On the sensitivity of Genetic Matching to the choice of balance measure

Adeola Oyenubi

ERSA working paper 840

November 2020

On the sensitivity of Genetic Matching to the choice of balance measure

Adeola Oyenubi[‡]

November 6, 2020

Abstract

This paper considers the sensitivity of Genetic Matching (GenMatch) to the choice of balance measure. It explores the performance of a newly introduced distributional balance measure that is similar to the KS test but is more evenly sensitive to imbalance across the support. This measure is introduced by Goldman & Kaplan (2008) (i.e. the GK measure). This is important because the rationale behind distributional balance measures is their ability to provide a broader description of balance. I also consider the performance of multivariate balance measures i.e. distance covariance and correlation. This is motivated by the fact that ideally, balance for causal inference refers to balance in joint density and individual balance in a set of univariate distributions does not necessarily imply balance in the joint distribution.

Simulation results show that GK dominates the KS test in terms of Bias and Mean Square Error (MSE); and the distance correlation measure dominates all other measure in terms of Bias and MSE. These results have two important implication for the choice of balance measure (i) Even sensitivity across the support is important and not all distributional measures has this property (ii) Multivariate balance measures can improve the performance of matching estimators.

JEL classification: I38, H53, C21, D13

Keywords: Genetic Matching, Balance measures, causal inference, Machine learning

1 Introduction

In an observational setting, treatment is not randomly assigned to units. The implication is that treatment exposure may be related to some covariates or/and

*School of Economics and Finance, University of the Witwatersrand, Johannesburg adeola.oyenubi@wits.ac.za

†The author is grateful to David Kaplan for his assistance in understanding the technical details of the GK measure. The authors is grateful to the Centre for High Performance computing, Rosebank, Cape Town, South Africa (https://www.chpc.ac.za/) for giving us access to their machine without which most of the simulations in this study would not have been possible.

the outcome. In the likely event that covariates are imbalanced across treatment arms, a naïve mean causal effect estimator will be biased. Rosenbaum & Rubin (1985) show that using the probability of exposure to treatment conditional on observed covariates (propensity scores) one can adjust for imbalance. This can be operationalized by matching based on propensity scores.

However, there have been several advances in the propensity score matching methodology. The key problem is that propensity scores are unknown and have to be estimated. These advances have highlighted the weakness in the traditional approach of using logit or probit model to estimate propensity scores. First, estimating propensity score involves a specification problem, while it is known that covariates that are correlated with both treatment and outcome should be included in the model (see Oyenubi (2020) for example), how these covariates should be included in terms of higher-order terms is ad hoc at best. Ho *et al* (2007) noted that to use nonparametric matching to avoid parametric modelling assumptions, a researcher must know the parametric functional form of the propensity score equation. However, until we have the specification of the propensity score that balances all covariates we have to keep repeating the process. Second, the choice of balance measure used to identify the "correct" propensity score specification can influence the result (Oyenubi & Wittenberg, 2020).

To avoid this problem Imai & Ratkovic, (2014) introduce Covariate Balancing Propensity Score that incorporates balancing condition directly into the estimation of propensity scores while Diamond & Sekhon (2013) side-line specification search by focusing on balance in relevant covariate using Genetic algorithm. Oyenubi & Wittenberg (2020) show that even under the methodology of Diamond & Sekhon (2013) the choice of balance measure is still important.

In this paper, I explore the findings of Oyenubi & Wittenberg (2020). The authors show through several simulation studies that the choice of balance measure influences the Bias and Mean Square Error (MSE) of matching estimates under genetic matching (GenMatch)[1]. In this paper, I provide further evidence in this respect by considering a (univariate) balance measures that were not considered by Oyenubi & Wittenberg (2020) and also extend their results to the case of multivariate measure of balance.

Primarily, two main analysis was performed. First, the fact that that the Kolmogorov Smirnov (KS) (i.e., the default distributional balance measure under GenMatch) suffers from low sensitivity at the tails (Buja and Rolke (2006); Goldman and Kaplan (2018)) is noted. This is problematic because balance measures that ignore imbalance irrespective of where they occur can introduce Bias and inefficiency in the estimation of treatment effect. Consequently, the current study proposes a way to solve this problem by considering the Dirichlet approach of Buja and Rolke (2006). This approach uses probability integral transform to attain more even sensitivity. Goldman and Kaplan (2016 & 2018) proposed a two-sample variant of this test that ease the computation of critical values (the original proposal uses permutation test). In this study, I refer to

[1] And by extension any matching estimator that seeks to optimize balance.

this measure as the GK measure[2]. The use of this measure is motivated by the fact that it is by design, more evenly sensitive to imbalance across the support of the distributions being compared.

This is in contrast to the KS measure which (as mentioned earlier) has lower sensitivity at the tails. This distinction is important in the context of the results of Oyenubi & Wittenberg (2020) (and for the choice of balance measure in general). Their result shows that the standardized mean difference (SMD) is more effective at minimizing Bias and MSE than the distributional measures (i.e. the KS test and the entropy measure) considered in their study (even under a realistic simulation design). However, in this study, I argue that this result is counter-intuitive given the arguments made in the literature in favour of distributional measures of balance (see Austin, 2009; Huber 2009; Belitser et al. 2011; Imai *et. al,* 2008). One of such arguments suggests that distributional measures of balance provide a broader description of imbalance (Austin, 2009; Huber 2009; Belitser et al. 2011; Imai *et. al,* 2008). The implication of this is that under the assumption that "better balance" leads to "better effect estimates", distributional measures should be more effective. In comparing the KS and the GK measures, it is important to note that "distributional measures" as a category of balance measures consist of different measures which vary in how they quantify the difference between distributions. For example, given the distinction between the KS and the GK measure, if the GK measure is found to outperforms the KS measure, this will imply that even sensitivity to departure from balance across distributions matters[3].

Second, the GK measure and all other balance measures considered by Oyenubi & Wittenberg (2020) assess balance at the individual covariate level. As noted by Iacus et. al (2012), the goal of measuring imbalance is to summarize the difference between the multivariate empirical distribution of pretreatment covariates. Furthermore, Andrei & McCarthy (2019) argue that adequate individual covariate balance does not necessarily imply balance in the multivariate sense. In general, marginal independence does not imply joint independence (Andrei & McCarthy, 2019). Following Andrei & McCarthy (2019) who propose the use of distance covariance (Székely et al, 2007) as a test of balance/independence, in the current study, the performance of this multivariate balance measure in the context of GenMatch is examined. Specifically, I use the distance covariance, correlation, and p-value of distance correlation (dCov, dCor, and dCorP respectively) as measures of balance under GenMatch.

For the first analysis, the simulation result shows that the GK measure outperforms the KS measure both in terms of Bias and MSE[4]. This suggests that

[2]Kaplan (2019) implement the measure in Stata, however the code used in this paper is the R version downloaded from the author's website "https://faculty.missouri.edu/~kaplandm/#distinf"

[3]I acknowledge that the entropy measure introduced by Oyenubi & Wittenberg (2020) can be argued to be a distributional balance measure that is sensitive across the support. However, this measure was dominated by a number of other measures in their simulation results, perhaps because it involves the estimation of kernel density which may affect its efficiency.

[4]Note that both measures (GK and KS) are used in conjunction with t-test of difference in means as suggested by (Diamond and Sekhon, 2013).

part of the reason for the poor performance of distributional balance measures in Oyenubi and Wittenberg (2020) is the kind of balance measure that was used[5]. On the other hand, while the GK measure does not outrightly outperform the SMD, its performance is more competitive than the other distributional measures considered by Oyenubi & Wittenberg (2020). Specifically, the SMD has a slightly lower Bias while the GK measure has a slightly lower RMSE. This suggests that the contrast between the two measures can be thought of as a trade-off between Bias and precision, where the SMD produces lower Bias at the expense of precision, and the GK measure does the opposite.

For the second analysis, result further show that the multivariate balance measures (dCov and dCor) dominate all the univariate measures in terms of MSE while dCov, in addition, dominates all other balance measures (both univariate and multivariate) in terms of Bias. This result implies that multivariate measures of balance are not just more effective in guiding the specification of propensity score (as shown by Andrei & McCarthy (2019)), but also perform better as after matching balance test (at least in this context). These results further strengthen the main argument in Oyenubi & Wittenberg (2020) that the choice of balance measures does matter. In general, the results of the current study suggest that multivariate balance measures are more promising than their univariate counterparts. However, since the use of multivariate measures in this context is relatively new, more evidence will be needed to check if this performance varies in different contexts.

The rest of the study is organized as follows, the next section reviews the literature to provide a justification for using the GenMatch approach. The GenMatch algorithm is reviewed and the GK and multivariate measures used in this study are introduced. This is followed by the sections that discuss the simulation design, presents and discusses the results, while the last section concludes the study.

2 Literature review

Oyenubi & Wittenberg (2020) argue that since balance measures vary in terms of what they are designed to capture (i.e. where some focus on certain parts of the distribution like mean or mean and variance, others compare distributions), one should expect variation in their performances when they are used as a yardstick to optimize balance for matching estimators. This variation is often ignored in the literature where in most cases the performance of different balance measures is taken as a given. For example, while researchers conduct robustness checks to see if the result is sensitive to the type of matching (e.g. one-to-one versus kernel matching) no such check is conducted to investigate the sensitivity of results to the choice of balance measure.

[5] Note that one can argue that the entropy distance measure proposed by Oyenubi & Wittenberg (2020) is also sensitive across the support, an important difference is that the entropy measure require the estimation of kernel density.

The authors noted that the choice of balance measure is particularly important in the case of GenMatch which is based on Genetic Algorithm (Mitchell, 1989; Carr, 2014). This is because it is well known that the performance of genetic algorithms depends on the fitness function which, in the case of GenMatch, is defined by the balance measure. Therefore, differences in fitness function (or balance measure) can lead to differences in optimal results. This argument can be extended to other matching methods since all matching methods seek to optimize balance. While Oyenubi & Wittenberg (2002) focused on showing this with different simulation studies under GenMatch, one can argue that this point is implied by other results in the literature. For example, Belitser et al. (2011) show that when different balance measures are used to select the specification of propensity score, the correlation between Bias and Imbalance vary with the choice of balance measure.

The proposed GK measure is introduced by Goldman & Kaplan (2018)[6], refining an idea from Buja and Rolke (2006). There are two versions of the test: the first version assesses equality of distribution quantile-by-quantile and displays the range of quantile values for which differences are statistically significant (Kaplan, 2019). The second version is a global goodness-of-fit (GOF) test similar to the KS test. Goldman & Kaplan (2018) show that this test may be preferred to the KS test because its sensitivity to deviations is more evenly spread across the distribution. The fact that the KS test suffers from low sensitivity at the tails is known in the literature (see Eicker (1979)). Even sensitivity across the support is important because it may better identify imbalance or cases where there are thin/no common support problem (in finite samples) irrespective of where they occur[7]. Therefore, differences in sensitivity of balance measures can be important in influencing the performance of matching estimators. This is well highlighted by the examples presented by Goldman & Kaplan (2018), where the authors noted that if the null hypothesis distribution is uniform (0,1), even a sample maximum of one million (which is impossible under the null) will not cause KS test to reject. Furthermore, with a sample size of 20, even 5 observations equal to one million cannot persuade KS test to reject at 10%. (see footnotes 3 and 4 in Goldman & Kaplan (2018) and Kaplan (2019) for details).

The distinction between multivariate measures of balance and balance measures that assess balance in individual covariates in order to declare balance overall is also important. This is because (as stated earlier) individual balance on many covariates does not necessarily imply joint balance. In a recent paper, Andrei & McCarthy (2019) show that the dCov measure (which is a multivariate measure of statistical dependence between random vectors analogous to product-moment correlation) detects significant differences between the joint distribution of covariates across treatment arms, even when the (absolute) SMD is less than 0.2 for each covariate. This is important because the literature suggests that values below 0.2 for SMD is indicative of adequate balance (Rosen-

[6] It is implemented in Stata by Kaplan (2019)

[7] Thin or no support problem can increase biases and variances of matching estimators (see Crump et al, 2009; Khan and Tamer, 2010)

baum & Rubin, 1985). This echoes the sentiment of Sekhon (2007) and Iacus et. al (2012) who have noted that ideally, measures that capture discrepancies in higher-order moments and are multidimensional should be preferred.

As shown in Oyenubi & Wittenberg (2020), depending on the Data Generating Process (DGP) the discrepancy in treatment effect estimates that stems from the choice of balance measure alone can be large enough for results to be drastically different. Therefore, the choice of balance measure has implication for inferences that come from analyses that relies on matching estimators. Existing simulation studies that examine the performance of matching estimators focus on other choices that must be made when matching estimators are employed. These choices include: the specification of the propensity score model (both in terms of distributional assumption (e.g. probit versus logit model) and the kind of covariates that should be included in the model, see Zhao (2004), Schmidt and Augurzky (2011) and Oyenubi(2020)); and the choice of matching algorithm or the type of matching to be used (i.e. Nearest neighbour, Radius, stratification or Kernel matching see Caliendo and Kopeinig (2008)) etc. While these choices are important because of their implication for Bias and MSE of the resulting matching estimate, my argument in this article is that the choice of balance measure is equally important. This is because the success, or lack thereof, of all these other choices, are measured by the balance measure (i.e. after matching balance)[8]. The result of Oyenubi & Wittenberg (2020) suggests that these choices interact with each other to determine the Bias and MSE of matching estimators.

3 Methodology

3.1 Why GenMatch?

First introduced by Hollard (1992), Genetic Algorithms belong to the class of approaches used in adaptive aspects of computation – search, optimization, machine learning, parameter adjustment, etc (Shapiro, 1999). Shapiro (1999) noted that Genetic Algorithms are essentially reinforcement learning algorithms, and like other learning algorithms, their performance is determined by the fitness function.

In the causal inference literature, matching is a popular approach to reduce Bias due to selection under the ignorability assumptions (see Rosenbaum and Rubin (1983)). The main idea is that under ignorability, matching units across treatment arms will allow for causal inference. All matching methods have one goal: to balance (ideally the joint) distribution of covariates across treatment arms after matching. To achieve this, matching algorithms depend on balance measures to decide when matching has achieved its goal.

The most popular matching approach – Propensity Score Matching (PSM) – has been shown to have many flaws (see Iacus *et al* (2012), Imai and Ratkovic

[8] This implies that an ineffective balance measure can erode the effect of other choices made earlier in the process.

(2014) and Goller *et al* (2020) for some discussion on this). Traditional PSM is based on manual optimization by iteratively tweaking matching parameters and checking balance. It, thus, often leads to suboptimal solutions (King et.al, 2017). To mitigate this problem a number of matching methods and alternative approaches to estimating propensity scores have been proposed in the literature. GenMatch is unique amongst the new matching approaches in that the choice of balance measure is left to the researcher. In view of the finding that balance measures can influence the performance of matching estimators, this is important[9].

3.2 Review of Genetic Matching

GenMatch is a general matching approach that combines the strengths of Mahalanobis distance matching and PSM. The algorithm searches a range of balancing scores to find the score that optimizes the covariate balance after matching (Diamond and Sekhon, 2013). The balancing scores are indexed by a weight matrix W such that each weight matrix corresponds to a different balancing score. GenMatch minimizes a generalized version of the Mahalanobis distance by incorporating an additional weight matrix W. It is given by

$$d\left(X_i, X_j\right) = \left\{ \left(X_i - X_j\right)' \left(S^{-1/2}\right)' \; W \, S^{-\frac{1}{2}} \left(X_i - X_j\right) \right\}^{1/2}$$

where X is the matrix of covariates and S is the sample covariance matrix of X. $S^{-1/2}$ is the Cholesky decomposition of S i.e. $S = S^{-\frac{1}{2}} \left(S^{-\frac{1}{2}}\right)^T$. W is a k by k positive definite matrix and k is the dimension of X (i.e., the number of covariates). All elements of W are zero except the main diagonal which consist of k parameters that must be chosen.

Note that estimated Propensity scores can be included as one of the covariates under GenMatch. In general, both propensity score and Mahalanobis distance matching can be thought of as limiting cases of GenMatch. If propensity scores contain all relevant information required to balance the covariates, all other variables will receive zero weight, or more appropriately, enough weight just to make sure that W is positive definite. In this case, GenMatch is equivalent to propensity score matching. On the other hand, GenMatch will converge to Mahalanobis distance (even when propensity score is included) if it is the more appropriate distance measure for the sample (i.e. when the propensity score fails to achieve the optimal level of balance in the covariates). In less extreme cases, GenMatch allocates weights to the propensity scores and all covariates (Diamond and Sekhon, 2013; Oyenubi, 2018 & 2019).

GenMatch requires the user to specify a loss function and a covariate balance measure. Note that the loss function (or fitness function) is a function of the balance measure. The default loss function under (the R implementation of)

[9] Other matching and weighting algorithm do not allow the user to choose their preferred balance measure. For example, the Covariate Balancing Propensity Score and Entropy balancing.

GenMatch minimizes the largest individual discrepancy based on P-values from KS tests and paired t-test for all the covariates. The algorithm uses P-values so that results from different tests can be compared on the same scale. Because the sample size is fixed within the optimization, the general concern that P-values depend on sample size does not apply (Imai, King, & Staurt, 2008; Diamond and Sekhon, 2013).

Interested readers should see Diamond & Sekhon (2013) for technical details concerning GenMatch.

3.3 Comparing the KS and the GK measures

As previously noted, there are two forms of the GK test. In this exposition, focus is given to the Global goodness-of-fit (GOF) version. Let $f(x)$ and $g(y)$ be the distributions to be compared and let $F(r)$ and $G(r)$ be the corresponding empirical cumulative distribution functions (CDF) – where r refers to points or quantiles on the CDF. The KS test hypothesis can be written as

$$H_0 : F(r) = G(r) \; for\, all\, r$$

If H_0 is true, then $\hat{F}(r)$ and $\hat{G}(r)$ should be "close" to each other; if not the test rejects the hypothesis. KS defines "close" as the maximum vertical distance between the CDFs, i.e.

$$KS = sup_r \left| F(r) - G(r) \right|$$

In other words, to compare the distributions, the KS statistic uses the "biggest gap" between the CDFs and consequently has uneven sensitivity to differences in different parts of the distribution (Goldman & Kaplan, 2018). Parizzi & Brcic (2011) note that the KS statistic tends to be more sensitive near the centre of the distribution relative the tails (also see Kvam & Vidakovic (2007)). This is because the KS distance does not distinguish between the shapes of the CDF curves, taking only the maximum distance (which occurs at one point) into account (Wang & Chang, 2012).

Buja and Rolke (2006) propose a Dirichlet-based approach to achieve more even sensitivity. Their approach uses probability integral transform to reduce the problem to that of the ordered statistic from the standard uniform distribution. The ordered statistic is known to follow a known ordered Dirichlet distribution in finite samples (Goldman and Kaplan, 2018). While the technical details behind this approach are beyond the scope of the current study, it suffices to say that the Dirichlet approach achieves even sensitivity relative to the KS test. Goldman & Kaplan (2018) improve this approach in terms of computational time and power. Interested readers should see Buja and Rolke (2006) and Goldman & Kaplan (2016 & 2018) for technical details. Also, see appendix B for a sketch or the argument presented in Goldman & Kaplan (2018) that describes how the GK measure works.

Under the Dirichlet approach, the GOF null can be written as

H_0 : all H_{0r} are true

This GOF test distinguishes whether all H_{0r} are true or at least one is false. This leads to greater sensitivity to imbalance at any r. The question about which specific r is false is handled by the second version of the test which is not of interest to this study.

3.4 Review of Distance covariance and correlation measures (multivariate measures of independence)

The fact that marginal balance for a set of covariates does not necessarily imply joint balance underscores the importance of considering measures of balance that capture discrepancy in joint distributions. Furthermore, assessing balance in marginal distributions for inference that should ideally be based on joint distribution introduces the possibility of increase type I error (see Lee (2013) for some discussion on multiple testing problem in the context of PSM).

Szekely *et al* (2007) introduce the dCov and dCor measures of statistical independence between two sets of vectors. This approach is based on Energy distance (i.e. a measure based on powers of Euclidean distance) and was introduced to replace the standard non-parametric goodness-of-fit tests in high dimensions. This measure can be used to compare distributions or the distributions of two samples e.g. treatment and control samples (Huling and Mak, 2020).

Let $X \in R^p$ and $M \in R^q$ be p and q dimensional vectors with finite first moments. Note that X and M need not be of the same size or dimension. Let $f_X(t) = E\left[e^{i(t,X)}\right]$, $f_M(s) = E[e^{i(s,M)}]$ and $f_{X,M}(t,s) = E\left[e^{i(t,s),(X,M)}\right]$ be characteristic functions of XM and (XM). The dCov of X and M is defined by

$$\nu^2(X,M) = \|f_{X,M}(t,s) - f_X(t) f_M(s)\|^2$$

where $\|.\|_w$ is the weighted L_2 norm defined as $\|\gamma(t,s)\|_w^2 = \int_R^{p+q} \left|\gamma(t,s)^2\right| w(t,s)dtds$; where the positive weight $w(ts)$ is such that the integral exists[10]. Therefore, ν can be used to measure the distance between the joint characteristic function and the product of marginal characteristic functions. It then forms the basis for test of independence between X and M i.e.,

$$H_0 : f_{X,M} = f_X f_M \quad vs \quad H_1 : f_{X,M} \neq f_X f_M$$

the distance covariance of X is

$$\nu^2(X) = \|f_{X,X}(t,s) - f_X(t) f_X(s)\|$$

So that the dCor is given by

$$R^2(X,M) = \begin{cases} \frac{\nu^2(X,M)}{\sqrt{\nu^2(X)\nu^2(M)}}, \nu^2(X)\nu^2(M) = 0 \\ 0, \quad \nu^2(X)\nu^2(M) = 0 \end{cases}$$

[10] There are a number of weights that can be used with this measure. In this study, the weight advocated by Székely and Rizzo (2020) is used because it has the property that the resulting distance will be both rigid motion invariant, and scale equivariant. Note that rigid motion can be explained as a way of moving points on a plane such that (a) relative distance between stays the same (b) the relative position of points stay the same.

dCov = 0 is equivalent to statistical independence (similar argument can be made for dCor).

Similar to Hainmueller (2012), who use entropy distance to recover a set of weights that balance the distribution of (individual) covariates, Huling and Mak, (2020) use the energy distance to recover the Energy Balancing Weights (EBW).

4 Simulation design

Monte Carlo studies are useful in examining the small sample properties of different matching estimators. Oyenubi & Wittenberg (2020) use 3 simulation studies, two of which can be argued to be unrealistic in practical settings. In this study, I focus on the simulation study that mimics conditions that can be found in reality.

In the spirit of Empirical Monte Carlo Study (EMCS), the simulation follows the work of Diamond & Sekhon (2013). The simulation is based on real data (i.e. the Dehejia and Wahba (1999) sample of the Lalonde (1986) experimental data set). The covariates include both discrete and continuous variables, there are eight covariates, six of which are binary variables. Since the design is not new, we present the details of the design in appendix A.

The parameters of the simulation are as follows; the treatment effect is 1000, the sample size is 445 (185 treated and 260 control observations) and the simulation is conducted 1000 times[11]. Note that the ratio of control to treated units is constant across simulations. Note that the value of the parameter $\varphi = -1.5$ i.e. the case where the initial imbalance is higher in Oyenubi & Wittenberg (2020) see the appendix for details.

5 Results

The result of the first simulation (univariate balance measures) is presented in table 1. Table 1 reports the percentage Bias relative to the true treatment effect and the Root Mean Square Error (RMSE). We present the result when the estimation of the treatment effect is done with and without Bias correction (Bias Correction refers to regression adjustment see Sekhon (2008)). To compare our results to the one in Oyenubi and Wittenberg (2020), we include the balance measures they used in their study (the exception is the entropy distance because it requires significantly more computation time).

In general, the result is similar to the one presented in Oyenubi & Wittenberg (2020) i.e. the performance of GenMatch varies with the choice of balance measure and the performance under Bias correction (BC) is worse than the alternative (No BC) in every case. Furthermore, the default measure has the

[11] Oyenubi & Wittenberg (2020) use 500 replication. However there is no difference in the ranking of the performance of different balance measures and the point estimates are very similar.

lowest Bias and RMSE when the estimation is with Bias correction. These results are consistent with what was found by Oyenubi & Wittenberg (2020). Note that this means the GK measure did not perform better than the KS measure under Bias correction. However, note that the result under BC is almost always worse than the result without Bias correction.

Focusing on the results without Bias correction (i.e., where all measures perform better in terms of Bias and RMSE[12]): the GK measure dominates the KS measure both in terms of Bias and RMSE. This suggests that distributional measures vary in their performance and measures that have even sensitivity should be preferred. While the SMD still dominates all other measures in terms of Bias, the GK measure dominates all other measures in terms of RMSE. In other words, there appears to be a trade-off between Bias and RMSE when it comes to the performance of SMD and GK.

We now consider the performance of multivariate measures of balance. For this analysis we consider the dCov, dCor and the p-value of the dCor measure (dCorP), the result is shown in Table 2.

The result shows that except for the case of the P-value of dCor, whose performance is the worst across the tables, the best performing multivariate measure dominates the best performing univariate measure both in terms of Bias and RMSE in both categories (i.e. BC and No BC). To put this in context note that the Bias of dCor is the lowest at 0.8% of the true treatment effect and this figure is almost half the size of the percentage Bias for the SMD, which dominates the univariate measures in terms of Bias. This result echoes the result of Andrei & McCarthy (2019) on the efficacy of multivariate measures.

The results of this study show that the SMD is still a competitive measure of balance but relative to it, (univariate) distributional balance measures that are evenly sensitive across the distribution can increase the precision of matching estimates. Furthermore, the distinction between univariate and multivariate balance measures is important in terms of Bias and precision of matching estimators. While multivariate measures are relatively new and more evidence will be needed to establish, their behaviour in different settings, the available evidence in the literature (Andrei & McCarthy, 2019; Huling and Mak, 2020), and the results of this study, suggest that they will provide useful alternatives to existing measures.

6 Conclusion

This study considers variation in performance of Genetic matching algorithm that stems from differences in the choice of balance measure. The result confirms and extends the result of Oyenubi & Wittenberg (2020). Specifically, the study finds that a distributional balance measure that is evenly sensitive to departure from balance is more effective than distributional balance measures that don't have this property. Results also indicated that such measures may reduce the MSE of matching estimate relative to the SMD. Lastly, the result shows

[12] Except the default measure that has higher RMSE when there is not bias correction

that multivariate balance measure largely performs better than their univariate counterparts under GenMatch.

The fact that the performance of balance measure might vary with the DGP and the matching algorithm is acknowledged. Therefore, there is need for more research to investigate the performance of the measures considered here (especially the multivariate measures) in different contexts. For example, because of the type of simulation design adopted with this study, it was impossible to investigate if the performance found here is sensitive to sample size.

Concerning the choice of balance measure this paper shows that (i) Even sensitivity across the support is important and not all distributional measures has this property (ii) Multivariate balance measures can improve the performance of matching estimators.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical approval:

This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest:

Author Adeola Oyenubi declares that he has no conflict of interest

References

[1] Andrei, A.C. and McCarthy, P.M., 2019. An omnibus approach to assess covariate balance in observational studies using the distance covariance. *Statistical Methods in Medical Research, 29*(7), 1846-1866.

[2] Buja, A. and Rolke, W., 2003. Calibration for simultaneity:(re) sampling methods for simultaneous inference with applications to function estimation and functional data, available at http://stat.wharton.upenn.edu/~buja/PAPERS/paper-sim.pdf

[3] Caliendo, M. and Kopeinig, S., 2008. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, *22*(1), pp.31-72.

[4] Dehejia, R. H., & Wahba, S., 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association, 94*(448), 1053-1062.

[5] Diamond, A., & Sekhon, J. S., 2013. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics, 95*(3), 932-945.

[6] Goldman, M. and Kaplan, D.M., 2018. Comparing distributions by multiple testing across quantiles or CDF values. *Journal of Econometrics*, *206*(1), pp.143-166.

[7] Goldman, M., & Kaplan, D. M., 2016. Evenly sensitive KS-type inference on distributions. Tech. rep., Working paper, available at http://www.paper.edu.cn/scholar/showpdf/MUT2AN2IOTD0Ex0h

[8] Goller, D., Lechner, M., Moczall, A. and Wolff, J., 2020. Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. *Labour Economics*, 101855.

[9] Hainmueller, J., 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 25-46.

[10] Holland, J.H., 1992. Adaptation in natural and artificial systems. 1975. Ann Arbor, MI: University of Michigan Press.

[11] Huber, M., 2009. Testing for covariate balance using nonparametric quantile regression and resampling methods. Unpublished Working and Discussion Papers.

[12] Huber, M., Lechner, M. and Wunsch, C., 2013. The performance of estimators based on the propensity score. *Journal of Econometrics, 175*(1), 1-21.

[13] Huling, J.D. and Mak, S., 2020. Energy Balancing of Covariate Distributions, available at https://arxiv.org/abs/2004.13962

[14] Ho, D.E., Imai, K., King, G. and Stuart, E.A., 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), pp.199-236.

[15] Iacus, S.M., King, G. and Porro, G., 2012. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 1-24.

[16] Imai, K & Ratkovic, M, 2014. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(1), 243–63

[17] Imai, K., King, G. and Stuart, E.A., 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society), 171*(2), 481-502.

[18] Kaplan, D.M., 2019. GK: Comparing distributions. *The Stata Journal*, *19*(4), 832-848.

[19] King, G., Lucas, C. and Nielsen, R.A., 2017. The balance-sample size frontier in matching methods for causal inference. American Journal of Political Science, 61(2), 473-489.

[20] Kinnear Jr, K.E., 1994. A perspective on the work in this book. Advances in Genetic Programming, 3-19.

[21] Kvam, P. H., & Vidakovic, B. (2007). Nonparametric statistics with applications to science and engineering (Vol. 653). John Wiley & Sons.

[22] LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 604-620.

[23] Lechner, M. and Strittmatter, A., 2019. Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews*, *38*(2), 193-207.

[24] Mebane Jr, W.R. and Sekhon, J.S., 2011. Genetic optimization using derivatives: the rgenoud package for R. Journal of Statistical Software, 42(11), 1-26.

[25] Oyenubi, A. and Wittenberg, M., 2020. Does the choice of balance-measure matter under genetic matching? *Empirical Economics*, 1-14.

[26] Oyenubi, A., 2018. Quantifying balance for causal inference: An information theoretic perspective (Doctoral dissertation, University of Cape Town).

[27] Oyenubi, A., 2020. A note on Covariate Balancing Propensity Score and Instrument-like variables. *Economics Bulletin*, *40*(1), pp.202-209.

[28] Parizzi, A., & Brcic, R., 2011. Adaptive InSAR stack multilooking exploiting amplitude statistics: A comparison between different techniques and practical results. *IEEE Geoscience and Remote Sensing Letters, 8*(3), 441-445.

[29] Rosenbaum, P.R. and Rubin, D.B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

[30] Schmidt, C.M. and Augurzky, B., 2001. The propensity score: A means to an end. *Tech. rep., IZA Discussion Paper Series*.

[31] Sekhon, J.S., 2007. Alternative balance metrics for Bias reduction in matching methods for causal inference. Survey Research Center, University of California, Berkeley.

[32] Sekhon, J.S., 2008. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software,* (Forthcoming).

[33] Shapiro, J., 1999, July. Genetic algorithms in machine learning. In *Advanced Course on Artificial Intelligence* (pp. 146-168). Springer, Berlin, Heidelberg.

[34] Székely, G.J., Rizzo, M.L. and Bakirov, N.K., 2007. Measuring and testing dependence by correlation of distances. *The annals of statistics*, *35*(6), 2769-2794.

[35] Székely, G.J. and Rizzo, M.L., 2012. On the uniqueness of distance covariance. *Statistics & Probability Letters*, *82*(12), 2278-2282.

[36] Wang, F. and Chan, C., 2012, June. Variational-distance-based modulation classifier. In 2012 IEEE International Conference on Communications (ICC), 5635-5639.

Table 1: Simulation Results (Univariate measures)

	Bias		RMSE	
	BC	No BC	BC	No BC
Mean	41.3	25.2	584.91	577.52
Standardized difference in means	43.2	1.7	607.31	585.69
Default	33.8	29.4	552.55	749.86
GK	42.8	2.6	592.41	571.71

BC: Bias correction by regression No BC: No Bias correction (1000 replications)

Table 2: Simulation Results (Multivariate measures)

	Bias		RMSE	
	BC	No BC	BC	No BC
dCov	30.4	4.0	445.09	442.39
dCor	31.2	0.8	453.16	442.52
dCorP	21.4	193	612.53	2402.15

BC: Bias correction by regression No BC: No Bias correction (1000 replications)