

A Bayesian Network for Predicting Body Composition of Human Body

Oloyede I.

Department of Statistics, University of Ilorin, Ilorin. Nigeria.

Corresponding Email: oloyede.i@unilorin.edu.ng

Abstract

This study investigates the predictive body composition adopting age, height, weight, waist circumference and body mass index as the covariates to predict three components that form the body weight: bone, fat and lean. Both saturated Bayesian network and convenient Bayesian network were compared using absolute bias, standard deviation and root mean square error to ascertain consistency and efficiency of the two types of Bayesian networks. The Directed Acyclic Graph were used to compare the two networks; nine different samples sizes were constructed using uniform distribution for both dependent variables and covariates. The study observed that saturated Bayesian network outperformed convenient Bayesian network both at small and medium sample sizes whereas the two networks estimators converged at higher sample sizes yielding the same standard deviation and root mean squares error. The study therefore recommends that saturated Bayesian network should be used at small sample sizes while any of the two types could be used at higher sample sizes.

Keywords: DAGs, Saturated Bayesian Network, Convenient Bayesian Network

Introduction

A probabilistic graphical model is known as belief network or Bayesian network which is a joint probability distribution over a set of variables and the corresponding local univariate distributions. It actually comes from the recursive use of Bayes theorem to decompose the joint distribution to individual distribution of the nodes given the direct acyclic graph. It denotes knowledge about an uncertain domain, each node in the graph represents a random variable whereas the edge between the nodes

represents probabilistic dependence among the corresponding random variables (Ruggen 2007).

Recently, Bayesian network had gained popularity amongst researchers; it had been used in many applications such as machine learning, text mining, weather forecasting, cellular networks, medical diagnosis, natural language processor, speech recognition, signal processing, and error control codes (Ben-gal 2007). Ben-gal 2007 further clarified that combination of Bayesian statistics and Bayesian networks prevent over fitting. This is because there is misconception about Bayesian networks as it is different from Bayesian statistics. The nodes in Bayesian networks denotes not only random variables but also beliefs, hypothesis and latent, manifest variables (Murphy, 1998 and Ben-gal, 2007). The structure obtained in Bayesian networks is ideal for combining prior knowledge with the observed data. The beauty of Bayesian network is that it can be used even when there exist missing data (David, 1999). Bayesian networks captured probability information from data using directed acyclic graphs and factorised probability function (Rafeal *et.al*, 2004). It provides a model that is associated with a large number of variables (Pearl, 1998 and Castino *et al*, 1997).

Bayesian network methods

Several types are involved, chiefly amongst which are multinomial Bayesian network where variables are assumed to be discrete such that each variable has a finite set of possible values; Gaussian Bayesian network where variables are assumed to have a multivariate normal distribution $N(\mu, \Sigma)$ thus assumed to be continuous; and hybrid Bayesian network where variables are assumed to be both the discrete and continuous.

In this study Gaussian Bayesian network (GBN) is adopted whose joint density is expressed as: $f(x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x - \mu)' \Sigma^{-1}(x - \mu)\}$

Thus, GBN has probability density function by $f(x_i/\pi_i) \sim N(\mu_i + \sum_{j=1}^{i-1} \beta_{ij}(x_j - \mu_j))^2$. Thus $Y = X\Theta + E$ is the model of choice with $V(\text{vec}(E)) = I_n \otimes \Sigma$

Data Analysis and Discussion

Data were splitted into training and validation sets; the training set was used to learn or calibrate the models while the validation set was used to select the model that performs the prediction of new items well, this is an attempt to minimise overfitting (Hastie et.al, 2009).

Table 1: Saturated and Convenience Bayesian Networks at sample size 16

Sample 16	Saturated BNs			Convenient BNs		
	$ Bias $	Sd.Dev	SEP	$ Bias $	Sd.Dev	SEP
TF	9.54	0.5	9.56	9.54	9.28	14.12
LF	3.67	0.32	3.7	3.67	5.05	6.51
AF	2.1	0	2.1	2.1	0.08	2.1
TL	6.71	0.11	6.71	6.71	1.92	7.15
LL	5.49	0.16	5.5	5.49	2.68	6.4
AL	2.12	0.04	2.12	2.12	0.73	2.31
TB	0.38	0.01	0.38	0.38	0.23	0.46
LB	0.26	0.02	0.26	0.26	0.35	0.45
AB	0.08	0.01	0.08	0.08	0.17	0.2
Global Score	30.35	1.19	30.42	30.35	20.49	39.7

The Bias that emanated from the Saturated Bayesian network is 30.35, the standard deviation is 1.19, the standard error of prediction is 30.42 and the variance is 1.4161. On the other hand, that of the Convenient Bayesian network's bias, standard deviation, standard error of prediction and variance are 30.35, 20.49, 39.7 and 419.8401 respectively. In terms of efficiency, saturated Bayesian network outperformed convenient Bayesian network. For all the covariates the study deduced that saturated BN has minimum standard error thereby indicating efficient estimator.

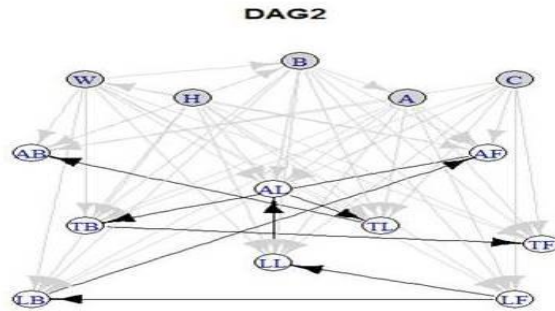


Figure 1: Direct Acyclic Graph at Sample Size 16

The weight, height, BMI, age and waist circumference influence the lean mass of the arm (AL), the lean mass of the trunk (TL), the lean mass of the leg (LL), the fat level of the trunk (TF), the fat level of the leg (LF), the bone of the leg (LB), the bone of the trunk (TB) and the fat level of the arm (AF). The weight, height and waist circumference are having influence on the bone of the arm (AB). The lean mass of the trunk, arm and leg (TL, AL and LL) influence the bone of the arm (AB), lean mass of the trunk (TL) and the lean mass of the arm (AL) respectively. The fat level of the arm and the leg (AF and LF) are having influence on the bone of the trunk, and the lean mass of the leg and the bone of the leg (TB, LL and LB) respectively. The bone of the trunk (TB) is having influence on the fat of the trunk (TF).

Table 2: Saturated and Convenience Bayesian Networks at sample size 25

Sample 25	Saturated BNs			Convenient BNs		
	Bias	Sd.Dev	SEP	Bias	Sd.Dev	SEP
TF	6.05	0.96	6.27	6.05	7.63	10.49
LF	5.45	0.43	5.49	5.45	3.92	7.16
AF	2.18	0.09	2.18	2.18	0.78	2.35
TL	3.46	0.78	3.58	3.46	6.57	7.63
LL	6.29	0.67	6.41	6.29	5.38	8.88
AL	2.83	0.11	2.84	2.83	0.91	3.05
TB	0.22	0.03	0.23	0.22	0.32	0.40
LB	0.37	0.03	0.38	0.37	0.28	0.49
AB	0.10	0.01	0.10	0.10	0.13	0.17
Global Score	26.95	3.12	27.46	26.95	25.92	40.63

The Bias that emanated from the Saturated Bayesian network is 26.95, the standard deviation is 3.12, the standard error of prediction is 27.46 and the variance is 9.7344, while that of the Convenient Bayesian network's bias, standard deviation, standard error of prediction and variance are 26.95, 25.92, 40.63 and 671.8464 respectively. The study inferred that saturated BN outperformed Convenient BN at sample size 25.

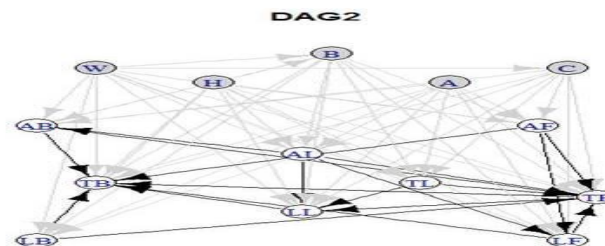


Figure 2: Direct acyclic graph at sample size 25

The weight, height, BMI, age and waist circumference influence the lean mass of the arm (AL), the lean mass of the trunk (TL), the lean mass of the leg (LL), the fat level of the trunk (TF), the fat level of the leg (LF), the bone of the leg (LB) and the bone of the trunk (TB). The weight, BMI, age and waist circumference are influencing the fat level of the arm (AF). The weight, height and waist circumference are having influence on the bone of the arm (AB).

Table 3: Saturated and Convenience Bayesian Networks at sample size 30

Sample 30	Saturated BNs			Convenient BNs		
	Bias	Sd.Dev	SEP	Bias	Sd.Dev	SEP
TF	7.49	5.65	9.91	7.49	5.65	9.91
LF	4.78	3.51	6.28	4.78	3.51	6.28
AF	1.15	1.25	1.84	1.15	1.25	1.84
TL	6.90	6.33	9.92	6.90	6.78	10.25
LL	6.26	5.36	8.62	6.26	5.36	8.62
AL	2.98	2.59	4.14	2.98	2.59	4.14
TB	0.13	0.28	0.32	0.13	0.29	0.33
LB	0.27	0.29	0.41	0.27	0.36	0.47
AB	0.12	0.10	0.16	0.12	0.13	0.18
Gross Score	30.08	25.36	41.62	30.08	25.91	42.02

The Bias that emanated from the Saturated Bayesian network is 30.08, the standard deviation is 25.36, the standard error of prediction is 41.62 and the variance is 643.1296 while that of the Convenient Bayesian network's bias, standard deviation, standard error of prediction and variance are 30.08, 25.91, 42.02 and 671.3281 respectively. The Saturated BN outperformed Convenient BN as explained in sample size 16.

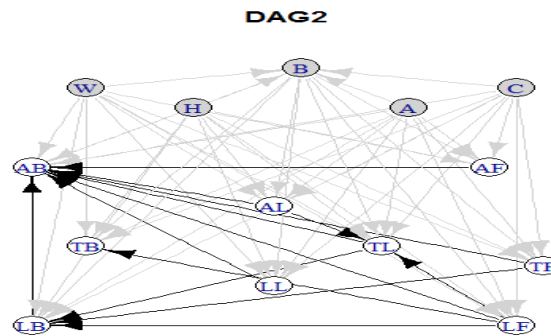


Figure 3: Direct Acyclic Graph at Sample Size 30

The weight, BMI, age and waist circumference are influencing the fat level of the arm (AF), the weight, height and waist circumference are having influence on the bone of the arm (AB). The weight, height, BMI, age and waist circumference influence the lean mass of the arm (AL), the lean mass of the trunk (TL), the lean mass of the leg (LL), the fat level of the trunk (TF), the fat level of the leg (LF), the bone of the leg (LB), the bone of the trunk (TB). Both the lean mass of the arm (AL) and the fat level of the leg (LF) are influencing the lean mass of the trunk (TL), the lean mass of the trunk (TL), fat level of the trunk (TF) and the fat level of the leg (LF) influence the bone of the leg (LB). The fat level of the arm (AF), lean mass of the arm (AL), lean mass of the trunk (TL), fat level of the trunk (TF), fat level of the leg (LF) and the lean mass of the leg (LL) are having influence on the bone of the arm (AB).

Table 4: Saturated and Convenience Bayesian Networks at sample size 50

Sample 50	Saturated BNs			Convenient BNs		
	Bias	Sd.Dev	SEP	Bias	Sd.Dev	SEP
TF	6.25	6.72	9.63	6.25	6.72	9.63
LF	2.74	4.97	5.93	2.74	4.97	5.93
AF	1.46	1.46	2.16	1.46	1.50	2.19
TL	5.49	7.16	9.52	5.49	7.16	9.52
LL	5.48	6.05	8.61	5.48	6.18	8.71
AL	1.92	2.51	3.33	1.92	2.51	3.33
TB	0.17	0.18	0.26	0.17	0.18	0.26
LB	0.21	0.24	0.33	0.21	0.25	0.34
AB	0.09	0.08	0.14	0.09	0.09	0.14
Global Score	23.80	29.38	39.90	23.80	29.56	40.0

The Bias that emanated from the Saturated Bayesian network is 23.80, the standard deviation is 29.38, the standard error of prediction is 39.90 and the variance is 863.1844 while that of the Convenient Bayesian network's bias, standard deviation, standard error of prediction and variance are 23.80, 29.56, 40.0 and 873.7936 respectively.

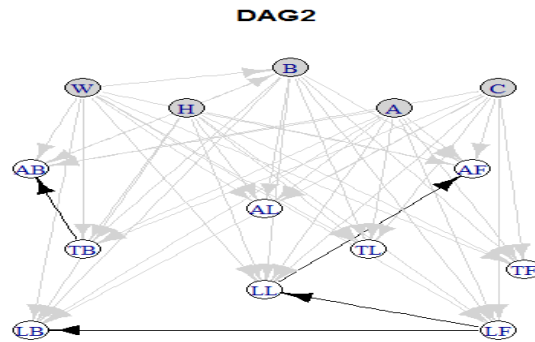


Figure 4: Direct Acyclic Graph at Sample Size 50

The weight, height and waist circumference are influencing the bone of the arm. The weight, BMI, age and waist circumference are having influence on the fat level of the arm. The weight, height, BMI, age and waist circumference influence the bone mass of the trunk (TB), the lean mass of the arm (AL), the lean mass of the trunk (TL), the fat level of the trunk (TF), the fat level of the leg (LF), influence the lean mass of the leg (LL), the bone of the leg (LB). The fat level of the leg (LF) is at the same time having influence on the lean mass of the leg (LL) and the bone of the leg (LB).

Table 5 Showing Saturated and Convenience Bayesian Networks at Sample Size 74

Sample 74	Saturated BNs			Convenient BNs		
	Bias	Sd.Dev	SEP	Bias	Sd.Dev	SEP
TF	4.80	6.70	8.63	4.80	6.70	8.63
LF	3.57	3.88	5.54	3.57	3.88	5.54
AF	1.03	1.15	1.62	1.03	1.15	1.62
TL	5.47	7.17	9.46	5.47	7.17	9.46
LL	5.48	5.11	7.75	5.48	5.11	7.75
AL	1.88	2.24	3.09	1.88	2.28	3.11
TB	0.18	0.18	0.27	0.18	0.19	0.28
LB	0.26	0.28	0.40	0.26	0.28	0.40
AB	0.09	0.09	0.14	0.09	0.09	0.14
Global Score	22.77	26.81	36.89	22.77	26.86	36.92

The Bias that emanated from the Saturated Bayesian network is 22.77, the standard deviation is 26.81, the standard error of prediction is 36.89 and the variance is 718.7761 while that of the Convenient Bayesian network's bias, standard deviation, standard error of prediction and variance are

22.77, 26.86, 36.92 and 721.4596 respectively.

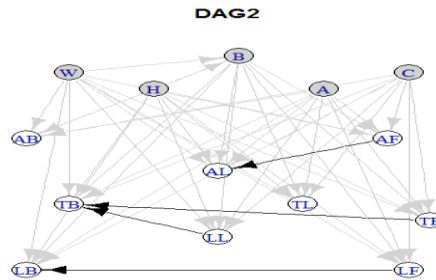


Figure 5: Direct Acyclic Graph at Sample Size 74

The weight, height and waist circumference have direct influence on the bone of the arm (AB). The weight, BMI, waist circumference and age have influence on the fat level of the arm (AF). The weight, height, BMI, age and waist circumference influence the bone of the trunk (TB), the lean mass of the arm (AL), the lean mass of the leg (LL), the fat level of the trunk (TF), the lean mass of the trunk (TL), the bone of the leg (LB) and the fat level of the leg (LF). The lean mass of the leg (LL) and the fat level of the trunk (TF) are having influence on the bone of the trunk (TB), the fat level of the leg (LF) is influencing the bone of the leg (LB).

Table 6: Saturated and Convenience Bayesian Networks at Sample Size 100

Samp100	Bias	Sd.Dev	SEP	Bias	Sd.Dev	SEP
TF	6.20	6.75	9.60	6.20	6.75	9.60
LF	3.27	4.11	5.53	3.27	4.11	5.53
AF	1.05	1.34	1.78	1.05	1.34	1.78
TL	5.62	7.15	9.62	5.62	7.15	9.62
LL	5.42	4.92	7.71	5.42	4.92	7.71
AL	1.89	2.41	3.23	1.89	2.41	3.23
TB	0.19	0.21	0.30	0.19	0.21	0.30
LB	0.23	0.26	0.37	0.23	0.26	0.37
AB	0.09	0.12	0.15	0.09	0.12	0.15
Global Score	23.95	27.27	38.28	23.95	27.27	38.28

The Bias that emanated from the Saturated Bayesian network is 23.95, the standard deviation is 27.27, the standard error of prediction is 38.28 and the variance is 743.6529 while that of the Convenient Bayesian network's bias, standard deviation, standard error of prediction and variance are 23.95, 27.27, 38.28 and 743.6529 respectively.

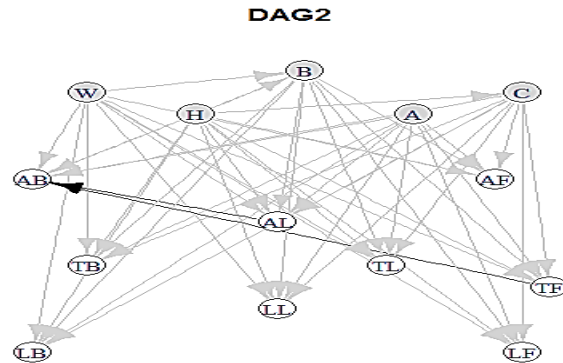


Figure 6: Direct acyclic graph at sample size 100

The weight, height and waist circumference influence the bone of the arm (AB). The weight, height, BMI, age and waist circumference influence the lean mass of the arm (AL), the lean mass of the trunk (TL), the fat level of the trunk (TF), the fat level of the leg (LF), the lean mass of the leg (LL), the bone of the leg (LB), the bone of the trunk (TB), the fat level of the arm (AF). Both the lean mass of the arm (AL) and the fat level of the trunk (TF) influence the bone of the arm (AB).

Table 7: Saturated and Convenience Bayesian Networks at sample size 200

Sampl200	 Bias 	Sd.Dev	SEP	 Bias 	Sd.Dev	SEP
TF	5.29	6.65	8.87	5.29	6.65	8.87
LF	3.33	3.99	5.46	3.33	3.99	5.46
AF	1.15	1.32	1.83	1.15	1.32	1.83
TL	6.15	6.62	9.44	6.15	6.62	9.44
LL	5.35	5.52	8.05	5.35	5.55	8.07
AL	1.95	2.53	3.34	1.95	2.53	3.34
TB	0.17	0.19	0.27	0.17	0.19	0.27
LB	0.22	0.28	0.38	0.22	0.28	0.38
AB	0.09	0.10	0.14	0.09	0.10	0.14
GlobalScore	23.71	27.21	37.76	23.71	27.24	37.78

The Bias that emanated from the Saturated Bayesian network is 23.71, the standard deviation is 27.21, the standard error of prediction is 37.76 and the variance is 740.3841. That of the Convenient Bayesian network's bias, standard deviation, standard error of prediction and variance are 23.71, 27.24, 37.78 and 742.0176 respectively.

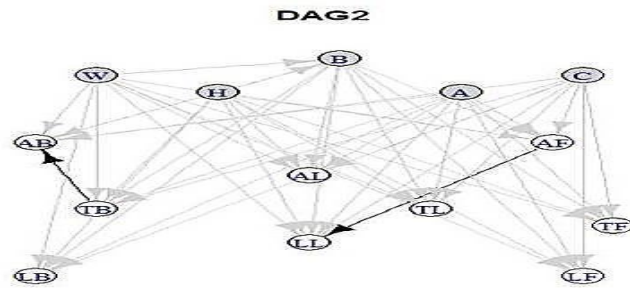


Figure 7: Direct Acyclic Graph at Sample Size 200

The weight, height and the waist circumference influence the bone of the arm (AB). The weight, height, BMI, waist circumference and age directly influence the bone of the trunk (TB) and the bone of the trunk (AL). The height, BMI, waist circumference and age directly influence the fat of the arm (AF). The weight, height, BMI, waist circumference and age directly influence the lean mass of the bone (LB), the lean mass of the leg (LL), directly influence the fat of the leg (LF), the lean mass of the trunk (TL), the fat level of the trunk (TF). The bone of the trunk (TB) influences the bone of the arm (AB) and fat of the arm (AF) influences the lean mass of the leg (LL).

Table 8: Saturated and Convenience Bayesian Networks at sample size 500

Samp500	 Bias 	Sd.Dev	SEP	 Bias 	Sd.Dev	SEP
TF	5.53	6.19	8.67	5.53	6.19	8.67
LF	3.43	3.96	5.49	3.43	3.96	5.49
AF	1.19	1.35	1.88	1.19	1.35	1.88
TL	6.16	6.87	9.63	6.16	6.87	9.63
LL	4.96	5.66	7.86	4.96	5.66	7.86
AL	2.11	2.45	3.36	2.11	2.45	3.36
TB	0.18	0.21	0.29	0.18	0.21	0.29
LB	0.23	0.28	0.38	0.23	0.28	0.38
AB	0.09	0.10	0.14	0.09	0.10	0.14
Global Scores	23.88	27.07	37.72	23.88	27.07	37.72

The Bias that emanated from the Saturated Bayesian network is 23.88, the standard deviation is 27.07, the standard error of prediction is 37.72 and the variance is 732.7849 while that of the Convenient Bayesian network's

bias, standard deviation, standard error of prediction and variance are 23.88, 27.07, 37.72 and 732.7849 respectively.

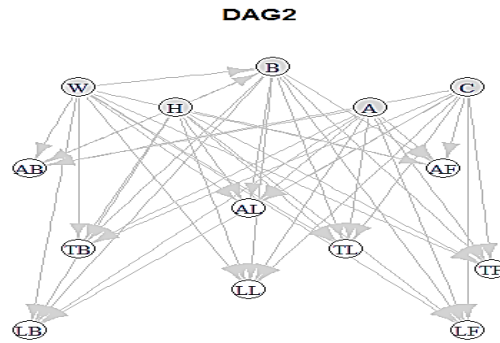


Figure 8: Direct Acyclic Graph at Sample Size 500

The weight, height, age and waist circumference influence the bone of the arm (AB). The weight, height, BMI, waist circumference and age directly influence the bone of the trunk (TB), the lean mass of the trunk (TL), the fat level of the trunk (TF), directly influence the bone of the leg (LB), directly influence the fat level of the leg (LF), influence the lean mass of the leg (LL), the fat of the arm (AF) and directly influence the lean of the trunk (AL).

Table 9: Saturated and Convenience Bayesian Networks at Sample Size 1000

Samp1000	Bias	Sd.Dev	SEP	Bias	Sd.Dev	SEP
TF	5.42	6.21	8.62	5.42	6.21	8.62
LF	3.4	3.98	5.46	3.4	3.98	5.46
AF	1.18	1.3	1.83	1.18	1.3	1.83
TL	5.8	6.85	9.39	5.8	6.85	9.39
LL	4.8	5.68	7.75	4.8	5.68	7.75
AL	2.01	2.38	3.27	2.01	2.38	3.27
TB	0.18	0.2	0.28	0.18	0.2	0.28
LB	0.23	0.26	0.36	0.23	0.26	0.36
AB	0.08	0.1	0.14	0.08	0.1	0.14
Global score	23.1	27	37.1	23.1	27	37.1

The Bias that emanated from the Saturated Bayesian network is 23.1, the standard deviation is 27, the standard error of prediction is 37.1 and the variance is 729 while that of the Convenient Bayesian network's bias,

standard deviation, standard error of prediction and variance are 23.1, 27, 37.1 and 729 respectively.

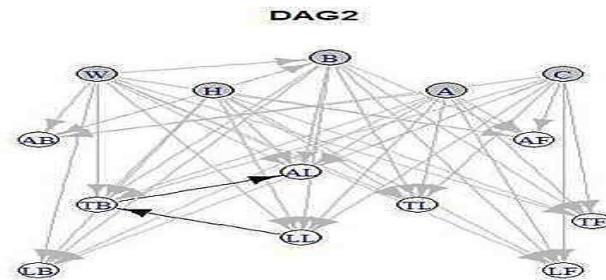


Figure 9: Direct Acyclic Graph at Sample Size 1000

The Weight, Height and Age are influencing the bone of the arm (AB). The Weight, Height, BMI and Age are having influence on the bone of the trunk (TB). The Weight, Height, BMI, Age and Waist circumference are having direct influence on the lean mass of the arm (AL), the lean mass of the trunk (TL), the fat level of the trunk (TF), the fat level of the leg (LF), the lean mass of the leg (LL), on the bone of the leg (LB), the fat level of the arm (AF). And because of the flexibility of the DAG, the covariates also influence one another, that is why the DAG above shows that the lean mass of the leg (LL) influences the bone of the trunk (TB) and that influences the lean mass of the arm (AL). The weight and height also influence the BMI.

Conclusion

By comparison, the estimate of the bias that emanated from the saturated Bayesian network when the sample is 74 is the minimum of all the sample categories and that makes it the most consistent while the standard deviation emanating from the saturated Bayesian network when the sample is 16, is said to make it the most efficient, being the minimum of all sample categories. The bias of the data at $n=16$ is the highest and the variance of the data at $n=50$ is the highest; these bring about the problem of underfitting and overfitting respectively. The study observed that saturated Bayesian network outperformed convenient Bayesian network both at small and medium sample sizes whereas the two networks estimators converged at higher sample sizes yielding the same standard deviation and root mean squares error. The study therefore recommends that saturated Bayesian network should be used at small sample sizes while any of the two types could be used at higher sample sizes.

References

- Ben-Gal, (2007). Bayesian Networks, in Ruggeri F., Faltin F. and Kenett R., Encyclopedia of Statistics in Quality and Reliability, Wiley and Sons, England.
- Castillo E., Gutierrez J.M, and Hadi A.S. (1997). Expert Systems and Probabilistic Network Models. Springer, New York.
- David, H. (1999). A tutorial on learning with Bayesian networks, in Learning in Graphical, M.J. Models, ed, MIT Press, Cambridge, also appears as Technical Report MSR-TR-95-06, Microsoft Research, March, 1995.
- Murphy, K. (1998). A brief introduction to graphical models and Bayesian networks. <http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html>.
- Pearl, J., (1998). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo.
- Rafael C., Carmen S., and Jose M. G. (2004). Applications of Bayesian Networks in Meteorology. In Advances in Bayesian Networks, Gámez et al. eds., 309-327, Springer.
- Ruggeri F., Faltin F. & Kenett R. (Eds.) (2007), Encyclopedia of Statistics in Quality and Reliability, John Wiley & Sons, England.