SoCal: supervised genotype calling via ellipsoidal separation for Affymetrix SNP microarray

Huwenbo Shi (603-778-363) shihuwenbo@ucla.edu

ABSTRACT

Background: Genotype calling for SNP microarray is an important step in array-based association studies. Most supervised genotype calling algorithms for SNP microarray fit generative models (e.g. Gaussian models) on log-transformed allele intensities of samples having reference genotype calls.
Methods: In this article, I present SoCal, a supervised genotype calling method that efficiently finds ellipsoidal decision regions for each genotype via ellipsoidal separation instead of fitting a generative model. By assigning different weights to the criteria of finding separating ellipsoids, SoCal can control the effect of outliers in training data.
Results: In leave-one-out cross-validation with HapMap reference calls, SoCal achieved a concordance rate of 98.94% at a call rate if 100% and 99.71% at a call rate of 95%, comparable to many state-of-the-art methods. By training SoCal using HapMap reference calls, SoCal achieved 95.10% concordance rate with genotype calls made by CRLMM at a call rate of 100% on a set of validation calls excluding the training calls. SoCal also shows more robustness than RLMM, a method that uses Gaussian decision regions to call genotypes, when there are outliers in training data. Overall, SoCal is a novel and promising genotype calling algorithm.

1 Introduction

Accurate genotyping of SNPs is essential to discovering true signals in association studies. Although next generation sequencing technology provides cheap whole-genome sequences for genotyping SNPs, SNP microarray is still a cost-effective genotyping technology for many specific association studies. In an Affymetrix SNP microarray, oligonucleotide probes are used to match and bind DNA fragments containing biallelic SNPs. Then a fluorescence scanner scans the microarray to quantify perfect match and mismatch of these fragments. Most genotype calling procedures for SNP microarray consists of two steps. In the first step, raw information from microarray is summarized to obtain the intensities, θ_A and θ_B, of the two alleles, denoted by A and B, of each SNP. In the second step, SNPs are classified into genotype AA, AB, or BB based on the allele intensities they generate. The focus of this article is on the second step of the genotype calling procedure—genotype calling using summarized allele intensities.

For a specific SNP, if a sample has genotype AA or BB, the allele intensity, θ_A or θ_B, will be higher respectively. If a sample has genotype AB, the intensities, θ_A and θ_B, will be similar. If one plots $log(\theta_A)$ versus $log(\theta_B)$ of a SNP for a number of samples, normally

3 ellipsoidal clusters can be observed, one for each genotype, as shown in Figure 1. Many genotype calling algorithms use model-based unsupervised clustering to identify these clusters and then assign genotypes to each cluster. Although these methods are applicable to a wide range of microarrays, they don't take advantage of genotype calls that are already available. Also, these methods use the EM algorithm to estimate model parameters, which is sensitive to starting parameters and slow to converge. To utilize reference genotype calls, Rabbee et al. proposed the RLMM algorithm, a supervised genotype calling method that uses reference genotype calls to form decision regions by fitting bivariate Gaussian distributions on observed allele intensities for each genotype. These Gaussian decision regions are then used to call SNPs with unknown genotype. However, fitting a Gaussian distribution is known to be sensitive to outliers, which, for SNP microarrays, can be caused by genomic structural variations.

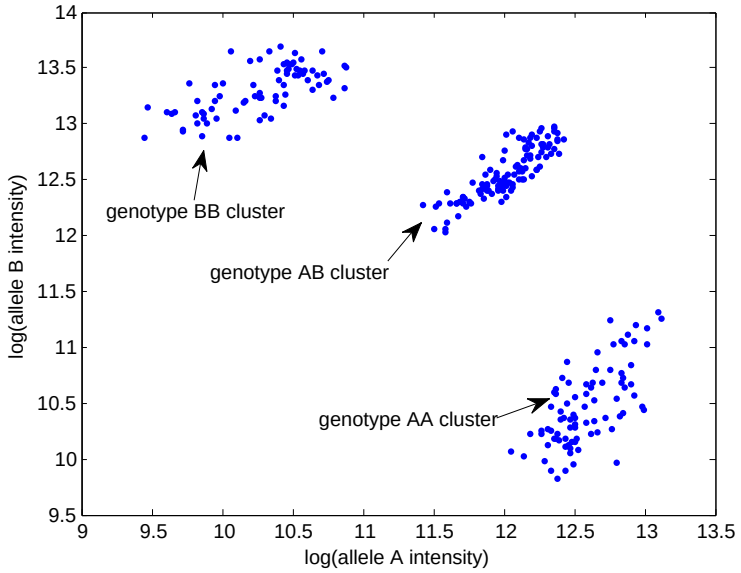


Figure 1: Genotype clusters obtained from Affymetrix SNP array allele intensity values

In this article, I present SoCal, a supervised genotype calling algorithm for Affymetrix SNP microarrays. Instead of fitting Gaussian distributions on allele intensities with reference genotype calls, SoCal efficiently finds ellipsoidal decision regions for each genotype of a SNP by solving a conic programming problem. SoCal can control the effect of outliers by assigning different weights to each of the criteria for finding the ellipsoids—separation ratio, ellipsoid volume, and number of enclosed points. After SoCal finds the ellipsoidal decision regions for each genotype of a SNP, it uses them to call the same SNP with unknown genotype.

Using reference genotype calls from the HapMap Project as training and validation data, SoCal achieved 99.71% accuracy at a call rate of 95% during leave-one-out cross-validation. Furthermore, SoCal shows more robustness than the RLMM method when there are outliers in the training data.

2 Methods

2.1 Overview of SoCal's genotype calling procedure

SNP allele intensities are first summarized from raw microarray data using SNPRMA, which removes non-biological effect from the data. After this step, SoCal calls genotypes in two steps. In the first step, SoCal finds ellipsoidal decision regions for each of the genotype of a SNP using reference genotype calls. In the second step, SoCal classifies SNPs with unknown genotypes using minimum distance classification.

2.2 Pattern separation by ellipsoid

An ellipsoid $\mathcal{E} \subseteq \mathbb{R}^n$ can be expressed as $\mathcal{E} = \{x \in \mathbb{R}^n | (x-c)^T E(x-c) \leqslant 1\}$, where c is the center of the ellipsoid, and E a positive definite matrix denoting the shape and orientation of the ellipsoid. Let $\{a_i\}$ be the points to be included in an ellipsoid, and $\{b_j\}$ be the points to be excluded, the problem of ellipsoidal separation is to find c and E such that $(a_i - c)^T E(a_i - c) \leqslant 1 \, \forall i$ and $(b_j - c)^T E(b_j - c) > 1 \, \forall j$.

2.3 Forming ellipsoidal decision regions for each genotype

Let $G = \{AA, AB, BB\}$ be the set of genotypes of a SNP, and J_{AA}, J_{AB}, J_{BB} the index set of samples with the corresponding genotype. Let $X = \{(log(\theta_A), log(\theta_B))_i | i = 1, \cdots, |J_{AA}| + |J_{AB}| + |J_{BB}|\}$ be the set of log transformed allele intensities of all the samples, and $X_{AA} = \{x_j | x_j \in X, j \in J_{AA}\}$, $X_{AB} = \{x_j | x_j \in X, j \in J_{AB}\}$, $X_{BB} = \{x_j | x_j \in X, j \in J_{BB}\}$ the set of log transformed allele intensities from samples having the corresponding genotype.

To find the ellipsoid that includes X_{AA} and excludes $X_{AB} \cup X_{BB}$, one sets $\{a_i\} = X_{AA}$ and $\{b_j\} = X_{AB} \cup X_{BB}$, and solves the following conic programming problem. The derivation of the problem formulation is largely followed from [7]. For the sake of space, detailed derivation of the problem formulation is not presented here.

$$
\begin{aligned}
\text{minimize} \quad & -\beta_1 k + \beta_2 trace(T) + \beta_3 \|u - \mathbb{1}\|_1 \\
\text{subject to} \quad & (1, a_i)^T \tilde{E} (1, a_i) \leqslant u_i \; \forall i \\
& (1, b_j)^T \tilde{E} (1, b_j) \geqslant k \; \forall j \\
& \tilde{E} = \begin{bmatrix} s & v^T \\ v & F \end{bmatrix} \succeq 0 \\
& \begin{bmatrix} F & I \\ I & T \end{bmatrix} \succeq 0
\end{aligned}
$$

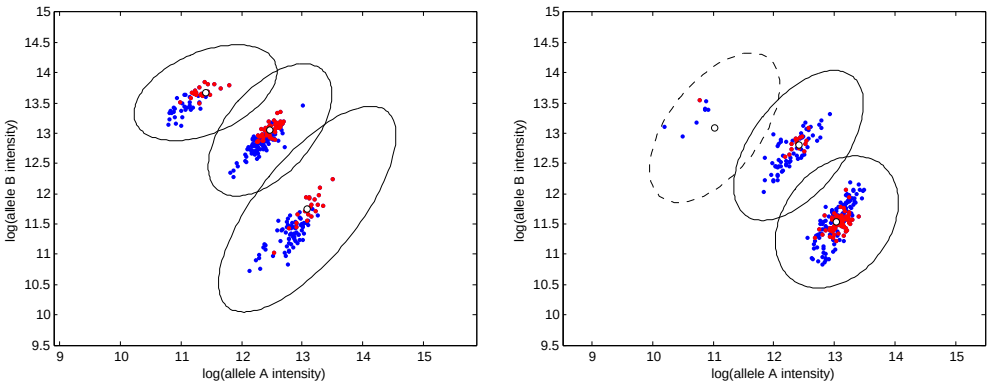
In the problem formulation above $\beta_i > 0$ are the weights assigned to each criteria of finding the ellipsoid—separation ratio, ellipsoid volume, and number of enclosed points. In SoCal, β_1, β_2, β_3 are empirically set to 1, 10^4, and 10^2 respectively.

Let $\tilde{E}^* = \begin{bmatrix} s & v^T \\ v & F \end{bmatrix}$ be the optimal solution to the problem above. The separating ellipsoid $\mathcal{E}^*$ is defined as $\mathcal{E}^* = \{x \in \mathbb{R}^n | (x - c^*)^T E^* (x - c^*) \leqslant 2(1 + k)\}$, where $c^* = -F^{-1} v$, $E^* = \frac{F}{(1 - s + c^{*T} F c^*)}$.

To find the ellipsoid that includes X_{AB} and excludes $X_{AA} \cup X_{BB}$, one sets $\{a_i\} = X_{AB}$ and $\{b_j\} = X_{AA} \cup X_{BB}$, and solves the above conic programming problem. The same procedure also applies to finding the ellipsoid that includes X_{BB} and excludes $X_{AA} \cup X_{AB}$.

2.4 Rescuing missing genotype clusters

If a SNP has moderate minor allele frequency (MAF), the genotype clusters of that SNP are well defined, and SoCal obtains three ellipsoidal decision regions for that SNP, one for each genotype (Figure 2a). However, if a SNP has lower MAF, some genotype cluster may not be well defined. For these SNPs, SoCal estimates the ellipsoid for the missing genotype cluster using the ellipsoids for the other two genotypes through simple geometric transformations (Figure 2b).



(a) SNP with well defined genotype clusters (b) SNP with sparse BB genotype cluster

Figure 2: Each dot in the plots above represents a sample, with samples having HapMap reference genotype calls marked red. The ellipsoids were obtained using only the reference calls.

2.4.1 Missing genotype AA or BB cluster

If genotype AA cluster of a SNP has less than 3 reference calls, SoCal first finds the ellipsoids for genotype AB and BB clusters, and then estimates that for genotype AA cluster through simple geometric transformations.

Let $\mathcal{E}_{AB} = \{x \in \mathbb{R}^n | (x - c_{AB})^T E_{AB} (x - c_{AB}) \leqslant 1\}$ and $\mathcal{E}_{BB} = \{x \in \mathbb{R}^n | (x - c_{BB})^T E_{BB} (x - c_{BB}) \leqslant 1\}$ be the ellipsoids obtained for genotype AB and BB clusters, and n_{AB}, n_{BB} the unit vectors pointing in the direction of the major axis of the corresponding ellipsoid. SoCal estimates the center of $\mathcal{E}_{AA}$, the ellipsoid for genotype AA cluster, by reflecting c_{BB}, the center of $\mathcal{E}_{BB}$, across the major axis of $\mathcal{E}_{AB}$. To estimate the orientation of $\mathcal{E}_{AA}$, SoCal first determines the angle between n_{AB} and n_{BB}, and then applies a rotation matrix of that angle on E_{AB}.

Formally, let $\mathcal{E}_{AA} = \{x \in \mathbb{R}^n | (x - c_{AA})^T E_{AA} (x - c_{AA}) \leqslant 1\}$ be the estimated ellipsoid for genotype AA cluster, and α the angle between n_{AB} and n_{BB}, then $c_{AA} = -c_{BB} + 2c_{AB} + 2n_{AB}((c_{BB} - c_{AB})^T n_{AB})$, and $E_{AA} = R^T E_{AB} R$, where R is a rotation matrix of angle α.

If genotype BB cluster is missing, the center and orientation of the ellipsoid for that cluster is estimated in a similar way. Formally, let $\mathcal{E}_{BB} = \{x \in \mathbb{R}^n | (x - c_{BB})^T E_{BB} (x - c_{BB}) \leqslant 1\}$ be the estimated ellipsoid for genotype BB cluster, and α the angle between n_{AB} and n_{AA}, then $c_{BB} = -c_{AA} + 2c_{AB} + 2n_{AB}((c_{AA} - c_{AB})^T n_{AB})$, and $E_{BB} = R^T E_{AB} R$, where R is a rotation matrix of angle $-\alpha$.

2.4.2 Missing genotype AB cluster

Although SNPs with genotype AB cluster missing were not observed in HapMap reference genotype calls, for completeness, for these SNPs SoCal first obtains, $\mathcal{E}_{AA}$ and $\mathcal{E}_{BB}$, the ellipsoids for genotype AA and BB cluster, and then estimates the center of $\mathcal{E}_{AB}$, the ellipsoid for the missing cluster, using the mid-point between the centers of $\mathcal{E}_{AA}$ and $\mathcal{E}_{BB}$. The orientation of $\mathcal{E}_{AB}$ is obtained by applying a rotation to the ellipsoid with the minimum volume among $\mathcal{E}_{AA}$ and $\mathcal{E}_{BB}$.

Formally, let $\mathcal{E}_{AB} = \{x \in \mathbb{R}^n | (x - c_{AB})^T E_{AB}(x - c_{AB}) \leqslant 1\}$ be the estimated ellipsoid for genotype AB cluster, and α the angle between n_{AA} and n_{BB}, then $c_{AB} = (c_{AA} + c_{BB})/2$, and $E_{AB} = R^T \hat{E} R$, where $\hat{E}$ is the matrix of the ellipsoid with the minimum volumne among $\mathcal{E}_{AA}$ and $\mathcal{E}_{BB}$, and R a rotation matrix of angle $\pm\alpha/2$. The sign of the angle of rotation is dependent on the choise of ellipsoid on which rotation is applied—positive for $\mathcal{E}_{AA}$ and negative for $\mathcal{E}_{BB}$.

2.5 Genotype calling

After the ellipsoidal decision regions, $\mathcal{E}_g = \{x \in \mathbb{R}^n | (x - c_g)^T E_g (x - c_g) \leqslant 1\}, \forall g \in \{AA, AB, BB\}$ of a SNP are obtained, SoCal uses them to classify SNPs with unknown genotypes using minimum distance classification.

If a sample has allele intensity θ_A and θ_B at SNP n, SoCal first computes $D_g = \sqrt{(x - c_g)^T E_g (x - c_g)}$, where $x = (log(\theta_A), log(\theta_B))$, for each $g \in \{AA, AB, BB\}$. SoCal then calls the genotype, $\mathcal{G}$, of that sample at SNP n as the genotype having the minimum D_g, that is, $\mathcal{G} = \arg\min_{g \in \{AA, AB, BB\}} D_g$.

SoCal defines $\lambda = 1 - D_{\mathcal{G}}/(D_{AA} + D_{AB} + D_{BB})$ to quantify the confidence of each genotype call. By increasing the threshold for λ, SoCal can achieve higher call accuracy at the cost of decreasing call rate.

3 Materials

The microarray used for evaluation in this project was the Affymetrix GeneChip Human Mapping 50K Xba Array, which contains 58,960 SNPs. Raw microarray data for 270 samples was obtained from HapMap FTP. And reference genotype calls were obtained from HapMap using HapMart.

After removing strand-ambiguous SNPs and SNPs not present on HapMart from the microarry, 16,387 SNPs were left. Figure 3 shows the minor allele frequency distribution for these SNPs.

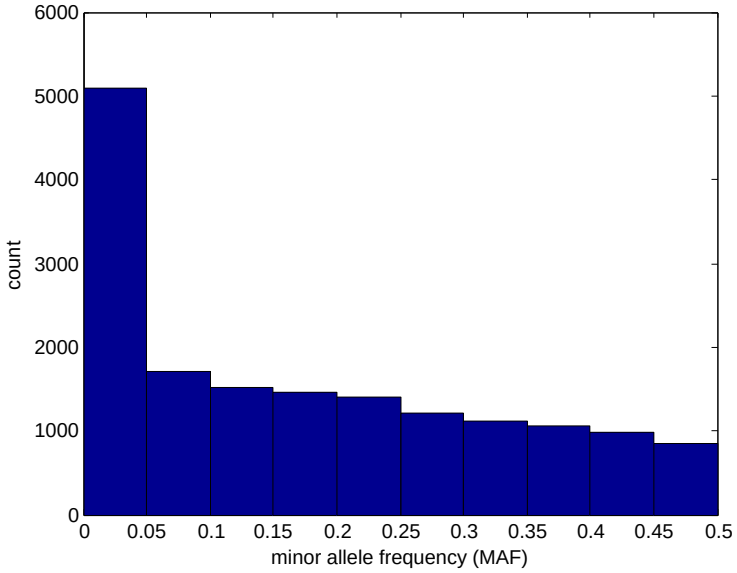


Figure 3: Minor allele frequency distribution for the 16,387 SNPs.

From the 16,387 SNPs, 4,064 SNPs with two genotype clusters having less than 3 reference genotype calls were further removed from the microarry. Among these SNPs, 3,596 are monomorphic SNPs. In total, 12,323 SNPs were left for evaluation. On average, each of these SNPs has 83 reference genotype calls.

4 Results

4.1 Cross-validation with HapMap reference calls

To evaluate the accuracy of SoCal, I compared the genotype calls made by SoCal with the reference calls from HapMap through leave-one-out cross-validation. For each SNP, I used one sample from the reference set as validation data and the rest as training data. I repeated this process until all the samples in the reference set were used as validation data exactly once.

First, I compared the accuracy of SoCal under different choices of β_i, which are the weights assigned to the criteria of finding the ellipsoidal decision regions for each genotype cluster. Figure 4 shows the concordance rate of SoCal in the leave-one-out cross-validation for a wide range of call rates when different values of β_i were used. Because the choice of weights, $\beta_1 = 1$, $\beta_2 = 10^4$, $\beta_3 = 10^2$, had the highest call rate at fixed concordance rates, it's set to be the default choice of SoCal. And all future experiments presented in this article used this choice of weights.

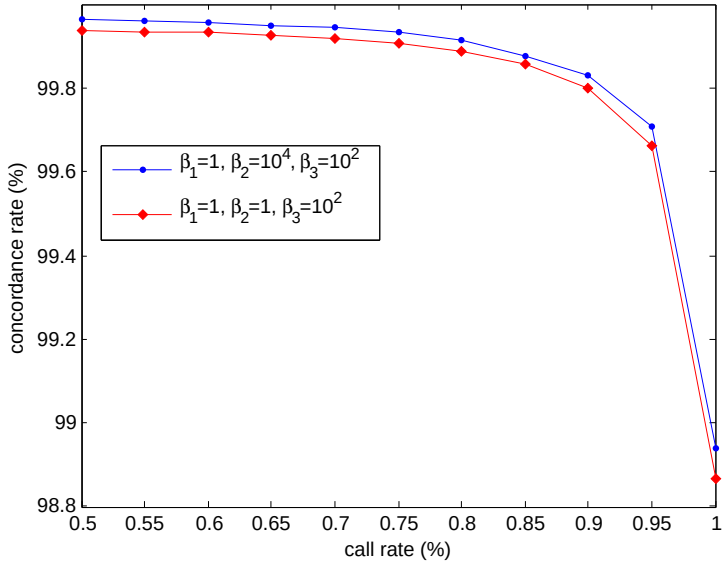


Figure 4: Concordance rate of SoCal in the leave-one-out cross-validation with HapMap reference calls as a function of call rate, for different choices of β_i

Table 1 shows the genotype calls made by SoCal and those available on HapMap during leave-one-out cross-validation. At a call rate of 100%, SoCal made 1,081,319 calls in total, out of which 1,069,857 were concordant with HapMap calls, achieving a concordance rate of 98.94%.

HapMap/SoCal	AA	AB	BB	No Call
AA	360,289	2,282	1,058	0
AB	2,667	341,012	2,257	0
BB	851	2,347	368,556	0

Table 1: At a call rate of 100%, SoCal achieved 98.94% concordance rate in the leave-one-out cross-validation with HapMap reference calls.

Table 2 shows detailed comparison between SoCal and HapMap calls at a call rate of 95%. At a call rate of 95%, SoCal made 1,028,258 calls in total, out of which 1,025,242 were concordant with HapMap calls, achieving a concordance rate of 99.71%. These results are comparable to those of many state-of-the-art methods.

HapMap/SoCal	AA	AB	BB	No Call
AA	348,221	390	298	14,720
AB	710	319,394	775	25,057
BB	410	427	357,627	13,290

Table 2: At a call rate of 95%, SoCal achieved 99.71% concordance rate in the leave-one-out cross-validation with HapMap reference calls.

4.2 Comparison with CRLMM calls

As another way of evaluating the accuracy of SoCal, I compared the genotype calls made by SoCal and those made by CRLMM, a widely used state-of-the-art genotype

calling method for SNP microarrays.

I first trained SoCal using all the samples with HapMap reference calls, and then I made genotype calls on the rest of the samples not in the reference set. I excluded the training samples when comparing SoCal with CRLMM, and compared these two methods only at samples not in the training samples.

Table 3 shows detailed comparison between SoCal and CRLMM at a call rate of 100%. In total, SoCal made 2,245,891 calls, out of which 2,134,868 were concordant with those made by CRLMM, achieving a concordance rate of 95.10%. The high concordance rate between SoCal and CRLMM implies that SoCal has the potential to become an alternative genotype caller.

CRLMM/SoCal	AA	AB	BB	No Call
AA	781,903	23,244	10,405	0
AB	22,340	564,280	22,533	0
BB	7,730	24,771	788,685	0

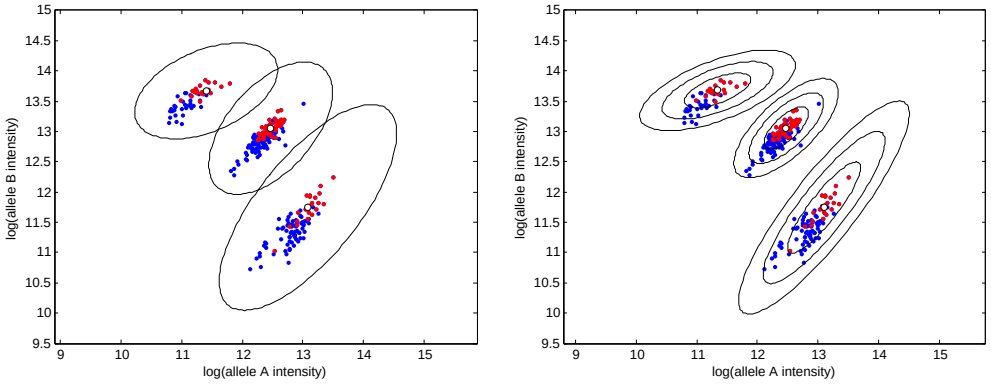
Table 3: At a call rate of 100%, SoCal achieved 95.10% concordance rate with the calls made by CRLMM. Training samples for SoCal were excluded during comparison.

4.3 Comparison with RLMM in the presence of outliers

I investigated how robust SoCal is when there are outliers in the training data. For comparison, I implemented the RLMM algorithm, which independently fits bivariate Gaussian distributions on the log transformed allele intensities of each genotype cluster and then classifies SNPs with unknown genotype into the distribution having minimum Mahalanobis distance.

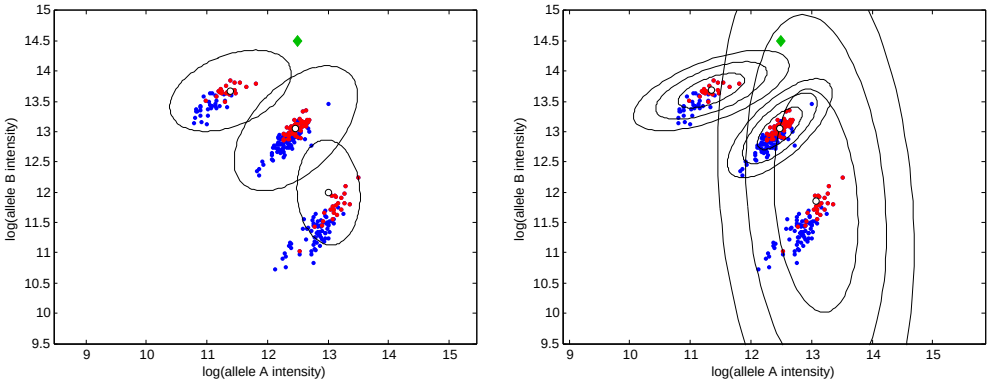
For accurate comparison, I selected a subset of 3,442 SNPs that have more than 10 reference calls for each of the genotype cluster from the set of 12,323 SNPs used for evaluation in the previous section. To simulate outliers, for each SNP, I first estimated μ_g, the mean of the log transformed allele intensities of each genotype cluster, and then drew one outlier for each genotype cluster from the Gaussian distribution $N(\mu_g, \gamma I)$, where I is the identity matrix and γ a positive constant controlling the variance of the distribution—by increasing γ, one increases the effect of outliers. In total, I simulated 3 outliers for each SNP, one for each genotype cluster.

As an example, Figure 5 shows the ellipsoidal decision regions obtained by SoCal and the level curves of the Gaussian decision regions obtained by RLMM before and after an outlier is introduced to the genotype AA cluster. Before an outlier is introduced, both SoCal and RLMM can make accurate genotype calls. However, after an outlier is introduced into the genotype AA cluster, the variance of the Gaussian decision region obtained by RLMM is significantly affected, making the classificatoin much less accurate. On the other hand, because SoCal not only puts penalty on outliers but also jointly uses data from other genotype clusters to form the decision region, it can still classify most SNPs into correct genotypes.



(a) Decision regions found by SoCal when there is no outlier

(b) Decision regions found by RLMM when there is no outlier



(c) Decision regions found by SoCal when there is no outlier

(d) Decision regions found by RLMM when there is no outlier

Figure 5: Each dot in the plots above represents a sample, with samples having HapMap reference genotype calls marked red. The ellipsoids were obtained using only the reference calls.

Figure 6 shows the decrease in concordance rate of SoCal and RLMM at call rate of 100% in leave-one-out cross-validation with HapMap reference genotype calls as the variance of simulated outliers varies from 0 to 10. Clearly, the concordance rate of SoCal decreases much more slowly than does the RLMM method. Thus, SoCal is in general more robust to outliers than RLMM.

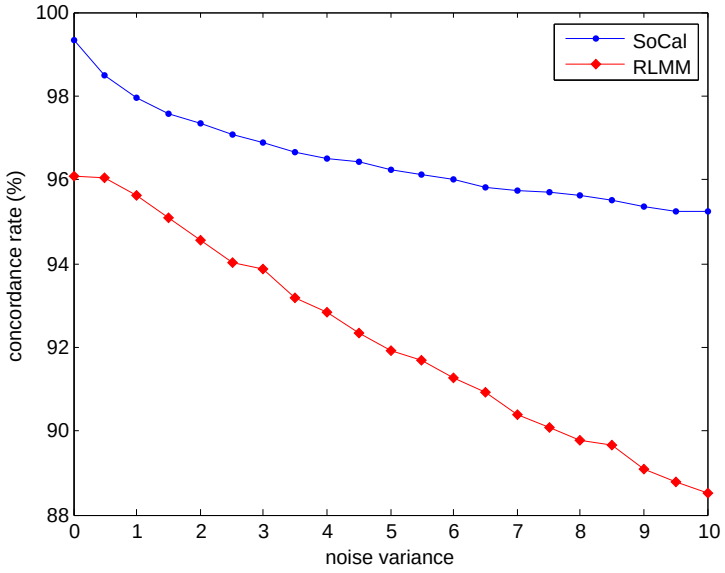


Figure 6: Concordance rate of SoCal and RLMM in the leave-one-out cross-validation with HapMap reference calls as a function of outlier variance.

4.4 Software implementation

SoCal is implemented in Python. To solve the conic programming problem of finding separating ellipsoids, SoCal uses CVXOPT. Source code of SoCal is available at `https://github.com/huwenboshi/wqe/tree/master/genotype_caller`.

5 Discussion

I have presented SoCal, a supervised genotype calling algorithm for Affymetrix SNP microarray. Unlike many existing supervised genotype calling algorithms that try to fit generative models (mostly Gaussian mixture models) on summarized allele intensities with reference genotype calls, SoCal uses these data to efficiently finds ellipsoidal decision regions for each genotype cluster by solving a conic programming problem. Both cross-validation of SoCal with HapMap reference calls and comparison with genotype calls made by CRLMM show that SoCal is comparable in accuracy to many of the state-of-the-art genotype calling methods. Also, when there are outliers in training data, SoCal outperforms RLMM, a genotype calling method that uses Gaussian decision regions to call genotypes, demonstrating the robustness of SoCal in the presence of outliers over existing methods. Overall, SoCal proves to be a promising alternative genotype caller for Affymetrix SNP microarray.

Like many supervised genotype callers, SoCal has its limitations. First, SoCal is not directly applicable to SNP microarrays that don't have reference genotype calls. Second, SoCal is, in general, also not directly applicable to the same SNP microarray across different laboratories because different laboratory protocols can generate different allele intensities for even the same SNP microarray. In these scenarios, one can first call genotypes from microarrays using unsupervised genotype calling algorithms. One can then use these calls as training data for SoCal and use SoCal to form refined ellipsoidal decision regions for each genotype cluster. Because SoCal is robust to outliers in the

training data, the refined ellipsoidal decision regions can be used to accurately call genotypes for future SNPs. Another limitation of SoCal is that users need to fine tune the weights (β_i) assigned to the criteria of finding the ellipsoids. However, experiments with SoCal show that SoCal is mostly sensitive to the choice of β_2 and β_3. And optimal values for parameters can be estimated efficiently through experiments.

SoCal is still in development, and can be improved and extended in many directions. First, the current method that SoCal uses to handle SNPs with missing genotype clusters is through simple and fixed geometric transformations. This method assumes that genotype AA cluter and genotype BB cluster are symmetric around genotype AB cluster. However, this is not true in general. Future efforts should be spent on how to estimate ellipsoids for missing genotype clusters more accurately. Second, currently SoCal only uses allele intensities data for reference genotype calls. However, allele intensities data for samples having structual variations is also available. A possible improvement for SoCal is to include these data in finding the ellipsoids for each genotype cluster to further refine the decision regions. These refined decision regions can then be used to detect possible structural variations.

To summarize, SoCal presents a novel and promising method for genotype calling. It's efficient in that it finds decision regions by solving a conic programming problem, which is solvable in polynomial time with guaranteed global optimum. SoCal is comparable in accuracy with many state-of-the-art methods. Although SoCal has its limitations, these limitations are also present in other supervised genotype callers, and have been addressed previously. SoCal can also be extended and improved to be more accurate and to have more functionality.

References

[1] Rho, S. W., Abell, G. C., Kim, K., Nam, Y., & Bae, J. (2010). Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends in Biotechnology, 28, 291-299.

[2] Norlén, H., Pettersson, E., Ahmadian, A., Lundeberg, J., & Sundberg, R. (2008). Classification of SNP genotypes by a Gaussian mixture model in competitive enzymatic assays. Mathematical Statistics Stockholm University Research Report, 3, 1-26.

[3] Lin, Y., Tseng, G. C., Cheong, S. Y., Bean, L. J., Sherman, S. L., & Feingold, E. (2008). Smarter clustering methods for SNP genotype calling. Bioinformatics, 24, 2665-2671.

[4] Wu, C. F. (1983). On the Convergence Properties of the EM Algorithm. The Annals of Statistics, 11, 95-103.

[5] Rabbee, N., & Speed, T. P. (2005). A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics, 22, 7-12.

[6] Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods, 11, 407-409.

[7] Glineur F. (1998). Pattern separation via ellipsoids and conic programming. (MS Thesis). Facult Polytechnique de Mons, Mons, Belgium.