OPEN ACCESS Freely available online

PLOS ONE

A New Distribution-Free Approach to Constructing the Confidence Region for Multiple Parameters

Zhiqiu Hu[1], **Rong-Cai Yang**[1,2]*

1 Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Alberta, Canada, **2** Research and Innovation Division, Alberta Agriculture and Rural Development, Edmonton, Alberta, Canada

Abstract

Abstract Construction of confidence intervals or regions is an important part of statistical inference. The usual approach to constructing a confidence interval for a single parameter or confidence region for two or more parameters requires that the distribution of estimated parameters is known or can be assumed. In reality, the sampling distributions of parameters of biological importance are often unknown or difficult to be characterized. Distribution-free nonparametric resampling methods such as bootstrapping and permutation have been widely used to construct the confidence interval for a single parameter. There are also several parametric (ellipse) and nonparametric (convex hull peeling, bagplot and HPDregionplot) methods available for constructing confidence regions for two or more parameters. However, these methods have some key deficiencies including biased estimation of the true coverage rate, failure to account for the shape of the distribution inherent in the data and difficulty to implement. The purpose of this paper is to develop a new distribution-free method for constructing the confidence region that is based only on a few basic geometrical principles and accounts for the actual shape of the distribution inherent in the real data. The new method is implemented in an R package, distfree.cr/R. The statistical properties of the new method are evaluated and compared with those of the other methods through Monte Carlo simulation. Our new method outperforms the other methods regardless of whether the samples are taken from normal or non-normal bivariate distributions. In addition, the superiority of our method is consistent across different sample sizes and different levels of correlation between the two variables. We also analyze three biological data sets to illustrate the use of our new method for genomics and other biological researches.

Citation: Hu Z, Yang R-C (2013) A New Distribution-Free Approach to Constructing the Confidence Region for Multiple Parameters. PLoS ONE 8(12): e81179. doi:10.1371/journal.pone.0081179

Editor: Rongling Wu, Pennsylvania State University United States of America

Received July 10, 2013; **Accepted** October 9, 2013; **Published** December 4, 2013

Funding: This research is supported by the Natural Sciences and Engineering Research Council of Canada discovery grant (Award #183983) to R-C Y. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rong-cai.yang@ales.ualberta.ca

Introduction

Confidence interval estimates of individual parameters are more informative than simple point estimates and thus they are widely used in statistical inference [1,2,3]. However, a joint confidence region (CR) for two or more parameters is often needed in practical applications. Classical applications include the joint CR for two or more regression coefficients in a typical multiple regression analysis [2]. More recently, there have been calls for the use of the joint CRs to ascertain superior genotypes identified for target environments in biplot analysis of genotype-by-environment interaction [4,5] or to unambiguously infer about population stratification in human admixtures [6,7,8,9,10].

Construction of the confidence intervals or regions for parameters often assumes that the data are from a normal distribution and they are balanced. For example, for bivariate normally-distributed data, the required CR is an ellipse whose shape depends largely on the level of the correlation between the two variables. However, when the distribution is unknown or hard to be characterized, several nonparametric procedures are available for construction of the confidence intervals or regions. Data peeling is a valuable approach to inspecting the structure of multivariate data [11]. The predominant implementation of data peeling is based on the convex hull of the data [12]. In convex hull peeling, the outmost convex hull is identified, the observations in the convex are assigned with index value of one and then these observations are removed from the data. This procedure is iterated but the index value is increased by one for each iteration until all observation are assigned with indexes. A CR can be determined by identifying the layer of peeling with the indexes higher than the threshold (preset significant level). The peeling approach is further developed by considering data depth [13,14] to address the inquiry to the effectiveness of the procedure [11,15]. HPDregionplot [16] is another nonparametric method for constructing CR. The fundamental behind the HPDregionplot is to use the contour that embraces the desired proportion of the capacity based on the two-dimensional kernel density estimates [17] as CR.

One of the key limitations with these parametric and non-parametric methods is the inaccurate estimation of the coverage rate by the CRs with the data of unknown distributions. All the non-parametric methods are computationally demanding [18] and some of them (e.g., HPDregionplot) are sensitive to small sample sizes. In this paper, we introduce a simple distribution-free geometry-based procedure that allows for constructing the CR for two or more parameters when there is no knowledge about the sampling distributions of the estimated parameters. We examine

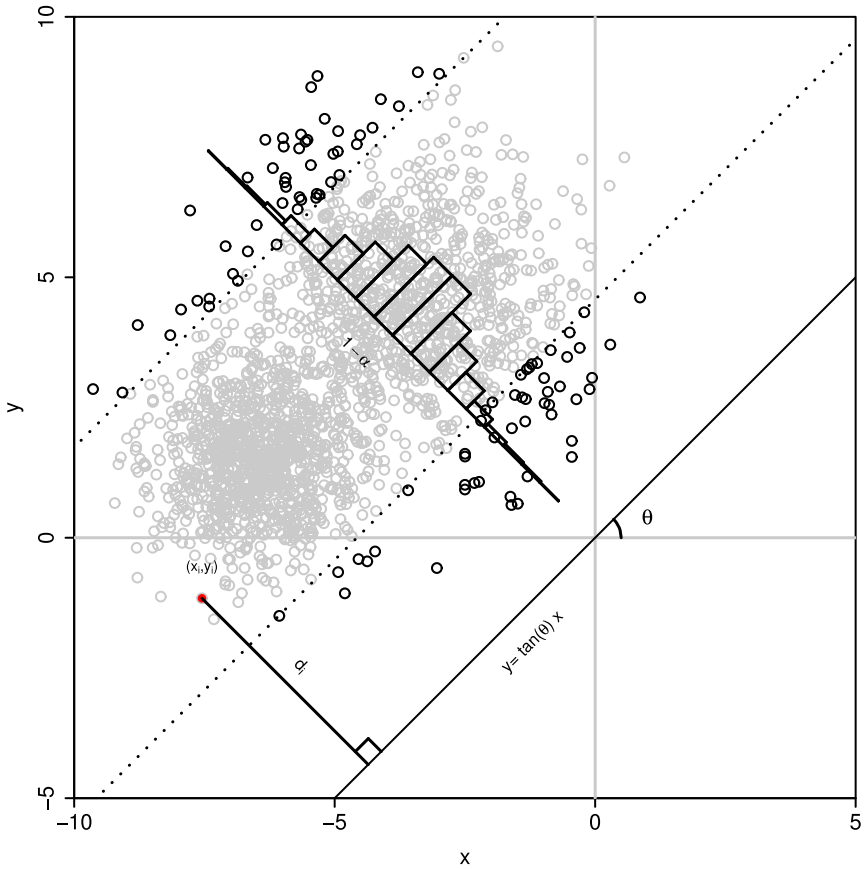

Figure 1. The confidence region constructed for an arbitrary reference line. The simulated population is an equal-proportional mixture of the observations sampled from two bivariate normal distributions which are given by $N\left(\begin{bmatrix}-6.5\\1.5\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$ and $N\left(\begin{bmatrix}-3.5\\4.5\end{bmatrix},\begin{bmatrix}2.25&0\\0&2.25\end{bmatrix}\right)$, respectively. The two parallel dashed lines are the boundaries of the confidence region for the reference line with angle θ. The black and gray open circles are points outside and within the boundaries, respectively. The histogram shows the distribution of the distances with respect to the reference line and the heights of the bars are the observed frequencies multiplied by 5.
doi:10.1371/journal.pone.0081179.g001

statistical properties of the new method through computer simulations and illustrate its use through two biological examples.

Materials and Methods

Quantile for a single parameter

For a single parameter, the distribution-free approach to computing a percentile is quite straightforward. Although different definitions for percentiles exist [19], all the definitions would lead to similar results given a large number of the random samples [20]. After obtaining estimates from individual random samples, three basic steps are followed to construct a distribution-free confidence interval: (1) to sort the N estimates in the ascending order; (2) to search for the nearest ranks for p^{th} percentile by picking up the closest integers to $N \times p$; and (3) to estimate the desired percentile by linear interpolation between the two consecutive ranks.

Quantiles for multiple parameters

Although the above procedure considers one variable only, it can be extended to the calculation of the CR simultaneously for two or more variables. For simplicity, let us consider the case of two variables. Let $\mathbf{x}$ and $\mathbf{y}$ be the two vectors of size $(N \times 1)$. The values in vector $\mathbf{x}$ are the Euclidean distances, in geometry, between the observed points and the vertical coordinate (i.e., the reference line at $x=0$). Similarly, the values in vector $\mathbf{y}$ are the Euclidean distances between the observed points and the horizontal coordinate (i.e., the reference line at $y=0$). Thus the quantiles estimated for a single parameter are also the quantiles of the relative distances between the observed points and the reference line at $x=0$ or $y=0$. However, with unknown joint sampling distribution of variables x and y, all potential reference lines across the entire plane need to be considered while constructing the distribution-free CR.

Here we describe a general geometry-based approach to constructing the CR for any bivariate data. As mentioned earlier, the confidence interval for one variable can be regarded as a special case in which the reference line has been set to either vertical or horizontal coordinate axis ($x=0$ or $y=0$). Now let us consider the confidence interval for an arbitrary reference line (cf. Figure 1). Since the positions of the observations in relation to a reference line, i.e., the distances with directions, are used to obtain the percentile, all reference lines have the same slopes but with different intercepts. We simplify the derivation by assuming all reference lines through the origin of the coordinates. The arbitrary reference line is expressed as

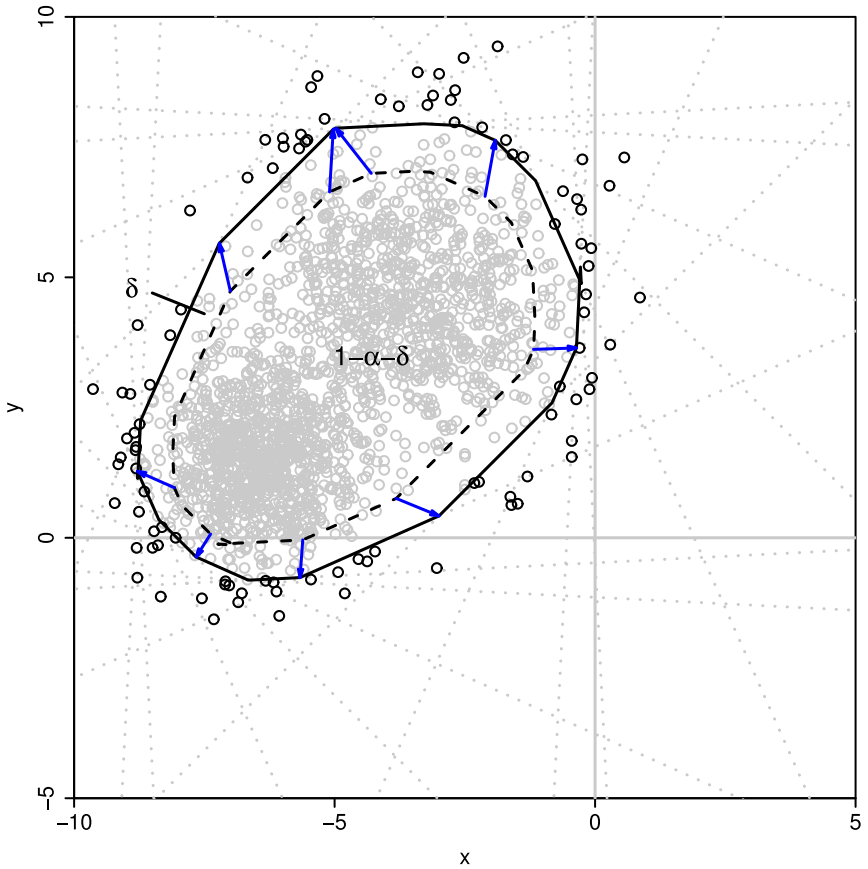

Figure 2. The confidence region and polygon boundaries obtained by rotating the reference line. The polygon is formed by the dash lines and the confidence region is constructed by setting the significant level for each test at α. The outer polygon with solid lines represents the expanded confidence region with observed significant level approximating to desired significant level α.
doi:10.1371/journal.pone.0081179.g002

$$\mathbf{y} = \tan(\theta)\mathbf{x}, \tag{1}$$

where θ is the angle between the reference line and the horizontal abscissa (see section A of Appendix S1 for detailed derivation). It is also evident from Figure 1 that the relative position (distance) of the i^{th} observations (x_i, y_i) to the reference line as given in eq (1), is calculated as (see section A of Appendix S1 for detailed derivation),

$$d_i = \frac{\tan(\theta)x_i - y_i}{\sqrt{\tan(\theta)^2 + 1}}. \tag{2}$$

Applying eq. (2) repeatedly for all N observations, we obtain the relative positions that are stored in vector **d**. If the **d** vector is viewed as a single variable, then the algorithm described earlier can be directly applied to calculate the required quantiles. Here we consider that the statistical inference is based on the two-tailed tests. For a specified significance level α, the confident interval of a single parameter is flanked by the observed lower- and upper-boundaries, i.e., the $(N \times \alpha/2)^{\text{th}}$ and $[N \times (1-\alpha/2)]^{\text{th}}$ percentiles. In geometry view, the boundaries $l_{\theta,1} = d_{N\times\alpha/2}$ and $l_{\theta,2} = d_{N\times(1-\alpha/2)}$ represent the distances between two parallel lines and the reference line to ensure that 95% of the total data points lie within the boundaries and 5% outside the boundaries in the direction $\theta + \pi/2$ (see Figure 1). The function of the i^{th} boundary line in an arbitrary direction in the plane is given as (see section B of Appendix S1 for detailed derivation)

$$y = \tan(\theta)x + l_{\theta,i}\sqrt{1 + \tan(\theta)^2}, i = 1,2. \tag{3}$$

Let us denote the subset of all out-of-boundary points in the direction with the angle of θ as P_θ. The observed significant level in this direction is expected to approximate the specified significant level for a single parameter (α),

$$\alpha'_\theta = \frac{n_\theta}{N} \approx \alpha \tag{4}$$

where n_θ is the number of out-of-boundary points in the direction with the angle of θ in P_θ. Using the same strategy, we obtain the boundary lines in all directions by rotating the reference line in all directions over the plane. By taking all boundaries jointly into consideration, we construct a CR as a polygon in the plane under the assumption that the significant level for each direction is α. To the newly constructed region, the observations outside the polygon are counted as

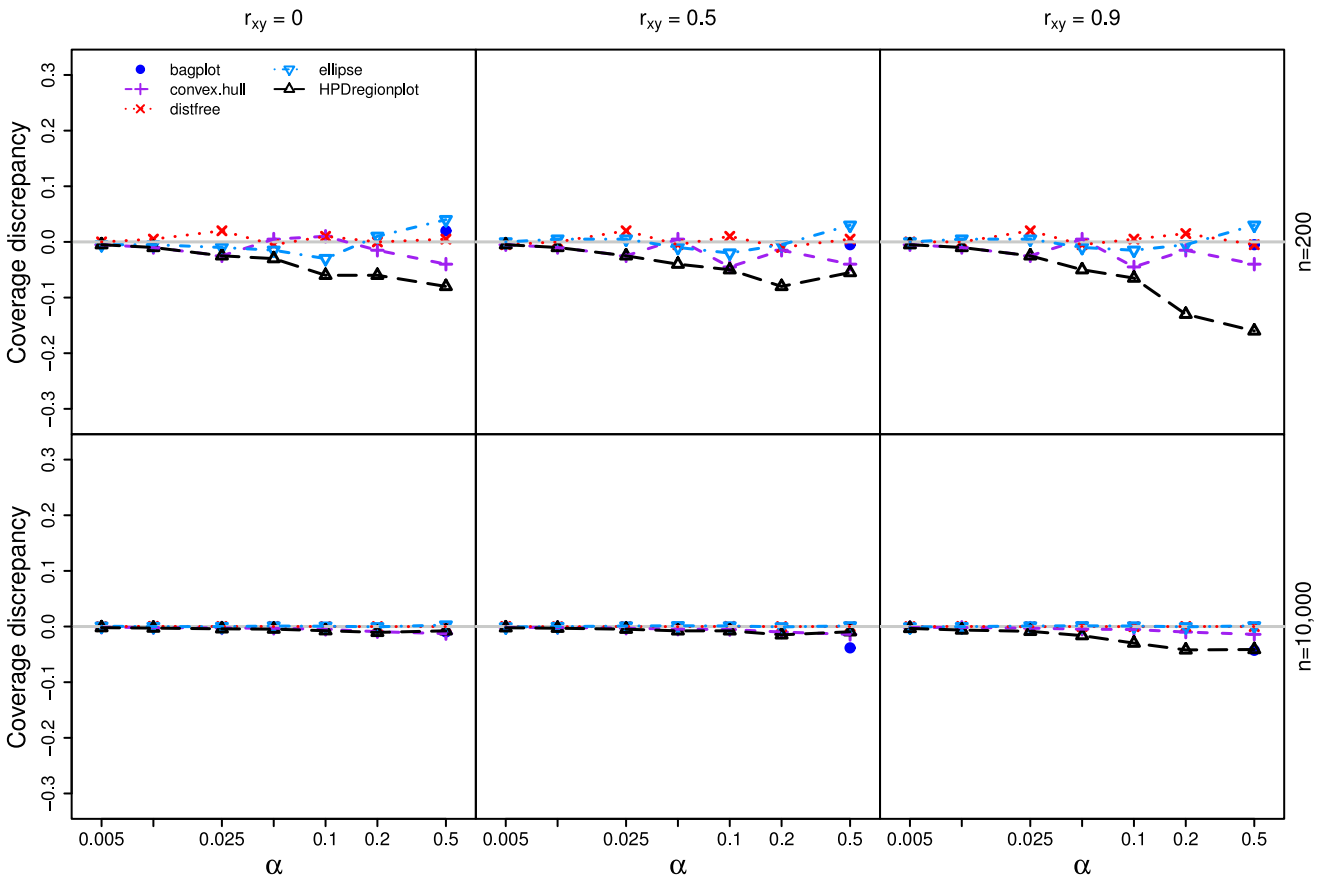

Figure 3. Coverage discrepancy of the empirical confidence regions in simulation design I. The empirical confidence regions for a range of probability levels ($\alpha = 0.005$ to 0.5) are constructed by five methods (distfree, ellipse, bagplot, convex hull peeling and HPDregionplot) based on a small sample (n = 200) and a large sample (n = 10,000) taken from a bivariate normal distribution with mean vector $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and variance-covariance matrix of $\Sigma = \begin{bmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{bmatrix}$, where r_{xy} takes 0, 0.5 and 0.9. The confidence region by bagplot is available only at $\alpha = 0.5$.
doi:10.1371/journal.pone.0081179.g003

$$P = \bigcup P_\theta. \tag{5}$$

Since the directions with angles of θ and $\theta + \pi$ are actually the same reference line, we require the slope of the reference lines to increase monotonically with the angle while rotating the reference line with the range of θ being $\theta \in [-\pi/2, \pi/2]$.

It should be noted that the method described above can also be viewed as a set of multiple tests and thereby the observed significant level for the CR is actually greater than the α level that is specified for each test, i.e.,

$$\alpha' = \frac{n}{N} = \alpha - \delta \tag{6}$$

where n is the number of observations in P, δ is the difference between the expected and the desired significant levels (Figure 2). Thus, the α value that is actually specified to calculate the CR for each test should be lower than the desired significant level for multiple tests. Although it is difficult to provide a general function to describe the relationship between the two values, the desired α value can be obtained iteratively from the follow equation

$$\alpha_{k+1} = \alpha_k + (\alpha_0 - \alpha_k') \frac{\alpha_k'}{\alpha_0} \tag{7}$$

where α_k is the assigned value of the significant level required for generating the CR in each direction, α_k' denotes the actually significant level for the CR bounded by the polygon as showed in eq (6), and α_0 is the desired significant level for the overall test.

In this study, we construct the CR that is approximated by a polygon in a two-dimensional plane for the two variables. In each direction, the polygon is bounded by the lower- and upper-boundaries as given in eq (3). The vertices of the polygon are the crossover points of all adjacent boundary lines. The vertice between two adjacent reference lines with the angle of δ is a point in the plane whose two coordinate values are given by,

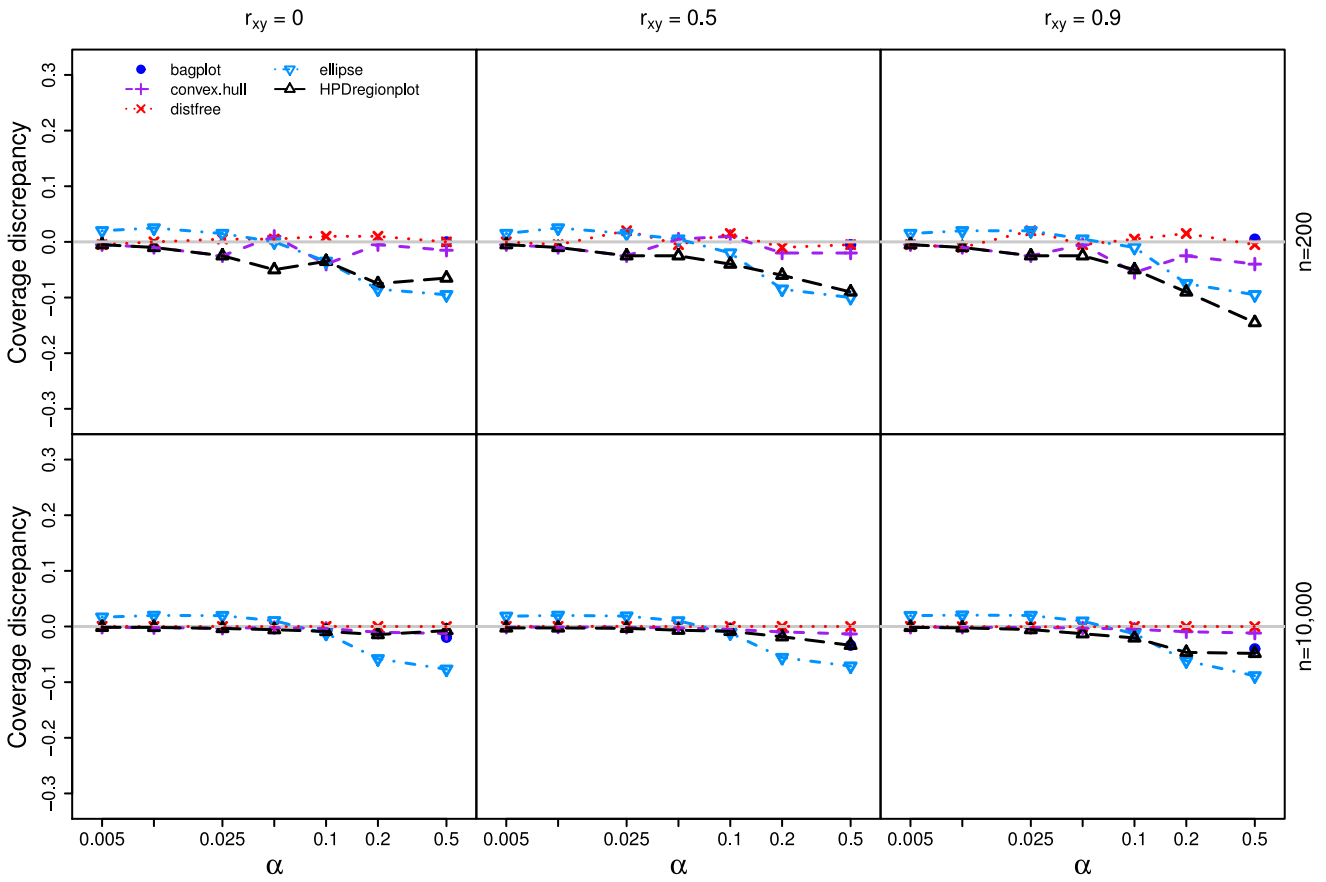

Figure 4. Coverage discrepancy of the empirical confidence regions in simulation design II. The empirical confidence regions for a range of probability levels ($\alpha = 0.005$ to 0.5) are constructed by five methods (distfree, ellipse, bagplot, convex hull peeling and HPDregionplot) based on a small sample (n = 200) and a large sample (n = 10,000) taken from a bivariate noncentral *F*-distribution with the correlation between two variables of r_{xy} = 0, 0.5 and 0.9. The confidence region by bagplot is available only at $\alpha = 0.5$.
doi:10.1371/journal.pone.0081179.g004

$$x = -\frac{l_{\theta,i}\sqrt{1+\tan(\theta)^2} - l_{\theta+d,i}\sqrt{1+\tan(\theta+\Delta)^2}}{\tan(\theta)-\tan(\theta+\Delta)} \quad (8)$$

$$y = -\frac{l_{\theta,i}\sqrt{1+\tan(\theta)^2}\times\tan(\theta+\Delta) - l_{\theta+d,i}\sqrt{1+\tan(\theta+\Delta)^2}\times\tan(\theta)}{\tan(\theta)-\tan(\theta+\Delta)},$$

where, $i = 1,2$ (see section C of Appendix S1 for detailed derivation).

Results

Simulation studies

The performance of our new method is evaluated by analyzing simulation data. We simulate bivariate data with two variables x and y. Three bivariate sampling distributions are considered in our simulations. In simulation I, x and y are sampled from a bivariate normal distribution $N(\mu,\Sigma)$, where $\mu = \begin{bmatrix}0\\0\end{bmatrix}$ and $\Sigma = \begin{bmatrix}1 & r_{xy}\\ r_{xy} & 1\end{bmatrix}$ with r_{xy} being the correlation between variables, x and y. In simulation II, the two variables (x and y) are generated from a bivariate noncentral *F*-distribution following the approach of Song and Hsiao [21]. The marginal *F*-distribution of each of the two variables is specified as $F(d_1,d_2 = 30,\lambda = 10)$, where d_1 and d_2 are degrees of freedom and λ is the noncentrality parameter. In simulation III, the two variables (x and y) are generated from a mixture of two bivariate normal distributions which is given by $\frac{2}{3}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & r_{xy}\\ r_{xy} & 1\end{bmatrix}\right) + \frac{1}{3}N\left(\begin{bmatrix}3\\3\end{bmatrix},\begin{bmatrix}1 & r_{xy}\\ r_{xy} & 1\end{bmatrix}\right)$. In all three simulations, the correlation r_{xy} takes three values of 0, 0.5 and 0.9. In each simulation, we take $n = 200$ and $n = 10{,}000$ pairs of $x-y$ observations from the distribution to represent small and large samples, respectively.

For each data, empirical CRs are constructed using our new method (distfree.cr/R, http://statgen.ualberta.ca), the classical ellipsoidal confidence region approach [2] implemented by the CAR package [22] in R [23] and other three nonparametric methods, the HPDregionplot in the emdbook/R package [16], the classic convex hull peeling [12], and data peeling based on the Tukey's depth [24]. The CRs are constructed for seven significance levels, $\alpha = 0.005$, 0.01, 0.025, 0.05, 0.1, 0.2 and 0.5. However, only one level of significance $\alpha = 0.5$ is used for the peeling approach based on the Tukey's depth because we use the bagplot approach [24], via the bagplot function in the aplpack/R package [25], to implement the peeling based on Tukey's depth, but both the method [24] and the software implementation [25] are developed exclusively for $\alpha = 0.5$ (Dr Peter Wolf, private communication). We develop an R code to implement the classical convex hull peeling approach based on its definition (available at http://statgen.ualberta.ca). The adequacy of the CRs is measured using coverage discrepancy plots [26] for each simulation run, i.e.,

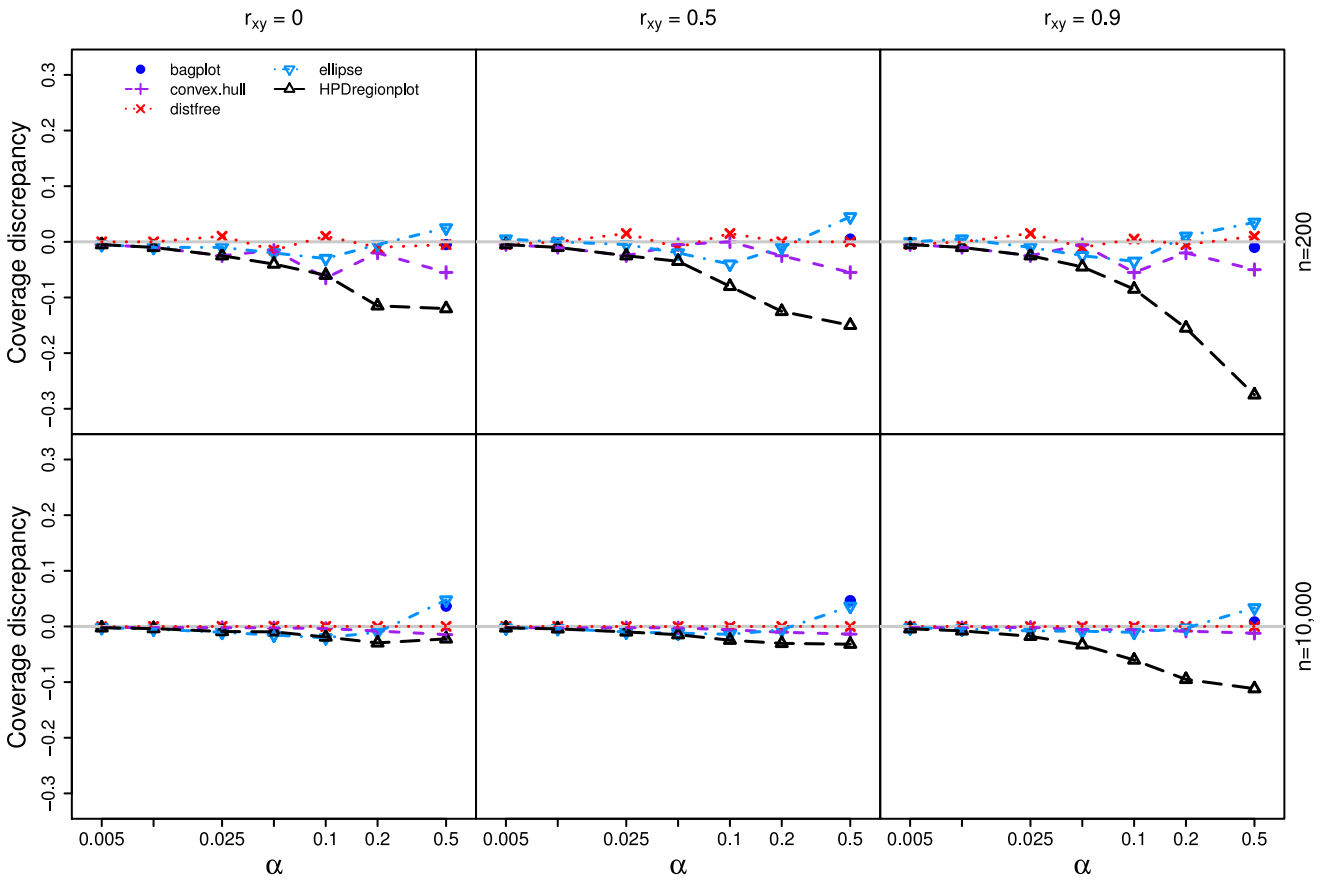

Figure 5. Coverage discrepancy of the empirical confidence regions in simulation design III. The empirical confidence regions for a range of probability levels ($\alpha = 0.005$ to 0.5) are constructed by five methods (distfree, ellipse, bagplot, convex hull peeling and HPDregionplot) based on a small sample (n = 200) and a large sample (n = 10,000) taken from a mixture of two bivariate normal distributions which is given by $\frac{2}{3}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & r_{xy}\\ r_{xy} & 1\end{bmatrix}\right)+\frac{1}{3}N\left(\begin{bmatrix}3\\3\end{bmatrix},\begin{bmatrix}1 & r_{xy}\\ r_{xy} & 1\end{bmatrix}\right)$, where r_{xy} takes 0, 0.5 and 0.9. The confidence region by bagplot is available only at $\alpha = 0.5$.
doi:10.1371/journal.pone.0081179.g005

the deviation of the realized-αestimate of each method to its real value. The realized-αis calculated as the proportion of the observations outside an empirical confidence polygon, which is determined using the pnt.in.poly function in the SDMTools/R [27].

In all three simulations, our method outperforms other methods (Figures 3, 4, and 5) as the realized-α estimates by our method is close to or coincides with the true significance levels for both small (n = 200) and large (n = 10,000) samples with all three r_{xy} values. The classic ellipsoidal method provides overestimation when α is low and underestimation when α is high. All methods including the ellipsoid approach produce similar 95% CRs for the data from the bivariate normal distribution as in simulation I (Figure 6). However, the CRs determined by the ellipsoid approach fail to account for the actual shapes of non-normal sampling distributions as in simulations II and III (Figures 7 and 8). The HPDregionplot is the most sophisticated strategy in capturing the shape of non-normal sampling distribution in all simulations. However, the realized-α estimates by the HPDregionplot approach are constantly lower than the true significance levels; the underestimation tends to increase with the significant level and the correlation (r_{xy}), and it is more pronounced for non-normal data in simulations II (Figure 4) and III (Figure 5) than for normal data in simulation I (Figure 3). It is somewhat surprising to note that the bagplot method performs as well as our method with small sample (n = 200) but it performs poorly with the large sample (n = 10,000) particularly when r_{xy} is high.

Empirical examples

We also analyze three empirical examples to illustrate the use of our new method for the analysis of real data sets. The first data set is taken from Table 4.3 of Rawlings et al. [2]. Since the data set was already described and analyzed by Rawlings et al. [2], we will only recapitulate the essential details of the data. The original data set consisted of physical fitness measurements on 31 men involved in a physical fitness program at the North Carolina State University. The variables measured were age (years), weight (kg), oxygen uptake rate (ml per kg body weight per minute), time to run 1.5 miles (minutes), heart rate while resting, heart rate while running (at the same time oxygen uptake was measured), and maximum heart rate while running. Rawlings et al. [2] carried out the multiple regression analysis to investigate the response of oxygen uptake to the change of time to run 1.5 miles (minutes), heart rate while resting, heart rate while running (at the same time oxygen uptake was measured), and maximum heart rate while running.

For illustration, we only show the CRs of the pairwise regression coefficients as constructed by our new method and the classic methods. The CRs are constructed using the convex hull data peeling approach [12], the classical ellipsoidal method as

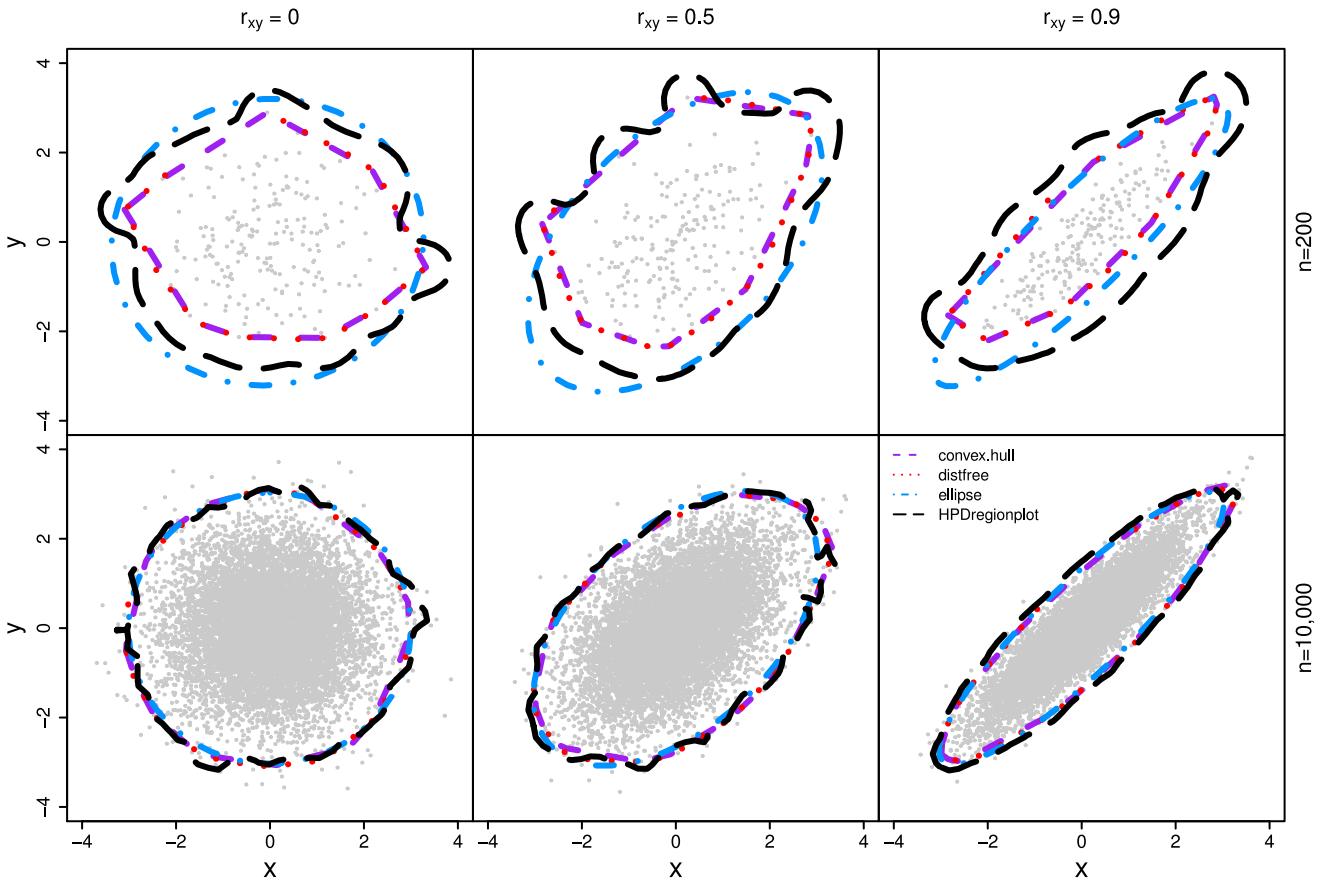

Figure 6. The 95% empirical confidence regions estimated by the four methods (distfree, ellipse, convex hull peeling and HPDregionplot) in simulation I which is detailed in Figure 3.
doi:10.1371/journal.pone.0081179.g006

implemented using the CAR package in R [22], the HPDregionplot in the emdbook/R package [16] and our new geometry-based method (distfree.cr/R, http://statgen.ualberta.ca). Bootstrapping is used to generate 10,000 random samples from the original data. The size of each bootstrap sample is set to 31, the number of individuals as used in the original study. The multiple regression analysis is done for each bootstrap sample. The pairwise regression coefficients as well as their CRs ($\alpha = 0.05$) calculated by the four approaches are plotted (Figure 9). The realized-αvalues are calculated as the proportions of the total observations that lie outside the CRs determined by our new method and the classical methods for all six pairs of regression coefficients. For each pair, the chi-square test statistics is computed to examine the significance of coverage discrepancies of the empirical CRs under the preset significance level of $\alpha = 0.05$. The testing results show the superiority of our new method over the classic methods because the deviations of the realized-αvalues from $\alpha = 0.05$ by our new method are not biased from 0.05 in all pairs whereas there are 4, 6, and 2 pairs with biased realized-α estimates for convex hull peeling, ellipse, and HPDregionplot, respectively.

The second data set is obtained from the 1000 Genomes project [28]. This data set consists of 1,092 human individual records from four super populations, which include 246 Africans (AFR), 181 Ad Mixed Americans (AMR), 286 East Asians (ASN), and 379 Europeans (EUR). For each record, there is an integrated haplotype map of 38 million single nucleotide polymorphisms (SNPs), 1.4 million short insertions and deletions and 14,000 larger deletions. Prior to the analysis, we use the PLINK software [29] to remove the SNPs with minor allele frequency (MAF) of <0.05 and the SNPs with interval sizes smaller than 50 k base pairs in order to have a manageable subset of data. After the removal, a total of 51,529 SNPs remain and we use this subset of the data for the subsequent analysis. Principal component analysis (PCA) as implemented in the EIGENSTRAT software [9] is carried out. The first two principal components are used to generate the scatter plots as well as to construct the 95% confidential regions for individual super populations using the new method as well as the classical methods (Figure 10).

It is evident from Figure 10 that the four methods generate distinctly different CRs particularly for the AFR and AMR populations. The four methods also reveal different patterns of population differentiation. The CRs constructed by the ellipse and HPDregionplot methods suggest that the EUR population is largely contained within the AMR population. In contrast, the CRs constructed by our new method and convex hull peeling approach suggest that the EUR population is somewhat distinguishable from the AMR population. In addition, the realized-α values derived from our new methods are always closer to the prescribed significance level of $\alpha = 0.05$ than those from the classical methods.

The third empirical example is the winter wheat (*Triticum aestivum* L.) data set that has been used (e.g., Yan et al.[30]) for the biplot analysis of genotype×environment interaction. We (Yang et al. [31] and Hu and Yang [32]) have recently analyzed this data

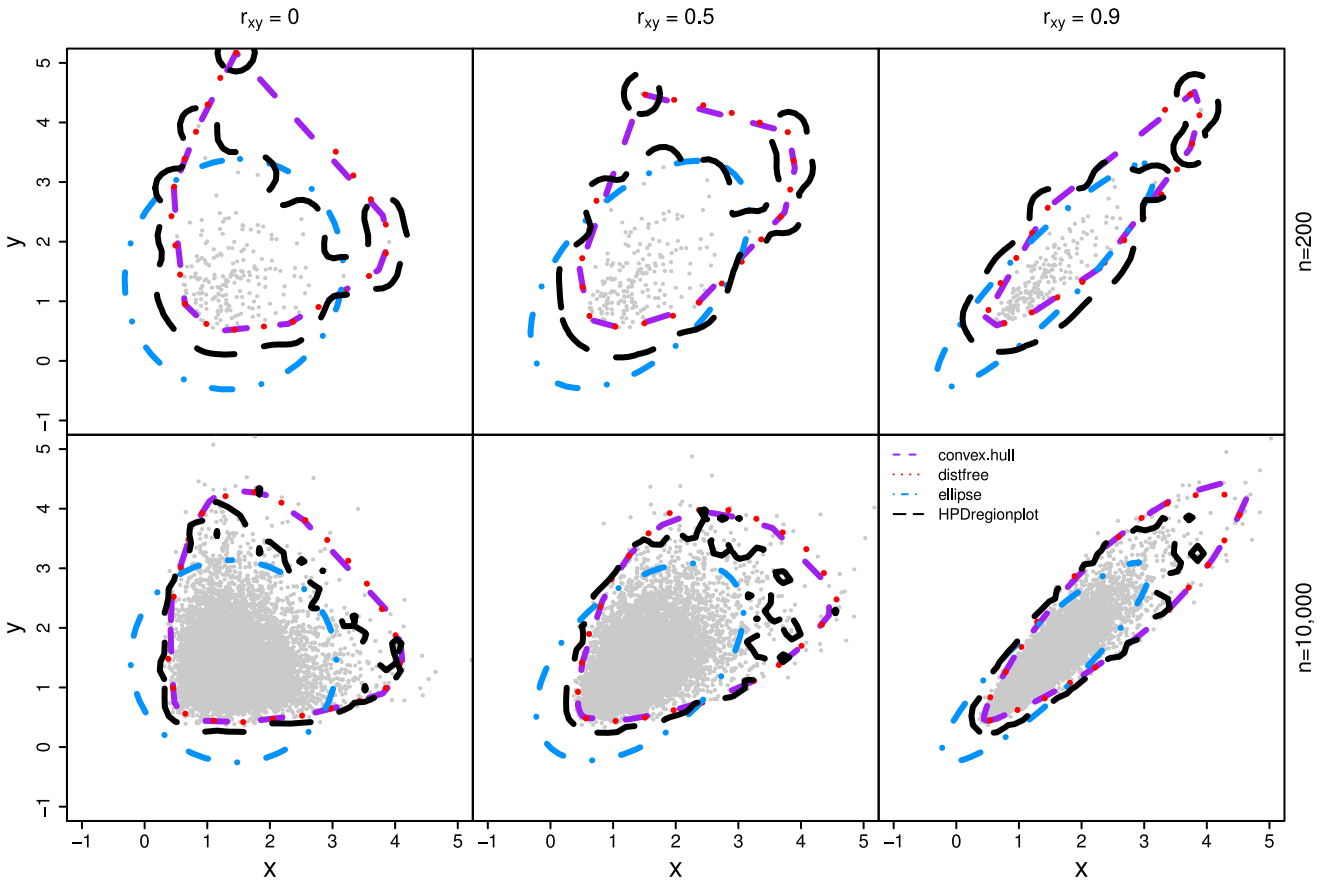

Figure 7. The 95% empirical confidence regions estimated by the four methods (distfree, ellipse, convex hull peeling and HPDregionplot) in simulation II which is detailed in Figure 4.
doi:10.1371/journal.pone.0081179.g007

set as well to illustrate the application of our bootstrapping approach to statistical inference about genotypic and environmental scores obtained from singular value decomposition (SVD) of the two-way genotype×environment table. Here the example serves to show how the CRs constructed for individual genotypic and environmental scores corresponding to the first two principal components (PC1 and PC2) are valuable in pointing out the uncertainty around the mega-environments delineated by the earlier studies. Briefly, the data set consists of the yields of 18 winter wheat genotypes (G1 to G18) tested at nine environments (E1 to E9) in Ontario, Canada. Prior to the analysis, the deviations of cell means for all 162 (18×9) genotype-environment combinations from location means are calculated. The resultant matrix is the basis for bidirectional bootstrapping, SVD and Procrustes rotation as explained in Hu and Yang [32].

The biplot of PC1 vs. PC2 genotypic and environmental scores along with the 95% CR is presented in Figure 11. The PC1 and PC2 account for about 78% of the total variability. To highlight key features in the biplot, the CR are displayed only for those scores that are significantly different from the origin of the biplot [i.e., the CR of the scores that do not include the point of (0,0)]. A hexagon is drawn to connect six genotypes (G3, G7, G8, G12, G13 and G18) that are located at the corners (i.e., vertices) of the hexagon in the biplot. To further facilitate the interpretation of the biplot, six line segments perpendicular to different sides of the polygon are drawn through the origin to subdivide the polygon into six sectors involving different subsets of environments and genotypes: the genotype at the corner of each sector is considered as the 'best' performer in the environments included in that sector as often claimed in the earlier studies (e.g., Yan et al. [30]). However, it is evident from the 95% CR of the scores that the 'best' genotypes are often not statistically different from other genotypes. For example, genotype G8 at the upright corner is indistinguishable from genotypes G4 and G10 in the same sector, judging from their overlapped CR. Simple visual inspection of the biplot [30] claimed that genotype G18 yielded more than genotype G8 in eastern Ontario (represented by E5 and E7) and G8 yielded more than G18 in southwestern Ontario (represented by the other seven environments). With the 95% CR being now attached to individual scores (Figure 11), this claim is no longer true because the CRs for G8 and G18 overlap. Thus, identification of superior genotypes or mega-environments based on the initial inspection of biplots is simply a curious visual observation only and it must be substantiated by subsequent parametric or non-parametric statistical assessments before being recommended for practical utility.

Discussion

In this study, we develop a new geometry-based, distribution-free approach to constructing the CR for two or more variables. Our new method is based only on a few basic geometrical principles and accounts for the actual shape of the distribution (Figures 1 and 2). Thus, it should be a significant complement to the existing parametric (ellipsoidal [2]) and nonparametric

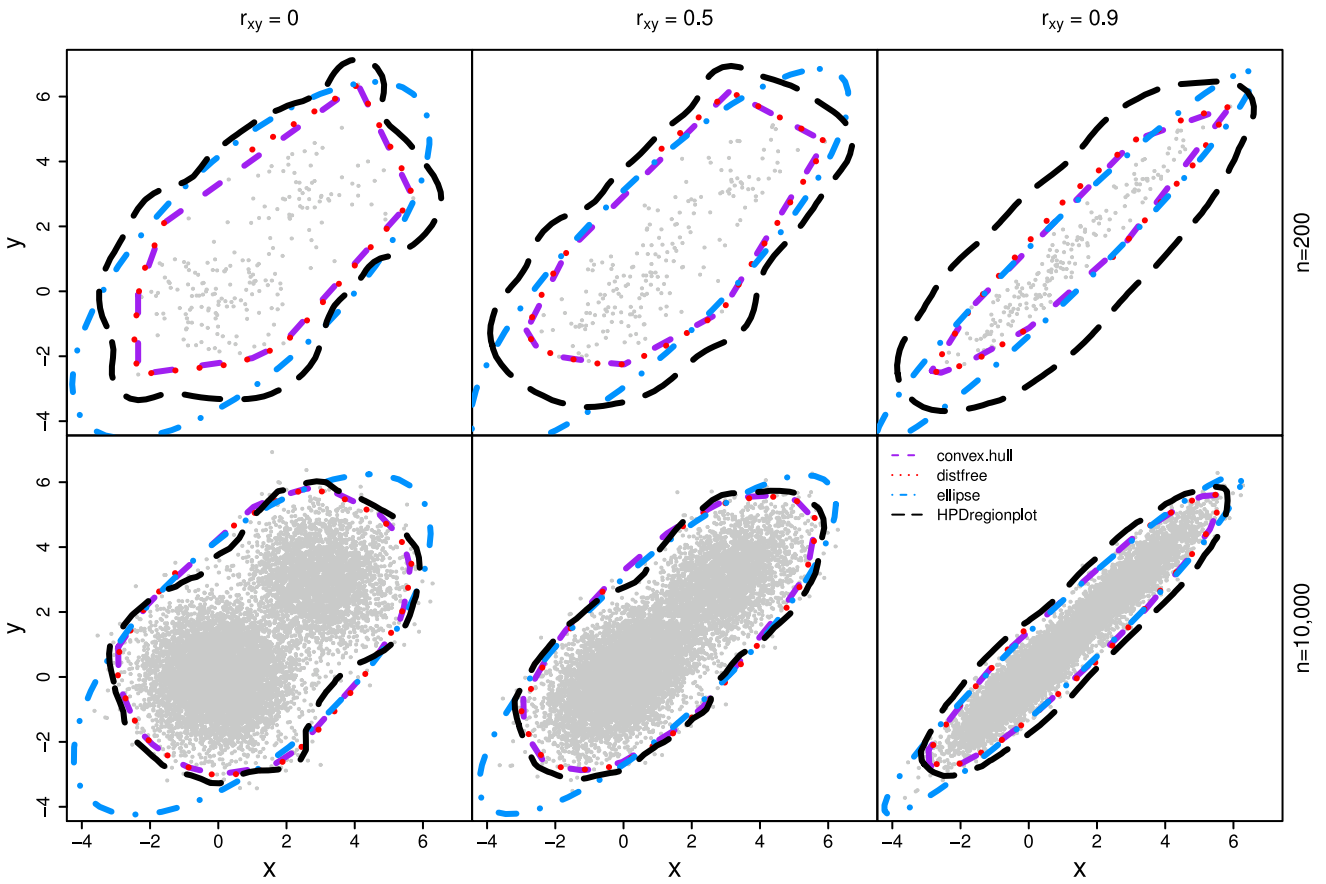

Figure 8. The 95% empirical confidence regions estimated by the four methods (distfree, ellipse, convex hull peeling and HPDregionplot) in simulation III which is detailed in Figure 5.
doi:10.1371/journal.pone.0081179.g008

methods including bagplot [16], convex hull peeling [12], and HPDregionplot [25]).

Our method outperforms other parametric and non-parametric approaches to constructing CRs judging from coverage discrepancy plots of realized-α estimates. It is evident from Figures 3, 4, and 5 that our method always provides more accurate estimates of α than the other methods regardless of whether the sampling distribution is normal (simulation I) or not (simulations II and III). In addition, the superiority of our method is consistent over different levels of correlation between the two variables. So why is our method better? Simply put, it is the only method that accommodate for the actual shape of the distribution and allows for adjusting the realized-αvalue to an individual data point level. While the convex hull peeling and data peeling based on Tukey's depth can also account for the shape of the actual distribution represented by the original data, the realized-αvalue may still be different from the true α because the CR is determined by a 'peeling' layer. Thus, all the data points on the same layer have to be included or excluded simultaneously once the layer is determined as the border of the CR. The true α value can be under- or over-estimated unless each peeling layer consists of only one data point, an unlikely scenario for not too small samples or unless, by chance, the peeling layer along with outer layers constitute the exact α value.

The realized-α estimates by the parametric ellipsoidal method and semi-parametric HPDregionplot may also be biased, but for a different reason. In these methods, the original data are used merely to estimate parameters. It is these estimated parameters along with assumed normal distribution, rather than the original data that are used for constructing CRs. If the data is normally distributed, an unbiased estimate of α can be achieved; if, on the other hand, the data is from a non-normal distribution, the estimate of α may be biased upward or downward. If the true CR is a concave polygon or a crescent moon or the union of disjoint convex areas, then the HPDregionplot is the *only* method that is capable of capturing the true shape of the CR(e.g., the shape of the simulated distribution in simulation III). However, the HPDregionplot may produce the CRs with multiple isolated polygons for small sample sizes (e.g., simulation II for n = 200). Furthermore, in the current version of the emdbook/r package (version 1.3.2.1) on CRAN [16], the HPDregionplot function may also generate unclosed rather than closed polygons for CRs. In an attempt to address this issue, Dr. Ben Bolker, the author of the emdbook/r package, provided us with a set of new parameters for HPDregionplot function (private communication). While the use of these new parameters guarantees the closed polygons by extending the regions for the kde2d function, the polygons derived by the new HPDregionplot function are slightly larger than that calculated by the previous version, thereby leading to the underestimation of the realized-α values. Unfortunately, there is currently no solution to the issue. The HPDregionplot approach works well with accurate estimates of the empirical kernel density.

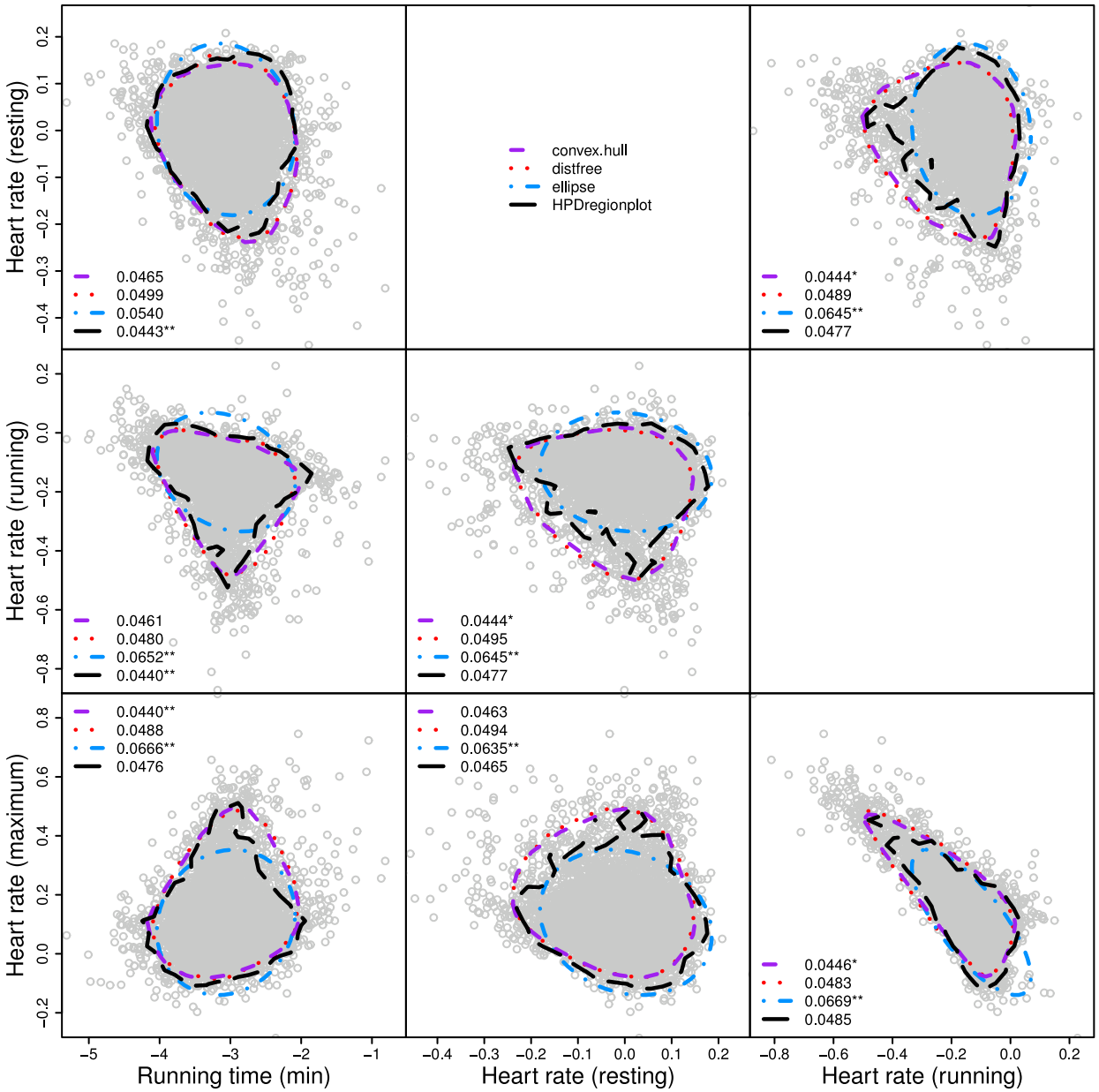

Figure 9. The joint confidence regions of the regression coefficient estimates of the physical fitness measurements on 31 men involved in a physical fitness program at the North Carolina State University. The numbers in the figure are the realized-α values of the corresponding confidence regions. * and ** indicate significant deviations of the realized-αvalues from $\alpha = 0.05$, according to chi-square tests, at $P<0.05$ and $P<0.01$, respectively.
doi:10.1371/journal.pone.0081179.g009

High information content in the original data would be especially important for accurate estimation. This is probably why higher correlation between the two variables has caused greater discrepancy between the realized and true α values (Figures 3, 4, and 5). However, no similar trend is observed when the autocorrelation within the variables is considered (Figures S1-S4).

As shown above, the coverage discrepancy is a necessary criterion for evaluating the performance of different methods for constructing CRs. Nevertheless, it is not a sufficient criterion. For example, it is evident from Figure 4 that, in simulation II, the realized α estimates by the ellipsoidal method are biased upward with low α, but downward with high α. An inflexion point exists near $\alpha = 0.05$ where there is little coverage discrepancy. However, this coincidence does not necessarily mean that the ellipsoidal-based CR can be used to approximate the CR for the sample taken from an F-distributed data because there is bias at all other α levels. It is shown (Figure S5) that the point of the transition from over- to under-estimation of α changes with the degrees of freedom for the F-distributions, but there is little dependence on the noncentrality parameter.

Since each curve in the coverage discrepancy plot (Figures 3, 4, and 5) is calculated from a single random sample, the repeatability of the coverage discrepancy patterns revealed by the plots may be questioned [26]. To confirm the results in Figures 3, 4, and 5, ten additional random samples are generated from the three simulated bivariate distributions described earlier. The coverage discrepancy

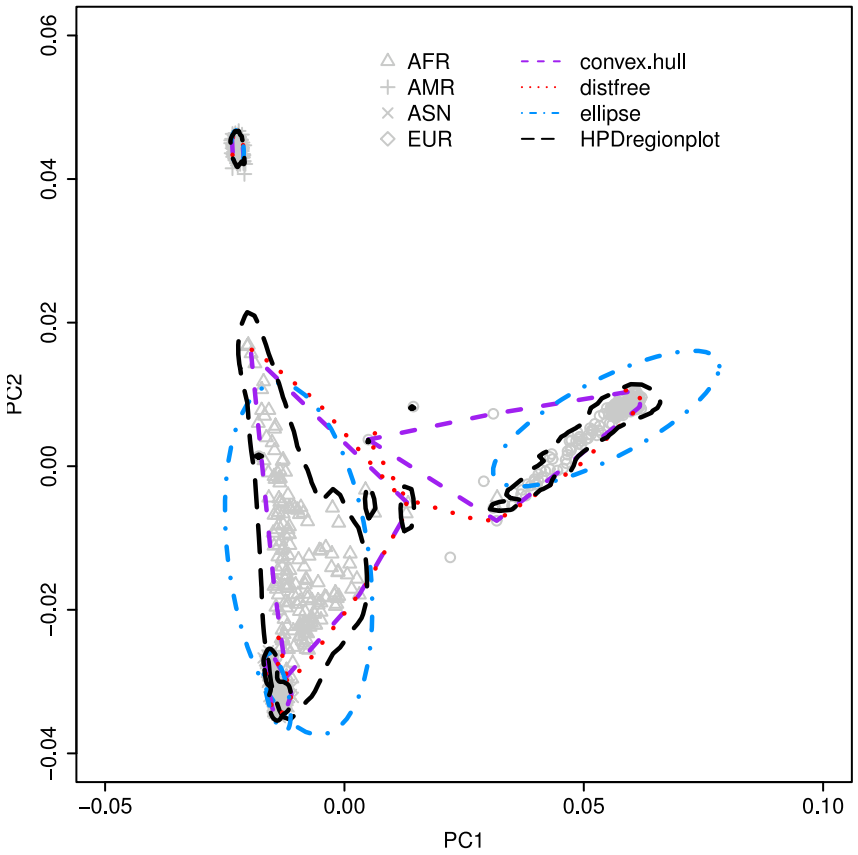

Figure 10. Plots of 1,092 human individuals in 2-D space using the scores of the first two principal components as calculated by EIGENSTRAT based on 51,529 SNP markers. The polygons represent the 95% confidential regions of four individual populations: AFR for African, AMR for Ad Mixed American, ASN for East Asian, and EUR for European.
doi:10.1371/journal.pone.0081179.g010

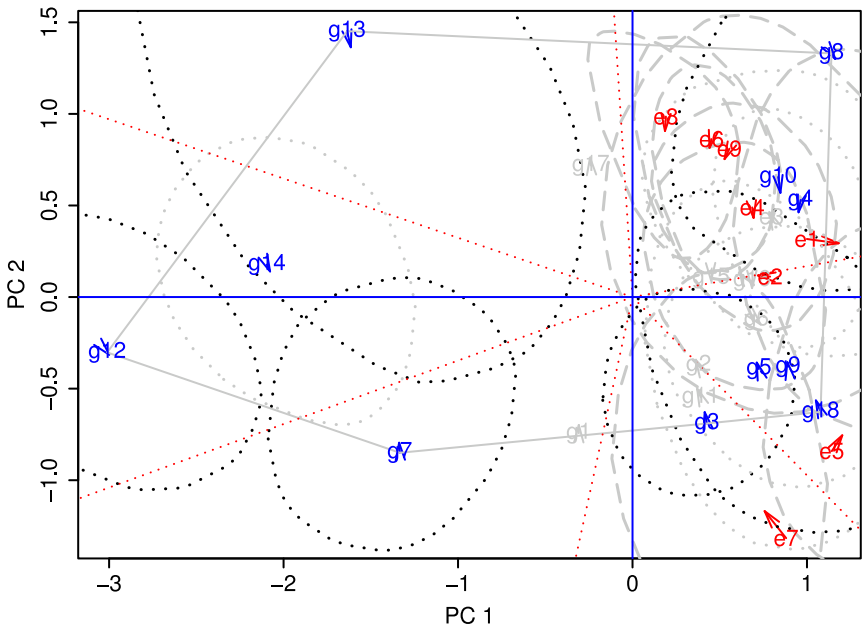

Figure 11. Biplot of 18 genotypic scores and nine environmental scores from the Ontario winter wheat data. The 95% confidence regions are constructed for the genotypic and environmental scores using 10,000 bootstrap samples.
doi:10.1371/journal.pone.0081179.g011

curves by the five methods are displayed in Figure S6. The plots show that the patterns revealed by the coverage discrepancy curves are fairly stable across different samples.

We provide detailed descriptions of our new distribution-free approach to constructing CR for two parameters only. This does not mean that it works only for the two-dimensional data. In fact, our method can be extended to higher-dimension situations. In constructing a CR for three or more parameters, we need to calculate the distances between the data points and reference planes (three variables) or reference hyperplanes (four or more variables). For example, the formula for the distance between the ith point in the three-dimensional space $\{x_i,y_i,z_i\}$ and the reference plane $(ax+by+cz+e=0)$ is given by Korn and Korn [33],

$$\begin{aligned} d_i &= \frac{|ax_i+by_i+cz_i+e|}{\sqrt{a^2+b^2+c^2}} \\ &= \frac{|\cos\phi_x x_i+\cos\phi_y y_i+\cos\phi_z z_i+e'|}{\sqrt{\cos^2\phi_x+\cos^2\phi_y+\cos^2\phi_z}} \end{aligned} \quad (9)$$

The second part of equation (9) is obtained using the 'normal' form of the reference plane (a normal line is the line perpendicular to the reference plane),

$$\cos\phi_x x_i+\cos\phi_y y_i+\cos\phi_z z_i+e'=0$$

where

$$\begin{aligned} \cos\phi_x &= \frac{a}{\sqrt{a^2+b^2+c^2}}, \quad \cos\phi_y = \frac{b}{\sqrt{a^2+b^2+c^2}}, \\ \cos\phi_z &= \frac{c}{\sqrt{a^2+b^2+c^2}}, \quad e' = \frac{e}{\sqrt{a^2+b^2+c^2}}, \end{aligned} \quad (10)$$

with ϕ_x, ϕ_y and ϕ_z being the angles between the normal line and axis x, axis y and axis z, respectively, and e' being the distance between the reference plane and the origin. The actual implementation requires the following two considerations: (1) the sample size required to construct a reliable CR is exponentially increased with the addition of variables; and (2) the amount of computation under higher dimension circumstances is escalating as more reference lines need to be taken into account while constructing the high-dimensional CR. Nevertheless, further research is needed for implementing and interpreting the multidimensional CRs.

Although the normal distribution has been widely assumed in the past [1,2], the joint sampling distribution of the pairwise regression coefficients that are obtained from the data of the oxygen intake experiment by bootstrapping is evidently deviated from a bivariate normal distribution (Figure 9). Thus the basic assumption required for constructing ellipsoidal CRs may often be incorrect and this might lead to distorted CRs and thus to incorrect practical uses.

The second empirical example serves to demonstrate the use of our new method for adding the statistical inference capability to one of the most popular tools currently used in human population genomics. The correction for population stratification is an essential step towards eliminating spurious genetic effects in the genome-wide association study (GWAS) of admixed populations [34]. Cavalli-Sforza et al. [6] proposed the use of the principal component analysis (PCA) for detecting the stratification among human populations. Recently, the strategy has been further developed and adopted in using genomic data for the analysis of population stratification in human [7,8,9,10]. The effectiveness of such PCA-based detection depends on correct inference about the ancestry and population structure. Currently, the commonly used means of inferring the population stratification is the use of scatter plots of the first few principal components known as "radiation of circular or elliptic clines from a specification area" or the "principal-component map" [6]. However, the determination of population sharing or membership based on these plots or maps is somewhat arbitrary because it is based solely on visual inspection. Since the sampling distributions of the principal component scores derived from SNP markers are unknown, the use of the classical ellipsoidal method for constructing the CRs may not be adequate. The third example shows further utility of our new method for strengthening the biplot analysis of genotype×environment interaction. Thus, our distribution-free approach to constructing any multivariate CRs provides a statistical basis for such determination.

Supporting Information

Figure S1 The impact of autocorrelations (0, 0.5 and 0.9) on the coverage discrepancy plots for small sample $n=200$ in simulation I which is detailed in Figure 3.
(EPS)

Figure S2 The impact of autocorrelations (0, 0.5 and 0.9) on the coverage discrepancy plots for large sample $n=10{,}000$ in simulation I which is detailed in Figure 3.
(EPS)

Figure S3 The impact of autocorrelations (0, 0.5 and 0.9) on the coverage discrepancy plots for small sample $n=200$ in simulation II which is detailed in Figure 4.
(EPS)

Figure S4 The impact of autocorrelations (0, 0.5 and 0.9) on the coverage discrepancy plots for large sample $n=10{,}000$ in simulation II which is detailed in Figure 4.
(EPS)

Figure S5 The effect of F distribution with equal degrees of freedom ($d_1 = d_2$) on the coverage discrepancy plots of the empirical confidence regions as approximated by a normal distribution.
(EPS)

Figure S6 Coverage discrepancy plots based on 10 independent simulated samples of sizes n = 200 and n = 10,000. The two variables are assumed independent.
(EPS)

Figure S7 The confidence region constructed for two reference lines.
(EPS)

Appendix S1 Derivation of equations.
(DOCX)

Acknowledgments

We thank Professor Ben Bolker of the Department of Mathematics & Statistics, McMaster University for his help with the use of emdbook/r

package, HPDregionplot. We thank Dr. Hans Peter Wolf of the Fakultät für Wirtschaftswissenschaften of the Universität Bielefeld for his help with the use of aplpack/r package, bagplot.

Author Contributions

Conceived and designed the experiments: ZH R-CY. Performed the experiments: ZH. Analyzed the data: ZH. Contributed reagents/materials/analysis tools: R-CY ZH. Wrote the paper: ZH R-CY.

References

1. Draper NR, Smith H (1998) Applied Regression Analysis. New York: John Wiley & Sons.
2. Rawlings JO, Pantula SG, Dickey DA (1998) Applied Regression Analysis: A Research Tool. New York, NY: Springer-Verlag.
3. Davidson R, MacKinnon JG (2004) Econometric theory and methods. New York: Oxford University Press.
4. Yang R-C, Crossa J, Cornelius PL, Burgueño J (2009) Biplot analysis of genotype x environment interaction: Proceed with caution. Crop Sci 49: 1564–1576.
5. Denis J-B, Gower JC (1996) Asymptotic confidence regions for biadditive models: Interpreting genotype-environment interactions. J R Stat Soc Ser C Appl Stat 45: 479–493.
6. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. Princeton, N.J.: Princeton University Press.
7. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40: 646–649.
8. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, et al. (2007) PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet 3: e160.
9. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
10. Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, et al. (2005) Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2: 81–89.
11. Porzio GC, Ragozini G (2000) Peeling multivariate data sets: a new approach. Quad Stat 2: 85–99.
12. Green PJ (1981) Peeling bivariate data. In: Barnett V, editor. Interpreting multivariate data. UK: Wiley.
13. Yeh AB, Singh K (1997) Balanced confidence regions based on Tukey's depth and the bootstrap. J Roy Stat Soc B Met 59: 639–652.
14. Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: Descriptive statistics, graphics and inference. Ann Stat 27: 783–840.
15. Donoho DL, Gasko M (1992) Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness. Ann Stat 20: 1803–1827.
16. Bolker B (2012) emdbook: Ecological Models and Data in R. R package version 1.3.4.
17. Venables WN, Ripley BD (2002) Modern Applied Statistics With S; Chambers J, Eddy W, Härdle W, Sheather S, Tierney L, editors. New York, NY: Springer.
18. Petitjean M, Saporta G (1992) On the performance of peeling algorithms. Appl Stoch Model D A 8: 91–98.
19. Hyndman RJ, Fan YN (1996) Sample quantiles in statistical packages. Am Stat 50: 361–365.
20. Schoonjans F, De Bacquer D, Schmid P (2011) Estimation of population percentiles. Epidemiology 22: 750–751.
21. Song WMT, Hsiao LC (1993) Generation of autocorrelated random variables with a specified marginal distribution. 1993 Winter Simulation Conference Proceedings 374–377.
22. Fox J, Weisberg S (2011) An R Companion to Applied Regression. Thousand Oaks CA: Sage.
23. R Core Team (2012) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
24. Rousseeuw PJ, Ruts I, Tukey JW (1999) The bagplot: A bivariate boxplot. Am Stat 53: 382–387.
25. Wolf P, Bielefeld U (2012) aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, and some slider functions. R package version 1.2.7.
26. de Peretti C, Siani C (2010) Graphical methods for investigating the finite-sample properties of confidence regions. Comput Stat Data Anal 54: 262–271.
27. VanDerWal J, Falconi L, Januchowski S, Shoo L, Storlie C (2012) SDMTools: Species distribution modelling tools: Tools for processing data associated with species distribution modelling exercises. R package.
28. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.
29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
30. Yan W, Kang MS, Ma B, Woods S, Cornelius PL (2007) GGE Biplot vs. AMMI Analysis of Genotype-by-Environment Data. Crop Science 47: 641–653.
31. Yang R-C, Crossa J, Cornelius PL, Burgueño J (2009) Biplot Analysis of Genotype×Environment Interaction: Proceed with Caution. Crop Science 49: 1564–1576.
32. Hu Z, Yang R-C (2013) Improved statistical inference for graphical description and interpretation of genotype×environment interaction. Crop Science doi: 10.2135/cropsci2013.2104.0218.
33. Korn GA, Korn TM (1968) Mathematical Handbook for Scientists and Engineers: Definitions,Theorems, and Formulas for Reference and Review. New York: McGraw-Hill.
34. Seldin MF, Pasaniuc B, Price AL (2011) New approaches to disease mapping in admixed populations. Nature reviews Genetics 12: 523–528.