Uncertainty Quantification In Difficult Risk Models Built on Large Genetic Vectors

Grace Wahba

Based on "Distance Covariance For Improved Feature Screening"
Jing Kong *PhD* thesis and JK *et al* work in progress

SIAM Conference on Uncertainty Quantification
March 31-April 3, 2014
Savannah, Georgia

Links to these slides in my website
`http://www.stat.wisc.edu/~wahba/ -> TALKS`
Jing Kong's website
`http://www.stat.wisc.edu/~kong`

Uncertainty Quantification in Difficult Risk Models Built on Large Genetic Vectors

Abstract: We describe an approach to risk modeling in a biostatistical context. An extremely large number of observed variables are potentially related to an outcome of interest (yes or no). If the chance of a correct prediction (yes or no) is too close to guessing, then it is desired not to predict, but if the odds of being correct are good, then a prediction will be reported, along with an estimate of its accuracy. There are three steps: (i) Use Distance Correlation (DCOR) to select a subset of the variables that, taken together are related to the outcome of interest. (ii) Use a Support Vector Machine with Reject Option (SVM-R) to build a prediction model. (iii) Use a form of multiple cross validation (MCV) to build and test multiple models that assess the reliability of the variable selection and prediction process.

Outline

1. Introduction, application to Ovarian Cancer data with a large number of candidate variables and weak signals.
2. Variable Selection using Distance Covariance (DCOV)
3. Classification for difficult problems using the Suport Vector Machine with Reject Option (SVM-R)
4. Application to The Cancer Genome Atlas Ovarian Cancer Data
5. Multiple Cross Validation (MCV) to assess uncertainty
6. Comments and Conclusions
7. References

Quantifying Uncertainty in Hard Classification Problems With Large Attribute Vectors

We discuss a three step approach to risk factor modeling (DCOV, SVM-R, MCV) and build a model based on a cancer example where the attribute vector consists of 12,042 gene expression values given for 279 subjects which are known either to respond or to be resistant to a particular treatment. The approach makes no distributional assumptions for the attributes, and accommodates itself to the possibility that some fraction of the population is hard to classify from the genetic data, a third category "do not classify" (a. k. a. "reject") is an option. In the cancer data, a large fraction are not classified, but for those that are, the results are firm. A multiple cross validation is used to quantify uncertainty, and we see a commonly observed conundrum when cross validating through variable selection given a humongous number of candidates (to be described).

Sample Distance Covariance (DCOV)

For a random sample $(X, Y) = \{(X_k, Y_k) : k = 1, ..., n\}$ of n i.i.d random vectors (X, Y) from the joint distribution of random vectors X in R^p and Y in R^q, the Euclidean distance matrices $(a_{ij}) = (|X_i - X_j|_p)$ and $(b_{ij}) = (|Y_i - Y_j|_q)$ are computed. Define the double centering distance matrices

$$A_{ij} = a_{ij} - \overline{a}_{i\cdot} - \overline{a}_{\cdot j} + \overline{a}_{\cdot\cdot}, \quad i, j = 1, ..., n,$$

where

$$\overline{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \overline{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, \quad \overline{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij},$$

similarly for $B_{ij} = b_{ij} - \overline{b}_{i\cdot} - \overline{b}_{\cdot j} + \overline{b}_{\cdot\cdot}, \quad i, j = 1, ..., n.$

Sample Distance Covariance (DCOV) (continued)

The sample distance covariance $\mathcal{V}_n(X, Y)$ is defined by

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij} B_{ij}.$$

The sample distance correlation (DCOR) $\mathcal{R}_n(X, Y)$ is defined by

$$\mathcal{R}_n^2(X, Y) = \begin{cases} \dfrac{\mathcal{V}_n^2(X, Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0; \\ 0, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0, \end{cases}$$

where the sample distance variance is defined by

$$\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X, X) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^2.$$

Population Distance Covariance

Szekely and Rizzo (2009) defined the population distance covariance between $X \in \mathrm{R}^p$ and $Y \in \mathrm{R}^q$ to be

$$\mathrm{V}^2(X,Y) = \frac{1}{c_p c_q} \int_{\mathrm{R}^{p+q}} \frac{|f_{X,Y}(s,t) - f_X(s) f_Y(t)|^2}{|s|_p^{1+p} |t|_q^{1+q}} dt\, ds$$

where $f_{X,Y}(s,t)$, $f_X(s)$, and $f_Y(t)$ are the characteristic functions of (X,Y), X, and Y, respectively, and c_p, c_q are constants chosen to produce scale free and rotation invariant measure that doesn't go to zero for dependent variables. The idea originates from the property that the joint characteristic function factorizes under independence of the two random vectors. This leads to the remarkable property that $V^2(X,Y) = 0$ if and only if X and Y are independent. The sample version of DCOV is an estimate of the population DCOV.

Population Distance Covariance: Theorem

Li, Zhong and Zhu (2012) proposed using distance correlation for feature screening, but did not provide a necessary stopping criterion. The following theorem will provide a principled way of choosing a stopping criterion:

Theorem (J. Kong) Suppose we have random vectors $X \in \mathrm{R}^{p_1}, Z \in \mathrm{R}^{p_2}$ and $Y \in \mathrm{R}^{q}$, and assume Z is independent of (X, Y), then

$$V^2(X : Z, Y) \leq V^2(X, Y),$$

where $X : Z \in \mathrm{R}^{p_1+p_2}$ and V is the population distance covariance defined above.

Note that there are no distributional assumptions on the variables and the components may have quite disparate distributions.

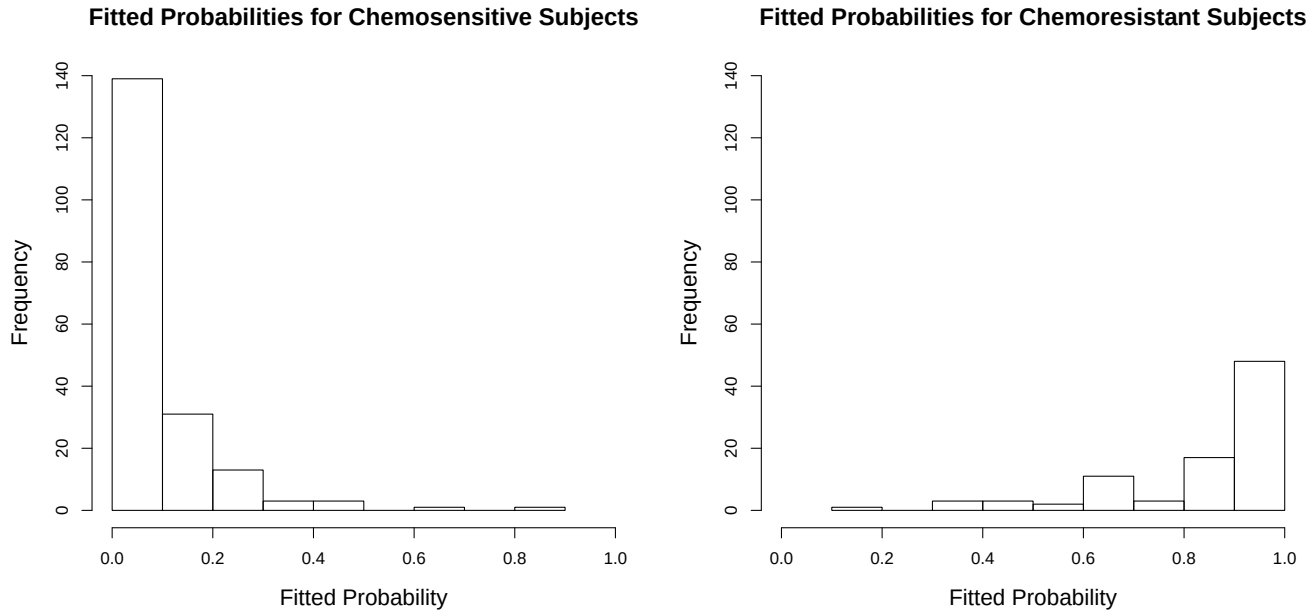
Select Variables using DCOV

1. Calculate marginal sample distance correlations for $x_i, i = 1, ..., p$ with the response.
2. Rank the variables in decreasing order of the sample distance correlations (DCOR). Denote the ordered variables as $x_{(1)}, x_{(2)}, ..., x_{(p)}$. Add $x_{(1)}$ to $x_{\mathcal{S}}$, defined as the set of variables included so far.
3. For i from 2 to p, keep adding $x_{(i)}$ to $x_{\mathcal{S}}$ if $\mathcal{V}_n^2(x_{\mathcal{S}}, y)$, the sample DCOV, does not decrease. Stop otherwise.

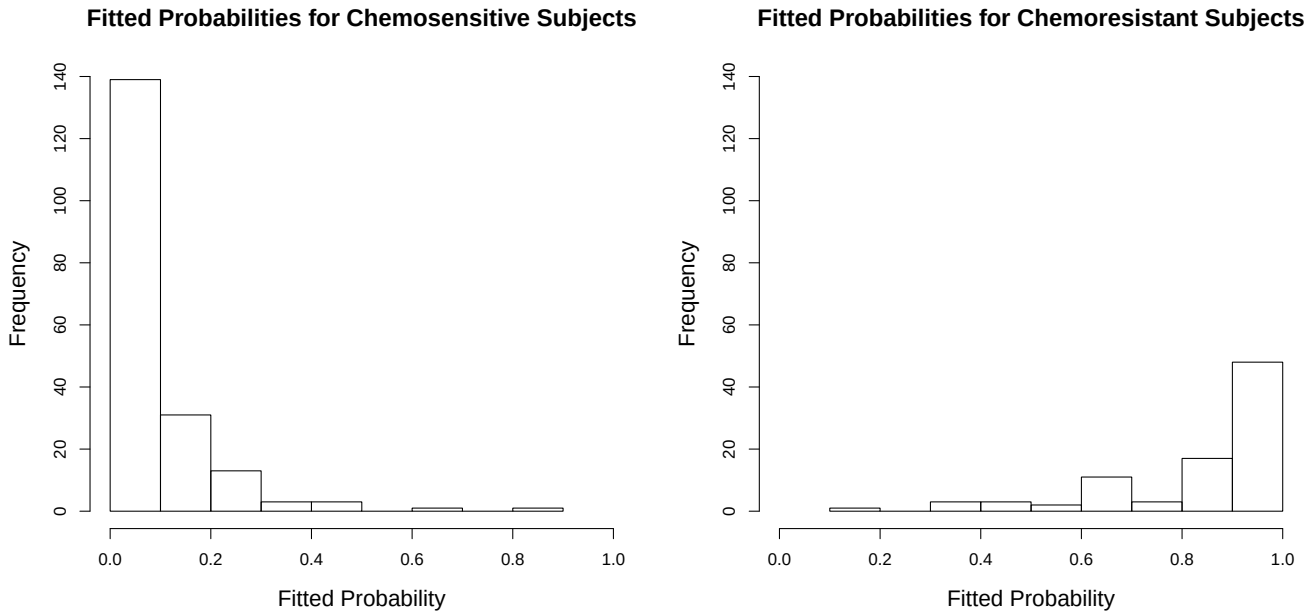
Apply the Procedure to TCGA Ovarian Cancer Data

We apply the above procedure to The Cancer Genome Atlas (TCGA) Ovarian cancer data. The TCGA collected high-quality, high-dimensional, and multi-modal genetic data from women with ovarian cancer. There were 279 samples with explicit chemostatus and gene expression (Affymetrix HT-HGU133a) data in the public set, among which 191 subjects are sensitive to chemotherapy and 88 are chemoresistant. Expression data for 12042 genes after log transformation are used for analysis. The issue is to explore whether there are genes whose expression pattern is strongly correlated with the response, i.e. chemotherapy status. Our feature screening procedure on the standardized log scale gene expression data for the 279 patients selects 82 genes, among which 5 were reported to be related to ovarian cancer in the literature, namely *IGFBP5, GPR3, MAPK4, FZD5*, and *FGF22*.

Fitted Probabilities of Being Chemoresistant, by Subject Label



Fitted probabilities of being chemoresistant for 191 chemosensitive subjects (left) and 88 chemoresistant subjects (right). (Bernoulli likelihood additive spline model on the 82 selected genes. R code gss with default tuning (GACV, Gu and Xiang 2001).



Fitted probabilities of being resistant have high density around small values for sensitive patients and large values for resistant patients respectively, with overlapping in the middle values. This suggests that we are less confident about the chemostatus for the patients in the middle range and may in practice want to withhold decision for such cases.

The Support Vector Machine With Reject Option (SVM-R)

The SVM-R was proposed in Bartlett and Wegkamp (2008) (see also Wegkamp and Yuan (2011), and is a practical way of solving the problem that there are some subjects for which you would like to make a decision and others for which you prefer not to, because the chance of making a mistake is not tolerable.

First, the usual two class SVM has data on n subjects, $y_i \in \{-1, 1\}$ with attribute vectors $x \in \mathcal{X}$, and it is desired to obtain a classifier $f \in \mathcal{F}$ which will provide a classification as $+1$ if $f(x)$ is positive and -1 if $f(x)$ is negative. Letting $\tau = yf$, the loss function is 1 if $\tau < 0$ and 0 otherwise. The so-called hinge function $(1 - \tau)$ for $\tau < 1$ and 0 otherwise (compactly written $(1 - \tau)_+$) is a convex upper bound to the desired loss function.

The Support Vector Machine With Reject Option (SVM-R) (continued)

The usual SVM finds f in some class $\mathcal{F}$ to minimize

$$\sum_{i=1}^{n}(1 - y_i f(x_i))_+ + \lambda J(f).$$

where $J(f)$ is some penalty functional. Popular examples include f in an RKHS with $J(f)$ the RKHS square norm, and f a linear combination of basis functions with $J(f)$ being the l_1 penalty on the coefficients. Under some general conditions, it is known that the SVM is actually estimating the sign of the log odds ratio (Lin *et al* 2002), which explains why it is so popular.

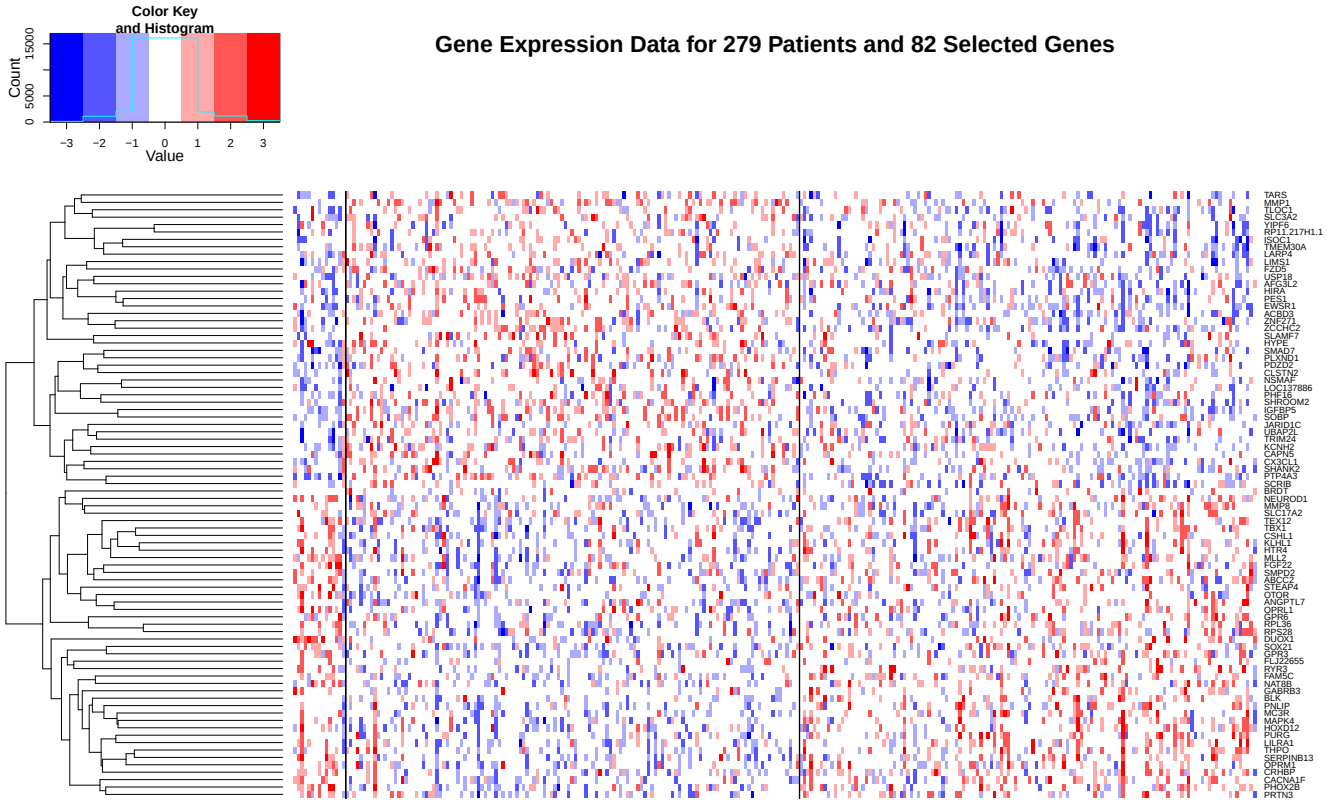
The Support Vector Machine With Reject Option (SVM-R) (continued)

Letting $d < 1/2$ be the cost of a reject, the SVM-R replaces the hinge function in the above equation with the generalized hinge function $(1 - a\tau), \tau < 0$, and $(1 - \tau)_+, 0 \leq \tau$, where $a = (1 - d)/d > 1$. It was shown that the SVM-R is a convex surrogate for the desired loss function of d for reject, 1 for a mistake and 0 otherwise.

For application to the Ovarian Cancer data, $\mathcal{F}$ consists of linear combinations of the log gene expressions, with the penalty functional being the l_1 norm of the coefficients.

The optimization problem is fast and easy to compute.

Gene Expression Data for 279 Patients and 82 Selected Genes



SVM-R results, $d = 1/4$. **l. to r. Resistant** 14 resistant subjects and 1 sensitive subject; **Sensitive**, 123 sensitives and 8 resistants, and **Do not classify**, 67 sensitives and 66 resistants.

Multiple Cross Validation (MCV) Train-Tune-Test Models

1. Randomly partition 279 samples: a $2/3 \times 4/5 = 8/15$ training set, a $1/5$ tuning set and a $1/3 \times 4/5 = 4/15$ testing set.

$$|---train----------|-tune-|--test--|$$

2. 12042 genes:, select genes using DCOV on the training set.
3. Build the SVM-R model on the training set with the selected genes for $d = 1/3, 1/4$ and $1/5$.
4. Use the tuning set to choose the tuning parameter for SVM-R.
5. Use the model with chosen tuning parameter to predict labels for the testing set.
6. Repeat 1.-5. 50 times.
7. Aggregate the prediction results for the 50 replications and apply majority votes for each subject.

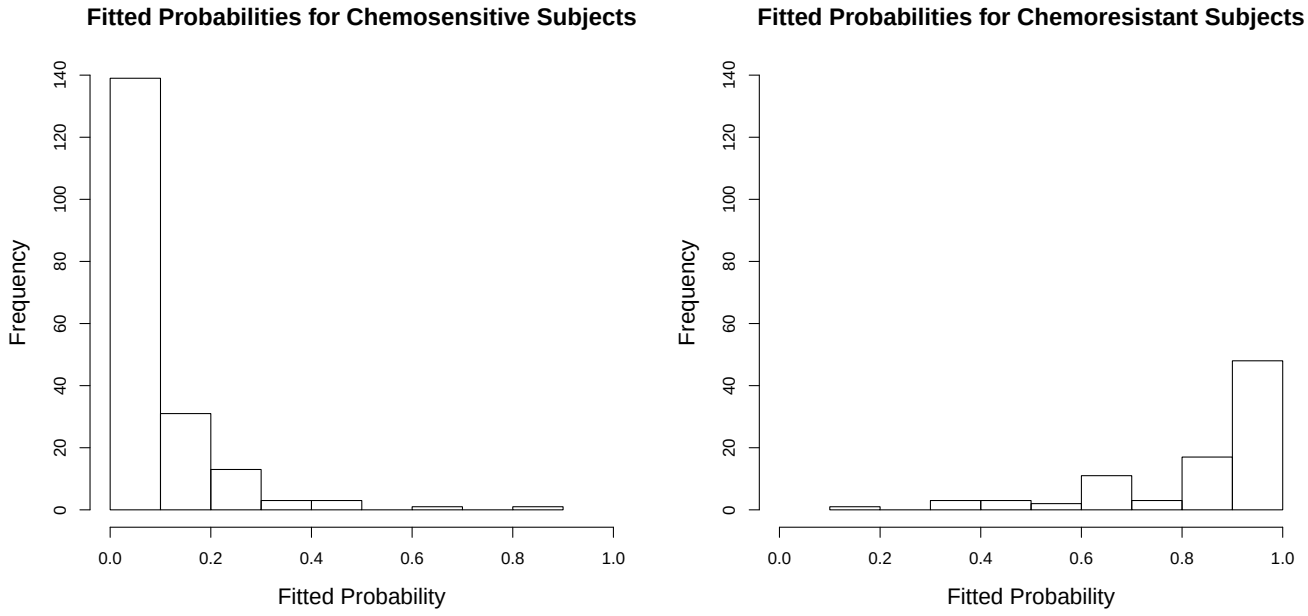
Results of train-tune-test and majority voting

	testing accuracy	# with decision
$d = 1/3$	0.6936	173
$d = 1/4$	0.9130	23
$d = 1/5$	0.9048	21

Testing accuracies for majority votes based on 50 random replications.

So, for $d = 1/4$ or $1/5$ we can get over 90% verified accuracy at a rather severe cost of making a decision for only about 7% of the data set. (These decisions were all chemosensitive)

Returning to the original penalized likelihood fit on 82 genes



Pretending this plot is ground truth, for about 90% accuracy, the rule would be:

Prob $\geq .9$ Chemoresistant

Prob $\leq .1$ Chemosensitive

Prob $\in (.1, .9)$ Do not classify

Then about 140 of the chemosensitive subjects and 50 of the chemoresistant subjects, or about 68% of the 279 subjects would be correctly identified.

Why not use that model? We don't really know how good it is. Model uncertainty due to variable selection in sampling 12042 variables with 279 subjects is not accounted for. This issue is common to selecting variables from a humongous number of candidates, in the not-low-hanging-fruit situation.

More details:

50 Sets of MCV Gene Expression Selections

The union of the 50 gene selections before SVM-R modeling consists of 1245 genes, and includes all 82 genes. 34 out of 1245 genes get selected at least 10 times, where 33 of them appear in the 82 genes, but very few appear in more than 25 runs. For $d = 1/5$, after the SVM-R models are run, 787 out of 1245 genes remain.

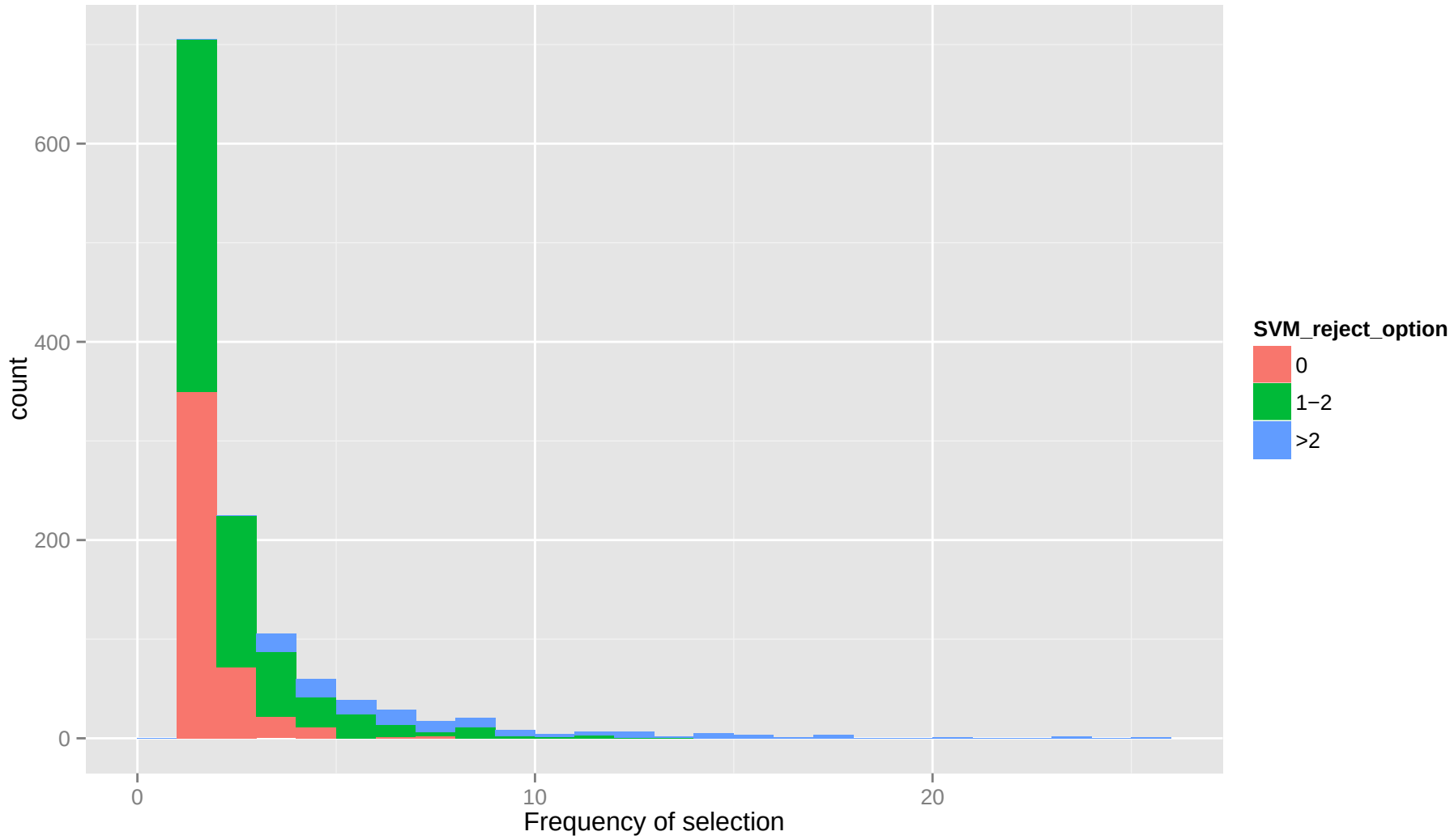


Figure 1: Frequency for 1245 genes being selected by DCOV method, pink color represents numbers deleted by SVM-R.

The following summarizes the mean training and testing accuracy and the mean number of subjects with decision in the 50 replicates, along with the previous voting scheme results.

	mean training accuracy(std)	mean testing accuracy(std)	mean number testing with decision(std)
$d = 1/3$	0.8606(0.0719)	0.7119(0.0547)	45.3400(12.1867)
$d = 1/4$	0.9356(0.0362)	0.8034(0.1320)	12.8400(10.2725)
$d = 1/5$	0.9593(0.0347)	0.8476(0.1181)	11.6600(9.6840)
$d = 1/3$	voting	0.6936	173
$d = 1/4$	voting	0.9130	23
$d = 1/5$	voting	0.9048	21

The voting scheme compares favorably to building an average model.

Comments and Conclusions

1. Is it worthwhile to attempt classification when only 7% of the model building population gets classified. The numbers here assume that the costs of the two kinds of errors are the same, something that surely depends on the application, and may depend on things that are subjective, like quality of life, or hard to predict, like longevity. Should insurance be required to pay for a new cancer pill that costs $100,000 per person? Science meets public policy, just as it does in climate modeling and global warming.

2. What about other possibly important variables? In Corrada Bravo *et. al.* 2009, in a paper based on the Beaver Dam Eye study, we argue that genetic, behavioral and clinical variables combine to predict the disease of interest. In the Ovarian Cancer study we have two additional variables with complete data: Cancer grade and cancer stage. Surprisingly, including these two in our models with genetic data did not affect the results in any practical way. Unfortunately, in human subjects data the more possibly-relevant variables are available, the easier it might be to identify deidentified subjects.

3. The large number of variables that appear only in a small number of runs suggests noise in the variable selection procedure. It could also suggest the conundrum that the "true" model consists of a large number of variables with modest effects of which different subsets give rise to roughly equal prediction ability. Options for further study in this and other difficult problems include allowing the DCOV stopping criteria to be modified by some amount δ, and allowing the greedy variable selection algorithm to be doubly greedy by testing the next best m of the remaining variables rather than just the next variable. It remains to obtain theoretical results to guide exploration in alternate scenarios.

4. Although we have applied these tools (DCOV, SVM-R, MCV) to biomedical data we argue that they are quite portable across disciplines, including Atmospheric and Earth sciences.

References

[1] G. Szekely and M. Rizzo. Brownian distance covariance. *Ann. Appl. Statist.*, 3:1236–1265, 2009.

[2] R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation. *J. Amer.Statist. Assoc.*, 107:1129–1139, 2012.

[3] C. Gu and D. Xiang. Cross-validating non-gaussian data: Generalized approximate cross-validation revisited. *JCGS*, 10:581–591, 2001.

[4] P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *JMLR*, 9:1823–1840, 2008.

[5] M. Wegkamp and M. Yuan. Support vector machines with a reject option. *Bernoulli*, 17:1368–1385, 2011.

[6] Y. Lin, G. Wahba, H. Zhang, and Y. Lee. Statistical properties and adaptive tuning of support vector machines. *Machine Learning*, 48:115–136, 2002.

[7] H. Corrada Bravo, K. E. Lee, B. E. K. Klein, R. Klein, S. K. Iyengar, and G. Wahba. Examining the relative influence of familial, genetic and environmental covariate information in flexible risk models. *Proceedings of the National Academy of Sciences*, 106:8128–8133, 2009. Open Source at `www.pnas.org/content/106/20/8128.full.pdf+html`, PMCID: 2677979.