Research Article

Statistics in Medicine

Received 1 September 2014, Accepted 14 January 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6441

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

Jing Kong,[a*†] Sijian Wang[a,b] and Grace Wahba[a,b,c]

Variable selection is of increasing importance to address the difficulties of high dimensionality in many scientific areas. In this paper, we demonstrate a property for distance covariance, which is incorporated in a novel feature screening procedure together with the use of distance correlation. The approach makes no distributional assumptions for the variables and does not require the specification of a regression model and hence is especially attractive in variable selection given an enormous number of candidate attributes without much information about the true model with the response. The method is applied to two genetic risk problems, where issues including uncertainty of variable selection via cross validation, subgroup of hard-to-classify cases, and the application of a reject option are discussed. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: distance correlation; variable selection; SVM with reject option; TCGA ovarian cancer data; penalized Bernoulli likelihood

1. Introduction

The idea of feature screening came along as high-dimensional data were collected in modern technology. It was aimed at dealing with the challenges of computational expediency, statistical accuracy, and algorithmic stability because of high dimensionality. Fan and Lv proposed the sure independence screening (SIS) [1] and showed that the Pearson correlation ranking procedure possessed a sure screening property for linear regression with Gaussian predictors and responses. A new feature screening procedure for high-dimensional data based on distance correlation [2], named DC-SIS, was presented in [3]. DC-SIS retained the sure screening property of the SIS and additionally possessed new advantages of handling grouped predictors and multivariate responses by using distance correlation. Moreover, because distance correlation was applicable to arbitrary distributions, DC-SIS could also be used for screening features without specifying a regression model between the response and the predictors and thus was robust to model mis-specification.

However, both SIS and DC-SIS relied on a user-specified model size d, which decided the number of predictors being selected. Let the sample size be n, d was chosen to be multipliers of the integer part of $n/\log n$ in [1] and [3], which did not depend on any other characteristics of the data. As pointed out by a referee of [3], the choice of d was of great importance in practical implementation and might influence the screening results significantly. Our study is aimed at fixing this shortcoming by including an automatic stopping criteria for DC-SIS based on the property of distance covariance (DCOV).

The screening procedures may fail if a feature is marginally uncorrelated, but jointly correlated with the response, or in the reverse situation where a feature is jointly uncorrelated but has higher marginal correlation than some important features. An iterative SIS was proposed in [1] to fix this problem. Current research interest involves dealing with this drawback, but this work is not related to this quest.

[a] *Department of Statistics, University of Wisconsin–Madison, 1300 University Avenue, Madison, WI, 53706, U.S.A.*
[b] *Department of Biostatistics & Medical Informatics, University of Wisconsin–Madison, 1300 University Avenue, Madison, WI, 53706, U.S.A.*
[c] *Department of Computer Sciences, University of Wisconsin–Madison, 1300 University Avenue, Madison, WI, 53706, U.S.A.*
[*] *Correspondence to: Jing Kong, Department of Statistics, University of Wisconsin–Madison, 1300 University Avenue, WI, 53706, U.S.A.*
[†] *E-mail: kong@stat.wisc.edu*

We demonstrate our improved method through two real examples. The small round blue cell tumors (SRBCTs) data were relatively easy to classify and had been studied extensively. The Cancer Genome Atlas (TCGA) ovarian cancer data, however, were much more challenging because of the large number of genes and limited sample size. The target was to identify the important genes that contribute to the sensitive or resistant status after receiving a particular chemotherapy treatment. A substantial fraction of the population was difficult to classify and a 'withholding decision' option is allowed in the support vector machine with reject (SVM-R) option model to adapt to this fact. A multiple cross validation (MCV) is used to quantify uncertainty given a humongous number of candidates, and we see a commonly observed dilemma that different variables are selected by using different subsets of the data. Comparison between the results from the original data and those from the data obtained by randomly permuting the response provide further justification on our conclusions. Furthermore, the MCV on the permuted data discloses the existence of spuriously correlated variables in high-dimensional data and thus failure of variable selection and model building based on training data.

2. Some preliminaries

2.1. Distance correlation

Reference [2] proposed distance correlation as a measurement of dependence between two random vectors. The method has been successfully applied to various problem, see [4] for example. To be specific, the authors defined the DCOV between $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ to be

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

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 does not go to 0 for dependent variables. The idea is originated 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 and distance correlation involves pairwise distances. For a random sample, $(X,Y) = \{(X_k, Y_k) : k = 1, \ldots, n\}$ of n i.i.d random vectors (X,Y) from the joint distribution of random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$, the Euclidean distance matrices $(a_{ij}) = (|X_i - X_j|_p)$ and $(b_{ij}) = (|Y_i - Y_j|_q)$ with $i,j = 1, \ldots, n$ are computed. Define the double centering distance matrices

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

where

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

similarly for $B_{ij} = b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}$, $i,j = 1, \ldots, n$. Then, the sample 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 $R_n(X,Y)$ 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

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

2.2. Feature screening via distance correlation (distance correlation-sure independence screening)

Reference [1] proposed SIS procedure based on the Pearson correlation for feature selection. The distance correlation version of this technique (DC-SIS) was studied in [3]. With a user-specific model size d, the variables whose distance correlations with the response ranking from first to dth in decreasing order were selected. The authors explored the theoretic properties of the DC-SIS and proved that the DC-SIS kept the desired sure screening property established in [1]. Moreover, because of the property of distance correlation, DC-SIS procedure was robust to model mis-specification, which was demonstrated in their simulations.

3. Improving distance correlation-sure independence screening using distance covariance

Theorem 1
Suppose random vectors $X, Z \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, and assume Z is independent of (X, Y), then

$$V^2(X+Z, Y) \leqslant V^2(X, Y), \tag{1}$$

where V is the population distance variance defined in [2].

Proof

$$\begin{aligned} V^2(X+Z, Y) &= \| f_{X+Z,Y}(t,s) - f_{X+Z}(t) f_Y(s) \|^2 \\ &= \frac{1}{c_p c_q} \int_{\mathcal{R}^{p+q}} \frac{1}{|t|_p^{1+p} |s|_q^{1+q}} \left| f_{X+Z,Y}(t,s) - f_{X+Z}(t) f_Y(s) \right|^2 dtds. \end{aligned}$$

The following fact follows from the definition of characteristic function and independence assumption.

$$\begin{aligned} &\left| f_{X+Z,Y}(t,s) - f_{X+Z}(t) f_Y(s) \right|^2 \\ &= \left| E e^{it^T(X+Z)+is^T Y} - E e^{it^T(X+Z)} E e^{is^T Y} \right|^2 \\ &= \left| E e^{it^T X + is^T Y} E e^{it^T Z} - E e^{it^T X} E e^{it^T Z} E e^{is^T Y} \right|^2 \\ &= \left| f_{X,Y}(t,s) f_Z(t) - f_X(t) f_Z(t) f_Y(s) \right|^2 \\ &= \left| f_Z(t) \right|^2 \left| f_{X,Y}(t,s) - f_X(t) f_Y(s) \right|^2, \end{aligned}$$

Because $|f_Z(t)| \leqslant 1$ by the property of characteristic function,[‡] we have

$$\left| f_{X+Z,Y}(t,s) - f_{X+Z}(t) f_Y(s) \right|^2 \leqslant \left| f_{X,Y}(t,s) - f_X(t) f_Y(s) \right|^2,$$

which implies

$$\begin{aligned} V^2(X+Z, Y) &\leqslant \frac{1}{c_p c_q} \int_{\mathcal{R}^{p+q}} \frac{1}{|t|_p^{1+p} |s|_q^{1+q}} |f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2 dtds \\ &= \| f_{X+Z,Y}(t,s) - f_{X+Z}(t) f_Y(s) \|^2 \\ &= V^2(X, Y). \end{aligned}$$

□

We know that if $E|X|_p < \infty$, $E|X+Z|_p < \infty$, and $E|Y|_p < \infty$, then almost surely

$$\lim_{n\to\infty} V_n^2(X+Z, Y) = V^2(X+Z, Y),$$
$$\lim_{n\to\infty} V_n^2(X, Y) = V^2(X, Y).$$

[‡]*In* [5]*, the author obtained equality here, which is incorrect.*

Thus, for the sample DCOV, if n is large enough, we should have

$$V_n^2(X+Z,Y) \leqslant V_n^2(X,Y),$$

under the assumption of independence between (X,Y) and Z.

In the case where (X,Z) is of interest, which is the usual situation for variable selection setting, we could use the preceding theorem by incorporating degenerated random vectors as follows. Suppose $X \in \mathbb{R}^{p_1}$ and $Z \in \mathbb{R}^{p_2}$, then we augment X and Z to be $\tilde{X} = (X, 0_{p_2})$ and $\tilde{Z} = (0_{p_1}, Z)$, respectively. $\tilde{X}$ and $\tilde{Z}$ are therefore of the same dimension and $\tilde{X} + \tilde{Z} = (X,Z)$.

We implemented the preceding theorem as a check for stopping for DC-SIS. For the original DC-SIS, it required a user-specified model size d, which was always chosen as multipliers of the integer part of $n/\log n$. For our improved screening procedure with distance correlation, we first ranked the importance of $x_i, i = 1, \ldots, p$ using the marginal distance correlations with the response as DC-SIS did and initialized $\mathcal{S}$ as the singleton including the index of the top one variable. Instead of selecting the top d variables, we kept adding variables to $x_{\mathcal{S}} = \{x_i : i \in \mathcal{S}\}$ according to the ordered list of variables until observing a decrease in the DCOV between $x_{\mathcal{S}}$ and y. The procedure took the following steps, and we denoted the procedure as DCOV method.

(1) Calculate marginal distance correlations for $x_i, i = 1, \ldots, p$ with the response.
(2) Rank the variables in decreasing order of the distance correlations. Denote the ordered variables as $x_{(1)}, x_{(2)}, \ldots, x_{(p)}$. Start with $x_{\mathcal{S}} = \{x_{(1)}\}$.
(3) For i from 2 to p, include $x_{(i)}$ to $x_{\mathcal{S}}$, that is, updating $x_{\mathcal{S}}$ by the concatenated variables $(x_{\mathcal{S}}, x_{(i)})$, if $V_n^2(x_{\mathcal{S}}, y)$ does not decrease. Stop otherwise.

4. Real application on small round blue cell tumor data

The SRBCTs are four different childhood tumors named so because of their similar appearance on routine histology, which makes correct clinical diagnosis extremely challenging. However, accurate diagnosis is essential because the treatment options, responses to therapy, and prognoses vary widely depending on the diagnosis. They include Ewing's family of tumors (EWS), neuroblastoma (NB), non-Hodgkin lymphoma (in our case, Burkitt's lymphoma, BL), and rhabdomyosarcoma (RMS). The SRBCT data being published in [6] included the expression of 2308 genes measured on 63 samples (23 EWS, 8 BL, 12 NB, and 20 RMS). These data are known as an easy-classified example and have been studied by many [7]. Using the multicategory SVM is one of several methods that have excellent classification results on this dataset. Hence, we focus more on the selected genes.

We applied our improved feature screening procedure on this dataset and compared our selection of genes with the 96 top genes reported in [6]. This is a multicategory classification, and the genes were screened in a one-versus-rest fashion. Specifically, for each of the four different types of tumors, we generated a response indicator vector taking a value of 0 if the sample came from the current interested type and 1 otherwise. This allowed us to implement our screening procedure and obtained the genes, which showed high distance correlation with the current type of tumor. The four groups of selected genes were combined as a whole collection of in total 176 genes. Forty-seven genes turned out to be in common for the DCOV selection and the top 96 genes used in [6].

We further examined the power of these two groups of genes in differentiating the four types of tumors by presenting the pairwise distances of the 63 samples (Figure 1). As shown in the plot, the samples were arranged in the order of EWS, BL, NB, and RMS. The pairwise distances resulted from the two selections of genes were scaled to maximum of 1, respectively, so that they shared the same magnitude. Both groups of genes could tell the four types of tumors apart. Compared with the 96 genes from [6], however, the 176 DCOV selected genes show better distinguishability and clearer contrast over the four classes. Moreover, the right panel almost missed the samples labeled from 57 to 62 in the class of RMS, but the 176 DCOV genes could recognize them with big differences between the in and out class pairwise distances. The dataset were known to be easy for classification, and both sets of genes were able to classify the testing set of 20 samples perfectly via k-nearest neighbor method with $k = 3$.

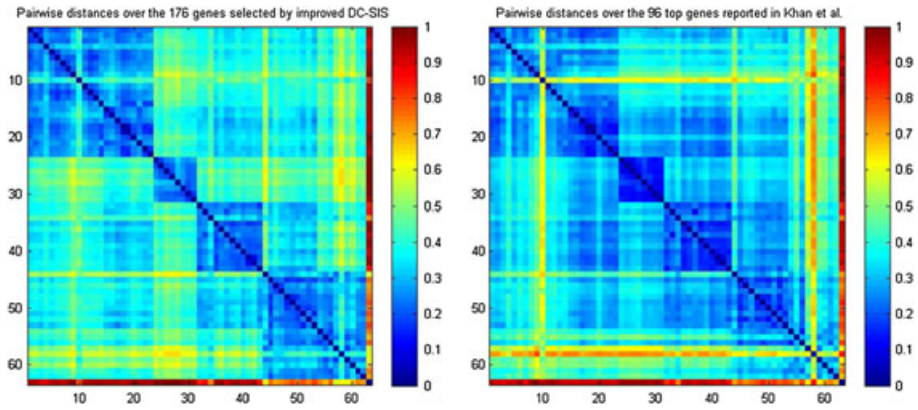

Figure 1. Comparison of pairwise distances between the two selections of genes. Left and right panels present the pairwise distances of the 63 samples over the improved distance correlation-sure independence screening (DC-SIS) selection of 176 genes and the 96 reported genes in [6], respectively.

5. Real application on The Cancer Genome Atlas ovarian cancer data

5.1. Data description

Ovarian cancer is the fifth leading cause of cancer deaths among women in the United States; 22,240 new cases and 14,030 deaths were estimated to have occurred in 2013 [8, 9]. The standard treatment for high-grade serous ovarian cancer is aggressive surgery followed by chemotherapy. Despite treatment, a vast majority of ovarian cancer patients eventually relapse and die of their disease with a major cause of chemotheraphy resistance [10]. Identification and prediction of patients with chemoresistant thus become important for improving the outcome of ovarian cancer.

The Cancer Genome Atlas collected high-quality, high-dimensional, and multimodal 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 were sensitive to chemotherapy, and 88 were chemoresistant. Expression data for 12,042 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 indicating chemotherapy status.

5.2. Distance covariance gene selection results based on all the observations

Our feature screening procedure on the gene expression data for the 279 patients selected 82 genes, among which five were reported to be related to ovarian cancer in the literature. IGFBP5 ranked as the fifth of the six members of insulin-like growth factor-binding protein (IGFBP) family and is known to be important for cell growth control, induction of apoptosis, and other IGF-stimulated signaling pathways. IGFBP5 expression is shown to prevent tumor growth and inhibited tumor vascularity in a xenograft model of human ovarian cancer and is suggested that IGFBP5 plays a role as tumor suppressor by inhibiting angiogenesis [11]. GPR3, the seventh, is a member of a family of G-protein couple receptors whose activation of protein kinase A (PKA) and subsequent increase of cyclic adenosine monophosphate (AMP) level promotes meiotic arrest in the oocyte [12]. Mice deficient in GPR3 display premature ovarian aging and loss of fertility [13]. MAPK4, the 18th, is a member of mitogen-activated protein kinases (MAPK) signaling pathway. MAPK signal transduction cascade is dysregulated in a majority of human tumors [14]. It is suggested to play an important role in molecular diagnostics and molecular therapeutics for low-grade ovarian cancer [15]. FZD5 ranked as the 22th encodes Frizzled-5 protein, which is believed to be the receptor for the Wnt5A ligand [16]. The Wnt5A ligand plays a context-dependent role in human cancers. It has been demonstrated that Wnt5a is expressed at significantly lower levels in human epithelial ovarian cancer cell lines and in primary human epithelial ovarian cancers compared with either normal ovarian surface epithelium or fallopian tube epithelium [17]. FGF22, the 56th, is a member of fibroblast growth factors (FGFs) family, whose members possess broad mitogenic and cell survival activities and are involved in a variety of biological processes, including embryonic development, cell growth, morphogenesis, tissue repair, tumor growth, and invasion. The inhibition of FGFR2, which is a member of this family, has been found to increase cisplatin sensitivity in ovarian cancer [18].

Thirty-nine pathways were found to be associated with the 82 genes, among which three pathways are known to be important for ovarian cancer. MAPK signaling pathway is suggested to play an important role in molecular diagnostics and molecular therapeutics for low-grade ovarian cancer [15]. Wnt signaling pathway is best known for its role in tumorigenesis. Bast and Mills [15] demonstrated the difference in Wnt signaling pathway between normal ovarian and cancer cell lines and between benign tissue and ovarian cancer. They also pointed out that those differences implicate that Wnt signaling leads to ovarian cancer development despite the fact that gene mutations are uncommon. Yin *et al*. [19] suggested that genetic variants in the transforming growth factor beta (TGF-β) signaling pathway are associated with ovarian cancer risk and may facilitate the identification of high-risk subgroups in the general population.

5.3. Support vector machine with reject option

We estimated the probabilities of being chemosensitive or chemoresistant by a penalized Bernoulli likelihood main effect spline model using the `R` package `gss` [20]. Aside from the additive expression effects of the selected 82 genes, we also included two more covariates, namely the cancer grade and cancer stage of the subjects. Cancer grade is an indicator for grades 2 and 3. Cancer stage indicates whether the subject is in stages IIIC and IV or not. As shown in Figure 2, the estimated probabilities have high density around small and large values for sensitive and resistant patients, respectively, with overlapping in the middle values. This suggested that we were less confident about the chemostatus for the patients in the middle range, and so we sought a principled approach, which withholds decision for such cases.

References [21,22] investigated the SVM-R for binary classification where the results of classification could be $-1, +1$ or withhold decision. Given a discriminant function $f : \mathcal{X} \to \mathbb{R}$, the method only reports $sgn(f(x)) \in \{-1, 1\}$ if $|f(x)| > \delta$ and withholds decision if $|f(x)| \leqslant \delta$. Suppose that the cost of making a wrong decision is 1 and that of rejecting to make a decision is $d \in [0, \frac{1}{2}]$, then an proper risk function is

$$L_{d,\delta}(f) = El_{d,\delta}(Yf(X)) = P\{Yf(X) < -\delta\} + dP(|Yf(X)| \leqslant \delta)$$

with the discontinuous loss function

$$l_{d,\delta}(z) = \begin{cases} 1, & \text{if } z < -\delta; \\ d, & \text{if } |z| \leqslant \delta; \\ 0, & \text{otherwise.} \end{cases}$$

The classifier associated with the discriminant function

$$f_d^*(x) = \begin{cases} -1, & \text{if } \eta(x) < d; \\ 0, & \text{if } d \leqslant \eta(x) \leqslant 1 - d; \\ +1, & \text{if } \eta(x) > 1 - d, \end{cases}$$

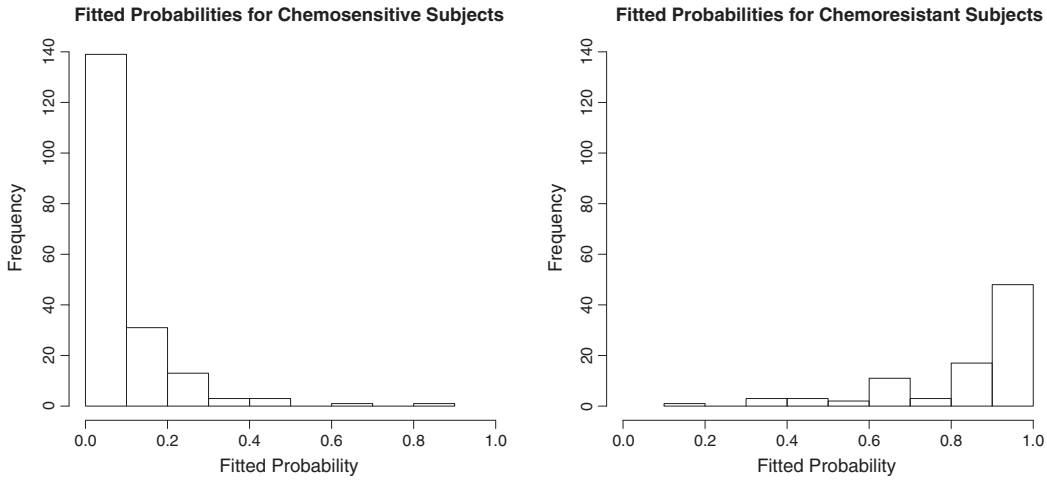

Figure 2. Fitted probabilities by penalized Bernoulli likelihood model with the 82 genes.

with $\eta(x) = P\{Y = +1|X = x\}$ minimizes the risk $L_{d,\delta}(f)$ with

$$L_d^* = L_{d,\delta}(f_d^*) = E\min\{\eta(X), 1-\eta(X), d\}.$$

To avoid working with the discontinuous loss $l_{d,\delta}$, [21,22] proposed a convex surrogate loss, which is the generalized hinge loss,

$$\phi_d(z) = \begin{cases} 1-az, & \text{if } z < 0; \\ 1-z & \text{if } 0 \leqslant z < 1; \\ 0, & \text{otherwise,} \end{cases}$$

where $a = (1-d)/d \geq 1$. It followed that f_d^* also minimizes the risk associated with ϕ_d over all measurable $f : \mathcal{X} \to \mathbb{R}$.

The discriminant functions f took the form $f_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x)$ based on a set of known functions $f_j : \mathcal{X} \to \mathbb{R}$ and coefficients $\lambda_j \in \mathbb{R}, 1 \leqslant j \leqslant M$. The coefficients were chosen to minimize the empirical risk

$$\hat{R}_\phi(f_\lambda) = \frac{1}{n}\sum_{i=1}^n \phi(Y_i f_\lambda(X_i)).$$

To reflect the preference for sparse solutions, which is desirable when M is large compared to the sample size n, an l_1 type restriction $\|\lambda\|_1 = \sum_{j=1}^M |\lambda_j|$ was incorporated in [21] and f_λ is estimated by $f_{\hat{\lambda}_r}$, where

$$\hat{\lambda}(r) = \arg\min_{\lambda\in\mathbb{R}^M} \hat{R}_\phi(f_\lambda) + r\|\lambda\|_1 \tag{2}$$

and $r > 0$ is a tuning parameter. We followed [21] to call this model SVM-R.

The authors in [22] also showed that the choice of $\delta = 1/2$ gives the optimal bound established by the excess risk of ϕ_d on the excess risk $L_{d,\delta} - L_d^*$ for any fixed $d \in [0, 1/2)$ and measurable function f. For this reason, we fixed $\delta = 1/2$ for our practical use of the method. Furthermore, we took the set of known functions $f_j : \mathcal{X} \to \mathbb{R}$ to be linear functions of the log transformation on the 12,024 genes. The optimization problem (2) was formulated into a linear programming task and solved using MATLAB.

Figure 3 presents the 82 genes for the 279 subjects in groups according to the SVM-R classification results. The results correspond to the particular choice of $d = 1/4$ and $r = 4$ to illustrate the benefits of SVM-R. As shown in the plot, there is a big difference in the gene expression between the subjects assigned to be resistant and sensitive. The behavior of the 82 genes for those without a certain decision tends to be in-between.

5.4. *Fivefold cross validation*

In order to choose the tuning parameter in SVM-R, we need to hold aside a tuning set before selecting the genes. Leaving out different observations leads to different gene selection results. Here we applied a fivefold cross validation analysis to examine the variations of selections of genes and SVM-R model fitting results across different partitions of the dataset. The implementation followed the steps in the succeeding text.

(1) Randomly partition the 279 subjects into five non-overlapping folds.
(2) Select genes from the 12,024 genes based on four folds as the training set.
(3) Build SVM-R model with the selected genes and the two cancer status variables based on the training set.
(4) Use the leaving-out fold as the tuning set to choose the tuning parameter for SVM-R with mean $l-$loss, defined in the succeeding text, as the criteria.
(5) Repeat 2–4 for the five folds.

The $l-$loss for a subject is 1 if a misclassification occurs, d if a withholding decision is made, and 0 otherwise. The mean $l-$loss is the average over the $l-$losses for all the subjects in a given set of data. We looked for tuning parameter values minimizing the mean $l-$loss.

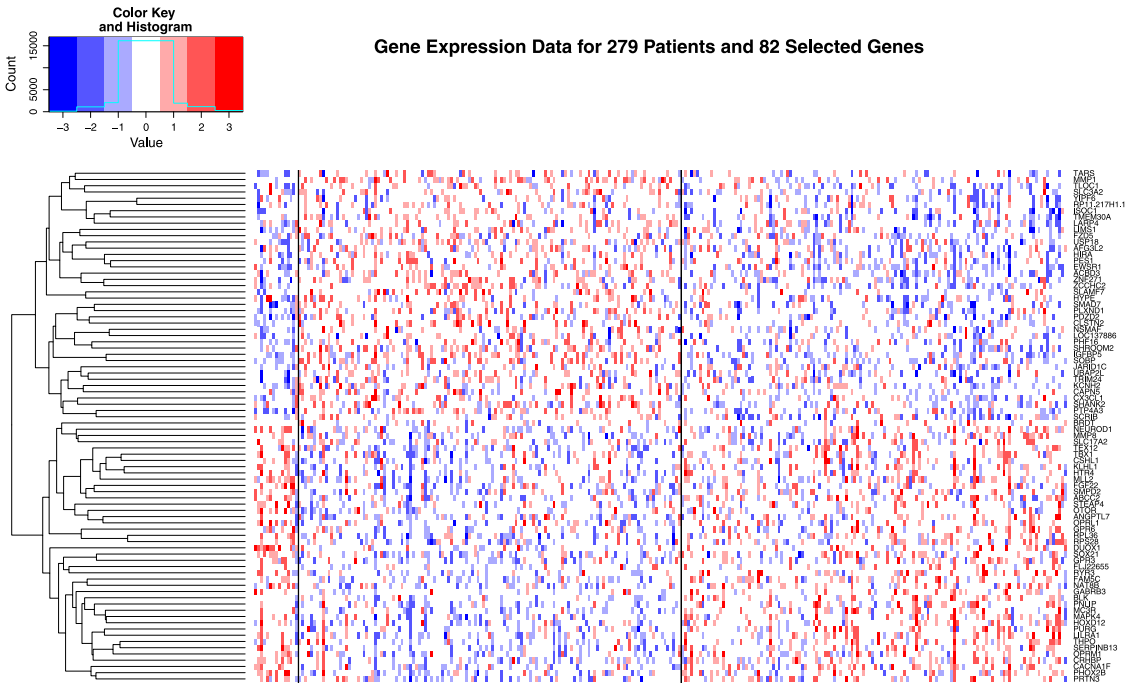

Figure 3. Gene expression data for the 82 selected genes and 279 subjects with support vector machine with reject a option classification for $d = 1/4$ and $r = 4$. The subjects are grouped according to their assigned decisions by the support vector machines with a reject option. The left group involves 15 patients (one sensitive and 14 resistant) classified to be resistant. The middle group is assigned to be sensitive and contains 123 sensitive and eight resistant subjects. Sixty-seven sensitive patients and 66 resistant patients with a withhold decision are shown in the right group.

Table I. Pairwise intersections of $S_1, \ldots, S_5$ and the 82 genes.

	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

The diagonal numbers are the numbers of selected genes in each S_i.

The aforementioned procedure produced five selections of genes before SVM-R, namely $S_1, \ldots, S_5$. In addition, we also have the 82 genes selected from all the subjects previously. Table I presents the pairwise intersections of these six sets with each other. The union of the five selections includes 211 genes, which covers 77 genes in the 82 genes. Seventy-three out of 211 genes have frequency more than 1 where 63 of them appear in the 82 genes. After implementing SVM-R, the union of genes is reduced to 98 genes. Figure 4 displays the histogram of these 211 genes colored by the frequency after SVM-R runs for $d = 1/5$. The pink region denotes the parts further eliminated by SVM-R, which is consistent with the DCOV selection in that SVM-R further rules out the genes with low frequency in the union.

5.5. Multiple cross validation

In order to consider uncertainty in variable selection and model building due to different partitions of the dataset, we further extended the fivefold cross validation to MCV and assessed the prediction power through the following procedure. The results were summarized in the upper part of Table II.

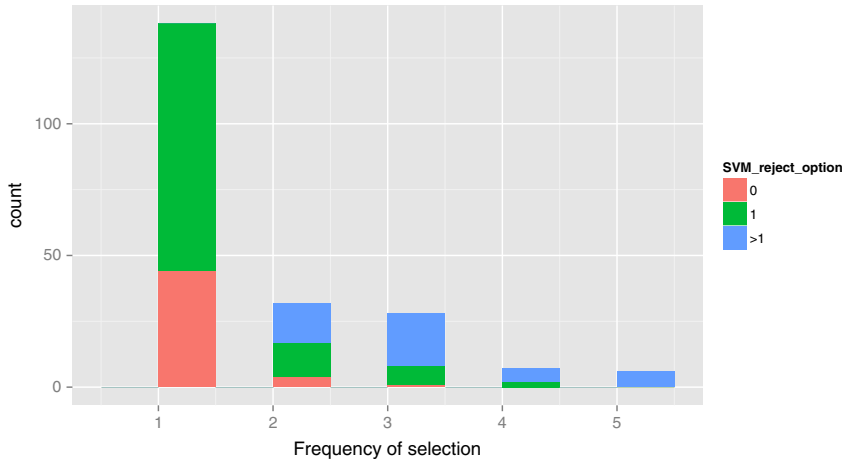

Figure 4. Frequency for the union of $S1, \ldots, S5$, colored by frequencies after support vector machine (SVM) with a reject option for $d = 1/5$.

Table II. Results for the 50 individual replications for $d = 1/3, 1/4$, and $1/5$.

	d	Number of replications with decision	Mean training accuracy(std)	Mean testing accuracy(std)	Mean number training with decision (std)	Mean num test with decision (std)
Original	1/3	50	0.8319(0.0914)	0.6943(0.0544)	101.9400(27.8043)	49.1800(15.4295)
	1/4	43	0.9371(0.0336)	0.7807(0.1250)	43.0698(27.0338)	20.1860(12.9638)
	1/5	37	0.9420(0.0358)	0.8215(0.1460)	24.4595(21.3874)	11.4865(10.0626)
Permute	1/3	49	0.7984(0.0078)	0.6910(0.0426)	112.1837(26.4352)	55.5918(13.9954)
	1/4	28	0.9225(0.0023)	0.6867(0.0810)	56.9643(21.4346)	25.2857(10.4132)
	1/5	9	0.9686(0.0015)	0.7071(0.1322)	53.4444(28.3333)	24.5556(13.0682)

The upper and lower parts are results for the original and permuted data, respectively. The third column shows the number of replications out of 50 with at least one definite decision made on the testing set. The fourth and fifth columns of the table conclude the mean training and testing accuracies with standard deviation (std) in the parenthesis, respectively, restricted to the repetitions with decision made. The last two columns display the mean and std for the number of patients assigned decisions for the training and testing sets, respectively, given the replications with decision made.

(1) Randomly partition 279 samples into a 1/5 tuning set, a $2/3 \times 4/5 = 8/15$ training set and a $1/3 \times 4/5 = 4/15$ testing set.
(2) Select genes from the 12,024 genes using the proposed method on the training set.
(3) Build SVM-R model using the selected genes and the two cancer status variables based on the training set.
(4) Use the tuning set to choose the tuning parameter for SVM-R with mean $l-$loss as the criteria.
(5) Use the model with chosen parameter to predict labels for the testing set.
(6) Repeat 1–5 for 50 times.

To understand more about the 50 models, we further explored the prediction results for $d = 1/5$. The prediction labels from the 50 models were aggregated, following the idea of ensemble methods. The result for each individual was recorded in a vector of three frequencies, namely the frequency of being classified as sensitive subjects, the frequency of obtaining a withholding, and the frequency of being assigned to be resistant out of the 50 models. Let (s_i, w_i, r_i) be the vector for the ith patient.

A finer analysis was conducted by looking at the strength of being sensitive or resistant according to (s_i, w_i, r_i). A voting score v_i was defined as $(s_i - r_i)/w_i$. Hence, a positive v_i indicated a tendency of being sensitive, whereas a negative v_i suggested more possibility of being resistant.

To understand the meaning of v_i, we first divided the voting scores into four consecutive intervals, each covering about 1/4 of the population. The last interval was further partitioned into two to identify a subgroup of patients with extremely large v_i's and high homogeneity in the class label. Table III (upper part) presents the five intervals and describes the distribution of v_i's as well as the proportion of sensitive subjects within each range of v_i, compared to the overall proportion of sensitive patients, that is,

Table III. Frequency of voting score v_i's and proportion of sensitive subjects in each subinterval for $d = 1/5$.

		Voting score				
		$(-0.1, 0]$	$(0, 0.1]$	$(0.1, 0.2]$	$(0.2, 0.4]$	$(0.4, 1.5]$
Original	Frequency	76	74	67	47	15
	Proportion	0.5658	0.6486	0.7164	0.8085	0.9333
Permuted	Frequency	145	67	43	24	0
	Proportion	0.6759	0.6866	0.7209	0.6667	NA

The upper and lower parts correspond to the original and permuted data, respectively.

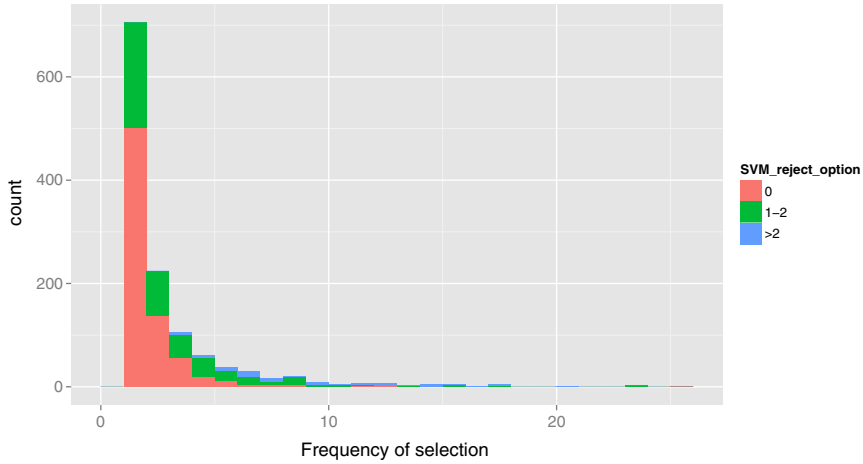

Figure 5. Frequency for 1245 genes being selected by DCOV method, colored by frequencies after support vector machine (SVM) with a reject option for $d = 1/5$.

191/279, in the population. It turned out that the trend of being sensitive weakened monotonically as the voting score decreased. The stratification specified a subgroup of 15 patients, who possessed the greatest voting scores, with very high accuracy to be chemosensitive. The next highest voting score subgroup of 47 subjects also showed strong confidence of being sensitive compared to the sample proportion. The conclusion from partitioning the voting scores was conservative but led to more convincing and steady classification results.

Each replication of the 50 MCVs gave rise to a different collection of selected genes. This issue is common to selecting variables from a humongous number of candidates, in the not-low-hanging-fruit situation. The union of the 50 gene selections before SVM-R consisted of 1245 genes and included all the 82 genes discussed previously. Thirty-four out of 1245 genes were chosen at least ten times, where 33 of them appeared in the 82 genes, but very few appeared in more than 25 runs. The l_1 penalty provided additional elimination, and for $d = 1/5$, 498 out of 1245 genes remained after SVM-R runs. Figure 5 displays the histogram for the 1245 genes before SVM-R. We distinguished their frequency after SVM-R by different colors. It is shown that a large number of genes with low frequency are further deleted by SVM-R model, that is, pink color.

5.6. Permutation of the response

Our method involved several components, including variable selection, SVM-R, MCV, and the voting score, which were interacting with each other and led to 15 patients with over 93% accuracy to be sensitive for $d = 1/5$. To further understand the mechanism and to demonstrate that the outcomes were not produced by noises, we randomly permuted the response and went through the whole procedure to compare the results with those for the original data.

It followed that the DCOV method selected genes spuriously correlated with the permuted response based on the training data in each replication of the 50 MCVs. The maximum distance correlation value of the selected genes in each repetition was very close to that for the original data. The highly corre-

lated genes appeared because of the high dimensionality of over 12,000 genes and less than 200 training samples.

However, the MCV step played the role of a safeguard against the fake signals. As Table II depicted, the mean training accuracies for the original and permuted data showed similar behavior for the original and permuted data, meaning that the selected genes were indeed important for the training data. Thus, the chosen genes in the permutation set should provide little prediction power for the tuning and testing data. Hence, the validation sets selected large tuning parameter values driving all the patients with no decision for many of the 50 replications for $d = 1/4$ and $1/5$. This did not happen for $d = 1/3$ because the sample ratio $191/279$ is slightly greater than $1/3$. For the rest of the replications with decision making, the mean testing accuracies for the permuted data remained at the level of the sample proportion of sensitive subjects for all three values of d, which deviated much from the increasing pattern in the original data.

These suggested that the MCV procedure was able to provide double fail-secure for fake signals. On the one hand, SVM-R placed a cap on the conditional probability of misclassfication and eliminated the replications where the selected genes could not produce results achieving the specified confidence on the validation set. On the other hand, the mean testing accuracies on the replications passing through the safeguard of tuning sets would be no better than assigning everyone to the sensitive class when there was no real signal.

Furthermore, the poor prediction performance of the 50 individual models ended up with unsurprisingly disappointing voting score results for the permuted data, as shown in the lower part of Table III. Many of the patients obtained a relatively small value of the voting score, and nobody obtained a score in the range where the original data had the highest 6accuracy, meaning that the confidence was quite low. Moreover, the stratified ratios of sensitive subjects for different ranges of the voting scores did not show anything insightful other than being around the sample proportion.

6. Summary of procedures in The Cancer Genome Atlas ovarian cancer data

Several components were included in the TCGA ovarian cancer data. Here we summarize the pieces in the algorithm in the succeeding text so that potential users are able to follow the procedure with new datasets.

(1) Set replication number N for MCV and $p_{train}, p_{test} \in (0, 1)$ with $p_{train} + p_{test} \leqslant 1$ for the proportion of training and testing sets.
(2) For i runs from 1 to N
    (a) Randomly partition the data into a p_{train} training set, a p_{test} testing set, and a $1 - p_{train} - p_{test}$ tuning set. (If the model in mind does not need parameter tuning, one can omit the tuning set with $p_{train} + p_{test} = 1$.)
    (b) Select variables using DCOV on the training set
        - Calculate marginal distance correlations for $x_j, j = 1, \ldots, p$ with the response on the training set.
        - Rank the variables in decreasing order of the distance correlations. Denote the ordered variables as $x_{(1)}, x_{(2)}, \ldots, x_{(p)}$. Start with $x_S = \{x_{(1)}\}$.
        - For j from 2 to p, include $x_{(j)}$ to x_S, that is, updating x_S by the concatenated variables $(x_S, x_{(i)})$, if $V_n^2(x_S, y)$ on the training set does not decrease. Stop otherwise.
    (c) Build models, for example, SVM-R in the TCGA ovarian case, using the selected variables on the training set.
    (d) Use the tuning set to select parameters for the model.
    (e) Use the model with chosen parameter to predict the response value for the testing set.
(3) Aggregate the predicted results for N replications
    - For classification task, one may use majority vote to obtain the final labels. The voting scores for the distinct labels can be used to evaluate the strength of being classified to each category for every observation.
    - For regression task, one may take the average of the N predicted values as the final prediction.

7. Discussion

The paper introduced a new variable selection procedure based on the property of DCOV and demonstrated the application through two examples. The SRBCTs data played a role of a toy example to show that the performance of the proposed method worked well in easy cases. The TCGA ovarian cancer data, however, were much more challenging to deal with because of the humongous number of variables and very limited sample size. The uncertainty of variable selection was discussed through gene selection results using random subsets of the data. The SVM-R option was used to withhold decision for subjects who were difficult to classify. An ensemble method of combining models built on random subsets of the data was implemented to assess the prediction performance. Although we had applied these tools (DCOV, SVM-R, and MCV) to biomedical data in the paper, we argue that they are quite portable across disciplines.

As shown in Table III, a small portion of the model building population obtained classified for $d = 1/5$. Is it worthwhile to attempt the classification in such cases? It depends on the application, for example, differential costs of two types of misclassification and subjective considerations including quality of life influenced by the treatment, therapy expense, and expected survival time.

Both the analysis of fivefold cross validation and MCV showed the uncertainty of gene selection results based on different subsets of the data. The large number of variables that appeared only in a small number of runs suggested noises in the data and the difficulty caused by limited training sample size in the high dimensional scenario. It could also suggest the conundrum that the 'true' model consists of a large number of variables with modest effects of which different subsets gives 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 ϵ 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.

The analysis of random permutation on the response served as both a validation of our results and a discussion of what one is likely to obtain without any true signal. If someone started with an entirely different dataset having the same proportions for the two classes with that in the original data but no real signal at all, as what one might obtain from scrambling, and went through every step, and finally obtained a subgroup of patients with large voting score values, the result was no better than just guessing that everyone was sensitive. This experiment was also a cautionary tale that if one had not held out validation sets, the analyst could be easily fooled by spurious correlated variables and perfect training accuracy. Our proposed MCV and analysis through the voting scores provided protection against finding fake signals.

Acknowledgements

J. K. and G. W. are supported by National Science Foundation (NSF) Grant DMS1308877 and National Institutes of Health (NIH) Grant EY09946. S. W is supported by NIH Grant 5R01HG007377-02.

References

1. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2008; **70**(5):849–911.
2. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 2008; **35**(6):2769–2794.
3. Li R, Zhong W, Zhu L. Feature screening via distance correlation. *Journal of the American Statistical Association* 2012; **107**(499):1129–1139.
4. Kong J, Klein BEK, Klein R, Lee KE, Wahba G. Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proceedings of the National Academy of Sciences* 2012; **109**(50):20352–20357.
5. Kosorok MR. Discussion of brownian distance covariance. *The Annals of Applied Statistics* 2009; **3**(4):1270–1278.
6. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 2001; **7**(6):673–679.
7. Lee Y, Lin Y, Wahba G. Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* 2004; **99**(465):67–81.
8. National Cancer Institute website. http://www.cancer.gov/cancertopics/types/ovarian [Accessed on 28 August 2013].
9. Network Cancer Genome Atlas Research and others. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011; **474**(7353):609–615.

10. Selvanayagam ZE, Cheung TH, Wei N, Vittal R, Lo KW, Yeo W, Kita T, Ravatn R, Chung TK, Wong YF, Chin KV. Prediction of chemotherapeutic response in ovarian cancer with DNA microarray expression profiling. *Cancer Genetics and Cytogenetics* 2004; **154**(1):63–66.
11. Rho SB, Dong SM, Kang S, Seo SS, Yoo CW, Lee DO, Woo JS, Park SY. Insulin-like growth factor-binding protein-5 (IGFBP-5) acts as a tumor suppressor by inhibiting angiogenesis. *Carcinogenesis* 2008; **29**(11):2106–2111.
12. Mehlmann LM, Saeki Y, Tanaka S, Brennan TJ, Evsikov AV, Pendola FL, Knowles BB, Eppig JJ, Jaffe LA. The GS-linked receptor GPR3 maintains meiotic arrest in mammalian oocytes. *Science* 2004; **306**(5703):1947–1950.
13. Ledent C, Demeestere I, Blum D, Petermans J, Hämäläinen T, Smits G, Vassart G. Premature ovarian aging in mice deficient for GPS3. *Proceedings of the National Academy of Sciences of the United States of America* 2005; **102**(25):8922–8926.
14. Basu S, Harfouche R, Soni S, Chimote G, Mashelkar RA, Sengupta S. Nanoparticle-mediated targeting of MAPK signaling predisposes tumor to chemotherapy. *Proceedings of the National Academy of Sciences* 2009; **106**(19):7957–7961.
15. Bast RC, Mills GB. Personalizing therapy for ovarian cancer. BRCAness and beyond. *Journal of Clinical Oncology* 2010; **28**(22):3545–3548.
16. Thiele S, Rauner M, Goettsch C, Rachner TD, Benad P, Fuessel S, Erdmann K, Hamann C, Baretton GB, Wirth MP, Jakob F, Hofbauer LC. Expression profile of WNT molecules in prostate cancer and its regulation by aminobisphosphonates. *Journal of Cellular Biochemistry* 2011; **112**(6):1593–1600.
17. Bitler BG, Nicodemus JP, Li H, Cai Q, Wu H, Hua X, Li T, Birrer MJ, Godwin AK, Cairns P, Zhang R. Wnt5a suppresses epithelial ovarian cancer by promoting cellular senescence. *Cancer Research* 2011; **71**(19):6184–6194.
18. Cole C, Lau S, Backen A, Clamp A, Rushton G, Dive C, Hodgkinson C, McVey R, Kitchener H, Jayson G C. Inhibition of FGFR2 and FGFR1 increases cisplatin sensitivity in ovarian cancer. *Cancer Biology & Therapy* 2010; **10**(5):495–504.
19. Yin J, Lu K, Lin J, Wu L, Hildebrandt MAT, Chang DW, Meyer L, Wu X, Liang D. Genetic variants in TGF-β pathway are associated with ovarian cancer risk. *PloS One* 2011; **6**(9):e25559.
20. Gu C. gss: general smoothing splines. *R package version 2.1-2*, 2007.
21. Wegkamp M, Yuan M. Support vector machines with a reject option. *Bernoulli* 2011; **17**(4):1368–1385.
22. Bartlett PL, Wegkamp MH. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research* 2008; **9**:1823–1840.