraic computations and the subsequent interpretations are no longer necessarily possible.

Parameter Estimation

When a GBN is free with respect to its DAG, that is the parameters of each equation (2.41) are not constrained, maximum likelihood (ML) estimates can be computed with the standard ML estimators of each equation.

When some additional simple equalities are assumed like those in Table 2.2, things are more difficult. ML estimators cannot be simply obtained from data frames stacking the variables and their corresponding parents since the same regression coefficient can be involved in different sets of regression coefficients. The situation is similar to that of dynamic GBNs, but without the additional complication of handling hidden variables (Murphy, 2002, Chapters 3 and 4). To overcome this difficulty, we suggest the following heuristic alternating least squares procedure:

- We define a score for the difference between two GBNs sharing the same DAG as the sum of squared discrepancies of all the parameters, including the standard deviation parameters.
- We initialise all the parameters using the unconstrained fit described above.
- We iterate until convergence with the defined score. Each iteration is a cycle over all expectation parameters. We update each expectation along with the parameters and the standard deviations of the involved nodes, while keeping all the others fixed. Estimation is performed by weighted least squares.
- We monitor convergence with predefined threshold on the score.

For all the examples we considered, convergence was very fast, typically after less than ten iterations. The implementation of this algorithm is available within an R package named /rbmn/ from the third author which will be made available on CRAN.

Prediction of the Body Composition

The human body composition is the allocation of body weight among three components: (L)ean, (F)at and (B)one. In detailed analyses, the body composition is investigated for each of the main segments of the body: (T)runk, (L)egs and (A)rms. Body composition is an important diagnostic since ratios of these masses can reveal regional physiological disorders. In the following, we will try to predict it from easily accessible covariables: the (A)ge, the (H)eight, the (W)eight and the waist (C)ircumference; more details can be found in Tian *et al.* (2013) where a saturated model was used. For this purpose, we have retained a data set of one hundred

white men. For each individual the variables to predict as well as the covariables have been recorded. Below are the first six.

```
   A   H     W   C   TF   LF  AF   TL   LL   AL  TB  LB  AB
1 83 182  92.6 117 17.1  8.9 3.0 31.2 18.5  6.6 0.6 1.1 0.5
2 68 169  74.7  93  8.3  5.4 2.0 28.0 16.2  7.5 0.7 1.0 0.5
3 28 182 112.2 112 17.7 11.3 3.1 36.7 24.5 10.1 0.8 1.1 0.5
4 41 171  82.6  96 10.6  6.5 1.8 29.2 19.6  7.8 0.8 1.1 0.5
5 85 169  71.1 102 10.9  4.7 1.9 26.2 14.5  5.8 0.6 1.1 0.4
6 29 176  88.4  96 11.2  7.5 2.7 31.4 19.8  8.3 0.7 1.0 0.4
```

Here, (LF) stands for the (L)eg (F)at, and so on. An additional covariable, the body mass index (B) has been calculated; it is a very popular score normalising the weight by the height. Overall we have five covariables plus nine (3×3) variables for $n = 100$ individuals.

Comparing a collection of crossed BNs

Using the crossed structure of the nine variables to perform the prediction, we will try to improve on the one given by the standard multivariate multiple regression model from Equation (2.40). To do so, we randomly split our data set into two subsets of size $n = 50$, using one for estimating any model and the second one to assess the prediction quality of the model, without repeating this process. As we are doing a multivariate prediction, there are several ways to score the prediction quality. We will use a simple one, more precisely for each individual to predict, we obtain a Normal distribution defined by its expectation (based on the regression formula and the corresponding covariable values) and its standard deviation. The difference between the observed value and the expectation is the bias (B_i^v, for the individual i and variable v); the squared standard deviation is the variance of the prediction ($(\sigma^v)^2$). To obtain global scores, we will define a global bias, a global standard deviation and a global standard error of prediction by summing them up as follows:

$$\begin{aligned} |Bias| &= \sum_{v=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} |B_i^v| \right), \\ Sd.Dev. &= \sum_{v=1}^{p} \sigma^v, \\ SEP &= \sum_{v=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} \sqrt{|B_i^v|^2 + (\sigma^v)^2} \right). \end{aligned} \tag{2.53}$$

In addition to these global quality scores, we measured the parametric dimensions with the number of arcs linking the covariables to the variables ($c2v$: the number of retained regression coefficients) and the number of arcs between pairs of variables ($v2v$: related to the complexity of the correlation structure within the variables). Also, we introduced $f.v2v$, the parametric dimension of the covariance matrix, which is smaller in case of equality constraints on the regression parameters.

A systematic series of non degenerated crossed BNs were attempted among all possible 25 DAGs within the three compartments (F, L, B) crossed with all possible 25 DAGs within the three segments (T, L, A), and for each one the four constraint types proposed in Table 2.2. Table 2.3 shows the results for the best twelve together with the corresponding results for the saturated model. Some interesting features can be noticed from this table:

- The selected crossed BNs perform better than the saturated model globally (SEP) and for both the bias and the variance, the improvement being greater for the variance.

Table 2.3: Quality prediction of the saturated model and the best 12 found crossed BNs. Bn-c and BN-s are the coding of the generating BNs for the two series (compartment and segment). The constraint type refers to Table 2.2. $|Bias|$, $Sd.Dev$ and SEP are defined in (2.53). $c2v$ is the number of arcs from a covariable to a variable; $v2v$ is the number of arcs from a variable to another variable; $f.v2v$ is the parametric dimension of the $v2v$ arcs (the number of constraints have been subtracted).

model	BN-c	BN-s	constraint type	$\|Bias\|$	$Sd.Dev$	SEP	$c2v$	$v2v$	$f.v2v$
saturated	-	-	-	5.97	7.51	10.19	45	36	36
1	22	14	F.C	5.82	6.98	9.66	21	15	9
2	12	14	F.C	5.82	6.99	9.67	19	12	8
3	3	14	F.C	5.82	7.00	9.68	21	9	7
4	9	14	F.C	5.83	7.01	9.69	20	12	8
5	12	6	F.C	5.80	7.05	9.70	20	9	5
6	22	6	F.C	5.80	7.05	9.70	22	12	6
7	12	14	C.C	5.81	7.05	9.72	19	12	4
8	22	14	C.C	5.82	7.06	9.73	21	15	5
9	3	6	F.C	5.82	7.07	9.73	23	6	4
10	9	6	F.C	5.83	7.07	9.74	22	9	5
11	11	14	F.C	5.88	7.03	9.74	19	12	8
12	12	6	C.C	5.81	7.09	9.74	20	9	3

- The reduction in the number of parameters with respect to the saturated model is striking, especially for the variance parameters.
- None of the selected models is without constraints (type F.F).
- For all the selected models, constraints on the three segments (Trunk, Legs, Arms) are present.
- For the segments, only two generating BNs (numbers 14 and 6) are present among the possible 25 ones. These two generating BNs are nested since 14 is (A $\rightarrow$ T $\rightarrow$ L) and 6 is (A; T $\rightarrow$ L).
- For the compartments, two generating BNs (numbers 22 and 12) are dominating. Also these two generating BNs are nested since 12 is (B $\leftarrow$ L $\rightarrow$ F) and 22 adds to it the arc (F $\rightarrow$ B).

For the sake of the example, consider model 2 from Table 2.3, obtained with the combination of the preponderant generating BNs, the twelfth for the compartments and the fourteenth for the segments. Here are its regression equations (the residuals' standard deviations are reported in parentheses):

```
 A = 54.68  (19.85)
 H = 175.78  (6.188)
 W = -107.974 + 1.095*H  (14)
 B = 53.331 + -0.302*H + 0.32*W  (0.263)
 C = -23.416 + 0.223*A + 0.25*H + 2.453*B  (4.077)
AL = 5.133 + 0.139*W + -0.093*C  (0.687)
TL = -13.854 + 0.131*H + 0.156*W + 1.039*AL  (1.308)
LL = 9.418 + -0.026*A + 0.206*W + -0.125*C + 0.2*TL  (1.244)
AB = -0.359 + 0.004*H + 0.011*AL  (0.049)
TB = 0.219 + -0.003*C + 0.011*TL + 0.915*AB  (0.072)
LB = 0.271 + 0.011*LL + 0.835*TB  (0.097)
AF = 1.899 + 0.003*A + -0.025*H + 0.091*W + -0.387*AL  (0.409)
TF = 104.054 + -0.686*H + 0.918*W + 0.283*C + -2.456*B + 0.432*AF + -0.387*TL  (1.247)
LF = -3.71 + -0.018*A + 0.136*W + 0.248*B + 0.027*TF + -0.387*LL  (1.279)
```

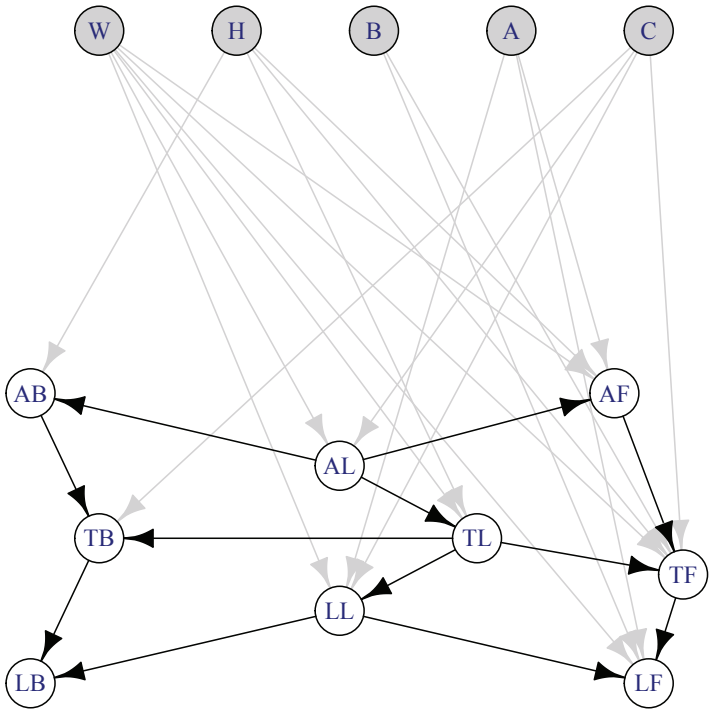


Figure 2.8: DAG associated to Model 2 of Table 2.3. Arcs within covariables were not drawn for the sake of clarity; they are of no importance when conditioning by the covariables.

The corresponding DAG is presented in Figure 2.8. Such a simple model displays good predictive power, and can also be used as a starting point for understanding the phenomenon under investigation. It makes sense for the (L)ean compartment be to the origin of most of the variation compared to the (F)at and (B)one compartments; and that given the (T)runk composition, there is no more correlation between the (A)rm and (L)eg segments.

Discussion

ANOVA models, regression models and their combinations presented in the framework of linear models are versatile tools for analysing complex data sets at the level of big trends, that is at the level of modelling the expectations of random variables (Graybill, 1976). To achieve better predictions, in this framework the common way is to reduce the number of used covariables looking for a small and efficient subset (see Miller (2002) and Celeux and Robert (2006) for a review). The next step is the modelling of variances and covariances. Random models are the natural extensions in that direction; many developments have been proposed in that direction, such as the introduction of variance components in hierarchical models and extensions. For instance, linear models have been imagined to fit the logarithm of the variances (Foulley *et al.*, 2004) in the univariate case. In the multivariate case, we think that BNs, not only GBNs, are appropriate candidates for further proposals. In this study, we showed that two-way structures can be introduced. Of course, there is no limitation to two ways, similar multi-factor approaches

can be devised as well.

We demonstrated that, at least for GBN modelling, it was possible to introduce a known structure on the set of variables of interest, and that can lead to very effective results to obtain an interpretable predictive formula. One of the advantages of the BN formulation is to allow non-statisticians, typically experts in some field, to contribute to model specification through the easy to understand DAG presentations. In our mind, such BNs must and can serve as thinking material for non-experts in BNs. In that respect, the availability of user-friendly and performing software is a prerequisite and we are happy to see that more and more R packages playing this role, are proposed: the most complete and versatile is BNLEARN (Scutari, 2010), but PCALG (Kalisch *et al.*, 2012), DEAL (Bottcher and Dethlefsen., 2012) and IGRAPH (Csardi and Nepusz, 2006) are also worth mentioning.

Besides the introduction of structures on the set of the variables of interest, our study underlines the distinction to make between variables and covariables. One could think that the ideal model would be a model such that the targeted variables be conditionally independent to the covariables. That is all the covariation between them could be explained by external variables. With this respect some of the exhibited models, having a very small parametric dimension (for instance Model 12 of Table 2.3 with 3 instead of 36) are appealing.

Many more ideas could be proposed to achieve the goals we were interested in. Among them, the use of distributions other than Normal ones. Probably mathematical properties will be much more difficult to obtain, but the advantage would be to achieve a more realistic model specification. Advanced numerical tools already exist to undertake such an investigation. Among them, even if not originally devised for this purpose, are the BUGS software packages (Lunn *et al.*, 2013). But also simpler approaches could also be worthwhile, like the use of transformations of the initial variables. More sophisticated constraints than the equalities could also be implemented. For instance, following again a two-way structure, some bilinear modelling could be thought about like those proposed to generalise additive models (Denis and Gower (1996)).

In conclusion, we would like to state that not only BNs are beautiful for mathematical aspects, they are also useful for plenty of applied questions within a classic statistics point of view.

Chapter 3

Application

Accurate predictions of body composition are useful in achieving a greater understanding of human energy metabolism in physiology and in different clinical conditions, and in evaluating interventions. Many disease processes affect bone and soft tissue at the same time. Therefore, the statistical prediction for body composition becomes an attractive technique for a variety of clinical research and practice applications in the large studies.

In section 3.1, we will describe available datasets for our study applications, including an introduction of predictor variables and predicted variables. In section 3.2, we will focus on the multivariate prediction for segmental body composition with a comparison study of locally weighted models. A published work related the multivariate prediction will be presented in subsection 3.2.2, and this work will discuss the usefulness of our proposed modeling in the physiological framework. In section 3.3, we are interested in predicting age-related changes in body composition. A Bayesian modeling method will be first introduced by taking into account anthropometric evolution with aging. Then a Frequentist modeling is investigated, and this method is applied to different cases of study, according to BMI, ethnicity and history of weight change during life. Moreover, the physiological aspects will be discussed.

3.1 Available datasets

In the present study, there are two main datasets at our disposal and they will be used to apply our proposed statistical modeling in body composition prediction. The two datasets are respectively the National Health and Nutrition Examination Survey (NHANES) and a French CHU DXA datasets. These two datasets involve only adult subjects aged 18 y and more. Besides, a third dataset from a medical examination center at Saint-Brieuc (France) is considered. It involves only gender, age, height, weight and waist circumference observations, and DXA-measured body compositions are not available. The NHANES dataset is the principal dataset in the application study because of its significance, moreover it contains waist circumference information. The French CHU dataset is used as an external validation dataset, this will help us to get more idea about the generalisability of our proposed model. With respect to the Saint-Brieuc dataset, as its sample size is very large (*i.e.*, 11500 men and 11962 women), we can obtain a relevant sample size for narrow age interval (see Table (3.10) for details); therefore it is intentionally used to assess age-related changes in the anthropometric variable, and to build the corresponding age-related functions. It is worth indicating that none of statistical predictions are performed on this datast. According to the significance of the datasets, we will first describe the NHANES dataset, following by the French CHU DXA dataset.

3.1.1 NHANES dataset

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. The NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation. Began in the early 1960s, the survey has been conducted as a series of surveys focusing on different population groups or health topics. The survey examines a nationally representative sample of about 5000 persons each year. These persons are located in states across the country, 15 of which are visited each year. To produce reliable statistics, NHANES oversamples persons 60 and older, as well as African Americans, Asians, and Hispanics.

The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel. Findings from this survey will be used to determine the prevalence of major diseases and risk factors for diseases. Information will be used to assess nutritional status and its association with health promotion and disease prevention. NHANES findings are also the basis for national standards for such measurements as height, weight, and blood pressure. Data from this survey will be used in epidemiological studies and health sciences research, which help to develop sound public health policy, direct and design health programs and services, and expand the health knowledge for the Nation.

The NHANES has been split into several datasets. Each dataset contains the information for a specific fied of two successive annual surveys. These datasets can be download from www.cdc.gov/nchs/ and allow a free access to another datasets, such as : anthropometry, biochemistry, body composition, demographics, clinical indicators (*e.g.*, cardiovascular diseases, infectious diseases, kidney or sexually transmitted diseases, physical activity, diabetes, nutrition, obesity, osteoporosis ...).

In the present study, we only use a NHANES DXA dataset from 1999–2004 period. Subjects are characterised by predictor variables, such as gender, ethnicity, age, height, weight and waist circumference. A preliminary data processing has been conducted. As a result, we select subjects aged 20-85 year, with BMI values ranging from 18 to 40 kg/m^2 and who belonged to one of the three considered ethnicity categories : White, Black and Hispanic. This selection results in a sample size of 3977 men (1984 White, 720 Black and 1273 Hispanic) and 3692 women (1830 White, 697 Black and 1165 Hispanic). The present study was always conducted separately on men and women, in addition, as we mentioned at beginning, the whole NHANES dataset is randomly split into three subsets; therefore, the complete NHANES dataset is first split by gender, then for each gender, we randomly split the corresponding NHANES dataset into three subsets : a training dataset (TRD); a test dataset (TED); a validation dataset (VAD). The training subset is considered as the reference dataset and it is particularly used to provide weighted subjects in the locally weighted approaches. The test subset is used to evaluate the parameters values in different locally weighted approaches, the optimal parameters values are determined by the accurate prediction in the test subset. Finally validation subset is used to make model selections between the models.

3.1.2 French CHU dataset

Now we will give a brief description of the French CHU DXA dataset. The French CHU dataset compris 1140 French subjects aged between 20 and 79 years and with BMI between 18.5 and 40 kg/m^2. The individual body composition is measured by DXA (Hologic QDR-4500) during routine examination at the Radiology Department of the Clermont-Ferrand University Hospital Centre (CHU) between 1998 and 2008. However, ethnicity was not mentioned and waist circumference was not measured during the examination. The French CHU DXA dataset was used as an external validation dataset. This French DXA dataset is independant of the NHANES validation subset, and the use of this external validation allows to get another idea of the prediction accuracy in a different population context, therefore we believe that it will make the comparison study more revelant.

3.1.3 Predicted variables and predictor variables

As the predicted variables, the segmental body composition (SBC) is a proportion of body weight, it is convenient to decompose the body weight into fat, lean and bone masses in different segments. Table (3.1) summarizes possible segmental body compositions to be assessed. Nevertheless we are mainly interested in 9 SBCs in the present study, and these 9 of the SBCs are red colored in Table (3.1).

Table 3.1: Summary of possible segmental body compositions (kg). We have appendicular=arm+leg, *e.g.*, apF = aF + lF.

	F(at)	**L**(ean)	**B**(one)	**F**(at)**F**(ree)	**W**(eight)
a(rm)	aF	aL	aB	aFF	aW
t(runk)	tF	tL	tB	tFF	tW
l(eg)	lF	lL	lB	lFF	lW
ap(pendicular)	apF	apL	apB	apFF	apW
b(ody)	bF	bL	bB	bFF	bW

As we mentioned previously, the statistical prediction will be separately conducted by men and women; therefore the gender is not considered as a predictor variable in the models. Height and weight have been justified as suitable predictor variables, and they can be used to well predict body composition together with age (Mioche *et al.*, 2011a,b). Waist circumference is well-known indicator of health risk, especially when used in combination with some height-to-weight indexs (*e.g.*,BMI) (Janssen *et al.*, 2002). Thus waist circumference will also be considered as a predictor variable. Table (3.2) presents the means and standard deviations of age, anthropometric variables and DXA segmental body compositions for the different datasets in both genders. It is necessary to remind that waist circumference is not available in the French CHU dataset, therefore when we apply the statistical models in this validation dataset, the statistical models will take only age, height and weight as predictor variables.

Table 3.2: Age, anthropometric variables and dual-energy X-ray absorptiometry body composition characteristics for men and women in the National Health and Nutrition Examination Survey (NHANES) training dataset (TRD), test dataset (TED) and validation dataset (VAD) and in the French CHU dataset (French CHU).

		NHANES TRD	NHANES TED	NHANES VAD	French CHU
Men	Number	1989	994	994	526
	Age(years)	50.6±18.91	51.16±19.33	50.6±18.63	46.61±17.02
	Height (cm)	174.02±7.81	174.12±7.68	174.06±8	174.62±6.79
	Weight (kg)	83.9±15.2	84.07±16.21	83.93±15.98	78.83±13.08
	Waist circumference (cm)	98.62±12.62	98.78±12.96	98.58±12. 48	-
	Arm fat (kg)	2.4±0.96	2.38±0.99	2.37±0.95	1.96±0.85
	Trunk fat (kg)	11.21±4.98	11.26±5.1	11.17±4.87	8.76±4.48
	Leg fat (kg)	6.55±2.5	6.55±2.61	6.45±2.62	5.56±2.16
	Appendicular fat(kg)	8.95±3.34	8.94±3.48	8.81±3.46	7.52±2.89
	Body fat (kg)	21.21±8.03	21.24±8.3	21.02±8.07	17.34±7.04
	Arm lean (kg)	7.61±1.46	7.59±1.51	7.66±1.52	6.87±1.19
	Trunk lean(kg)	29.74±4.46	29.86±4.78	29.79±4.61	29.27±4.4
	Leg lean (kg)	19.14±3.43	19.18±3.58	19.23±3.67	19.01±3.09
	Appendicular lean (kg)	26.75±4.74	26.77±4.97	26.89±5.04	25.87±3.95
	Body lean (kg)	60.06±9.03	60.2±9.65	60.26±9.56	58.82±7.64
	Body bone (kg)	2.63±0.44	2.63±0.44	2.64±0.45	2.67±0.41
	Body fat-free mass (kg)	62.7±9.35	62.83±9.97	62.9±9.89	61.49±7.9
Women	Number	1846	923	923	569
	Age (years)	51.64±18.92	52.09±18.79	51.11±18.08	49.28±14.84
	Height (cm)	160.71±6.86	160.77±6.77	160.57±6.86	161.93±6.64
	Weight (kg)	71.56±14.24	72.86±14.5	72.62±14.31	67.66±13.46
	Waist circumference (cm)	92.42±12.69	93.50±12.99	93.46±12.88	-
	Arm fat (kg)	3.33±1.25	3.48±1.31	3.42±1.26	2.79±1.12
	Trunk fat (kg)	12.83±4.92	13.27±5.21	13.32±5.11	10.36±4.88
	Leg fat(kg)	10.03±3.35	10.3±3.41	10.07±3.35	9.24±3.07
	Appendicular fat (kg)	13.36±4.34	13.78±4.46	13.48±4.34	12.03±3.89
	Body fat (kg)	27.08±8.74	27.93±9.15	27.69±8.95	23.3±8.23
	Arm lean (kg)	4.35±0.86	4.39±0.87	4.41±0.9	3.98±0.84
	Trunk lean (kg)	21.56±3.2	21.86±3.15	21.84±3.18	21.68±3.97
	Leg lean (kg)	13.55±2.73	13.64±2.74	13.63±2.67	13.53±2.78
	Appendicular lean (kg)	17.9±3.48	18.02±3.49	18.04±3.44	17.51±3.35
	Body lean (kg)	42.47±6.59	42.89±6.56	42.89±6.52	42.29±6.65
	Body bone (kg)	2.01±0.37	2.04±0.37	2.03±0.35	2.07±0.34
	Body fat-free mass (kg)	44.49±6.84	44.93±6.81	44.93±6.75	44.36±6.87

3.2 Segmental body composition prediction

Body composition is closely related to health in both individuals and populations. The ongoing epidemic of obesity in adults (and children) has highlighted the importance of body fat for short term and long term health. Moreover, other components of body composition also influence health outcomes, and their measurements are increasingly considered valuable in clinical practice. Accurate measurements of body composition can be obtained from different methods, such as BIA and DXA. However, they require fixed equipment, and particularly they are time consuming and expensive. As a result, they are not convenient relevant for use as a part of routine clinical examinations or population studies.

In this part, we will assess different statistical models for segmental body composition prediction. In subsection 3.2.1, three locally weighted models will be described, following by a comparison study of prediction performance with the published models. The usefulness of waist circumference as predictor variable is also investigated, but the results will be reported directly in a published paper (see section 3.2.2). In subsection 3.2.2, a global multivariate model will be presented, and this work led to the publication of one research paper in the *British Journal of Nutrition*.

3.2.1 Comparison of different locally weighted approaches

In the previous section 2.2, we introduced the principes of our proposed locally weighted approaches, such as locally weighted SVMR. Each locally weighted approach has its proper advantage in different applications. To our knowledge, until now, no locally weighted approach with regression models like our proposals has been studied for body composition assessment, except Mioche *et al.* (2011a)'s work using a more likely non-parametric way. The aim of this part is to investigate the prediction ability of our proposals for segmental body composition (SBC), and to compare these approaches between themself, and also with the reference published models. The reference published models are considered because they are used to provide the baseline of the prediction accuracy for our proposals, hence this could make the comparison study more suitable. However it is worth mentioning that these models are univariate models, and they predict mainly percent of body fat. We will organize this part in the following way : the reference models are first described, then three proposed locally weighted aproaches will be presented, finally a comparison study is performed by using two different validation datasets.

The potential uses of statistical methods for body composition assessment have been highlighted (Snijder *et al.*, 2006), and several attempts to predict body composition, particularly body fat percentage (bF%), using linear models with simple predictor variables have been made. A summary of the body composition prediction models published between 1985 and 2003 has been given by Sun and Chumlea (2005). They pointed out that (1) a general model for both two genders, different ethnicities and wide age ranges may lose its accuracy due to increased heterogeneity; (2) cross-validation of prediction models was needed to assess the degree of their validality; (3) for validation studies, accuracy should be standardised for the mean of the predicted variable; (4) few prediction models were derived from datasets using DXA.

In the present study, three published models are retained and considered as the reference models. The first two models are linear models and the third one is a non-parametric model. The common aspect of these three models consist in using age and simply acquired anthropometric covariates to predict body composition. The first model was proposed by Gallagher *et al.* (2000a), and their model was derived from a DXA dataset. It is worth noting that they used a surrogate of obesity, BMI, as a covariate rather than the anthropometric covariates. The

model reads :

$$\begin{aligned} bF\% &= 76 - 1097.8 \times \frac{1}{BMI} - 20.6 \times SEX + 0.053 \times AGE \\ &+ 95 \times Asian \times \frac{1}{BMI} - 0.044 \times Asian \times AGE \\ &- 0.044 \times Asian \times AGE + 154 \times SEX \times \frac{1}{BMI} \\ &+ 0.034 \times SEX \times AGE \end{aligned} \tag{3.1}$$

where $SEX = 1$ for men and 0 otherwise; $Asian = 1$ for Asian subjects and 0 otherwise.

Later, to investigate the effects of sex, age and race on the relation between BMI and measured bF%, Jackson *et al.* (2002) established another linear model. More precisely, their model was derived from a densitometry dataset gathering from four clinical centres, and the model took into account only the covariate BMI :

$$\begin{cases} bF\% = 3.76 \times BMI - (0.04 \times BMI^2) - 47.80 & \text{for men} \\ bF\% = 4.35 \times BMI - (0.05 \times BMI^2) - 46.24 & \text{for women} \end{cases} \tag{3.2}$$

Recently, Mioche *et al.* (2011a) developped a non-parametric model derived from a DXA dataset (NHANES 1994-2004). This non-parametric model was based on a probabilistic Bayesian network (BN), and it included sex, age, body weight and height as covariates. The key idea is to, for each subject to be predicted, (1) select a subset of similar subjects in the reference dataset with respect to covariables, (2) and to make the prediction based on the subset, instead of the whole reference dataset. There are two main differences from the previous linear models. The first one is that the prediction made from this non-parametric model is at the individual level. Specifically, for the subject to be predicted, from his/her given values of the four covariates (sex, age, weight and height), a BN model was obtained from an extracted dataset where training subjects had similar covariate values. The second difference is that Mioche *et al.* (2011a)'s model has potential to perform simultaneously a multivariate prediction. The more comprehensive model description and its further applications can be found in Mioche *et al.* (2011a,b). Hence we only give a brief formulation of their model. Assume that Y be a segmental body composition to be predicted, BN model consists in establishing a relationship as follows :

$$Y = f_{BN}(SEX, AGE, HGT, WGT, D) \tag{3.3}$$

where $f_{BN} : (SEX, AGE, HGT, WGT) \times (w, d) \rightarrow R$. The selection of similar subjects in the reference dataset is controled by parameters w and d : $w = (w_a, w_h, w_w)$ is a vector of associated weighting parameters for age, height and weight, respectively; $d = (d_a, d_h, d_w)$ is a vector of the absolute values of the difference for age, height and weight, respectively, between the subject to predict and the training subjects of the same gender. Also a maximal distance (D_{max}) is defined as the maximal selection limit :

$$D_{max} = \frac{\max(w_a d_a, w_h d_h, w_w d_w)}{w_a + w_h + w_w} \tag{3.4}$$

Only the subjects in the reference dataset with $D < D_{max}$ are retained as candidates for prediction. When the closest subject in the reference dataset has a distance greater than D_{max}, the predictive subset is empty, there no prediction is made. In Mioche *et al.* (2011a), the optimal value of w and d are given based on a simulation study.

Now we will present the proposed three locally weighted approaches. The key idea of locally weighted concept is described in section 2.2. Briefly speaking, the locally weighted process allows to attribute varying importance of each subject in the reference dataset, and this importance value (between 0 and 1) is based on the discrepancies between the reference subjects and the subject to be predicted with respect to the covariables. After weighting each subject by its own importance value, a weighted dataset is obtained, then a statistical model is built for the body composition prediction at the individual level. The statistical models that we consider are respectively the linear, SVM and Bayesian linear model. We denote them locally weighted linear model (LWL), locally weighted SVM regression model (LWSVMR) and locally weighted Bayesian linear model (LWB), respectively.

Assume that, for the subject to be predicted, $\mathbf{Y}$ corresponds to a vector of q segmental body compositions (SBCs) to be predicted, LWL is formulated :

$$\mathbf{Y} = f_{LWL}(AGE, HGT, WGT, WAI, w) = f_{LWL}(\mathbf{X}, w) \tag{3.5}$$

where $\mathbf{X}$ observation matrix ($n \times 4$), n number of observations; $f_{LWL} : (AGE, HGT, WGT, WAI) \times w \to R^q$; w a n vector of weighting associated to each subject in the reference dataset.

As mentioned in section 2.3.3, it is possible to include the weighting values in the SVM regression (Lin and Wang, 2002). Assume that, for the subject to be predicted, Y is a component in $\mathbf{Y}$, $\mathbf{x} = (AGE, HGT, WGT, WAI)$ the vector of the covariate values. LWSVMR reads :

$$\begin{aligned} Y &= f_{LWSVMR}(AGE, HGT, WGT, WAI, w) \\ &= \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) + b \text{ , SVMR with a kernel function } K(\cdot) \end{aligned} \tag{3.6}$$

where $f_{LWSVMR} : (AGE, HGT, WGT, WAI) \times w \to R$ and i the number of subjects in the reference dataset. First of all, a grid research for locally weighted SVM model parameters was performed (not shown) using the NHANES test subset, this study allowed to select the optimal kernel function and its associated parameters, as well as the cost and the ν value in the ν-SVMR. An optimal combination of parameters is chosen based on the most accurate prediction in the NHANES test subset. After all, the polynomial kernel is retained depending on three parameters (γ, c, d) :

$$K(\mathbf{u}, \mathbf{v}) = (\gamma \times \mathbf{u}^T \mathbf{v} + c)^d \tag{3.7}$$

where $\mathbf{u}$ and $\mathbf{v}$ are two vectors. Table (3.3) summarizes the retained combination of parameters for our proposed locally weighted SVM model.

Table 3.3: Summary of parameters in locally weighted SVM model.

Parameter	Value
SVMR type	ν-SVMR
Kernel function	Polynomial kernel
γ	0.0008
c	0
d	2
cost (C)	5
ν	0.8

Finally, locally weighed Bayesian linear model is proposed. Here we will give a formal formulation of LWB. Assume that, for the subject to be predicted, $\mathbf{Y}$ follows a multinormal distribution :

$$\mathbf{Y} \sim \mathcal{N}(\mu_{\mathbf{Y}}, \Sigma_{\mathbf{Y}}) \tag{3.8}$$

$\mu_{\mathbf{Y}}$ is mean vector of $\mathbf{Y}$ and $\Sigma_{\mathbf{Y}}$ is variance-covariance matrix. Also we suppose that the q SBCs are independent from each other, because there are the predictor variables will be introduced, and they are assumed to take into account most of correlation between the SBCs; therefore $\Sigma_{\mathbf{Y}}$ is diagonal matrix. For each component Y in $\mathbf{Y}$, LWB reads :

$$\begin{array}{rcl} Y & = & f_{LWB}(WGT, AGE, HGT, WAI, w) \\ p(Y|WGT, AGE, HGT, WAI, w) & \sim & N(\alpha_W \cdot WGT + \rho_A \cdot AGE \\ & & +\rho_H \cdot HGT + \rho_C \cdot WAI, \quad \omega \times \sigma_Y^2) \end{array} \tag{3.9}$$

where $\alpha_W, \rho_A, \rho_H, \rho_C$ and σ_Y^2 are the parameters, and ω is the weighting value associated to each subject. In the Bayesian framework, we have to specify the prior distribution of the parameters. Frequently, the coefficient parameters are supposed to follow a normal distribution, and the standard deviation parameter σ follows a uniform distribution, we retained :

$$\begin{array}{rcl} \alpha_W \sim \mathcal{N}(\mu_{\alpha_W}, s^2_{\alpha_W}) & , & \rho_A \sim \mathcal{N}(\mu_{\rho_A}, s^2_{\rho_A}) \\ \rho_H \sim \mathcal{N}(\mu_{\rho_H}, s^2_{\rho_H}) & , & \rho_C \sim \mathcal{N}(\mu_{\rho_C}, s^2_{\rho_C}) \\ \sigma_Y \sim \mathcal{U}(a, b) & & \end{array}$$

For the sake of brevity, we only give the retained prior distributions of the parameters for some SBC. For instance, Table (3.4) shows the prior distributions of the parameters for predicting body fat mass. As we see here, means of ρ_A, ρ_H and ρ_C are zero, that is because we want to make these three parameters few informative, while mean of α_W is not zero, because the SBC is part of body weight, but it is not important to give an accurate ratio, therefore we feel suitable to suppose 0.5 as expectation for α_W.

Table 3.4: Proposed prior distributions of the parameters for predicting body fat mass.

Node	Parents	Associate parameters	Prior distribution
bF	WGT	α_W	$\mathcal{N}(0.5, 10^2)$
	AGE	ρ_A	$\mathcal{N}(0, 25^2)$
	HGT	ρ_H	$\mathcal{N}(0, 20^2)$
	WAI	ρ_C	$\mathcal{N}(0, 17^2)$
		σ_{bF}	$\mathcal{U}(0, 4)$

In our locally weighted approach, to calculate the similarity between training subjects and the subject to be predicted, we retained the following distance function :

$$\mathcal{D}_{sum}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{p} \omega_k^{cov} \times |x_k - y_k| \tag{3.10}$$

where $\mathbf{x}, \mathbf{y} \in R^p$ and they are vectors of covariates for two distinct subjects, $\omega^{cov} = (\omega_1^{cov}, \ldots, \omega_p^{cov})$ is a vector of associated weighting parameters for p covariates. Table (3.5) summarizes the retained weighting parameter values associated to each locally weighted model.

Table 3.5: Optimal covariate weightings.

Covariates	LWL	LWSVMR	LWB
Age (ω_a^{cov})	0.1	0.1	0.002
Height (ω_h^{cov})	0.01	0.01	0.001
Weight (ω_w^{cov})	0.01	0.01	0.025
Waist (ω_c^{cov})	0.035	0.035	0.035

Gallagher *et al.* (2000a)'s and Jackson *et al.* (2002)'s model predict only bF%, their prediction of bF is then obtained by multiplying the predicted value of bF% by body weight. Moreover, we can deduce their indirect prediction for bFF by substracting body weight by predicted value of bF ($bFF = bW - bF$). To be fair for Gallagher *et al.* (2000a)'s and Jackson *et al.* (2002)'s models estimated from different datasets, adjusted formulas were derived by re-estimating the parameters of their models using the NHANES training. Then original and adjusted formulas are applied to the NHANES validation and French CHU datasets. Besides, it is worth noting that for Gallagher *et al.* (2000a)'s and Jackson *et al.* (2002)'s models, either original or adjusted, only the prediction accuracy for bF and bFF are calculated. As waist circumference is not available in the French CHU dataset, therefore we can only use three covariates to predict these 9 SBCs. The accuracy of a prediction for a given SBC is globally assessed by the SEP1 and REP1 criteria[1].

Figure 3.1 and 3.2 show SEP1 for 9 SBCs from the reference original, adjusted published models and the proposed locally weighted models in NHANES men and women, respectively. In both genders, the three locally weighted models provide a better quality of the prediction than two original published models. When comparing with the reference adjusted models, for both gender, the three locally weighted models enable to provide more accurate prediction for bF and bFF, except that for women in the NHANES validation subset, the adjusted Gallagher *et al.* (2000a)'s model yields a close SEP1 value to that from LWL, LWSVMR and LWB. Furthermore, despite adjusted Mioche *et al.* (2011a)'s model has similar SEP1 value with locally weighted models for body bone, appendicular lean and trunk lean masses, for other compartments, the three locally weighted models still have more accurate prediction. Globally, within the three locally weighted models, their prediction accuracies are at the same level, nevertheless a little more accurate predictions are provided by LWL and LWSVMR in bF and apL for the NHANES validation men.

SEP1 value gives the absolute prediction accuracy for each SBC, and it is not possible to tell which SBC has the best prediction accuracy within them. To investigate this question, we can use a relative criterion, REP1. Figure 3.4 shows REP1 for 9 SBCs from different models for the NHANES validation subset. The performances of prediction between models are very similar to that resulted of SEP1 criterion, nevertheless the discrepancies are reduced on account of a relative expression of SBCs. When comparing between the 9 SBCs, we observe that all models yield the best prediction accuracy in body lean mass, following by trunk and appendicular lean masses. Less accurate predictions are found in body and trunk fat masses, as well as in body bone mass, even tougth it has a small amount.

Similar results for SEP1 are found in the French CHU dataset (Figure 3.3), though, for men in the French CHU dataset, the differences of SEP1 are not so important as those in other cases. This could be explained by the absence of waist circumference as predictor variable. Within the three locally weighted models, LWL and LWSVMR are more accurate than LWB, particularly

[1]Definition and more details about the criteria can be found in section 1.1.2.

for bL and bFF prediction. Interestingly, we observe that for women in the French CHU dataset, LWL yields the best accuracy for most of SBCs, especially for tF. In addition, Figure 3.5 shows corresponding REP1 criteria for 9 SBCs in the French CHU dataset. The discrepancies for bF prediction between the reference original and the locally weighted models almost disappear, particularly in men. This could be explained by the absence of waist circumference as predictor variable.

Figure 3.1: For men in NHNAES validation subset, the prediction accuracy criterion SEP1 for 9 SBCs from different models.

Different colors, line types and point characters are associated to different models. Detailed legend is given on the bottom subfigure. LWL : Locally Weighted Linear model; LWB : Locally Weighted Bayesian model; LWSVMR : Locally Weighted SVM Regression model. The SEP1 values for 9 SBCs are connected to have better presentation, because for some SBCs, the SEP1 values from models are very close and it is difficult to identify the differences.

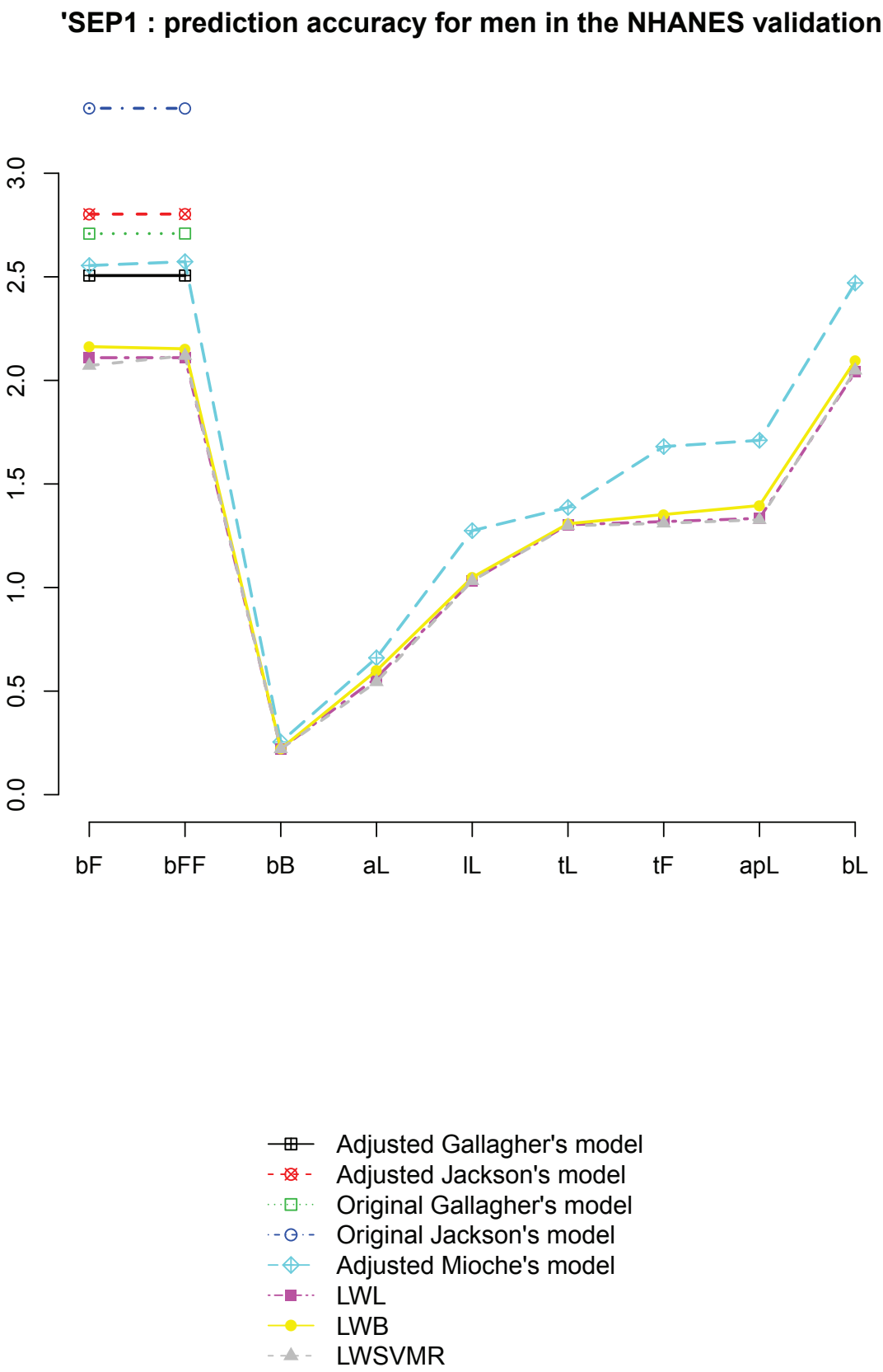


Figure 3.2: For women in NHNAES validation subset, the prediction accuracy criterion SEP1 for 9 SBCs from different models. c.f. Figure 3.1 for legend details.

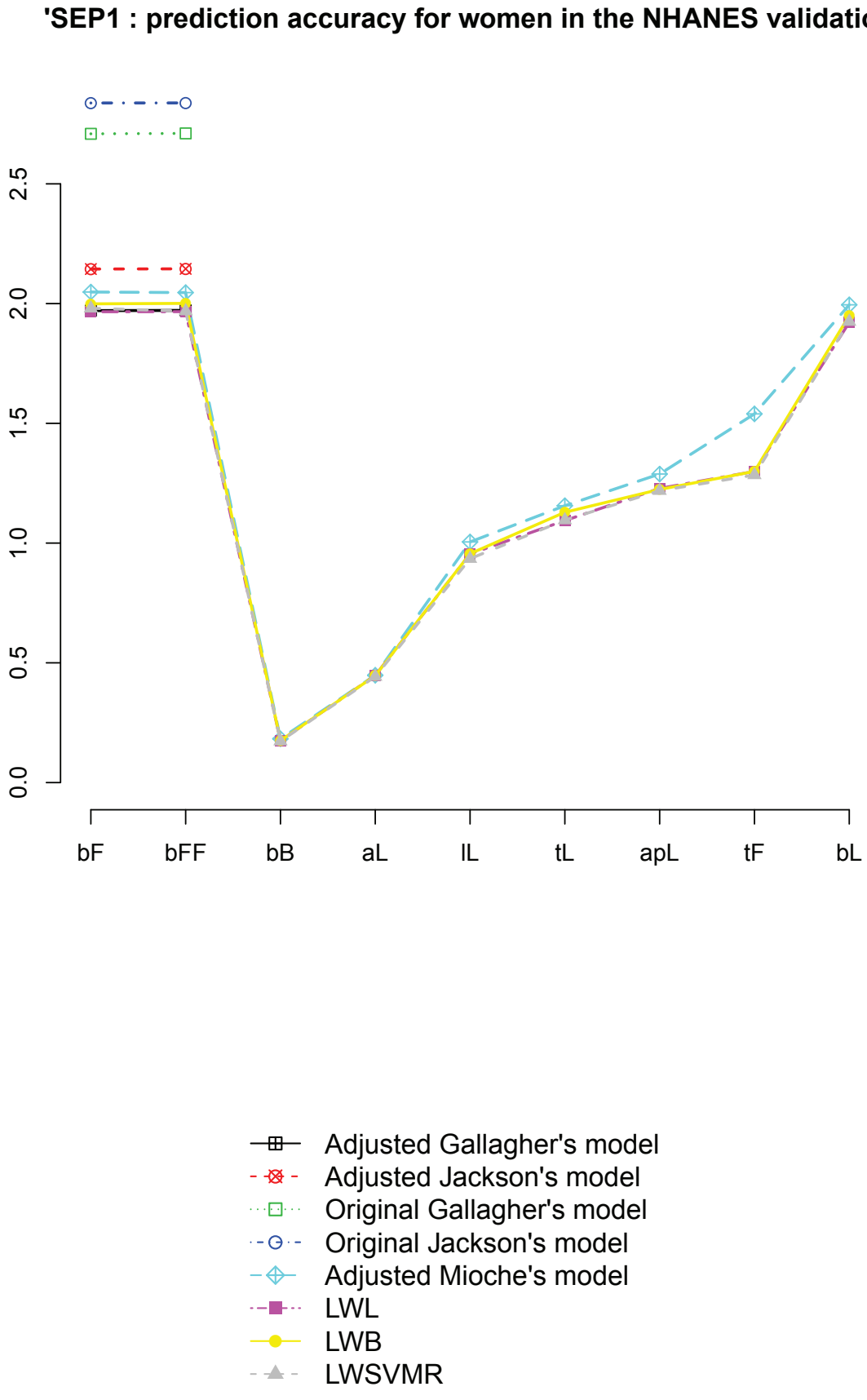
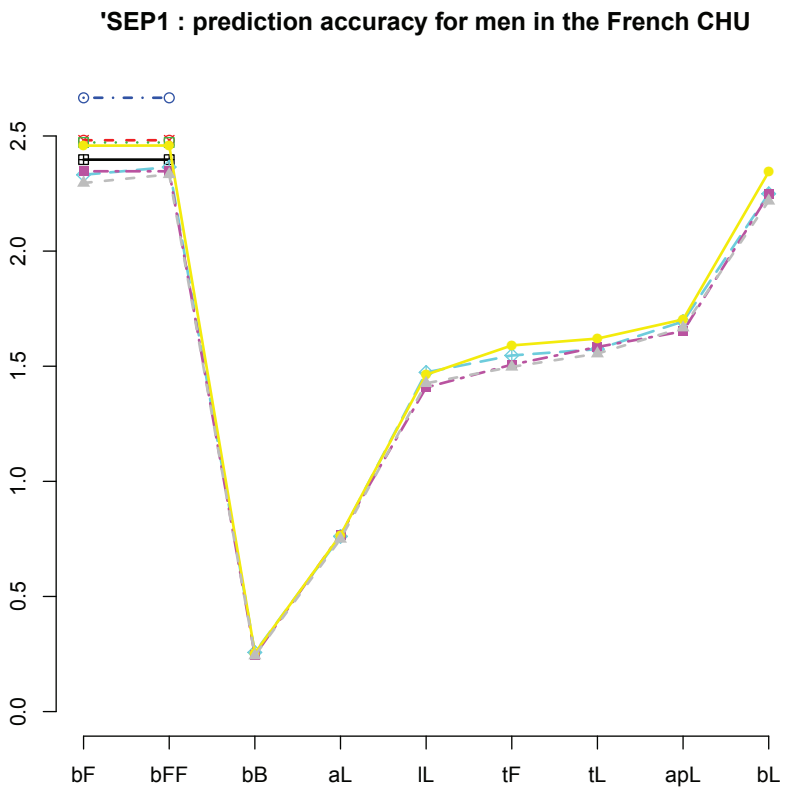


Figure 3.3: For both genders, the prediction accuracy criterion SEP1 for 9 SBCs from different models in French CHU dataset.



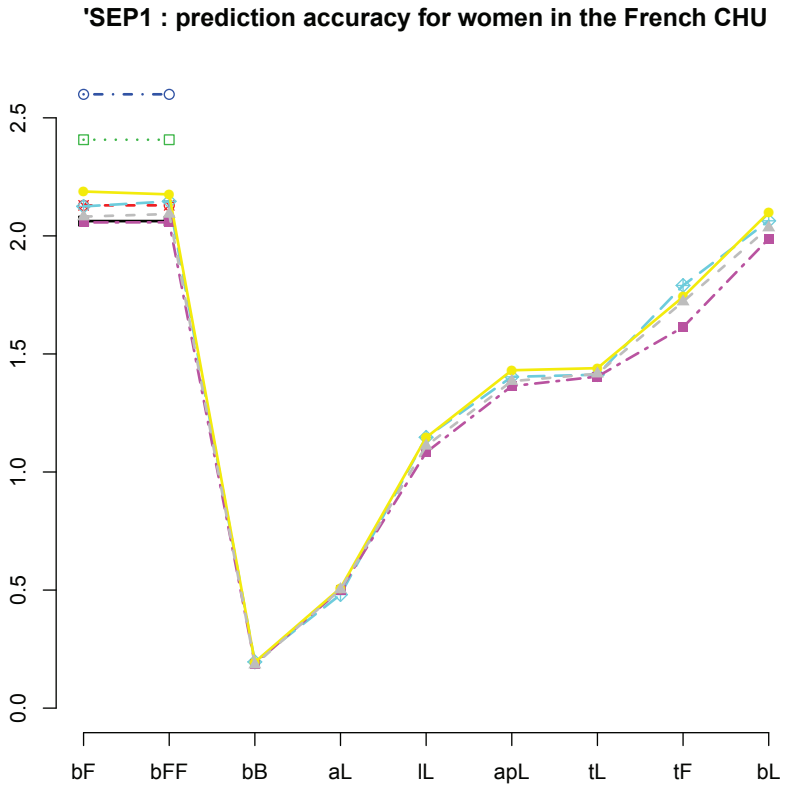
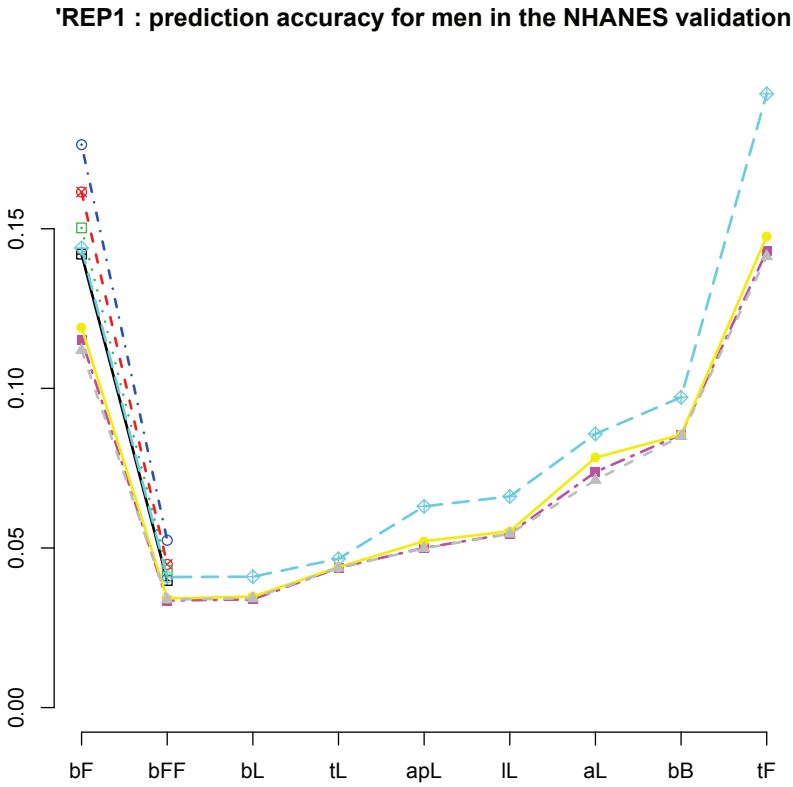


Figure 3.4: For both genders, the prediction accuracy criterion REP1 for 9 SBC from different models in the NHANES validation subset. c.f. Figure 3.1 for legend details.



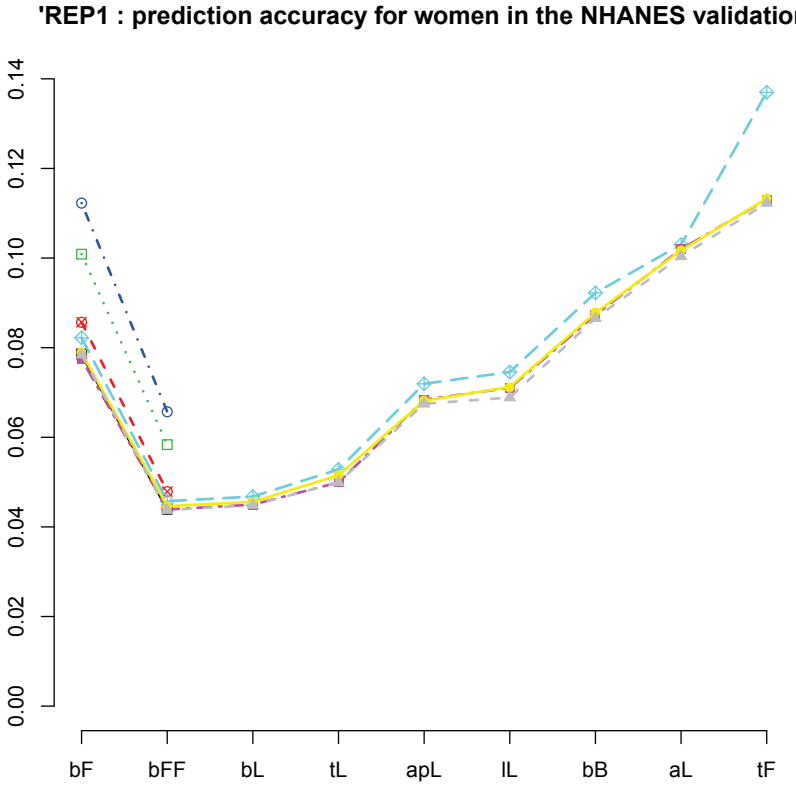
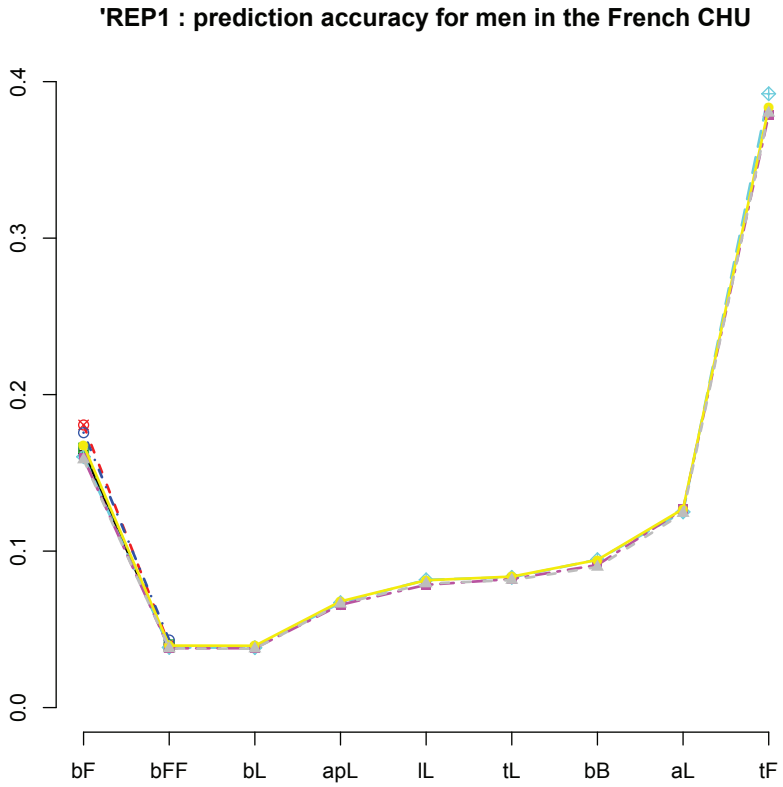
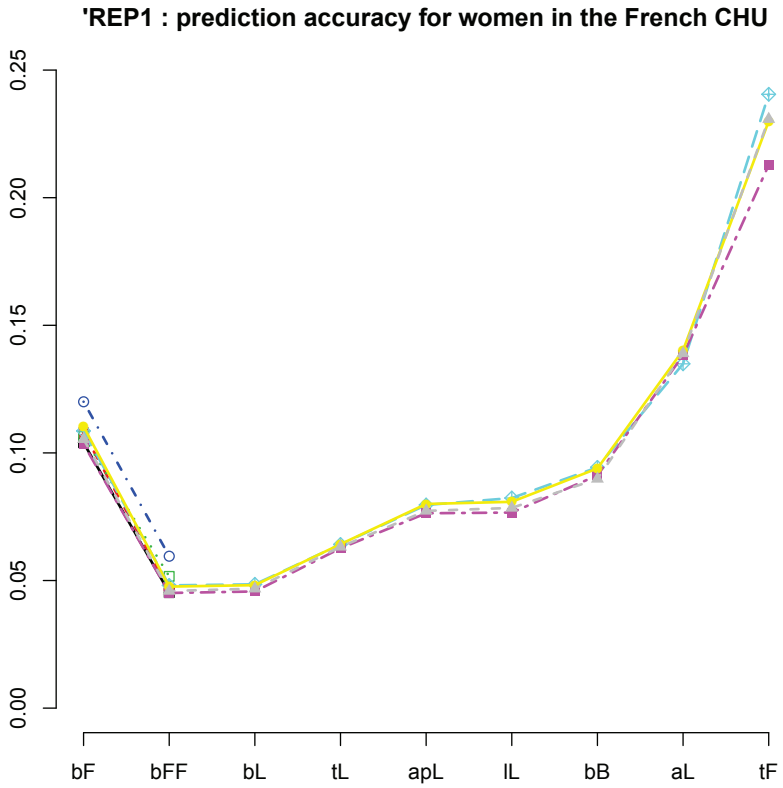


Figure 3.5: For both genders, the prediction accuracy criterion REP1 for 9 SBC from different models in the French CHU dataset. c.f. Figure 3.1 for legend details.





We also extended the comparison study from a global level into a categorical levels. More precisely, we attempted to investigate the prediction accuracy in three BMI and four age categories. The corresponding three BMI categories are respectively [18,25[, [25,30[and [30,40[kg/m^2, and the four age categories are [20,35[, [35,50[, [50,65[and [65,80[years. The accuracy of prediction in different BMI and age categories are shown in Figure 3.6 - 3.7. In each figure, SEP1 and REP1 are simultaneously represented, and it is necessary to remind that the y-axis of REP1 subfigure is inverse, which means that lower value is on the top and the higher value is on the bottom.

For the sake of understanding of these complicated figures, we will give a detailed comprehensive discussion on Figure 3.6 :

- It shows bF prediction of different models in different BMI and age categories for the NHANES validation men.
- With respect to the BMI categories, for all models, the SEP1 values increase when BMI increasing, that emphasizes that a less accurate predictions are provided for obese subjects (*i.e.*, BMI category is $[30, 40[$).
- Moreover, for each model, the SEP1 values in the first two BMI categories are smaller than the global level.
- With respect to the age categories, the original and adjusted published models provide mainly better prediction accuracy in the older age categories than in the younger age categories. In particular for subjects aged 50 y and over, the published models (either original or adjusted) enable to give a smaller SEP1 values than the global level.
- By contrary, the three locally weighted models yield more accurate bF prediction in the younger age categories.
- When comparing with the published models in different BMI and age categories, the three locally weighted models still provide better bF prediction accuracy, as done at the global level.
- Interestingly, according to REP1 criteria, we find a best bF prediction accuracy in the high BMI category, as well as in the older age categories for the three locally weighted models.

Regarding to other SBCs in the two validation datasets for both genders (Figure 3.7 - 3.10), the results are summarized as follow :

- Regarding to men in the NHANES validation subset, an accurate predictions are found for BMI < 30 and younger age categories for all SBC concerned.
- For the women in the NHANES validation subset, the better quality is found in young female subjects for fat and lean masses.
- Compared with the prediction accuracy at the global level, for men in the French CHU dataset, all locally weighted models yield a better quality for BMI < 30 in body and trunk fat masses, whereas only for BMI < 25 in body, trunk and appendicular lean masses.
- Moreover, a better quality is observed for younger male subjects (less than 50 year) in the French CHU dataset.

- With respect to women in the French CHU dataset, all locally weighted models provide a better quality than the global level for BMI < 30 and older age categories (more than 50 year).

Summary

Our proposed locally weighted models enable to yield a suitable quality of the prediction, more importantly they allow to achieve a multivariate prediction for several SBCs. The locally weighted process may have a relevant use when size of the dataset is large (data-rich situation), because for the subject to be predicted, more the candidat subjects in the reference dataset are closer, more they will be taken into account in the prediction, therefore the "outlier" subjects will have less effect of leverage in the prediction. Furthermore, in the multivariate framework, we are able to take into account the dependencies between the predicted variables, then to express them in certain structure form. This consideration leads to incorporate our knowledge into dependency structure, also it has potential to make more accurate prediction. Indeed, section 2.4.3 described this approach and introduced a novel method, Crossed Linear Gaussian Bayesian Networks, in detail. Otherwise, an additional comparison study (not shown) was performed between these three locally weighted models and a global multivariate linear model. The results show that even though our proposed locally weighted models are more accurate than the published linear model (Gallagher *et al.*, 2000a; Jackson *et al.*, 2002), the global multivariate linear model is as relevant as the locally weighted models for body composition prediction. We published our proposed global multivariate linear model for predicting segmental body composition in *Journal of British Nutrition*. The full paper can be found in section 3.2.2.

Figure 3.6: For men in the NHANES validation subset, accuracy of the prediction for body fat (bF) mass from different models in the three BMI and four age categories. Detailed legend is given below.

Figure is divided into two parts : the top part corresponds to the SEP1 criterion and the bottom part corresponds to the REP1 criterion. On the y-axis of SEP1 subfigure, the lower value is on the bottom and the higher value is on the top; on the y-axis of REP1 subfigure, the lower value is on the top and the higher value is on the bottom. On the top of SEP1 subfigure, the labels for the three BMI categories, twice global level and the four age categories are displayed. The global level are shown twice to have a more comprehensive baseline representation. The numbers on the bottom of REP1 subfigure indicate sample size for each category.

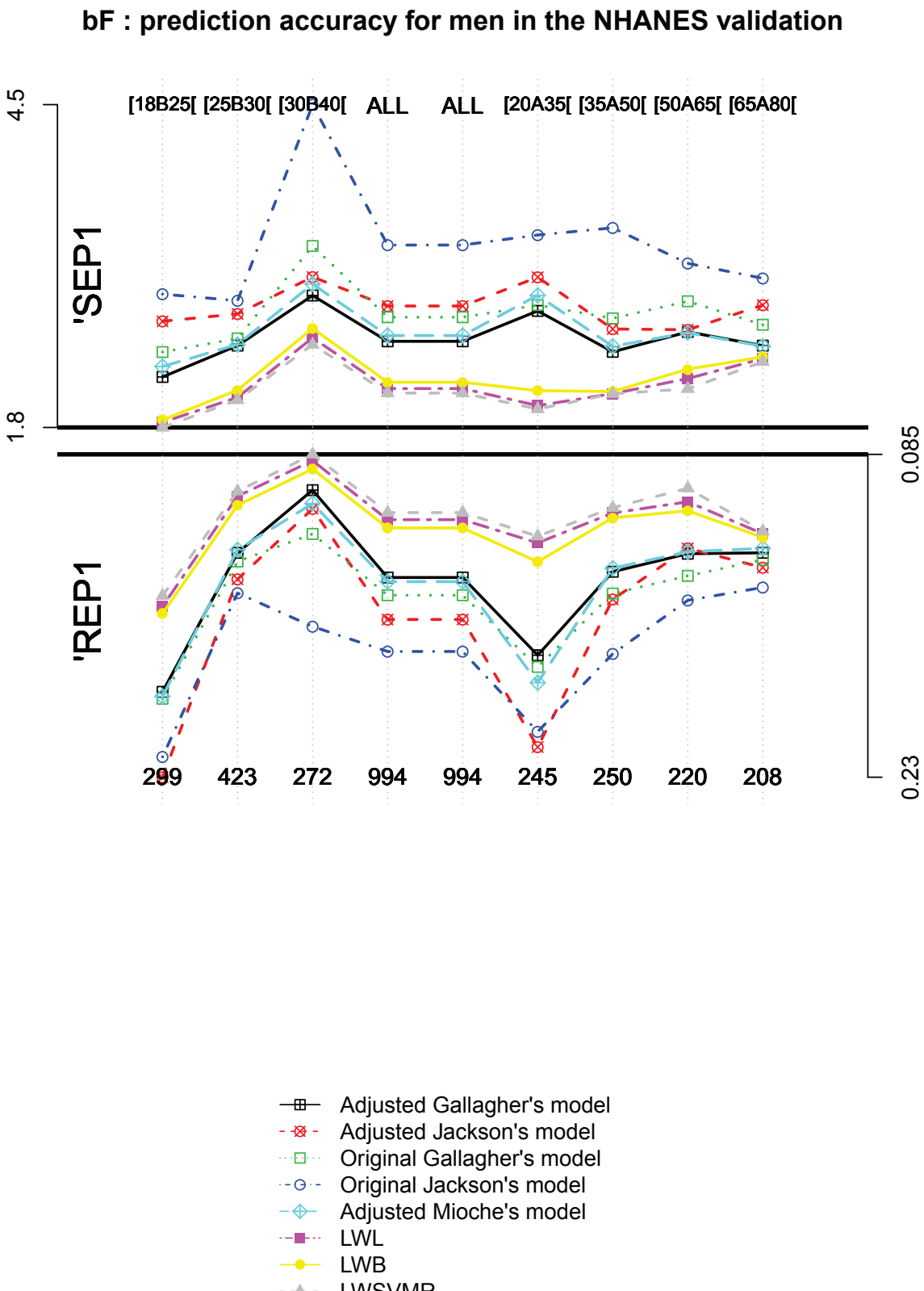


Figure 3.7: For men in the NHANES validation subset, accuracy of the prediction for 9 SBCs from different models in the three BMI and the four age categories. c.f. 3.6 for figure description.

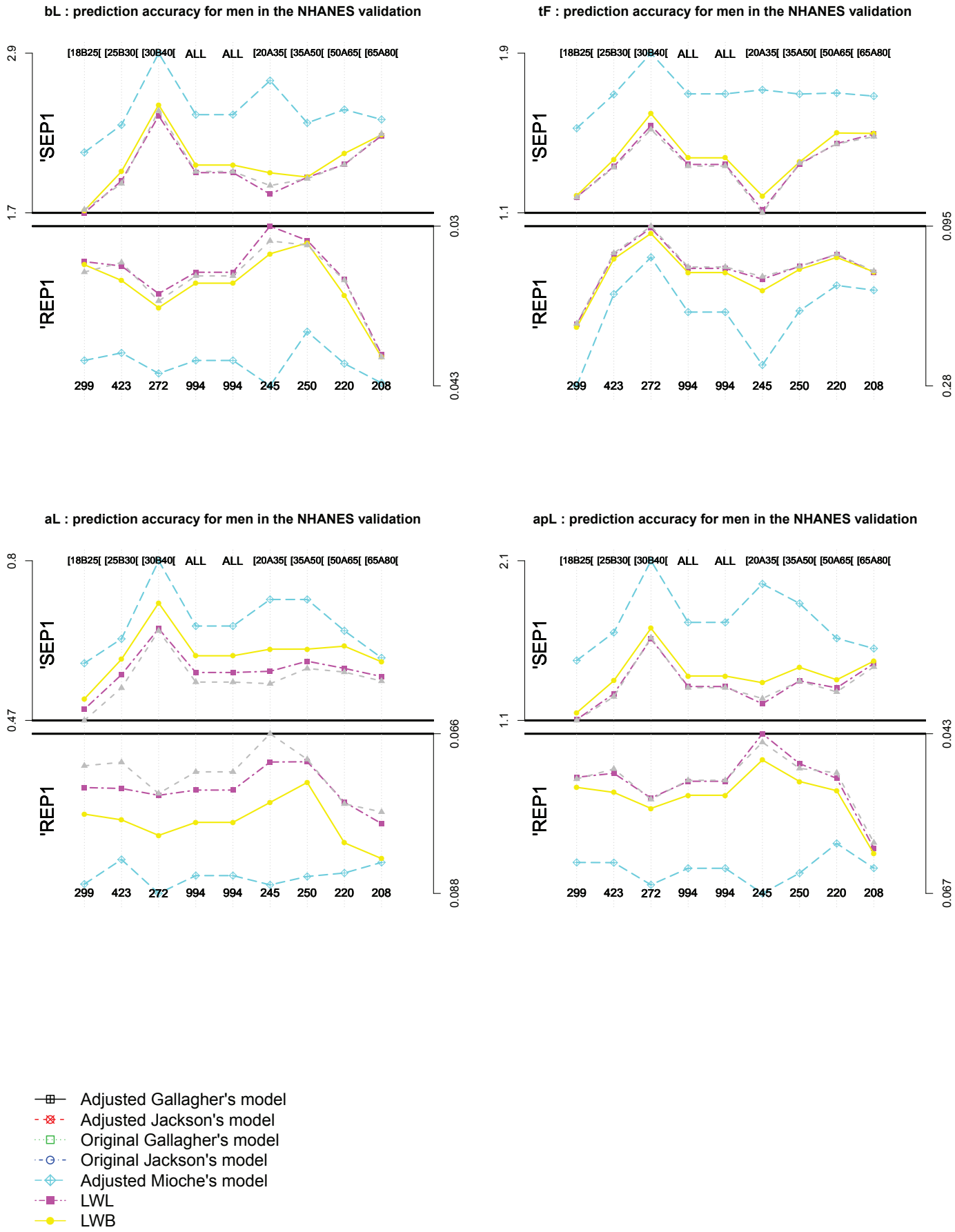


Figure 3.8: For women in the NHANES validation subset, accuracy of the prediction for 9 SBC from different models in the three BMI and the four age categories. c.f. 3.6 for figure description.

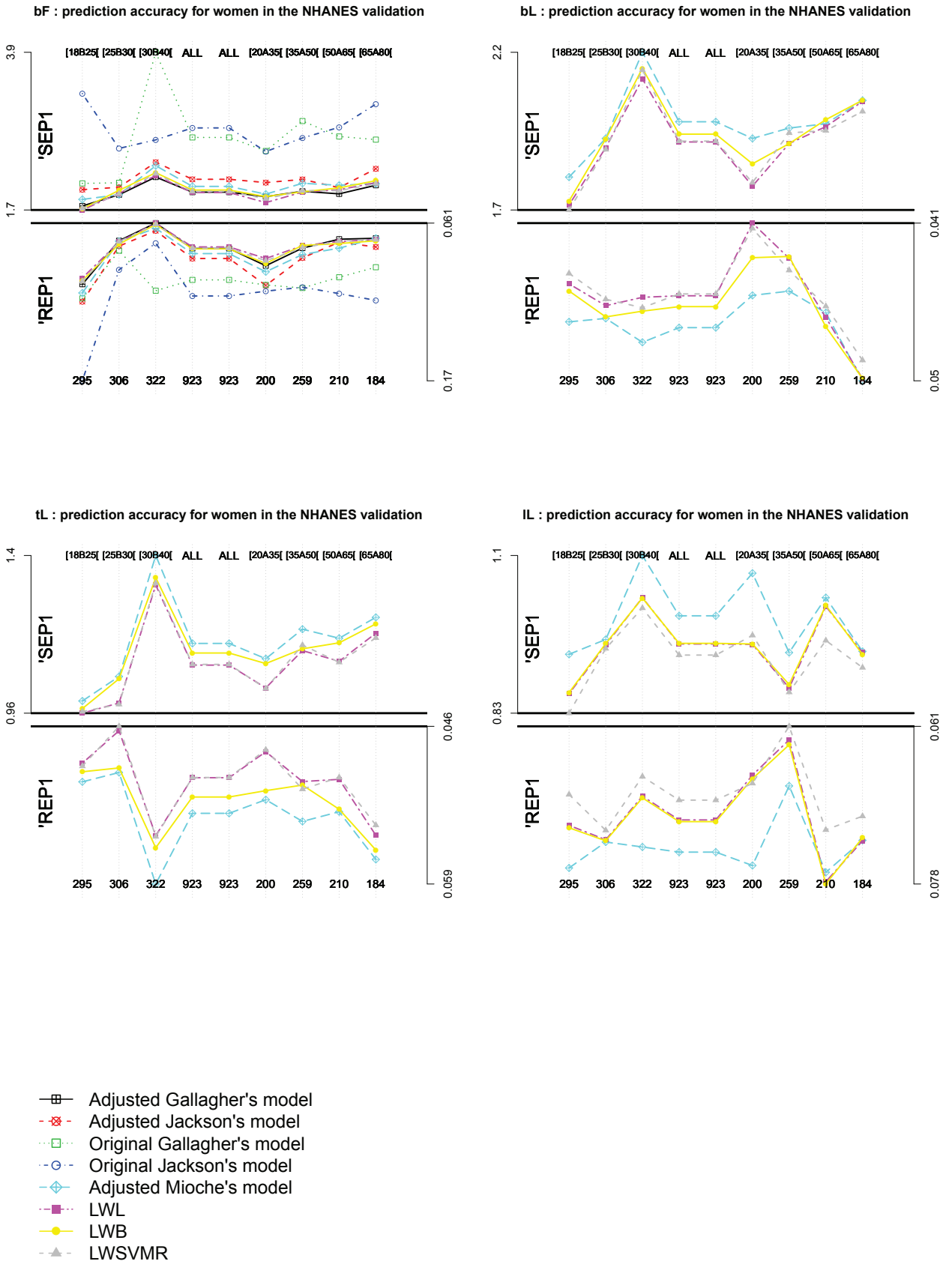


Figure 3.9: For men in the French CHU dataset, accuracy of the prediction for 9 SBC from different models in the three BMI and the four age categories. c.f. 3.6 for figure description.

Figure 3.10: For women in the French CHU dataset, accuracy of the prediction for 9 SBC from different models in the three BMI and the four age categories. c.f. 3.6 for figure description.

3.2.2 A multivariate model for predicting segmental body composition (Paper accepted in *British Journal of Nutrition*)

British Journal of Nutrition, page 1 of 11

doi:10.1017/S0007114513001803

A multivariate model for predicting segmental body composition

Simiao Tian[1,2,3]*, Laurence Mioche[3], Jean-Baptiste Denis[1] and Béatrice Morio[2,3]

[1]*INRA, Unité de Recherche MIA, F-78352 Jouy-en-Josas, France*
[2]*Unité de Nutrition Humaine, Clermont Université, Université d'Auvergne, BP 10448, F-63000 Clermont-Ferrand, France*
[3]*INRA, UMR 1019, UNH, F-63000 Clermont-Ferrand, France*

(Submitted 14 December 2012 – Final revision received 2 May 2013 – Accepted 2 May 2013)

Abstract

The aims of the present study were to propose a multivariate model for predicting simultaneously body, trunk and appendicular fat and lean masses from easily measured variables and to compare its predictive capacity with that of the available univariate models that predict body fat percentage (BF%). The dual-energy X-ray absorptiometry (DXA) dataset (52 % men and 48 % women) with White, Black and Hispanic ethnicities (1999–2004, National Health and Nutrition Examination Survey) was randomly divided into three sub-datasets: a training dataset (TRD), a test dataset (TED); a validation dataset (VAD), comprising 3835, 1917 and 1917 subjects. For each sex, several multivariate prediction models were fitted from the TRD using age, weight, height and possibly waist circumference. The most accurate model was selected from the TED and then applied to the VAD and a French DXA dataset (French DB) (526 men and 529 women) to assess the prediction accuracy in comparison with that of five published univariate models, for which adjusted formulas were re-estimated using the TRD. Waist circumference was found to improve the prediction accuracy, especially in men. For BF%, the standard error of prediction (SEP) values were 3·26 (3·75) % for men and 3·47 (3·95) % for women in the VAD (French DB), as good as those of the adjusted univariate models. Moreover, the SEP values for the prediction of body and appendicular lean masses ranged from 1·39 to 2·75 kg for both the sexes. The prediction accuracy was best for age <65 years, BMI $<30\,kg/m^2$ and the Hispanic ethnicity. The application of our multivariate model to large populations could be useful to address various public health issues.

Key words: Multivariate models: Body composition: Dual-energy X-ray absorptiometry: Predictions

The assessment of human body composition is important for evaluating health and nutritional status. Among health issues, overweight and obesity are worldwide problems. Increased fat mass, especially in the trunk location[1–4], has been associated with an increased risk of metabolic diseases, such as type 2 diabetes and CVD. The amount of lean body mass, especially of appendicular muscle mass, is also directly correlated with health and particularly with the mortality rate[3,4]. Accurate measurements of body composition can be obtained from different methods, such as underwater weighing, dilution techniques and dual-energy X-ray absorptiometry (DXA). However, their applications are not always convenient for large populations, because they require fixed equipment and they are also time consuming and expensive.

The potential uses of statistical methods for body composition assessment have been highlighted[5], and several attempts to predict body composition, particularly body fat percentage (BF%), using linear models with simple predictor variables have been made. A summary of the body composition prediction models published between 1985 and 2003 has been given by Sun & Chumlea[6]. They pointed out that (1) a general model for two sexes, different ethnicities and wide age ranges may lose its accuracy due to increased heterogeneity; (2) cross-validation of prediction models was needed to assess their generalisability; (3) for validation studies, accuracy should be standardised for the mean of the predicted variable; (4) few prediction models were derived from datasets using DXA.

The advantages of using sex, age, ethnicity and easily accessible anthropometric measurements, such as body weight and height, are simplicity and cost efficiency. Their use would allow access to large datasets to describe body composition characteristics. Previous published linear models have made univariate predictions[7–11]. Alternatively, a non-parametric model based on Bayesian networks that uses the same predictor variables has been proposed[12,13]. This Bayesian networks

Abbreviations: APF, appendicular fat; APL, appendicular lean; BF, body fat; BF%, body fat percentage; BFF, body fat-free mass; BL, body lean; DXA, dual-energy X-ray absorptiometry; French DB, French dual-energy X-ray absorptiometry dataset; MWC, models with waist circumference; MWoC, models without waist circumference; NHANES, National Health and Nutrition Examination Survey; RSD, relative standard deviation; SEP, standard error of prediction; TED, test dataset; TF, trunk fat; TRD, training dataset; VAD, validation dataset.

* **Corresponding author:** S. Tian, email simiao.tian@jouy.inra.fr

British Journal of Nutrition

approach consists in selecting a subset of individuals so that their predictor variable characteristics are similar to those of the individuals to be predicted. This model allows simultaneous prediction of segmental compartments, but requires the availability of a reference dataset. To our knowledge, until now, no multivariate linear prediction model has been proposed for body composition assessment. The aim of the present study was, therefore, to develop sex-specific multivariate models for estimating some segmental compartments of metabolic importance (i.e. lean body mass, appendicular muscle mass and trunk fat (TF)) from age and easily accessible anthropometric variables. The usefulness of waist circumference was also investigated and combined with age, height and weight as predictor variables. These multivariate models, based on the reference dataset National Health and Nutrition Examination Survey (NHANES), were validated with two different populations in agreement with the principles proposed by Sun & Chumlea[6].

Subjects and methods

Databases

All body composition values related to predictions were extracted from the NHANES website (http://www. cdc.gov/nchs/about/major/nhanes/) from the 1999–2004 period. Subjects were characterised by predictor variables, such as sex, ethnicity, age, height, weight and waist circumference. For the present study, we selected subjects aged 20–85 years, with BMI values ranging from 18 to 40 kg/m^2 and who belonged to one of the three considered ethnicity categories: White, Black and Hispanic. This selection resulted in a sample size of 3977 men (1984 White, 720 Black and 1273 Hispanic) and 3692 women (1830 White, 697 Black and 1165 Hispanic).

The study was conducted separately on men and women; therefore, the complete NHANES dataset was split by sex. For each sex, we randomly split the corresponding NHANES dataset into three sub-datasets: a training dataset (TRD); a test dataset (TED); a validation dataset (VAD).

As the number of individuals was high, the splitting was done at random as suggested by Hastie *et al.*[14] and Nivre[15] in data-rich situations. The TRD was used as a reference dataset to fit the parameters of a series of possible models. The test dataset was used to estimate the prediction error of each fitted model to make model selection, and the VAD was used to perform a one-round validation calculation and to assess the prediction accuracy of the final chosen models.

An independent external dataset (French DB, French DXA dataset) was used to assess the performance of the prediction models in a different population context. The French DB was obtained from a routine examination at the Radiology Department of the Clermont-Ferrand University Hospital Centre between 1998 and 2008. It contains data on 1095 French subjects, 526 men and 569 women, aged between 20 and 85 years and with BMI values ranging between 18 and 40 kg/m^2. However, ethnicity was not mentioned and waist circumference was not measured during the examination.

The study carried out using the NHANES dataset complies with the Declaration of Helsinki, the National Center for Health Statistics Ethics Review Board approved the protocols, and written informed consent was obtained from each participant. Moreover, the study using the French dataset was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving human subjects were approved by the Clermont-Ferrand University Hospital Centre, France, and by the local ethics committee. Written informed consent was obtained from all subjects at recruitment after being informed of the nature, purpose and possible risks of the protocols.

Measurement of body composition

Whole-body and segmental body compositions were assessed using DXA (Hologic QDR 4500A fanbeam densitometer for the NHANES dataset and Hologic QDR-4500 densitometer for the French DB; http://www.gmecorp-usa.com/IM/XR/BD/HOLLOGIC/4500/SV/Qdr4500dos.pdf). For the NHANES dataset, detailed descriptions have been published earlier[16]. Briefly, whole-body DXA scans were taken at the NHANES mobile examination centre for eligible participants during the 6-year period from 1999 to 2004; the participants with certain physical conditions were excluded from the DXA examination[17]. The DXA scans allow the quantification of multiple whole-body and regional components, including bone mineral content, fat and lean soft tissue. Body fat (BF) and body lean (BL) masses and TF and trunk lean masses were thus determined[18]. Appendicular composition was the sum of arm and leg fat (APF, appendicular fat) and lean (APL, appendicular lean) masses[19]. Body fat-free mass (BFF) was calculated as the sum of the BL mass and bone mineral content.

Statistical methods

Non-parametric approaches. First, several non-parametric approaches were evaluated to make absolute body composition predictions. The term 'non-parametric' implies that the number and nature of the parameters are flexible and not fixed in advance[20]. These non-parametric approaches followed the statistical methodology described by Mioche *et al.*[12]. The local prediction models included weighted linear regression, support vector machine regression[21,22] and Bayesian regression[23]. For a given individual to be predicted, these methods follow three steps: (1) dissimilarities are calculated between the individual to be predicted and each individual of the TRD based on the values of the predictor variables; (2) the dissimilarities are transformed into weights to give more importance to similar individuals; (3) a prediction model is developed from this weighted dataset. When weights are constrained to be 0 or 1, the method corresponds to the selection of a sub-dataset as performed by Mioche *et al.*[12].

Multivariate linear regression. In the present study, a multivariate multiple linear regression, supposed to satisfy linear model assumptions, was also used as a possible alternative to these sophisticated prediction models. Multiple univariate linear regression is easily extended to deal with situations

British Journal of Nutrition

Table 1. Formulas of the five published prediction models for body fat percentage (BF%) for men and women*

References		Models	
		Men	Women
Gallagher *et al.*[7]†	Original	55·49 − 43·8/BMI + 0·087 age	76 − 1097·8/BMI + 0·053 age
	Adjusted	45·65 − 708·3/BMI + 0·104 age	58·72 − 675·2/BMI + 0·069 age
Jackson *et al.*[8]	Original	3·76 BMI − 0·04 BMI^2 − 47·8	4·35 BMI − 0·05 BMI^2 − 46·24
	Adjusted	2·29 BMI − 0·023 BMI^2 − 20·8	3·27 BMI − 0·042 BMI^2 − 20·55
Larsson *et al.*[9]‡	Original	–	–
	Adjusted	46·8 (1 − exp(−0·047 (BMI − 11·18)))	47·2 (1 − exp(−0·099 (BMI − 10·95)))
Levitt *et al.*[10]	Original	48·1 − 952·38/BMI + 0·176 age	63·2 − 948/BMI + 0·135 age
	Adjusted	45·65 − 708·3/BMI + 0·104 age	58·72 − 675·2/BMI + 0·069 age
Gómez-Ambrosi *et al.*[11]	Original	− 44·988 + 0·503 age + 3·172 BMI − 0·026 BMI^2 − 0·02 BMI age + 0·00 021 BMI^2 age	− 34·299 + 0·503 age + 3·1353 BMI − 0·031 BMI^2 − 0·02 BMI age + 0·00 021 BMI^2 age
	Adjusted	− 26·627 + 0·259 age + 2·211 BMI − 0·019 BMI^2 − 0·008 BMI age + 0·0001 BMI^2 age	− 16·206 − 0·011 age + 2·53 BMI − 0·025 BMI^2 − 0·009 BMI age − 0·0002 BMI^2 age

* Adjusted formulas were estimated from the training dataset.
† For Gallagher's model, only the non-Asian model is reported.
‡ For Larsson's model, the parameter values are not provided; only the statistical formula is provided, which is as follows: $y = a \times (1 - e^{-b(\mathrm{BMI}-\mathrm{BMI}_0)})$.

where the response consists of $P > 1$ different variables; this is termed 'multivariate linear regression'[24]. Estimates of the regression parameters are determined by the least squares method. The fitting model in a multivariate model for each variable will be the same as that which would result from a univariate model. However, the constraint in the multivariate model consists in using identical predictor variables for all the predicted variables. The advantage of using the multivariate approach is that it takes the correlation structure between the responses into account, which is useful for a number of inference tasks, e.g. to give simultaneous confidence regions for all the responses together.

Validation analysis. The selection of models from previously described multivariate approaches was based on the prediction accuracy and complexity of the models. The accuracy was measured by the standard error of prediction (SEP) and the relative standard deviation (RSD, two criteria defined below), and the complexity of the models was assessed by the number of parameters and computing time.

Waist circumference usefulness analysis. The usefulness of waist circumference for prediction was investigated. To do so, prediction accuracy was checked for some categories of BMI (18–25, 25–30 and 30–40 kg/m^2), age (20–35, 35–50, 50–65 and 65–80 years) and ethnicity (White, Black and Hispanic). This categorical analysis was performed only on the VAD. The prediction accuracy in this categorical study was expressed by a 100-scale score. A score of 100 denotes a baseline, i.e. the average level of prediction quality for all the categories; a score less than 100 denotes a better quality than the average level; and in contrast a score greater than 100 indicates a worse quality.

Comparison with published univariate models. In the literature, univariate linear regressions have been developed to primarily predict BF% from BMI, age and, occasionally, waist circumference or ethnicity as predictor variables[7–11]. Of the univariate models published between 2000 and 2012, five were retained with different combinations of predictor variables (Table 1). Gallagher's[7] and Larsson's[9] models were derived from a DXA dataset, Jackson's[8] model from a densitometry dataset from four clinical centres, Levitt's[10] model from a densitometry and water dilution dataset, and Gómez-Ambrosi's[11] model from an air-displacement plethysmography dataset. Original and adjusted formulas were applied to the VAD and French DB. The adjusted formulas were derived by re-estimating the parameters of the published models using the TRD. Their prediction accuracies were considered as baseline values to evaluate those of our proposed combination of predictor variables in the multivariate models. The prediction of BF% from our multivariate model was calculated by dividing the predicted value of BF by body weight, multiplied by 100.

Assessment of the prediction accuracy. The accuracy of a prediction for a given variable was globally assessed using the SEP:

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(\mathrm{measured}_i - \mathrm{predicted}_i)^2}{n}},$$

where n is the number of subjects in the VAD or French DB. The unit of SEP is the same as the unit of the predicted variable (kg or %). The SEP is then detailed into bias and standard deviation: $\mathrm{SEP}^2 = \mathrm{bias}^2 + \mathrm{SD}^2$ to investigate the trade-off between model bias and variance in prediction. The RSD provides another assessment of the prediction accuracy. It was calculated by dividing 100 × SEP by the mean of the predicted variable, and it is expressed in percentage of the global mean. Finally, the coefficient of determination R^2 was used to assess the goodness of fit in the validation procedure.

Statistical test analyses. Population characteristics are expressed as means and standard deviations. Differences between each of the three subsets of the NHANES dataset and French DB were analysed using Student's t tests. These t tests aimed to assess the differences between the American and French samples. Only the SEP difference was analysed by a permutation test[25]. Furthermore, paired t tests and Bland–Altman plots[26] were used to determine the difference and the limits of agreement between the published univariate models and the multivariate model. A CI for the mean of the difference was also calculated under a normality assumption. Statistical calculations and analyses were performed using

British Journal of Nutrition

version 2.12.2 of the R software (http://cran.r-project.org/doc/contrib/Lam-IntroductionToR_LHL.pdf)[27], a language and an environment for statistical computing.

Results

Sample characteristics

The means and standard deviations of age, anthropometric variables and DXA body composition for the different datasets are presented in Table 2 for men and women. Within the three subsets of the NHANES dataset, the men and women were of the same age, but some difference was observed for the French subjects. For men, except for height, all the variables were significantly different between the French DB and the three NHANES dataset subsets. For women, most of the variables were significantly different between the French DB and the three NHANES dataset subsets, except for trunk lean, BL and BFF.

Prediction models

The study of the selection of models from the test dataset showed that the non-parametric approaches did not provide a significantly better SEP than the multivariate linear regression. Moreover, the non-parametric approaches need more parameters and computing times. Therefore, multivariate linear regression is the only model mentioned in the paper to predict segmental compartments. The parameters of this multivariate model are given in Table 3 for models with and without waist circumference (MWC and MWoC, respectively) as a predictor variable.

Inclusion of waist circumference

Tables 4 and 5 summarise the prediction accuracy for three categories of BMI and ethnicity and four age ranges when MWC and MWoC are applied to the VAD. Predictions were more accurate when waist circumference was included, especially for men with a BMI value that ranged from 18 to 30 kg/m^2 and whose age ranged from 25 to 65 years.

Table 2. Age, anthropometric variables and dual-energy X-ray absorptiometry body composition characteristics for men and women in the National Health and Nutrition Examination Survey (NHANES) training dataset (TRD), test dataset (TED) and validation dataset (VAD) and in the French dataset (French DB)

(Mean values and standard deviations)

	NHANES TRD		NHANES TED		NHANES VAD		French DB	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Men (*n*)	1989		994		994		526	
Ethnicity (*n*)								
White	983		492		509		–§	
Black	367		171		182		–§	
Hispanic	639		331		303		–§	
Age (years)	50·60	18·91	51·16	19·33	50·6	18·63	46·61*†‡	17·02
Height (cm)	174·02	7·81	174·12	7·68	174·06	8·00	174·62	6·79
Weight (kg)	83·90	15·2	84·07	16·21	83·93	15·98	78·83*†‡	13·08
Waist circumference (cm)	98·62	12·62	98·78	12·96	98·58	12·48	–	–
Trunk fat (kg)	11·21	4·98	11·26	5·10	11·17	4·87	8·76*†‡	4·48
Appendicular fat (kg)	8·95	3·34	8·94	3·48	8·81	3·46	7·52*†‡	2·89
Body fat (kg)	21·21	8·03	21·24	8·30	21·02	8·07	17·34*†‡	7·04
Trunk lean (kg)	29·74	4·46	29·86	4·78	29·79	4·61	29·27*†‡	4·40
Appendicular lean (kg)	26·75	4·74	26·77	4·97	26·89	5·04	25·87*†‡	3·95
Body lean (kg)	60·06	9·03	60·20	9·65	60·26	9·56	58·82*†‡	7·64
Body fat-free mass (kg)	62·70	9·35	62·83	9·97	62·90	9·89	61·49*†‡	7·90
Women (*n*)	1846		923		923		569	
Ethnicity (*n*)								
White	911		475		444		–§	
Black	368		163		166		–§	
Hispanic	567		285		313		–§	
Age (years)	51·64	18·92	52·09	18·79	51·11	18·08	49·28*†‡	14·84
Height (cm)	160·71	6·86	160·77	6·77	160·57	6·86	161·93*†‡	6·64
Weight (kg)	71·56	14·24	72·86	14·50	72·62	14·31	67·66*†‡	13·46
Waist circumference (cm)	92·42	12·69	93·50	12·99	93·46	12·88	–	–
Trunk fat (kg)	12·83	4·92	13·27	5·21	13·32	5·11	10·36*†‡	4·88
Appendicular fat (kg)	13·36	4·34	13·78	4·46	13·48	4·34	12·03*†‡	3·89
Body fat (kg)	27·08	8·74	27·93	9·15	27·69	8·95	23·3*†‡	8·23
Trunk lean (kg)	21·56	3·20	21·86	3·15	21·84	3·18	21·68	3·97
Appendicular lean (kg)	17·90	3·48	18·02	3·49	18·04	3·44	17·51*†‡	3·35
Body lean (kg)	42·47	6·59	42·89	6·56	42·89	6·52	42·29	6·65
Body fat-free mass (kg)	44·49	6·84	44·93	6·81	44·93	6·75	44·36	6·87

* Mean values were significantly different from those of the TRD ($P<0·05$; *t* test).
† Mean values were significantly different from those of the TED ($P<0·05$; *t* test).
‡ Mean values were significantly different from those of the VAD ($P<0·05$; *t* test).
§ Ethnicity was not mentioned in the French DB.

Table 3. Multivariate prediction model estimates of parameters for the seven segmental compartments (kg) including or not including waist circumference as a predictor variable*

	With waist circumference					Without waist circumference			
	Intercept (kg)	β_A (kg/year)	β_H (kg/cm)	β_W	β_C (kg/cm)	Intercept (kg)	β_A (kg/year)	β_H (kg/cm)	β_W
Men									
TF	−1354·89	0·67	−6·28	8·19	28·87	1597·00	6·55	−19·8	31·44
APF	−538·53	−0·52	−3·06	11·90	10·07	491·26	1·53	−7·77	20·01
BF	−1796·17	0·17	−9·53	20·84	38·72	2162·87	8·06	−27·67	52·02
TL	−191·23	0·84	7·48	29·36	−6·51	−856·43	−0·48	10·52	24·12
APL	1635·96	−1·30	1·12	44·26	−28·43	−1270·45	−7·09	14·43	21·37
BL	1776·70	−0·41	8·15	76·07	−36·00	−1903·77	−7·74	25·01	47·08
BFF	1795·49	−0·17	9·54	79·16	−38·72	−2163·66	−8·05	27·67	47·98
Women									
TF	96·99	1·30	−11·11	21·19	15·01	1460·01	3·67	−17·22	33·55
APF	1343·80	2·80	−10·62	38·42	−12·94	168·76	0·76	−5·35	27·76
BF	1503·60	4·06	−21·78	60·01	2·17	1701·07	4·40	−22·67	61·80
TL	−1104·06	−0·96	11·87	14·23	4·15	−726·93	−0·30	10·18	17·65
APL	−513·85	−2·43	8·07	22·88	−5·47	−1010·63	−3·29	10·30	18·37
BL	−1420·31	−3·54	20·03	38·26	−1·14	−1524·23	−3·73	20·49	37·32
BFF	−1504·58	−4·06	21·78	40·00	−2·17	−1701·97	−4·40	22·67	38·21

TF, trunk fat; APF, appendicular fat; BF, body fat; TL, trunk lean; APL, appendicular lean; BL, body lean; BFF, body fat-free mass, calculated as the sum of the BL mass and bone mineral content.

*The parameters are, respectively, associated with the intercept, age (β_A), height (β_H), weight (β_W) and waist circumference (β_C). For the sake of presentation, all values have been multiplied by 100.

Regarding ethnicity categories, a remarkable improvement in accuracy was found for Black men when waist circumference was used as a predictor variable. Compared with that of MWoC, the prediction accuracy of MWC was improved by a 45% unit (in a 100-scale score) for TF and APL masses and by a 30% unit (in a 100-scale score) for BF, BL and BFF masses. By contrast, for women in all the BMI, age and ethnicity categories, the quality of the predictions was similar between MWC and MWoC.

For subjects of both sexes, the prediction by MWC was less reliable with BMI values in the range 30–40 kg/m^2 than for the other BMI categories. Indeed, the prediction accuracy of MWC was reduced by 35 and 25% units for a BMI >30 kg/m^2 than for the BMI categories of 18–25 and 25–30 kg/m^2.

Regarding the three ethnicity categories, MWC provided the best quality of fit for Hispanic individuals, followed by White and Black individuals. More precisely, for the BF, BL and appendicular compartments in Hispanic individuals, the prediction accuracy of MWC was improved by 20% unit in Hispanic men than in White and Black men. Similarly, it was improved by 10 and 30% units in Hispanic women than in White and Black women, respectively.

Table 4. Accuracy of the proposed prediction models with waist circumference (MWC) and without waist circumference (MWoC) as a predictor variable for the seven segmental compartments in different BMI, age and ethnicity categories for men in the National Health and Nutrition Examination Survey validation dataset*

		BMI categories (kg/m^2)			Age categories (years)				Ethnicity categories		
Compartments		18–25	25–30	30–40	20–35	35–50	50–65	65–80	White	Black	Hispanic
TF	MWC	90	95	116	86	97	105	113	104	102	90
	MWoC	105	120	136	126	114	125	119	120	145	103
APF	MWC	77	94	127	108	98	101	99	105	109	84
	MWoC	81	98	129	115	100	106	98	108	117	86
BF	MWC	87	94	120	99	96	102	110	104	105	88
	MWoC	100	113	133	129	108	120	112	117	137	98
TL	MWC	88	93	119	97	99	107	105	100	108	95
	MWoC	89	95	119	100	98	108	105	102	107	95
APL	MWC	84	98	118	98	99	95	111	101	125	80
	MWoC	105	121	142	138	124	121	114	116	172	95
BL	MWC	86	94	121	98	96	102	111	105	104	88
	MWoC	98	112	134	127	108	119	112	116	134	98
BFF	MWC	87	94	120	99	96	102	110	104	105	88
	MWoC	100	113	134	129	109	120	112	117	137	98

TF, trunk fat; APF, appendicular fat; BF, body fat; TL, trunk lean; APL, appendicular lean; BL, body lean; BFF, body fat-free mass, calculated as the sum of the BL mass and bone mineral content.

*The accuracy is assessed by a 100-scale score: the smaller the score, the better the prediction. A value of 100 corresponds to the global standard error of prediction for all the categories with waist circumference as a predictor variable.

British Journal of Nutrition

Table 5. Accuracy of the proposed prediction models with waist circumference (MWC) and without waist circumference (MWoC) as a predictor variable for the seven segmental compartments in different BMI, age and ethnicity categories for women in the National Health and Nutrition Examination Survey validation dataset*

		BMI categories (kg/m^2)			Age categories (years)				Ethnicity categories		
Compartments		18–25	25–30	30–40	20–35	35–50	50–65	65–80	White	Black	Hispanic
TF	MWC	87	97	113	97	95	102	105	101	115	89
	MWoC	95	108	131	111	110	111	120	110	138	100
APF	MWC	83	91	121	92	100	103	108	96	111	100
	MWoC	81	98	131	94	107	110	118	102	116	107
BF	MWC	89	96	112	94	100	100	108	97	122	90
	MWoC	90	96	112	95	100	100	108	97	123	90
TL	MWC	88	88	119	94	104	101	108	99	113	95
	MWoC	88	90	124	96	106	104	110	100	116	98
APL	MWC	89	97	111	101	95	101	101	93	137	85
	MWoC	92	98	115	105	99	103	103	94	143	88
BL	MWC	89	96	113	93	100	101	109	98	121	91
	MWoC	89	96	113	93	100	101	109	98	121	91
BFF	MWC	89	96	112	94	100	100	108	97	122	90
	MWoC	90	96	112	94	101	100	108	97	123	90

TF, trunk fat; APF, appendicular fat; BF, body fat; TL, trunk lean; APL, appendicular lean; BL, body lean; BFF, body fat-free mass, calculated as the sum of the BL mass and bone mineral content.

*The accuracy is assessed by a 100-scale score: the smaller the score, the better the prediction. A value of 100 corresponds to the global standard error of prediction for all the categories with waist circumference as a predictor variable.

Multivariate prediction models

The validation scores for the multivariate model were calculated using the VAD, and they are given in Table 6. For the prediction of BF and BL masses, a SEP value less than 2·8 kg was found for both men and women (men: 2·75 and 2·66 kg; women: 2·52 and 2·47 kg, respectively). By contrast, because of the differences in the compartment masses, the RSD values were much lower for the BL prediction than for the BF prediction (men: 4·41 and 13·08 %; women: 5·76 and 9·01 %, respectively). The corresponding R^2 values averaged 0·9 for both the sexes (men: 0·88 and 0·92; women: 0·92 and 0·86, respectively).

Regarding other segmental compartments such as trunk and APF and APL masses and BFF, the SEP values ranged from 1·65 to 2·75 kg for men and from 1·39 to 2·52 kg for women. Similarly, in both the sexes, because of the differences in the compartment sizes, the RSD values were lower for trunk and APL masses than for the corresponding fat masses. They varied from 5·54 to 8·76 % for trunk and APL masses and from 12·54 to 19·18 % for trunk and APF masses. The corresponding R^2 values ranged from 0·8 to 0·9 for both the sexes.

The bias ranged approximately from 0·50 to 0·90 kg for both men and women, which were low in comparison with the model variance (Table 6). Comparisons of the predictions by models and the observations are shown in Fig. 1 for men and women. For men, segmental body compositions were globally well predicted, even if for extreme parts, some bias appeared: an underestimation for a high fat mass and an overestimation for low lean mass. For women, an underestimation for high APF and APL masses was observed.

When the multivariate prediction model without waist circumference was applied to the French DB, the predictions were still good (table not shown). For men, the SEP values were 2·95 kg (R^2 0·84) for BF mass and 2·84 kg (R^2 0·87) for BL mass, with the RSD values being equal to 17·01 and 4·83 %, respectively. For women, the corresponding SEP values were 2·86 kg (R^2 0·89) and 2·80 kg (R^2 0·84) with the respective RSD values of 12·27 and 6·62 %.

Table 6. Accuracy of the multivariate prediction model calculated using the National Health and Nutrition Examination Survey validation dataset using waist circumference for the seven segmental compartments*

	Men					Women				
Compartments	SEP (kg)	Bias (kg)	SD (kg)	RSD (%)	R^2	SEP (kg)	Bias (kg)	SD (kg)	RSD (%)	R^2
TF	1·73	0·50	1·65	15·49	0·87	1·67	0·71	1·51	12·54	0·90
APF	1·69	0·90	1·43	19·18	0·76	1·99	0·85	1·80	14·76	0·79
BF	2·75	0·92	2·59	13·08	0·88	2·52	0·83	2·38	9·10	0·92
TL	1·65	0·52	1·56	5·54	0·87	1·39	0·61	1·25	6·36	0·81
APL	1·76	0·62	1·65	6·55	0·88	1·58	0·70	1·41	8·76	0·79
BL	2·66	0·70	2·56	4·41	0·92	2·47	0·90	2·30	5·76	0·86
BFF	2·75	0·73	2·65	4·37	0·92	2·52	0·90	2·36	5·61	0·86

SEP, standard error of prediction; RSD, relative standard deviation; TF, trunk fat; APF, appendicular fat; BF, body fat; TL, trunk lean; APL, appendicular lean; BL, body lean; BFF, body fat-free mass, calculated as the sum of the BL mass and bone mineral content.

*The absolute value of the total weight of the segmental compartments is predicted. The accuracy is assessed by the SEP in kg and the RSD in %. RSD = 100 × (SEP/$\overline{y}$) for a predicted variable Y and its mean $\overline{y}$, and it is expressed as a percentage. For example, for BF of men, RSD = 100 × 2·75/21·02 = 13·08 %.

British Journal of Nutrition

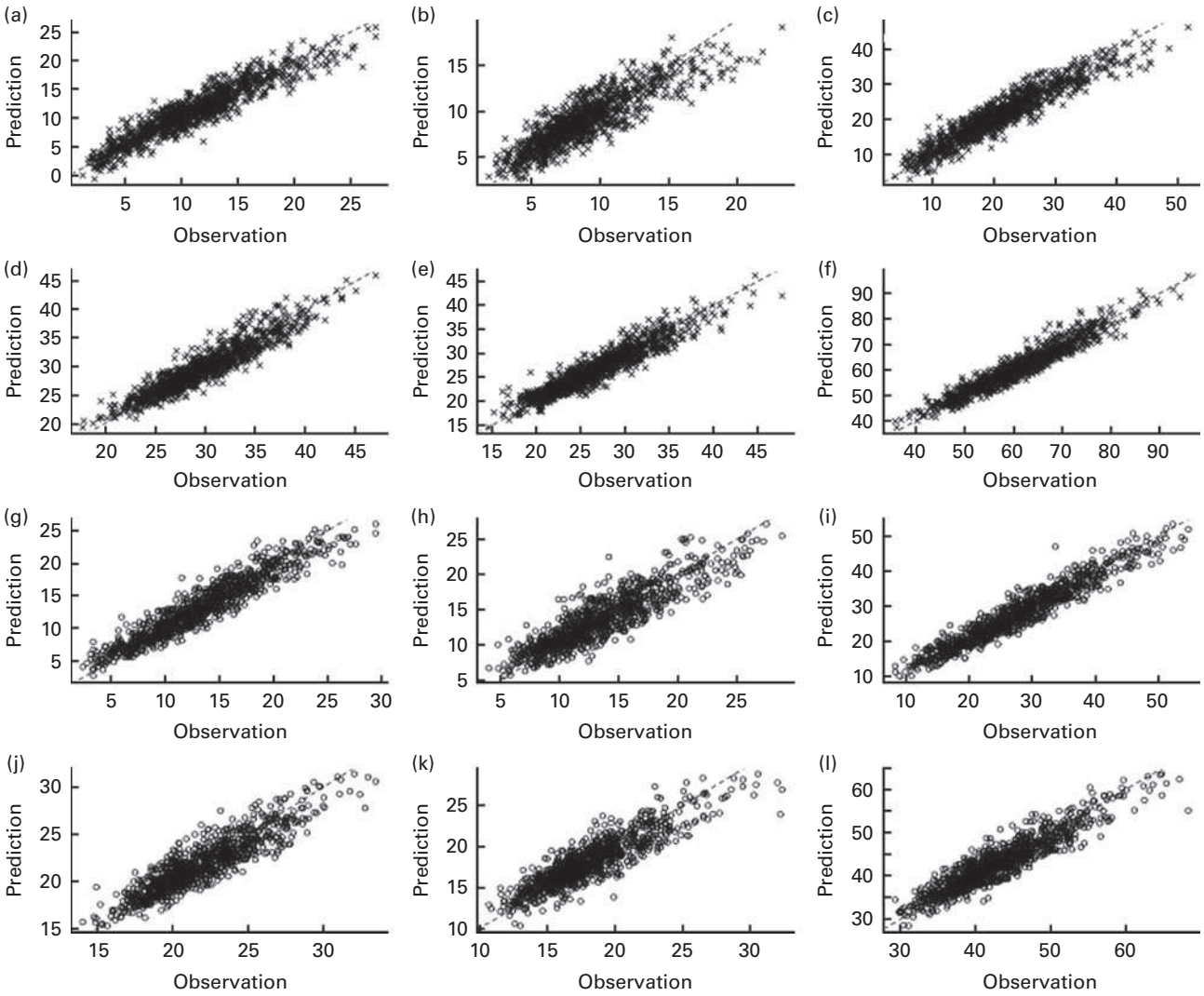


Fig. 1. Scatter plot of the multivariate model for the prediction of different segmental body compositions against their observations in the validation dataset. Men are represented by × and women by ○. The first bisectors are drawn (---). Men: (a) trunk fat (TF); (b) appendicular fat (APF); (c) body fat (BF); (d) trunk lean (TL); (e) appendicular lean (APL); (f) body lean (BL). Women: (g) TF; (h) APF; (i) BF; (j) TL; (k) APL; (l) BL.

Comparison with published prediction models

When the published formulas with their predictor variables were re-adjusted in the TRD and then applied to the VAD and French DB, the quality of fit was improved in comparison with that of their original formulas. For the BF% of men, the prediction accuracy of the adjusted formula was increased in the VAD by 0·5% unit for Gallagher's and Jackson's prediction models and by 1% unit for Levitt's and Gómez-Ambrosi's prediction models. For the BF% of women, the prediction accuracy was improved, on average, by 1% unit for all the models (Table 7). For the same compartment in French men and women, only a slight improvement in accuracy was found for the univariate models, except for Gómez-Ambrosi's prediction model, for which the prediction accuracy was improved by 1·5% unit.

The accuracy of our multivariate prediction model, based on age, height, weight and waist circumference, was compared with that of the five adjusted published prediction models. In the VAD, the multivariate prediction of BF% yielded one of the best accuracies, with SEP values of 3·26 and 3·74%, respectively, for men and women. For men, our SEP values were 0·5% unit better than those of Gallagher's, Levitt's and Gómez-Ambrosi's prediction models and 1% unit better than those of Jackson's and Larsson's prediction models. By contrast, for women, the differences between SEP values of the various models were small (Table 7). The Bland–Altman plots are shown in Fig. 2 for men and women in the VAD. It appeared that the agreement between our model and the adjusted published models was better for women than for men. In addition, for all the paired t tests, P values ranged from 0·51 to 0·95 and 0·18 to 0·89 for men and women, respectively. Therefore, the difference in predictions was not statistically significant. With respect to the CI of the mean of the difference, it ranged from −0·17 to 0·19 for men and −0·03 to 0·13 for women. These results show that there is no systematic difference between our multivariate

Table 7. Accuracy of the five published models, original and adjusted, and our proposed model for body fat percentage prediction calculated using the National Health and Nutrition Examination Survey (NHANES) validation dataset (VAD) and the French dataset (French DB)*

			Original			Adjusted		
Datasets	Sex	Models	SEP (%)	RSD (%)	R^2	SEP (%)	RSD (%)	R^2
NHANES VAD	Men	Gallagher *et al.*[7]	4·05	16·61	0·59	3·77	15·45	0·61
		Jackson *et al.*[8]	4·89	20·05	0·51	4·23	17·33	0·51
		Larsson *et al.*[9]	−‡	−‡	−‡	4·23	17·33	0·51
		Levitt *et al.*[10]	5·11	20·94	0·60	3·77	15·45	0·61
		Gómez-Ambrosi *et al.*[11]	5·52	22·62	0·61	3·75	15·37	0·61
		Multivariate prediction	–	–	–	3·26	13·26	0·71
	Women	Gallagher *et al.*[7]	4·46	11·95	0·67	3·47	9·29	0·68
		Jackson *et al.*[8]	5·09	13·62	0·62	3·77	10·10	0·63
		Larsson *et al.*[9]	−‡	−‡	−‡	3·75	10·04	0·63
		Levitt *et al.*[10]	4·48	11·99	0·68	3·47	9·29	0·68
		Gómez-Ambrosi *et al.*[11]	4·90	13·13	0·67	3·75	10·40	0·63
		Multivariate prediction	–	–	–	3·47	9·29	0·68
French DB	Men	Gallagher *et al.*[7]	3·95	18·45	0·62	3·76	17·56	0·61
		Jackson *et al.*[8]	4·23	19·74	0·58	3·97†	18·54	0·58
		Larsson *et al.*[9]	−‡	−‡	−‡	3·97†	18·54	0·58
		Levitt *et al.*[10]	5·23	24·42	0·60	3·76	17·56	0·61
		Gómez-Ambrosi *et al.*[11]	5·39	25·17	0·64	3·63	16·95	0·63
		Multivariate prediction	–	–	–	3·74	17·47	0·62
	Women	Gallagher *et al.*[7]	4·43	13·18	0·66	3·93	11·68	0·67
		Jackson *et al.*[8]	5·02	14·91	0·63	4·12	12·25	0·64
		Larsson *et al.*[9]	−‡	−‡	−‡	4·10	12·19	0·64
		Levitt *et al.*[10]	4·28	12·73	0·67	3·93	11·68	0·67
		Gómez-Ambrosi *et al.*[11]	5·37	15·97	0·66	3·93	11·68	0·67
		Multivariate prediction	–	–	–	3·95	11·74	0·67

SEP, standard error of prediction; RSD, relative standard deviation.
* The accuracy is assessed by the SEP and RSD, and both are expressed in percentage. The R^2 is also calculated.
† There is a significant difference in SEP values between the adjusted univariate prediction models and the multivariate prediction model with the permutation test ($P < 0·05$).
‡ The original parameter coefficients are not available.

prediction model and each of the adjusted published prediction models.

In the French DB, the prediction of BF% was based on age, height and weight. The SEP values of our multivariate prediction model were 3·74 and 3·95 %. They were slightly higher than those of Gómez-Ambrosi's prediction model (3·63 %) in men and than those of Gallagher's and Levitt's prediction models (3·93 %) in women.

Discussion

BF, TF and other segmental compartments, such as appendicular muscle mass, are useful factors for assessing predisposition to metabolic risks; therefore, examinations of these segmental compartments provide interesting information. The proposed multivariate model aimed at simultaneously predicting them from age and easily measured anthropometric predictor variables, with a particular focus on the importance of waist circumference. It was built using a US dataset and validated independently using two different datasets. The present results showed that, with the proposed combination of four predictor variables, including waist circumference, the multivariate model enabled accurate predictions for segmental body compositions.

Waist circumference is a well-known predictor of abdominal accumulation of subcutaneous and visceral adipose tissues. In 2001, the National Cholesterol Education Program – Adult Treatment Panel III included waist circumference as a risk factor for the metabolic syndrome[1]. Waist circumference was then widely used to improve the prediction of BF% in combination with a weight-for-height index, such as BMI[28,29]. In the study by Lean *et al.*[30], BF%, which was assessed by densitometry, was more closely related to waist circumference than to BMI, particularly for men. In another study related to BFF, Bosty-Westphal *et al.*[31] found that waist circumference was a risk factor for decreased BFF and that it was a good anthropometric index for health risk assessment. Similarly in the present study, the accuracy of our multivariate model was improved when waist circumference was entered as a predictor variable. This was particularly meaningful for men for the segmental compartments, such as TF, APL, total BF and total BL masses. For men, a significant improvement in accuracy was observed in all the BMI categories and in the age categories of 20–35, 35–50 and 50–65 years. In addition, waist circumference was especially required to improve the prediction accuracy for Black men in comparison with the other two ethnicity categories. We thus concluded that waist circumference should be included in the multivariate prediction model for normal, overweight and obese subjects, although it is known in clinical practice that there is a physical difficulty in measuring waist circumference of the latter subjects.

One important aspect of our proposed model is that it is capable of predicting simultaneously several segmental compartments; to our knowledge, this is the first proposal made for a multivariate model. The joint use of several

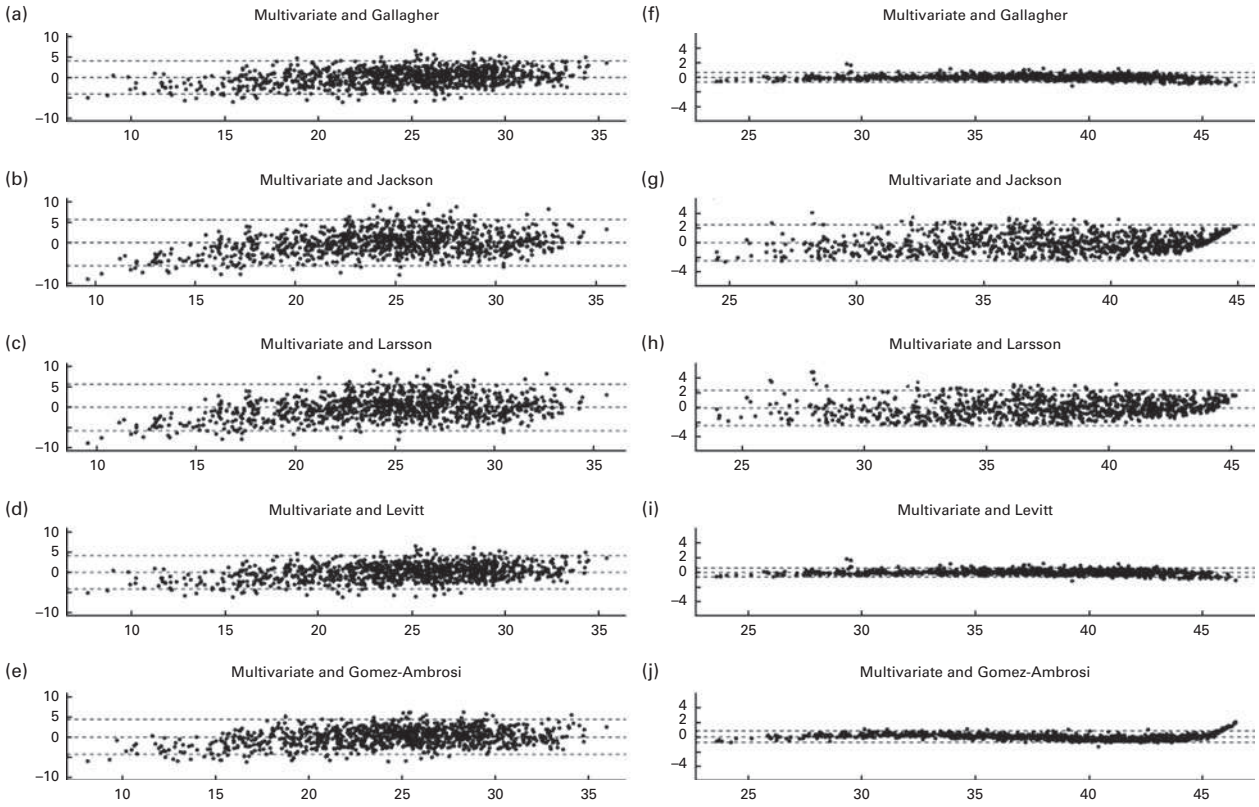


Fig. 2. Bland–Altman plots for the difference between body fat percentage (BF%) prediction by the multivariate model and that by the five adjusted published models *v.* average BF% prediction by the two models. The three dashed lines represent the mean difference and the mean and 1·96 SD. (a–e) Men and (f–j) women.

segmental body compositions has been justified in some metabolic disease risk studies. Indeed, an excess amount of TF is associated with a higher cardiometabolic risk, but in addition, after TF mass is controlled, a higher APF mass can be shown to be associated with a more favourable metabolic profile, particularly in women[32,33]. In a study on subjects aged 60–80 years, Saunders *et al.*[34] found that the absolute amount of TF and APF masses influenced the metabolic risk in elder men and women. Moreover, based on a study using DXA, BF was shown to be a complementary significant contributor to BMR in addition to BFF[35]. Some longitudinal studies in cohorts of older subjects[36,37] have highlighted that the loss of APL mass, measured using DXA, was associated with a greater risk of all-cause mortality compared with individuals with stable APL mass. Furthermore, Kilgour *et al.*[38] found that in advanced cancer patients, an APL mass-for-height index, measured by DXA, had a significant impact on cancer-related fatigue in men. Therefore, in order to better assess the health status or the metabolic risks of individuals, it is beneficial to predict simultaneously several segmental compartments from the statistical models. In the present study, the results for different populations underline that our proposed model enables the accurate assessment of several segmental compartments for the three ethnicities studied. The reliable prediction for body, trunk and appendicular components may be used for further studies related to pathophysiological and metabolic issues.

Of the already published models, five were retained for evaluating the usefulness of the proposed combination of four predictor variables in the multivariate model. These published models mainly integrated BMI and age as predictor variables; some were derived from either densitometry-based or air-displacement plethysmography-based datasets. Original and adjusted formulas, derived from the TRD, were applied to the VAD and their prediction accuracies were used as baseline values for comparison. The results show that the prediction of BF% of predictor variables used in our multivariate model yields a competing accuracy in comparison with the five adjusted published models. This finding justifies the relevance of using age, height, weight and waist circumference for predicting body composition.

Measurements of body composition can be obtained using a variety of methods, each of which provides a different amount of information about body compartments. Each method has specific limitations and measurement errors[39,40]. DXA and the four-compartment models are usually designated as reference methods for assessing body composition[41–43]. For BF%, the precision is approximately 3% for DXA and even lower than 3% for the four-compartment models[42]. If we take into account these measurement errors combined with the prediction accuracy of our model, we can calculate the

model precision using the following formula:

$$\sqrt{\text{DXA precision } (\%)^2 + \text{model prediction accuracy } (\%)^2}.$$

In our model, the SEP values for BF% were 3·2 % for men in the VAD and less than 4 % for women in the VAD and French DB. Our model thus yields an interesting precision of 4·4 and 4·8 % for men and women in the VAD and 5·0 % in the French DB. Interestingly, Lohman[44] developed standards for evaluating prediction errors (SEP) for BF%. He proposed that an ideal prediction would be denoted by a SEP value less than 2 %, a good prediction by a SEP value ranging from 3·5 to 4 % and a poor prediction by a SEP value greater than 5 %. According to these standards, our multivariate model with the four predictor variables yielded a good prediction error. Indeed, the SEP values for BF% were equal to 3·26 and 3·74 % in men and women, respectively. Even if our prediction model was shown to be good, it cannot replace a direct measurement such as DXA. Nevertheless, due to its easy application and cost efficiency, it appears to be a convenient tool to evaluate the need of DXA prescription. Besides this, the multivariate model enables to suggest a pathophysiological situation or detect a dangerous evolution in case of follow-up. Moreover, such applications could be of interest to educate patients with chronic metabolic diseases. Finally, from a research perspective, such a model could be highly relevant in predicting specific risks in large populations.

The present study was limited in some aspects. First, while working with the NHANES dataset, ethnic groups were limited to White, Black and Hispanic subjects for whom accurate predictions were provided. Furthermore, only subjects aged from 20 to 85 years with BMI values ranging from 18 to 40 kg/m^2 were examined. Subjects with a BMI >40 kg/m^2 were excluded because they are morbidly obese. Already for a BMI >30 kg/m^2, the accuracy of our model was lower than that for the other two BMI categories. Moreover, waist circumference has little incremental predictive power of disease risk for subjects with a BMI >35 kg/m^2 [45]. Thus, a particular study should be conducted to predict body composition of morbidly obese individuals. Finally, since data on waist circumference were not available in the French DB, the prediction of body composition for this database only used the three other predictor variables, with the result being a lower accuracy compared with that of the VAD. This result strengthens the conclusions regarding the importance of including waist circumference as a predictor variable.

In summary, waist circumference is an important predictor variable for the prediction of segmental body composition, especially in men. When using age, height, weight and waist circumference, our multivariate model yields a competing accuracy compared with other published univariate models for the prediction of BF%. Compared with these published formulas, the originality and advantage of the proposed model consist in predicting simultaneously several segmental compartments (such as TF mass or APL mass) with a good accuracy; the multivariate outcomes might then be used in studies necessitating the assessment of metabolic risk factors in large populations.

Acknowledgements

We thank the Human Nutrition Department and Applied Mathematics and Informatics unit of the French National Institute for Agricultural Research for a fellowship that permitted us to conduct the study. The authors are grateful to Dr Ristori from the Radiology Department of the Clermont-Ferrand University Hospital for providing DXA data from the Clermont-Ferrand University Hospital dataset. The authors' responsibilities were as follows: S. T. was responsible for model computations, statistical analysis and the first draft of the manuscript; L. M. was responsible for data acquisition, design of the study and physiological interpretation; J.-B. D. was responsible for the design of the study, model computations and statistical analysis; B. M. was responsible for the design of the study and physiological interpretation. All authors read and agreed with the contents of the manuscript. None of the authors has any conflict of interest concerning the manuscript.

References

1. Carr DB, Utzschneider KM, Hull RL, *et al.* (2004) Intra-abdominal fat is a major determinant of the National Cholesterol Education Program Adult Treatment Panel III criteria for the metabolic syndrome. *Diabetes* **53**, 2087–2094.
2. Vega GL, Adams-Huet B, Peshock R, *et al.* (2006) Influence of body fat content and distribution on variation in metabolic risk. *J Clin Endocrinol Metab* **91**, 4459–4466.
3. Greenlund LJ & Nair KS (2003) Sarcopenia – consequences, mechanisms, and potential therapies. *Mech Ageing Dev* **124**, 287–299.
4. Vandervoot AA & Symons TB (2001) Functional and metabolic consequences of sarcopenia. *Can J Appl Physiol* **26**, 90–101.
5. Snijder MM, Van Dam RM, Visser M, *et al.* (2006) What aspects of body fat are particularly hazardous and how do we measure them? *Int J Epidemiol* **35**, 83–92.
6. Sun SS & Chumlea WC (2005) Statistical methods. In *Human Body Composition*, 2nd ed., pp. 151–160 [SB Heymsfield, TG Lohman and Z Wang, *et al.*, editors]. Champaign, IL: Human Kinetics.
7. Gallagher D, Heymsfield SB, Heo M, *et al.* (2000) Healthy percentage body fat ranges: an approach for developing guidelines based on body mass index. *Am J Clin Nutr* **72**, 694–701.
8. Jackson AS, Stanforth PR, Gagnon J, *et al.* (2002) The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. *Int J Obes* **26**, 789–796.
9. Larsson I, Henning B, Lindroos AK, *et al.* (2006) Optimized predictions of absolute and relative amounts of body fat from weight, height, other anthropometric predictors, and age. *Am J Clin Nutr* **83**, 252–259.
10. Levitt DG, Heymsfield SB, Pierson RN Jr, *et al.* (2007) Physiological models of body composition and human obesity. *Nutr Metab (Lond)* **4**, 19–32.
11. Gómez-Ambrosi J, Silva C, Catalán V, *et al.* (2012) Clinical usefulness of a new equation for estimating body fat. *Diabetes Care* **35**, 383–388.
12. Mioche L, Bidot C & Denis JB (2011) Body composition predicted with a Bayesian network from simple variables. *Br J Nutr* **105**, 1265–1271.

British Journal of Nutrition

13. Mioche L, Brigand A, Bidot C, *et al.* (2011) Fat-free mass predictions through a Bayesian network enable body composition comparisons in various populations. *J Nutr* **1411**, 573–580.
14. Hastie T, Tibshirani R and Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer.
15. Nivre J (2006) *Inductive Dependency Parsing*. Dordrecht: Springer.
16. Centers for Disease Control and Prevention (2000) National Health and Nutrition Examination Survey: body composition procedures manual. http://www.cdc.gov/nchs/data/nhanes/BC.pdf (accessed 27 September 2008).
17. Centers for Disease Control and Prevention (2008) The 1999–2004 dual energy X-ray absorptiometry (DXA) multiple imputation data files and technical documentation. http://www.cdc.gov/nchs/about/major/nhanes/dxx/dxa.html (accessed January 2008).
18. Mazess RB, Barden HS, Bisek JP, *et al.* (1990) Dual-energy X-ray absorptiometry for total body and regional bone mineral and soft tissue composition. *Am J Clin Nutr* **51**, 1106–1112.
19. Wang ZM, Visser M, Ma R, *et al.* (1996) Skeletal muscle mass: evaluation of neutron activation and dual-energy X-ray absorptiometry methods. *J Appl Physiol* **80**, 824–831.
20. Sprent P & Smeeton NC (2001) *Applied Nonparametric Statistical Methods*, 3rd ed. Boca Raton, FL: Chapman, Hall/CRC.
21. Vapnik VN (1998) *Statistical Learning Theory*. New York, NY: Wiley.
22. Lin CF & Wang SD (2002) Fuzzy support vector machines. *IEEE Trans Neural Net* **13**, 464–471.
23. Gelman A, Carlin JB, Stern HS, *et al.* (2003) *Bayesian Data Analysis*, 2nd ed. Boca Raton, FL: Chaplan & Hall/CRC.
24. Anderson TW (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann Math Statist* **22**, 327–351.
25. Röhmel J (1996) Precision intervals for estimates of the difference in success rates for binary random variables based on the permutation principle. *Biometrical J* **38**, 977–993.
26. Bland JM & Altman DG (1996) Statistical method for assessing agreement between two methods of clinical measurement. *Lancet* **i**, 307–310.
27. R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org (accessed January 2011).
28. Jassen I, Heymsfield SB, Allison DB, *et al.* (2002) Body mass index and waist circumference independently contribute to prediction of nonabdominal, abdominal subcutaneous, and visceral fat. *Am J Clin Nutr* **75**, 683–688.
29. Aeberli I, Gut-Knabenhans M, Kusche-Ammann RS, *et al.* (2012) A composite score combining waist circumference and body mass index more accurately predicts body fat percentage in 6- to 13-year-old children. *Eur J Nutr* **52**, 247–253.
30. Lean ME, Han TS & Deurenberg P (1996) Predicting body composition by densitometry from simple anthropometric measurements. *Am J Clin Nutr* **63**, 4–14.
31. Bosty-Westphal A, Danielzik S, Geisler C, *et al.* (2006) Use of height^3:waist circumference^3 as an index for metabolic risk assessment. *Br J Nutr* **95**, 1212–1220.
32. Snijder MB, Dekker JM, Visser M, *et al.* (2004) Trunk fat and leg fat have independent and opposite association with fasting and postload glucose levels: the Hoorn study. *Diabetes Care* **27**, 372–377.
33. Van Pelt RE, Evans EM, Schechtman KB, *et al.* (2002) Contribution of total and regional fat mass to risk for cardiovascular disease in older women. *J Physiol Endocrinol Metab* **282**, 1023–1028.
34. Saunders TJ, Davidson LE, Janiszewski PM, *et al.* (2009) Association of the limb fat to trunk fat ratio with makers of cardiometabolic risk in elderly men and women. *J Gerontol A Biol Sci Med Sci* **64**, 1066–1070.
35. Johnstone AM, Murison SD, Duncan JS, *et al.* (2005) Factors influencing variation in basal metabolic rate include fat-free mass, fat mass, age, and circulating thyroxine but not sex, circulating leptin, or triiodothyronine. *Am J Clin Nutr* **82**, 941–948.
36. Szulc P, Munoz F, Marchand F, *et al.* (2010) Rapid loss of appendicular skeletal muscle mass is associated with higher all-cause mortality in older men: the prospective MINOS study. *Am J Clin Nutr* **91**, 1227–1236.
37. Lee CG, Boyko EJ, Nielson CM, *et al.* (2011) Mortality risk in older men associated with changes in weight, lean mass, and fat mass. *J Am Geriatr Soc* **2**, 233–240.
38. Kilgour RD, Vigano A, Trutschnigg B, *et al.* (2010) Cancer-related fatigue: the impact of skeletal muscle mass and strength in patients with advanced cancer. *J Cachexia Sarcopenia Muscle* **1**, 177–185.
39. Ellis KJ (2000) Human body composition: *in vivo* methods. *Physiol Rev* **80**, 649–680.
40. Lee SY & Gallagher D (2008) Assessment methods in human body composition. *Curr Opin Clin Nutr Metab Care* **11**, 566–572.
41. Wellens R, Chumlea WC, Guo S, *et al.* (1994) Body composition in white adults by dual-energy X-ray absorptiometry, densitometry, and total body water. *Am J Clin Nutr* **59**, 547–555.
42. Plank LD (2005) Dual-energy X-ray absorptiometry and body composition. *Curr Opin Clin Nutr Metab Care* **8**, 305–309.
43. Lohman TG & Chen Z (2005) Dual-energy X-ray absorptiometry. In *Human Body Composition*, 2nd ed., pp. 63–78 [SB Heymsfield, TG Lohman and Z Wang, *et al.*, editors]. Champaign, IL: Human Kinetics.
44. Lohman TG (1992) *Advances in Body Composition Assessment. Current Issues in Exercise Science Series (Monograph 3)*. Champaign, IL: Human Kinetics.
45. National Institutes of Health (2000) *The Practical Guide: Identification, Evaluation, and Treatment of Overweight and Obesity in Adults*. Bethesda, MD: National Institutes of Health.

British Journal of Nutrition

3.3 Body composition changes in aging

Age-related changes in body composition have been increasingly recognized as a potentially modifiable factor in the quest for optimal health, function, and longevity. The shifts with aging in body composition toward more body fat mass, especially the accumulation of more internalized fat deposits, and the loss of muscle mass. Previous studies showed that after age 60, total skeletal muscle mass declined by 0.8 kg and 0.4 kg over a 5-year period in men and women, respectively. As changes in body composition are associated with increase the risk of a wide range of chronic disorders, it is useful to monitor these age-related changes. Indeed, describing the aging effect on body composition provides important epidemiologic information for identifying possible etiologic mechanisms as targets for treatment and prevention of such adverse health outcomes, given the worldwide growth rates of obesity and the increase of the aging population. Furthermore, in the clinical setting, this kind of studies can give healthcare professionals insight on future state of individual health, and can be used to support a preventative health programme, helping to educate patients about the importance of maintaining healthy fat levels for life-long good health. By understanding what is going on with aging, healthcare professionals can make better-informed decisions for individual behaviour, as well as can develop a more personalized exercise and nutrition program.

In this part, we will develop two modeling methods for assessing age-related changes. In subsection 3.3.1, a Bayesian modeling will be described. As changes in body size and shape (*i.e.*, height, weight and waist circumference) also occur with aging, we think that it is suitable to consider these anthropometric variables firstly, then to study changes in body composition with aging and the anthropometric variables. Ethnicity-related differences in body composition have been recognized (Wu *et al.*, 2007; Aleman-Mateo *et al.*, 2009), also are reported an ethnicity effect on age-related changes in body fat-free masses (Obisesan *et al.*, 2005). Nevertheless there are few studies on other segmental body compositions. In subsection 3.3.2, we attempt to determine age-related changes in trunk fat, appendicular lean masses and other compartments, according to BMI at the age of 20 y, ethnicity and history of weight change during life. To do so, a frequentist modeling will be investigated by assessing different cases of study, and this approach leads to a research paper submitted to *British Journal of Nutrition*.

3.3.1 Bayesian modeling for age-related changes in body composition

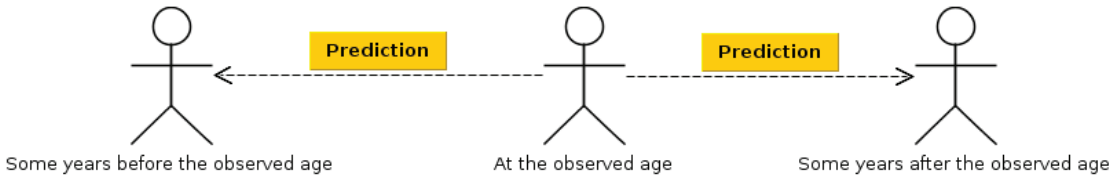
We have a cross-sectional dataset[2] NHANES, where each subject is observed once. For each subject, height (H), weight (W), waist circumference (WC) and segmental body compositions (SBC) are available at the age (A) when the subject was recruited in the NHANES. The main aim of study is that : for each subject, given subject's single observation (the little man in the middle of Figure 3.11), how SBC change during the lifespan (from 20 to 85 years)? The key idea of prediction processus is consisting of two assumptions :

- we are able to model covariate change;
- variable changes are mainly related to covariate changes.

In the first step, at a given time, we follow the order displayed in Figure 3.12 for covariate assessment. First of all, height is investigated and it depends on age. Secondly, weight is

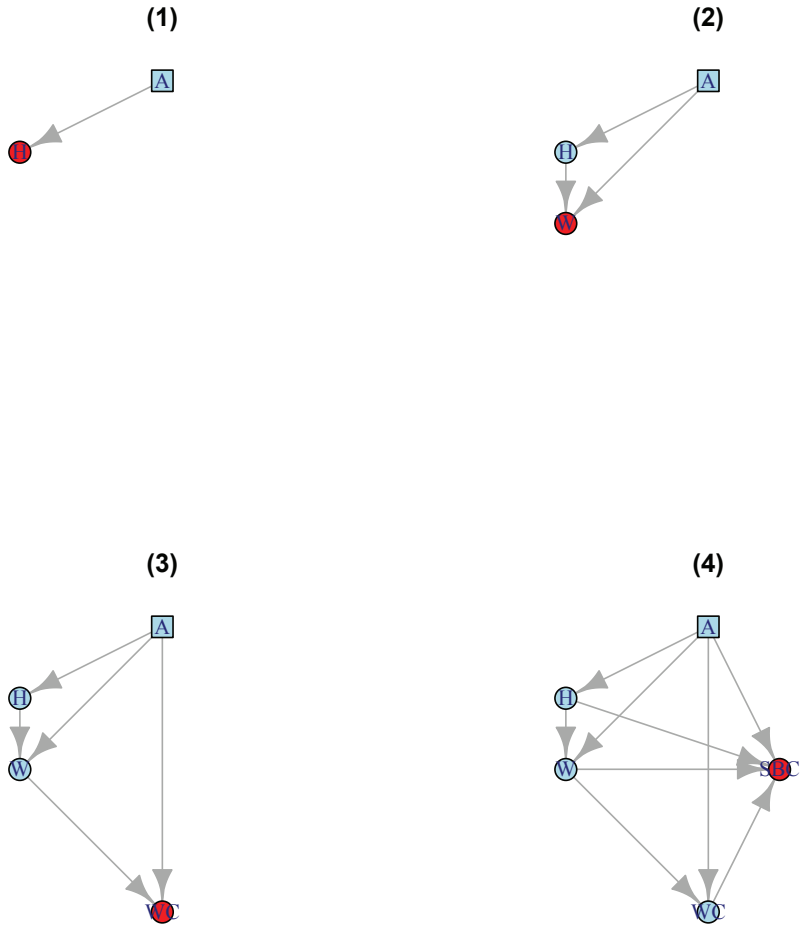
[2]Cross-sectional dataset refers to observations of many different subjects at a given time, each observation belonging to a different subject. For instance, a simple cross-sectional dataset is height measurement for 100 randomly chosen patients in hospital for 2013. Cross-sectional dataset is distinguished from longitudinal dataset, where there are multiple observations for each subject, over time.

Figure 3.11: Scheme of prediction processus.



assessed by age and estimate of height. Finally waist circumference is studied in turn, and its estimate is directly obtained from age and weight, as well as from information passing from height to waist through weight. In the second step, four covariates are considered together to predict segmental body compositions.

Figure 3.12: A Bayesian network presents conditional dependencies between covariates and SBC at a given time. The four subfigures presents the order of covariates/variable assessment. The red node indicate the variable to be modelled and the arrows indicate the dependency.

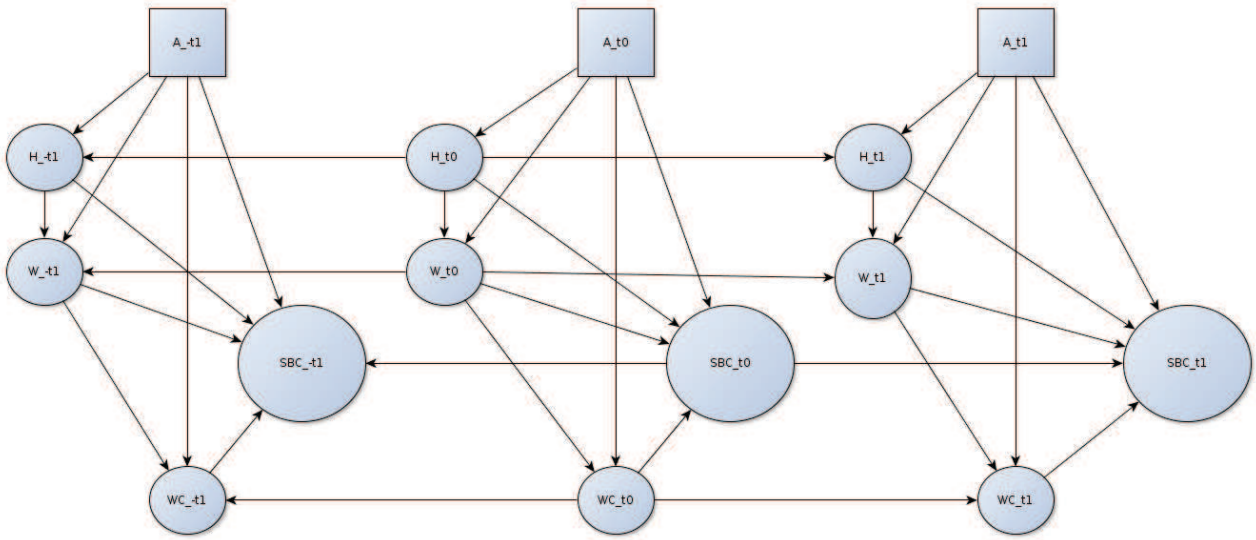


Now we will extend our modeling from the static structure to a dynamic structure. Before doing so, we will clarify some definition and assumptions (details and arguments are given in the next section) :

1. Definition of t_0 : the time (or the age) when subjects are observed in the dataset;
2. Given t_0, $AGE_{t_0}, HGT_{t_0}, WGT_{t_0}$, WAI_{t_0} and $SBC_{t_0}$are known;
3. Within each gender, assume that age-related changes in height, weight and waist circumference are identical for each subject;
4. Assume that height, waist circumference and SBC follow a Normal distribution, whereas weight follows a lognormal distribution.

Figure 3.13 illustrates the prediction scheme of proposed dynamic structure. Available information at t_0 has two main uses : firstly it will be used as a random effect, which implies individual characteristic; secondly it will be used as the starting point, which makes predictions downstream and upstream. More details about reasoning and mathematical formulations will be discussed in the next subsections. The order of assessment for covariate variables and SBC follows the order in Figure 3.12.

Figure 3.13: Dynamic structure of Bayesian network presents conditional dependencies between covariates and SBC during different time intervals. t_0 is the time or the age when subjects are recruited in the dataset. The horizontal arrows between two time intervals mean the predictions either downstream or upstream.



3.3.1.1 Age-related Normal Distribution for Height

The decline in human height with age in the datasets such the ours is well recognized. It generally results from (1) the physiology of aging in the individuals and (2) so called birth cohort effect. Birth cohort effect is the fact that previous generations were on average shorter than more recent generations. Studies showed that average human height increased about 1 *cm/decade* over the century from the 1850s for many countries of Western Europe (Floud, 1994; Steckel, 1995). This remark will be used when assessing height change in the cross-sectional datasets, because subjects, who are collected in the same moment, are not born in the same year. Thus, when studying a height change during the lifespan with cross-sectional datasets, we should take into account birth cohort effect. An intuitive way to correct individual height is by correcting the dataset such that all subjects were born in the same year. In the present study, we are interested in assessing age-related change in height, however the available datasets are rather cross-sectional than longitudinal, therefore before modeling height on age, an adjustment is conducted to get rid of birth cohort effect.

A cross-sectional dataset from a medical examination center at Saint-Brieuc (France) is used to assess age-related change in height. Height characteristics are shown in Table (3.10). At each age interval, the number of observations is relevant, accordingly we ensure that average of height is well estimated. From Table (3.10), a decrease of height is found when aging, but this decline is not only due to age effect, since there is a *birth cohort effect* that makes the elderly shorter than the younger. As mentioned above, average of height increases about 1 *cm/decade*, by transforming the 1 *cm/decade* increase to 0.1 *cm/year*, we thus obtain an adjusted height, denoted HGT.adj and it is calculated by :

$$HGT.adj = HGT + (AGE - 20) \times 0.1 \tag{3.11}$$

The adjusted height (HGT.adj) implies height value if individual were born in the same year of the individuals aged 20 year in the cross-sectional dataset. Mean and standard deviation of adjusted height for age intervals are also summarized in Table (3.10). After removing birth cohort effet, height reaches approximately its maximum at 35-40 years, then decreases in some rate. Somehow, we modify height values from transversal type to longidutinal type, and these adjusted heights will be considered in the following age-related change study.

Now we focus on the main objective of this section : age-related change in height. For this topic, both cross-sectional and longitudinal datasets were conducted in the literature to assess the change. The results from cross-sectional studies might be less reliable than longitudinal studies, but the conclusions come out the same. Otherwise the statistical models of height from age cover from linear form to quadratic (Galloway, 1988; Sorkin *et al.*, 1999; Morgan, 2010). In this framework, we attempted to construct an age-related probability distribution for height, also denoted as conditional distribution of height given age. For each subject i, $i \in \{1, \ldots, N\}$ (N sample size), the general form of conditional distribution of height is :

$$HGT_i|(AGE_i, \alpha_i^H) \sim \mathcal{N}(\alpha_i^H + \alpha_1 \times AGE_i + \alpha_2 \times AGE_i^2, \sigma_H^2) \tag{3.12}$$

$$\alpha_i^H \sim \mathcal{N}(\mu_H, \sigma_\alpha^2) \tag{3.13}$$

It is worth noting that the notation α is especially reserved for parameters in the height conditional distribution[3]. Mean of height in equation (3.12) is viewed as two parts: part of random effect α_i^H and part of fixed age effet $\alpha_1 \times AGE_i + \alpha_2 \times AGE_i^2$. Random effect is associated to each subject and it implies individual characteristics, while fixed age effect refers to population

[3]other greek letters will be used for other covariables and variables in the following subsections.

characteristics which are similar for all subjects, more precisely, all subjects share the same change pattern with age. Variance component σ_α^2 measures the between-subjects variability, while σ_H^2 accounts for measurement error for individual i. A quantitative description of the previous conditional distribution and additional parameters are given later. It is worth noting that the parameters in the conditional distribution will be estimated by Bayesian methods and to do so, we have to define the prior distribution for the parameters.

Bayesian methods allow to integrate experts' knowledges or previous findings into the prior distribution, then by incorporating the prior informations with datasets, we enable to update a new estimated distribution, called posterior distribution. Following this logic, we will take into account the published results to define the prior distribution. Therefore we think suitable to mention several interesting works in the literature.

Galloway (1988) put forward a simple model to use in forensic anthropology, and the author used a sample of 550 white individuals from southern Arizona aged 50 - 92 years. Their height was measured and they were asked to report their present height and their height at 25, which was assumed to be their maximal height. From Galloway's regression estimates, the results showed that decline in height did not begin until 45 years, with a loss of 0.172 cm/year for men and 0.155 cm/year for women. Overall, estimated height of an individual in his/her older age can be calculated by :

$$\widehat{HGT} = \begin{cases} HGT_{AGE=25} - 0.172 \times (AGE - 45)_+, & \text{For men} \\ \\ HGT_{AGE=25} - 0.155 \times (AGE - 45)_+, & \text{For women} \end{cases} \tag{3.14}$$

where

$$(AGE - 45)_+ = \begin{cases} (AGE - 45), & (AGE - 45) \geq 0 \\ \\ 0, & \text{otherwise} \end{cases} \tag{3.15}$$

Sorkin *et al.* (1999) studied age-related change in height from a longitudinal dataset and they used both linear and quadratic formulation to describe the relationship between height change rate (loss or gain) and age. Only linear form of change rate is presented here, because it allows to obtain a quadratic form of height on age by a simple multiplication. More precisely, the general function of height on age is written as $Height = \beta_0 + \beta_{Subject} AGE$ with $\beta_{Subject}$ change rate (Sorkin *et al.*, 1999, p. 970). Their linear formula of the change rate $\beta_{Subject}$ is given :

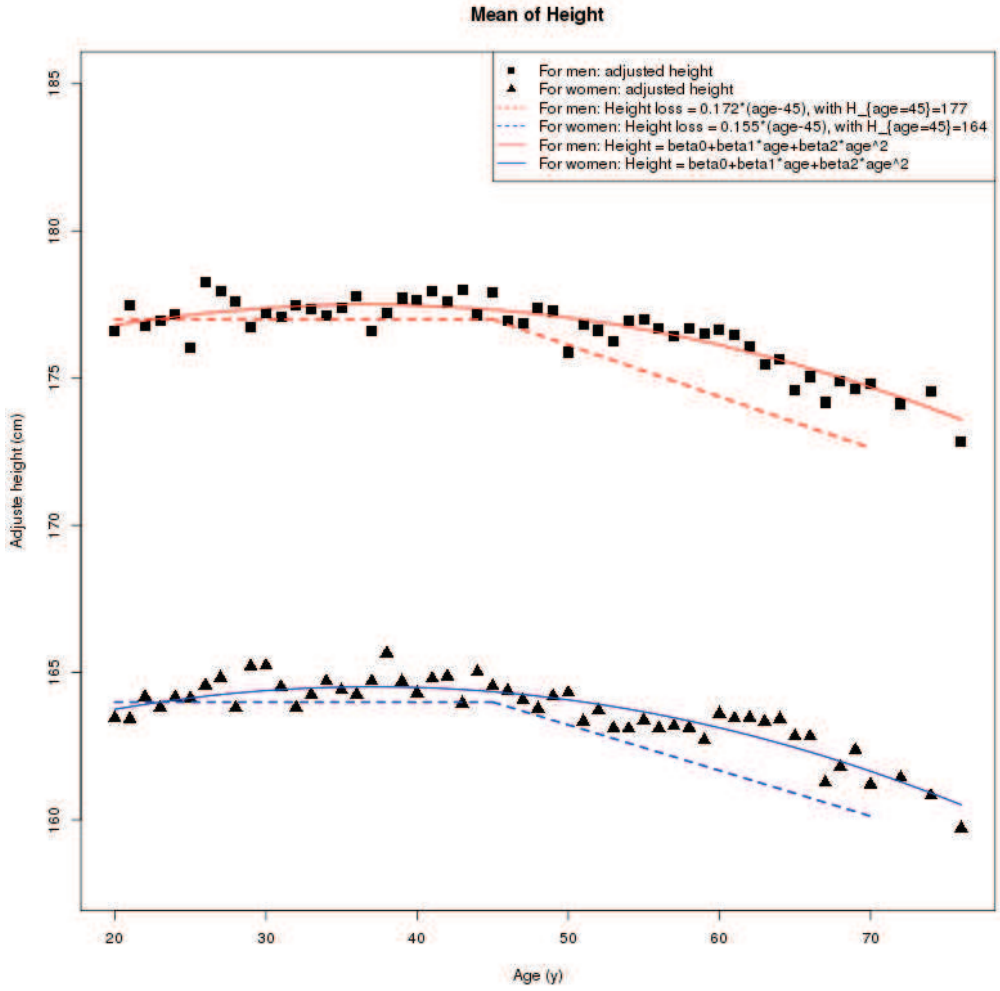
$$\beta_{Subject} = \begin{cases} -0.0912 - 0.00478 \times (AGE - 50.6), & \text{for men} \\ \\ -0.1605 - 0.00654 \times (AGE - 53.5), & \text{for women} \end{cases} \tag{3.16}$$

Injecting the change rate expression (3.16) into the general function of height, we thus deduce age-related model for height :

$$\widehat{HGT} = \begin{cases} \beta_0 + 0.151 \times AGE - 0.00478 \times AGE^2, & \text{For men} \\ \\ \beta_0 + 0.189 \times AGE - 0.00654 \times AGE^2, & \text{For women} \end{cases} \tag{3.17}$$

For the purpose of choosing the polynomial degree in the height model, we perform a graphical comparison study. Figure 3.14 shows times series plots of mean of adjusted height (HGT.adj) by age, Galloway's model and a fitted Sorkin *et al.*'s quadratic model are also drawn in the same figure. We observe that the quadratic form represents a better approximation of age-related change in height, therefore we expressed mean of height in the conditional distribution (3.12) with a quadratic form.

Figure 3.14: Comparison study for choosing polynomial degree in height model. Adjusted height (on y axis) is plotted against age (on x axis). The dashed line represents Galloway (1988)'s model and the solid line represents a fitted quadratic model (inspired by Sorkin *et al.* (1999)). Men : (■); Women : (▲). The red lines (dashed and solid) : models for men; the blue lines (dashed and solid) : models for women.



Now we can give a full description of the conditional distribution of height given age (expression (3.12)). We assume that $\sigma_H^2 = 0.3^2$ for both gender and we use the approximate values of coefficients in equation (3.17) as prior mean of α_1 and α_2. Overall, the prior distributions for $\alpha_1, \alpha_2, \mu_H$ and σ_α for both genders are summarized :

Table 3.6: Prior distributions for the parameters in the conditional distribution of height.

Parameters	Men	Women
α_1	$\mathcal{N}(0.15, 0.1^2)$	$\mathcal{N}(0.18, 0.2^2)$
α_2	$\mathcal{N}(-0.004, 0.01^2)$	$\mathcal{N}(-0.006, 0.01^2)$
μ_H	$\mathcal{N}(175, 20^2)$	$\mathcal{N}(165, 15^2)$
σ_α	$\mathcal{U}(1, 40)$	$\mathcal{U}(1, 40)$

Conditional distribution with time series structure

Previously we introduced a conditional distribution of height on age in the static scheme, this approach allows to estimate individual height distribution when his/her age is available. Now we will extend this approach in a dynamic scheme, that is how individual height changes when aging. Individual height value when observed in the dataset is a very useful information for this dynamic approach, because it will be considered as individual characteristics, which is fixed rather than random in equation (3.12) elsewhere. More precisely, the dynamic structure for conditional distribution of height given age is inspired by a time series approach : the height in the future is determined by the height today, plus an age-related change during this time interval. The formulation is given as follows :

$$HGT_{t+\Delta t} \sim \mathcal{N}(HGT_t + \lambda^H_{t+\Delta t}, \quad \sigma^2_H), \quad \lambda^H_{t+\Delta t} = \alpha_1 \times \Delta t + \alpha_2 \times [(t+\Delta t)^2 - t^2] \tag{3.18}$$

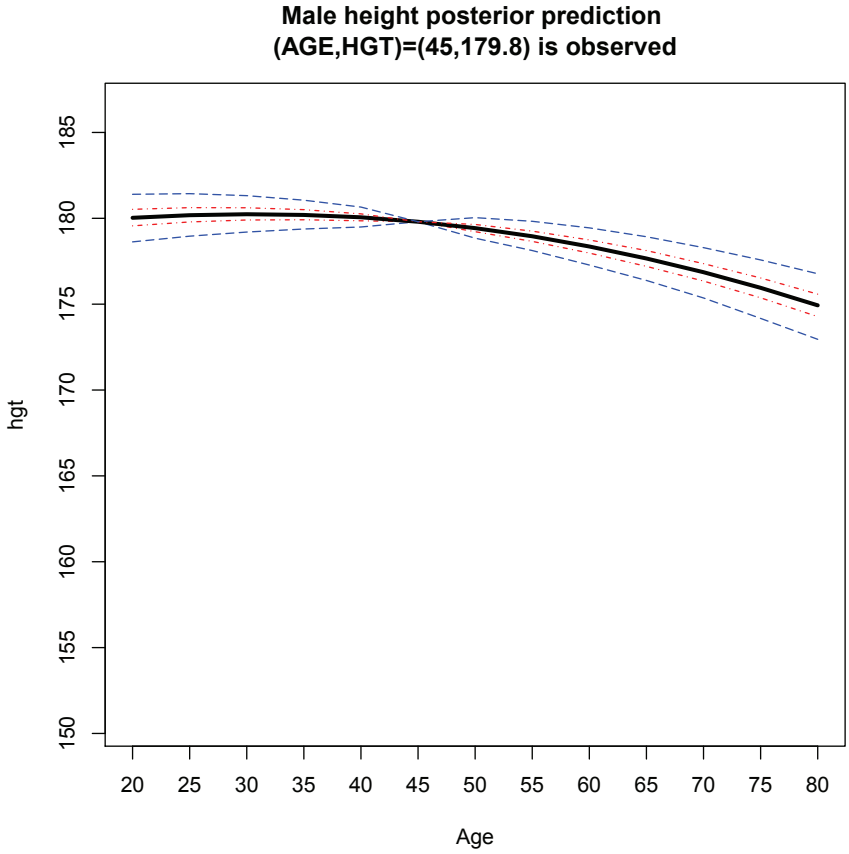
where Δt is time interval either between today and a past (downstream prediction) or between today and a future (upstream prediction). Δt can take all integers from $[-65, 65]\backslash\{0\}$, because this interval will cover observed age range in our dataset (*i.e.*, the observed age is between 20 and 85 year). The mean in this distribution has also two components which imply individual and population characteristics respectively : HGT_t is individual characteristics because it corresponds to observed value of height when the subject was included in the dataset; $\lambda^H_{t+\Delta t}$ is population characteristics because α_1 and α_2 are common for all subjects, given t and $t + \Delta t$. Equation (3.18) is more realistic to reflect height change during aging, because we update height prediction from one moment to the next using subject's proper observed information. In practice, we start from the age when subjects are observed, for example, a male subject is collected in the dataset at his age of 20 y, the observed value of HGT_{20} is used to predict height in the following years, say 25 y ($\Delta t = 5$), then predicted value of HGT_{25} is used for the next prediction HGT_{30}, and so on in this sequential way.

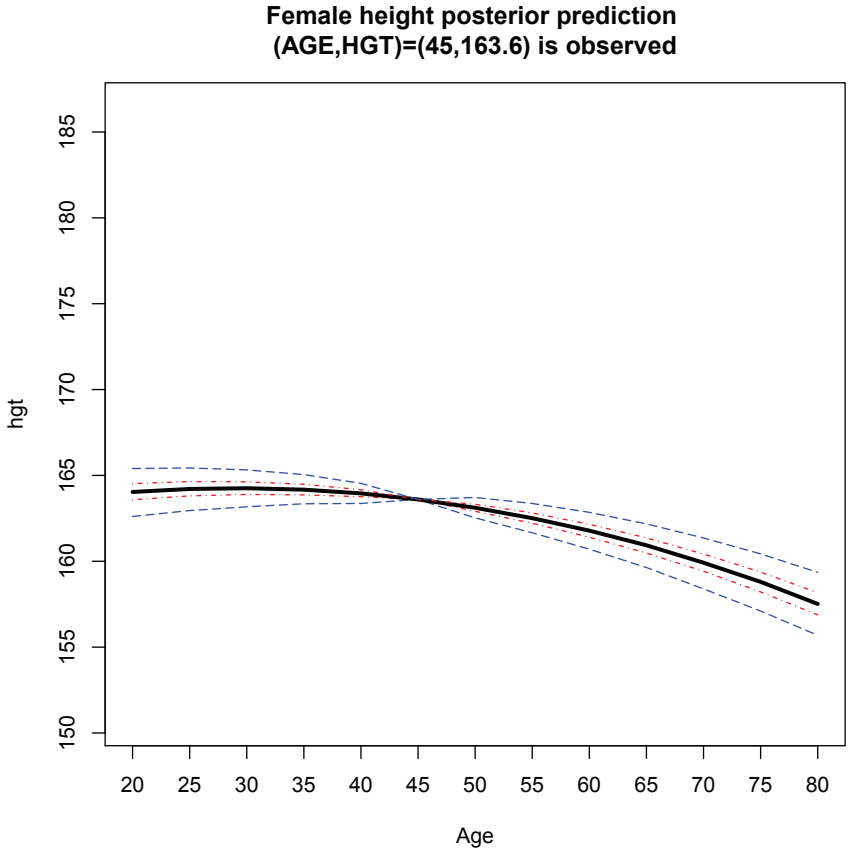
A simple example is simulated following the previous dynamic modeling (*i.e.*, equation (3.18)). We take respectively a male and a female subject whose observed age is 45 year when observed in the dataset; the corresponding height is 179.8 and 163.6 cm, respectively. By combinating the prior distribution and a dataset[4], we thus obtain the age-related change in height for both subjects, shown in Figure 3.15. Height is more likely stable until 45 year, then declines afterwards. The credible intervals of the mean reduce to zero at 45 year, because at that age, individual height is observed therefore known. More upstream and downstream prediction are far away 45 year, more the width of credible intervals are bigger. This implies a greater uncertainty in the prediction. Interestingly, in the subject-level of this example, the age-related change in height matches closely to that in the population level with an adjustment of height (Figure 3.14).

[4]We use the NHANES training dataset.

Figure 3.15: Age-related change in height for a male and female subjects, respectively.

For each gender, the observed age is 45 year, and the corresponding height is 179.8 and 163.6 cm for the male and the female subjects. The black solid line represents the mean of height during aging, the red and blue dashed line represent 95% and 50% credible intervals of the mean, respectively.





3.3.1.2 Age-related Lognormal Distribution for Weight

Several cross-sectional studies illustrate an increase in body weight throughout early and middle adulthood until approximately age 60 at which point the weight trajectory begins to decline (Guo *et al.*, 1999; Chumlea *et al.*, 2002; Lewis *et al.*, 1997). Longitudinal studies have confirmed these cross-sectional observations and have shown a decline in body weight in both genders after 60 (Visser *et al.*, 2003) or 70 (Gallagher *et al.*, 2000b; Raguso *et al.*, 2006; Hughes *et al.*, 2004) years of age. While it is clear that body weight decreases in elderly populations, the overall severity and consistency of this decline is not fully understood. These previous findings will help us to propose a conditional distribution of weight given age.

Early studies with NHANES datasets showed that body weight tends to follow a lognormal distribution (Burmaster and Crouch, 1997; Portier *et al.*, 2007). To confirm this finding in our study, a graphical investigation with the Saint-Brieuc dataset was performed within each gender. More precisely, a normal quantile-quantile plot (Q-Q plot) was drawn respectively for weight and log-transformed weight values. Clearly Figure 3.16 indicates that the lognormal assumption provides a more reasonable fit, therefore we retained the lognormal distribution form for weight.

Now we will focus on the mean form in the conditional distribution of weight. Burmaster and Crouch (1997) and Portier *et al.* (2007) used spline regression to model mean and standard deviation of the lognormal distribution of weight. In fact, they studied individuals from birth to 70+ years, and the spline regression enabled to take into account a particular weight change during adolescenthood. However in present study, only adult subjects (age $\geq$ 20 year) are studied, age-related change in weight is sufficiently fitted with parabolical curve. Besides, DAG proposed in Figure 3.12, we suppose a dependency of weight on age and height, therefore the mean of weight based on age and height is written :

$$\begin{aligned} log(WGT) &= \mathcal{W}(AGE, HGT) \\ &= \beta_0 + \beta_1 AGE + \beta_2 AGE^2 + \beta_3 HGT + \varepsilon_W \end{aligned} \tag{3.19}$$

Furthermore, from a preliminary comparison study of different mean forms (not shown), we found that equation (3.19) yielded a similar SEP value as did a second-order spline regression model with two spline knots. That suggests that we can simplify the spline model to our proposed weight model with age and height. Mean and standard deviation of weight are shown in Table (3.10). Time series plots of mean weight estimates in each age interval are displayed in Figure 3.17. The Burmaster and Crouch's smoothing curves and our fitted curves are also drawn in the same figure for graphical comparison. It is worth noting that our fitted curve is more unsteady, that is because of the height-related effect is taken into account in weight estimate during aging.

After elaborating the weight mean form, we can now deduce the general form of the conditional distribution of weight given age and height :

$$Log(WGT_i)|AGE_i, HGT_i \sim \mathcal{N}(\beta_i^W + \beta_1 \times AGE_i + \beta_2 \times AGE_i^2 + \beta_3 \times HGT_i, \sigma_W^2) \tag{3.20}$$

$$\beta_i^W \sim \mathcal{N}(\mu_W, \sigma_\beta^2) \tag{3.21}$$

where $i = 1, \ldots, N, N$ subjects. Analogous reasoning is made as in the previous section of height, we assume that $\sigma_W^2 = 0.005^2$ for both gender. Inspired by the coefficient values in Burmaster and Crouch (1997), we summarized the prior distribution for $\beta_k, k = 1, 2, 3, \mu_W$ and σ_β for men and women respectively :

Figure 3.16: Normal quantile-quantile plot (Q-Q plot) for weight and log-transformed weight values. Men are on the top, women are on the bottom. The left panels : Normal Q-Q plots; the right panels : Lognormal Q-Q plots.

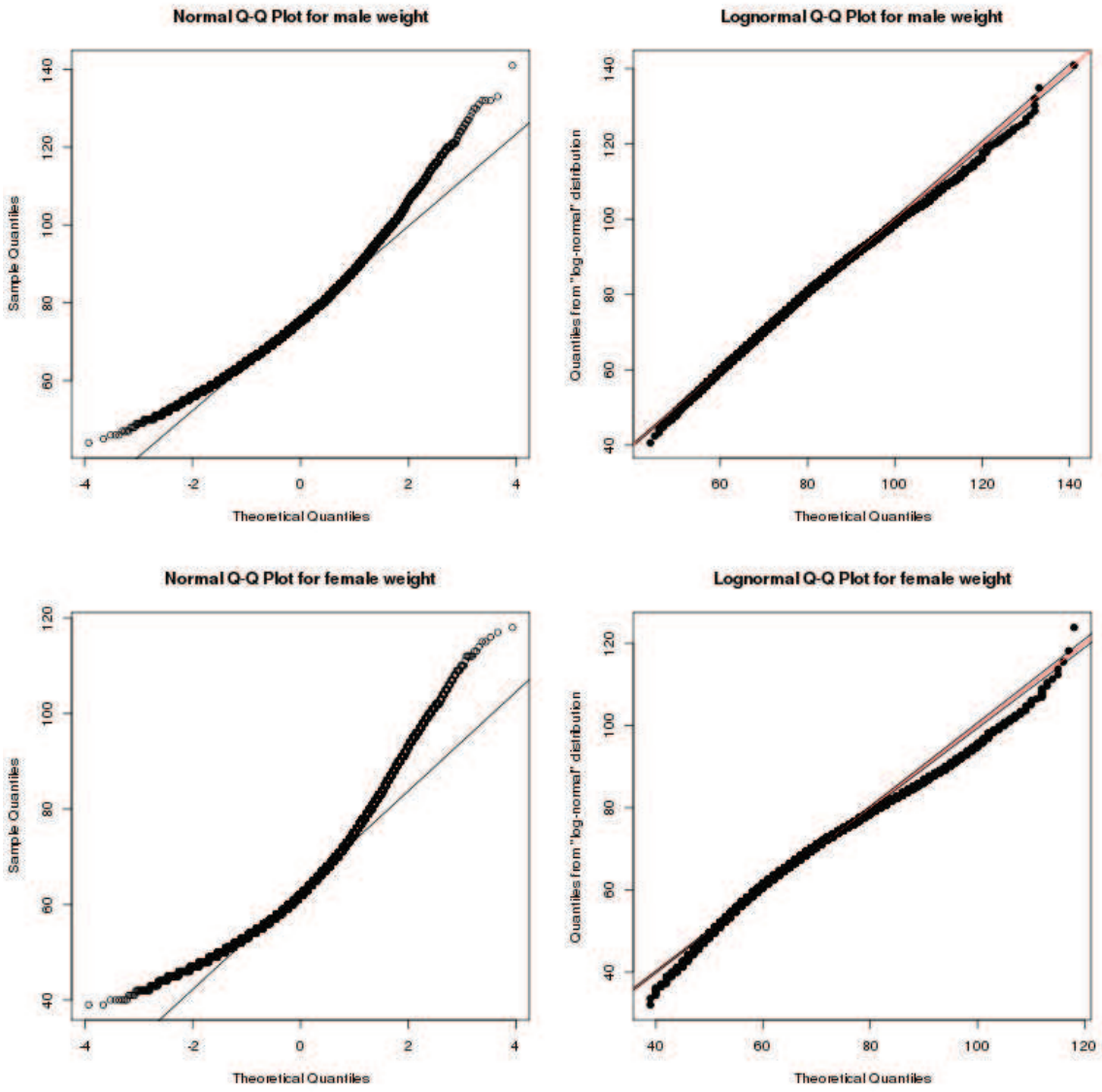
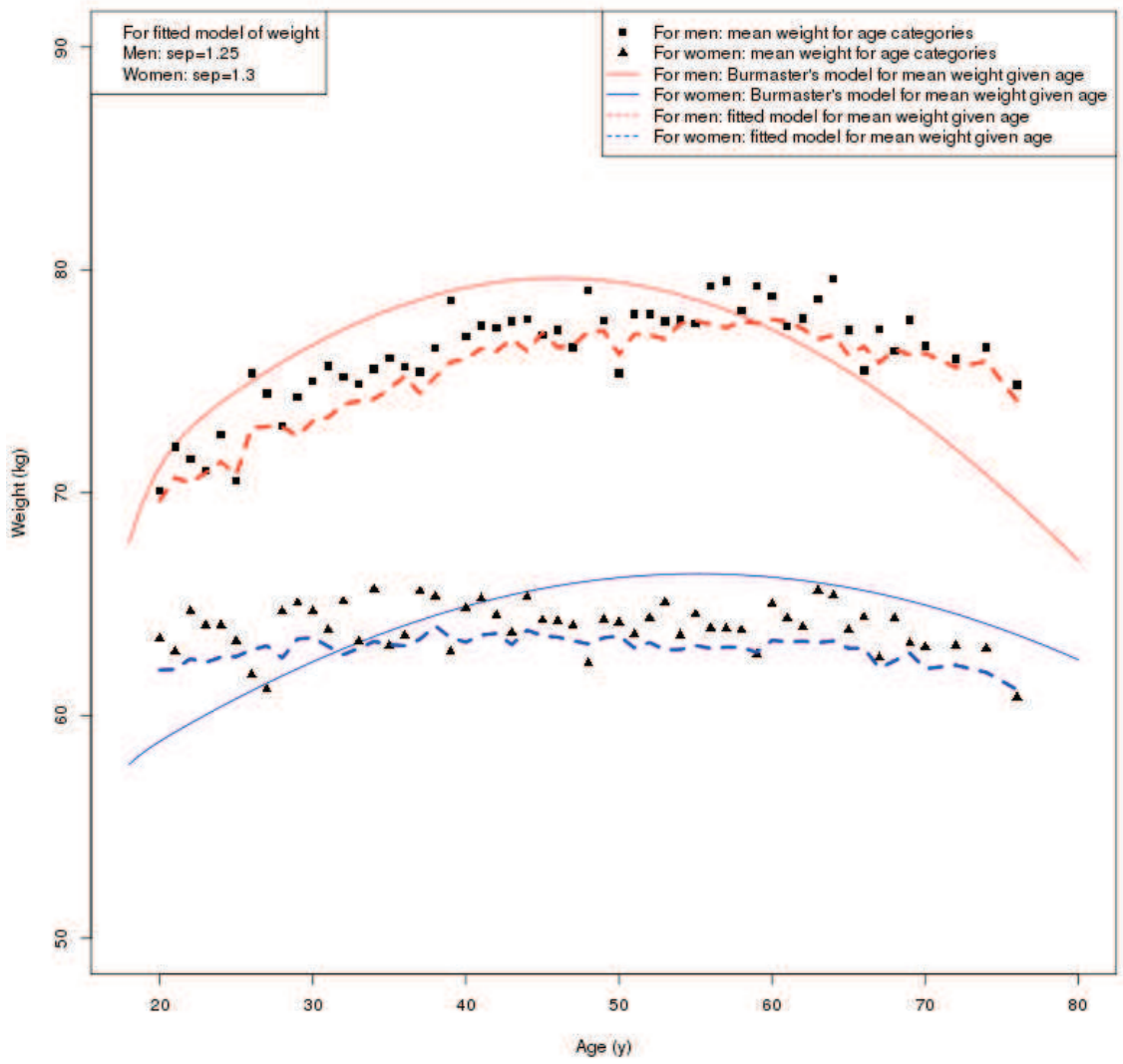


Table 3.7: Prior distributions for the parameters in the conditional distribution of weight for the age and the height in equation (3.20).

Parameters	Men	Women
β_1	$\mathcal{N}(0.015, 0.05^2)$	$\mathcal{N}(0.01, 0.01^2)$
β_2	$\mathcal{N}(-0.00015, 0.001^2)$	$\mathcal{N}(-0.000095, 0.001^2)$
β_3	$\mathcal{N}(0.01, 0.1^2)$	$\mathcal{N}(0.009, 0.01^2)$
μ_W	$\mathcal{N}(4.5, 2^2)$	$\mathcal{N}(4.1, 2.5^2)$
σ_β	$\mathcal{U}(0.1, 3.5)$	$\mathcal{U}(0.1, 3)$

Figure 3.17: Comparison study for choosing mean of weight expression in the conditional distribution. Weight (on y axis) is plotted against age (on x axis). The solid line represents Burmaster and Crouch (1997)'s model, and the dashed line represents our proposed model. Men : (■); Women : (▲). The red lines (dashed and solid) : models for men; the blue lines (dashed and solid) : models for women.



Conditional distribution with time series structure

Following the same reasoning done for the height, we propose a dynamic structure for the conditional distribution of weight given age and height. More precisely, the formulation is expressed :

$$Log(WGT)_{t+\Delta t} \sim \mathcal{N}(Log(WGT)_t + \lambda^W_{t+\Delta t}, \quad \sigma^2_W), \tag{3.22}$$

$$\lambda^W_{t+\Delta t} = \beta_1 \times \Delta t + \beta_2 \times [(t+\Delta t)^2 - t^2] + \beta_3 \times (HGT_{t+\Delta t} - HGT_t) \tag{3.23}$$

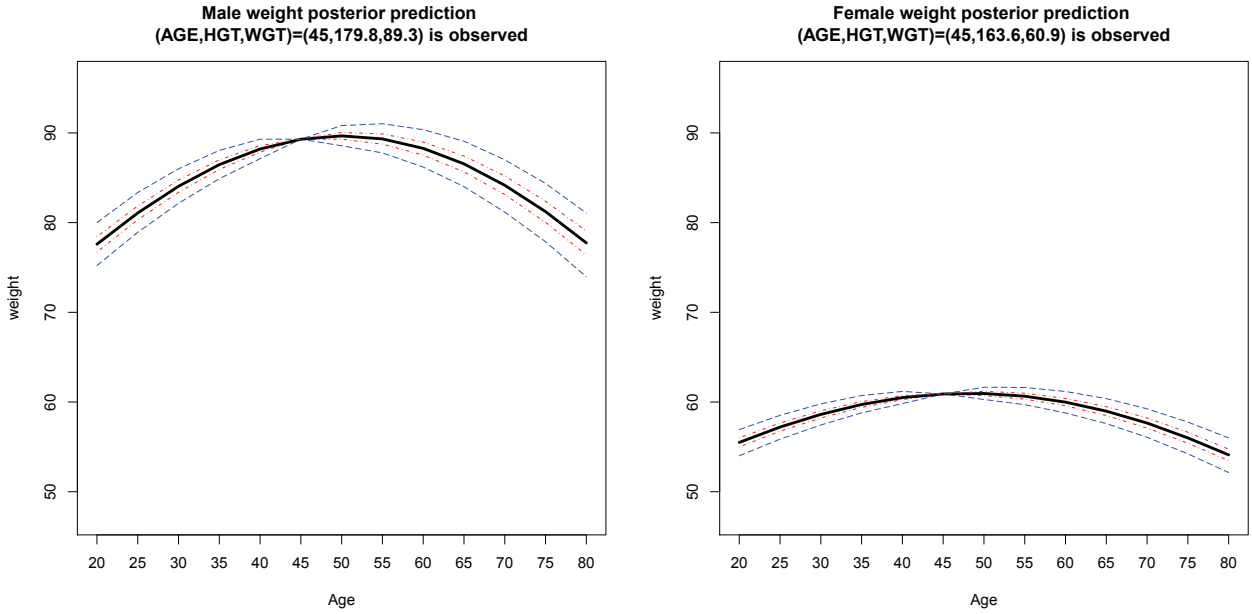
where Δt is time interval between today and the time for the prediction. WGT_t (or $Log(WGT)_t$) is used to emphasize individual characteristics, because it corresponds to observed value of

weight when the subject was included in the dataset. $\lambda^{W}_{t+\Delta t}$ corresponds to population characteristics because β_1, β_2 and β_3 are common for all subjects, given t and $t + \Delta t$. By using this dynamic structure, we are able to update weight prediction based on subject's proper observed information, either downstream or upstream. In practice, we will perform a sequential way to predict weight values during aging, as explained in the previous section for height.

Also, we conducted a simulation study with Bayesian dynamic modeling (equation (3.22)). Assume that a male and female subjects whose observed age is 45 y when collected in the dataset; the corresponding height and weight are 179.8 cm and 89.3 kg for the male subject, and 163.6 cm and 60.9 kg for the female subject. By combinating the prior distribution and a dataset, we thus obtain the age-related change in weight for both subjects, displayed in Figure 3.18. In both genders, weight increases until 50-55 year, then decreases afterwards. Compared with the female subject, the male peer has a greater rate of increase in weight, as well as a greater rate of loss after 55 year. Moreover, the credible intervals in the male subject are larger, particularly in the extremities of age intervals, that implies more variations and that may penalize the prediction.

Figure 3.18: Age-related change in weight for a male and female subjects, respectively.

For both genders, the observed age is 45 year; the corresponding height and weight are 179.8 cm and 89.3 kg for the male subject, and 163.6 cm and 60.9 kg for the female subject. The black solid line represents mean of weight during age, The black solid line represents mean of weight during aging, the red and blue dashed line represent 95% and 50% credible intervals , respectively.



3.3.1.3 Age-related Normal Distribution for Waist circumference

Data from NHANES show that waist circumference increases with age, and is larger in older than in younger adults of both genders up to the age of 70 years (Ford *et al.*, 2003). Similarly, in the Baltimore Longitudinal Study of Aging, age-related differences in waist-hip ratio were also reported in all BMI categories examined in both genders (Shimokata *et al.*, 1989). This study found that changes in waist circumference correlated directly with changes in weight. On average, with a 4.5 kg weight gain, men had a 4 cm increase in waist circumference and women had a 3.3 cm. Stevens *et al.* (2010) provided a good overview of age-related studies about waist circumference.

Based on these previous findings, we assume that waist circumference is affected by age and weight. Therefore, we suggested the following formulation of waist circumference on age and weight :

$$WAI = \gamma_0 + \gamma_1 \times AGE + \gamma_2 WGT \tag{3.24}$$

Figure 3.19 shows times series plots of waist circumference when aging, and a fitted linear model of waist (using equation (3.24)) is also drawn in the same figure. A good fitting of proposed model is found, thus it will be used to express the mean form in the conditional distribution of waist circumference later.

Here we describe the general form of the conditional distribution of waist circumference given age and weight :

$$WAI_i|AGE_i, WGT_i \sim \mathcal{N}(\gamma_i^C + \gamma_1 \times AGE_i + \gamma_2 WGT_i, \quad \sigma_C^2) \tag{3.25}$$

$$\gamma_i^C \sim \mathcal{N}(\mu_C, \sigma_\gamma^2) \tag{3.26}$$

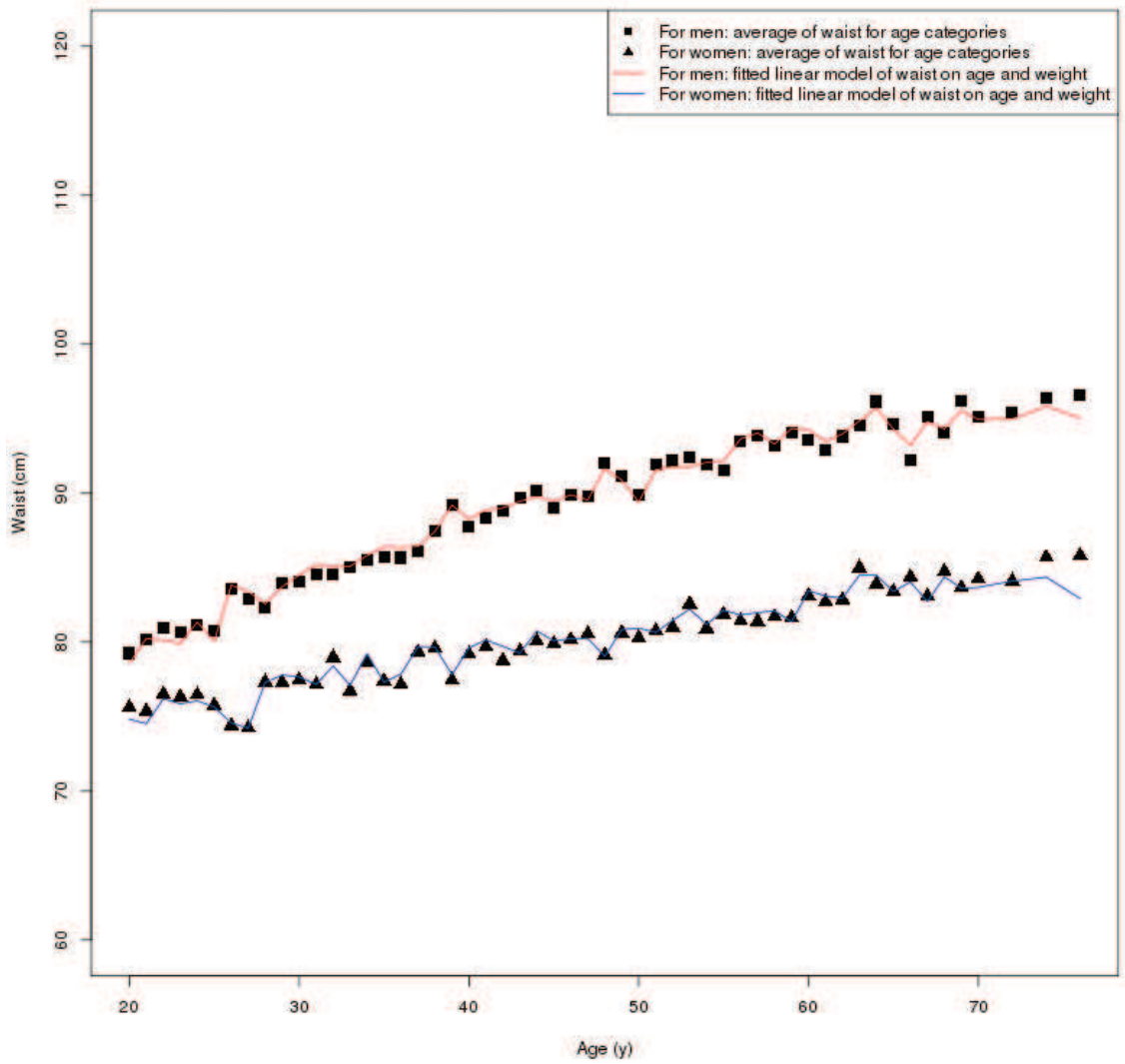
where $i = 1, \ldots, N, N$ subjects. γ_i^C represents random effect and $\gamma_1 \times AGE_i + \gamma_2 WGT_i$ fixed effect. It is worth mentioning that there is no height effect on waist circumference. Moreover we suppose that $\sigma_C^2 = 0.1^2$ for both gender, because σ_C is considered as a measurement error in the clinical setting.

The prior distributions for γ_1 and γ_2 are determined by using the published findings that *"On average, with a 4.5 kg weight gain, men had a 4 cm increase in waist circumference and women had a 3.3 cm"*. A waist-age and waist-weight ratio are calculated, and these two values will help us give prior mean for γ_1 and γ_2 respectively. International Diabetes Federation (IDF) published cut points to define a healthy waist circumference, and IDF guidelines that waist circumference cut points are 94 cm for European men and 80 cm for European women. Therefore these cut point values are considered as prior mean of μ_C for men and women. After all, the prior distributions for $\gamma_1, \gamma_2, \mu_C$ and σ_γ are summarized below :

Table 3.8: Prior distributions for the parameters in the conditional distribution of waist circumference defined in (3.25) and (3.26).

Parameters	Men	Women
γ_1	$\mathcal{N}(0.4, 2^2)$	$\mathcal{N}(0.3, 1.5^2)$
γ_2	$\mathcal{N}(0.9, 2^2)$	$\mathcal{N}(0.7, 1.5^2)$
μ_C	$\mathcal{N}(94, 20^2)$	$\mathcal{N}(80, 15^2)$
σ_γ	$\mathcal{U}(1, 20)$	$\mathcal{U}(1, 15)$

Figure 3.19: Time series plot for mean of waist circumference. Waist (on y axis) is plotted against Age (on x axis). The solid line represents our proposed model. Men : (■); Women : (▲). The red solide line : models for men; the blue solid line : models for women.



Conditional distribution with time series structure

Analogousely to the previous covariables, a dynamic structure could be proposed to emphasize the conditional distribution of waist circumference given age and weight. The formulation reads :

$$WAI_{t+\Delta t} \sim \mathcal{N}(WAI_t + \lambda^C_{t+\Delta t}, \quad \sigma^2_C), \quad \lambda^C_{t+\Delta t} = \gamma_1 \times \Delta t + \gamma_2 \times (WGT_{t+\Delta t} - WGT_t) \quad (3.27)$$

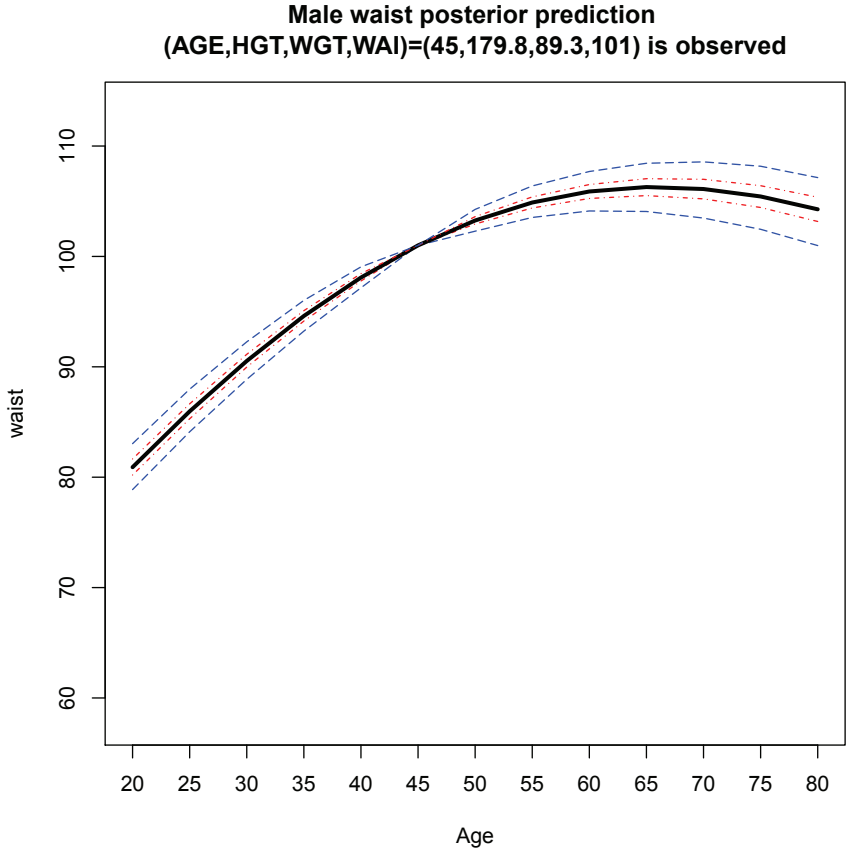
WAI_t implies individual characteristics and $\lambda^C_{t+\Delta t}$ implies population characteristics because γ_1 and γ_2 are common for all subjects, given t and $t + \Delta t$.

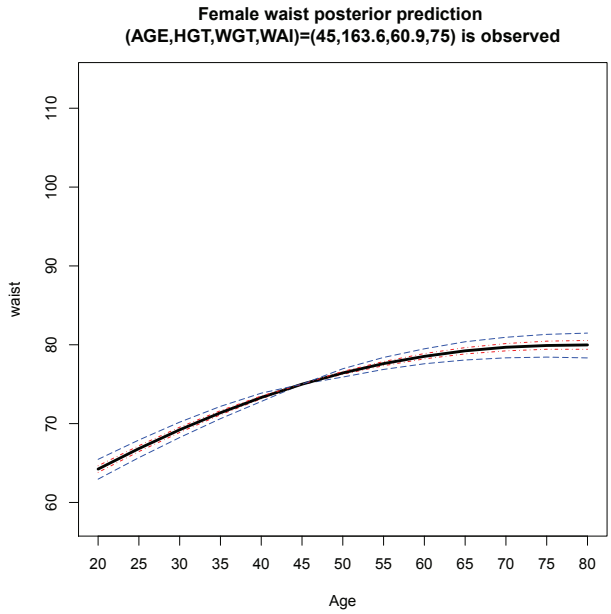
Here a simulation study with Bayesian dynamic modeling is performed for waist circumference. We took respectively a male and female subjects whose observed age is 45 y when collected in the dataset; the observed age is 45 year; the corresponding height, weight and waist circumference are 179.8 cm, 89.3 kg and 101 cm for the male subject, 163.6 cm, 60.9 kg

and 75 cm for the female subject. When integrating the prior distributions into a dataset, we are able to obtain the age-related change in waist circumference for both subjects, illustrated in Figure 3.20. The male subject has a greater increase of waist than the female peer from 20 year. In the male subject, waist increases consistently until 65-70 year, then declines slightly, whereas in the female subject, the increase occurs until 80 year. Furthermore, the credible intervals are larger in the older age of the male subject, this may result of the fact that waist circumference prediction depends on weight, and a greater variability in weight prediction affects the credible intervals of waist.

Figure 3.20: Age-related change in waist circumference for a male and female subjects, respectively.

For both gender, the observed age is 45 year; the corresponding height, weight and waist circumference are 179.8 cm, 89.3 kg and 101 cm for the male subject, 163.6 cm, 60.9 kg and 75 cm for the female subject. The black solid line represents mean of waist circumference during aging, the red and blue dashed line represent 95% and 50% credible intervals , respectively.





3.3.1.4 Age-related Normal Distribution for segmental body compositions

In the present subsection, we are interested in five segmental body compositions (SBC). They are trunk fat (tF), bdoy fat (bF), trunk lean (tL), appendicular lean (apL) and body lean (bL) masses. SBC can be predicted by age and anthropometric measurements, such as body weight and height. The advantages of using these covariates are simplicity and cost eficiency. Moreover their use would allow access to large datasets to describe body composition characteristics. Recently, Tian *et al.* (2013) justified usefulness of waist circumference for SBC prediction and proposed a multivariate modeling with age, height, weight and waist circumference as covariates. By following Tian *et al.* (2013)'s proposal of the covariates, we extend the frequentist approach into a Bayesian framework. More precisely, the previous four covariates are included to estimate mean of SBC in the conditional distributions. Also a random effect is integrated in the mean form to take into account individual characteristics. The general form of the conditional distribution of SBC is written as :

$$\begin{aligned} SBC_i^j | AGE_i, HGT_i, WGT_i, WAI_i &\sim \mathcal{N}(\eta_i^{SBC^j} + \eta_1^j AGE_i + \eta_2 HGT_i^j + \eta_3^j WGT_i + \eta_4^j WAI_i, \sigma^2_{SBC^j}) \\ \eta_i^{SBC^j} &\sim \mathcal{N}(\mu_{SBC^j}, \sigma^2_\eta) \end{aligned} \tag{3.28}$$

where $i = 1, \ldots, N, N$ subjects and $j = 1, \ldots, 5, 5$ SBC. Segmental body compositions can be seen as a proportion of weight, but it is not important to get the accurate value of proportion for each SBC, therefore we suggest a value of 0.5 as the prior mean for η_3, which is the parameter associated with weight. With respect to age, height and waist circumference, we assume a priori where there is no associated effect of these three covariates[5], thus 0 is taken as the prior mean for all of η_1, η_2 and η_4. The prior distributions for the parameters in equation (3.28) are given in Table 3.9 :

Table 3.9: Prior distributions for the parameters in the conditional distribution of the five SBC.

Parameters	Men	Women
η_1	$\mathcal{N}(0, 2^2)$	$\mathcal{N}(0, 2^2)$
η_2	$\mathcal{N}(0, 2^2)$	$\mathcal{N}(0, 2^2)$
η_3	$\mathcal{N}(0.5, 2^2)$	$\mathcal{N}(0.5, 2^2)$
η_4	$\mathcal{N}(0, 2^2)$	$\mathcal{N}(0, 2^2)$
μ_{tF}	$\mathcal{N}(11.21, 15^2)$	$\mathcal{N}(12.83, 15^2)$
μ_{bF}	$\mathcal{N}(21.21, 15^2)$	$\mathcal{N}(27.08, 15^2)$
μ_{tL}	$\mathcal{N}(29.74, 15^2)$	$\mathcal{N}(21.56, 15^2)$
μ_{apL}	$\mathcal{N}(26.75, 15^2)$	$\mathcal{N}(17.90, 15^2)$
μ_{bL}	$\mathcal{N}(60.06, 15^2)$	$\mathcal{N}(42.47, 15^2)$
σ_η	$\mathcal{U}(1, 30)$	$\mathcal{U}(1, 30)$

It is important to indicate that we attempted to impose a non-informative priors for parameters η, thus for each $\eta_i, i = 1, \ldots, 4$, we found the same prior distribution for the five predicted SBCs.

Conditional distribution with time series structure

We just propose the conditional distributions of SBC in a population-level. The prior distributions of parameters are provided mainly according to the published studies. Now we will give a

[5]As matter of fact, there exists covariate effects, however for the sake of simplicity, we consider the previous prior mean. Another way to appropriately define the prior mean for $\eta_j, j = 1, \ldots, 4$, would be use estimates of parameters in Tian *et al.* (2013).

dynamic structure of the conditional distribution in the subject-level. This dynamic structure is close to AR(1) type and can be used to predict age-related changes in SBC. Conceptually, the random effect parameter in equation (3.28) is replaced by individual observed SBC's value at the age when the subject was recruited in the dataset, while the fixed effect parameters remained invariant during the whole age interval. The dynamic conditional distribution of $SBC^j_{t+\Delta t}$, given its dependent variables (AGE, HGT, WGT, WAI), is as follows :

$$SBC^j_{t+\Delta t} \sim \mathcal{N}(SBC^j_t + \lambda^{SBC^j}_{t+\Delta t}, \sigma^2_{SBC^j}) \tag{3.29}$$

$$\begin{aligned} \lambda^{SBC^j}_{t+\Delta t} &= \eta^j_1 \times \Delta t \\ &\quad + \eta^j_2 \times (HGT_{t+\Delta t} - HGT_t) \\ &\quad + \eta^j_3 \times (WGT_{t+\Delta t} - WGT_t) \\ &\quad + \eta^j_4 \times (WAI_{t+\Delta t} - WAI_t) \end{aligned} \tag{3.30}$$

SBC^j_t is the observed value at age $= t$ and represents individual characteristics; $\lambda^{SBC^j}_{t+\Delta t}$ is the changes affected by a linear combination of the four covariates during t and $t + \Delta t$.

Analogously to the previous covariable studies, we conducted a simulation study to assess age-related change in SBC for a male and female subjects. In both genders, the observed age is 45 year, and the corresponding height, weight and waist circumference are those used in the previous sections. After fitting these dynamic Bayesian models with the Markov chain Monte Carlo (MCMC) algorithm, we obtained samples from the posterior distribution of model parameters, reflecting the sources of uncertainty, which are in turn used to get the posterior distribution of SBC. The posterior prediction for the five SBC are provided in Figure 3.21 - **??**. The black solid line represents mean of SBC during aging, the red and blue dashed line represent 95% and 50% credible intervals, respectively. Generally speaking, SBC vary more in the male subjects than in the female peers during aging. For the male subject concerned, the apL declines from about 30 kg at 45 year to less than 25 kg at 80 year, approximately 16% loss of apL, whereas the percent of loss is just 13% in the female subject. With respect to trunk lean mass, a loss of 3 kg is found in the male subject, while almost 1 kg in the female peer. Furthermore, body lean mass decreases from 66 kg at 45 year to about 55 kg at 80 year in the male subject, which implies 16% loss. However there is just 10% loss of bL in the female subject.

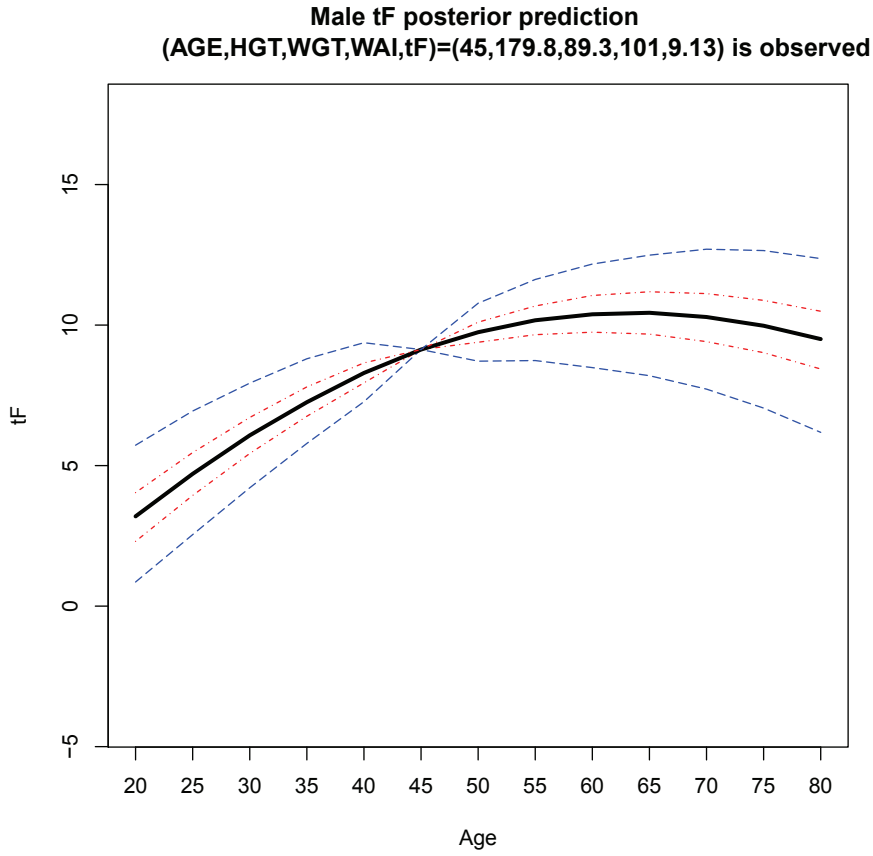
It might be less interesting to underline the predictions downstream before 45 year, the observed age of subjects. Nevertheless, it is still interesting to follow the change patterns in fat masses, such as trunk and body fat. Fat masses progress consistently until 60-65 year, then decline to the level at 45-50 year. From 20 year to 45 year, the male subject almost double his body fat and triple his trunk fat mass. In addition, fat masses arrive their maximum later than do body weight (Figure 3.18), thus this implies that even though an individual keeps weight stable in the middle age, there is still an accumulation of fat masses.

3.3.1.5 Conclusion

In summary, we propose two Bayesian modeling methods for age-related change study. The main advantage of these methods is to allow conducting a longitudinal analysis from the cross-sectional datasets. Otherwise, the Bayesian modeling enables to provide a prediction distribution, rather than a simple value, this is more relevant for exploring the uncertainty or accuracy problems. Also we can incorporate the previous findings in the prior distribution, by combining it with the datasets, we could obtain more suitable conclusions.

Figure 3.21: Age-related change in trunk fat (tF) for a male and female subjects, respectively.

For both genders, the observed age is 45 year; the corresponding height, weight, waist circumference and tF are 179.8 cm, 89.3 kg 101 cm and 9.13 kg for the male subject, 163.6 cm, 60.9 kg 75 cm and 7.72 kg for the female subject. The black solid line represents mean of tF during aging, the red and blue dashed line represent 95% and 50% credible intervals, respectively.



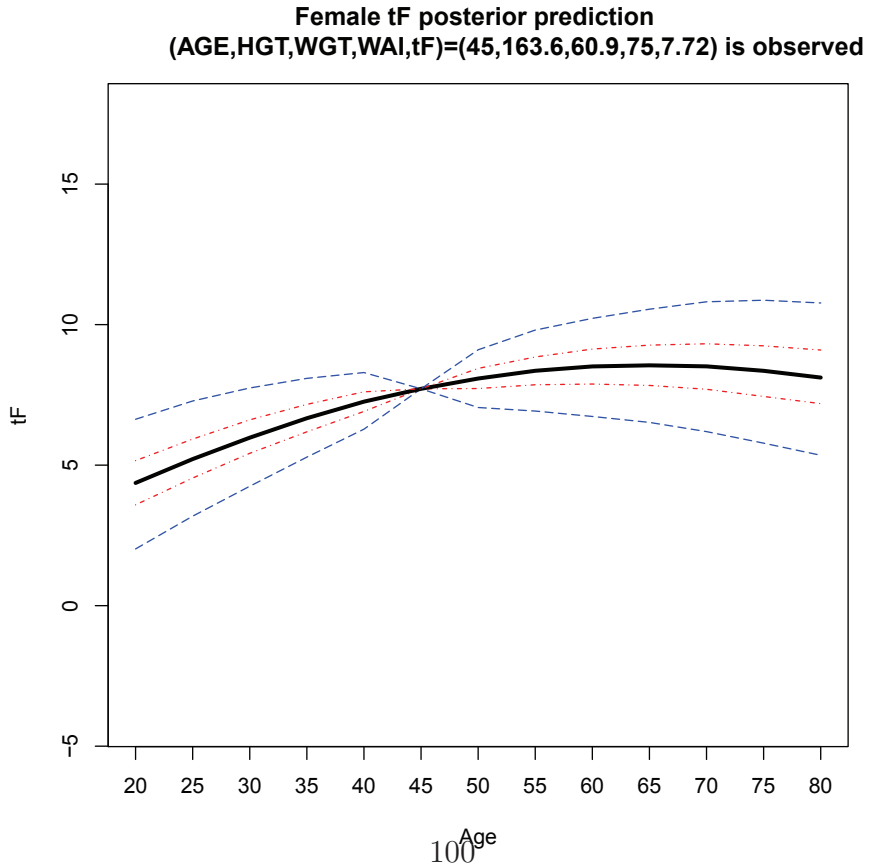


Figure 3.22: Age-related change in body fat (bF) an trunk lean (tL) for a male and female subjects, respectively. c.f. Figure 3.21.

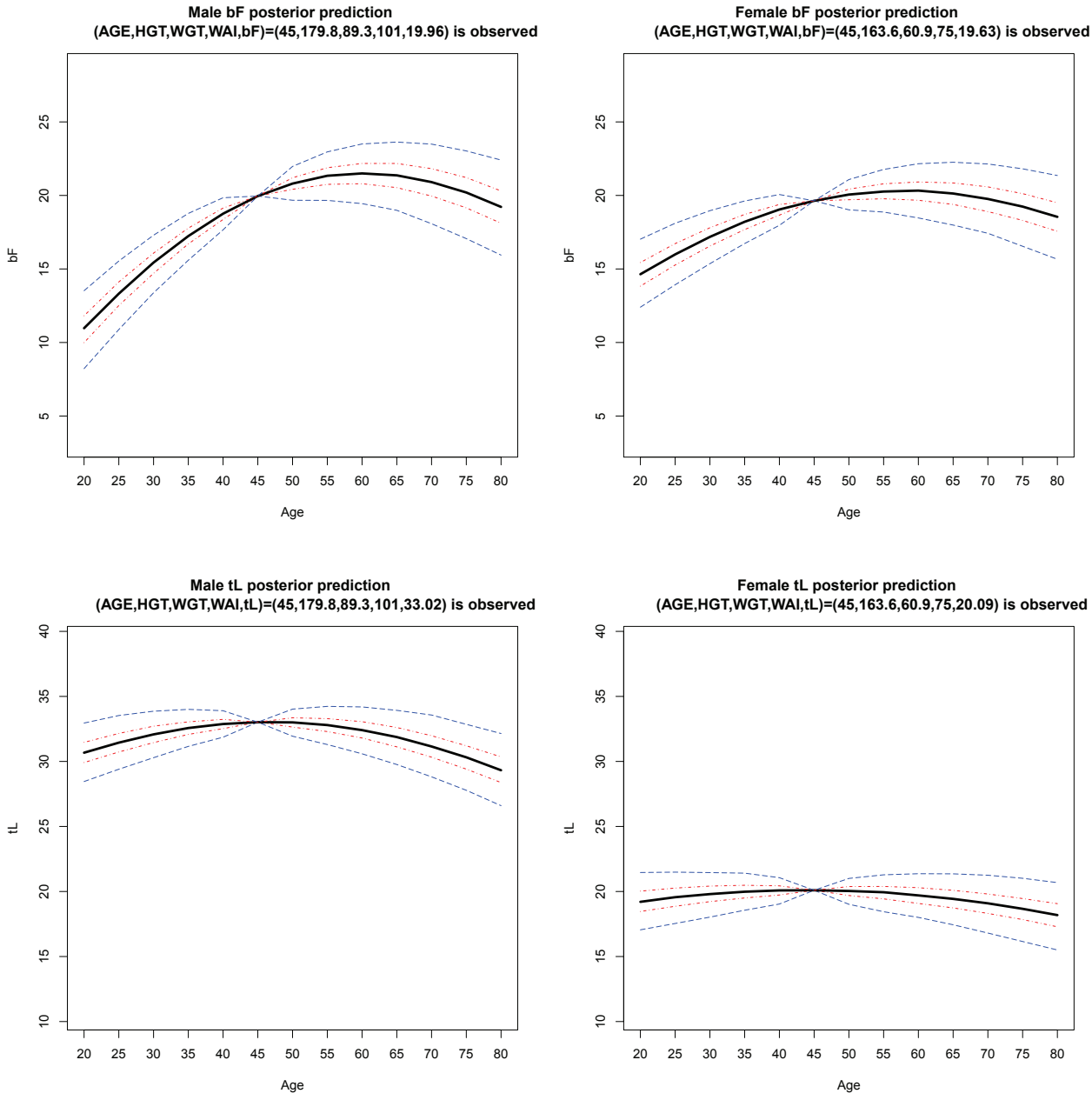


Figure 3.23: Age-related change in appendicular lean (apL) and body lean (bL) for a male and female subjects, respectively. c.f. Figure 3.21.

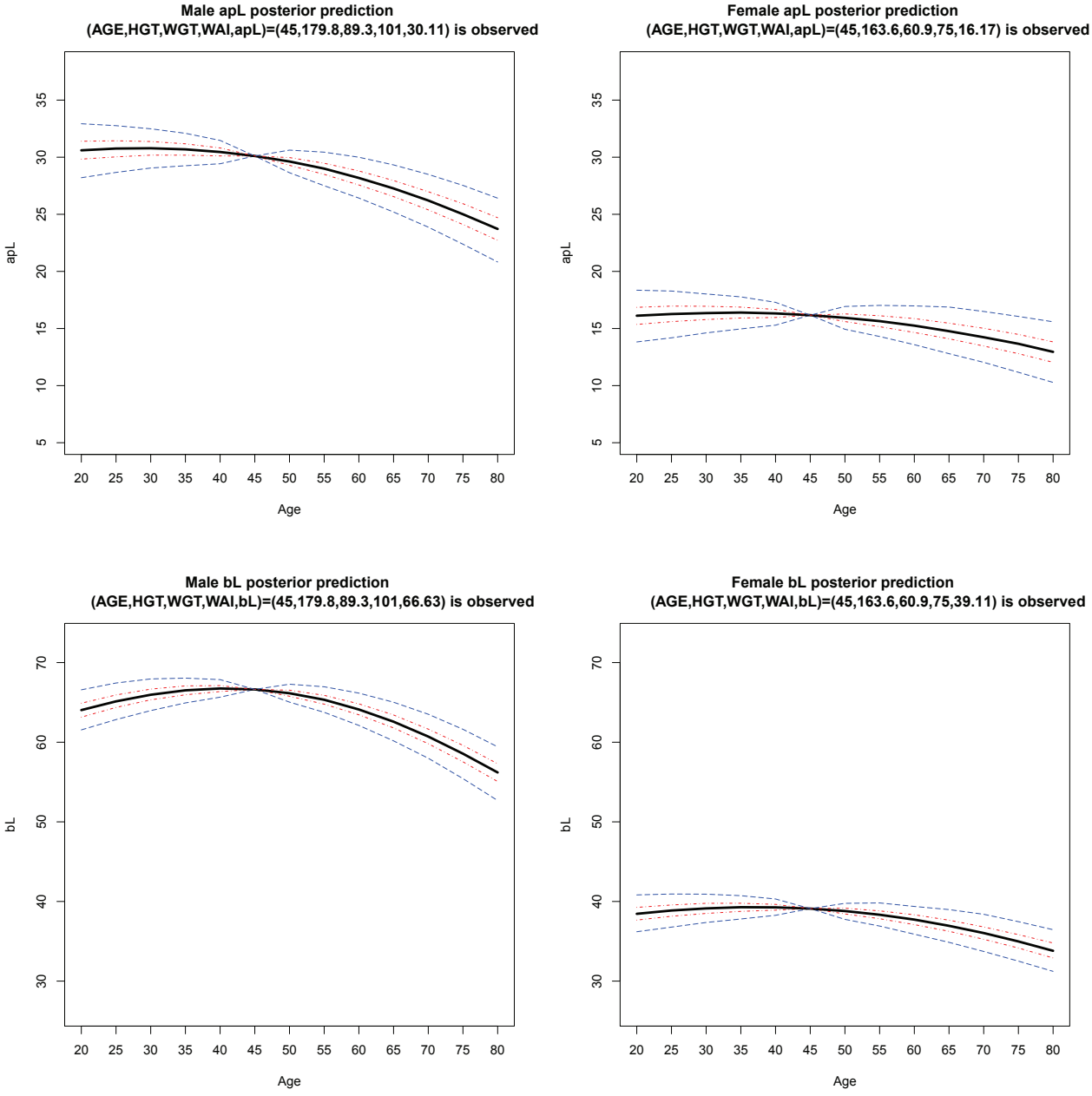


Table 3.10: In each age interval, age, anthropometric variables and adjusted height characteristics for men and women in the medical examination center at Saint-Brieuc, France (Mean values and standard deviations). HGT.adj is the adjusted height and it is calculated by equation (3.11). The age intervals $[a, b)$ mean $\{x \in R | a \leq x < b\}$.

		Men					Women				
	Age	n	HGT	HGT.adj	WGT	WAI	n	HGT	HGT.adj	WGT	WAI
1	[20,21)	222	176.59±6.44	176.59±6.44	70.08±12.62	79.26±9.45	418	163.45±6.35	163.45±6.35	63.46±13.09	75.61±10.04
2	[21,22)	237	177.38±7.09	177.48±7.09	72.05±12.72	80.16±9.24	349	163.33±5.52	163.43±5.52	62.88±12.31	75.36±10.49
3	[22,23)	226	176.56±6.73	176.76±6.73	71.53±12.4	80.99±9.44	357	163.98±6.51	164.18±6.51	64.68±12.8	76.5±10.01
4	[23,24)	219	176.64±6.91	176.94±6.91	70.97±10.91	80.69±8.63	288	163.51±6.61	163.81±6.61	64.05±12.56	76.32±10.56
5	[24,25)	210	176.75±6.96	177.15±6.96	72.59±12.13	81.17±8.76	242	163.76±6.54	164.16±6.54	64.06±13.31	76.45±11.45
6	[25,26)	190	175.54±7.18	176.04±7.18	70.55±12.37	80.75±8.38	239	163.63±6.36	164.13±6.36	63.35±12.36	75.75±10.89
7	[26,27)	153	177.68±6.65	178.28±6.65	75.37±12.85	83.54±10.13	167	163.97±6.77	164.57±6.77	61.83±9.77	74.38±8.3
8	[27,28)	154	177.27±6.84	177.97±6.84	74.45±14.71	82.91±10.34	159	164.13±6.73	164.83±6.73	61.17±10.02	74.22±8.48
9	[28,29)	140	176.8±6.49	177.6±6.49	73±12.07	82.34±9.69	147	163.01±5.89	163.81±5.89	64.68±12.85	77.32±11.15
10	[29,30)	131	175.83±6.43	176.73±6.43	74.28±12.71	83.96±9.99	137	164.31±6.25	165.21±6.25	65.05±12.29	77.27±9.84
11	[30,31)	128	176.23±7.19	177.23±7.19	75.02±13.36	84.07±9.68	145	164.23±6.09	165.23±6.09	64.69±11.02	77.45±9.45
12	[31,32)	124	175.98±7.41	177.08±7.41	75.67±12.69	84.57±9.68	131	163.42±5.98	164.52±5.98	63.85±11.98	77.15±10.6
13	[32,33)	152	176.27±7.06	177.47±7.06	75.19±12.1	84.57±9.57	149	162.62±6.67	163.82±6.67	65.13±13.02	78.97±11.56
14	[33,34)	161	176.03±6.18	177.33±6.18	74.87±12.47	85.05±10	164	162.95±6.28	164.25±6.28	63.33±12.7	76.73±11.44
15	[34,35)	143	175.73±6.31	177.13±6.31	75.55±11.26	85.55±9.4	175	163.31±6.52	164.71±6.52	65.68±12.7	78.61±10.3
16	[35,36)	173	175.9±6.89	177.4±6.89	76.03±12.82	85.74±10.06	170	162.9±5.72	164.4±5.72	63.14±11.43	77.4±9.68
17	[36,37)	174	176.18±6.88	177.78±6.88	75.64±12.03	85.66±9.08	161	162.66±5.87	164.26±5.87	63.58±11.24	77.15±10.03
18	[37,38)	144	174.89±6.54	176.59±6.54	75.43±13.28	86.08±10.88	178	163.01±6.25	164.71±6.25	65.58±12.52	79.33±10.53
19	[38,39)	172	175.41±6.48	177.21±6.48	76.49±13.26	87.5±10.77	166	163.86±6.01	165.66±6.01	65.34±11.54	79.59±10.83
20	[39,40)	187	175.82±7.16	177.72±7.16	78.61±12.74	89.22±10.44	162	162.8±6.23	164.7±6.23	62.86±11.26	77.43±10.74
21	[40,41)	197	175.64±6.54	177.64±6.54	77.01±12.87	87.78±10.02	189	162.31±6.27	164.31±6.27	64.82±11.57	79.18±11.43
22	[41,42)	194	175.85±7.25	177.95±7.25	77.49±12.55	88.31±9.99	185	162.7±6.35	164.8±6.35	65.24±11.84	79.68±11.22
23	[42,43)	208	175.38±6.87	177.58±6.87	77.38±12.27	88.83±10.11	212	162.67±5.64	164.87±5.64	64.5±11.7	78.76±10.62
24	[43,44)	228	175.68±6.38	177.98±6.38	77.68±12.61	89.65±10.42	198	161.66±5.53	163.96±5.53	63.73±11.58	79.38±10.63
25	[44,45)	190	174.75±6.35	177.15±6.35	77.79±12.24	90.16±9.62	214	162.62±6.11	165.02±6.11	65.31±11.76	80.05±11.8
26	[45,46)	227	175.4±7.01	177.9±7.01	77.07±12.49	89.04±10.07	229	162.04±6.17	164.54±6.17	64.28±10.34	79.87±10.25
27	[46,47)	240	174.33±6.92	176.93±6.92	77.3±12.18	89.87±9.86	232	161.79±6.3	164.39±6.3	64.24±12.49	80.16±12.51
28	[47,48)	269	174.17±6.83	176.87±6.83	76.51±12.2	89.75±9.83	276	161.37±6.7	164.07±6.7	64.03±11.04	80.54±11.55
29	[48,49)	271	174.59±6.8	177.39±6.8	79.1±12.85	91.98±10.54	226	160.98±5.61	163.78±5.61	62.35±9.7	79.13±10.09
30	[49,50)	255	174.38±6.8	177.28±6.8	77.72±13.65	91.14±11.15	230	161.3±6.01	164.2±6.01	64.31±10.44	80.63±11
31	[50,51)	273	172.88±6.82	175.88±6.82	75.37±12.1	89.92±10.08	259	161.32±5.78	164.32±5.78	64.2±11.08	80.28±11.33
32	[51,52)	270	173.7±6.76	176.8±6.76	78.02±12.41	91.93±10.26	273	160.25±5.66	163.35±5.66	63.63±11.84	80.75±11.74
33	[52,53)	283	173.42±6.82	176.62±6.82	78.02±13.05	92.17±10.73	249	160.52±6.2	163.72±6.2	64.39±11.79	81±11.32
34	[53,54)	171	172.96±6.75	176.26±6.75	77.7±14.26	92.41±11.74	117	159.81±5.51	163.11±5.51	65.09±11.5	82.56±12.58
35	[54,55)	247	173.56±6.09	176.96±6.09	77.75±12.13	91.96±10.72	288	159.7±5.56	163.1±5.56	63.6±11.27	80.91±11
36	[55,56)	271	173.49±6.21	176.99±6.21	77.6±12.55	91.5±10.87	301	159.88±6	163.38±6	64.55±11.85	81.86±11.79
37	[56,57)	269	173.1±6.25	176.7±6.25	79.29±12.57	93.49±10.72	297	159.51±5.72	163.11±5.72	63.93±11.45	81.41±11.79
38	[57,58)	275	172.74±6.17	176.44±6.17	79.51±12.66	93.86±10.68	295	159.48±5.76	163.18±5.76	63.91±11.41	81.39±11.55
39	[58,59)	287	172.89±6.36	176.69±6.36	78.16±11.21	93.17±9.93	298	159.32±5.68	163.12±5.68	63.83±10.83	81.71±11.48
40	[59,60)	263	172.62±5.92	176.52±5.92	79.29±11.41	94.09±10.51	327	158.8±6.25	162.7±6.25	62.72±10.9	81.61±10.97
41	[60,61)	398	172.65±6.37	176.65±6.37	78.83±11.9	93.57±10.56	377	159.6±5.88	163.6±5.88	65.01±10.56	83.11±10.93
42	[61,62)	275	172.38±6.13	176.48±6.13	77.47±10.85	92.9±9.66	269	159.34±5.37	163.44±5.37	64.37±10.96	82.71±11.08
43	[62,63)	184	171.86±6.4	176.06±6.4	77.82±10.31	93.82±9.64	235	159.27±5.81	163.47±5.81	63.97±10.83	82.85±11.19
44	[63,64)	286	171.17±6.23	175.47±6.23	78.69±12.05	94.56±10.64	188	159.04±5.23	163.34±5.23	65.62±10.82	84.98±10.99
45	[64,65)	395	171.25±6.43	175.65±6.43	79.59±11.62	96.15±10.43	209	159.03±5.27	163.43±5.27	65.39±11.56	83.91±12.23
46	[65,66)	170	170.11±5.89	174.61±5.89	77.29±11.78	94.64±9.92	177	158.34±5.07	162.84±5.07	63.85±10.58	83.4±11.64
47	[66,67)	115	170.45±5.11	175.05±5.11	75.48±9.84	92.23±9.53	145	158.23±5.48	162.83±5.48	64.41±11.03	84.37±10.99
48	[67,68)	157	169.48±5.88	174.18±5.88	77.34±12.33	95.15±11.26	164	156.55±5.56	161.25±5.56	62.59±9.93	83.07±10.95
49	[68,69)	157	170.1±5.89	174.9±5.89	76.35±10.57	94.06±9.43	170	156.99±5.18	161.79±5.18	64.37±10.51	84.72±11.36
50	[69,70)	148	169.74±5.9	174.64±5.9	77.76±12.8	96.16±10.94	155	157.47±5.42	162.37±5.42	63.23±10.24	83.65±10.53
51	[70,72)	317	169.75±5.93	174.8±5.93	76.59±10.86	95.09±10.49	296	156.15±5.63	161.2±5.64	63.07±10.23	84.24±10.41
52	[72,74)	208	168.88±6.13	174.13±6.12	76.01±10.76	95.38±9.84	221	156.18±5.52	161.43±5.51	63.14±10.69	84.1±11.1
53	[74,76)	174	169.12±5.64	174.55±5.64	76.52±10.5	96.36±10.16	173	155.38±6.15	160.82±6.15	63.01±10.7	85.69±10.33
54	[76,100)	188	166.95±5.92	172.84±5.87	74.83±11.25	96.59±10.74	184	153.79±6.08	159.71±6.02	60.83±10.56	85.81±11.38

3.3.2 Frequentist modeling for age-related changes in body composition (Paper submitted to *British Journal of Nutrition*)

Title: Age-related changes in segmental body composition according to ethnicity and history of weight change during life

Author: Simiao Tian[1,2,3], Béatrice Morio[2,3], Jean-Baptiste Denis[1], Laurence Mioche[3]

[1] INRA, Unité de Recherche MIA, F-78352 Jouy-en-Josas

[2] Université d'Auvergne, Unité de Nutrition Humaine, BP 10448, F- 63000 Clermont-Ferrand

[3] INRA, UMR 1019, UNH, F- 63000 Clermont-Ferrand

Authors' contact: Beatrice.Morio@clermont.inra.fr;

Address correspondence to Simiao Tian, INRA, Unité de Recherche MIA, 78352 Jouy-en-Josas, France. Telephone: +33 (0)1 34 65 22 36. E-mail: Simiao.Tian@gmail.com

Last names of authors for PubMed: Tian, Mioche, Denis, Morio

Running head: body composition changes during aging

Key words: Multivariate model, aging, body composition

Abstract:

This study aimed at assessing age-related changes in body composition and more specifically in trunk fat and appendicular lean masses, according to BMI at the age of 20 y (BMIref), ethnicity and history of weight change during life. A cross-sectional DXA dataset was extracted from National Health and Nutrition Examination Survey 1999-2004, and only European-American and African-American subjects were retained (2705 men, 2527 women in total). For each gender and ethnicity, 6 different study cases were considered based on three BMIref categories (normal, overweight and obese: BMIref=22, 27 and 30 kg/m^2, respectively) and two weight contexts across ages: 1) stable weight (RP) and 2) weight gain (GP) across lifespan. A nonparametric modeling was first built to study age-related changes in body composition. Then a parametric modeling was developed for assessing BMIref- and ethnicity-specific effect during aging. For RP context in both gender and ethnicities, trunk fat (TF) increased gradually; body fat (BF) remained stable until 40 y and increased thereafter; trunk lean (TL) remained stable but appendicular lean (APL) and body lean (BL) declined from 20 y. For GP context, TF and BF progressed with likely constant rate; APL, TL and BL increased until 40-50 y, then decline slightly. Compared with European-American subjects in both gender, African-American subjects had lower trunk fat and body fat masses. Moreover, ethnic differences in body composition were constant across life. To conclude, for our specified study cases, ethnicity-related difference was found in body composition, but the magnitude of difference was consistent in aging.

Introduction

Aging is associated with substantial changes in body composition. Reduction in body lean (BL) or body fat-free (BFF) mass occurs during aging[1] together with an increase of body fat (BF) related to accumulation of adipose tissue, particularly in the abdominal region[2]. These changes are closely linked with muscle strength reduction during aging[3]. The loss of muscle mass, known as sarcopenia, may have a negative impact on physical function, and lead to functional impairment and disability[4-6]. Meanwhile, accumulation of body fat may associate with a number of metabolic risk factors and lead to an increased prevalence of chronic metabolic diseases[7]. Many studies have shown that increased abdominal fat mass is an independent risk factor for hypertension, stroke, and type 2 diabetes[8-10]. Other reports suggest that upper body fat (truncal fat)[11-12] has been strongly associated with insulin resistance, metabolic risk factors, and their disease outcomes.

Although body composition, as well as its age-related change, has a strong genetic component[13], it is likely influenced by external factors such as social environment and physical activities[14]. Assessing these changes in segmental body composition with aging may be important because the study will lead to a pre-diagnosis on prevention of morbidity and mortality risk[15].

Most studies on age-related changes in body composition were derived from cross-sectional dataset[16-17]. One weakness of such studies is not to take into account the possible birth cohort effect[18-20]. In Ding *et al.*'s longitudinal study of aged 70-79 subjects using DXA[18], they reported that (1) at the same age, later birth cohorts had a greater body fat and lean masses than did earlier cohorts in both genders; (2) within each cohort, BF initially increased with age and decreased after 80 year, while BL decreased with age, additionally, the decrease was more rapid in men than in women; (3) the amount of BF was much less than that of BL, nevertheless the increase in BF was greater than that in BL, which led to an increase of BF%. Recently, Mioche *et al*[21] proposed a nonparametric modeling originally to predict segmental body composition from easily acquired covariates. Furthermore, they validated their nonparametric model in an original comparison of various body composition studies to highlight the respective influence of other variables, such as ethnicity and method for BFF assessment[22]. After all, their proposed statistical methodology could be readjusted to overcome drawbacks of cross-sectional dataset for age-related study[21].

In the present study, we were interested in the age-related changes in segmental body compositions in different ethnic and BMI context. By using a nonparametric modeling, we conducted a longitudinal analysis from a cross sectional dataset. The aims of our study were (1) to

appraise the mean age-related changes in segmental body composition (SBC) for different study cases; (2) to develop a parametric modeling from nonparametric models for a smoother graphical presentation and easy interpretation; and (3) to assess BMI-, ethnicity-related difference in body composition changes with aging.

Subjects and methods

Samples

Samples for this study were extracted from the NHANES website within the 1999-2004 period dataset (http://www.cdc.gov/nchs/about/major/nhanes/). Subjects were characterized by covariates, such as gender, ethnicity, age, height, weight and waist circumference. A preliminary study related to anthropometric values across the lifespan was conducted among Hispanic-American (HA), European-American (EA) and African-American (AA) subjects. It turned out that in the same BMI level and age interval, HA subjects had different height, weight and waist circumference compared to EA and AA peers (Table not shown). As a consequence for the present study, we only retained EA and AA subjects aged 20-85 years, with BMI ranging from 18 to 40 kg/m^2, This selection resulted in a sample size of 2705 men (1984 EA men and 721 AA men) and 2527 women (1830 EA women and 697 AA women). Height, weight and waist circumference were particularly considered as similarity criterion among different subjects. Height was supposed constant during the whole lifespan; weight and waist circumference had two different change contexts (detail below). The study was separately conducted on men and women.

Segmental body composition

Whole-body and segmental body compositions were assessed using dual-energy X-ray absorptiometry (DXA) (Hologic QDR 4500A fanbeam densitometer for NHANES). For the NHANES dataset, detailed descriptions have been published elsewhere[23]. Briefly, whole-body DXA scans were administered in the NHANES mobile examination center to eligible participants during the 6-y period from 1999 to 2004; the participants with certain physical conditions were excluded from the DXA examination[24]. The DXA scans permit quantification of multiple whole body and regional components, including bone mineral content, fat, and lean soft tissue. Body fat (BF) and lean (BL) masses, trunk fat (TF) and lean (TL) masses were thus determined[25]. The appendicular lean mass (APL) was the sum of arms and legs lean masses[26]. In the present study, we are interested in the previously mentioned five segmental body compositions, as they are significant in health assessment.

Study cases

For each gender, we considered two ethnicity categories (EA and EE), two weight contexts: reference profile (RP) and gain profile (GP) to mimic two life situations and three BMI categories (normal weight, overweight and obese categories), a total of 12 study cases.

For each BMIref, height, weight and waist circumference profile at 20 year were the same for the two ethnicities, and the starting values in men and women were given in **Table 1**. RP assumed that height, weight and waist circumference were constant when aging; GP assumed that weight remained constant until 40 y, then increased by 5%/decade to 60 y and stabilized thereafter, and that waist circumference remained constant until 40 y, then increased by 3cm/decade to 60 y and stabilized thereafter[27-29].

Age

In the present study, we conducted both of a nonparametric and parametric modeling (described below). In the nonparametric modeling, age was converted into a categorical covariate and categorized into six intervals: 20-29, 30-39, 40-49, 50-59, 60-69, and ≥70 y old. A preliminary study showed that for each ethnicity, this categorization ensured adequate subset sizes for the nonparametric modeling. In contrast in the parametric modeling, age was considered as a continue covariate.

Statistical modeling and analysis

Nonparametric modeling. A nonparametric approach was first used to assess segmental body composition changes in aging. This nonparametric approach followed the idea of Mioche *et al.* 2011a. With respect to each study case, the nonparametric modeling follows five steps : (1) individual height, weight and waist circumference (stature value) at 20 y is specified, (2) then change in weight and waist circumference profile (either RP or GP) was applied to obtain stature changes for all age interval after 20 year, (3) for each age, based on individual stature value and ethnicity, a dispersion tolerance is defined to select the candidates of the same ethnic category with similar stature values in the NHANES dataset[17], (4) the prediction of SBC was calculated by average value in the selected subset, (5) the predictions at each age interval were connected linearly to construct a body composition trends in aging. It is worth noting that the nonparametric approach allowed to generally emphasizing how SBC change in aging for each study case.

Parametric modeling. In purpose to smooth graphical representation, several parametric modeling were proposed. First of all, BMI category effect was considered, and this effect was combined with age effect to model body composition changes. To study this effect, the proposed parametric models covered from a simplest one $M0_B(A)$ to the most complicated $M3_B(A)$ (**Table 2**). Then the same methodology was conducted to assess ethnicity effect on age-related changes in body composition.

For a given stature trends, the parametric modeling follows four steps: (1) all subsets of candidates associated to the age classes using in nonparametric approach were combined, (2) the obtained dataset was then used to fit parametric models, (3) for BMI category and ethnicity effect, several proposed parametric models were fitted and a standard error of the estimate (SEE) was calculated for each model as follows:

$$SEE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p}}$$

where n sample size and p number of parameters in the model, (4) the model selection was done by considering the trade-off between SEE and the number of parameters in the models.

Statistical calculations and analyses were performed using version 2.12.2 of the R software[30], a language and environment for statistical computing.

Results

Sample characteristics

Means and standard deviation for the five considered segmental body compositions are provided for all study cases in **Table 3** and **4** for men and women. Generally for each gender and ethnicity studied, segmental and body fat masses increased until 60 y, and declined afterwards, whereas lean masses are likely stable until 60 y, then decreased gradually.

More precisely, for EA men, TF and BF increased gradually until 60-69 y, then they declined, while TF and BF progressed consistently for AA peers. With respect to lean masses for EA men, APL, TL and BL increased to 40-49 y, then they decreased. Nevertheless, for AA men, APL, TL and BL stabilized until 60-69 y, and declined afterwards. Similar results were found in EA and AA women, except that for AA women, TF and BF increased until 40-49 y, then stabilized to 60-69 y, and declined afterwards.

Model selection

For BMI category effect, the SEE values of nonparametric and parametric models are shown in **Table 5**. For the two weight trends contexts in white men and women, $M3_B(A)$ yielded the fewest SEE values, but had the greatest number of parameters (not shown). However, in comparison with nonparametric models and other parametric models, $M1_B(A)$ enabled similar SEE values with $M3_B(A)$, and especially required fewer parameters than $M3_B(A)$. In addition, ANOVA test showed that $M1_B(A)$ were significantly different than $M0_B(A)$, this underlined an additive model with the covariates BMI and age. $M1_B(A)$ indicated that there was a BMI-related difference in SBC at each age class, but all of BMI categories shared the same trends in aging. Thus $M1_B(A)$ was retained to model age-related changes in body composition for different BMI categories.

On the basis of $M1_B(A)$, ethnicity effect was then studied following the same methodology. Proposed parametric models were described in Table 2. The SEE values of nonparametric and parametric models were showed in **Table 6** for men and women. Nonparametric model and $M14_{B,E}(A)$ yielded the fewest SEE values, but the difference was not important in comparison with $M11_{B,E}(A)$. Also the statistical test ANOVA showed $M11_{B,E}(A)$ was significantly different than $M10_{B,E}(A)$; therefore $M11_{B,E}(A)$ was retained to study ethnicity-, BMI- and age-related changes in body composition.

SBC trends in aging

For the sake of simplicity, only the age-related trend curves for EA normal weight subjects (reference curve) are drawn. Based on the additivity of the retained model $M11_{B,E}(A)$, the curves corresponding to other BMI- and ethnicity-specific subjects are simply vertical translation of the reference curve.

The smooth curves are shown in **Figure 1** and **Figure 2** respectively for EA normal men and women. It is worth noticing that the starting value at 20 y in RP context was higher than that in GP context. Indeed, this is due to the leverage effect associated with lighter subject's weight in other age intervals for RP context. More precisely, given the same subset of subjects at 20-29 y between RP and GP contexts, the parametric model has to compromise the overall fit by heightening at left extremity in RP context. However in the nonparametric modeling framework, the starting value at 20 y was the same for both RP and GP context (not shown).

In EA normal men, for RP context, TF increased consistently with aging from 20 y, whereas BF stabilized to the age of 40 y and increased thereafter. Regarding lean masses, TL was more stable than APL and BL, nevertheless APL and BL declined from 20 y. For GP context, TF and BF increase likely with constant growth rate. APL progressed and reached its maximum value at age of 40-50 y, then decline slightly. Moreover, TL and BL increased until 50 y, then TL stabilized, whereas BL declined afterwards.

In EA normal women, for RP context, similar results were found as in men. For GP context, TF and BF increased from 20 y, and the growth rate was likely linear, whereas APL was almost stable with aging. TL and BL increased until 50 y, then TL stabilized and BL declined slightly thereafter.

Profile contexts with BMI and ethnicity effect

Reference profile. As mentioned above, $M11_{B,E}(A)$ was retained to study BMI-, ethnicity- and age-related changes in body composition. Since the retained model was an additive model, the different BMI and ethnicity categories shared the same trends in body composition across age intervals, but with vertical linear translations. For the present study, the baseline for BMI and ethnicity category was respectively BMI=Normal and Ethnicity=EA. **Table 7** summarized the parameter differences between other BMI and ethnicity categories and their baseline category.

In men, for ethnicity effect, AA men had lower trunk fat, body fat and trunk lean masses than EA peers (the differences were -1.48, -1.31 and -0.83 kg, respectively), had greater appendicular and body lean masses (the differences were 2.19 and 1.52 kg, respectively). For BMI effect, overweight and obese men had always greater segmental body compositions than normal weight men. The difference varied from 3.53 to 7.36 kg respectively for APL and BL between overweight and normal category, whereas from 5.49 to 11.77 kg between obese and normal category.

Similar results were found in women. AA women had -0.97, -0.58 and -0.6 kg lower of TF, BF and TL than EA peers, while 1.9 and 1.55 kg greater of APL and BL. Within BMI categories, overweight women had 4.44 and 8.12 kg higher in BL and BF than normal weight women, and these differences widened to 8.09 and 14.17 kg for obese women.

With comparison with men for ethnicity effect, the absolute value of parameter differences in women was also greater in body and segmental fat masses, and lower in lean masses. With respect to BMI effect, the parameter differences in women was higher than in men for appendicular and body fat masses, while lower for appendicular, trunk and body lean masses.

Gain profile. The same modeling was conducted for weight gain profile study case with retained $M11_{B,E}(A)$. Because of the property of additive models, the different BMI and ethnicity categories followed the similar age-related trends in body composition. **Table 7** showed parameter differences between other BMI and ethnicity categories and their baseline.

In both men and women, for BMI and ethnicity effect, the similar results were observed as in the RP context: (1) AA subjects had lower TF, BF and TL than EA peers, while had greater APL and BL masses; (2) for BMI effect, overweight and obese subjects had always higher segmental body composition than normal weight subjects. Nevertheless, compared with the RP context, the parameter differences were greater in GP context.

Discussion

In the present study, we assessed the age-related changes in body composition following different BMI, ethnicities and weight changes. Although there are several studies that address similar topics with mainly the cross sectional datasets, to our knowledge, our proposed modeling is the first to conducted a predictive longitudinal statistical analysis by using a cross sectional dataset. In fact, one weakness of the cross sectional datasets is the lack of follow-up information, whereas the longitudinal studies are not cost efficiencies, and sometimes not practical for long adulthood period. Our proposed methodology enables to overcome the previous weakness of cross sectional dataset. Indeed, the key idea of our proposed methodology is to select a subset of candidates of the same gender and ethnicity, who have similar anthropometric values at a certain age interval. As the selected subset is representative, average value can be considered as the body composition prediction for this age interval. Also it is worth noting that, rather than focusing on the prediction accuracy, the aim was to assess age-related changes in body composition, and to understand how ethnicity influences these changes.

Age-related trends in segmental body composition

The age-related changes in segmental body compositions are affected by a variety of factors, such as physical activity, menopausal status, nutrition and disease[31]. Understanding these associated factors will help to assist in the prevention of functional limitation. Also establishing scope of the age-related changes will be profitable in the management of health status into old age. In theory, the longitudinal studies are more reliable than cross-sectional studies in the aging framework, but also with presence of some drawbacks[32]. Recently, Ding *et al.*[18] used another methodological approach which integrated cross-sectional and longitudinal study, and they found that, at the same age, youngest cohorts had greater body fat mass than oldest cohorts.

In the present study, we conducted a predictive longitudinal statistical analysis by using a cross sectional dataset. Primarily, we specified three BMI categories at 20 y (BMIref) with proper initial stature values, and two representative weight trend contexts (RP and GP) in aging within EA and AA subjects. Then we studied a nonparametric modeling to predict body composition changes in these various contexts and a parametric modeling for a better graphical representation and easy interpretation. In the RP context, total and segmental lean masses declined from the age of 20 y, while body and trunk fat masses increased consistently. Following a GP context, total and

segmental fat masses increased in aging, whereas lean masses increased to the age of 50 y, and decreased slightly thereafter. Our previous findings confirm similar results in the literature.

The topic of body composition in aging was widely discussed. In an observational study on healthy subjects aged from birth to 80 y, Henche *et al.*[33] reported the evolution of total and regional fat masses and percentage of certain lean masses. They showed that in both genders, BF increased until 70 year then slightly declined. With respect to regional compartments, TF increased until 55 year of age in men, while to 70 year of age in women, and stabilized therefore in both gender, respectively. In another observational study in women aged 18-94 using DXA, carried on by Welch et Sowers[34], BF increased gradually with increasing age until 56 years, then decreased afterward. Nevertheless in our study, for both RP and GP contexts, we found that BF and TF increased consistently from 20 to 70 years. These dissimilar findings may be the result of a decreasing weight after 40-60 year in Henche *et al.*[33] and Welch et Sowers[34], whereas we supposed either a constant or an increasing weight.

In a prediction study by Chumlea et al.[35], the estimated body fat-free (BFF) masses progressed until age 60 and 45 year respectively for men and women, after which the estimated BFF declined. Also based on a dataset of men aged 35-81 y using DXA, Atlantis et al.[36] showed that compared with the baseline age group values 35-44 y, BL decreases with age. Also in the study of Welch and Sowers[34], BL stabilized until 57 year, then decreased with aging. In our study, the prediction for women confirmed Welch and Sowers's finding about age-related changes in BL, however the age at which time BL declined was younger, about 50 y for GP context and 20 y for RP context. In addition in the GP context, APL increased until 40-50 year in men, and decreased slightly, while it was more likely stable in women. With respect to TL, it progressed consistently in both genders.

The magnitude of BFF change in aging was also studied by using longitudinal small datasets. The results showed that the age-related longitudinal changes in BFF were -1.2 and -0.9 kg/decade for men and -0.1 and -0.4 kg/decade for women, respectively in Hughes *et al.*[37] and Kyle et al.[38]. The differences in BFF change rate might result in sample characteristics. Indeed, Kyle et al.[38] studied a BIA-based dataset aged 20-73 y, whereas Hughes *et al.*[37] studied a hydrodensitometry-based dataset of elderly men and women (initial age of 60.7 $\pm$7.8 y). By approximating some non-linear curve to linear in RP context, we found a rate of -0.8 and -0.5 kg/decade in BL (BL has very close changes with BFF) for men and women, respectively. Also the results showed, in RP context, a rate of $\pm$0.8 and $\pm$0.5 kg/decade in TF and of -0.8 and -0.5 kg/decade in APL for men and women, respectively. Regarding GP context, the rate rose to 2 kg/decade in BF for both men and women. To

summarize, our results about BL change rates confirm Hughes *et al.*'s findings, that emphasizes a promising use of our proposed statistical methodology in the longitudinal analysis based on a cross sectional dataset. Moreover, other compartment change rates are provided, from a physiological point of view, these findings may be interesting when long-term longitudinal studies are lacking.

Ethnicity effect

Ethnic differences in body composition have been recognized in US[39]. The accumulation of fat masses, in particular trunk fat masses, is strongly related to age and ethnicity in both men and women[40]. Secondly, the ethnicity-related differences in body composition may occur primarily in early adulthood[41]. In the present study for our study cases, we found that (1) ethnicity-related differences occur in segmental body compositions with aging; (2) based on the retained additive model, these differences are constant within each age interval. Moreover, our study suggest that across age intervals, AA subjects had lower trunk fat, trunk lean and body fat masses than EA peers, and greater total and appendicular lean masses. Our findings are consistent with previous conclusion that ethnic differences in body composition exist during the overall lifespan, and these differences remain constant across age intervals. Indeed, Casas *et al.*[41] showed using DXA dataset among healthy Mexican-American (MA) and European-American (EA) women aged 20-75 y, that EA people may have modestly lower body and trunk fat masses and slightly higher amounts of body fat-free mass, trunk region in particular compared to MA. Moreover, their findings showed that the ethnicity-related differences in body composition resulted primarily of dissimilarities in the young adult to early middle-age. Thus, they suggested that ethnicity-related differences may occur in early adulthood.

Some early studies have shown that ethnicity is an important factor to explain the relationship between body fat and BMI[16]. In another study related to ethnicity effect, Fernandez *et al.*[42] reported that the prediction of BF percentage from BMI in MA women differs from that of EA and AA women, however no significant differences between EA and AA women or between any combination of these three ethnicity categories in men. Contrary to previous findings in BF percentage, we studied directly amount of BF in kilo and the present study demonstrates a significant difference in BF and other segmental body compositions between EA and AA subjects. More precisely, at the same BMI level, EA subjects had a lower body fat mass than EA peers within overall age intervals.

In addition, by conducting an elderly cohort aged 60-98 y with MA, EA and AA subjects, Aleman-Mateo *et al.*[43] showed that after controlling BMI and age, there was an ethnicity-related difference in body composition across ethnicity categories: AA subjects have significantly lower body and

trunk fat masses than EA peers, and greater total and appendicular lean masses. Similarly in the present study, our results support these previous findings with quantified difference values. More precisely, in the GP context, AA subjects had respectively about 2 and 1.5 kg greater APL and BL than EA peers, whereas 1 kg lower TF than EA peers.

Besides, in a study using BIA dataset and multicomponent model-derived prediction formulae, Chumlea et al.[35] found that the means for body fat-free mass within different ethnicity categories had similar patterns across age groups. The structure of our retained additive model agreed with this finding, and for AA category, its age-related trends in body composition can be translated vertically from EA's.

There are several limitations of this study. First of all, the weight trend contexts were based on the published findings; therefore, it may lack of precision. For a further study, an independent weight trends function could be developed to ensure the precision issue. Secondly, this study used a cross-sectional dataset for a longitudinal analysis. Nonparametric modeling enabled to extract a subset of similar subjects for a given age interval, however it cannot take into account the effect of birth cohort, particularly effect on height (previous generations are shorter than more recent generations, because height increases about 1cm/decades). Thus, it is more sensible to use an independent height function with aging. Moreover, the retained additive model was used to describe overall panoramic age-related changes rather than to accurately estimate body composition values. For further clinical use of this model, another validation study should be conducted.

In summary, we assessed age-related changes in segmental body composition associated with BMI and ethnicity effect. A nonparametric modeling was proposed to address a longitudinal analysis from a cross sectional dataset. Furthermore, we developed a parametric modeling to smooth graphical presentation of age-related changes in body composition. Similar to other studies, ethnical differences were found in body fat and lean masses, also in appendicular and trunk regions. We provided additional quantitative information for ethnic differences, and we also reported that these ethnic differences were constant across the age intervals.

Acknowledgments

All co-authors have seen and agreed with the contents of the manuscript and none of co-authors had any conflict of interest concerning the manuscript. The authors' responsibilities were as follows: ST: model computations, statistical analysis and the first draft of the manuscript; BM: design of the study and physiological interpretation; JBD: design of the study, model computations and statistical analysis; LM: data acquisition, design of the study and physiological interpretation.

References

1. Kyle U, Genton L, Hans D *et al.* (2001) Age-related differences in fat-free mass, skeletal muscle, body cell mass and fat mass between 18 and 94 years. *Eur J Clin Nutr* 55, 663-672.

2. Kuk J, Saunders T, Davidson L *et al.* (2009) Age-related changes in total and regional fat distribution. *Ageing RES REV* **8**, 339-348.

3. Baumgartner RN, Waters DL, Gallagher D *et al.* (1999) Predictors of skeletal muscle mass in elderly men and women. *Mech Ageing Dev* **107**, 123-136.

4. Baumgartner RN, Koehler KM, Gallagher D *et al.* (1998) Epidemiology of sarcopenia among the elderly in New Mexico. *Am J Epidemiol* **147**, 755-763.

5. Morley JE, Baumgartner RN, Roubenoff R *et al.* (2001) Sarcopenia. *J Lab and Clin Med* **137**, 231-243.

6. Roubenoff R (2001) Origins and clinical relevance of sarcopenia. *Can J Appl Physiol* **26**, 78-89.

7. Carr MC & Brunzell JD (2004) Abdominal obesity and dyslipidemia in the metabolic syndrome: importance of type 2 diabetes and familial combined hyperlipidemia in coronary artery disease risk. *J of Clin Endo & Meta* **89**, 2601-2607.

8. Larsson B, Svardsudd K, Welin L *et al.* (1984) Abdominal adipose tissue distribution, obesity, and risk of cardiovascular disease and death: 13 year follow up of participants in the study of men born in 1913. *Br Med J* **288**, 1401–1404.

9. Lapidus L, Bengtsson C, Larsson B *et al.* (1984) Distribution of adipose tissue and risk of cardiovascular disease and death: a 12 year follow up of participants in the population study of women in Gothenburg, Sweden. *Br Med J* **289**, 1257–1261.

10. Ducimetiere P, Richard J, Cambien, F (1986) The pattern of subcutaneous fat distribution in middle-aged men and the risk of coronary heart disease: the Paris Prospective Study. *Int J Obes* **10**, 229–240.

11. Kissebah AH, Vydelingum N, Murray R *et al.* (1982) Relation of body fat distribution to metabolic complications of obesity. *J Clin Endocrinol Metab* **54**, 254–260.

12. Abate N, Garg A, Peshock RM *et al.* (1995) Relationship of generalized and regional adiposity to insulin sensitivity in men. *J Clin Invest* **96**, 88–98.

13. Wulan SN, Westerterp KP, Plasqui G. (2010) Ethnic differences in body composition and the associated metabolic profile: a comparative study between Asians and Caucasians. *Maturiats* **65**, 315-319.

14. Franzini L, Elliott MN *et al.* (2009) Influences of physical and social neighborhood environments on children's physical activity and obesity. Am J of Public Health **99**, 271-278.

15. Visscher TLS & Seidell JC (2001) The public health impact of obesity. *Ann Rev Pub Health* **22**, 355-375.

16. Gallagher D, Heymsfield SB, Heo M *et al.* (2000) Healthy percentage body fat ranges: an approach for developing guidelines based on body mass index. *Am J Clin Nutr* **72**, 694-701.

17. Tian S, Mioche L, Denis J-B *et al.* (2013) A multivariate modeling for predicting segmental body composition. *Bri J Nutr* ISSN 1475-2662, 1-11.

18. Ding J, Kritchevsky S, Newman A *et al.* (2007). Effects of birth cohort and age on body composition in a sample of community-based elderly. *Am J Clin Nutr* **85**, 405-410.

19. Floud R (1994) Heights of Europeans since 1750: A New Source for European Economic History. Stature, living Standards, and Economic Development: Essays in Anthropometric History. Chicago, *Univ. of Chicago Press*, 9-24.

20. Steckel R (1995) Stature and the Standard of Living. *J Eco Literature* **33**, 1903-1940.

21. Mioche L, Bidot C, Denis JB (2011) Body composition predicted with a Bayesian network from simple variables. *Br J Nutr* **105**, 1265-1271.

538 22. Mioche L, Brigand A, Bidot C *et al.* (2011) Fat-Free Mass Predictions through a Bayesian
539 Network Enable Body Composition Comparisons in Various Populations. *J Nutr* **1411**, 573-
540 580.

541 23. Centers for Disease Control and Prevention. National Health and Nutrition Examination
542 Survey: body composition procedures manual. Available from:
543 http://www.cdc.gov/nchs/data/nhanes/BC.pdf.

544 24. Centers for Disease Control and Prevention. The 1999-2004 dual energy X-ray
545 absorptiometry (DXA) multiple imputation data files and technical documentation. Available
546 from: http://www.cdc.gov/nchs/about/major/nhanes/dxx/dxa.html.

547 25. Mazess RB, Barden HS, Bisek JP *et al*. (1990) Dual-energy x-ray absorptiometry for total
548 body and regional bone mineral and soft tissue composition. *Am J Clin Nutr* **51**, 1106-1112.

549 26. Wang ZM, Visser M, Ma R *et al*. (1996) Skeletal muscle mass: evaluation of neutron
550 activation and dual-energy X-ray absorptiometry methods. *J Appl Physiol* **80**, 824-831.

551 27. Liese AD, Doring A, Hense H *et al.* (2001) Five year changes in waist circumference, body
552 mass index and obesity in Augsburg, Germany. *Eur J of Nutr* **40**, 282-288.
553
554 28. Ford E, Mokdad A, Giles W (2003) Trends in waist circumference among US adults. *Obesity*
555 **11**, 1223-1231.
556
557 29. Balkau B, Picard P, Vol S *et al.* (2007) Consequences of change in waist circumference on
558 cardiometabolic risk factors over 9 years. *Diabetes Care* **30**, 1901-1903.
559
560 30. Development Core Team. R: A language and environment for statistical computing. Vienna:
561 R Foundation for Statistical Computing; 2006 [cited 2011]. ISBN 3-900051-07-0. Available
562 from: http://www. R-project.org.
563
564 31. Guo SS, Zeller C, Chumlea WC, Siervogel RM (1999) Aging, body composition, and
565 lifestyle: the Fels Longitudinal Study. *Am J Clin Nutr* **70**, 405-411.
566
567 32. Buffa R, Floris GU, Putzu PF et al. (2011) Body composition variation in ageing. *Collegim*
568 *Antropologicum* **35**, 259-265.
569

570 33. Henche SA, Torres RR, et Pellico LG (2007) A evaluation of patterns of change in total and
571 regional body fat mass in healthy Spanish subjects using dual-energy X-ray absorptiometry
572 (DXA). *Eur J of Clin Nutr* **62**, 1440-1448.
573
574 34. Welch GW & Sowers MR (2000) The interrelationship between body topology and body
575 composition varies with age among women. *J Nutr* **130**, 2371-2377.
576
577 35. Chumlea WC, Guo SS, Kuczmarski RJ *et al.* (2002) Body composition estimates from
578 NHANES III bioelectrical impedance data. *Int J obesity* 26, 1596-1609.
579
580 36. Atlantis E, Martin SA, Haren MT *et al.* (2008) Lifestyle factors associated with age-related
581 differences in body composition: the Florey Adelaide Male Aging Study. *Am J Clin Nutr* **88**,
582 95-104.
583
584 37. Hughes VA, Frontera, WR, Roubenoff, RE *et al.* (2002) Longitudinal changes in body
585 composition in older men and women: role of body weight change and physical activity. *Am*
586 *J Clin Nutr* **76**, 473-481.
587
588
589 38. Kyle UG, Melzer K, Kayser B *et al.* (2006) Eight-year longitudinal changes in body
590 composition in healthy Swiss adults. *J Am Coll Nutr* **25**, 493-501.
591
592 39. Okosun IS, Liao Y, Rotimi CN *et al.* (2000) Abdominal adiposity and clustering of multiple
593 metabolic syndrome in White, Black and Hispanic Americans. *Ann Epidemiol* **5**, 263-270.
594
595 40. Wu C-H, Heshka S, Wang J et al. (2007) Truncal fat in relation to total body fat: inuences of
596 age, sex, ethnicity and fatness. *Int J Obesity* **31**, 1384-1391.
597
598 41. Casas YG, Schiller BC, DeSouza CA *et al.* (2001) Total and regional body composition
599 across age in healthy Hispanic and white women of similar socioeconomic status. *Am J Clin*
600 *Nutr* **73**, 13-18.
601
602 42. Fernandez JR, Heo M, Heymsfield SB et al. (2003) Is percentage body fat differentially
603 related to body mass index in Hispanic American, African Americans, and European
604 Americans? *Am J Clini Nutri* **77**, 71-75.

43. Alema-Mateo H, Lee SY, Javed F et al. (2009) Elderly Mexicans have less muscle and greater total and truncal fat compared to African-Americans and Caucasians with the same BMI. *J Nutr Heal & Aging* **13**, 919-923.

Table 1. Starting value of covariates (Age, Height, Weight, Waist circumference) for the three BMI categories at 20 y (BMIref) in men and women.

Gender	Normal (BMI= 22 kg/m^2)	Overweight (BMI= 27 kg/m^2)	Obese (BMI= 30 kg/m^2)
	(Age, Height, Weight, Waist)	(Age, Height, Weight, Waist)	(Age, Height, Weight, Waist)
Men	(20 y, 175 cm, 67 kg, 85 cm)	(20 y, 175 cm, 85 kg, 95 cm)	(20 y, 175 cm, 95 kg, 105 cm)
Women	(20 y, 165 cm, 60 kg, 83 cm)	(20 y, 165 cm, 75 kg, 95 cm)	(20 y, 165 cm, 85 kg, 105 cm)

Table 2. Proposed parametric models for assessing BMI category and ethnicity effect, respectively, for a given gender and a given weight trend context. BMI effect is first studied, then ethnicity effect. $Mi_B(A)$ denotes models associated to the BMI effect. $M1i_{B,E}(A)$ the extension of $M1_B(A)$.

Label	Model	Number of free parameters
For BMI category effect		
$M0_B(A)$	$SBC = \eta + \alpha \times A + \beta \times A^2$	3
$M1_B(A)$	$SBC = \eta_B + \alpha \times A + \beta \times A^2$	5
$M2_B(A)$	$SBC = \eta_B + \alpha_B \times A + \beta \times A^2$	7
$M3_B(A)$	$SBC = \eta_B + \alpha_B \times A + \beta_B \times A^2$	9
For ethnicity effect		
$M10_{B,E}(A)$	$SBC = \eta_B + \alpha \times A + \beta \times A^2$	5
$M11_{B,E}(A)$	$SBC = \mu_E + \eta_B + \alpha \times A + \beta \times A^2$	7
$M12_{B,E}(A)$	$SBC = \mu_E + \eta_{B,E} + \alpha \times A + \beta \times A^2$	11
$M13_{B,E}(A)$	$SBC = \mu_E + \eta_{B,E} + \alpha_E \times A + \beta \times A^2$	13
$M14_{B,E}(A)$	$SBC = \mu_E + \eta_{B,E} + \alpha_E \times A + \beta_E \times A^2$	15

Table 3. Men: mean and standard deviation of segmental body composition variables obtained by DXA in NHANES for different age intervals and ethnicity categories.

Ethnicity		20-29 y	30-39 y	40-49 y	50-59 y	60-69 y	>70 y
EA	**n**	285	289	297	270	312	531
	TF	9.08±4.99	10.46±5.12	11.90±4.80	13.18±5.23	14.33±5.15	12.28±4.49
	BF	19.03±8.84	20.37±8.49	22.27±7.70	23.87±8.31	25.43±8.25	22.66±7.39
	APL	28.52±4.24	28.71±4.58	28.80±4.43	27.55±3.95	26.87±4.23	23.60±3.76
	TL	30.73±4.41	31.11±4.64	32.09±4.52	31.57±4.17	31.62±4.46	28.60±4.15
	BL	62.83± 8.65	63.41± 9.25	64.49± 8.93	62.74± 8.10	62.08± 8.67	55.63± 7.96
AA	**n**	130	131	150	98	117	95
	TF	7.05±4.72	8.11±4.27	9.51±4.65	10.70±5.11	11.11±5.13	11.12±4.55
	BF	16.62±9.46	17.43±7.76	19.42±8.07	20.82±8.73	21.49±8.47	21.88±7.87
	APL	31.06±5.39	30.97±5.02	30.49±5.02	29.71±5.00	28.69±4.33	25.38±4.61
	TL	29.36±4.76	29.93±4.47	30.29±4.56	30.68±4.69	30.56±4.55	27.46±4.04
	BL	64.17±10.27	64.57± 9.42	64.46± 9.59	64.15± 9.74	63.07± 8.82	56.47± 8.66

TF, trunk fat; BF, body fat; TL, trunk lean; APL, appendicular lean; BL, body lean.
EA means European-American; AA means African-American.

Table 4. Women: Mean and standard deviation of segmental body composition variables obtained by DXA in NHANES women for different age intervals and ethnicity categories.

Ethnicity		20-29 y	30-39 y	40-49 y	50-59 y	60-70y	>70 y
EA	**n**	228	278	271	245	292	516
	TF	10.14±4.95	11.46±5.52	12.83±5.85	13.97±5.24	14.67±4.83	12.66±4.36
	BF	23.27± 8.53	25.41± 9.80	27.50±10.41	28.77± 8.86	29.90± 8.15	26.45± 7.81
	APL	18.39±2.63	18.49±3.08	18.70±3.34	17.98±2.82	17.46±2.90	15.80±2.56
	TL	21.72±2.70	22.29±3.10	22.98±3.36	22.57±3.13	22.01±3.09	20.45±2.79
	BL	43.05±5.28	43.70±6.18	44.63±6.72	43.50±5.90	42.42±5.97	39.09±5.30
AA	**n**	107	130	157	89	121	93
	TF	10.92±5.60	12.60±5.69	14.11±4.76	14.41±4.94	14.74±5.03	12.99±4.59
	BF	25.61±10.16	28.26±10.23	31.01± 8.90	31.29± 8.75	31.09± 9.57	28.63± 9.06
	APL	21.16±3.49	22.04±4.08	21.50±3.28	20.12±3.25	20.45±3.50	18.99±3.47
	TL	21.50±3.00	22.73±3.57	23.08±3.01	22.50±3.23	23.01±3.20	21.61±3.25
	BL	45.88±6.52	48.05±7.72	47.84±6.21	45.83±6.38	46.76±6.66	43.75±6.67

TF, trunk fat; BF, body fat; TL, trunk lean; APL, appendicular lean; BL, body lean.
EA means European-American; AA means African-American.

Table 5. For BMI effect, standard error of the estimate (kg) of different models in EA men and women. In columns, the five segmental body compositions, in rows, the different models for two weight trend contexts.

Gender		Model	TF	BF	APL	TL	BL
Men	RP	Nonpara	2.43	3.88	2.37	2.17	4.23
		$M0_B(A)$	3.54	5.74	3.09	3.09	6.05
		$M1_B(A)$	2.43	3.88	2.35	2.16	4.2
		$M2_B(A)$	2.43	3.88	2.35	2.16	4.2
		$M3_B(A)$	2.43	3.87	2.34	2.15	4.18
	GP	Nonpara	2.51	3.97	2.46	2.26	4.39
		$M0_B(A)$	3.91	6.42	3.25	3.36	6.51
		$M1_B(A)$	2.52	3.99	2.45	2.29	4.42
		$M2_B(A)$	2.49	3.94	2.45	2.27	4.4
		$M3_B(A)$	2.48	3.91	2.45	2.27	4.4
		Model	TF	BF	APL	TL	BL
Women	RP	Nonpara	2.47	4.16	1.74	1.76	3.28
		$M0_B(A)$	3.96	6.91	2.26	2.36	4.52
		$M1_B(A)$	2.5	4.19	1.75	1.78	3.3
		$M2_B(A)$	2.49	4.19	1.75	1.78	3.3
		$M3_B(A)$	2.49	4.19	1.75	1.78	3.3
	GP	Nonpara	2.57	4.38	1.75	1.79	3.32
		$M0_B(A)$	4.2	7.41	2.36	2.45	4.71
		$M1_B(A)$	2.59	4.44	1.76	1.81	3.34
		$M2_B(A)$	2.59	4.42	1.76	1.8	3.33
		$M3_B(A)$	2.59	4.42	1.76	1.8	3.33

RP means reference profile; GP means gain profile.

TF, trunk fat; BF, body fat; TL, trunk lean; APL, appendicular lean; BL, body lean.

Table 6. For ethnicity effect, standard error of the estimate (kg) of different models in men and women (cf. Table 5).

Gender	Weight change context	Model	TF	BF	APL	TL	BL
Men	RP	Nonpara	2.34	3.76	2.36	2.14	4.16
		$M10_{B,E}$	2.45	3.83	2.55	2.18	4.27
		$M11_{B,E}(A)$	2.35	3.78	2.34	2.14	4.15
		$M12_{B,E}(A)$	2.35	3.79	2.34	2.14	4.15
		$M13_{B,E}(A)$	2.34	3.78	2.34	2.14	4.15
		$M14_{B,E}(A)$	2.34	3.78	2.34	2.14	4.15
	GP	Nonpara	2.43	3.88	2.43	2.23	4.32
		$M10_{B,E}(A)$	2.53	3.95	2.66	2.3	4.5
		$M11_{B,E}(A)$	2.44	3.91	2.44	2.27	4.38
		$M12_{B,E}(A)$	2.43	3.91	2.43	2.27	4.37
		$M13_{B,E}(A)$	2.43	3.91	2.44	2.27	4.37
		$M14_{B,E}(A)$	2.43	3.91	2.43	2.27	4.37
		Model	TF	BF	APL	TL	BL
Women	RP	Nonpara	2.41	4.16	1.79	1.74	3.29
		$M10_{B,E}(A)$	2.52	4.22	1.98	1.78	3.41
		$M11_{B,E}(A)$	2.44	4.2	1.8	1.76	3.32
		$M12_{B,E}(A)$	2.44	4.2	1.8	1.76	3.32
		$M13_{B,E}(A)$	2.44	4.2	1.8	1.76	3.31
		$M14_{B,E}(A)$	2.44	4.2	1.8	1.76	3.31
	GP	Nonpara	2.51	4.36	1.8	1.78	3.32
		$M10_{B,E}(A)$	2.61	4.41	2.02	1.83	3.47
		$M11_{B,E}(A)$	2.53	4.4	1.82	1.81	3.37
		$M12_{B,E}(A)$	2.53	4.4	1.82	1.81	3.37
		$M13_{B,E}(A)$	2.53	4.4	1.82	1.8	3.36
		$M14_{B,E}(A)$	2.53	4.4	1.81	1.8	3.36

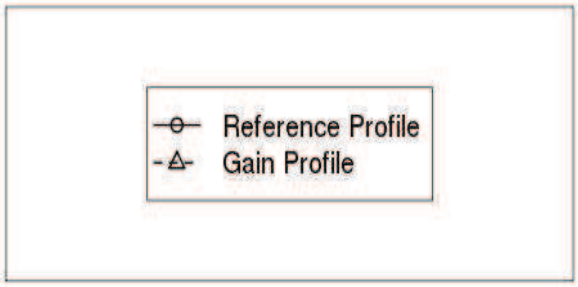
Table 7. For the retained model $M11_{B,E}(A)$ in men and women, parameter differences of ethnicity and BMI categories from their baseline category, which are Ethnicity=EA and BMI=normal weight, respectively.

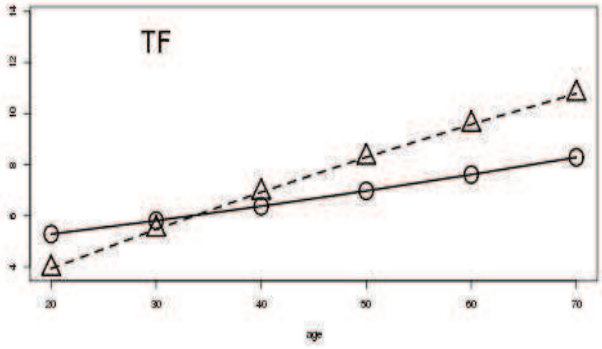
Gender	Weight change context		TF	BF	APL	TL	BL
Men	RP	$\mu_{AA} - \mu_{EA}$	-1.48	-1.31	2.19	-0.83	1.52
		$\eta_{OW} - \eta_N$	3.8	6.37	3.53	3.57	7.36
		$\eta_{OB} - \eta_N$	6.87	11.54	5.49	5.87	11.77
	GP	$\mu_{AA} - \mu_{EA}$	-1.34	-1.06	2.36	-0.76	1.78
		$\eta_{OW} - \eta_N$	4.45	7.67	3.63	3.89	7.8
		$\eta_{OB} - \eta_N$	7.67	13.02	5.76	6.39	12.6
			TF	BF	APL	TL	BL
Women	RP	$\mu_{AA} - \mu_{EA}$	-0.97	-0.58	1.9	-0.6	1.55
		$\eta_{OW} - \eta_N$	4.46	8.12	2.04	2.23	4.44
		$\eta_{OB} - \eta_N$	7.78	14.17	3.8	4.01	8.09
	GP	$\mu_{AA} - \mu_{EA}$	-0.96	-0.44	2.01	-0.65	1.63
		$\eta_{OW} - \eta_N$	4.68	8.73	2.38	2.47	5.01
		$\eta_{OB} - \eta_N$	8.32	15.14	4.17	4.27	8.72

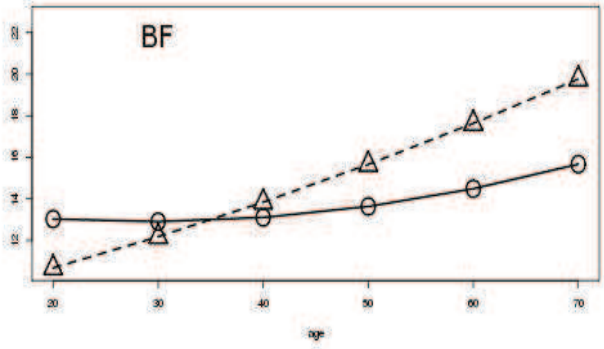
.

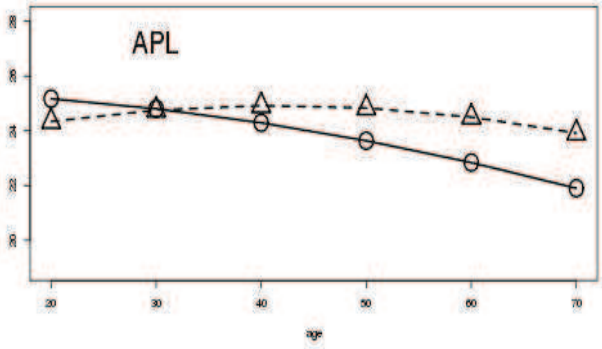
Figure 1: Age-related changes in segmental body composition for NHANES European-American normal men. Age is on the x-axis, and estimations of segmental body compositions from $M11_{B,E}(A)$ are on y-axis. Reference Profile is represented by (o) and Gain Profile by (Δ).

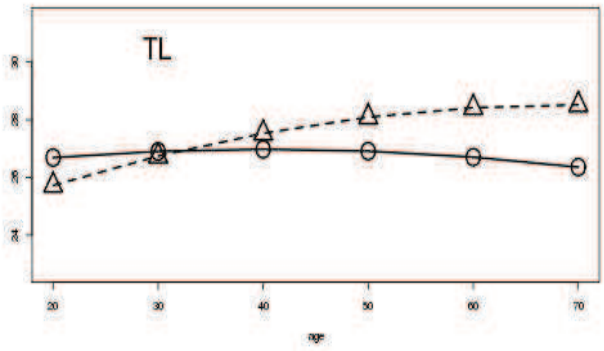
Figure 2: Age-related changes in segmental body composition for NHANES European-American normal women. Age is on the x-axis, and estimations of segmental body compositions from $M11_{B,E}(A)$ are on y-axis. Reference Profile is represented by (o) and Gain Profile by (Δ).

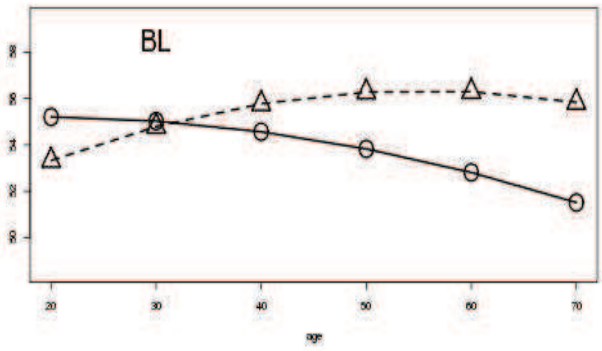

Reference Profile
Gain Profile

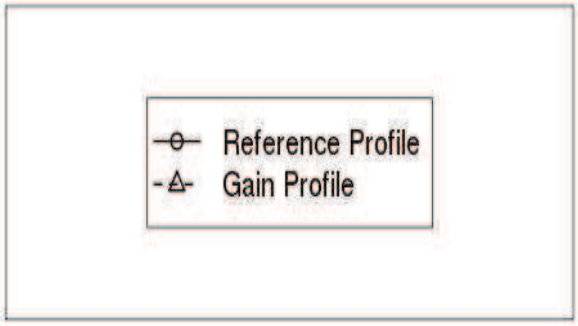

TF
20
30
40
50
60
70
age

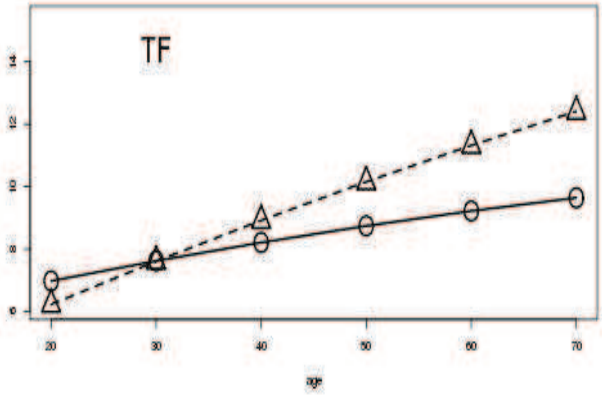

BF
20
30
40
50
60
70
age

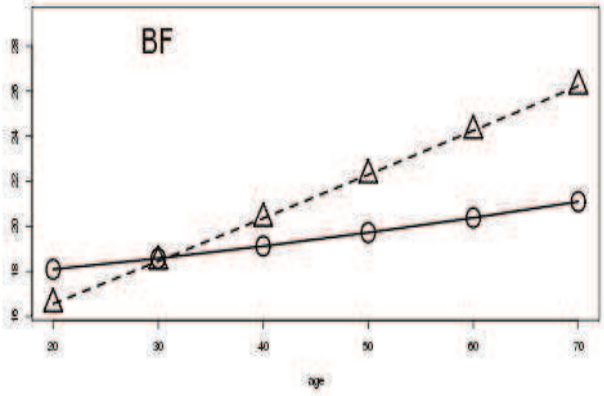

APL
20
30
40
50
60
70
age

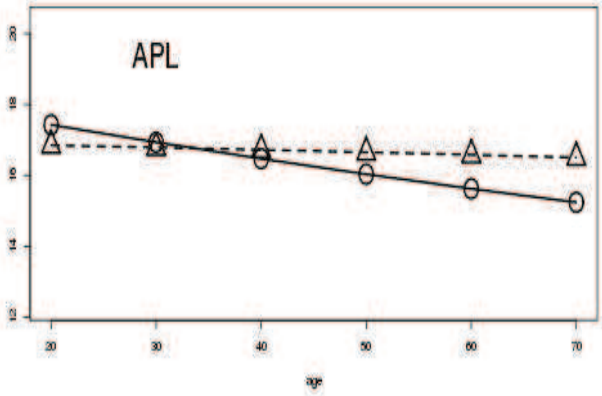

TL
20
30
40
50
60
70
age

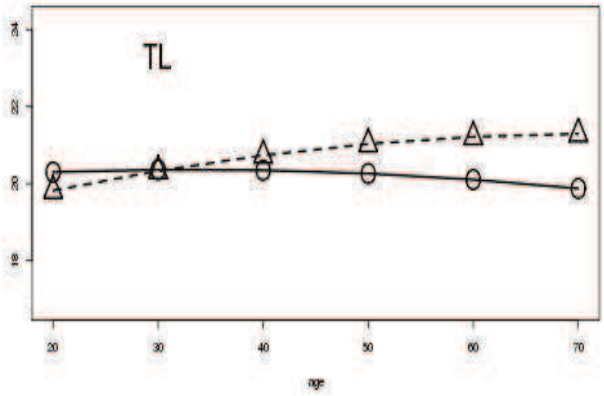

BL
20
30
40
50
60
70
age

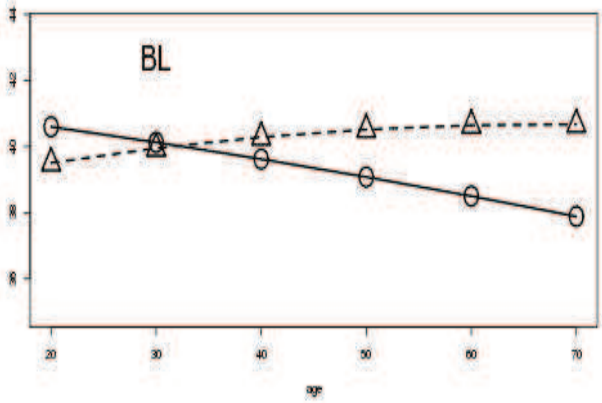

Reference Profile
Gain Profile


TF
20
30
40
50
60
70
age


BF
20
30
40
50
60
70
age


APL
20
30
40
50
60
70
age


TL
20
30
40
50
60
70
age


BL
20
30
40
50
60
70
age

Chapter 4

Conclusions

4.1 Contributions

This thesis presents body composition prediction by multivariate locally weighted and Bayesian networks modeling. The related work led to three scientific papers : one research paper has been published in the *British Journal of Nutrition*, the other two have been submitted to *British Journal of Nutrition* and *Journal de la Société Française de Statistique*, respectively. Valuable accomplishments of our work can be summarized in three main contributions :

- First of all, several multivariate models have been proposed, such as locally weighted linear, SVM regression and Bayesian linear models. One main advantage of these multivariate models consists of predicting simultaneously segmental body compositions. As a matter of fact, segmental compartments like body fat, trunk fat, and appendicular lean masses, are useful factors for assessing pre-disposition to metabolic risks; therefore, the joint examinations of these segmental compartments provide interesting information. Moreover a particular focus on the importance of waist circumference has been given in the published paper (section 3.2.2). The accuracy of the multivariate model was improved when waist circumference was entered as a predictor variable. This was particularly meaningful for men for the segmental compartments, such as trunk fat, appendicular lean and total body fat and lean masses. For men, a significant improvement in accuracy was observed in all the BMI categories and in the age categories of 20-35, 35-50 and 50-65 years.

- The second contribution of this thesis was to study age-related changes in segmental body compositions, associated with anthropometric covariables. In a general framework, two Bayesian modeling methods are proposed for age-related change study. The reasoning of building the models starts from a static scheme, then extending it to a dynamic scheme. A set of static DAGs is introduced, and these DAGs are used to describe the dependencies between the anthropometric covariables and segmental body compositions, as well as the order of assessment of covariables and SBC. Each anthropometric covariable is associated to a proper age-related function. These functions allow to estimate covariable values in the dynamic scheme (*i.e.*, when aging). The main advantage of these methods is to allow conducting a longitudinal analysis from the cross-sectional datasets with the help of modeling trajectories. Also, the Bayesian modeling enables to provide a prediction distribution, rather than a simple value, this is more relevant for exploring the uncertainty or accuracy problems.

- The third contribution of this thesis was to propose a family of Crossed Gaussian Bayesian Networks. One of the advantages of the BNs formulation is to allow non-statistician, typically expert of one domain, to enter into their mechanism through the easy to understand

DAG presentations. In our mind, such BNs must and can serve as thinking material for non expert in BNs. The idea is to use structured DAGs when the set of nodes is the product of two series of items. A parsimonious sub-model of multivariable model is described by a Gaussian Bayesian network, and the purpose consists of obtaining a better multivariable prediction than with a plain linear regression model. We demonstrated that, at least for gaussian Bayesian networks modelling, it was possible to introduce a known structure on the set of variables of interest, and that can lead to very effective results to obtain an interpretable predictive formula. This novel statistical method is applied to the prediction of segmental body compositions adding simple easy acquired covariables. The results show that crossed BNs globally perform better than the saturated model (SEP). In addition, the reduction of the parametric dimension, with respect to the saturated model is striking, especially for the variance parameters.

4.2 Limits

Along with these achievements, we still feel necessary to point out some difficulties encountered and solved during these studies :

- The statistical models are data-based, therefore, when comparing with some published models based on different measurement datasets and cohorts, we re-adjusted their models by using the same training DXA datasets in order to propose fair comparisons.

- Available datasets were cross-sectional, but we attempted to conduct a study of age-related changes in segmental body compositions. Thanks to literature contributions, we were able to build an age-related function associated to each anthropometric covariable in the present study. These functions allowed to estimate covariable values in a dynamic scheme (*i.e.*, when aging). By using these estimated covariate values, segmental body composition was finally predicted in a longitudinal framework.

- Most of the computations have been made with the **R** software (R Development Core Team, 2009), a very convenient statistical tool for its power, flexibility and available algorithms. Nevertheless, the fuzzy SVM (subsection 2.3.3) is not available in **R** environment, but in **MatLab**[1]. The re-implementation in **R** environment was not a major aim of this thesis, also it is time-consuming, therefore we attempted to find an alternative to perform fuzzy SVM in **R** environment. Fortunately, there is a R package[2] by which it is feasible to call matlab functions from **R** enviroment. Pseucoding has been writen to call corresponding **MatLab** functions to perform locally weighted SVM model. Furthermore, another rationale to use available SVM functions in **MatLab** is because they provide a good software warranty and they are widely used in different applications.

- Non-parametric approaches are promising and appealing, however one difficulty in locally weighted approaches is to select an appropriate subset of candidate subjects with similar covariable values for a predicted subject. At the beginning, we used a discrete way based on a single cut-off distance value to decide whether or not the candidate subjects should be included (either inclusion or exlusion), thus there is no smooth weighting, even for the included candidate subjects[3]. This way could have disadvantages for unusual predicted subjects, such as a short obese subject, because the prediction will be penalized by presence of less similar included subjects, however it is better to take into account

[1] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[2] The package is 'R.matlab', and the user manuel can be found in http://cran.r-project.org/web/packages/R.matlab/R.matlab.pdf.

[3] All included subjects have the same contribution to the prediction process.

more similar candidates. To consider these issues, we thus proposed a continuous way with smooth weighting such that each candidate has his/her own attribution depending on the distance value.

- Multivariate regression model proposed to answer the first thesis aim had the inconvenience in too many of parameters. More precisely, it had $p(q+1)$ parameters for the expectation and $\frac{p(p+1)}{2}$ parameters for the covariance matrix, where p is the number of variables and q is the number of covariables. To overcome this point, we proposed to use Bayesian networks which are all the more efficient that the number of variables is high. But there is another difficulty in elaborating the Bayesian networks concept. Generally, there is no distinction in Bayesian networks between the predictor variables and the predicted variables, but in the present study, we believed that SBCs were correlated between them, even together with inclusion of the predictor variables. If we assumed an edge from each predictor variable to SBCs, the modeling would become a satured linear model and it will be not cost-efficient in parameter dimensions. By performing a heuristic research over a selected number of possible Bayesian networks, we found an equal structure between fat, lean and bone masses within different compartments; therefore we deduced the Crossed Bayesian networks idea with a possible structure between the predicted variables.

Due to these difficulties, we were able to advance progressively in our work. However, these are still some limitations that we should highlight :

- First of all, the prediction models were based on datasets only including BMI ranged from 18-40 kg/m^2, thus the reliability of the models is doubtful when applied to subjects whose BMI value is out of this range.

- Body composition can vary significantly based on several genetic factors, such as gender and ethnicity (Okosun *et al.*, 2000), however, the proposed prediction models only account for the gender factor, which seems for us to be the most important factor.

- In section 3.3 related to age-related change study, due to the lack of the longitudinal datasets, the prediction accuracy of our models has not been validated.

- The prediction models were built from DXA-measured datasets (NHANES datasets). Although DXA is considered as a standard technique for body composition measurement, its application has also limitations (Roubenoff *et al.*, 1993). The scanning bed has an upper weight limit and the whole-body field-of-view can not accommodate very large persons; DEXA cannot distinguish clearly between soft tissue and bone in certain segments. In the present study, the aim was not to evaluate DXA measurement, but to make a good use of the available DXA-measured datasets to predict segmental body compositions from easily acquired anthropometric variables. Thus, we were not interested in the absolute error related to DXA measurement, but in the prediction error of the models, which inherited the DXA measurement error.

Besides, we found some unexpected results that we want to underline. The sophisticated locally weighted approaches have been proposed, and they were expected to have a better prediction performance than the global multivariate modeling. Nevertheless, the results showed a very close prediction performance between the locally weighted approaches and the global multivariate model. We have even conducted an error research study at a subject-level, *i.e.*, we examinated the prediction error for each predicted subject, and it turned out that there was always a consistency of prediction error between the locally weighted approaches and the global approach. Although without a comphrensitive explanation about this point, we strongly feel this is due to the datasets. Indeed, for most of the predicted subjects, there was a relevant

number of similar candidates, that might lead to a good prediction accuracy even from a simple model. Therefore, even if the locally weighted approaches correctly identify an exact neighborhood for a predicted subject, the prediction accuracy will not be improved as much as we expected.

4.3 Perspectives

Body composition prediction by our multivariate proposal has been shown accurate and useful, we think necessary to mention new potential of our work in clinical applications :

- Our body composition prediction is useful in achieving an improved understanding of how segmental body compositions influence overall health and disease, and in evaluating interventions. Several studies showed that body composition was related to some disease risk factors (Snijder *et al.*, 2004; Van Pelt *et al.*, 2002; Johnstone *et al.*, 2005). The present study aimed at predicting segmental body compositions and estimating the differential changes between them with aging. Thus, we are more likely in preclinical stage. With the help of segmental body composition prediction provided by our models, experts could identify individual health risk associated with total body fat or excessive trunk fat masses. Our proposed models, accurate and straightforward, could be useful to clinicians and medical practitioners. Only by inputting age and anthropometric informations, they can estimate a patient's body composition because the formulae can be easily implemented in an spreadsheets-like engine. They could assess or exploit some relevant indices related to health risks, such as a ratio of trunk fat to leg lean mass, and mimic their changes with aging.

- Our body composition assessment could be helpful in studying nutritional status. Expert could use predicted segmental body composition, obtained from our models, to evaluate nutritional status by (1) comparing individuals with themselves or with reference values; (2) determining whether individuals or groups fall within the population range; (3) promoting further research on segmental body composition index for health and nutritional assessment. Based on our multivariate model, the estimation of segmental body composition is as simple as a calculation of a BMI value, because, besides height and weight value, only age and waist circumference are needed. Then physiologists could use these predicted segmental body composition to develop more relevant and meaningful indices.

- Another contribution of the present study is that we developed statistical modeling methods for assessing age-related changes in segmental body composition, the interest of the proposal was to be able to use a cross-sectional dataset for a longitudinal analysis thank to assumptions and literature contributions. Segmental body composition, such as fat-free mass, trunk fat and leg lean masses, is associated with risk factors for a variety of chronic diseases from middle to old age. Therefore, understanding the scope of the age-related changes in body composition will help to investigate the relationship between body composition and increased morbidity and mortality in elderly, moreover, to assist in the management of health status into old ages.

- Due to a lack of the longitudinal datasets, we have not been able to achieve all the work. In section 3.3, we studied age-related changes in body composition, and we aimed to provide a general overview about how SBCs evolve with aging. Nevertheless, the prediction accuracy has not been investigated. For further research, it is necessary to have longitudinal datasets at our disposal to verify the prediction performance of the proposed modeling.

- Moreover, in subsection 3.3.1, the dynamic Bayesian modeling was conducted at the subject-level, *i.e.*, we focused on predicting body composition evolution for a particular anthropometric profile. In further works, we could extend it at the population-level prediction. For instance, when applying this modeling on a group of subjects aged 20 years in a dataset, we will be able to obtain the prediction from 30 years to 80 years, and investigate the variations related to age effect.

- In the methodological framework of the Crossed Gaussian Bayesian Network, a potential further research is the use of other distributions than the Normal one, such as beta distribution, log-normal distribution or a mixed distribution. Mathematical properties will be much more difficult to obtain, but the advantage would be to go closer to the reality.

- In the application framework of our proposed models, another potential research direction is to investigate usefulness of other covariables, such as ethnicity, physical activity or anthropometric ratios. In addition, these new covariables could help to estimate a differential age-related changes in segmental body compositions.

Bibliography

Acid, S. et de Campos, L. M. (2001). A hybrid methodology for learning belief networks: BENEDICT. *International Journal of Approximate Reasoning*, 27(3): 235–262.

Acid, S. et de Campos, L. M. (2003). Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs. *J. Artif. Intell. Res.(JAIR)*, 18: 445–490.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6): 716–723.

Aleman-Mateo, H., Lee, S., Javed, F., Thornton, J., Heymsfield, S., Pierson, R., Pi-Sunyer, F., Wang, Z., Wang, J., et Gallagher, D. (2009). Elderly Mexicans have less muscle and greater total and truncal fat compared to African-Americans and Caucasians with the same BMI. *The journal of nutrition, health & aging*, 13(10): 919–923.

Aliferis, C. F., Tsamardinos, I., et Statnikov, A. (2003). HITON: a novel Markov Blanket algorithm for optimal variable selection. In *AMIA Annual Symposium Proceedings*, volume 2003, page 21. American Medical Informatics Association.

Anderson, T. (2003). *An introduction to multivariate statistical analysis.* Wiley, 3rd edition edition.

Atkeson, C. G., Moore, A. W., et Schaal, S. (1997). Locally weighted learning. *Artificial intelligence review*, 11(1-5): 11–73.

Bacciu, D., Etchells, T., Lisboa, P., et Whittaker, J. (2013). Efficient identification of independence networksusing mutual information. *Comput Stat*, 28: 621–646.

Bang-Jensen, J. et Gutin, G. (2009). *Digraphs: Theory, Algorithms and Applications.* Springer, 2nd edition.

Bedogni, G., Brambilla, P., Bellentani, S., et Tiribelli, C. (2006). The assessment of body composition in health and disease. *Journal of Human Ecology Spe*, Special Issue(14): 21–25.

Behnke, A. R. (1942). Physiologic studies pertaining to deep sea diving and aviation, especially in relation to the fat content and composition of the body: the Harvey lecture, March 19, 1942. *Bulletin of the New York Academy of Medicine*, 18(9): 561.

Bergeron, C., Cheriet, F., Ronsky, J., Zernicke, R., et Labelle, H. (2005). Prediction of anterior scoliotic spinal curve from trunk surface using support vector regression. *Engineering Applications of Artificial Intelligence*, 18(8): 973–983.

Bleeker, S., Moll, H., Steyerberg, E., Donders, A., Derksen-Lubsen, G., Grobbee, D., et Moons, K. (2003). External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology*, 56(9): 826–832.

Bottcher, S. G. et Dethlefsen., C. (2012). *deal: Learning Bayesian Networks with Mixed Variables.*

Bouckaert, R. R. (1995). *Bayesian belief networks: from construction to inference*. PhD thesis.

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

Brožek, J., Grande, F., Anderson, J. T., et Keys, A. (1963). DENSITOMETRIC ANALYSIS OF BODY COMPOSITION: REVISION OF SOME QUANTITATIVE ASSUMPTIONS*. *Annals of the New York Academy of Sciences*, 110(1): 113–140.

Brožek, J. et Keys, A. (1951). The evaluation of leanness-fatness in man: norms and interrelationships. *British Journal of Nutrition*, 5(02): 194–206.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2): 121–167.

Burmaster, D. et Crouch, E. (1997). Lognormal distributions for body weight as a function of age for males and females in the United States, 1976–1980. *Risk Analysis*, 17(4): 499–505.

Calle, E. E., Rodriguez, C., Walker-Thurmond, K., et Thun, M. J. (2003). Overweight, obesity, and mortality from cancer in a prospectively studied cohort of US adults. *New England Journal of Medicine*, 348(17): 1625–1638.

Celeux, G., M. J.-M. et Robert, C. (2006). Sélection bayésienne de variables en régression linéaire. *Journal de la Société Française de Statistique*, 147(1): 59–79.

Chan, D., Watts, G., Barrett, P., et Burke, V. (2003). Waist circumference, waist-to-hip ratio and body mass index as predictors of adipose tissue compartments in men. *Qjm*, 96(6): 441–447.

Chang, C. et Lin, C. (2002). Training v-support vector regression: theory and algorithms. *Neural Computation*, 14(8): 1959–1977.

Chickering, D., Geiger, D., et Heckerman, D. (1995). Learning Bayesian networks: Search methods and experimental results. In *Fifth International Workshop on Artificial Intelligence and Statistics*, pages 112–128.

Chickering, D. M. (2003). Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3: 507–554.

Chumlea, W., Guo, S., Kuczmarski, R., Flegal, K., Johnson, C., Heymsfield, S., Lukaski, H., Friedl, K., et Hubbard, V. (2002). Body composition estimates from NHANES III bioelectrical impedance data. *International journal of obesity and related metabolic disorders journal of the International Association for the Study of Obesity*, 26(12): 1596–1609.

Chumlea, W., Guo, S. S., *et al.* (2000). Assessment and prevalence of obesity: application of new methods to a major problem. *Endocrine*, 13(2): 135.

Colditz, G. A., Willett, W. C., Rotnitzky, A., et Manson, J. E. (1995). Weight gain as a risk factor for clinical diabetes mellitus in women. *Annals of internal medicine*, 122(7): 481–486.

Committee, W. E. *et al.* (1995). Physical status: the use and interpretation of anthropometry. Technical Report 121.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence*, 42(2): 393–405.

Cooper, G. F. et Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4): 309–347.

Cruz-Ramírez, N., Acosta-Mesa, H.-G., Barrientos-Martínez, R.-E., et Nava-Fernández, L.-A. (2006). How good are the Bayesian information criterion and the minimum description length principle for model selection? A Bayesian network analysis. In *MICAI 2006: Advances in Artificial Intelligence*, pages 494–504.

Csardi, G. et Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems: 1695.

Daly, R., Shen, Q., et Aitken, S. (2011). Learning Bayesian networks: approaches and issues. *The knowledge Engineering Review*, 26(3): 99–157.

de Campos, C. et Ji, Q. (2010). Properties of Bayesian Dirichlet scores to learn Bayesian network structures. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

De Campos, L. M., Fernandez-Luna, J. M., Gámez, J. A., et Puerta, J. M. (2002). Ant colony optimization for learning Bayesian networks. *International Journal of Approximate Reasoning*, 31(3): 291–311.

de Koning, L., Merchant, A. T., Pogue, J., et Anand, S. S. (2007). Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: meta-regression analysis of prospective studies. *European heart journal*, 28(7): 850–856.

Denis, J.-B. et Gower, J. C. (1996). Asymptotic confidence regions for biadditive models: interpreting genotype-environment interactions. *Applied Statistics*, 45(4): 479–493.

Dobbelsteyn, C., Joffres, M., MacLean, D., et Flowerdew, G. (2001). A comparative evaluation of waist circumference, waist-to-hip ratio and body mass index as indicators of cardiovascular risk factors. The Canadian Heart Health Surveys. *International journal of obesity*, 25(5): 652–661.

Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S., et Murray, C. J. (2002). Selected major risk factors and global and regional burden of disease. *The Lancet*, 360(9343): 1347–1360.

Fahrmeir, L. et Tutz, G. (1994). *Multivariate Statistical Modelling based on Generalized Linear Models*. Springer-Verlag.

Faraway, J. (1995). Data Splitting Strategies for Reducing the Effect of Model Selection on Inference.

Fedorov, V. V., Hackl, P., et Müller, W. G. (1993). Moving local regression: The weight function. *Journaltitle of Nonparametric Statistics*, 2(4): 355–368.

Fezeu, L., Minkoulou, E., Balkau, B., Kengne, A.-P., Awah, P., Unwin, N., Alberti, G. K., et Mbanya, J.-C. (2006). Association between socioeconomic status and adiposity in urban Cameroon. *International journal of epidemiology*, 35(1): 105–111.

Floud, R. (1994). Heights of Europeans since 1750: ANewSource for EuropeanEconomicHistory. *Stature, Living Standards, and Economic Development: Essays in Anthropometric History. Chicago, Univ. of Chicago Press*, pages 9–24.

Ford, E., Mokdad, A., et Giles, W. (2003). Trends in waist circumference among US adults. *Obesity*, 11(10): 1223–1231.

Foulley, J.-L., Sorensen, D., Robert-Granié, C., et Bonaïti, B. (2004). Heteroskedasticity and Structural Models for Variances. *Jour. Ind. Soc. Ag. Slatistics*, 57: 64–70.

Freedman, D. S., Khan, L. K., Dietz, W. H., Srinivasan, S. R., et Berenson, G. S. (2001). Relationship of childhood obesity to coronary heart disease risk factors in adulthood: the Bogalusa Heart Study. *Pediatrics*, 108(3): 712–718.

Frey, L., Fisher, D., Tsamardinos, I., Aliferis, C. F., et Statnikov, A. (2003). Identifying Markov blankets with decision tree induction. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 59–66. IEEE.

Friedman, J., Hastie, T., et Tibshirani, R. (2007). Sparse Inverse Covariance Estimation With the Graphical Lasso. *Biostatistics*, 9: 432–441.

Fu, S. et Desmarais, M. C. (2008). Tradeoff analysis of different Markov blanket local learning approaches. pages 562–571.

Gallagher, D., Heymsfield, S., Heo, M., Jebb, S., Murgatroyd, P., et Sakamoto, Y. (2000a). Healthy percentage body fat ranges: an approach for developing guidelines based on body mass index. *The American Journal of Clinical Nutrition*, 72(3): 694–701.

Gallagher, D., Ruts, E., Visser, M., Heshka, S., Baumgartner, R., Wang, J., Pierson, R., Pi-Sunyer, F., et Heymsfield, S. (2000b). Weight stability masks sarcopenia in elderly men and women. *American Journal of Physiology- Endocrinology And Metabolism*, 279(2): 366–375.

Gallagher, D., Visser, M., Sepulveda, D., Pierson, R., Harris, T., et Heymsfield, S. (1996). How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups? *American journal of epidemiology*, 143(3): 228.

Galloway, A. (1988). Estimating actual height in older individuals. *Journal of Forensic Sciences*, 33(1): 126–136.

Gao, J., Gunn, S. R., et Harris, C. J. (2003). Mean field method for the support vector machine regression. *Neurocomputing*, 50: 391–405.

Gasse, M., Aussem, A., et Elghazel, H. (2012). An Experimental Comparison of Hybrid Algorithms for Bayesian Network Structure Learning. *Machine Learning and Knowledge Discovery in Databases*, 7523: 58–73.

Glover, F. (1989). Tabu searchpart I. *ORSA Journal on computing*, 1(3): 190–206.

Graybill, F. A. (1976). *Theory and application of the linear model*. Duxbury Press.

Gunn, S. (1998). Support vector machines for classification and regression. Technical report.

Guo, S., Zeller, C., Chumlea, W., et Siervogel, R. (1999). Aging, body composition, and lifestyle: the Fels Longitudinal Study. *The American journal of clinical nutrition*, 70(3): 405–411.

Hansen, M. H. et Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454): 746–774.

Hastie, T., Tibshirani, R., et Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag.

Heckerman, D. *et al.* (1998). A tutorial on learning with Bayesian networks. *Nato Asi Series D Behavioural And Social Sciences*, 89: 301–354.

Heckerman, D., Geiger, D., et Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3): 197–243.

Hercberg, S., Galan, P., Preziosi, P., Bertrais, S., Mennen, L., Malvy, D., Roussel, A., Favier, A., et Briancon, S. (2004). The SU. VI. MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Archives of internal medicine*, 164(21): 2335–2342.

Heyward, V. et Stolarczyk, L. (1996). *Applied Body Composition Assessment*. Champaign, IL: Human Kinetics.

Hofmann, T., Scholkopf, B., et Smola, A. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3): 1171–1220.

Hsieh, S. D. et Yoshinaga, H. (1995). Waist/height ratio as a simple and useful predictor of coronary heart disease risk factors in women. *Internal medicine (Tokyo, Japan)*, 34(12): 1147–1152.

Hughes, V. A., Roubenoff, R., Wood, M., Frontera, W. R., Evans, W. J., et Singh, M. A. F. (2004). Anthropometric assessment of 10-y changes in body composition in the elderly. *The American journal of clinical nutrition*, 80(2): 475–482.

Hyndman, R. J. et Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4): 679–688.

Ice, C. L., Cottrell, L., et Neal, W. A. (2009). Body mass index as a surrogate measure of cardiovascular risk factor clustering in fifth-grade children: Results from the coronary artery risk detection in the Appalachian Communities Project. *International Journal of Pediatric Obesity*, 4(4): 316–324.

Ivanciuc, O. (2007). Applications of support vector machines in chemistry. *Reviews in Computational Chemistry*, 23: 291–400.

Jackson, A., Stanforth, P., Gagnon, J., Rankinen, T., Leon, A., Rao, D., Skinner, J., Bouchard, C., et Wilmore, J. (2002). The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. *International Journal of Obesity*, 26(6): 789–796.

Jackson, A. S. *et al.* (1984). Research design and analysis of data procedures for predicting body density. *Med Sci Sports Exerc*, 16(6): 616–620.

Janssen, I., Heymsfield, S., Allison, D., Kotler, D., et Ross, R. (2002). Body mass index and waist circumference independently contribute to the prediction of nonabdominal, abdominal subcutaneous, and visceral fat. *The American Journal of Clinical Nutrition*, 75(4): 683–688.

Jensen, F. et Nielsen, T. (2007). *Bayesian networks and decision graphs*. Springer Verlag.

Jensen, F. V. (1996). *An introduction to Bayesian networks*, volume 210. UCL press London.

Johnstone, A., Murison, S., Duncan, J., Rance, K., et Speakman, J. (2005). Factors influencing variation in basal metabolic rate include fat-free mass, fat mass, age, and circulating thyroxine but not sex, circulating leptin, or triiodothyronine. *The American journal of clinical nutrition*, 82(5): 941–948.

Justice, A. C., Covinsky, K. E., et Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Annals of internal medicine*, 130(6): 515–524.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., et Bühlmann, P. (2012). Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software*, 47(11): 1–26.

Keys, A. et BroŽek, J. (1953). Body fat in adult man. *Physiological Reviews*, 33(3): 245–325.

Keys, A., Fidanza, F., Karvonen, M. J., Kimura, N., et Taylor, H. L. (1972). Indices of relative weight and obesity. *Journal of chronic diseases*, 25(6): 329–343.

Kim, J. H. et Pearl, J. (1987). CONVINCE: A conversational inference consolidation engine. *Systems, Man and Cybernetics, IEEE Transactions on*, 17(2): 120–132.

Ko, G., Chan, J., Woo, J., Lau, E., Yeung, V., Chow, C.-C., Wai, H., Li, J., So, W.-Y., et Cockram, C. (1997). Simple anthropometric indexes and cardiovascular risk factors in Chinese. *International journal of obesity*, 21(11): 995–1001.

Koller, D. et Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Koller, D. et Sahami, M. (1996). Toward Optimal Feature Selection. In *Proceedings of 13th conference on machine learning: 3-6 July 1996*, pages 284–292.

Korb, K. B. et Nicholson, A. E. (2011). *Bayesian Artificial Intelligence*. CRC press, 2nd edition.

Kuk, J., Saunders, T., Davidson, L., Ross, R., *et al.* (2009). Age-related changes in total and regional fat distribution. *Ageing research reviews*, 8(4): 339.

Kyle, U., Bosaeus, I., De Lorenzo, A., Deurenberg, P., Elia, M., Heitmann, B., Kent-Smith, L., Melchior, J., Pirlich, M., Scharfetter, H., *et al.* (2004). Bioelectrical impedance analysis–part I: review of principles and methods. *Clinical Nutrition*, 23(5): 1226–1243.

Kyle, U., Genton, L., Hans, D., Karsegard, L., Slosman, D., Pichard, C., *et al.* (2001). Age-related differences in fat-free mass, skeletal muscle, body cell mass and fat mass between 18 and 94 years. *European journal of clinical nutrition*, 55(8): 663–672.

Lancet, P. S. (2009). Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *The Lancet*, 373(9669): 1083–1096.

Larrañaga, P., Poza, M., Yurramendi, Y., Murga, R. H., et Kuijpers, C. M. H. (1996). Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(9): 912–926.

Lauritzen, S. L. et Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224.

Leonard, C. M., Roza, M. A., Barr, R. D., et Webber, C. E. (2009). Reproducibility of DXA measurements of bone mineral density and body composition in children. *Pediatric radiology*, 39(2): 148–154.

Leray, P. (2006). *Réseaux bayésiens : apprentissage et modélisation de systèmes complexes*. PhD thesis, Université de Rouen.

Levitt, D., Heymsfield, S., Pierson Jr, R., Shapses, S., et Kral, J. (2007). Physiological models of body composition and human obesity. *Nutrition & Metabolism*, 4: 19–32.

Lewis, C., Smith, D., Wallace, D., Williams, O., Bild, D., et Jacobs Jr, D. (1997). Seven-year trends in body weight and associations with lifestyle and behavioral characteristics in black and white young adults: the CARDIA study. *American Journal of Public Health*, 87(4): 635–642.

Lin, C. et Wang, S. (2002). Fuzzy support vector machines. *Neural Networks, IEEE Transactions on*, 13(2): 464–471.

Lohman, T. et Chen, Z. (2005). *Human body composition*, chapter Dual-Energy X-Ray Absorptiometry, pages 63–77. Human Kinetics, 2nd edition.

Lunn, D., Jackson, C., Best, N., Thomas, A., et Spiegelhalter, D. (2013). *The* BUGS *Book. A practical introduction to Bayesian analysis*. CRC press.

Mangasarian, O. L. et Musicant, D. R. (2000). Robust linear and support vector regression. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(9): 950–955.

Margaritis, D. (2003). *Learning Bayesian network model structure from data*. PhD thesis, University of Pittsburgh.

Margaritis, D. et Thrun, S. (1999). Bayesian network induction via local neighborhoods. In *In Proceedings of the Neural Information Processing Systems 12*, pages 505–511.

McLaren, L. (2007). Socioeconomic status and obesity. *Epidemiologic reviews*, 29(1): 29–48.

Miller, A. J. (2002). *Subset selection in regression*. Boca Raton: Chapman & Hall / CRC, 2d edition.

Mioche, L., Bidot, C., et Denis, J. (2011a). Body composition predicted with a Bayesian network from simple variables. *British Journal of Nutrition*, 105: 1265–1271.

Mioche, L., Brigand, A., Bidot, C., et Denis, J. (2011b). Fat-Free Mass Predictions through a Bayesian Network Enable Body Composition Comparisons in Various Populations. *The Journal of Nutrition*, 141: 1573–1580.

Moore, A. et Schneider, J. (2002). Real-valued all-dimensions search: Low-overhead rapid searching over subsets of attributes. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 360–369. Morgan Kaufmann Publishers Inc.

Moore, A. et Wong, W.-K. (2003). Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. In *ICML*, volume 3, pages 552–559.

Morgan, S. (2010). Adjustment of age-related height decline for Chinese–a 'natural experiment'longitudinal survey using archival data,'. In *Economic History Society Annual Conference, University of Durham*, pages 26–28.

Murphy, K. P. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3): 691–692.

Neapolitan, R. (2004). *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ.

Nielsen, J. D., Kočka, T., et Pena, J. M. (2002). On local optima in learning Bayesian networks. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 435–442.

Obisesan, T. O., Aliyu, M. H., Bond, V., Adams, R. G., Akomolafe, A., et Rotimi, C. N. (2005). Ethnic and age-related fat free mass loss in older Americans: the Third National Health and Nutrition Examination Survey (NHANES III). *BMC Public Health*, 5(1): 41.

Okosun, I., Liao, Y., Rotimi, C., Prewitt, T., et Cooper, R. (2000). Abdominal adiposity and clustering of multiple metabolic syndrome in White, Black and Hispanic Americans. *Annals of epidemiology*, 10(5): 263–270.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Peña, J. M., Nilsson, R., Björkegren, J., et Tegnér, J. (2007). Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45(2): 211–232.

Portier, K., Keith Tolson, J., et Roberts, S. (2007). Body weight distributions for risk assessment. *Risk analysis*, 27(1): 11–26.

Pouliot, M.-C., Després, J.-P., Lemieux, S., Moorjani, S., Bouchard, C., Tremblay, A., Nadeau, A., et Lupien, P. J. (1994). Waist circumference and abdominal sagittal diameter: best simple anthropometric indexes of abdominal visceral adipose tissue accumulation and related cardiovascular risk in men and women. *The American journal of cardiology*, 73(7): 460–468.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raguso, C. A., Kyle, U., Kossovsky, M. P., Roynette, C., Paoloni-Giacobino, A., Hans, D., Genton, L., et Pichard, C. (2006). A 3-year longitudinal study on body composition changes in the elderly: role of physical exercise. *Clinical Nutrition*, 25(4): 573–580.

Reeves, G. K., Pirie, K., Beral, V., Green, J., Spencer, E., et Bull, D. (2007). Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. *BMJ: British Medical Journal*, 335(7630): 1134.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5): 465–471.

Roubenoff, R., Kehayias, J., Dawson-Hughes, B., et Heymsfield, S. (1993). Use of dual-energy x-ray absorptiometry in body-composition studies: not yet a "gold standard". *American Journal of Clinical Nutrition*, 58(5): 589–591.

Russell, S. et Norvig, P. (2009). *Artificial Intelligence: A Modern Approach, 3rd edition*. Prentice Hall.

Schölkopf, B., Smola, A. J., Williamson, R. C., et Bartlett, P. L. (2000). New support vector algorithms. *Neural computation*, 12(5): 1207–1245.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3): 1–22.

Sedgewick, R. (2011). *Algorithms*. Addison-Wesley, 4th edition.

Singh, M. et Valtorta, M. (1993). An algorithm for the construction of Bayesian network structures from data. pages 259–265.

Smola, A. et Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3): 199–222.

Snijder, M., Van Dam, R., Visser, M., et Seidell, J. (2006). What aspects of body fat are particularly hazardous and how do we measure them? *International Journal of Epidemiology*, 35(1): 83.

Snijder, M. B., Dekker, J. M., Visser, M., Bouter, L. M., Stehouwer, C. D., Yudkin, J. S., Heine, R. J., Nijpels, G., et Seidell, J. C. (2004). Trunk fat and leg fat have independent and opposite associations with fasting and postload glucose levels the Hoorn study. *Diabetes care*, 27(2): 372–377.

Sobal, J. et Stunkard, A. J. (1989). Socioeconomic status and obesity: a review of the literature. *Psychological bulletin*, 105(2): 260.

Sorkin, J., Muller, D., et Andres, R. (1999). Longitudinal change in height of men and women: implications for interpretation of the body mass index: the Baltimore Longitudinal Study of Aging. *American Journal of Epidemiology*, 150(9): 969–977.

Spirtes, P., Glymour, C. N., et Scheines, R. (2000). Causation Prediction & Search 2e. 81.

Srinivasaraghavan, J. et Allada, V. (2006). Application of Mahalanobis distance as a lean assessment metric. *The International Journal of Advanced Manufacturing Technology*, 29(11): 1159–1168.

Steckel, R. (1995). Stature and the Standard of Living. *Journal of Economic Literature*, 33(4): 1903–1940.

Stevens, J., Katz, E., et Huxley, R. (2010). Associations between gender, age and waist circumference. *European journal of clinical nutrition*, 64(1): 6–15.

Steyerberg, E. W., Harrell Jr, F. E., Borsboom, G. J., Eijkemans, M., Vergouwe, Y., et Habbema, J. D. F. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, 54(8): 774–781.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147.

Sun, S. et Chumlea, W. (2005). *Human body composition*, chapter Statistical methods, pages 151–160. Human Kinetics, 2nd edition.

Thissen, U., Pepers, M., Ustun, B., Melssen, W., et Buydens, L. (2004). Comparing support vector machines to PLS for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems*, 73(2): 169–179.

Tian, S., Mioche, L., Denis, J.-B., et Morio, B. (2013). A multivariate model for predicting segmental body composition. *British Journal of Nutrition*, pages 1–11.

Timm, N. (2002). *Applied Multivariate Analysis*. Springer.

Tsamardinos, I., Aliferis, C., et Statnikov, A. (2003a). Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 673–678. ACM.

Tsamardinos, I., Aliferis, C., Statnikov, A., et Statnikov, E. (2003b). Algorithms for large scale Markov blanket discovery. In *Proceedings of the 16th international Florida artificial intelligence research society conference, AAA press*, pages 376–381.

Tsamardinos, I., Brown, L., et Aliferis, C. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1): 31–78.

Van Pelt, R., Evans, E., Schechtman, K., Ehsani, A., et Kohrt, W. (2002). Contributions of total and regional fat mass to risk for cardiovascular disease in older women. *American Journal of Physiology-Endocrinology and Metabolism*, 282(5): 1023–1028.

VanItallie, T., Yang, M.-U., Heymsfield, S., Funk, R., et Boileau, R. (1990). Height-normalized indices of the body's fat-free mass and fat mass: potentially useful indicators of nutritional status. *The American journal of clinical nutrition*, 52(6): 953–959.

Vapnik, V. (1998). *Statistical learning theory*. John Wiley & Sons.

Vapnik, V. (2000). *The nature of statistical learning theory; Second edition*. Springer.

Vapnik, V., Golowich, S. E., et Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*, pages 281–287.

Verma, T. et Pearl, J. (1991). Equivalence and synthesis of causal models. *Uncertainty in Artificial Intelligence*, 6(6): 255–268.

Visser, M., Pahor, M., Tylavsky, F., Kritchevsky, S. B., Cauley, J. A., Newman, A. B., Blunt, B. A., et Harris, T. B. (2003). One-and two-year change in body composition as measured by DXA in a population-based cohort of older men and women. *Journal of Applied Physiology*, 94(6): 2368–2374.

Wang, H.-F. et Tsaur, R.-C. (2000). Insight of a fuzzy regression model. *Fuzzy Sets and Systems*, 112(3): 355–369.

Wasserman, L. (2004). *All of statistics: a concise course in statistical inference*. Springer Verlag.

Wei, M., Gaskill, S. P., Haffner, S. M., et Stern, M. P. (1997). Waist Circumference as the Best Predictor of Noninsulin Dependent Diabetes Mellitus (NIDDM) Compared to Body Mass Index, Waist/hip Ratio and Other Anthropometric Measurements in Mexican AmericansA 7-Year Prospective Study. *Obesity research*, 5(1): 16–23.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Winkleby, M. A., Jatulis, D. E., Frank, E., et Fortmann, S. P. (1992). Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *American journal of public health*, 82(6): 816–820.

Wong, M. L., Lee, S. Y., et Leung, K. S. (2004). Data mining of Bayesian networks using cooperative coevolution. *Decision Support Systems*, 38(3): 451–472.

Wu, C.-H., Heshka, S., Wang, J., Pierson, R., Heymsfield, S., Laferrere, B., Wang, Z., Albu, J., Pi-Sunyer, X., et Gallagher, D. (2007). Truncal fat in relation to total body fat: influences of age, sex, ethnicity and fatness. *International Journal of Obesity*, 31(9): 1384–1391.

Yang, H. et Na, M. (2008). Fuzzy support vector regression model for the calculation of the collapse moment for wall-thinned pipes. *Nuclear engineering and technology*, 40(7): 607–614.

Yaramakala, S. et Margaritis, D. (2005). Speculative Markov blanket discovery for optimal feature selection. In *ICDM '05: Proceedings of the 5th IEEE International Conference on Data Mining*, pages 809–812. IEEE.

Yen, I. H. et Moss, N. (1999). Unbundling education: a critical discussion of what education confers and how it lowers risk for disease and death. *Annals of the New York Academy of Sciences*, 896(1): 350–351.