Learning Genetic Risk Models Using Distance Covariance

Grace Wahba

Based on Jing Kong's PhD dissertation, to appear

and

"Using distance covariance for improved variable selection with application to learning genetic risk models"

Statistics in Medicine, online before print 29 January 2015

Jing Kong, Sijian Wang and Grace Wahba

Center for Statistics and Machine Learning at Princeton University

April 7, 2015

Princeton, New Jersey

Links to these slides in my website

`http://www.stat.wisc.edu/~wahba/` $->$ `TALKS`

Jing Kong's website

`http://www.stat.wisc.edu/~kong`

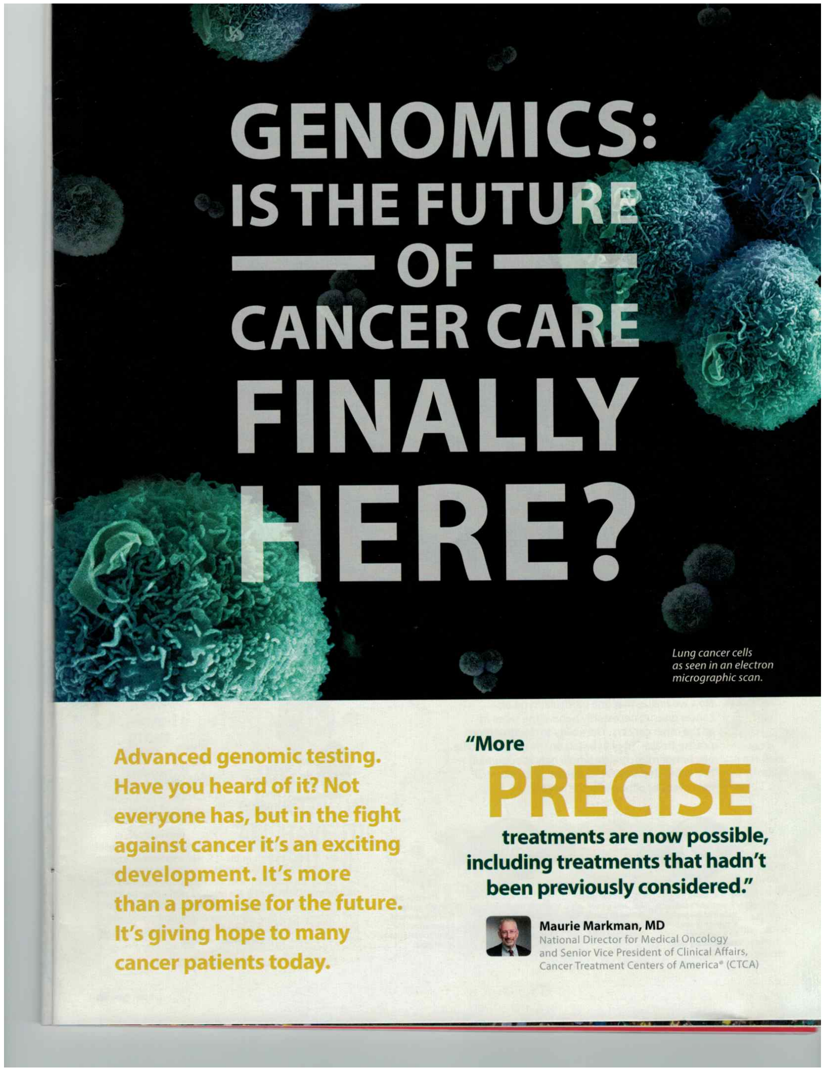
MARCH 30, 2015
TIME
Both of these women have brain tumors.
One is beating the odds.
CLOSING THE CANCER GAP
BY ALICE PARK
time.com

Woman on the left is being treated based on genetic information.

Both women have a glioblastoma.

- Woman on the left has her DNA sequenced for any of the few hundred mutations known to cause healthy cells to grow uncontrollably.
- Her mutations are checked against a database of ongoing clinical trials at Memorial Sloan Kettering to see if she might be eligible for experimental drugs that might do some good. One was found.
- Drug not available to woman on the right in Bismarck ND.

(From Time Magazine, March 30 2015)



GENOMICS:
IS THE FUTURE
OF
CANCER CARE
FINALLY
HERE?

Lung cancer cells as seen in an electron micrographic scan.

Advanced genomic testing. Have you heard of it? Not everyone has, but in the fight against cancer it's an exciting development. It's more than a promise for the future. It's giving hope to many cancer patients today.

"More PRECISE treatments are now possible, including treatments that hadn't been previously considered."

Maurie Markman, MD
National Director for Medical Oncology
and Senior Vice President of Clinical Affairs,
Cancer Treatment Centers of America® (CTCA)

For some, maybe.

Successes:
Herceptin-for Breast Cancer with a particular gene mutation (HER+)
Gleevec-for Leukemia with a particular genetic signature (Ph +CML)

For others, it may not be so simple. Complex patterns of mutations, possibly in pathways (Oncogenes, Tumor Suppressor Genes), or other genetic information, may be relevant to whether a particular subject responds to a particular treatment. Lots of candidate patterns.

"Cancer: The Emperor of All Maladies" on PBS

Learning Genetic Risk Models Using Distance Covariance (DCOV)

Abstract

We extend an approach suggested by Li, Zhong and Zhu (2012) to use distance covariance as a variable selection method by providing the DCOV Variable Selection Theorem, which gives a principled stopping rule for a greedy variable selection algorithm. We apply the resulting DCOV Variable Selection Method in two genetic based classification problems with small sample size and large vectors of gene expression data. The first problem involves the well known SBRCT (Small Blue Round Cell Tumor) childhood Leukemia data, which involves gene expression data from four different types of Leukemia, and it is well known that these data are easy to classify. The second involves Ovarian Cancer data from The Cancer Genome Atlas, and involves Ovarian Cancer patients that are either sensitive or resistant to a platinum based cancer chemotherapy. The Ovarian Cancer data presents a difficult classification problem.

Outline:

1. Review of DCOV
2. The DCOV Variable Selection Theorem
3. Easy Case, application to the SBRCT data
4. Hard Case, application to the Ovarian Cancer Data, results validated, but modest.
5. Summary and Comments

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(DCOV) $V_n(X,Y)$ is defined by

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

The sample distance correlation (DCOR) $R_n(X,Y)$ is defined by

$$R_n^2(X,Y) = \begin{cases} \dfrac{V_n^2(X,Y)}{\sqrt{V_n^2(X)V_n^2(Y)}}, & V_n^2(X)V_n^2(Y) > 0; \\ 0, & V_n^2(X)V_n^2(Y) = 0, \end{cases}$$

where the sample distance variance is defined by

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

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: The DCOV Variable Selection Theorem

Li, Zhong and Zhu (2012) proposed using distance correlation for feature screening, (thanks!) 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 $X \in \mathrm{R}^{p_1}, Z \in \mathrm{R}^{p_2}$, $Y \in \mathrm{R}^{q}$, and $(X : Z) \in \mathrm{R}^{p_1+p_2}$ and assume Z is independent of (X, Y), then

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

where V^2 is the population distance covariance.

This says that if variables Z independent of X and Y are added to X, then the DCOV of $(X : Z)$ with Y will decrease, or, at least not increase.

The DCOV Variable Selection Method

Let $\mathcal{S}$ be some set of variables and let $x_{\mathcal{S}} = \{x_i : i \in \mathcal{S}\}$ be the set of variables in $\mathcal{S}$.

1. Calculate marginal sample distance correlations for $x_i, i = 1, ..., p$ with the response.

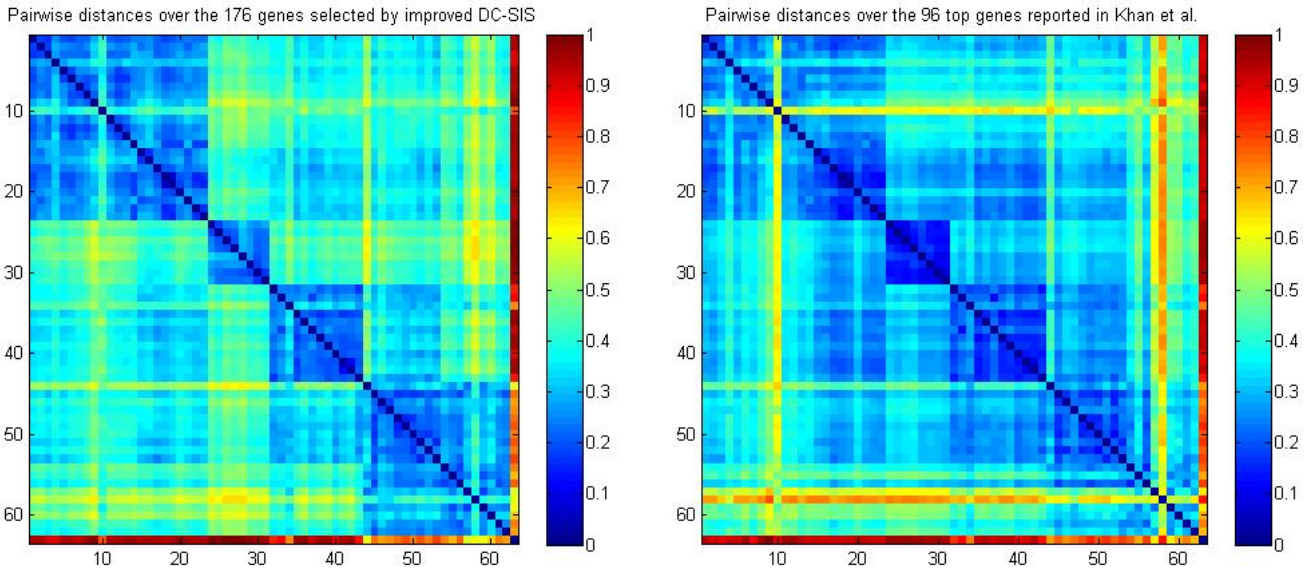
2. Rank the variables in decreasing order of the sample distance correlations. Denote the ordered variables as $x_{(1)}, x_{(2)}, ..., x_{(p)}$. Start with $x_{\mathcal{S}} = \{x_{(1)}\}$.

3. For i from 2 to p, keep adding $x_{(i)}$ to $x_{\mathcal{S}}$ if $V_n^2(x_{\mathcal{S}}, y)$, the sample DCOV, does not decrease. Stop otherwise.

Easy case: SRBCT childhood Leukemia data (Kahn *et al* (2001))

This data set consists of 83 subjects labeled with four types of Leukemia, and 2308 gene expression values. It is well known to be an easy classification problem and many authors have used the same 63 subjects as a training set, with generally near perfect classification on the test set of the remaining 20. For demonstration purposes we used the DCOV Variable Selection Method in a a one-vs-rest proceedure to obtain four sets of genes, which combined as 176 genes, 47 of which were in common with the 96 genes reported in Kahn *et al.*

Easy case: SRBCT childhood Leukemia data (cont.)

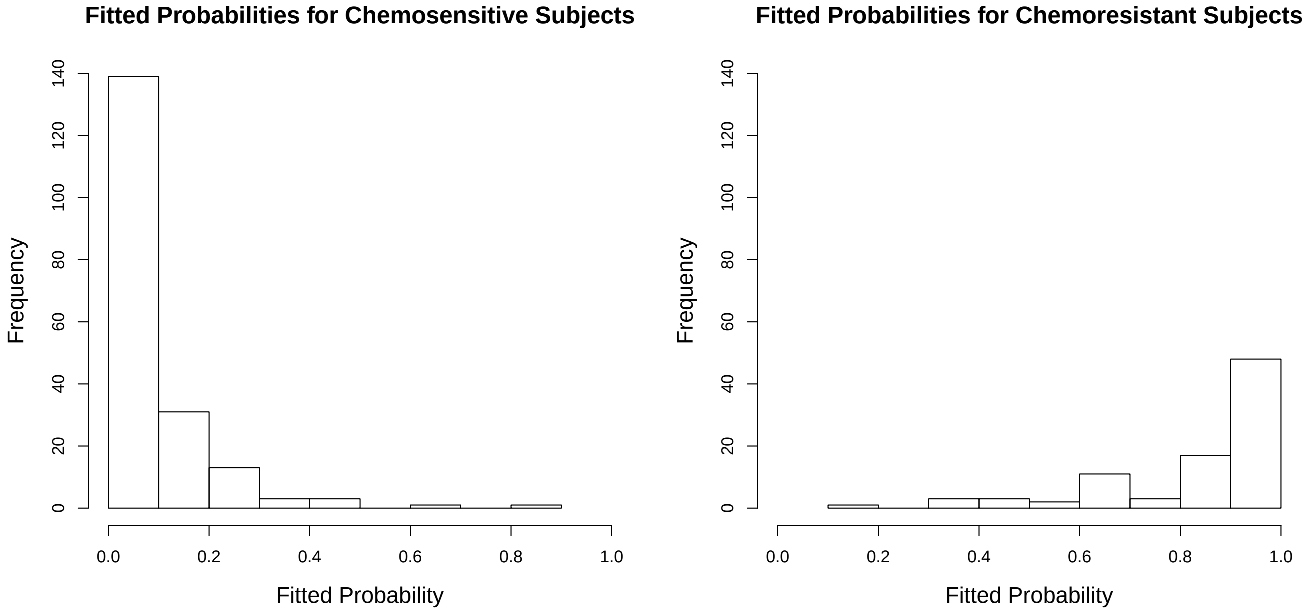


Left and right panels show pairwise distances of the 63 training samples using the 176 genes with the DCOV Variable Selection method and the 96 genes reported in Kahn *et al.* Four clusters are seen in both, and both sets can classify the 20 test set members perfectly via 3-nearest neighbors, although the number of genes in common is not large.

Hard case: The Cancer Genome Atlas (TCGA) Ovarian Cancer data (cancergenome.nih.gov)

The Ovarian Cancer data set contains a large amount of information on the ovarian cancer tumors of about 500 subjects, we focus on 279 subjects with 12,042 gene expressions, along with cancer grade and cancer stage. Subjects were recorded as either sensitive (191, or 68.5%) or resistant (88) to platinum-based combination chemotherapy. The goal is to see if a classification model can be built to assist in determining whether future cases will be sensitive. The DCOV Variable Selection method selected 82 genes from the 279 subjects.

Fitted probabilities of being resistant, by subject label

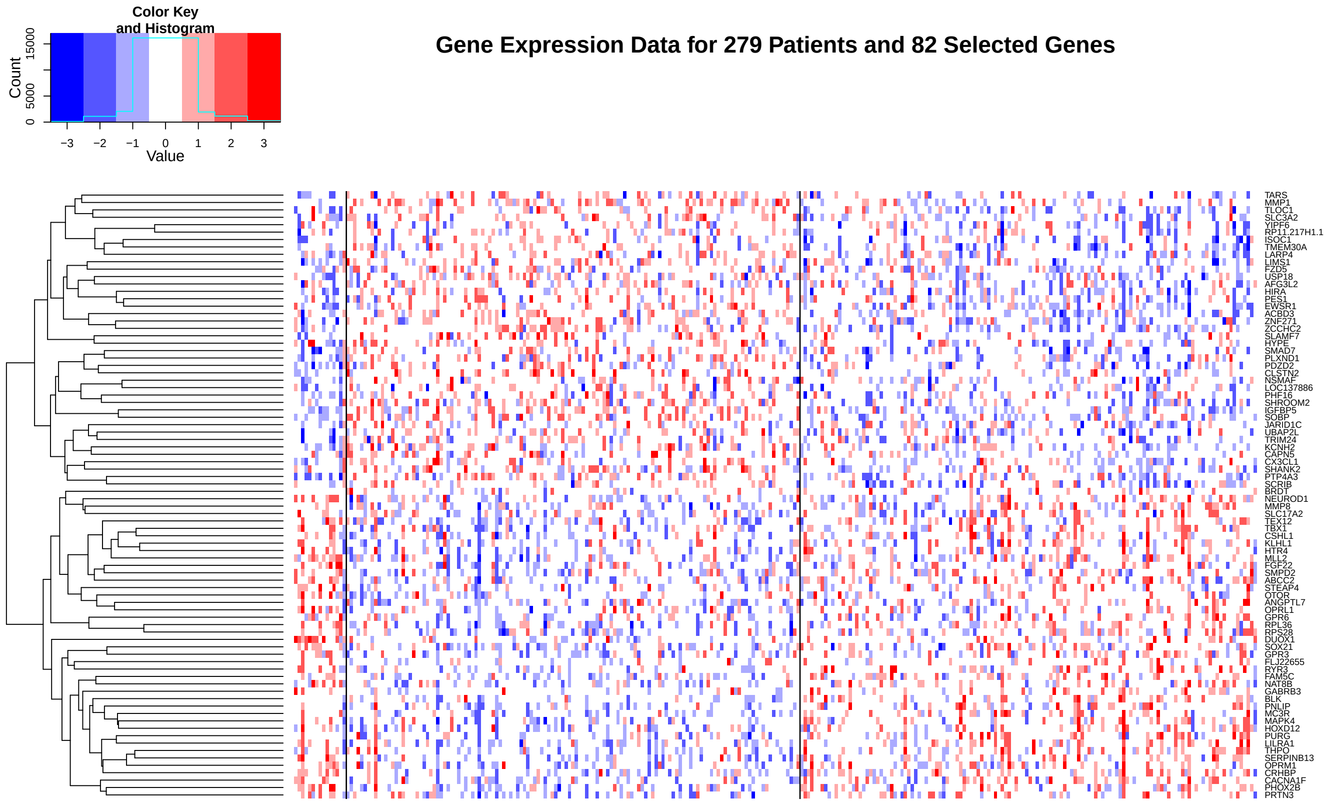


Estimated probabilities of being resistant for 191 sensitive subjects (left) and 88 resistant subjects (right). (Bernoulli likelihood additive spline model on the 82 selected genes. R code gss with default tuning (GACV, Gu and Xiang (2001)).

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

The SVM-R was proposed by Bartlett and Wegkamp (2008), see also Wegkamp and Yuan (2011). The usual two class SVM has data on n subjects, $y_i \in \{-1, 1\}$ with attribute vectors, and leads to a model which assigns every subject to either $+1$ or -1, with a cost of 1 if an erroneous decision is made. The SVM-R assigns every subject to either $+1, -1$, or "Reject", meaning that no assignment will be made. If an erroneous decision is made then the cost is 1 but if no assignment is made, the cost is d, where d is a parameter chosen by the user.

Gene expression data for 279 patients and 82 selected genes



SVM-R results, $d = 1/4, \lambda = 4$. **l. to r. Resistant** 14 resistant subjects and 1 sensitive subject; **Sensitive**, 123 sensitives and 8 resistants, and **Reject**, 67 sensitives and 66 resistants.

Five fold CV to choose λ in the SVM-R (l_1 penalty)

	S_1	S_2	S_3	S_4	S_5
S_1	53				
S_2	16	77			
S_3	23	21	87		
S_4	18	16	15	33	
S_5	27	30	31	21	94
82 genes	38	38	44	28	50

Pairwise intersections of $S_1, \ldots, S_5$ and the 82 genes. The diagonal numbers are the numbers of selected genes in each fold S_i. A small number of genes appeared in all five folds.

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. 12,042 genes:, select genes using DCOV on the training set.
3. Build the SVM-R model on the training set with the selected genes for $d = 1/4$ 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.

		num of reps with decision	mean training accuracy (std)	mean testing accuracy (std)	mean % training with decision	mean % test with decision
original	1/4	43	0.93(0.03)	0.78(0.12)	28%	27%
	1/5	37	0.94(0.04)	0.82(0.15)	16%	15%
permute	1/4	28	0.92(0.00)	0.68(0.08)	38%	34%
	1/5	9	0.97(0.00)	0.71(0.13)	35%	32%

First two rows: original data. Last two rows: data which has been permuted, so there should be no signal. Less cost for no decision means fewer decisions. Training accuracy for permuted data is about the same as the training accuracy for the original data. Testing accuracy for the permuted data is about the same as guessing. Testing accuracy on the original data here is higher, but with low decision rates.

A finer scoring system

Using results that vary across 50 CV replications a voting score v_i for each subject was computed as $v =$ (sensitive - resistant) divided by rejected, where the entries in v refer to counts among the 50 runs, for each person. $d = 1/5$.

		voting score				
		$(-0.1, 0]$	$(0, 0.1]$	$(0.1, 0.2]$	$(0.2, 0.4]$	$(0.4, 1.5]$
orig.	frequency	76	74	67	47	15
	proportion	0.5658	0.6486	0.7164	0.8085	0.9333
permu.	frequency	145	67	43	24	0
	proportion	0.6759	0.6866	0.7209	0.6667	NA

Newcomers could be classified by their voting scores. Partitioning the voting scores is conservative but leads to more convincing classification results.

Summary and Comments

- We have provided the DCOV Variable Selection Theorem and Method and applied the method to two classification problems, one easy, and one hard.
- For the easy case, a graphical display shows that classification will be easy, and that the DCOV Variable Selection Method (DCOV-VSM) provides relatively tight class clusters.
- For the hard case, a preliminary analysis suggests that a relatively modest portion of the data can be classified with some accuracy but a substantial fraction cannot. The DCOV Variable Selection Method is combined with SVM-R to classify some portion of a training set but not all. These techniques are combined with MCV to tune and validate the results. A finer voting scheme is proposed for (validated) classification, but only a fraction of the test cases are actually classified.

Summary and Comments (cont.)

In the Ovarian Cancer data, it became evident that there was a high level of variability among the folds in the number and choice of genes that were selected. This is a phenomenon that has been noted by many authors, in particular in the large p small n case. In the MCV training sets, $p = 12,042$, while n is only 148. In this data set it was, however, noted that the maximum distance correlation of the selected genes in each fold was actually close to that in the original data set. It is a challenge to understand the various scenarios that could explain the phenomena in such a large p very small n data set, but an important one, since such data sets are becoming common. It's possible that the "truth" has to do with the very complex nature of the relationship between genotype and phenotype, where, say, a very large number of variables with small but cumulative effect could occur, or complex pathways are reinforced or disrupted. Very interesting challenges remain.

References

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

[2] J. Kong, S. Wang, and G. Wahba. Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in Medicine*, 34:online before print, 29 January, 2015. PMID 25640961.

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

[4] J. Kahn *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artifical neural networks. *Nature Medicine*, 7(6):673–679, 2001.

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

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

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

Two related references:

[8] W. Shi, K. Lee, and G. Wahba. Detecting disease-causing genes by LASSO-patternsearch algorithm. *BMC Proceedings*, 1(Suppl 1):S60, 1–5, 2007. PMCID: PM2367607.

[9] F. Lu, S. Keles, S. Wright, and G. Wahba. A framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences*, 102:12332–12337, 2005. Open Source at www.pnas.org/content/102/35/12332, PMCID: PMC118947.