Two-stage variable selection for molecular prediction of disease

Hamed Firouzi
Department of EECS,
University of Michigan, Ann Arbor

Bala Rajaratnam
Department of Statistics,
Stanford University

Alfred O. Hero III
Department of EECS,
University of Michigan, Ann Arbor

***Abstract*—A two-stage predictor strategy is introduced in the context of high dimensional data (large p, small n). Here the focus application is a medical one: prediction of symptomatic infection given molecular expression levels in blood. The first stage of the two-stage predictor uses the previously introduced method of Predictive Correlation Screening (PCS) to select a subset of genes that are *important* in the prediction of symptom scores. Selected genes are used in the second stage to learn a predictor for the prediction of symptom scores. Under sampling budget constraints we derive the optimal sample allocation rules to the first and second stages of the two-stage predictor. Superiority of the proposed predictor relative to the well known method of LASSO is shown via experiment.**

I. Introduction

Consider the problem of under-determined multivariate linear regression in which training data $\{\mathbf{Y}_i, X_{i1}, ..., X_{ip}\}_{i=1}^n$ is given and a linear estimate of the q-dimensional response vector $\mathbf{Y}_i$, $1 \leq i \leq n < p$, is desired:

$$\mathbf{Y}_i = \mathbf{a}_1 X_{i1} + \ldots + \mathbf{a}_p X_{ip} + \epsilon_i, \ 1 \leq i \leq n, \qquad (1)$$

where X_{ij} is the ith sample of regressor variable (covariate) X_j, $\mathbf{Y}_i$ is a vector of response variables, $\mathbf{a}_j$ is the q-dimensional vector of regression coefficients corresponding to X_j, $1 \leq i \leq n, 1 \leq j \leq p$, and ϵ_i is the noise vector. In many applications the number of regressors p is significantly larger than the number of available samples n. Such applications arise in gene expression array analysis, text processing of internet documents, combinatorial chemistry, and others [1], [2]. Due to rank deficiency of the normal equations, overfitting errors and high computational cost, learning a linear predictor is difficult in such applications. Recently we introduced a method called Predictive Correlation Screening (PCS) that is specifically designed for selecting a subset of predictive regressors in cases where $p \gg n$ [3]. A generalization of hub screening method of [4], [5], PCS is a highly scalable technique for screening for connected variables in a correlation graph. However, unlike the correlation and partial correlation screening methods [4], [5], PCS screens for connectivity in a bipartite graph between the regressor variables $\{X_1, \ldots, X_p\}$ and the response variables $\{Y_1, \ldots, Y_q\}$. An edge exists in the bipartite graph between regressor variable j and response variable k if the thresholded min-norm regression coefficient matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]$ has a non-zero kj entry. The main idea behind PCS is that when the j-th column of this thresholded matrix is identically zero the j-th regressor variable is thrown out.

This research was supported in part by AFOSR grant FA9550-13-1-0043.

In this paper we provide additional results on the application of the two-stage prediction method to disease prediction, discussed briefly in [3]. The first stage of the two-stage predictor applies PCS to a few samples to select a subset of genes that are important for prediction of symptom scores. Selected genes are then used at the second stage of the two-stage predictor to learn a linear predictor using all of the available samples. The two-stage predictor is motivated by applications where the cost of samples increases with p. This is true, for example, in gene microarray experiments: a high throughput "full genome" gene chip with $p = 40,000$ gene probes can be significantly more costly than a smaller assay that tests fewer than $p = 15000$ gene probes (see Fig. 1). In this situation a cost-effective approach would be to use a two-stage procedure: first select a smaller number of variables on a few expensive high throughput samples and then construct the predictor on additional cheaper low throughput samples. The cheaper samples assay only those variables selected in the first stage.

The optimal sample allocation for the first and second stages of the two-stage predictor to minimize the Mean Squared Error (MSE) of the prediction under a sampling budget constraint is obtained. Specifically, we show that under the assumption of sparsity of active regressors (genes), if a total number of t samples are available, it is optimal to allocate only $\Theta(\log t)$ samples to the first stage.

The rest of the paper is organized as follows. Section II briefly describes the PCS method. In Sec. III we describe our two-stage predictor in the context of flu symptom prediction and we specify the optimal sample allocation rule. Finally, Sec. IV presents the experimental results and a comparison with LASSO.

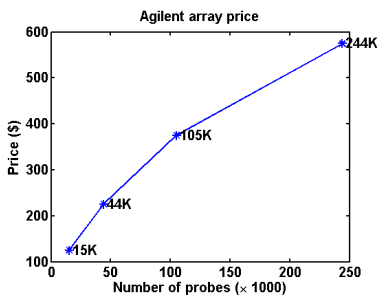


Fig. 1. Pricing per slide for Agilent Custom Micorarrays G2309F, G2513F, G4503A, G4502A (Feb 2013). The cost increases as a function of probeset size. Source: BMC Genomics and RNA Profiling Core. See also [3].

II. Variable Selection via Predictive Correlation Screening

Assume $\mathbf{X} = [X_1, \ldots, X_p]$ and $\mathbf{Y} = [Y_1, \ldots, Y_q]$ are random vectors of regressor and response variables, from which n observations are available. We represent the $n \times p$ and $n \times q$ data matrices as $\mathbb{X}$ and $\mathbb{Y}$, respectively.

The $p \times p$ sample covariance matrix $\mathbf{S}^x$ for data $\mathbb{X}$ is defined as:

$$\mathbf{S}^x = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_{(i)} - \overline{\mathbf{X}})^T (\mathbf{X}_{(i)} - \overline{\mathbf{X}}), \tag{2}$$

where $\mathbf{X}_{(i)}$ is the ith row of data matrix $\mathbb{X}$, and $\overline{\mathbf{X}}$ is the vector average of all n rows of $\mathbb{X}$.

Consider the $n \times (p+q)$ concatenated matrix $\mathbb{Z} = [\mathbb{X}, \mathbb{Y}]$. The sample cross covariance matrix $\mathbf{S}^{yx}$ is defined as the lower left $q \times p$ block of the $(p+q) \times (p+q)$ sample covariance matrix obtained by (2) using $\mathbb{Z}$ as the data matrix instead of $\mathbb{X}$. Assume that $p \gg n$. We define the ordinary least squares (OLS) estimator of $\mathbf{Y}$ given $\mathbf{X}$ as the min-norm solution of the underdetermined least squares regression problem

$$\min_{\mathbf{B}} \|\mathbb{Y}^T - \mathbf{B}\mathbb{X}^T\|_F^2, \tag{3}$$

where $\|\mathbf{A}\|_F$ represents the Frobenius norm of matrix $\mathbf{A}$. The min-norm solution to (3) is the $q \times p$ matrix of regression coefficients

$$\mathbf{B} = \mathbf{S}^{yx} (\mathbf{S}^x)^\dagger, \tag{4}$$

where $\mathbf{A}^\dagger$ denotes the Moore-Penrose pseudo-inverse of matrix $\mathbf{A}$. If the ith column of $\mathbf{B}$ is zero then the ith variable is not included in the OLS estimator. This is the main motivation for the proposed partial correlation screening procedure.

It can be shown that [3]:

$$\mathbf{B} = (\mathbf{H}^{xy})^T \mathbf{D}, \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix with non-zero diagonal entries and

$$\mathbf{H}^{xy} = (\mathbb{U}^x)^T \mathbb{U}^y. \tag{6}$$

in which $\mathbb{U}^x$ and $\mathbb{U}^y$ are $(n-1) \times p$ and $(n-1) \times q$ matrices whose columns lie on the unit sphere in $\mathbb{R}^{n-1}$. Therefore, screening for non-zero columns of $\mathbf{B}$ is equivalent to screening for non-zero rows $\mathbf{H}^{xy}$.

Now for a degree threshold $1 \le \delta \le q$ and a correlation threshold $0 \le \rho \le 1$, define the graph $\mathcal{G}_\rho(\mathbf{H}^{xy})$ as the undirected bipartite graph with parts labeled x and y, vertices $\{X_1, \cdots, X_p\}$ in part x and $\{Y_1, \cdots Y_q\}$ in part y (Fig. 2, see also [3]). For $1 \le i \le p$ and $1 \le j \le q$, there is an edge connecting X_i and Y_j if $|h_{ij}^{xy}| \ge \rho$, where h_{ij}^{xy} is the (i,j)th entry of $\mathbf{H}^{xy}$. Denote by d_i^x the degree of vertex X_i in $\mathcal{G}_\rho(\mathbf{H}^{xy})$. For each value $\delta \in \{1, \cdots, \max_{1 \le i \le p} d_i^x\}$, and each i, $1 \le i \le p$, denote by $\rho_i(\delta)$ the maximum value of the correlation threshold ρ for which $d_i^x \ge \delta$ in $\mathcal{G}_\rho(\mathbf{H}^{xy})$. $\rho_i(\delta)$ is in fact equal to the δth largest value $|h_{ij}^{xy}|, 1 \le j \le q$. $\rho_i(\delta)$ can be computed using Approximate Nearest Neighbors (ANN) type algorithms [6], [7]. Now for each i define the modified threshold $\rho_i^{\text{mod}}(\delta)$ as:

$$\rho_i^{\text{mod}}(\delta) = w_i \rho_i(\delta), \quad 1 \le i \le p, \tag{7}$$

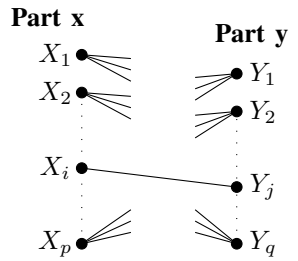


Fig. 2. Predictive correlation screening thresholds the matrix $\mathbf{H}^{xy}$ in (6) to find variables X_i that are most predictive of responses Y_j. This is equivalent to finding sparsity in a bipartite graph $\mathcal{G}_\rho(\mathbf{H}^{xy})$ with parts x and y which have p and q vertices, respectively. For $1 \le i \le p$ and $1 \le j \le q$, vertex X_i in part x is connected to vertex Y_j in part y if $|h_{ij}^{xy}| > \rho$.

where $w_i = D(i) / \sum_{j=1}^{p} D(j)$, in which $D(i)$ is the ith diagonal element of the diagonal matrix $\mathbf{D}$ in the representation (5).

Due to Propositions 1 and 2 in [3], representation (6) of $\mathbf{H}^{xy}$ allows us to assign approximate p-values to hubs in the graph $\mathcal{G}_\rho(\mathbf{H}^{xy})$ under the null hypothesis of sparse covariance matrices.

Assume $\mathbf{V}_1$ and $\mathbf{V}_2$ are two independent uniformly distributed vectors on the unit sphere in $\mathbb{R}^{n-1}$. The quantity $P_0(\rho)$ is defined as the probability that either $\|\mathbf{V}_1 - \mathbf{V}_2\| \le r$ or $\|\mathbf{V}_1 + \mathbf{V}_2| \le r$ for $r = \sqrt{2(1-\rho)}$. Also let:

$$\xi_{p,q,n,\delta,\rho} = p \binom{q}{\delta} P_0(\rho)^\delta. \tag{8}$$

Using definitions above, the approximate p-value assigned to vertex X_i for being a hub of degree at least δ in $\mathcal{G}_\rho(\mathbf{H}^{xy})$ is:

$$pv_\delta(i) \approx 1 - \exp(-\xi_{p,q,n,\delta,\rho_i^{\text{mod}}(\delta)}). \tag{9}$$

Finally selecting variables (genes) is performed by thresholding the p-values assigned to the genes at the desired significance level.

Next we introduce a bound on Familiy-Wise Error Rate (FWER) of PCS. Consider the following ground truth model:

$$\mathbf{Y} = \mathbf{a}_{i_1} X_{i_1} + \mathbf{a}_{i_2} X_{i_2} + \cdots + \mathbf{a}_{i_k} X_{i_k} + \epsilon, \tag{10}$$

in which ϵ is a noise vector that is statistically independent of $\mathbf{X}$. Assume that $\mathbf{X}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = [\sigma_{ij}]_{1 \le i,j \le p}$ which satisfies the following condition:

$$\sigma_{ij} = \sigma_{ji} = 0, \quad \forall\, i \in \{i_1, \cdots, i_k\}, j \notin \{i_1, \cdots, i_k\}. \tag{11}$$

Therefore, active (respectively inactive) variables are only correlated with the other active (respectively inactive) variables. Also, we assume that ϵ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma \mathbf{I}_{q \times q}$. The following theorem bounds the probability of selection error for PCS method with $\delta = 1$.

Theorem 1: Assume that k is known and we set PCS algorithm to select k variables with the smallest p-values. If the number of samples used for PCS is $\Theta(\log p)$, then PCS selects the true variables $X_{i_1}, \cdots, X_{i_k}$, with probability greater than $1 - q/p$.

Theorem 1 is proven in [3]. Note that the constant involved in the Θ above, depends on the coefficient matrix $\mathbf{A}$.

III. Two-stage symptom score predictor

Assume that there are a total of t samples $\{\mathbf{Y}_i, \mathbf{X}_i\}_{i=1}^{t}$ available. We propose the following two-stage predictor for prediction of symptom scores $\mathbf{Y}_i$ as a function of gene expression levels $\mathbf{X}_i$.

Stage 1. Perform PCS using $n \leq t$ samples to select a subset of genes that are important in prediction of symptom scores.

Stage 2. Use all t samples to obtain the OLS estimation of symptom scores as a function of selected gene expressions levels.

In stage 1 the predictor assays the whole genome on n samples to select a small subset of genes using PCS. To reduce the sampling cost, the second stage subsequently assays only the selected genes on all available samples to learn the predictor coefficients. We approximate the sampling cost at first and second stages with the quantities np and $(t-n)k$, respectively. Therefore, under a sampling budget μ, the following constraint must be satisfied:

$$np + (t-n)k \leq \mu. \tag{12}$$

The following theorem states the optimal sample allocation rule in order to minimize the asymptotic expected MSE of the two-stage predictor as a function of n, as $t \to \infty$. The assumptions on the data are similar to those of theorem 1.

Theorem 2: The sample allocation rule for MSE optimal two-stage predictor introduced above is:

$$n = \begin{cases} O(\log t), & c(p-k)\log t + kt \leq \mu \\ 0, & o.w. \end{cases} \tag{13}$$

Theorem 2, which is proven in [3], implies that under a generous budget limit μ, it is optimal to allocate only $n = O(\log t)$ samples to the first stage. However, if the budget is tight it is better to skip the first stage of the predictor. Figure 3 shows the allocation region as a function of sparsity coefficient $1 - k/p$.

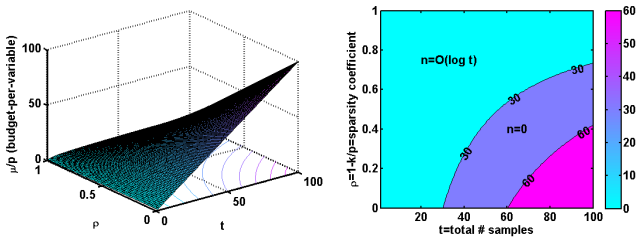


Fig. 3. Left: surface $\mu/p = \rho \log t + (1-\rho)t$. Right: contours indicating optimal allocation regions for $\mu/p = 30$ and $\mu/p = 60$, where $\rho = 1 - k/p$. See also [3].

IV. Experimental results

We illustrate the two-stage predictor on the Predictive Health and Disease dataset to predict the flu symptoms scores as a function of gene expression levels. The data was collected from 37 individuals enrolled in two challenge studies during which some subjects become symptomatically ill with the H3N2 flu virus [8]. For each subject, the gene expression levels and the symptoms were recorded at a large number of time points that include pre-inoculation and post-inoculation sample times. At each time point $p = 12023$ gene expression levels and 10 different symptom scores were measured. Each symptom score takes an integer value from 0 to 4, which measures the severity of that symptom at the corresponding time. The goal here is to learn a predictor that can accurately predict the symptom scores of a subject based on his measured gene expression levels. We applied our two-stage predictor to perform this task. The number of predictor variables (genes) selected in the first stage is restricted to 50. Since the symptom scores take integer values the second stage uses multinomial logistic regression instead of the OLS predictor. The performance is evaluated by leave-one-out cross validation. To do this, the data from all except one subject are used as training samples and the data from the remaining subject are used as the test samples. The final MSE is then computed as the average over the 38 different leave-one-out cross validation trials. In each of the experiments 18 out of the 37 subjects of the training set, are used in first stage and all of the 37 subjects are used in the second stage. Table I shows the result of this experiment for two-stage PCS predictor versus two-stage LASSO predictor [9]. We implemented LASSO using an active set type algorithm [10]. Note that, in this experiment, each symptom is considered as a one dimensional response and the two-stage algorithm is applied to each symptom separately. It is notable that the average symptom MSE of the two-stage PCS method performs better than that of LASSO. Furthermore, except for the first two symptoms, PCS performs better in predicting the symptom scores. The inferior performance of LASSO can be attributed to the fact that, unlike PCS, LASSO's stage 1 variable selection is not optimized for variable detection.

TABLE I. MSE of the proposed two-stage PCS predictor and the two-stage LASSO predictor used for symptom score prediction [3]. The data come from a challenge study experiment that collected gene expression and symptom data from human subjects [8].

Symptom	MSE: PCS	MSE: LASSO
Runny Nose	0.3537	0.3346
Stuffy Nose	0.5812	0.5145
Sneezing	0.3662	0.4946
Sore Throat	0.3026	0.3602
Earache	0.0761	0.0890
Malaise	0.3977	0.4840
Cough	0.2150	0.2793
Shortness of Breath	0.1074	0.1630
Headache	0.3299	0.3966
Myalgia	0.3060	0.3663
Average for all symptoms	**0.3036**	**0.3482**

Table II shows the 50 most frequent genes selected by PCS and LASSO. Note that for each of the 10 symptoms 38 different sets of 50 genes are selected by leaving each subject out. Therefore, the maximum possible frequency for selection of any gene is $10 \times 38 = 380$. It is observable that genes selected by PCS are generally more frequent than genes selected by LASSO. Hence, PCS tends to perform more consistent in selecting genes over different subjects and symptoms.

The 9 genes 'DMPK', 'CLIP3', 'GAPDHS', 'TFCP2L1', 'ANXA2P3', 'RGR', 'KLHL25', 'SRC' and 'C9orf45' are

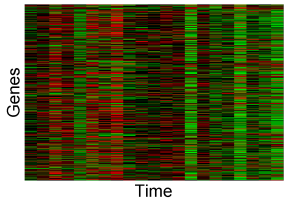
TABLE II. TOP 50 MOST FREQUENT GENES SELECTED BY PCS AND LASSO METHODS FOR THE TASK OF FLU SYMPTOM PREDICTION. WE CAN SEE THAT PCS TENDS TO BE MORE STABLE, WITH HIGHER GENE SELECTION FREQUENCIES, AS COMPARED TO LASSO. THERE ARE 9 GENES THAT ARE SELECTED FREQUENTLY BY BOTH METHODS WHICH ARE SHOWN WITH BOLD FONT IN THE TABLE.

Genes (PCS)	Frequency	Genes (LASSO)	Frequency
'TRAF2'	207	**'GAPDHS'**	144
'CTPS2'	192	'MARCO'	115
'NR2C2'	155	'SNORA64'	113
'ZNHIT2'	153	'728985_at'	103
'C9orf45'	152	'MAFK'	78
'HMGA1'	148	**'ANXA2P3'**	73
'GMPPB'	143	'LGALS2'	73
'RCAN1'	139	'ADAMTS5'	72
'CLIP3'	115	**'SRC'**	71
'KLHL25'	115	'C4orf18'	70
'KIAA0133'	115	**'KLHL25'**	68
'PAFAH1B3'	113	'C1orf115'	64
'TXNL4B'	113	'CAMK4'	64
'GAPDHS'	112	'CLEC10A'	62
'ZFP64'	112	'EPS8'	62
'PTDSS2'	112	'TRIB3'	62
'F10'	111	**'CLIP3'**	59
'PCYT2'	111	'KLC2'	59
'RGR'	111	**'C9orf45'**	59
'ELMO3'	111	'TOMM40'	58
'AKR7A3'	110	'CLIP4'	58
'THSD4'	110	'ZFHX3'	55
'SQSTM1'	110	'ALDH2'	54
'GLRX3'	107	'NOL6'	53
'SRC'	106	'BTBD2'	52
'XYLT2'	103	'SMG5'	51
'ASH2L'	102	'TRIM32'	48
'TADA3L'	94	'C3AR1'	46
'BRCA1'	92	'RUFY3'	45
'TPM3'	91	'BICD1'	45
'SYNJ2'	91	'CD300C'	44
'MYO1E'	88	'SGK3'	44
'HCFC1R1'	85	**'TFCP2L1'**	44
'NDST2'	83	'NR2F1'	44
'CPT2'	82	'GRB7'	43
'TFCP2L1'	82	'ANKRD7'	43
'ZNF446'	80	**'DMPK'**	42
'GSTT1'	79	'IFI27'	42
'TPPP3'	79	'PCTK3'	42
'ZNF576'	79	'SIGLEC1'	42
'DMPK'	78	'GJA9'	42
'CROCCL2'	76	'KIAA0556'	41
'ALAD'	76	'GPR20'	41
'INPP5B'	76	**'RGR'**	41
'MAPK7'	76	'USP46'	41
'CXCL13'	75	'LOC643332'	40
'ANXA2P3'	75	'ZP2'	40
'SOCS7'	75	'FZD1'	40
'CCDC88A'	75	'INSL3'	39
'SYN1'	75	'SERPINB5'	39

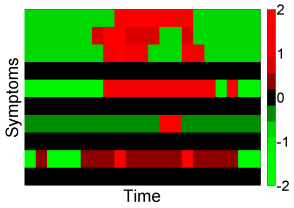
selected frequently by both methods. These genes are shown with bold font in table II. Also, Fig. IV shows the heat map of the 50 most frequent genes shown in table II over time, for a randomly selected subject.

V. CONCLUSION

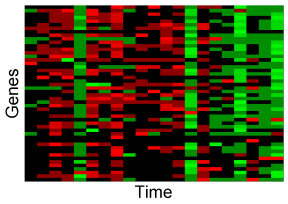
Using the previously proposed method of Predictive Correlation Screening (PCS) we developed a two-stage predictor of flu symptoms based on measured gene expression levels in the blood. The first stage incorporated the PCS method to select a subset of predictive genes. The second stage performed multinomial logistic regression on the selected genes to learn a predictor of the symptom scores. Experimental results established the advantages of the two-stage PCS predictor compared to the two-stage LASSO predictor.



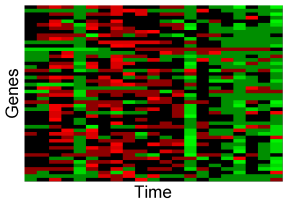
(a) Heat map of the complete genome over time for a randomly selected subject (subject #25).



(b) Heat map of the standardized symptom scores over time for a randomly selected subject (subject #25). The subject shows severe symptoms, shown in red.



(c) Heat map of the top 50 frequent genes selected by PCS over time for a randomly selected subject (subject #25).



(d) Heat map of the top 50 frequent genes selected by LASSO over time for a randomly selected subject (subject #25).

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] L. Wasserman and K. Roeder, "High dimensional variable selection," *Annals of statistics*, vol. 37, no. 5A, p. 2178, 2009.

[3] H. Firouzi, B. Rajaratnam, and A. Hero, "Predictive correlation screening: Application to two-stage predictor design in high dimension," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, 2013.

[4] A. Hero and B. Rajaratnam, "Large-scale correlation screening," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1540–1552, 2011.

[5] ——, "Hub discovery in partial correlation graphs," *Information Theory, IEEE Transactions on*, vol. 58, no. 9, pp. 6064–6078, 2012.

[6] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 1, pp. 117–128, 2011.

[7] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 891–923, 1998.

[8] Y. Huang, A. K. Zaas, A. Rao, N. Dobigeon, P. J. Woolf, T. Veldman, N. C. Øien, M. T. McClain, J. B. Varkey, B. Nicholson *et al.*, "Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection," *PLoS genetics*, vol. 7, no. 8, p. e1002234, 2011.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[10] J. Kim and H. Park, "Fast active-set-type algorithms for L_1-regularized linear regression," *Proc. AISTAT*, pp. 397–404, 2010.