Benford's Law Based Outliers Detection for Population Stratification in Genotype Data

Yuan Yuan[1], Nedret Billor[2], Asuman Turkmen[3]
[1]Auburn University, Auburn, AL 36830
[2]Auburn University, Auburn, AL 36830
[3]The Ohio State University, Newark, OH 43055

Abstract
The issue of population stratification refers to that, individual's ancestry groups, if not corrected, would confound GWAS results. To control for this issue, researchers include ancestry information into the association analysis. Yet the existence of outliers could threaten the correct identification of subgroups. In this study, we propose Benford's Law based outlier detection method for genotype data and examine its performance in identifying population structures.

Key Words: Benford's law, SNP data, GWAS, outliers, population stratification.

1. Introduction

Population-based association studies of unrelated individuals are powerful for gene mapping of complex traits (Amos, 2007). However, these studies are susceptible to potential confounding by population. Population stratification stems from the fact that populations are typically heterogeneous in terms of genetic ancestry. In genome-wide association studies (GWAS), the purpose is to identify the alleles that could most differentiate the case group from the control group. Since the GWAS sample is usually composed of people from different nationalities, the structure of population can be a confounder that would impact the results of association test if disregarded. The population stratification either causes spurious association between disease and marker or masks the true association. Therefore, it is important to infer the population structure.

There are a number of methods that have been proposed to overcome confounding issues due to population stratification. Many have proven useful in certain situations, such as genomic control approach (Devlin and Roeder, 1999) and the structured association approach (Pritchard et al., 2000; Rosenberg et al., 2002). These methods have been found to either over-adjust or under-adjust certain SNPs, depending on the ancestral information of individual SNPs. The Principal Components Analysis (PCA) by Price et al. (2016) (EIGENSTRAT) has been proven very useful for correction of population stratification since the first few principal components (PCs) contain the geographical information of the subjects and has low false positive error. Li and Yu (2008) have also proposed combining the multidimensional scaling (MDS) and clustering to deal with population stratification, which performs better than EIGENSTRAT. All of these methods are based on PCA and clustering, assuming that the genotype data are homogeneous—that is, free of outliers. However, genotype data may contain outliers, which would result in misclassification of subpopulations. If classical PCA were applied on the data, in return it would yield identification of spurious associations in GWAS. Therefore, Liu et al. (2013)

proposed combining clustering with robust PCA as an improved approach for correcting for artifacts arising from population stratification. This proposed method consists of three steps: 1, use robust principal component analysis (PCA) (Hubert et al., 2005) method for outlier detection in genotype data and remove them; 2, apply PCA to the clean data to get the top10 PC components and use clustering to deal with population stratification; 3, test each SNP's association with the outcome of interest by building a logistic regression model that includes the specific SNP as one factor, the selected PCs as covariates, and the cluster membership indicators as additional factors.

In order to identify population structure correctly in GWAS, detection of existing outliers and their removal from the dataset are essential steps. In this study, we propose to use an outlier detection method based on the Benford's law (Benford, 1938) via PCA to improve the correct identification of population structure in GWAS studies. Although the Benford's law is known for many years, its application in biological systems was barely investigated until recently (Costas et al. 2008; Karthik et al. 2016). To our best knowledge, this is the first paper that utilizes Benford's Law for outlier detection in GWAS.

A description of the Benford's law and how this method can be utilized for detecting outliers will be given in Section 2. Since our proposed method is implemented through PCA, we then introduce PCA based outlier-detection method for genotype data in Section 3. In order to assess the performance of the proposed method, a simulation study and real data application are given in Section 4. Finally, Section 5 consists of conclusion and discussion.

2. Benford's Law Based Outlier Detection Method

The Benford's law (BF), also known as the first-digit law, states that the larger digits have a lower likelihood to occur in the first digit position in naturally occurring datasets (Newcomb, 1881; Benford, 1938). The idea of using BF to screen data is based on the observation that regular, "naturally generated" data usually follow a logarithmic distribution, while contaminated data show abnormalities in the distribution (Hill, 1995). The leading digits (i.e., first nonzero digit) *d* of a random variable *x* of many real-life sets of numerical distribution had a cumulative probability distribution of

$$P(d) = log_{10}(d+1) - log_{10}(d) = log_{10}\left(1 + \frac{1}{d}\right)$$

in which $\mathrm{x} = \mathrm{y} \times 10^{n}$, y range from [1,10], and *d* is the integer part of *y*. This distribution of d is referred to as the Benford distribution (Newcomb, 1881; Benford, 1938). If a dataset is free from error or fabrication, it would follow the Benford's law; and violation of the Benford's law indicates abnormality. The conformation of a dataset to the Benford's law could be measured with goodness of fit tests (e.g. Chi-square test and Kolmogorov-Smirnov) called Benford's Tests (BF Tests) in general.

The Benford's law has been applied to test error, fraud, and fabrication in real-life datasets. Although the Benford's law is known for many years, its application in biological systems was barely investigated until recently (Costas et al., 2008; Karthik et al., 2016). To the best of our knowledge, this is the first study utilizing Benford's law in GWAS for population stratification.

3. Principal Component Based Outlier Detection Methods for Genotype Data

We want to make sure that we identify population structure correctly in GWAS. Therefore we will make sure that we identify all outliers that may affect the identification of population structure and remove them to improve the correct identification of population structure in GWAS. There are many outlier detection methods in the literature. We will focus on one of the most widely used outlier detection methods, which is based on PCA.

PCA provides insight for detecting multivariate outliers in the presence of collinearity. The detection of outliers via PCA is achieved through the confidence ellipse, which can be determined using the Mahalanobis distance (Mahalanobis, 1936) of the PC scores. This method consists of two steps: 1. Apply PCA to a dataset and retain the first-k dimensions; 2. Compute Mahalanobis distance for each individual using the first-k PC scores, and compare which against a cutoff chi-square score at 97.5 percentile with df=k. The individuals whose Mahalanobis distances are greater than the cutoff value are identified as PCA outliers.

3.1. Robust PCA Based Outlier Detection

Since the classical PCA is not resistant to outliers since it uses the correlation matrix, it yields misleading results in detecting outliers correctly. Therefore, robust version of PCA (Hubert et al., 2005) is suggested. We use the robust score and the orthogonal distances obtained from Robust PCA to identify different type of outliers. For instance, the individuals whose score distances are greater than the chi-square at 97.5% with df=10 are identified as outliers (see the details in Hubert et al., 2005). In current study, the Robust PCA was implemented via R package “ROSPCA” (Reynkens, 2018).

3.2. PCA based Benford’s Law Outlier Detection

When working with a numerical dataset, the Benford’s Test can be directly applied to the first digit distribution of either a whole dataset, or each row of the dataset. However, applying BF Test to a genotype (SNP) dataset whose leading digits are 0, 1, or 2 is trivial. To avoid this issue, we apply the Benford’s Test to the genotype (SNP) dataset using a PCA based procedure.

The purpose of the current study is to apply the Benford’s Test to identify subject outliers on the top 10 PC scores. To achieve this goal, we extract the first digit of individual’s PC scores, and compare it against the Benford’s distribution via goodness-of-fit test, such as Chi-square or Kolmogorov-Smirnov tests. We use R package “BenfordTests” (Joenssen, 2015) for the extraction of the first digits and the goodness of fit tests. The individuals whose the goodness of fit test are significant are identified as Benford’s outliers. The algorithm for the outlier detection procedure described above can be summarized in the following steps:

1. Obtain PC scores of the SNP dataset using classical PCA.
2. Retain first k dimensions of PC scores (i.e. We choose *k*=10 in the current study).
3. Detect outliers based on Chi-square test (i.e. We choose Chi-square test as the goodness of fit test for Benford’s law). The individuals whose Chi-square test results are significant are identified as outliers.

We use R package “BenfordTests” (Joenssen, 2015) for the extraction of the first digits and the goodness of fit tests. We choose a 0.05 alpha level to determine the significance of the Chi-square test.

4. Numerical Examples

4.1. Simulation Study

We conducted a simulation study to assess whether Benford’s law-based outlier detection method could improve the correct identification of population structure in GWAS studies. The simulation study is composed of 6 scenarios (i.e., 2 contamination rates for each of the three types of outliers), each scenario include 100 simulation runs. For each simulation run, we simulate a SNP dataset, contaminate it with different types of outliers, identify outliers using four different outlier detection methods, and test the 10-fold cross-validation classification accuracy before and after the removal of outliers. For each scenario, the simulation steps are as follows:

1. Simulate a 500 by 8000 SNP dataset with two separate groups based on Price et al. (2006) and Xu et al. (2013)’s simulation setup.
2. Contaminate the data with three types of outliers (Extreme, Gaussian, or Arbitrary outliers) with a varying contamination proportions, 5% or 10%.
3. Apply PCA, PCA based Benford’s law (i.e., BF), and ROBPCA to detect outliers.
4. Determine classification accuracy by using 10-fold cross validation via Support Vector Machine (SVM) before and after removing outliers detected in the third step.
5. Compare their performances based on the following criteria:

$$\text{AP} = \text{Accuracy\_after} - \text{Accuracy\_before}$$

If AP>0 outlier removal improves the classification accuracy. If AP<0 outlier removal does not improve the classification accuracy.

For a better understanding of the simulation steps, a flowchart for the simulation steps is provided in Figure 1.

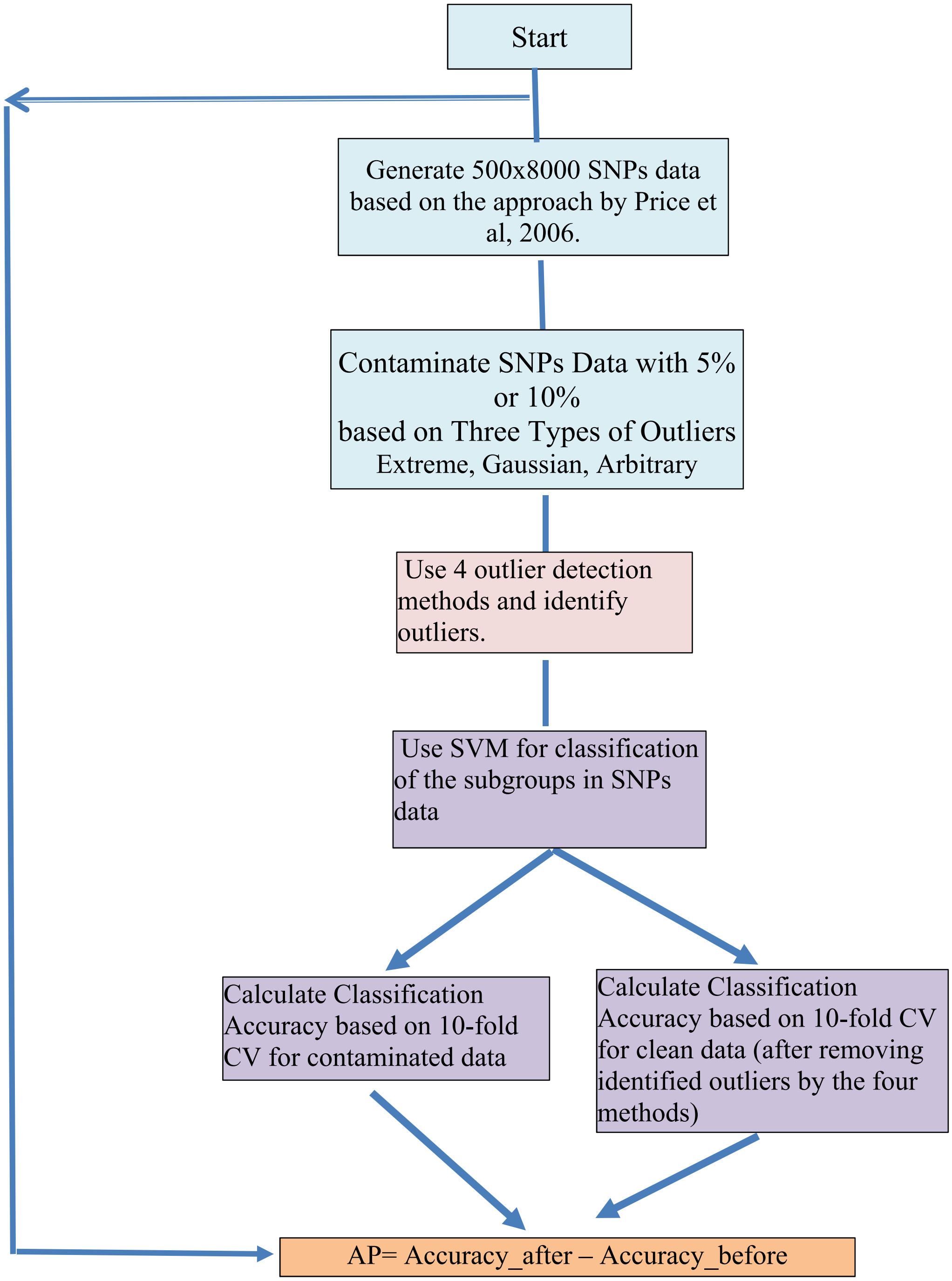

Start
Generate 500x8000 SNPs data based on the approach by Price et al, 2006.
Contaminate SNPs Data with 5% or 10% based on Three Types of Outliers Extreme, Gaussian, Arbitrary
Use 4 outlier detection methods and identify outliers.
Use SVM for classification of the subgroups in SNPs data
Calculate Classification Accuracy based on 10-fold CV for contaminated data
Calculate Classification Accuracy based on 10-fold CV for clean data (after removing identified outliers by the four methods)
AP= Accuracy_after – Accuracy_before

Figure 1. Flowchart of the simulation steps.

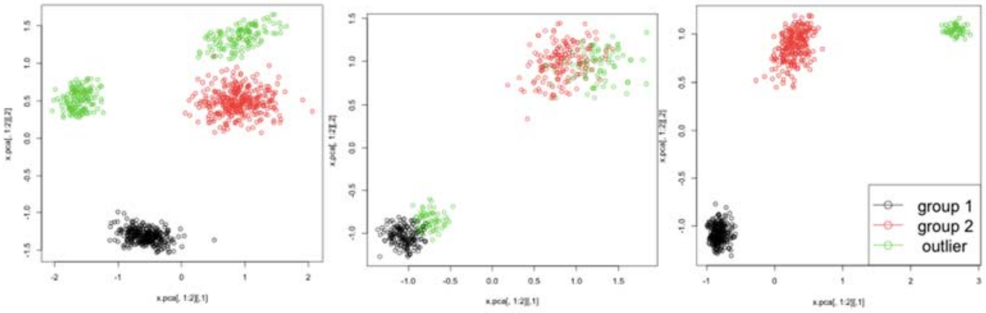
4.1.1. Simulation of SNP dataset
For each simulation run, we generate a 500 by 8000 SNP dataset based on Price et al. (2006) and Xu et al. (2013) using the following procedure.

1. Simulate a 500-by-150 SNP dataset as the case 1 described in Price et al. (2006). Specifically, the SNP dataset is composed of two separate population groups (i.e., 250 individuals for each group), with three categories of SNPs: the common SNPs which has

no difference across population groups, differential SNPs whose MAFs differ between population groups, and casual SNPs whose MAF depends on the case/control conditions.

2. Mapping the SNP dataset from low dimension to high dimensions. We bring the 500-by-150 SNPs dataset (Z) to high dimension (i.e., d=8000) following the steps in Xu et al. (2013). To achieve this goal, we use a 150-by-8000 random matrix A and compute: X = Z *A, in which Z is the original 500-by-150 SNP dataset. We then categorize X in to 0, 1, 2, by setting all entries less than 0 to 0; all entries that greater than 2 were set to 2, and set those between (0, 2) to 1.

4.1.2 Contamination of outliers.
We contaminate data to the high-dimensional SNP matrix X using three methods: Extreme outliers, Gaussian outliers, and Arbitrary outliers which are displayed in Figure 2.



Extreme Gaussian Arbitrary

Figure 2. Outlier types.

Extreme outliers based on the procedures in Liu et al. (2013), in which we add p proportion (i.e., p = 5% or 10 %) extreme values to the 2nd eigenvector of the SNPs dataset. The procedures are as following:

1. Obtain singular value decomposition (SVD) of the X matrix as X= UD V, where U and V are orthogonal matrices of the eigenvectors of XX^T and X^TX, respectively and D is the diagonal matrix of singular values of X.
2. Let u be the 2nd eigenvector of X.
3. Randomly sample *p**500 individuals from this vector and replace those values with randomly generated extreme values from a uniform distribution with [$\boldsymbol{\bar{u}+2s_u}$, $\boldsymbol{\bar{u}+3\sigma s_u}$]. The modified vector u was referred to as u_{mod}.
4.replace the second column of the eigenvector matrix, U, with the modified u_{mod} (i.e., obtain the modified matrix U_{mod}), and compute the SVD of the contaminated matrix X: $X_{mod} = U_{mod}$ D V.
5. Categorize X_{mod} to 0, 1, 2 by setting all entries less than 0 to 0; all entries that greater than 2 were set to 2, and set those between (0, 2) to 1.

Gaussian outliers based on the procedures in Xu et al. (2013):
1. Randomly sample *p**500 individuals from X. Add Gaussian noise to the sampled individuals, in which the Gaussian signal is a *p**500 by 8000 matrix following multivariate normal distributed with mean 0 and covariance matrix I_d.

2. Categorize the modified data matrix X_{mod} by setting all entries less than 0 to 0; all entries that greater than 2 were set to 2, and set those between (0, 2) to 1.

Arbitrary outliers based on the procedures in Xu et al. (2013):
1. Generate a random data matrix of p*500 by 8000 following the uniform distribution.
2. Categorize the random data matrix to 0,1,2 by setting all entries less than 0 to 0; all entries that greater than 2 were set to 2, and set those between (0, 2) to 1,
3. Combine the random data matrix and the X matrix by rows.

4.1.3 Assessment of Classification Accuracy of the PCA-based Outlier Detection Methods in Identification of Population Structure
We compared the classification accuracy of the methods based on PCA, BF, and ROBPCA outlier detection methods using the following steps:

1. For each simulated data with added contamination (5% and 10%), detect outliers based on each method.
2. Use support vector machine (SVM; implemented via R package “e1071”, Meyer et al. 2017) to determine how well the subgroups classified by using 10 fold CV technique for the contaminated data and clean data (i.e., after removing identified outliers)
3. Compute the classification accuracy for the contaminated and clean data and the criteria AP to assess the performance of the methods. If AP>0, outlier removal improves the classification accuracy. If AP<0, outlier removal does not improve the classification accuracy.

4.1.4 Simulation Results
To compare the performance of the methods, we generated side-by-side boxplot (Figure 3) for the AP measure for each scenario of the simulation.

We also conducted t-test to compare the AP of each method against 0. The BF achieved significant improvement with Extreme outliers 5% contamination, Guassian outliers 5% contamination, Arbitrary outliers at 5% and 10% contamination, while Robpca achieved significant improvement with both 5% and 10% Extreme outliers, and both 5% and 10% Arbitrary outliers. PCA achieved no significant improvement under none of scenarios.

Both BF and ROBPCA achieved significant improvement under 4 out of 6 scenarios. The results indicate that Benford’s based outlier detection approach is promising, and could be applied to genotype dataset.

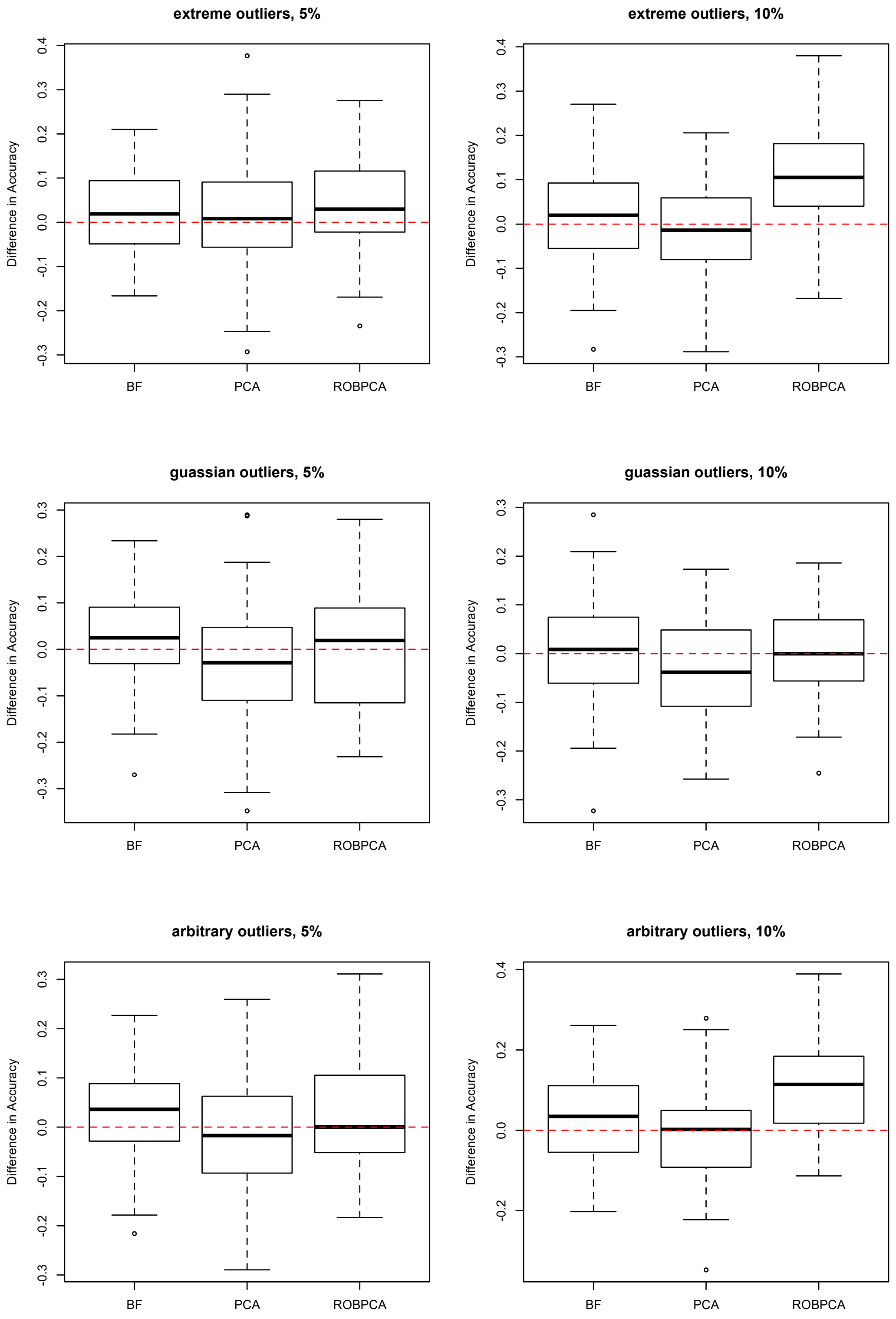


Figure 3. Comparison of relative classification accuracies of three outlier detection methods

4.2. Real Data Application

We further applied Benford's law based genotype outlier detection approach to two real datasets. The Hapmap3 dataset (International HapMap 3 Consortium, 2010) include 957 human individuals across 11 populations. We collapse the groups into four groups (i.e., Africa, Europe, Asia1, Asia2) here (Figure 4). The dataset assayed for 14,389 SNPs.

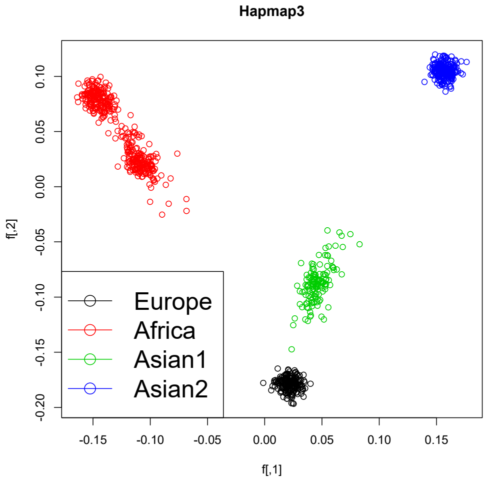


Figure 4. PC scores plot for HapMap 3 dataset

The Genetic Analysis Workshop 17 (GAW17) dataset (Almasy et al., 2011) include 697 independent individuals across 9 populations. We collapse the populations into three groups (i.e., Africa, Asian, Europe) (Figure 5). The dataset assayed for 24,487 SNPs.

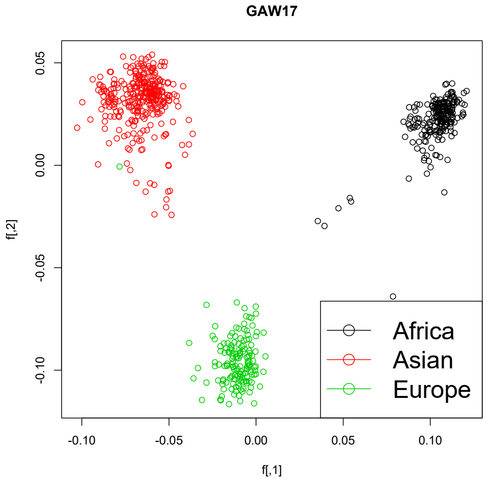


Figure 5. PC scores plot for GAW17 dataset

For each dataset, we first detect outliers with PCA, BF, and ROBPCA methods. We then conducted a bootstrapped cross-validation for B=50 times. The bootstrapping procedure is similar to the cross-validation procedure described in simulation section.

- Apply PCA, BF, and ROBPCA to detect outliers
- Apply 10-fold cross validation via SVM on the data before and after removing observations detected as above outliers.

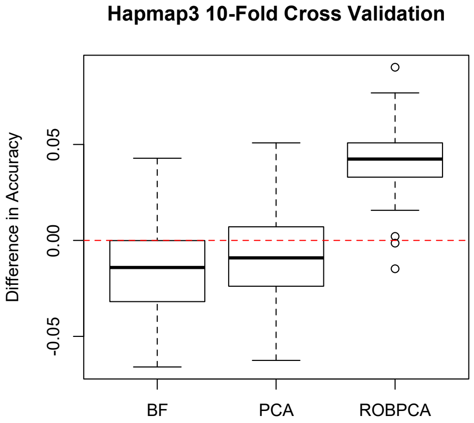


Figure 6. Comparison of methods for Hapmap3 data; ROBPCA outperforms the other methods

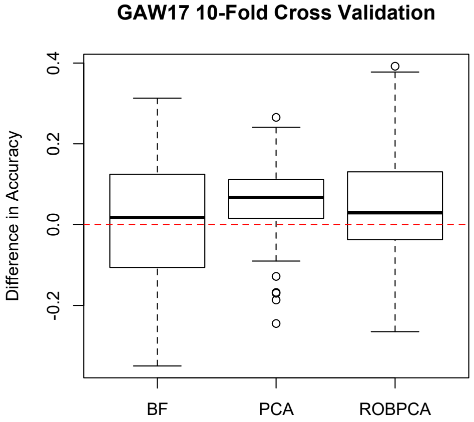


Figure 7. Comparison of methods for GAW17 data; BF, PCA and ROBPCA seem to separate the existing groups in the data

In summary, ROBPCA based outlier detection achieved significant improvement in both Hapmap3 and GAW17 dataset, while PCA and BF based outlier detection achieved improvement in GAW17 dataset.

5. Discussion and Limitation

The application of Benford's law has been popular in social and economic fields. In recent years, Benford's law has also been applied to Bio-system fields. As the first study to apply Benford's law in GWAS, the current study attempted a primitive trial to transfer the BF based methods into a genotype dataset. Our results reveal that removing Benford's outlier could improve population identification. In simulation study, the Benford's approach performed close to the Robust PCA. In real data application, our approach also improved population identification in GAW17 dataset. The results indicate that the BF based outlier detection is promising and acquires further investigation.

Benford’s law based outlier detection approach is distributional free and easy to compute. Unlike Mahalanobis distance or Robust PCA, it does not require the numerical dataset to be normally distributed. The computation load is also reduced to a goodness-of-fit test. The application of Benford’s law approach could potentially bring benefits to GWAS and analysis of high-dimensional dataset in general.

One important issue related to the Benford’s law and its application is its theoretical foundation. Despite of the widespread application of Benford’s law in fraud detection, it needs explanation why disobeying of the Benford’s law would indicate abnormality, and what types of dataset would confirm or not confirm to the Benford’s law. Those are important research questions relating to the application of the Benford’s law.

The current study is limited in several aspects. Firstly, we only included two-group setup when simulating the SNP dataset (i.e., the case 1 of Price et al. simulation setup). In the future, we can include more simulation settings such as the admixture population structure, or population structure of more than two groups. Secondly, we can also include more available goodness-of-fit tests for testing the Benford’s law, such as the Kolmogorov–Smirnov, Mantissa-Arc Test (Alexander, 2009). Different test may have different sensitivity/specificity, which could also impact the results of outlier detection.

The application of Benford’s law in GWAS is a novel and promising research topic. The current study found that removing Benford’s outliers would improve classification accuracy in identifying sub-populations. In the future, Benford’s law based robust-PCA procedure could be developed which does not require the removal of outliers. Research could also be conducted regarding whether Benford’s law based outlier-detection method could improve the power of association analysis.

References

Alexander, J. (2009). Remarks on the use of Benford's Law. Working paper, Case Western Reserve University, Department of Mathematics and Cognitive Science.

Almasy, L., et al. (2011). Genetic Analysis Workshop 17 mini-exome simulation. BM Proceedings. 5:9.

Amos, C.I. (2007). Successful design and conduct of genome-wide association studies, Human Molecular Genetics, 16:2, 2007, 220–225.

Benford, F.(1938). The law of anomalous numbers. Proceedings of the American philosophical society, 551–572.

Costas, E., López-Rodas, V., Toro, J.F., Flores-Moya, A. (2008). The number of cells in colonies of the cyanobacterium Microcystis aeruginosa satisfies Benford's law Aquatic Botany, 89(3),341–343.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071

Devlin, B., and Roeder, K. (1999), Genomic control for association studies, Biometrics, 55, 997-1004.

Dieter William Joenssen (2015). BenfordTests: Statistical Tests for Evaluating Conformity to Benford's Law. R package version 1.2.0. https://CRAN.R-project.org/package=BenfordTests

Hill, T. P. (1995). A statistical derivation of the significant-digit law. Statist. Sci., 10, 354-363.

Hubert, Mia, Peter J. Rousseeuw, and Karlien Vanden Branden. "ROBPCA: a new approach to robust principal component analysis." Technometrics 47.1 (2005): 64-79.

International HapMap 3 Consortium (2010), Integrating common and rare genetic variation in diverse human populations, Nature, 467, 52–58.

Joenssen, D.W. (2015). BenfordTests: statistical tests for evaluating conformity to Benford's Law. R package version 1.2.0. https://CRAN.R-project.org/package=BenfordTests

Karthik, D., Stelzer, G.,Gershanov,S., Baranes,D., and Salmon-Divon, M. (2016). Elucidating tissue specific genes using the Benford distribution.
BMC Genomics, 17, 595:609

Li, Q., Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures.
Genetic Epidemiology, 32(3), 215-226.

Liu, L., Zhang, D., Liu, H., and Arendt, C. (2013). Robust methods for population stratification in genome wide association studies. BMC Bioinformatics, 14: 132.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. Am J Math, 4: 39-40

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies.Nat Genet, 38, 904–909.

Pritchard, J.K., Stephens, M., Donnelly, P. (2000). Inference of population structure using multi-locus genotype data. Genetics. 155, 945–959.

Rosenberg, N.A., et al. (2002). Genetic structure of human populations. Science. 298, 2381–2385.

Tom Reynkens (2018). rospca: Robust Sparse PCA using the ROSPCA Algorithm. R package version 1.0.4. https://CRAN.R-project.org/package=rospca

Xu, H., Caramanis, C., and Mannor, S. (2013). Outlier-robust PCA: the high-dimensional case. IEEE transactions on information theory, 59(1), 546-572.