Semi-unsupervised quantization of strawberry shape diversity in modern germplasm

Mitchell J. Feldmann & Steven J. Knapp

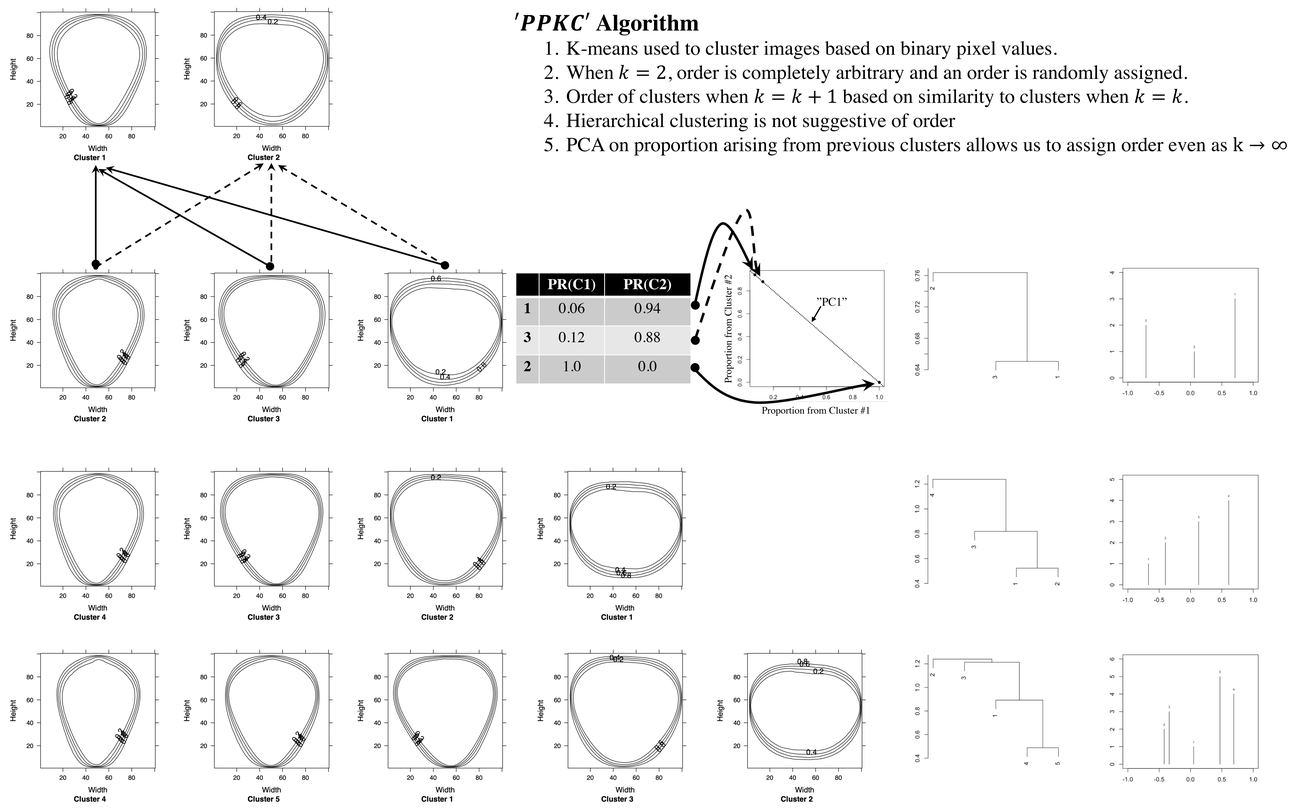
Department of Plant Sciences, University of California – Davis, Davis CA, 95616

@MitchFeldmann
mjfeldmann@ucdavis.edu
mjfeldmann

1. Summary

The shape of strawberry fruit is intimately linked, along with other visual attributes, to their marketability. To explore the existing morphospace in an unbiased method, K-means clustering was used to determine appropriate number of fruit categories at varying levels. A new method for transforming nominal descriptions onto an ordinal scale was applied to afford the use of robust statistical analyses common in quantitative genetics. Variance components estimated from multinomial mixed models indicated broad sense heritability (H^2) ranging from 0.65 for $k = 9$ and 0.98 for $k = 2$. Using a complex set of comprehensive and linear phenotypes, random forest regression was used to determine the most important quantitative phenotypes for classifying shapes. Of the 97 original quantitative descriptors only 15 were kept for classification. Mean accuracies from 10-fold cross validation ranged from 88-97% for discriminant models and 96-99% for support vector machines. These analyses suggest that there is a significant amount of genetic variation attributable to shape classifications and that few variables are needed for classification.

2. K-means Clustering & Principal Progression of *K* Clusters



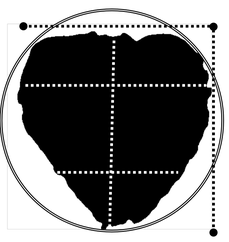
	PR(C1)	PR(C2)
1	0.06	0.94
3	0.12	0.88
2	1.0	0.0

$'PPKC'$ Algorithm

1. K-means used to cluster images based on binary pixel values.
2. When $k = 2$, order is completely arbitrary and an order is randomly assigned.
3. Order of clusters when $k = k + 1$ based on similarity to clusters when $k = k$.
4. Hierarchical clustering is not suggestive of order
5. PCA on proportion arising from previous clusters allows us to assign order even as $k \to \infty$

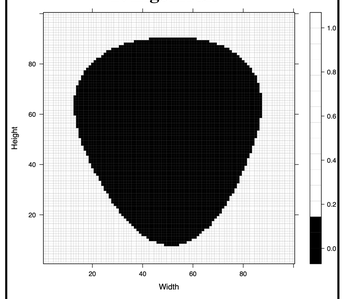
3. Quantitative Phenotypes

1. Linear & Geometric

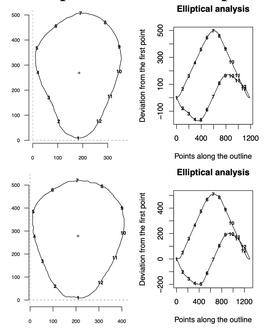


- Maximum Height & Width.
- Width at 25, 50 & 75% of height
- Aspect Ratio, Circularity, Solidity

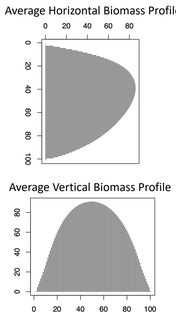
2. Eigenfruit



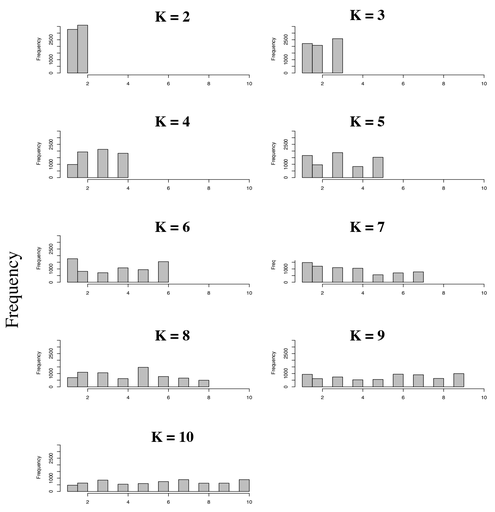
3. Elliptic Fourier Descriptors

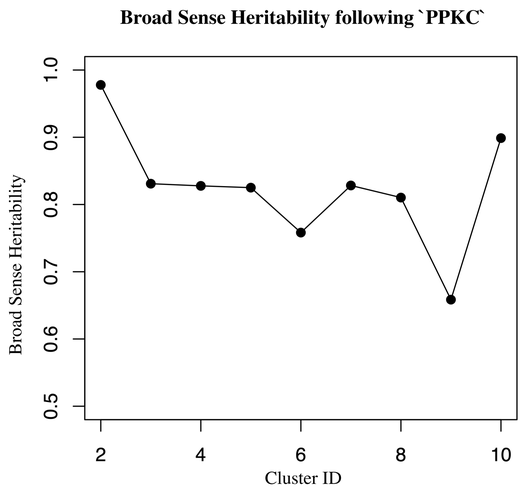


4. Biomass Profile



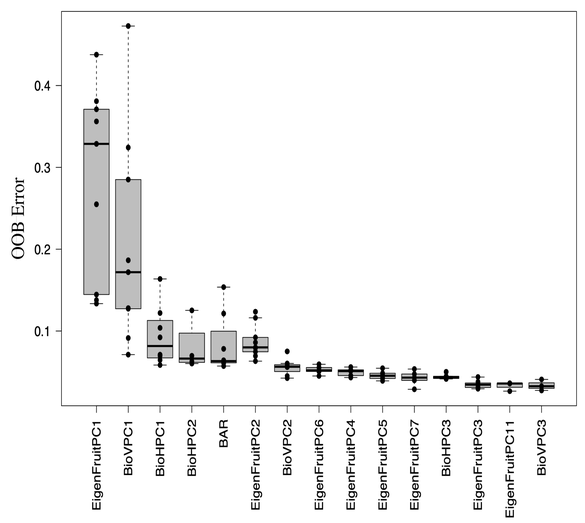
4. Distribution and H^2 of Ordered Clusters

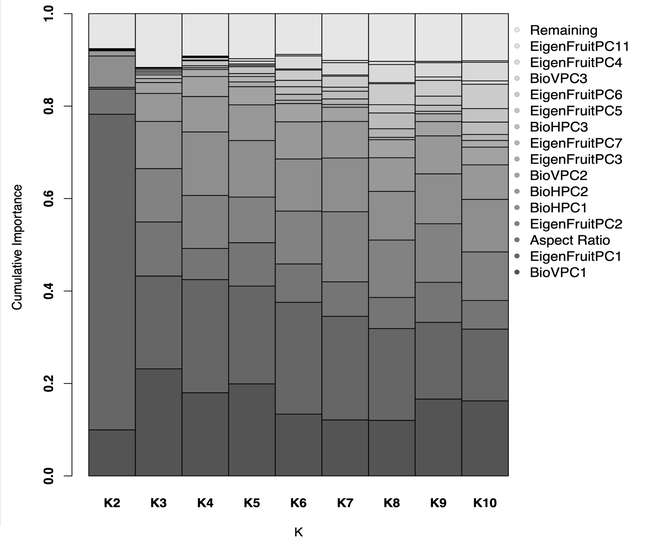




Left. Distribution of ordered clusters $k = 2{:}10$. **Right.** H^2 for each of the cluster values $k = 2{:}10$. Variance components estimated using ordinal::clmm() (Christensen, 2018)

5. Variable Selection using Random Forests

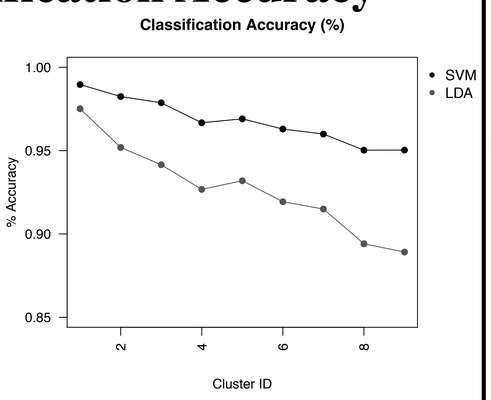




Left. OOB error of selected variables. **Right.** Importance of variables in classifying when $k = 2{:}10$. Random forest variable selection performed using vsurf::VSURF() (Genuer *et al*, 2018)

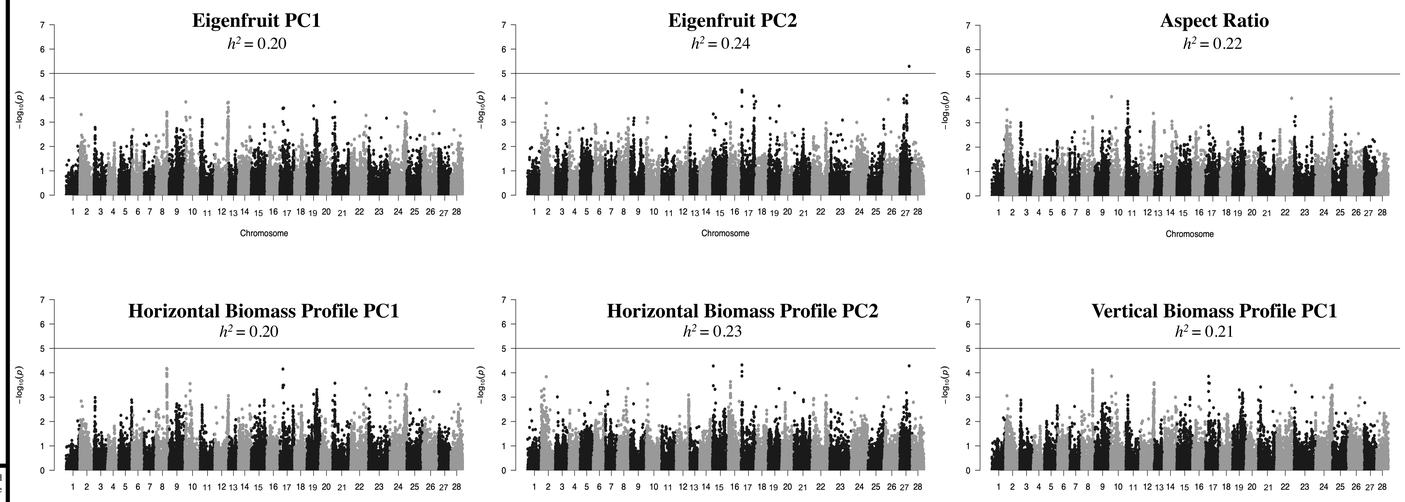
6. Heritability and Classification Accuracy

Variable	Times Selected	H^2
Eigen Fruit PC1	9	0.76
Eigen Fruit PC2	9	0.57
Eigen Fruit PC3	8	0.04
Eigen Fruit PC4	5	0.69
Eigen Fruit PC5	6	0.52
Eigen Fruit PC6	5	0.60
Eigen Fruit PC7	5	0.36
Eigen Fruit PC11	3	0.44
Vertical Biomass Profile PC1	9	0.75
Vertical Biomass Profile PC2	7	0.44
Vertical Biomass Profile PC3	3	0.05
Horizontal Biomass Profile PC1	8	0.70
Horizontal Biomass Profile PC2	4	0.62
Horizontal Biomass Profile PC3	5	0.73
Aspect Ratio	8	0.74



Left. Number of times a variable was selected and H^2 of selected variables. **Right.** Classification accuracy using the 15 selected variables. H^2 estimated using lme4::lmer() (Bates *et al*, 2015). Discriminant functions fit with MASS::lda() (Venables & Ripley, 2002). Support vectors fit with e1071::svm() (Meyer *et al*, 2018).

7. Genome Wide Association Scan



Top left to bottom right. Manhattan plots for the 6 variables with the largest OOB error. PC1 and PC2 of EigenFruit, Aspected ratio, PC1 and PC2 of horizontal biomass profile, and PC1 of vertical biomass profile. GWAS performed using rrblup::gwas() (Endelman, 2011).

Acknowledgments: This work is supported by the Specialty Crops Research Initiative (#2017-51181-26833) from the USDA National Institute of Food and Agriculture. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture. This work is also supported by the Henry A. Jastro Graduate Student Research Award.

References:

Turk, MA and Pentland, AP. 1991. Face recognition using eigenfaces. IEEE.

Turner *et al*. 2018. An automated image analysis pipeline enables genetic studies of shoot and root morphology in carrot (*Daucus carota* L.). Frontiers in Plant Science.