# A three-stage framework for gene expression data analysis by $L_1$-norm support vector regression

Hyunsoo Kim

Department of Computer Science, University of Minnesota,
Twin Cities, 200 Union Street S.E., Minneapolis, MN 55455, USA
E-mail: hskim@cs.umn.edu

Jeff X. Zhou and Herbert C. Morse III

Laboratory of Immunopathology,
National Institute of Allergy and Infectious Diseases,
National Institutes of Health, 5640 Fishers Lane,
Rockville, MD 20852, USA
E-mail: jezhou@niaid.nih.gov E-mail: hmorse@niaid.nih.gov

Haesun Park*

Department of Computer Science, University of Minnesota,
Twin Cities, 200 Union Street S.E., Minneapolis, MN 55455, USA
E-mail: hpark@cs.umn.edu
*Corresponding author

**Abstract:** The identification of discriminative genes for categorical phenotypes in microarray gene expression data analysis has been extensively studied, especially for disease diagnosis. In recent biological experiments, continuous phenotypes have also been dealt with. For example, the extent of programmed cell death (apoptosis) can be measured by the level of caspase 3 enzyme. Thus, an effective gene selection method for continuous phenotypes is desirable. In this paper, we describe a three-stage framework for gene expression data analysis based on $L_1$-norm support vector regression ($L_1$-SVR). The first stage ranks genes by recursive multiple feature elimination based on $L_1$-SVR. In the second stage, the minimal genes are determined by a kernel regression, which yields the lowest ten-fold cross-validation error. In the last stage, the final non-linear regression model is built with the minimal genes and optimal parameters found by leave-one-out cross-validation. The experimental results show a significant improvement over the current state-of-the-art approach, i.e., the two-stage process, which consists of the gene selection based on $L_1$-SVR and the third stage of the proposed method.

**Keywords:** gene expression data analysis; apoptosis; support vector regression; recursive multiple feature elimination; gene selection; continuous phenotype.

**Reference** to this paper should be made as follows: Kim, H., Zhou, J.X., Morse III, H.C. and Park, H. (2005) ‘A three-stage framework for gene expression data analysis by $L_1$-norm support vector regression’, *Int. J. Bioinformatics Research and Applications*, Vol. 1, No. 1, pp.51–62.

**Biographical notes:** Hyunsoo Kim received his PhD degree in Computer Science from the University of Minnesota, Twin Cities, MN, in 2004 and his Doctoral thesis was Machine Learning and Bioinformatics. His research interests include bioinformatics, computational biology, and systems biology. He has published papers in several major machine learning and bioinformatics journals. He is currently working at the Computational Biology and Bioinformatics Laboratory, Department of Cell Biology, NYU School of Medicine in NYC, NY.

Jeff X. Zhou obtained his PhD degree in Biochemistry in 1996. His expertise includes functional genomics and biomedical data mining. He is currently a research fellow at the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD.

Herbert C. Morse, III, is Chief of the Laboratory of Immunopathology in the National Institute of Allergy and Infectious Diseases, NIH. His interests are in patterns of normal B cell differentiation and the mechanisms involved in development of B cell lineage lymphomas. He was instrumental in the development of a consensus nomenclature that relates hematopoietic neoplasms of mice to those in humans. He has established gene expression profiling as a powerful tool for distinguishing subsets of mouse B cell lymphomas.

Haesun Park has been Program Director (IPA) in the Division of Computing and Communications Foundations (CCF) at the Directorate for Computer and Information Science and Engineering, National Science Foundation, since November 2003 and has been on faculty at the Department of Computer Science and Engineering, University of Minnesota. She has been a professor in Twin Cities since 1987. Haesun Park received her BS degree in Mathematics (Summa Cum Laude) from Seoul National University, Seoul Korea, in 1981 with the university president's medal for the top graduate of the university. She received her MS and PhD degrees in Computer Science from Cornell University, Ithaca, NY in 1985 and 1987, respectively. Her current research interests include numerical linear algebra, pattern recognition, large-scale data analysis, bioinformatics, and information retrieval. Currently, she is on the editorial board of SIAM *Journal on Matrix Analysis and Applications by Society for Industrial and Applied Mathematics*, *Mathematics of Computation* by the American Mathematical Society and *BIT Numerical Mathematics*, and on the founding editorial board of *International Journal of Bioinformatics Research and Applications*.

## 1 Introduction

Gene expression microarray analysis is one of the fundamental technologies in genomics research. A microarray data set usually contains categorical phenotypes, such as different diseases, or continuous phenotypes, such as quantitative measurements of cellular function. An effective and reliable data analysis method is essential for the identification of genes responsible for the phenotypes. There have been many approaches for identifying discriminative genes. For example, correlation coefficients were used to rank genes for evaluating how well an individual gene contributes to discrimination between diseased and normal patients (Golub et al., 1999). Elements of the weight vector of a classifier can be used as feature ranking coefficients. By training a multivariate classifier such as a support vector machine, the most informative features can be chosen by ranking the components of the weight vector. The classical Fisher's linear discriminate is also a multivariate classifier that is optimised during training to handle multiple features simultaneously. But it fails when the number of features is larger than the number of data points due to singularity problems (Fukunaga, 1990). Recursive feature elimination with support vector machines was proposed for gene selection (Guyon et al., 2002). Unfortunately, it requires a very high computational cost since it is based on backward feature elimination and a wrapper method.

For continuous phenotypes, regression approaches for feature selection have emerged (Roth, 2004; Segal et al., 2003) where classification is considered as a special case of regression. Ridge regression tends to retain all elements of the weight vector since it uses an $L_2$ penalty. The least absolute shrinkage and selector operator (LASSO) method uses an $L_1$ penalty instead so that it zeroes out all but an optimal feature subset. The originally proposed solution of LASSO (Tibshirani, 1996) could not be applied to undersampled problems, where the number of features is greater than the number of samples. Osborne et al. (2000) developed a method based on convex programming to handle the undersampled case, but it requires high computational and memory costs. Least Angle RegreSsion (LARS), which is a less greedy version of forward stepwise regression, was developed to obtain all LASSO solutions in a highly efficient fashion (Efron et al., 2004). A more general feature selection algorithm based on $L_1$-norm support vector regression ($L_1$-SVR) has been developed (Bi et al., 2003). This method consists of two stages. The first stage is searching significant genes by $L_1$-SVR and the second stage is building a non-linear regression model by kernel $L_1$-SVR. Although this two-state process is the current state-of-the-art, it is still not a fully optimised procedure. We replace the first stage by the recursive multiple feature elimination of $L_1$-SVR that provides consistent genes for different training/test splits. An additional stage is also introduced in order to identify a smaller number of more discriminative genes.

In this paper, a three-stage framework for building regression model for continuous phenotypes is established. During the first stage, a set of candidate genes is chosen and ranked by a linear feature selection algorithm. During the second stage, the minimal genes are determined by a kernel regression, which yields the lowest ten-fold cross-validation error. In the last stage, the final non-linear regression model is built with the minimal genes and the optimal parameters found by leave-one-out cross-validation. This three-stage method has shown significantly better performance than the two-stage method in terms of the leave-one-out cross-validation error and the number of chosen genes when we identified genes relevant to programmed cell death (apoptosis).

## 2 Methods

### *2.1 The first stage: feature ranking*

Although there are many linear feature selection algorithms available, for simplicity we describe three feature selection methods for regression problems, which are based on

(1) Pearson correlation coefficient

(2) ν-SVR linear programming (Bi et al., 2003)

(3) recursive multiple feature elimination (RMFE) with ν-SVR linear programming.

Method (3) is our feature selection method for continuous phenotypes in the first stage. Method (1) is introduced only for comparison and explanation of the reason why it is not always a good idea to use the Pearson correlation coefficient in order to select discriminative minimal genes. In the state-of-the-art two-stage method, method (2) is directly used for the final gene selection. On the other hand, in our proposed three-stage framework, method (3) is used to select and rank genes and the final minimal genes are determined in the second stage.

Here, we describe these three linear feature selection algorithms one by one. The feature selection based on the Pearson correlation coefficient ranks genes in order of magnitude of the coefficient between a feature and an observed continuous phenotype. By the magnitude without the sign, the negative correlated genes can be selected as well. To choose a subset of the ranked genes, a threshold value is necessary. The threshold value is determined by the mean value of the corresponding magnitude of the Pearson correlation coefficients for artificially appended three random gauge variables.

The gene selection based on ν-SVR linear programming is now reviewed. For the training data ($a_i; y_i$) for $1 \leq i \leq m$, where $\mathbf{a}_i \in \mathrm{R}^{n\times 1}$, the linear regression function is given by

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \tag{1}$$

where **w** and $b$ are obtained by solving a linear $\varepsilon$-insensitive $\nu$-SVR optimisation problem (Bi et al., 2003)

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi^*,\varepsilon} & \sum_j^n |w_j| + C\frac{1}{m}\sum_i^m(\xi_i + \xi_i^*) + C\nu\varepsilon, \\ s.t. \quad & y_i - \mathbf{w}^T\mathbf{a}_i - b \leq \varepsilon + \xi_i, \quad i = 1,\ldots,m \\ & \mathbf{w}^T\mathbf{a}_i + b - y_i \leq \varepsilon + \xi_i^*, \quad i = 1,\ldots,m \\ & \xi_i, \xi_i^*, \nu \geq 0, \quad i = 1,\ldots,m, \end{aligned} \tag{2}$$

where $\xi_i$ and $\xi_i^*$ for $i = 1, \ldots, m$ are slack variables, $C$ is a parameter that controls the trade-off between margin and training errors, and $\varepsilon$ is a tube parameter. Since the tube parameter $\varepsilon$ is difficult to determine, $\upsilon$-SVR was developed to automatically adjust the tube size, $\varepsilon$, by using a parameter $\upsilon \in (0, 1]$ (Schölkopf et al., 2000; Smola et al., 1999). This $L_1$-norm SVR is applied for the gene selection problem since it can be used to zero out multiple irrelevant genes at the same time. A similar idea has been applied in LASSO (Efron et al., 2004; Osborne et al., 2000). Equation (2) can be reformulated by introducing $u_j \geq 0$ and $v_j \geq 0$ and denoting $w_j = u_j - v_j$ and $|w_j| = u_j + v_j$ as

$$\begin{aligned} \min_{\mathbf{u},\mathbf{v},b,\xi,\xi^*,\varepsilon} & \sum_j^n (u_j + v_j) + C\frac{1}{m}\sum_i^m(\xi_i + \xi_i^*) + C\nu\varepsilon, \\ s.t. \quad & y_i - (\mathbf{u} - \mathbf{v})^T\mathbf{a}_i - b \leq \varepsilon + \xi_i, \quad i = 1,\ldots,m \\ & (\mathbf{u} - \mathbf{v})^T\mathbf{a}_i + b - y_i \leq \varepsilon + \xi_i^*, \quad i = 1,\ldots,m \\ & \xi_i, \xi_i^*, \nu \geq 0, \quad i = 1,\ldots,m, \\ & u_j, v_j \geq 0, \quad j = 1,\ldots,n. \end{aligned} \tag{3}$$

After obtaining the solution $\mathbf{w} = \mathbf{u} - \mathbf{v}$, the insignificant genes whose corresponding component of $\mathbf{w}$ is lower than a threshold are eliminated. The threshold value is determined by a mean value of the corresponding components of $\mathbf{w}$ for artificially appended three random gauge variables. The parameters $C$ and $v$ are determined by ten-fold cross-validation within a reasonably large range, i.e., $C \in [2^{-5};\ 2^{15}]$ and $v \in [0.1,\ 0.7]$. The ranking of chosen genes is determined by the magnitude of the corresponding elements of $\mathbf{w}$.

Finally, the recursive multiple feature elimination with $v$-SVR linear programming is introduced by applying recursive feature elimination to $v$-SVR linear programming. The threshold value for each iteration is determined by the mean value of the corresponding components of $\mathbf{w}$ for artificially appended three random gauge variables. This method is not only accurate but also computationally efficient since it can remove several irrelevant features at the same time for the identification of significant minimal genes for continuous phenotypes. It also produces consistent genes for different training/test splits.

## *2.2 The second stage: feature selection by a kernel regression*

The second stage is finding minimal genes by a kernel regression, which yields the lowest ten-fold cross-validation error. In the two-stage method (Bi et al., 2003), there is no stage like this second stage since the minimal genes are already determined in its first stage.

Even though there are many possible kernel regression methods, a kernel $v$-SVR linear programming is described here. For the training data $(\mathbf{a}_i, y_i)$ for $1 \le i \le m$, where $\mathbf{a}_i \in \mathrm{R}^{n\times 1}$, the non-linear regression function is given by

$$f(\mathbf{x}) = \sum_{i}^{m} \alpha_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) + b, \tag{4}$$

where $\phi(\cdot)$ is a non-linear function that maps a vector in the input space to a vector in a feature space. The inner product is computed by a radial basis kernel function

$$\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2), \tag{5}$$

where $\gamma$ is a kernel parameter. The $\alpha$ and $b$ in equation (4) are obtained by solving a non-linear $\varepsilon$-insensitive $\upsilon$-SVR optimisation problem (Bi et al., 2003)

$$\begin{aligned}
&\min_{\mathbf{u},\mathbf{v},b,\xi,\xi^*,\varepsilon} \sum_{j}^{n} (u_j + v_j) + C\frac{1}{m}\sum_{i}^{m} (\xi_i + \xi_i^*) + Cv\varepsilon, \\
&s.t. \quad y_i - \sum_{j}^{m} (u_j - v_j)\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) - b \le \varepsilon + \xi_i, \quad i = 1,\dots,m \\
&\qquad \sum_{j}^{m} (u_j - v_j)\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) + b - y_i \le \varepsilon + \xi_i^*, \quad i = 1,\dots,m \\
&\qquad u_j, v_j, \xi_i, \xi_i^*, v \ge 0, \quad i, j = 1,\dots,m.
\end{aligned} \tag{6}$$

In this formulation, $\alpha_j = u_j - v_j$ is used, where $u_j, v_j \geq 0$. The solution has either $u_j$ or $v_j$ equal to 0, depending on the sign of $\alpha_j$, so $|\alpha| = u_j + v_j$. Therefore, the optimal solution is given by $\alpha = \mathbf{u} - \mathbf{v}$. For this linear programming formulation, the kernel parameter $\gamma$ should be selected as well. The parameters $C$, $\gamma$, and $v$ are determined by ten-fold cross-validation within a reasonably large range, i.e., $C \in [2^{-5}; 2^{15}]$, $\gamma \in [2^{-40}; 2^{5}]$, and $v \in [0.1, 0.7]$.

In order to find the minimal genes, ten-fold cross-validation tests are repeated by adding a gene one by one from the highest rank (most relevant) gene to the lowest rank (most irrelevant) gene. As a consequence, a subset of ranked genes found in the first stage is identified, which produces the lowest root mean squared error (RMSE) of a ten-fold cross-validation. Detailed formulation of the RMSE is described in results/evaluation subsection.

### 2.3 The third stage: the final non-linear regression model

With the chosen genes in the second stage, a non-linear regression model can be built. As the second stage, although there are many possible kernel regression methods, a kernel $v$-SVR linear programming is also used in the third stage. Through leave-one-out cross-validation for parameter selection of $C$, $\gamma$, and $v$, the final non-linear regression model is built by training all samples with the optimal parameters.

## 3 Results

### 3.1 Microarray experiment and apoptosis measurement

A variety of mouse B-cell lymphoma samples were subjected to oligonucleotide microarray analysis as described previously (Zhang et al., 2004). Briefly, tumour cell RNA from 134 samples were prepared and hybridised to microarray chips comprising ~6,700 mouse gene targets represented by 70-mer oligonucleotides. The raw dataset is available at the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/, accession # GSE1908). Twenty-nine out of the 134 samples were also used for the measurement of caspase 3 levels, an indicator of apoptosis. Using immunohistochemistry staining with a specific antibody to mouse caspase 3, the average number of positive cells per high-power field was recorded and used as the level of caspase 3 in the tumour tissue.

### 3.2 Evaluation

We built a matrix that contains 29 samples (rows, data points) used for the measurement of caspase 3 levels and 119 genes (columns, features) reportedly related to apoptosis, and tried to find a minimal set of genes that well represents the level of caspase 3 for 29 samples. For this given problem, the corresponding final regression model obtained by each method described in the 'Methods' section is evaluated by the following procedure. For each training/test split during $n'$-fold cross-validation, the root mean squared error (RMSE) is computed for test data points, i.e.,

$$\text{RMSE} = \sqrt{\frac{1}{\text{m}'}\sum_{j}^{m'}(z_j^{*}-z_j)^2} \tag{7}$$

where $m'$ is the number of test data points in the training/test split, $z_j^{*}$ is the regression model predicted value, and $z_j$ is the corresponding value of the response variable (phenotype) obtained by experiments. Then, the $n'$-fold cross-validated RMSE value is obtained by the mean value of the $n'$ RMSE values obtained from $n'$-fold cross-validation. When performing leave-one-out cross-validation, the RMSE value for each fold is just an absolute value of the difference between $z_1^{*}$ and $z_1$ since $m' = 1$. The leave-one-out cross-validation error ($\text{LOOCV}_{\text{err}}$) is the mean value of the distances $d_i$, for $1 \le i \le m$, i.e.,

$$\text{LOOCV}_{\text{err}} = \frac{1}{m}\sum_{i}^{m} d_i = \frac{1}{m}\sum_{i}^{m}|\, y_i^{*} - y_i \,|, \tag{8}$$

where $m$ is the number of observations, $y_i^{*}$ is the regression model predicted value, and $y_i$ is the $i$th element of the response variable obtained by experiments for the $i$th data point left out during leave-one-out cross-validation.

One two-stage method and three three-stage methods are compared in terms of the accuracy of the final non-linear regression model for a continuous phenotype. The components of each multi-stage gene selection method are described in Table 1. Table 2 shows the number of selected genes at the first and second stages and the leave-one-out cross-validation rate for each method. The two-stage method based on $L_1$-SVR is the same method as Bi et al. (2003), which uses $\upsilon$-SVR linear programming in the first stage and kernel $\upsilon$-SVR linear programming in the second stage, which corresponds to the third stage in the other three-stage methods. Although it is possible to design many combinations of feature selection methods and kernel regression, three three-stage methods are introduced in this paper, i.e., three-stage/PC, three-stage/$L_1$-SVR, and three-stage/RMFE-$L_1$-SVR, in order to show the significance of the three-stage approach. The difference among these methods lies only in the choice of feature selection method in the first stage. In the second and third stages, these methods use kernel $\upsilon$-SVR linear programming for finding a minimal set of genes and building a final non-linear regression model. The three-stage/PC, three-stage/$L_1$-SVR, and three-stage/RMFE-$L_1$-SVR methods use feature selection methods based on the Pearson correlation coefficient, $\upsilon$-SVR linear programming, and RMFE with $\upsilon$-SVR linear programming, respectively, in the first stage.

**Table 1** Components of multi-stage gene selection methods

| *Method* | *1st stage* | *2nd stage* | *3rd stage* |
|---|---|---|---|
| Two-stage/$\text{L}_1$-SVR* | $\nu$–SVR | – | kernel $\nu$-SVR |
| Three-stage/PC | Pearson correlation | kernel $\nu$-SVR | kernel $\nu$-SVR |
| Three-stage/$\text{L}_1$-SVR | $\nu$-SVR | kernel $\nu$-SVR | kernel $\nu$-SVR |
| Three-stage/RMFE-$\text{L}_1$-SVR | RMFE-$\nu$-SVR | kernel $\nu$-SVR | kernel $\nu$-SVR |

*The two-stage method based on $L_1$-SVR is the same method as Bi et al. (2003).

**Table 2** Comparison of gene selection methods. The number of selected genes at the first and second stages and the leave-one-out cross-validation error are presented

| | *No. of genes* | *No. of genes* | *LOOCVerr* |
|---|---|---|---|
| *Method* | *(1st stage)* | *(2nd stage)* | *($3^{rd}$ stage)* |
| Two-stage/L$_1$-SVR* | 49 | – | 10.51 |
| Three-stage/PC | 69 | 9 | 7.51 |
| Three-stage/L$_1$-SVR | 49 | 8 | 5.45 |
| Three-stage/RMFE-L$_1$-SVR | 29 | 13 | 3.66 |

*The two-stage method based on $L_1$-SVR is the same method as Bi et al. (2003).

### *3.3 Comparison between a two-stage method and three-stage methods*

The major difference between the two-stage method and the three-stage methods is that the two-stage method does not have an additional optimisation stage to obtain minimal discriminative genes like the second stage in the three-stage framework. Figures 1 and 2 show the procedure of the second stage to find a minimal set of genes among chosen genes in the first stage.

In the left panel of Figure 1, there is a global minimum at 55 genes. However, at nine genes, the second minimum is selected for biological analysis since the difference of RMSEs between the global minimum and the second minimum is not large while the number of genes (nine) is much less than 55. In the right panel of Figure 1, there is a sharp global minimum at eight genes. These results show the important role of the second stage. A larger number of genes does not always produce better accuracy than a smaller number of genes the regression model. In Figure 2, there is a global minimum at 13 genes among the total of 29 genes chosen by RMSE with $L_1$-SVR in the first stage. The 13 genes are a much smaller number of genes compared with 49 genes selected by the two-stage method. Moreover, in Table 2, the LOOCV$_{\mathrm{err}}$ of three-stage/RMFE-$L_1$-SVR is only 3.66, while that of two-stage method is 10.51.

**Figure 1** The second stage of the three-stage method based on Pearson correlation coefficient (left) and $L_1$-SVR (right). The ten-fold cross-validated RMSE is presented for each subset of genes ranked in the first stage

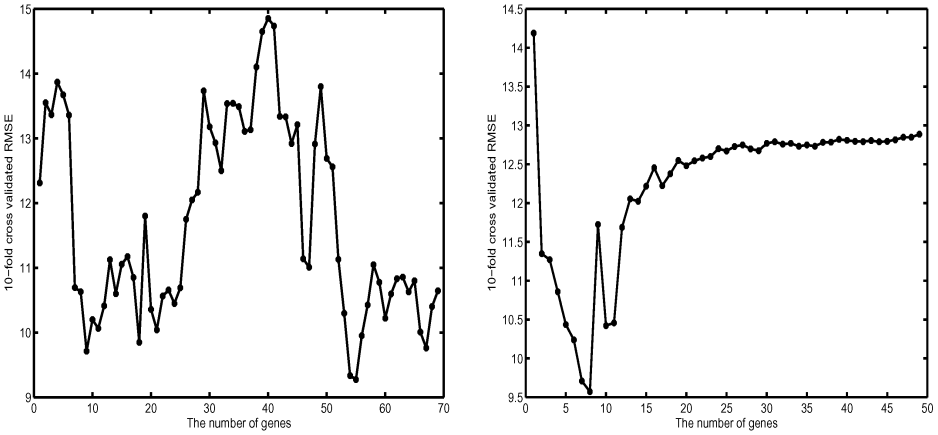
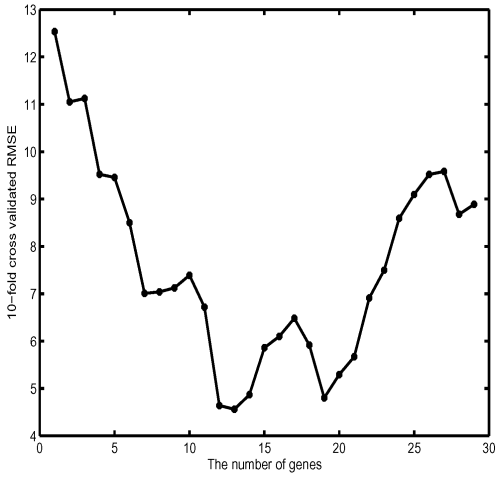


**Figure 2** The second stage of the three-stage method based on RMFE-$L_1$-SVR. The ten-fold cross-validated RMSE is presented for each subset of genes ranked in the first stage



We now describe the reason why the three-stage method shows better performance than the current state-of-the-art two-stage method. The two-stage method selects minimal discriminative genes by linear $L_1$-SVR, which are used to build a non-linear model by kernel $L_1$-SVR. However, this two-stage method does not take into account the full performance of the kernel $L_1$-SVR, even though a smaller number of genes could be sufficient when we utilise kernel $L_1$-SVR instead of linear $L_1$-SVR in order to obtain similar performance results. In the proposed method, the second stage optimises the gene selection by using kernel $L_1$-SVR, which is the major reason that the proposed three-stage framework identifies a much smaller number of genes. We can also explain this in terms of the LOOCV$_{\text{err}}$. In the two-stage method, the kernel $L_1$-SVR is performed on the genes found by linear $L_1$-SVR without any optimisation stage (i.e., the second stage of the three-stage framework). On the other hand, the final stage of the three-stage framework takes advantage of the chosen genes that were already optimised by kernel $L_1$-SVR in the second stage. The data in the right panel of Figures 1 and 2 show that we are able to obtain smaller RMSE values when we ignored some low ranked genes in the first stage.

### *3.4 Comparison among three-stage methods*

Using RMFE with $\upsilon$-SVR linear programming, we found a smaller number of genes (29) than the other feature selection methods in the first stage. In the second stage, the final number of selected genes by three-stage/RMFE-$L_1$-SVR (13) is larger than that of three-stage/$L_1$-SVR (eight). However, Table 2 shows that the 13 genes are more appropriate genes for obtaining an accurate regression model for its lower LOOCV$_{\text{err}}$. This means that the RMFE has the advantage of selecting a small number of genes with good feature ranking. This is an expected result since the RMFE tends to find consistent genes for different training/test splits (Kim, 2004).

On the other hand, the Pearson correlation coefficient may not produce good feature ranking even though a higher magnitude of Pearson correlation coefficient means a higher correlation between a feature and an observed continuous phenotype. This may be related to a collinearity problem in the regression. Collinearity is a situation where there is a close to near perfect linear relationship among some or all of the independent variables in a regression model. In practical terms, this means there is some degree of redundancy or overlap among variables. It is sometimes described as multicollinearity, near collinearity, or ill conditioning. If there are highly correlated genes between chosen genes in order of magnitude of the Pearson correlation coefficients, choosing the correlated genes may not be as good a choice as choosing independent genes in spite of their small magnitude of Pearson correlation coefficient.

### *3.5 Biological analysis*

Table 3 displays the 13 genes that were selected by three-stage/RMFE-$L_1$-SVR. Among the 13 genes, Ripk3, Tnfrsf12, Faf1, Bmp7, Traf1, Tnip1, and Litaf were also found by three-stage/$L_1$-SVR. Ripk3, Tnfrsf12, Faf1, and Rela were chosen by three-stage/PC as well. Notably, Ripk3, Tnfrsf12, and Faf1 were identified by all three-stage methods tested. Ripk3, the mouse receptor interacting protein 3, is a component of the tumour necrosis factor receptor-1-mediated apoptosis cascade (Pazdernik et al., 1999). Although caspase 3 is a further downstream effector of apoptosis and there is no information available to show any interaction between Ripk3 and caspase 3, our results suggest that Ripk3 holds one of the strongest correlation relationships with the level of caspase 3. This finding, if biologically confirmable, would provide an important insight for the functions and mechanisms of Ripk3 protein. Tnfrsf12 (also called Ws1, Apo3, TRAMP, LARD, TR3, and DR3) is one of six death-domain containing TNFR family members (the others are TNFR1, CD95/FAS, DR4, DR5, and DR6). Members of the mammalian tumour necrosis factor receptor (TNFR) family are cell-surface proteins that interact with a corresponding TNF-related ligand family. Tnfrsf12 is the one most closely related to TNFR1 (Kitson et al., 1996), and mediates a variety of developmental events including the regulation of cell proliferation, differentiation, and apoptosis (Wang et al., 2001). A Fas-associated protein factor, Faf1, potentiates Fas-mediated apoptosis, although it cannot initiate Fas-induced apoptosis (Chu et al., 1995). All selected genes were biologically confirmed that they are highly relevant to apoptosis.

**Table 3** Genes relevant to programmed cell death (apoptosis). Total 13 genes that were selected by three-stage/RMFE-$L_1$-SVR are displayed from the most relevant gene (Ripk3)

| *Gene* | *Description* |
|---|---|
| Ripk3* | Receptor-interacting serine-threonine kinase 3 |
| Tnfrsf12* | Tumour necrosis factor receptor superfamily, member 12 |
| Faf1* | Fas-associated factor 1 |
| Bmp7 | Bone morphogenetic protein 7 |
| Tnfsf4 | Tumour necrosis factor (ligand) superfamily, member 4 |
| Traf1 | Tnf receptor-associated factor 1 |
| Tnip1 | TNFAIP3 interacting protein 1 |
| Rela | v-rel reticuloendotheliosis viral oncogene homologue A (avian) |
| Litaf | LPS-induced TN factor |
| Tnfsf10 | Tumour necrosis factor (ligand) superfamily, member 10 |
| Tnfrsf6 | Tumour necrosis factor receptor superfamily, member 6 |
| Tnfaip2 | Tumour necrosis factor, alpha-induced protein 2 |
| Casp6 | Caspase 6, apoptosis-related cysteine protease |

*Top three genes were chosen by all three-stage methods.

## 4 Conclusion

In the present work, we found that the two-stage approach is not the best method for building an accurate non-linear regression model of continuous phenotypes for gene expression data analysis. For this problem, the three-stage framework has been proposed and evaluated by leave-one-out cross-validation. The three-stage methods that were tested in this paper showed much lower $\text{LOOCV}_{\text{err}}$ than the two-stage method. The proposed recursive multiple feature elimination based on $L_1$-SVR in the first stage played an important role in obtaining the final good regression model due to its selectivity and consistency. It was also successfully shown that the second stage of the three-stage methods is an essential step for increasing the accuracy of the non-linear regression model. In summary, a novel three-stage framework has been introduced for gene expression data analysis based on continuous phenotypes.

Our future work will include the development of three-stage methods that optimise the feature selection algorithms and/or other non-linear regression methods. Any future improvement of feature selection algorithms or non-linear regression methods can be adapted to the three-stage framework.

## Acknowledgments

We thank Thomas McCarty and Xingpei Hao for performing the microarray experiment and making the caspase 3 measurement. This material is based upon work supported in part by the National Science Foundation Grants CCR-0204109 and ACI-0305543.

## References

Bi, J., Bennett, K.P., Embrechts, M., Breneman, C.M. and Song, M. (2003) 'Dimensionality reduction via sparse support vector machines', *Journal of Machine Learning Research*, Vol. 3, pp.1229–1243.

Chu, K., Niu, X. and Williams, L.T. (1995) 'A Fas-associated protein factor, FAF1, potentiates Fas-mediated apoptosis', *Proc. Natl Acad. Sci.*, USA, Vol. 92, pp.11894–11898.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) 'Least angle regression', *Annals of Statistics*, Vol. 32, No. 2, pp.407–499.

Fukunaga, K. (1990) *Introduction to Statistical Pattern Recognition,* 2nd ed., Academic Press, Boston.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, Vol. 286, pp.531–537.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) 'Gene selection for cancer classification using support vector machines', *Machine Learning*, Vol. 46, pp.389–422.

Kim, H. (2004) *Machine Learning and Bioinformatics*, PhD Thesis, University of Minnesota, Twin Cities, MN, USA.

Kitson, J., Raven, T., Jiang, Y-P., Goeddel, D.V., Giles, K.M., Pun, K-T., Grinham, C.J., Brown, R. and Farrow, S.N. (1996) 'A death-domain-containing receptor that mediates apoptosis', *Nature*, Vol. 384, pp.372–375.

Osborne, M., Presnell, B. and Turlach, B. (2000) 'On the LASSO and its dual', *J. Comput. Graph. Statist.*, Vol. 9, pp.319–337.

Pazdernik, N.J., Donner, D.B., Goebl, M.G. and Harrington, M.A. (1999) 'Mouse receptor interacting protein 3 does not contain a caspase-recruiting or a death domain but induces apoptosis and activates NF-kappaB', *Cel. Biol.*, Vol. 19, pp.6500–6508.

Roth, V. (2004) 'The generalized LASSO', *IEEE Trans. Neural Networks*, Vol. 15, No. 1, pp.16–28.

Schölkopf, B., Smola, A., Williamson, R.C. and Bartlett, P.L. (2000) 'New support vector algorithms', *Neural Computation*, Vol. 12, pp.1207–1245.

Segal, M.R., Dahlquist, K.D. and Conklin, B.R. (2003) 'Regression approaches for microarray data analysis', *J. Comp. Biol.*, Vol. 10, No. 6, pp.961–980.

Smola, A., Schölkopf, B. and Rätsch, G. (1999) 'Linear programs for automatic accuracy control in regression', *Processing ICANN'99, Int. Conf. on Artificial Neural Networks*, Springer, Berlin.

Tibshirani, R. (1996) 'Regression shrinkage and selection via LASSO', *J. Roy. Statist. Soc. B*, Vol. 58, pp.267–288.

Wang, E.C., Thern, A., Denzel, A., Kitson, J., Farrow, S.N. and Owen, M.J. (2001) 'DR3 regulates negative selection during thymocyte development', *Molecular and Cellular Biology*, Vol. 21, No. 10, pp.3451–3461.

Zhang, J.Q., Okumura, C., McCarty, T., Shin, M.S., Mukhopadhyay, P., Hori, M., Torrey, T.A., Naghashfar, Z., Zhou, J.X., Lee, C.H., Roopenian, D.C., Morse III, H.C. and Davidson, W.F. (2004) 'Evidence for selective transformation of autoreactive immature plasma cells in mice deficient in Fasl.', *J. Exp. Med.*, Vol. 200, No. 11, pp.1467–1478.