ABSTRACT: The plot of the first two principal coordinates as often derived from the analysis of large-scale SNP data is essential for identifying the population stratification that may lead to high false positive rate in genome-wide association studies. The current use of bivariate plot is for visualization without confidence regions (CRs) needed to justify the observed population structures. Parametric approaches are often used but sampling distributions of parameters estimated from SNPs are unknown. The objective of this paper is to describe a new nonparametric method for constructing CRs without knowing the sampling distributions. The statistical properties of the new method are evaluated using simulation data and its application is illustrated with the analysis of a human SNP data. Several parametric and nonparametric methods of constructing CRs are also compared with the new method in both simulation and real data analysis.

Keywords: confidence region, nonparametric statistics, population genomics.

INTRODUCTION

The information of population structure provides an essential basis to identify the stratification that needs to be accounted for in inferring about demographic and evolutionary factors or minimizing false positive rates in genome-wide association studies (The 1000 Genomes Project Consortium (2012)). However, such population structure is often identified based only on visual judgment of bivariate plots (biplots) using the first two principal coordinates from the principal component analysis (PCA) of SNP or other marker data. No statistical inference based on confidence regions (CRs) of biplots is available for delineation.

Construction of CRs often assumes that the data are from a normal distribution and they are balanced (Rawlings, et al. (1998)). For example, for bivariate normally-distributed data, the required confidence region is an ellipse whose shape depends largely on the level of the correlation between the two variables. When the distribution is unknown or hard to be characterized, construction of CRs is based on several nonparametric procedures including data peeling (Porzio and Ragozini (2000)), convex hull and data depth (Liu, et al. (1999); Yeh and Singh (1997)) and HPDregionplot (Bolker (2012)).

One of the key limitations with these parametric and nonparametric methods is the inaccurate estimation of the coverage rate by the confidence regions with the data of unknown distributions. All the nonparametric methods are computationally demanding (Petitjean and Saporta (1992)) and some of them (e.g., HPDregionplot) are sensitive to small sample sizes. Hu and Yang (2013) recently developed a simple nonparametric geometry-based procedure that allows for constructing CRs for two or more parameters with unknown sampling distributions. In this paper, we will illustrate its use for delineating spatial population structure of genomic variation using a SNP data set from the 1000 Genome Project (The 1000 Genomes Project Consortium (2012)).

MATERIALS AND METHODS

The geometry model. Here we briefly describe the conceptual model of the geometry-based approach to constructing CRs for any bivariate data as developed by Hu and Yang (2013). This can be best done through a couple of figures (Figures 1 and 2). In Figure 1, the two parallel dashed lines are the boundaries of the confidence interval for the reference line that has the angle (θ) with the horizontal abscissa. 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. It should be pointed out that the relative position (distance) of the ith observation (xi, yi) to the reference line is given by

\[ d_i = \frac{\tan(\theta) x_i - y_i}{\sqrt{\tan(\theta)^2 + 1}} \]  

for i = 1 to N (the total number of observations). For a specified significance level α, the confident interval with respect to the reference line is flanked by the observed lower- and upper-boundaries, i.e., the \((N \times \alpha / 2)\)th and \([N \times (1 - \alpha / 2)]\)th percentiles. In geometry view, the boundaries \(l_{\theta_1} = N \times \alpha / 2\) and \(l_{\theta_2} = N \times (1 - \alpha / 2)\) represent two parallel lines 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 ith boundary line in an arbitrary direction in the plane is given as

\[ y = \tan(\theta)x + l_i \sqrt{1 + \tan(\theta)^2}, i = 1, 2 \]  

In Figure 2, two polygons are constructed by a set of confidence intervals with each confidence interval being obtained through repeatedly rotating the reference line (Figure 1) in all directions over the plane. The inner polygon is formed using the overall significant level α as an initial value for determining the confidence interval at each direction. However, this polygon would lead to an underestimation of the desired CR by the amount δ. An iterative procedure is used to obtain the outer polygon with
the observed significant level approximating to the desired significant level $\alpha$. The initial value for obtaining the inner polygon can be arbitrary. For example, we can also use the Bonferroni correction to get an initial value. In this case, the resulting polygon may be outside the current outer polygon and thus compressing instead of expanding the polygon is needed to achieve the desired significant level $\alpha$.

Simulation. We simulate bivariate data with two variables $x$ and $y$. Three simulation scenarios are considered. 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 (1993). 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 $\lambda$ 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 $\frac{2}{3} N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{bmatrix} \right) + \frac{1}{3} N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 & r_{xy} \\ r_{xy} & 3 \end{bmatrix} \right)$. In all three scenarios, 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. The R code for simulation is available at our website (distfree.cr/R, http://statgen.ualberta.ca).

**Empirical data.** The data set is obtained from The 1000 Genomes Project Consortium (2012). 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 (Purcell, et al. (2007)) to remove the SNPs with minor allele frequency (MAF) of $<0.05$ and the SNPs with interval sizes smaller than 50k 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 (Price, et al. (2006)) is carried out.

**RESULTS AND DISCUSSION**

**Simulation results.** For illustration, we only present part of the simulation results for $n = 10000$ and $\alpha = 0.10$. In all three simulation scenarios, our method outperforms other methods as the realized-$\alpha$ estimates by our method is close to or coincides with the true significance level ($\alpha = 0.10$) with all three $r_{xy}$ values (Table 1). The classic ellipsoidal method provides overestimation when $\alpha$ is low and underestimation when $\alpha$ is high. In addition, the confidence regions determined by the ellipsoid approach fail to account for the actual shapes of non-normal sampling distributions as in simulations II and III. The HPDregionplot is the most sophisticated strategy in capturing the shape of non-normal sampling distribution in all simulations. However, the realized-$\alpha$ 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

**Figure 1.** A conceptual model for constructing the confidence interval with an arbitrary reference line.

**Figure 2.** A conceptual model for building the confidence region and polygon boundaries as obtained by rotating the reference line.
Empirical results. The first two principal components are used to generate the scatter plots as well as to construct the 95% confidential regions for individual superpopulations using the new method as well as the classical methods (Figure 3). The four methods generate distinctly different CRs particularly for the AFR and AMR populations and they reveal different patterns of population differentiation. The CRs by the ellipse and HPDregionplot methods show that the EUR population is largely contained within the AMR population. In contrast, the CRs by our new method and convex hull peeling approach show that the EUR population is somewhat distinguishable from the AMR population. The overall rates of misclassification are 4.95% for our method, 3.21% for convex hull, 32.78% for ellipse and 25.82% for HPDregionplot. Obviously, the misclassification rate by our method is very close to the prescribed significance level of 5%.

Conclusion
Both the simulation and empirical examples demonstrate that the use of CRs by our new method can add to the statistical inference capability to delineate spatial genomic variation for the analysis of population stratification in humans. The correction for population stratification is an essential step towards eliminating spurious genetic effects in the genome-wide association study (GWAS) of admixed populations. In the past, the use has been made of the principal component analysis (PCA) for detecting the stratification among human populations and more recently, the strategy has been further developed and adopted in using genomic data for the analysis of population stratification in human (Price, et al. (2006)). 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" (Cavalli-Sforza, et al. (1994)). 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. Thus, our nonparametric approach to constructing any bivariate CRs provides a statistical basis for such determination.

LITERATURE CITED
Bolker B. (2012). emdbook: Ecological Models and Data in R.