Predicting Country of Origin from Genetic Data

G. David Poznik

Introduction

Genetic variation in Europe is spatially structured; similarity decays with geographic distance. The most striking visual manifestation of this is the fact that when the first two principal components of the genotype matrix are plotted against one another with samples labeled by country of origin, the map of Europe is recapitulated [1]. As a result, Principal Components Analysis (PCA) has been used to predict the country of origin of genetic samples. Yang, *et al.* [2] achieve similar results with a model-based approach in which they assume a logistic slope function for each genetic variant. But these unsupervised methods do not learn decision boundaries. How well can we predict country of origin from genetic data using supervised machine learning approaches? In this work, I (1) investigate four methods to build a classifier that labels a sample with its country of origin based solely on genetic data; (2) frame the question as a continuous response regression problem to predict the geographic centroid of a collection of unlabeled samples. I achieve impressive accuracy, sensitivity, specificity, and positive predictive value in the classification problem and extraordinary root mean square error (RMSE) in the regression.

Data Preparation

For this work, I used the Population Reference Sample (POPRES) data set [3], for which 3,192 European individuals were genotyped at 500,568 single nucleotide polymorphisms (SNPs). Because the data were generated on a first-generation whole-genome SNP array, considerable care was taken to pre-process and clean the data in order to ensure against spurious results. I used `PLINK` software [4] to apply basic filters to both SNPs and samples.

SNPs

I excluded SNPs that met any of the following filter conditions: (1) minor allele frequency (MAF) < 1%, (2) missingness across samples > 10%, (3) Hardy-Weinberg equilibrium (HWE) p-value < 10^{-5}, (4) located on a sex chromosome. The first of these filters is due to the fact that the allele calling algorithm is unreliable when few alleles of a given type are present in the data. The second indicates poor sample quality, and the third is an indicator of genotyping error. For an A/B SNP where the frequency of the A allele is p, the HWE test flags variants for which the genotype frequencies of AA, AB, BB, depart significantly from those expected under equilibrium (p^2, $2pq$, q^2). Finally, the fourth filter is in place to enable joint analyses of males and females. These four filters left 363,810 high quality SNPs.

At a sampling density such as that for which these data were generated, the SNPs do not represent independent measurements. Rather, local correlation, known as linkage disequilibrium, exist along a chromosome. To account for this, I thinned the data such that the remaining SNPs were in approximate linkage equilibrium. I used `PLINK` to compute correlations amongst SNPs. In windows of 50 SNPs, sliding by 5 SNPs at a time, I constructed a list of SNPs to exclude due to an $r^2 > 0.8$ with an included SNP. The thinned set of SNPs, upon which all analyses are based, numbered 208,899.

Samples

First, I excluded 22 samples with non-European or unknown ancestry and one with missing data at more than 10% of sites. The next step was to identify cryptic relatedness amongst samples, as unmodeled excess genetic sharing would violate sample independence assumptions of downstream analyses. I subsampled approximately 20,000 high frequencies (MAF > 0.2) SNPs and again used `PLINK` to compute identity-by-descent (IBD) statistics for all pairs of individuals. By computing average sharing with all others for each individual, I identified two instances of contamination. These outliers were removed along with 67 individuals found to exhibit excess sharing with a retained sample (*piHat* > 0.075).

I then conducted a PCA to identify outlying samples. To do so, I used `EIGENSTRAT` software [5] in outlier identification mode. Specifically, in each of five iterations, I computed the top ten eigenvectors and flagged for removal a total of 26 samples seen to possess a PC score more than six standard deviations beyond the mean for any of the ten components. I then removed 232 samples from ancestrally heterogeneous countries (Australia, Canada, USA) and 492 whose reported grandparental origins did not meet a strict consensus. Samples for whom grandparental data were unavailable, were retained. Five individuals with inferred Sardinian ancestry were excluded along with 130 individuals from countries with fewer than 15 representatives. Sample counts by county at this stage are illustrated in **Figure 1**. Finally, 162 Swiss individuals lacking linguistic data were excluded. Thus, analysis commenced on a set of 2053 individuals. Of these, 1453 were used exclusively for PC construction, and 600 were carried forward for analysis. The 1453 included 203 samples from countries with 15–70 representatives, as preliminary analyses indicated that class imbalance issues were preventing accurate prediction for these under-represented countries. 131 Portuguese samples were excluded due to my inability to separate them from the Spanish. I also removed 738 Swiss French and 84 Swiss German due to excessively strong commonality with France and Germany, and due to the fact that linguistic data on grandparents was unavailable, thus casting doubt on the labelings. I downsampled representatives of the United Kingdom and Spain from 382 to 151 and from 217 to 151, respectively. This left 600 samples for supervised machine learning analyses. The 600 were split into a randomly sampled training set of 405 (70%) and a hold-out test set of 180. This initially designated training set was subsequently pruned by 15 samples (see next section). Thus, training was conducted on a set of 385 samples (89 French, 72 Germans, 148 Italians, 132 Spanish, and 144 from the United Kingdom), and testing was performed on 25 French, 20 Germans, 45 Italians, 38 Spanish, and 52 British.

Imputation

Off-the-shelf supervised machine learning methods generally assume complete data. Therefore, the final stage of pre-processing was to impute missing values. I used `BEAGLE` software [6] to do so. `BEAGLE` infers haplotype phasing from genotype data. That is, it determines which alleles from neighboring SNPs in a given sample reside on a single chromosome. From these inferences, a population pool of haplotypes is constructed, and missing genotypes can be inferred by comparing an individual's haplotypes to the reference pool.

Feature Extraction

With this cleaned and pre-processed data set, I could progress to the machine learning. After pruning for correlation between SNPs, I was left with over 200,000 features, so I decided that dimensionality reduction via PCA was warranted. When one elects this path, one is essentially placing a bet: that the response variable is most strongly correlated with the directions of greatest variation within the predictors. Inspired by the work of Novembre, *et al.* and others, I felt this bet worth taking, so I computed the top 100 principal components in the set of 2053 samples and used these as the features for all learning. A biplot of PC1 vs. PC2 recapitulates the map of Europe (**Figure 2**), echoing the findings of Novembre *et al.* and building confidence in my cleaned data set. In this plot, samples are colored by country. A small dot represents a sample used in PC construction but not in downstream analysis. A larger filled circle represents a training set sample, and an open circle represents a hold-out test set example.

Initially, I thought it more appropriate to exclude the test set samples from the construction of the PCs and to then project these samples onto the basis determined exclusively from the 1873 reference and training samples. However, manual inspection of PC biplot thusly derived clearly indicated that such an approach would not yield optimal separation. Specifically, the test samples from this run were all pulled toward the center of the biplot with respect to the clouds defined by their compatriots. That is, though the samples themselves were uniformly sampled from within countries, they were clearly *not* uniformly sampled within PC space. Consequently, the training set would not have been truly representative of the test set[1]. This makes sense, since the PC loadings are most strongly defined by those SNPs constituting the major variation within the samples used to construct them. Thus, excluding the test set from PC construction amounts to ignoring much of the information contained within the test samples that would be most helpful toward classification! I reasoned that in building a classifier or regression model, one is essentially constructing a recipe, and there is no harm in including in that recipe a prescription to run PCA on a new unlabelled set of samples along with a reference panel such as POPRES. In so doing, one can maximally leverage the *feature* information completely blind to the true and unknown labels.

I found that I could further improve performance by identifying country-specific outliers along the top two principal components in the training set. That is, for each of the five countries, I flagged for removal any sample whose PC1 or PC2 score represented a *Z*-score greater than 2.5 ($p = 0.01$) for that country. These most likely constituted labeling errors. For example, a British national of French ancestry for whom grandparental data was unavailable would have been thusly flagged. 15 samples (two, zero, three, three, and seven from the respective countries) were removed. These are marked by a "×" in **Figure 2**. This procedure requires looking at the labels, so the test set was left as-is, despite the possible presence of labeling errors therein.

Classification

I applied four supervised learning methods toward the classification problem of labeling individuals with a national identity. For each method, I investigated the effects of including the top *c* principal components over a range of *c* values, and for each, I optimized over the respective model parameters. I wrote a module that, given the test set labelings and a prediction vector, constructs a confusion matrix, and computes accuracy, and country-specific positive predictive value (PPV; precision), sensitivity (recall), and specificity, where I define the latter two in terms of a two-by-two contingency table (e.g. labeled French and labeled non-French versus predicted French and predicted non-French). Results were extremely impressive. Due to space limitations, I include a table for a single model only. Finally, I compare the accuracies obtained from the best model obtained from each approach (**Figure 4**). All analyses were conducted in `R`.

[1] Interestingly, my first attempt to resolve this issue was to remove all variants with MAFs in the 1% – 5% range. Unexpectedly, this approach appeared to have exacerbated rather than ameliorated the effect.

Naïve Bayes

The simplest model I applied was Naïve Bayes. To do so, I used the `naiveBayes` function from the `e1071` package. As the model is parameter-free, no parameter optimization was required. I investigated inclusion of the top 2–25 PCs, and the algorithm performed best, as measured by the accuracy of test set prediction, when the top nine PCs were included in the model (**Figure 4**). Differences, however, were slight, and this model was the most insensitive to the number of PCs included. I imagine this is due to the conditional independence assumption. The predictive value of each feature is assessed independently. Thus, if ancestry were independent of a given feature, inclusion of this feature as a predictor would have no effect on the model.

k-Nearest Neighbors

I applied the *k*-Nearest Neighbors algorithm via the `knn` function of the `class` package and jointly optimized over odd values of *k* from 1–15 and the inclusion of the top 2–25 PCs. A surface plot of this optimization is presented in **Figure 3a**. It reveals that accuracy was optimized at *k* = 13 when the top three PCs are included in the model. Though the biplot of PC1 and PC2 indicates strong correlation with geography, clearly PC3 contains additional discriminative information. In fact, this model achieved the highest level of accuracy of any. As such, a confusion matrix, and summary statistics are presented below (**Table 1**). A steep drop-off in performance was observed with the inclusion of additional PCs. In contrast to Naïve Bayes, *k*-NN assesses all features jointly. If, for example, PC4 was due to a laboratory effect, its inclusion in the model could preclude the influence of ancestral information in subsequent PCs.

	Label				
Prediction	**France**	**Germany**	**Italy**	**Spain**	**UK**
France	21	1	0	1	0
Germany	2	17	0	0	2
Italy	0	0	45	0	0
Spain	0	0	0	37	0
UK	2	2	0	0	50

Accuracy: 0.944 (170/180)

	France	**Germany**	**Italy**	**Spain**	**UK**
PPV (Precision)	0.913	0.810	1	1.000	0.926
Sensitivity (Recall)	0.840	0.850	1	0.974	0.962
Specificity	0.987	0.975	1	1.000	0.969

Table 1. Confusion Matrix and prediction summary statistics for *k*-NN with k=13 using the top 13 principal components as features.

Multinomial Logistic (Softmax) Regression

For my third analysis, I applied multinomial logistic (softmax) regression, a generalization of logistic regression to the multi-class classification problem. To do so I utilized the `multinom` function from the `nnet` package. Again, this model required no parameter optimization. Performance was similar to that of Naïve Bayes for small feature sets but declined with the inclusion of nine or more PCs.

Support Vector Machine

Finally, I turned to support vector machines (SVMs) with the `svm` and `tune` functions from the `e1071` package. Though SVMs are formulated for binary classification problem, `svm` achieve multi-class prediction through serial binary (one versus the rest) comparisons. For the analyses presented here, I conducted *C*-classification on scaled PCs utilizing a radial basis kernel. I investigated other kernel choices, but results were similar. I jointly optimized over a wide range of two parameters: *C*, the cost (*i.e.*, the Lagrange multiplier for the slack variables measuring the degree of misclassification), and γ, the scaling constant in the radial basis function. To do so, I used 5-fold cross validation within the training set. An example surface plot of optimization, where each axis is defined by the base two logarithm of a parameter, is presented in **Figure 3b**. For a given set of top PCs, I predicted test set countries of origin based on the model defined by the optimal (C, γ). I examined sets of the top 2, 5, 10, 15, and 20 PCs, and accuracy was maximized with the inclusion of ten (**Figure 4**).

Regression

For the second part of this project, I sought to predict not just a class label, but a continuous measure of geography. For this problem, I independently modeled latitude and longitude with linear models in an ordinary least squares regression. To do so, I utilized the `lm` function from the `stats` package. This assumption of independence between the two responses would be expected to hold under a model of continuous genetic exchange across the continent, and previous work has demonstrated this to be a quite reasonable assumption for Europe.

Since the precise geographic origins of samples were unknown, I used the geographic center of each country as the response for the training set samples, as in Novembre, *et al.* Results from regressing upon the top two PCs were unsurprisingly good. But what blew me away was that accuracy, as measured by root mean square error, improved more or less continuously with the stepwise addition of the top 87 PCs (**Figure 5a**). In **Figure 5b**, I plot the results achieved by this optimal RMSE model. The large filled circles represent a given country's geographic center—that which was used for the pair of responses for each training sample from the country. Small open circles represent the predicted geographic coordinates of the test samples, and large open circles represent the centroid of these predictions. RMSE was measured between the prediction centroid and the geographic center of the country. The results are simply extraordinary. There is almost no error for Spain, France, and Italy. I posit that the errors that exist for the United Kingdom and Germany are due to the imprecision in the training values rather than a lack of information within the genetic data. London is in the far south of the United Kingdom, so it stands to reason that the majority of samples were likely drawn from south of the geographic center, and I would venture to guess that the majority of Germans were sampled from the former West Germany.

Conclusion and Future Work

Upon careful framing of the problem to include distinct and well-represented countries and meticulous pre-processing of the data, I have constructed a means of classifying country of origin from genetic data with extraordinary accuracy, sensitivity, specificity, and positive predictive value. In addition, I have built a regression model that does a remarkable job predicting the geographic origin in continuous space of genetic samples.

The following analyses are the most immanently compelling directions in which I would like to have taken this work given more time:

1. *More sophisticated model selection*. In all analyses I used a simple stepwise addition of PCs. I imagine, for example that applying elastic net regression on the PCs could yield even better performance.
2. *Predicting the location of additional countries*. There is no reason not to include in the test set samples from the nine additional moderately sampled countries in an attempt to predict the geographic center of these countries from the genetic data of their citizens.
3. I would like to further investigate the intriguing finding that removing rare variation seemed exacerbate the non-representativeness in PC space of projected samples. The first thing to do would be to check the frequencies (overall and stratified by country) of the most highly loading SNPs for each PC.
4. *Crammer multi-class formulation of SVM*. Whereas my SVM application utilized serial binary classification, I learned at the poster session of a method that bakes multi-class labeling directly into the objective function. I played around with `ksvm` from the `kernlab` package, but did not have time to include the results in this report.
5. *More discriminative features*. The PCs are just a summary of the true variation within the samples. I would like to see the effect of using the correlation-thinned SNPs as features on their own. I thought about using tests of association for model selection. This approach would miss interaction effects, but may be worth a look. I would like to apply decision tree and random forest approaches to this problem. These methods should do a great job at identifying recent variants that have risen to appreciable frequencies in specific geographic regions. SNPs such as these should have great discriminative power. I would also like to apply penalized logistic regression and penalized linear regression to these features.
6. *SVM with a variable cost objective*. Country is not merely a categorical class label. Rather, the set of countries possesses a two-dimensional ordinality, and it might be worth investigating whether leveraging this information could improve performance. This could be accomplished by modifying the SVM objective function to take as input the geographic centroid of the training set country labels in order that it reflect not just whether a predicted label is incorrect, but how far off it is. For example, it would be better to confuse a British sample for German than for Italian. This will require altering the inequality condition of the SVM optimization [7].

Acknowledgements

The initial set of questions that led to this work was conceived in a discussion with Suyash Shringarpure and Simon Gravel. Brian Maples ran `BEAGLE` to impute sporadic missing data.

References

1. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98–101. Available:http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2735096&tool=pmcentrez&rendertype=abstract. Accessed 26 October 2012.
2. Yang W-Y, Novembre J, Eskin E, Halperin E (2012) A model-based approach for analysis of spatial structure in genetic data. Nature genetics 44: 725–731. Available:http://www.ncbi.nlm.nih.gov/pubmed/22610118. Accessed 6 October 2012.
3. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, et al. (2008) The Population Reference Sample , POPRES : A Resource for Population , Disease , and Pharmacological Genetics Research: 347–358. doi:10.1016/j.ajhg.2008.08.005.
4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M a R, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559–575. Available:http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950838&tool=pmcentrez&rendertype=abstract. Accessed 25 October 2012.
5. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS genetics 2: e190. Available:http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1713260&tool=pmcentrez&rendertype=abstract. Accessed 12 June 2011.
6. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American journal of human genetics 81: 1084–1097. Available:http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2265661&tool=pmcentrez&rendertype=abstract. Accessed 4 March 2012.
7. Tsochantaridis I, Hofmann T (2005) Large Margin Methods for Structured and Interdependent Output Variables. 6: 1453–1484.

Fig. 1

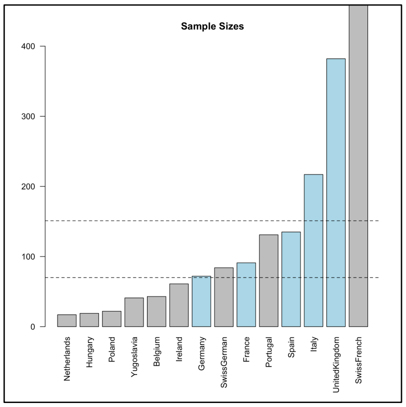


Fig. 2a

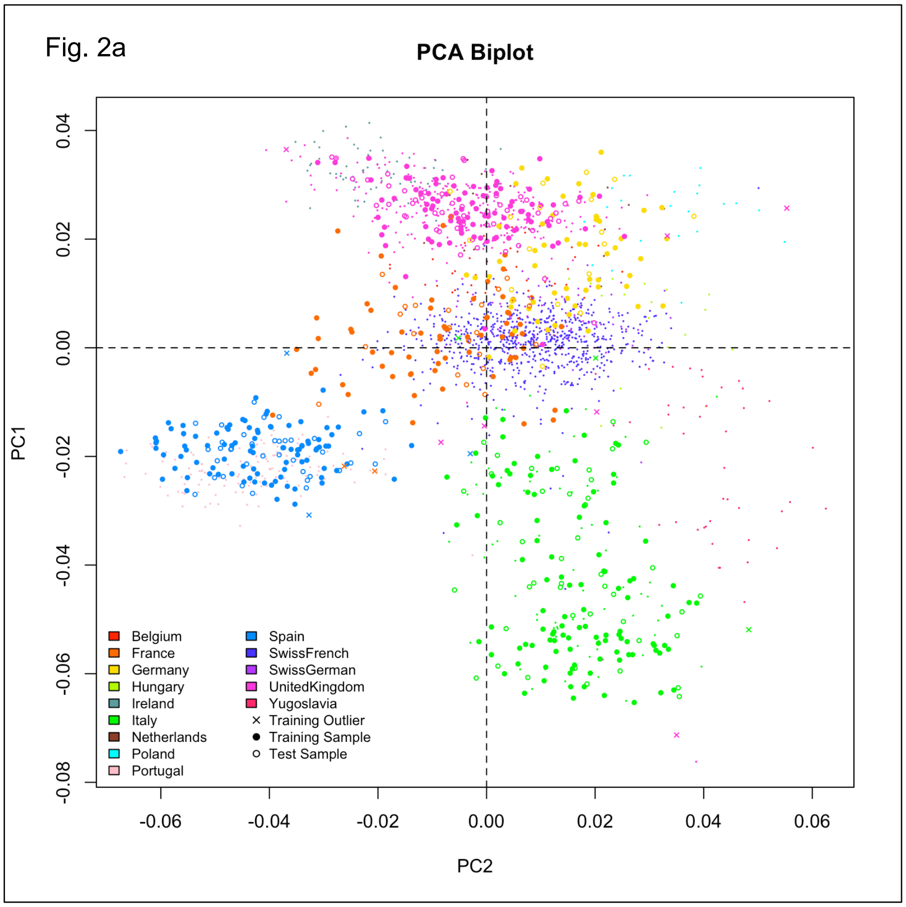


Fig. 2b

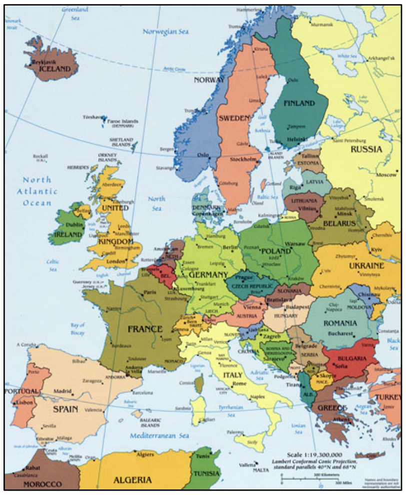


Fig. 3a

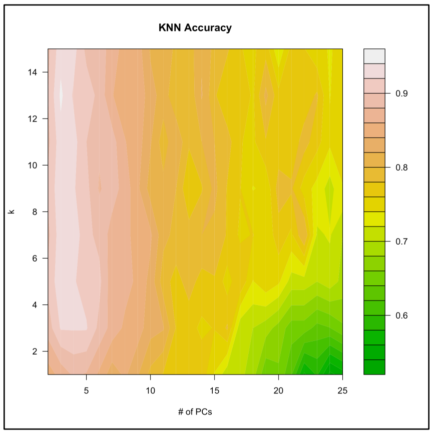


Fig. 3b

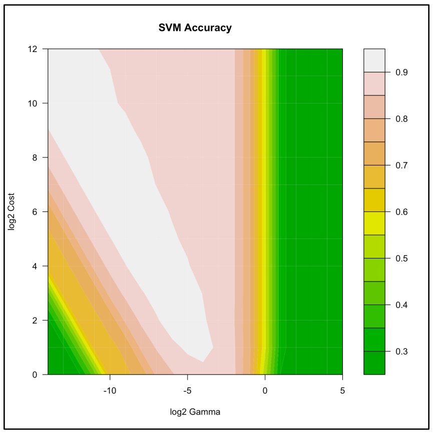


Fig. 4

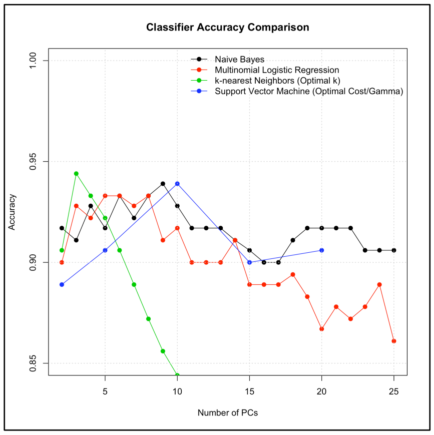


Fig. 5a

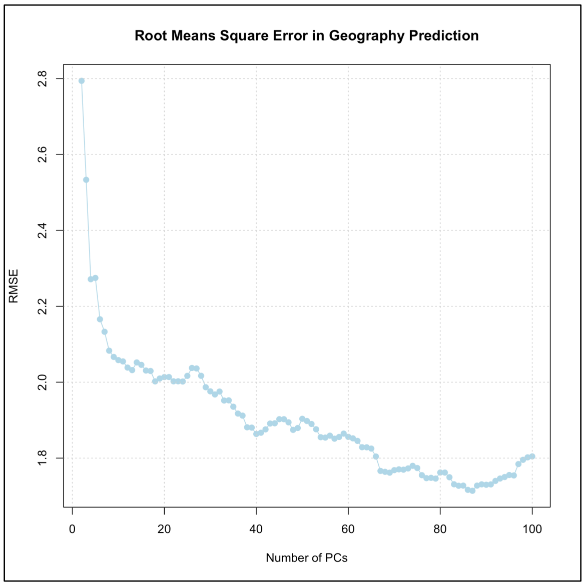
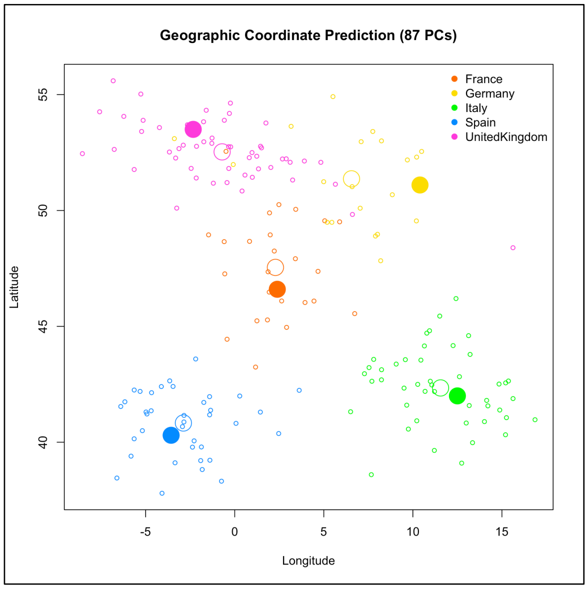


Fig. 5b


Geographic Coordinate Prediction (87 PCs)
France
Germany
Italy
Spain
UnitedKingdom
Latitude
Longitude