Journal of Statistical Theory and Practice (2021) 15:45
https://doi.org/10.1007/s42519-021-00167-2

ORIGINAL ARTICLE

Understanding the Sampling Bias: A Case Study on NBA Drafts

Polychronis Economou[1] · **Apostolos Batsidis**[2] · **George Tzavelas**[3] · **Sonia Malefaki**[4]

Accepted: 19 January 2021 / Published online: 23 March 2021

Abstract

In several real data applications a biased sample arises naturally from the selection procedure. Recently, Economou et al. (Biom J 62: 238–249, 2020) used the concept of bivariate weighted distributions and proposed four different families of weight functions to describe cases in which the bias in a bivariate sample is caused by adopting sampling schemes that result in over- or under-representation of individuals with specific properties in the sample. The current paper focuses on revealing the contribution of each variable to the bias in the bivariate sample. More specifically, under the Bayesian perspective, Approximate Bayesian Computation methods are used to sample approximately from the posterior distribution, and the Deviance Information Criterion is employed to compare the fit of the models obtained by using different weight functions. The proposed method is illustrated to a real data set concerning NBA draft players.

Keywords ABC rejection algorithm · Biased sample · Bias adjustment · Model comparison · Weighted distributions

✉ Polychronis Economou
peconom@upatras.gr

[1] Department of Civil Engineering, University of Patras, 265 00 Rion-Patras, Greece

[2] Department of Mathematics, University of Ioannina, 45 110 Ioannina, Greece

[3] Department of Statistics and Insurance Science, University of Piraeus, 80, M. Karaoli and A. Dimitriou St., 18534 Piraeus, Greece

[4] Department of Mechanical Engineering and Aeronautics, University of Patras, 265 00 Rion-Patras, Greece

1 Introduction

In statistics, sampling bias, also referred as sample selection bias, is a bias introduced in a sample by the used selection procedure. Under biased sampling schemes, individuals are selected in such a way that the observed sample cannot be considered as a representative sample of the population intended to be analyzed. This may happen either by applying unintentionally a non-random sampling scheme or due to the nature of the problem [1, 11, 19]. In any case, whenever certain members of the population are under- or over-represented in a sample, then a bias sample is obtained.

A characteristic example of a biased sample that arises naturally from the selection procedure is the NBA drafts. It is obvious that not all the basketball players who meet the formal requirements to be eligible for the NBA draft have the same odds to be indeed drafted, i.e., to be among the sixty players that consist the annual draft. In particular, players that excel, for instance, in strength or height are more likely to be included in the drafts than players that do not meet such high standards.

The use of a biased sample from a specific population may cause serious problems if treated as a random one from that population, since any statistic computed on the basis of a non-representative sample is systematically erroneous. This fact can lead to a systematic over- or under-estimation of the population parameters [29]. Moreover, a biased sampling scheme can be the reason for an observed false (positive or negative) correlation between two random variables (r.v.), say X and Y, that are either not correlated or correlated with a different direction. A classic example of the latter situation and the misleading results that may produce, was first noticed in 1946, in a case-control study linking diabetes with cholecystitis amongst inpatients who seek care. In [3] is illustrated that the two diseases were positive correlated although they are independent in the population. Berkson himself explained this spuriously finding by recognizing that a patient with more than one disease was more likely to be hospitalized than a patient with only a single disease. Since then, such a false observed correlation due to a biased sample, is known as Berkson's paradox or Berkson's bias or fallacy.

Since the ignorance of a bias leads to spurious findings in many fields, aspects of adjusting these findings have been considered by many authors. For instance, the so-called Bias Breaking Model was introduced in [9], while [20, 28] described the structure of the biases by using causal diagrams known as directed acyclic graphs. For more details on causal diagram theory and selection bias, see for instance [10, 12, 30]. Some other methods used to adjust selection bias are the poststratification [25] and the inverse probability weighting [24].

Recently, extending the basic ideas of [8, 21], the concept of the bivariate weighted distributions was used by Economou et al. [6] not only as a method of describing Berkson's paradox but also as an adjustment methodology applicable to many situations in which the recorded observations cannot be considered as a random sample from the parent distribution. In this frame four different families of weight functions were proposed to describe cases in which the bias in a bivariate sample is caused since certain members of the population are under- or over-represented in the sample. The selection of a specific weight function relies on the correct recognition of the sampling mechanism nature. The main purpose of the current work is to discuss how the weight functions proposed in Economou et al. [6] can be used to examine if the observed bias in a bivariate sample is caused either by both random variables or by one of them. Thus, the proposed method, which combines the concept of weighted distributions to represent bias, the Approximate Bayesian Computation (ABC) algorithm to approximately draw samples from Bayesian posteriors and the Deviance Information Criterion to compare the fit of different models, aims to reveal the contribution of each variable to the bias in the bivariate sample.

In this context, the rest of the paper is organized as follows. In Sect. 2, we recall the families of the weight functions introduced in Economou et al. [6]. These weight functions can formulate different sampling scenarios, in which certain members of the bivariate population are over-represented in the sample and at the same time some others are under-represented. Furthermore, we discuss how these weight functions can be used to describe cases in which the bias in a bivariate sample is caused not by both the random variables but by only one of them. Since in most cases the likelihood of the proposed model is complex, ABC methods, likelihood-free methods, are used for statistical inference in Sect. 3. Also, the Deviance Information Criterion (DIC), presented in [27] and further discussed in [4], is used for comparing the fit of the models obtained by using different weight functions. In Sect. 4, a real data application illustrates the proposed methodology using a bivariate sample from NBA draft players. Finally, some concluding remarks are given in the last section.

2 Weighted Distributions and Bias Adjustment Method

The concept of the bivariate weighted distributions was used in Economou et al. [6] not only as a method of describing Berkson's paradox but also as a bias adjustment method applicable to many situations in which the recorded observations cannot be considered as a random sample from the original distribution. Next, for completeness purposes, the definition of the bivariate weighted distributions is given (see for instance [26]).

Definition 1 Let (X, Y) be a two-dimensional random vector with joint probability density function (p.d.f.) $f(x,y;\theta)$, where θ is an unknown s-dimensional parameter, which belongs on a parameter space Θ, where $\Theta \subseteq R^s$ with $s \geq 1$. If the probability of selecting a population unit in a sample is proportional to a nonnegative weight function $w(x,\ y)$ i.e., $w : R^2 \to R^+$ with $E_f[w(X,Y;\theta)] < \infty$, the expectation of $w(X,Y;\theta)$ with respect to $f(x,y;\theta)$, then the observed biased sample from $(X,\ Y)$ can be interpreted as a random sample from a population with p.d.f.

$$f_w(x,y;\theta) = \frac{w(x,y;\theta)}{E_f[w(X,Y;\theta)]} f(x,y;\theta). \tag{1}$$

The two-dimensional random vector (X_w, Y_w) with joint p.d.f. $f_w(x, y;\theta)$ is called the vector of the weighted random variables corresponding to (X, Y), associated with $w(x, y;\theta)$. The p.d.f. $f_w(x, y;\theta)$ is called the bivariate weighted p.d.f. corresponding to $f(x, y;\theta)$, associated with w. Note that $E_f[w(X, Y;\theta)]$ serves as a normalizing constant for the bivariate weighted distribution.

Some properties of the bivariate and multivariate weighted distributions can be found, among others, in [2, 13, 17, 18].

The basic assumption behind the procedure presented in Economou et al. [6] is that a bivariate weighted distribution can be used to model situations in which some members of the population are over-represented and some other are under-represented in the sample causing significant impact in the population inference. Depending on which parts of the population are more likely to be included in the sample, four different families of weight functions were proposed in Economou et al. [6] to model such bivariate biased data and adjust the inference. Next, these four general cases, labeled as Cases 1–4, are briefly presented.

Case 1 corresponds to situations in which units with large values on X and/or large values on Y are more likely to be selected. This means that pairs with small values on X and small values on Y are under-represented in the sample. Case 2 corresponds to the non-random sampling under which units with small values on X and/or small values on Y are more likely to be observed, while in Case 3 (Case 4) pairs (x, y) with large (small) values on X and/or small (large) values on Y are more likely to be included in the sample. In this frame, Economou et al. [6] proposed the following reasonable and convenient functions to be used under the concept of weighted distributions in order to model Cases 1–4, respectively:

$$w_1(x, y;\theta, \gamma_X, \gamma_Y) = 1 - [1 - F_X(x;\theta_X)]^{\gamma_X}[1 - F_Y(y;\theta_Y)]^{\gamma_Y}, \tag{2}$$

$$w_2(x, y;\theta, \gamma_X, \gamma_Y) = 1 - [F_X(x;\theta_X)]^{\gamma_X}[F_Y(y;\theta_Y)]^{\gamma_Y}. \tag{3}$$

$$w_3(x, y;\theta, \gamma_X, \gamma_Y) = 1 - [1 - F_X(x;\theta_X)]^{\gamma_X}[F_Y(y;\theta_Y)]^{\gamma_Y}, \tag{4}$$

and

$$w_4(x, y;\theta, \gamma_X, \gamma_Y) = 1 - [F_X(x;\theta_X)]^{\gamma_X}[1 - F_Y(y;\theta_Y)]^{\gamma_Y}, \tag{5}$$

where θ_X and θ_Y are functions of θ, $\gamma = (\gamma_X, \gamma_Y)$ is a vector of extra parameters with $\gamma_X, \gamma_Y \geq 0$, while $F_X(x;\theta_X)$ and $F_Y(y;\theta_Y)$ are the cumulative distribution functions (c.d.f.) of X and Y, respectively. By letting γ_X, γ_Y to take strictly positive values the bias drives from both variables. Let us denote this model with weight function w_i, as $Model_{if}$ $(\mathcal{M}_{if})$, $i = 1, ..., 4$.

The choice of the weight function, in an application, is based on the nature of the problem and the characteristics inherited to the observed sample by each weight function by taking into account which values of the random variables are more likely to be observed in the sample.

2.1 The Role of the Parameters γ_X and γ_Y

Due to the form of the weight functions $w_i, i = 1, ..., 4$, the values of γ_X and γ_Y define the variable that drives the bias. More specifically, the severity of the bias depends on the values of the parameters γ_X and γ_Y. Figure 1 demonstrates the effect of the parameters γ_X, γ_Y on the severity of the bias under the weight function $w_1(x, y;\gamma_X, \gamma_Y)$ (Case 1). The weight function is plotted (the gray surface) using different values of γ_X and γ_Y, assuming an underlying bivariate normal distribution with mean vector $(\mu_X, \mu_Y) = (75, 35)$, standard deviations $\sigma_X = 4$, $\sigma_Y = 4.5$ and correlation coefficient $\rho = -0.1$. This means that we assume that $(X, Y)^t \sim N_2((75, 35)^t, \Sigma)$, with variance–covariance matrix

$$\Sigma = \begin{pmatrix} 16 & -1.8 \\ -1.8 & 20.25 \end{pmatrix}. \tag{6}$$

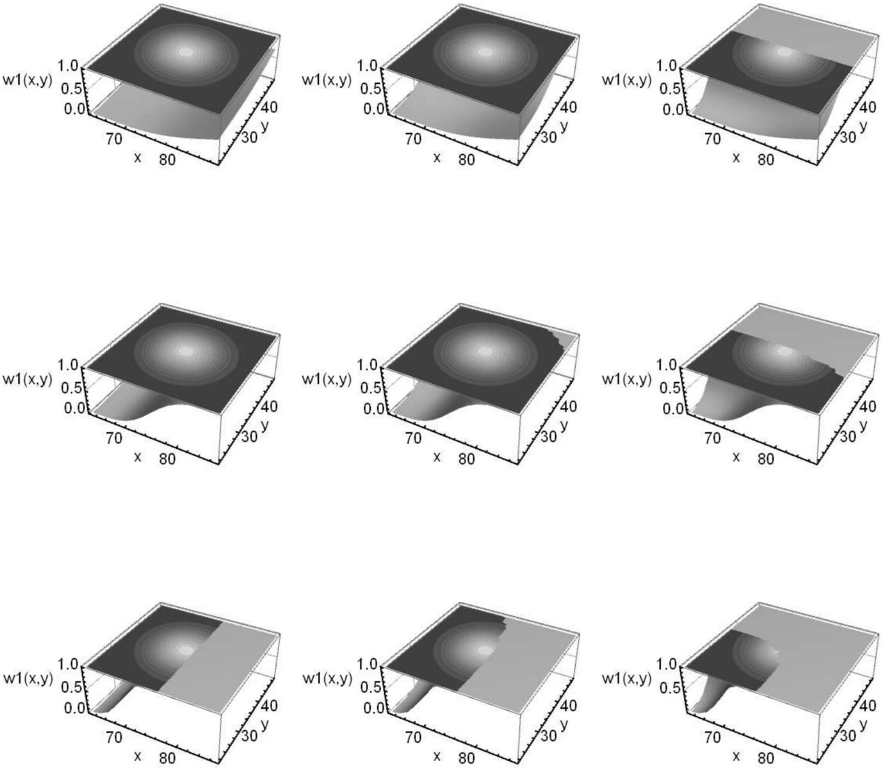


Fig. 1 The contour plot of a bivariate normal distribution with mean vector $(\mu_X, \mu_Y) = (75, 35)$ and variance–covariance matrix $\mathbf{\Sigma}$ defined in (6) along with the weight function $w_1(x, y;\gamma_X, \gamma_Y)$ (the gray surface) for different values of γ_X and γ_Y. The plots in the rows correspond to $w_1(x, y;\gamma_X, \gamma_Y)$ for γ_X equal to 0.1, 1 and 10, respectively, while the plots in the columns differ with respect to the parameter γ_Y, being 0.1, 1 and 10, respectively

The contour plots in Fig. 1 correspond to the aforementioned population distribution. The values 0.1, 1 and 10 were used for the parameters γ_X, γ_Y and each plot in Fig. 1 was made using a different combination of these values.

From the plots it is clear that as both parameters are relatively small (upper left plot) the weight function increases quite slowly making the ratio of the probability of a pair (x, y) with both large values of x and y to be observed over the probability of a pair (x, y) with both small values of x and y extremely large. At the same time as one of the parameters increases, like in the first row of Fig. 1 where the value of γ_X remains constant and equal to 0.1 while γ_Y increases, it seems that the bias in the sample is driven mainly by Y. Moreover, it seems that for large values of the parameter γ_Y—see upper right plot—a large area of the underlying bivariate normal distribution with mean vector $(\mu_X, \mu_Y) = (75, 35)$ and variance–covariance matrix $\mathbf{\Sigma}$ given in (6), is determined by an almost straight line, parallel to x-axis, in which every pair (x, y) is assigned the same probability to be selected in the sample. Similar remarks can be made whenever the parameter γ_Y remains constant and equal to 0.1, while γ_X increases (see the plots in the first column).

It is interesting to note that whenever one of the parameters or both of them increase, a larger area of the underlying p.d.f. of the population is assigned almost the same high probability to be selected in the sample—this is demonstrated by the gray area which is almost constant and equal to one for large values of X and/or Y. For example, at the bottom right plot, in which parameters γ_X and γ_Y are equal to 10, one can say that more of the three quarters of the population (defined by large values of X and/or Y) have the same probability of being observed. At the same time, the probability assigned to the part of the population with small values of X and Y becomes smaller and decreases rapidly as x or/and y decreases. This behavior results, in practice, in exclusion—or at least in heavy under-representation—of this part of the population in the sample.

Note that similar observations can be also made for the rest of the weight functions. Moreover, it is worth mentioning that the parameters γ_X and γ_Y define also two alternative models, obtained by setting either γ_X or γ_Y equal to zero. These models will be denoted as $\mathcal{M}_{iy}$ and $\mathcal{M}_{ix}$, respectively. The interpretation of $\mathcal{M}_{ix}$ ($\mathcal{M}_{iy}$) is that the bias in the observed bivariate sample is caused mainly, or even exclusively in the case of independent random variables, by $X(Y)$.

Remark 1 In practice, in order to study the bias in a bivariate sample, initially, the most suitable weight function $w_i, i = 1, \dots, 4$ has to be chosen, based on the prior information of the sampling procedure. Then, the three different models $\mathcal{M}_{ij}$, $j = f, x, y$ induced by this choice can be examined and compared by proper criteria to determine the most appropriate one.

Remark 2 The parameters γ_X and γ_Y cannot be both equal to zero since in this case all the weight functions correspond to the zero function which contradicts to the definition of the weighted distributions. On the other hand, if γ_X and/or γ_Y tend to infinity then all the proposed weight functions degenerate to the unit function, which correspond to a random sample.

3 ABC Rejection Algorithm and Model Comparison

The likelihood function of a biased bivariate sample $D = (x_j, y_j), j = 1, \dots, n$ from a parent population with known pdf $f(x, y;\theta)$ where θ is an unknown parameter vector, when the bias in the sample is described by the weight function $w_i(x, y;\theta, \gamma_X, \gamma_Y)$ can be written as the product

$$\prod_{j=1}^{n} f_{w_i}(x_j, y_j;\theta, \gamma_X, \gamma_Y) = \prod_{j=1}^{n} f_{w_i}(x_j, y_j;\zeta) = \frac{\prod_{j=1}^{n} w_i(x_j, y_j;\zeta) f(x_j, y_j;\theta)}{E_f^n[w(X, Y;\zeta)]},$$

where $\zeta = (\theta, \gamma_X, \gamma_Y)$. Note that due to the form of the weight function (i.e., involving the marginal cdfs and the exponent parameters γ_X, γ_Y) the normalizing constant $E_f^n[w(X, Y;\zeta)]$ contains all the unknown parameters making the evaluation of the likelihood function computationally costly or even infeasible. Thus, likelihood-free methods should be adopted to perform inference. Such a method is the ABC method. ABC is a very powerful likelihood-free technique that can be used in many different applications and fields. A short list of recent applications can be found, among others, in [14–16, 22]. The specific sweep strategy of the ABC algorithm is described next:

1. Adopt a prior distribution for each parameter involved in the expression of the weighted distribution, i.e., for each component of the vector of parameters θ of the bivariate random variable and γ_X, γ_Y. Since available information for the s-dimensional parameter θ of the population is expected to be available, informative priors are selected for it. On the other hand, this is not the case for γ_X and γ_Y, where non-informative priors are used.
2. Simulate θ^*, γ_X^* and γ_Y^* from their prior distributions. Let $\zeta^* = (\theta^*, \gamma_X^*, \gamma_Y^*)$.
3. Plug ζ^* in the proposed weighted distribution and simulate a sample $D^* = (x_j^*, y_j^*)$, $j = 1, \dots, n$ from the model with p.d.f. $f_{w_i}(x, y, \zeta^*)$.
4. Compute a discrepancy measure $d(D, D^*)$ between the observed D and the simulated D^* dataset and if $d(D, D^*) < \epsilon$, for some $\epsilon > 0$, accept ζ^*, otherwise repeat steps 2–4.

The aforementioned steps 2–4 are repeated until a predetermined number of non-rejected samples is obtained.

Remark 3 Note that for special cases of the weight functions, the vector of parameters $\zeta = (\theta, \gamma_X, \gamma_Y)$ may include some degenerated, fixed parameters. For example, for the models $\mathcal{M}_{iy}$ and $\mathcal{M}_{ix}$, $i = 1, \dots, 4$ the γ_X or γ_Y are set equal to zero, respectively. In these cases only the non-fixed parameters of ζ are simulated in step 2. Note that in such cases the weight functions $w_i(x, y;\theta, \gamma_X, \gamma_Y)$ are actually functions of only x or y and denoted as $w_{ix}(x;\theta, \gamma_X)$ or $w_{iy}(y;\theta, \gamma_Y)$, respectively.

In the sequel some necessary details for implementing the previous algorithm (see also [6]) in practice are discussed.

1. To keep the proposed method as simple as possible a general recommendation for the choice of the prior distributions could be the one given in Economou et al. [6]. Thus, we recommend the use of truncated at zero normal distribution (TN, hereafter) as prior distribution for any positive parameter, the normal distribution for any real parameter and the uniform distribution for any parameter bounded on a finite interval [a,b]. As explained in Economou et al. [6] such priors may not be the optimal ones, since any other prior distribution that incorporates in a more informative way, any prior expert knowledge from the problem domain is of course preferable. For instance, in the application presented in Sect. 4, a linear transformation of a Beta distribution is used as a prior for the correlation coefficient, while gamma distributions are used as priors for the variances of the normal distribution. Even if non-optimal priors are used the method can still yield reasonable parameter estimates. Adopting such priors may result in an increased rejection rate and therefore to a larger number of simulations until the predetermined number of ζ^* to be obtained. The posterior distribution is reported in the Appendix for the general case and in detail for the special case of the application presented in the following section.
2. As explained in detail in Economou et al. [6] sampling from $f_w(x, y; \zeta^*)$ is not an easy task. However, a random sample from $f_w(x, y; \zeta^*)$ is a biased sample from $f(x, y; \theta^*)$ where the probability of selecting a population unit in a sample is proportional to the weight function w. Based on this argument, [6] proposed to generate a large number N of observations from $f(x, y; \theta^*)$, which in most of the cases is an easy task. Then, the N observations can be considered as the "population" units. Finally a sample of size n from $f_w(x, y; \zeta^*)$ is obtained by applying a weighted sampling without replacement procedure to the "population" units with weights proportional to w. However, this procedure, which keeps the proposed method as simple as possible in order to be easy for practitioners, adds another layer of approximation to the posterior distribution of the parameters but this should not be that significant if the "population" size N is relatively large compared to the sample size n. This has been already demonstrated in Economou et al. [7] when a biased sample is drawn from a finite population without replacement with probability proportional to some function of its size, which does not depend on θ and can be easily extended to other cases.
3. A convenient discrepancy measure $d(D, D^*)$ between the observed D and the simulated D^* sample can be the absolute value of the standardized version of

$$T = \int_X \int_Y \left(f_D(x, y) - f_{D^*}(x, y)\right)^2 \mathrm{d}y\mathrm{d}x$$

where $f_D(x, y)$ and $f_{D^*}(x, y)$ are the kernel density estimates based on the observed D and the simulated D^* sample, respectively. Since under the null hypothesis that both samples share the same density, T is asymptotically normally distributed (for more details see [5]), a reasonable value for ϵ could be 1.96. This value corresponds to an asymptotic significance level 0.05 for testing the null hypothesis that both samples share the same density. In this frame, someone could perform the `kde.test` in R and accept ζ^* if its p value is less than 0.05.

Finally, in order to compare the three candidate models, $\mathcal{M}_{if}$, $\mathcal{M}_{iy}$ and $\mathcal{M}_{ix}$, induced by the selected weight function, the DIC, presented in [27] and further discussed in [4], is used. DIC is a generalization of Akaike information criterion (AIC), useful in Bayesian model selections problems and is defined as

$$\mathrm{DIC} = -4\, E_{\zeta}\left(\log f_w(x, y;\zeta)|(x, y)\right) + 2 \log f_w(x, y;\hat{\zeta})$$

where $\hat{\zeta}$ is usually the posterior mean of ζ. The posterior mode or median could be alternative choices for $\hat{\zeta}$. The $E_{\zeta}\left(\log f_w(x, y;\zeta)|(x, y)\right)$ can be estimated by using the ABC output by taking the sample mean of the simulated values of the log-likelihood. The model with the lowest value in DIC has the best performance on the data set under consideration and should be preferred.

The entire procedure, from the selection of the proper weight function to the choice of the prior distribution and the comparison of the three candidate models, is presented in detail in the following section using a real bivariate data set.

4 Application

4.1 NBA Draft

The NBA draft is an annual event in which the 30 teams in the National Basketball Association (NBA) select prospective players who are eligible and wish to join the league. Since 1989, the draft has consisted of two rounds and sixty players are selected in each draft. No player may sign with the NBA until he has been eligible for at least one draft. If a team chooses a player, that player cannot sign a contract to play for any NBA team other than that one.

Eligibility rules for players have changed several times during the history of the league but the main rules are that the players must be at least 19 years old at some point during the year of the draft and that they have never played in the NBA before. Players that have played at least one year of college basketball are eligible for the NBA draft. For that reason eligible players are typically college basketball players from NCAA (National Collegiate Athletic Association), international players and high school players one year after their high school graduation and if they are at least 19 years old as of the end of the calendar year of the draft.

The draft usually takes place near the end of June, during the NBA off-season, after more than 2-month rigorously training of the players that wish to join NBA. Usually, training starts after March Madness tournament, a single-elimination tournament, featuring 68 college basketball teams from the Division I level of the NCAA, which determine the national championship. The training includes not only individualized skill programs by replication of NBA workouts and body composition analysis and reconstruction, but also exposure workouts for national media outlets. Usually, players are training at various locations and are exposed on-site to every NBA team.

From the above it is clear that the drafted players consist a highly demanded group of players who not only outperform in most, if not all, of the statistics categories but also possess some important physical characteristics. For instance, their height and vertical jump, sometimes, guarantee a successful career in the NBA league. Therefore, drafted players are a biased sample and not a random one from the population of the players who are eligible for the NBA draft. The goal of the present section is to implement in detail the methodology previously proposed. In this frame, the population of interest encompasses all the players who are eligible to be drafted, while a bivariate biased data, where the height with no shoes and the max standing vertical jump (both in inches) of the 2017 draft players are given, will be considered. After a discussion about the prior knowledge related to the height and vertical jump of the members of the population of interest, taking into account the selection procedure of the drafted players and the characteristics of the four weight functions defined in relations (2)–(5), three possible models will be identified to analyze these bivariate biased data. More specifically, the models $\mathcal{M}_{1f}$, $\mathcal{M}_{1y}$ and $\mathcal{M}_{1x}$ (see Remark 3 for the notation) will be fitted, while using the DIC the candidate models were compared in order to choose the most suitable one.

4.2 Model Building—Prior Knowledge

For the 60 drafted players of 2017 included in the present application we considered the data regarding their height with no shoes X and their max standing vertical jump Y (both in inches). The data were extracted from https://data.world/achou/nba-draft-combine-measurements/workspace/file?filename=nba_draft_combine_all_Years.csv.

It is clear, as previously explained, that this bivariate sample is a biased and not a random one from the population of interest, i.e., the population which consists from all the players who are eligible to be drafted. Some of the descriptive statistics of the 60 drafted players of 2017 are presented in Table 1. The Spearman's correlation coefficient r between the two variables in the drafted players is equal to −0.397 (two-tailed p value = 0.004) and indicates a clear negative correlation. A plausible explanation is that taller players are also heavier which implies lower jump values, while at the same time shorter players in order to survive in a highly demanding environment should jump higher.

In order to apply the proposed methodology, initially the joint distribution of height and vertical jump in the population of interest should be determined. The bivariate normal distribution is a natural choice because the distribution of both

Table 1 Descriptive statistics for the height (with no shoes) and the max standing vertical jump (both in inches) of the 60 drafted players of 2017

	n Missing	Mean	SD	Minimum	Q1	Median	Q3	Maximum
Height (no shoes)	0	77.850	3.282	70.750	75.313	78.000	80.688	84.750
Vertical jump (max)	10	35.380	3.675	27.500	32.500	35.250	38.000	44.500

height and vertical jump can be approximated by a normal distribution in the population of interest. Height and vertical jump are also expected to be negatively correlated, probably not at the same degree as in the draft players but there will be no surprise if these characteristics were also negatively correlated in the entire population of interest. To summarize, we assume that the population distribution of the bivariate random vector (X, Y), where X denotes the height and Y the vertical jump, is the bivariate normal distribution with location parameter $(\mu_X, \mu_Y)'$, scale parameters $\sigma_X > 0$ and $\sigma_Y > 0$ and ρ the correlation between X and Y. Based on prior knowledge ρ is expected to be non-positive ($\rho \leq 0$) due to the negative influence of height on vertical jump.

In the sequel, one should not only choose between the available weight functions w_i, $i = 1, 2, 3, 4$, which were defined for dealing with different cases of bias sampling, but also to assign some prior distributions to the models parameters. For both issues, the prior knowledge based on previous studies related to the height and vertical jump in inches of the population of interest will be helpful.

Unfortunately, to the best of our knowledge, there are not available studies concerning the height and vertical jump of the international players that are eligible for the NBA drafts. On the other hand there are some studies for the NCAA Division I college basketball players. The fact that the percentage of international players in the NBA drafts is constantly increasing in the recent years and at the same time their rank in the drafts gets higher, allow us to assume that the eligible international players, who in average are not that young as the NCAA Division I college basketball players, share with the U.S. born players the same, if not better, characteristics. Therefore, the prior knowledge for the eligible players can be extracted from the studies regarding the NCAA Division I college basketball players.

The average height and vertical jump in inches of the NCAA Division I college basketball players was reported in [23] to be approximately 76.5 (range 66.5–90) and 28 (range 10–41.5), respectively. The extremely small value of the minimum vertical jump can be viewed as an outlier or as a measurement taken using a different protocol. As it is stated in [31] there are a number of protocols that are available for measuring the vertical jump that may result in notable differences in the measurements resulting to as much as 9 inches differences in the estimated mean value. In more recent unofficial reports the average vertical jump of NCAA Division I college basketball players is reported to be between 27 and 30 inches, while the average height is reported to be remarkably smaller than 76.5 inches. For example, in certain websites concerning basketball scholarships (see for example https://www.athleticscholarships.net/basketballscholarships.htm) the average height is reported to be 75 inches.

Even old studies (see [23]) show that there is no significant difference between the average height of draft players and NCAA Division I college basketball players, recent studies reveal that the mean height of the drafted players is higher than that of the NCAA Division I college basketball players. On the other hand, the jump ability of the drafted players is larger than all the average reported values. Actually almost all the drafted players have larger values than the average player in the NCAA Division I college basketball players. Summarizing, there is a clear tendency in the NBA drafts procedure to avoid players that are not tall enough and do not have large vertical jump. Therefore, it is natural to adopt the $w_1(x, y;\theta, \gamma_X, \gamma_Y)$ weight function, since this weight function allows to model situations in which large values of one or both variables are over-represented in the sample, as in the case of the NBA drafts.

As it was explained in the previous section for the parameters μ_X, μ_Y, σ_X^2, σ_Y^2 and ρ one should assign some prior distributions in such a way to incorporate all the available prior information, while since there is no available knowledge for γ_X and γ_Y no informative priors will be used. For these reasons the prior distributions of μ_X and μ_Y were set to be $N(76.5, 4.167^2)$ and $N(30, 4^2)$, respectively. The standard deviation, 4.167, of the height was roughly estimated by the reported range of [23], while the corresponding value for the vertical jump by assuming a range of values from 18 to 42, which is indeed slightly different from the one reported by [23] but more realistic based on more recent unofficial reports. The *inverse gamma* distribution was used as prior distribution for the σ_X^2, σ_Y^2 parameters with the shape parameter equal to 2 for both distributions and scale parameters equal to 4.167^2 and 4^2, respectively, in order the mean of the prior distribution to be the variance assumed for the priors of μ_X and μ_Y. For the parameter ρ the linear transformation $2W - 1$ of the $W \sim beta(\alpha, \beta)$ distribution was used as prior distribution, since $\rho \in [-1, 1]$. The parameters α and β of the transformed beta distribution were set equal to 25 and 30, respectively, in order the mean and the mode to be negative (−0.09 and −0.094, respectively), demonstrating the fact that we expect some negative correlation between these two variables, and to concentrate the values of the prior in a more reasonable region, i.e., between – 0.5 and 0.5. Finally, for the γ_X and γ_Y the truncated normal distribution was used as prior distributions. Since no information is available for these parameters the prior distributions are selected is such a way to explore the parameter space in a large range of values. More specifically, the *TN*(1, 10) was used for both γ_X and γ_Y, whenever the parameters γ_X and γ_Y was not set equal to zero (see models $\mathcal{M}_{1y}$ and $\mathcal{M}_{1x}$). The posterior distribution for the $\mathcal{M}_{1f}$ model under the aforementioned framework is presented in the Appendix.

4.3 Results—Models Comparison

Based on the previously discussed characteristics of the selection procedure of the drafted players, three possible models can be identified to analyze these bivariate biased data. The first, $\mathcal{M}_{1f}$, describes the case in which the bias in the observed sample is driven by both variables. The other two models, $\mathcal{M}_{1y}$ and $\mathcal{M}_{1x}$, describe the selection procedure under which the bias in the sample is due to only one of the variables.

Table 2 presents the descriptive statistics of the posterior marginal distributions of the parameters of the three models based on 5000 non-rejected samples. In the same table the value of DIC is also reported for each model. The model $\mathcal{M}_{1x}$ has the lowest value in DIC and presents the best performance on the studied data set thus it should be preferred. This suggests that the bias in the selection procedure of the drafted players is driven mainly by the height.

Table 2 Descriptive statistics of the posterior distributions of the parameters of the bivariate normal distribution for the 2017 NBA draft data under the three candidate models

Model	Descr. stat.	γ_X	γ_Y	μ_X	σ_X	μ_Y	σ_Y	ρ
$\mathcal{M}_{1f}$	2.5%	0.3036	0.3941	74.8464	2.0668	31.4468	2.1355	– 0.3443
DIC = −1500.462	25%	3.2303	3.3117	76.8228	2.8485	33.2662	2.9157	– 0.1873
Acc. rate: 0.0559	50%	7.0210	6.9647	78.0406	3.4575	34.3640	3.5146	– 0.0992
	75%	11.7486	11.9793	79.1214	4.2134	35.4960	4.3626	– 0.0097
	97.5%	22.8996	22.4645	80.8947	6.1956	37.7135	6.4965	0.1536
	Mean	8.1908	8.2174	77.9794	3.6279	34.4173	3.7395	– 0.0983
	SD	6.1373	6.1046	1.6023	1.0792	1.6182	1.1368	0.1281
	Mode	2.2615	2.3012	78.2635	3.0965	34.3004	3.1326	– 0.0947
$\mathcal{M}_{1x}$	2.5%	0.3790		73.8816	2.2278	31.5657	2.1555	– 0.3496
DIC = −1927.356	25%	3.5190		76.2024	3.0246	33.3796	2.9160	– 0.1897
Acc. rate: 0.0495	50%	7.3277		77.3955	3.6746	34.4537	3.5313	– 0.1011
	75%	12.1619		78.5182	4.4892	35.6260	4.3222	– 0.0116
	97.5%	23.7889		80.3527	6.6940	37.8122	6.4699	0.1639
	Mean	8.4517		77.3173	3.8683	34.5344	3.7336	– 0.0990
	SD	6.2277		1.6747	1.1617	1.6216	1.1223	0.1313
	Mode	2.5349		77.8856	3.2214	34.3605	3.0661	– 0.0760
$\mathcal{M}_{1y}$	2.5%		0.2585	74.9844	2.0943	30.2737	2.2748	– 0.3495
DIC = −1315.774	25%		2.9722	76.9306	2.8556	32.5462	3.1266	– 0.1902
Acc. rate: 0.0590	50%		6.8026	78.0786	3.4412	33.6388	3.8032	– 0.1008
	75%		11.7785	79.2020	4.2461	34.8123	4.6838	– 0.0116
	97.5%		22.8773	81.0750	6.1736	37.0908	7.1357	0.1607
	Mean		8.0052	78.0609	3.6318	33.6594	4.0381	– 0.0998
	SD		6.1819	1.5943	1.0566	1.7288	1.2579	0.1302
	Mode		1.8475	78.0709	3.1356	33.5334	3.3490	– 0.1090

The results for each model are computed based on 5000 non-rejected samples

The convergence of the sampler is depicted in Fig. 2 where the trace (history) and the ergodic mean plots are presented. The plots in Fig. 2 were obtained under the $\mathcal{M}_{1x}$. Similar plots can be also obtained for the other models.

In the upper plot of Fig. 3 are shown the contour plot of the estimated bivariate normal distribution and the corresponding weight function (using the mean of the posterior distributions)—the gray surface—under the $\mathcal{M}_{1x}$, the corresponding weight function and the scatterplot of the 2017 NBA drafted players. The figure suggests that the selection procedure indeed favors the players that are taller than others. More specifically, due to the posterior mean value of γ_X it seems that players that have notable lower height are almost excluded by the NBA drafts while players that possess values larger than the population mean value have larger probability of being selected in the drafts. To make this more clear at the lower plot of Fig. 3, the plot of the weight function as a function of height is also presented. Note that under $\mathcal{M}_{1x}$, which has the lowest DIC value and presents the

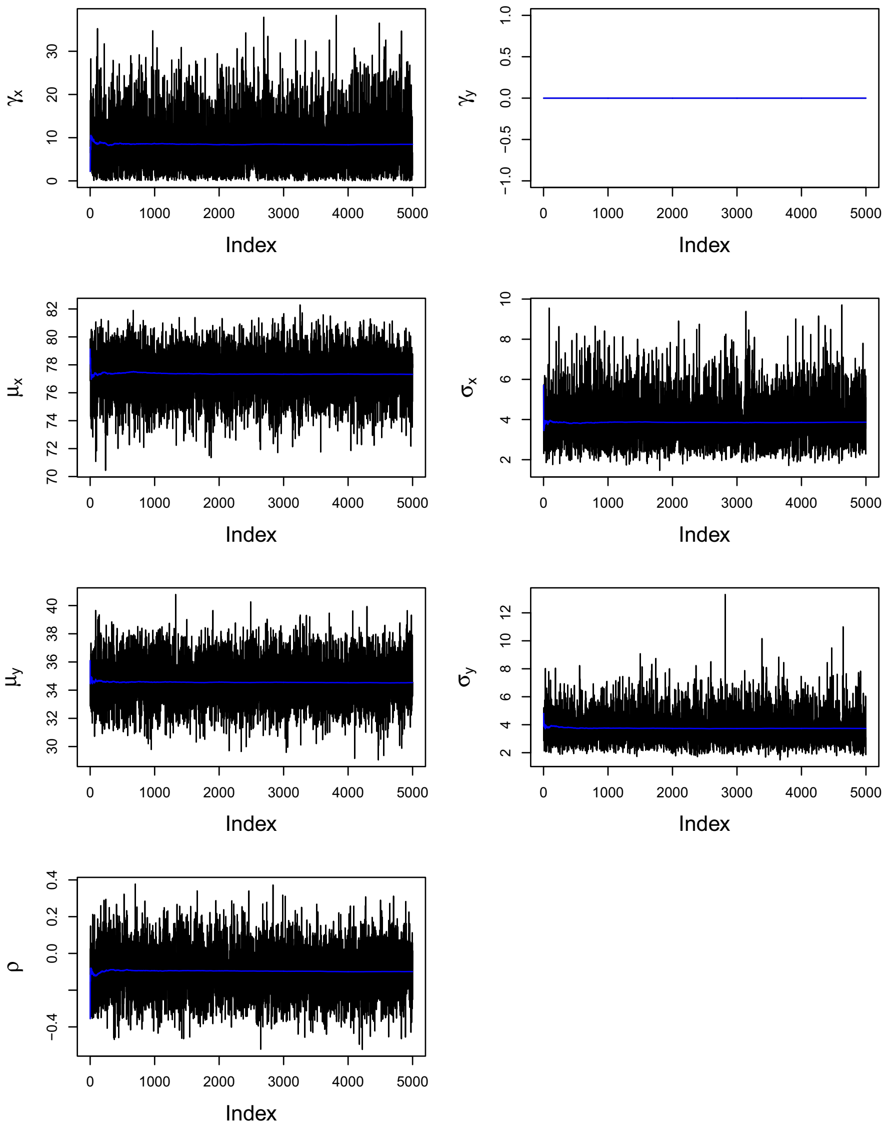
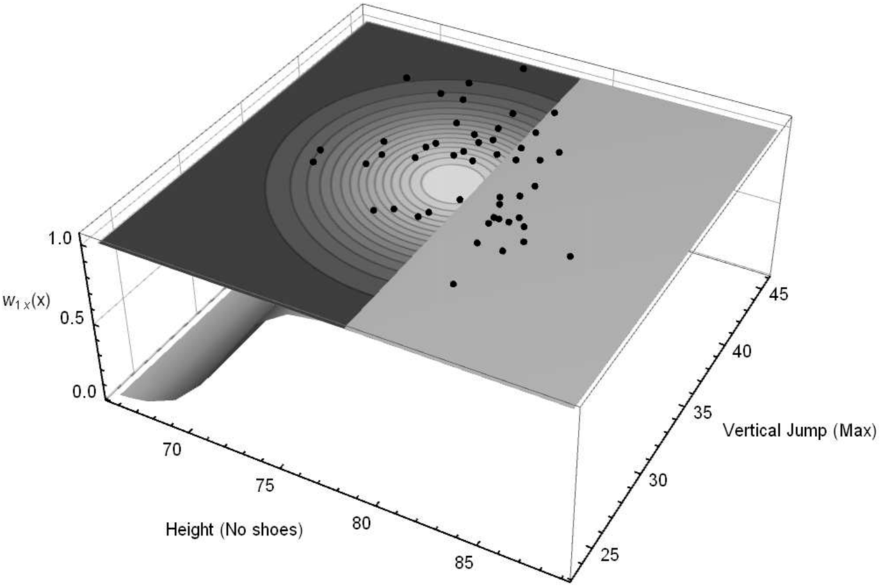


Fig. 2 The trace plot and the ergodic mean (blue line) under the $\mathcal{M}_{1x}$ for the 2017 NBA draft data for all the parameters. Note that the trace plot and the ergodic mean for γ_Y is constant since this parameter is zero under $\mathcal{M}_{1x}$

best performance on the data under study, the weight function varies only with respect to height. The plot suggests, since the probability of being selected in the drafts is proportional to the weight function, that the probability of a player being selected in the draft with 77.85 (the mean of the bias sample—right point in the graph) is two times higher than the probability of a player with 71.85 inches



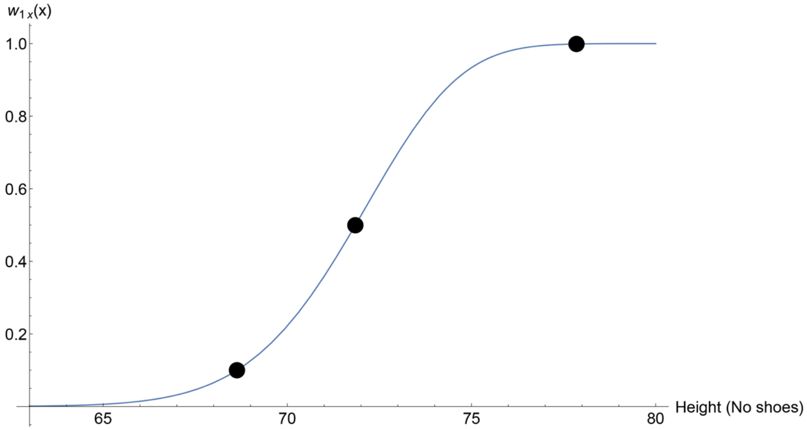


Fig. 3 Upper plot: The contour plot of the estimated bivariate normal and the corresponding weight function (using the mean of the posterior distributions) under the $\mathcal{M}_{1x}$ for the 2017 NBA draft data (black points). Lower plot: The weight function as a function of height with some characteristics points

height (middle point) and 10 times higher than the corresponding probability of a player with 68.63 inches height (left point).

The previous analysis may result in some practical recommendations involving not only the players who want to get drafted but also the agents who represent athletes and they are usually paid by a percentage of the money that the player earns. Taking into account that no significant change in the height can be observed in the short pre-draft period since height is difficult if not impossible to control by that age, the agent should initially select to represent players that are taller than others. Height may be enough to place a candidate player in the high probability drafted zone as indicated in Fig. 3. Afterwards, if an agent wants to increase the probability of his/her player to be selected in the drafts, he/she should provide him with individual nutrition and individualized skill programs to strengthen other factors like his vertical jump. The aforementioned conclusion coincides with the opinion of many of the well-known professional basketball agencies that a player is usually not drafted due to his lack of size, strength, or athleticism for the NBA.

The R code used in implementing the proposed method to the NBA 2017 draft players data set is available online as a supplementary file.

5 Conclusions

The main purpose of the current paper was to discuss how the four weight functions proposed in Economou et al. [6] can be used to understand if the observed bias in a bivariate sample is caused either by both random variables or by one of them. The proposed methodology can be applied to any bivariate biased data set in which specific members of the population are over- or under-represented in the sample. The background theory comprises the concept of weighted distributions to represent bias, the ABC algorithm to approximately draw samples from Bayesian posteriors and the Deviance Information Criterion to compare the fit of different models. More specifically, this contribution focused on the properties of the weight functions and the role of the extra parameters γ_X and γ_Y that governed their behavior is revealed. Special cases were introduced that describe situations in which the bias in the sample is caused only by one of the two random variables. A real data application for NBA draft players was also presented to illustrate the proposed methodology. This case study probably can serve as an example for practitioners potentially interested in using such methods for their own problems.

Appendix

In this Appendix the posterior density is reported for the general case and in detail for the special case of the application.

The likelihood function of a biased bivariate sample $D = (x_j, y_j), j = 1, \dots, n$ from a parent population with known pdf $f(x, y;\theta)$ where θ unknown parameters' vector, when the bias in the sample is described by the weight function $w_i(x, y;\theta, \gamma_X, \gamma_Y)$ is

$$\prod_{j=1}^{n} f_{w_i}(x_j, y_j;\theta, \gamma_X, \gamma_Y) = \prod_{j=1}^{n} f_{w_i}(x_j, y_j;\zeta) = \frac{\prod_{j=1}^{n} w_i(x_j, y_j;\zeta) f(x_j, y_j;\theta)}{E_f^n[w_i(X, Y;\zeta)]}.$$

Let $\pi(\zeta)$ be the joint prior density of the parameters of the model, where $\zeta = (\theta, \gamma_X, \gamma_Y)$. Then, the posterior density of the model has the form:

$$\pi(\zeta|\text{data}) \propto \frac{\prod_{j=1}^{n} w_i(x_j, y_j;\zeta) f(x_j, y_j;\theta)}{E_f^n[w_i(X, Y;\zeta)]} \cdot \pi(\zeta).$$

Based on the discussion of Sect. 4.2, the joint distribution of height and the vertical jump in the population of interest is a bivariate normal. Moreover, independence of the parameters of the model is assumed and a prior distribution is adopted for each parameter μ_X, μ_Y, σ_X^2, σ_Y^2, ρ, γ_X and γ_Y. Then, the posterior density takes the form:

$$\pi(\zeta|data) \propto \frac{\prod_{j=1}^{n} w_i(x_j, y_j;\zeta) f(x_j, y_j;\theta)}{E_f^n[w_i(X, Y;\zeta)]} \\ \pi(\mu_X)\pi(\mu_Y)\pi(\sigma_X^2)\pi(\sigma_Y^2)\pi(\rho)\pi(\gamma_X)\pi(\gamma_Y).$$

Using the priors described in Sect. 4.2 the following relation is obtained:

$$\begin{aligned}\pi(\zeta|data) \propto & \frac{\prod_{j=1}^{n} w_i(x_j, y_j;\zeta)}{E_f^n[w_i(X, Y;\zeta)]} \cdot \\ & \exp\left[-\frac{1}{2(1-\rho^2)} \sum_{j=1}^{n} \left[\frac{(x_j-\mu_X)^2}{\sigma_X^2} + \right.\right. \\ & \left.\left. \frac{(y_j-\mu_Y)^2}{\sigma_Y^2} - 2\rho \frac{(x_j-\mu_X)(y_j-\mu_Y)}{\sigma_X \sigma_Y}\right]\right] \\ & \exp\left[-\frac{1}{2}\left(\frac{(\mu_X-76.5)^2}{4.167^2} + \frac{(\mu_Y-30)^2}{4^2}\right)\right] \\ & \cdot (1+\rho)^{25-1}(1-\rho)^{30-1}(1-\rho^2)^{-n/2} \\ & \left(\frac{1}{\sigma_X^2}\right)^{2+1+n/2} \exp\left[-\frac{4.167^2}{\sigma_X^2}\right] \left(\frac{1}{\sigma_Y^2}\right)^{2+1+n/2} \exp\left[-\frac{4^2}{\sigma_Y^2}\right] \cdot \\ & \exp\left[-\frac{1}{2}\left(\frac{(\gamma_X-1)^2}{10} + \frac{(\gamma_Y-1)^2}{10}\right)\right] I(\gamma_X > 0) \cdot I(\gamma_Y > 0)\end{aligned}$$

which can be expressed equivalently as

$$\pi(\zeta|data) \propto \frac{\prod_{j=1}^{n} w_i(x_j, y_j;\zeta)}{E_f^n[w_i(X,Y;\zeta)]} \cdot \\ \exp\left[-\frac{1}{2(1-\rho^2)}\sum_{j=1}^{n}\left[\frac{(x_j-\mu_X)^2}{\sigma_X^2} \right.\right. \\ \left.\left. + \frac{(y_j-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x_j-\mu_X)(y_j-\mu_Y)}{\sigma_X\sigma_Y}\right]\right] \\ \exp\left[-\frac{1}{2}\left(\frac{(\mu_X-76.5)^2}{4.167^2} + \frac{(\mu_Y-30)^2}{4^2}\right)\right] \cdot \\ (1+\rho)^{24-n/2}(1-\rho)^{29-n/2} \\ \left(\frac{1}{\sigma_X^2\sigma_Y^2}\right)^{3+n/2} \exp\left[-\frac{4.167^2}{\sigma_X^2} - \frac{4^2}{\sigma_Y^2}\right] \cdot \\ \exp\left[-\frac{1}{2}\left(\frac{(\gamma_X-1)^2}{10} + \frac{(\gamma_Y-1)^2}{10}\right)\right] I(\gamma_X > 0) \cdot I(\gamma_Y > 0).$$

For the model $\mathcal{M}_{1f}$, i.e., $i = 1$ and γ_X, γ_Y strictly positive, the posterior density has the form

$$\pi(\zeta|data) \propto \frac{\prod_{j=1}^{n}\left(1-\left(1-\Phi\left(\frac{x_j-\mu_X}{\sigma_X}\right)^{\gamma_X}\right)\left(1-\Phi\left(\frac{y_j-\mu_Y}{\sigma_Y}\right)^{\gamma_Y}\right)\right)}{E_f^n\left[\left(1-\left(1-\Phi\left(\frac{X-\mu_X}{\sigma_X}\right)^{\gamma_X}\right)\left(1-\Phi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^{\gamma_Y}\right)\right)\right]} \cdot \\ \exp\left[-\frac{1}{2(1-\rho^2)}\sum_{j=1}^{n}\left[\frac{(x_j-\mu_X)^2}{\sigma_X^2} \right.\right. \\ \left.\left. + \frac{(y_j-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x_j-\mu_X)(y_j-\mu_Y)}{\sigma_X\sigma_Y}\right]\right] \\ \exp\left[-\frac{1}{2}\left(\frac{(\mu_X-76.5)^2}{4.167^2} + \frac{(\mu_Y-30)^2}{4^2}\right)\right] \cdot \\ (1+\rho)^{24-n/2}(1-\rho)^{29-n/2} \\ \left(\frac{1}{\sigma_X^2\sigma_Y^2}\right)^{3+n/2} \exp\left[-\frac{4.167^2}{\sigma_X^2} - \frac{4^2}{\sigma_Y^2}\right] \cdot \\ \exp\left[-\frac{1}{2}\left(\frac{(\gamma_X-1)^2}{10} + \frac{(\gamma_Y-1)^2}{10}\right)\right] I(\gamma_X > 0) \cdot I(\gamma_Y > 0).$$

Due to the posterior's form direct sampling from it or even sampling from a standard MCMC method is not an easy task. Thus, ABC methods are used.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42519-021-00167-2.

Compliance with ethical standards

Conflict of interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Afonso L, Corte Real P (2016) Using weighted distributions to model operational risk. ASTIN Bull 46(2):469–485
2. Arnold B, Nagaraja H (1991) On some properties of bivariate weighted distributions. Commun Stat Theory Methods 20(5–6):1853–1860
3. Berkson J (1946) Limitations of the application of fourfold table analysis to hospital data. Biom Bull 2:47–53
4. Celeux G, Forbes F, Robert CP, Titterington DM (2006) Bayesian Anal 1(4):651–673
5. Duong T, Goud B, Schauer K (2012) Closed-form density-based framework for automatic detection of cellular morphology changes. Proc Nat Acad Sci 109(22):8382–8387
6. Economou P, Batsidis A, Tzavelas G, Alexopoulos P (2020) ADNI: Berkson's paradox and weighted distributions: An application to alzheimer's disease. Bioml J 62:238–249
7. Economou P, Tzavelas G, Batsidis A (2020) Robust inference under r-size-biased sampling without replacement from finite population. J Appl Stat 47(13–15):2808–2824
8. Fisher R (1934) The effect of methods of ascertainment upon the estimation of frequencies. Ann Eugen 6(1):13–25
9. Geneletti S, Best N, Toledano MB, Elliot P, Richardson S (2013) Uncovering selection bias in case-control studies using Bayesian post-stratification. Stat Med 32:2555–2570
10. Greenland S (2003) Quantifying biases in casual models: classical confounding vs collider-stratification bias. Epidemiology 14:300–306
11. Gupta RC, Kirmani S (1990) The role of weighted distributions in stochastic modeling. Commun Statist 19(9):3147–3162
12. Hernan M, Hernandez-Diaz S, Robins J (2004) A structural approach to selection bias. Epidemiology 15:615–625
13. Jain K, Nanda A (1995) On multivariate weighted distributions. Commun Stat Theory Method 24(10):2517–2519
14. Kacprzak T, Herbel J, Amara A, Réfrégier A (2018) Accelerating approximate Bayesian computation with quantile regression: application to cosmological redshift distributions. J Cosmol Astropart Phys 2018(02):042
15. Kavetski D, Fenicia F, Reichert P, Albert C (2018) Signature-domain calibration of hydrological models using approximate Bayesian computation: theory and comparison to existing applications. Water Resour Res 54(6):4059–4083
16. McKinley T, Vernon I, Andrianakis I, McCreesh N, Oakley J, Nsubuga R, Goldstein M, White R (2018) Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. Stat Sci 33(1):4–18. https://doi.org/10.1214/17-STS618
17. Nanda A, Jain K (1999) Some weighted distribution results on univariate and bivariate cases. J Stat Plan Inference 77(2):169–180
18. Navarro J, Ruiz J, Aguila YD (2006) Multivariate weighted distributions: a review and some extensions. Statistics 40(1):51–64
19. Patil G, Rao C (1978) Weighted distributions and size-biased sampling with applications to wildlife populations and human families. Biometrics 34(2):179–189
20. Pearl J (1995) Casual diagrams for empirical research. Biometrika 82(4):669–688
21. Rao C (1965) On discrete distributions arising out of methods of ascertainment. Sankhya Indian J Stat Ser A (1961–2002) 27(2/4):311–324
22. Raynal L, Marin J, Pudlo P, Ribatet M, Robert CP, Estoup A (2018) ABC random forests for Bayesian parameter inference. Bioinformatics 35(10):1720–1728

23. Richard L, Berg K, Thomas B (1994) Physical and performance characteristics of ncaa division i male basketball players. J Strength Cond Res 8(4):214–218
24. Rotnitzky A, Robins J (2005) Inverse probability weighted estimation in survival analysis. In: Encyclopedia of Biostatistics. Wiley, London
25. Samuelsen S, Anestad H, Skrondal A (2007) Stratified case-cohort analysis of general cohort sampling designs. Scan J Stat 343:103–119
26. Sarabia JM, Gomez-Deniz E (2008) Construction of multivariate distributions: a review of some recent results. SORT 32(1):3–36
27. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B (Stat Methodol) 64(4):583–639
28. Spirtes P, Glymour C, Scheines R (1993) Causation, prediction, and search. The MIT press, Cambridge
29. Tzavelas G, Douli M, Economou P (2017) Model misspecification effects for biased samples. Metrika 80(2):171–185
30. VanderWeel T, Herman M, Robins J (2008) Casual directed acyclic graphs and the direction of unmeasured confoundin bias. Epidemiology 19:720–728
31. Ziv G, Lidor R (2010) Vertical jump in female and male basketball players-a review of observational and experimental studies. J Sci Med Sport 13(3):332–9

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.