Maximum likelihood estimation of GEVD: Applications in Bioinformatics

Minta Thomas, Anneleen Daemen, and Bart De Moor

Abstract—We propose a method, maximum likelihood estimation of generalized eigenvalue decomposition (MLGEVD) that employs a well known technique relying on the generalization of singular value decomposition (SVD). The main aim of the work is to show the tight equivalence between MLGEVD and generalized ridge regression. This relationship reveals an important mathematical property of GEVD in which the second argument act as prior information in the model. Thus we show that MLGEVD allows the incorporation of external knowledge about the quantities of interest into the estimation problem. We illustrate the importance of prior knowledge in clinical decision making/identifying differentially expressed genes with case studies for which microarray data sets with corresponding clinical/literature information are available. On all these three case studies, MLGEVD outperformed GEVD on prediction in terms of test area under the ROC curve (test AUC). MLGEVD results in significantly improved diagnosis, prognosis and prediction of therapy response.

Index Terms—eigenvalue decomposition, generalized eigenvalue decomposition, maximum likelihood generalized eigenvalue decomposition, generalized singular value decomposition

1 Introduction

Microarray technology is a significant tool in gene expression analysis and cancer diagnosis. These technologies are typically used for class discovery [1], [2] and prediction [3], [4]. Clinical data such as age, gender and medical history offer a proper care and treatment of patients for most diseases. The effective management of these data always lead to better clinical prognosis. Microarray data analysis are in general much more difficult and expensive to collect while clinical parameters are routinely measured by clinicians. A vital study on the prediction of breast cancer outcome has suggested that despite the emergence of these high-throughput technologies, clinical markers and profiles have similar power for prognosis [5]. We previously analyzed the influence of clinical and microarray data on prediction and observed that proper integration of these two data sets improved the prediction accuracy [6].

Biomarker discovery and prognosis prediction are essential for improved personalized treatment of cancer. Principal component analysis (PCA) and PCA-based approaches for example were used for the identification of differentially expressed genes (DEG) in pulmonary adenocarcinoma [7] and E coli [8]. Troyanskaya and colleagues developed nonparametric methods to identify DEG in microarray data [9]. Besides well-known statistical tests such as the chi-square test [10], Chun and Colleagues proposed a new test, the 'half Student's t-test', specifically for detecting DEG in heterogeneous diseases [11]. Singular Value Decomposition (SVD) and generalized SVD (GSVD) have been shown to have great potential within bioinformatics for extracting common information from data sets such as genomics and proteomics data [12], [13]. Maximum likelihood principal component analysis (MLPCA) is an error-in-variables modeling method in that it accounts for measurement errors in the estimation of model parameters. Wentzell *et al.* [14] generalized PCA method to MLPCA [15], [16]. The tight equivalence between MLPCA and total least squares (TLS) is explored in [17]. Finally, several studies have developed methods for integrating literature and microarray data sets for identifying disease related genes [18], [19].

In this paper we propose a method which incorporates external knowledge of interest in the analysis of microarray and clinical data sets. The main aim of the paper is to show the tight equivalence of maximum likelihood estimation of generalized eigenvalue decomposition (MLGEVD) with generalized ridge regression. This reveals an important mathematical property of GEVD/GSVD in which the second input acts as the prior information in the model. We incorporate microarray/literature information as prior information in the model to improve the performance in clinical decision making.

- *M. Thomas and B. De Moor are with KU Leuven Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics / iMinds Future Health Department, Kasteelpark Arenberg 10, B-3001, Leuven,Belgium. E-mail: minta.thomas@esat.kuleuven.be*
- *A. Daemen is with Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, CA, USA.*

2 Data Sets and Methods

2.1 Case Study - I

Breast cancer is one of the most extensively studied cancer types for which many microarray data sets are publicly available. Among them, we selected three cases for which also clinical information was available [20], [21], [22].

2.1.1 Microarray Data

The microarray data were obtained with the Affymetrix technology and preprocessed with MAS5.0, the GeneChip Microarray Analysis Suite 5.0 software (Affymetrix). However, as probe selection for the Affymetrix gene chips relied on earlier genome and transcriptome annotation that are significantly different from current knowledge, an updated array annotation was used for the conversion of probes to Entrez Gene IDs, lowering the number of false positives [23]. Finally, the low signal-to-noise [24] ratio of microarray data was taken into account by unsupervised exclusion of genes with low variation (variance less than the 20th percentile), retaining 4997, 5997 and 12633 most varying genes for the first, second and third microarray data respectively.

2.1.2 Clinical Data

The first case study of 129 patients contained information on 17 available clinical variables. Five variables were excluded [20]: two redundant variables that were least informative based on univariate analysis in those variable pairs with a correlation coefficient exceeding 0.7, and three variables with too many missing values. After exclusion of patients with missing clinical information, this data set (represented in Table 1) consisted of 110 patients, 85 in whom disease did not recur whilst in 25 patients disease recurred [25].

The second case study, in which response to treatment was studied, entailed 12 variables for 133 patients [21]. Patient and variable exclusion as described above resulted in 129 patients and 8 variables, shown in Table 2. Of the 129 remaining patients, 33 showed complete response to treatment while 96 patients were characterized by residual disease.

In the last case study, relapse was studied in 187 patients [22]. After preprocessing, this data set retained information on 5 variables for 177 patients with detailed information shown in Table 3. In 112 patients, no relapse occurred while 65 patients had a relapse.

Clinical data contains three different types of variables: continuous (C), ordinal (O) and nominal (N). Normalization is required to make these variables comparable to each other. Rank order, min-max and square root transformations were applied to the ordinal, continuous and nominal variables, respectively.

TABLE 1
Clinical variables data set I (breast cancer - recurrence)

Variable	Type	Range
1. Age(years)	C	31-88
2. Ethnicity	N	0, 1, 2
3. ER status	N	0, 1
4. PR status	N	0, 1
5. Radiation treatment	N	0, 1
6. Chemotherapy	N	0, 1
7. Hormonal therapy	N	0, 1
8. Nodal status (N)	O	0-2
9. Metastasis (M)	N	0, 1
10. Tumor stage	O	1-4
11. Tumor size	C	0.3-7.5
12. Tumor grade	O	1-3

ER, estrogen receptor; PR, progesterone receptor.

TABLE 2
Clinical variables data set II (breast cancer - treatment response)

Variable	Type	Range
1. Age(years)	C	28-79
2. Ethnicity	O	0, 1, 2, 3, 4
3. Pretreatment tumor stage	O	1-4
4. Nodal status (N)	O	0-3
5. Nuclear grade	O	1-3
6. ER status	N	0, 1
7. PR status	N	0, 1
8. HER2 status	N	0, 1

ER, estrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2.

TABLE 3
Clinical variables data set III (breast cancer - relapse)

Variable	Type	Range
1. Age(years)	C	32-86
2. Tumor size(cm)	C	0.2-8.2
3. Nodal status (N)	N	0, 1
4. ER status	N	0, 1
5. Tamoxifen treatment	N	0, 1

ER, estrogen receptor.

2.2 Case Study - II

2.2.1 Microarray Data

The colon cancer data set investigated in this manuscript was taken from the Bioinformatics Research Group Repository [26]. It contains 62 samples, among them 40 colon tumor samples and 22 normal colon samples, with 1,988 genes and 12 controls. Data were standardized to a mean of 0 and standard deviation of 1.

2.2.2 Literature Information

We used a well defined cancer vocabulary with 2406 terms from NCI Dictionary of Cancer Terms[27]. Pubmed abstracts with the terms in the vocabulary were extracted using Perl, version 5.10.1 for windows. We defined literature information as a matrix with rows corresponding to genes and columns to cancer related terms. Each entry in the matrix corresponds to the number of Pubmed abstracts in which the gene and term co-occur. We chose to retrieve entries containing the official gene name, abbreviations or aliases in the corresponding field, following the same strategy as used by Gevaert et al [19]. Finally, the cosine similarity measure was used to obtain gene-to-gene distances between 0 and 1, derived from the literature information.

2.3 Methods

2.3.1 Principal Component Analysis:

Principal component analysis (PCA) [28] of $m \times n$ matrix A equals

$$A = TP,$$

where score T is $m \times n$ matrix and coefficients P is orthogonal $n \times n$ matrix.

2.3.2 Maximum Likelihood Estimation of PCA:

In PCA, if $rank(A) = r$, then the best approximation of A is

$$\hat{A} = LF$$

where L (score matrix) is m$\times r$ matrix, $F \approx \mathcal{N}(0, Ir)$, and Ir is the identity matrix with r-dimension.

The maximum likelihood estimation of PCA [17] is as follows:

$$\min_F vec^T(A - LF)Q^{-1}vec(A - LF),$$

with Q the error covariance matrix of $vec(A - LF)$, where $vec(A - LF)$ stands for the vectorized form $(A - LF)$, i.e., a vector constructed by stacking the consecutive columns of $(A - LF)$ in one vector. Let the error covariance matrix $Q = blkdiag(Q1, \ldots, Q_n)$, where $Q_i \in \mathbb{R}^{m\times m}$. The solution of the problem can be computed efficiently in this case as follows:

$F_i = (L^T Q_i^{-1} L)^{-1} L^T Q_i^{-1} A_i$ where F_i is the i^{th} column of F.

2.3.3 Maximum Likelihood Estimation of Generalized Eigenvalue Decomposition:

The Generalized Singular Value Decomposition (GSVD)[29] of $m \times n$ matrix A and $p \times n$ matrix B is

$$A = U\Sigma_A X^T \quad (1)$$

$$B = V\Sigma_B X^T \quad (2)$$

where U, V are orthogonal matrices, the columns of X are generalized singular vectors and Σ_A, Σ_B are diagonal matrices with entries corresponding to the generalized singular values of matrices A and B respectively.

If $B^T B$ is invertible, the GEVD of $A^T A$ and $B^T B$ can be obtained from equations (1) and (2) as follows:

$$A^T A(X^T)^{-1} = B^T B(X^T)^{-1}\Lambda. \quad (3)$$

where Λ is a diagonal matrix with entries $\Lambda_{ii} = (\frac{\Sigma_{A_{ii}}}{\Sigma_{B\,ii}})^2$, $i = 1 \ldots$, n and the columns of $(X^T)^{-1}$ are generalized eigenvectors (GEVs).

The equation (3) can now be rewritten as a standard eigenvalue problem:

$$(B^T B)^{-1/2} A^T A(B^T B)^{-1/2} U = U\Lambda. \quad (4)$$

where $U = (B^T B)^{1/2}(X^T)^{-1}$, which is in the form of EVD of the matrix $(B^T B)^{-1/2} A^T A(B^T B)^{-1/2}$.

SVD of $A(B^T B)^{-1/2}$ is given by

$$A(B^T B)^{-1/2} = V\Lambda U^T \quad (5)$$

The matrix $(B^T B)^{-1/2}$ is defined [30] as follows: Let eigenvalue decomposition (EVD) of $(B^T B) = T\Sigma T^T$, where columns of T are eigenvectors and Σ is a diagonal matrix. $(B^T B)^{1/2} = T\Sigma^{1/2}T^T$ and $(B^T B)^{-1/2} = TQT^T$, where Q is a diagonal matrix with diagonal entries $Q_{ii} = (\Sigma_{ii})^{-1/2}$, $i = 1, \ldots, N$.

Let D be the projections of $A(B^T B)^{-1/2}$ on eigenvectors U, which is equivalent to the projections of A on GEVs $(X^T)^{-1}$ (See Equation 4). Thus we have,

$$D = A(B^T B)^{-1/2} U = A(X^T)^{-1}, \quad (6)$$

with $U^T U = I$ and $(X)^{-1}(B^T B)(X^T)^{-1} = I$, where I is the identity matrix.

In MLGEVD, we have to estimate GEVs $(X^T)^{-1}$ which maximize the variance of projected variables $A(X^T)^{-1}$ under the constraint that $(X)^{-1}B^T B(X^T)^{-1} = I$.

Maximum Likelihood Estimation of GEVD can be formulated as follows:

$\tilde{r}_{MLGEVD} = min_{\tilde{r}} \Sigma_{i=1}^n (D_i - \tilde{e}_i)^T Q_{\epsilon x}^{-1} (D_i - \tilde{e}_i) + \Sigma_{i=1}^n \tilde{r}_i^T (B^T B)\tilde{r}_i$ s. t. $\tilde{e}_i = A\tilde{r}_i$, where $Q_{\epsilon x}$ is the error covariance matrix and r_i is ith column of $(X^T)^{-1}$. In

the maximum likelihood interpretation of GEVD, we have to minimize the reconstruction error which can be formulated as follows:

Solution: Define the Lagrangian

$\mathcal{L} = \Sigma_{i=1}^{n}(e_i - \tilde{e_i})^T Q_{\epsilon x}^{-1}(e_i - \tilde{e_i}) + \Sigma_{i=1}^{n} r_i^T B^T B r_i + \Sigma_{i=1}^{n}\alpha_i(\tilde{e_i} - Ar_i)$.

with the optimality conditions,

$\frac{\partial \mathcal{L}}{\partial \tilde{e_i}} = -2Q_{\epsilon x}^{-1}(e_i - \tilde{e_i}) + \alpha_i^T = 0.$

$\frac{\partial \mathcal{L}}{\partial r_i} = -A^T\alpha^T + 2(B^TB)r_i = 0.$

$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \tilde{e_i} - Ar_i = 0,\ i = 1\ldots, n.$

Eliminations of $\tilde{e_i}, \tilde{r_i}$, and α_i yields an equation in the form of

$$\tilde{r_i} = (A^TQ_{\epsilon x}^{-1}A + B^TB)^{-1}A^TQ_{\epsilon x}^{-1}D_i, i = 1\ldots, n. \quad (7)$$

Thus, the Maximum Likelihood estimation of GEVD is

$$\tilde{r_{i}}_{MLGEVD} = (A^TQ_{\epsilon x}^{-1}A + B^TB)^{-1}A^TQ_{\epsilon x}^{-1}D_i, i = 1\ldots, n. \quad (8)$$

which is in the form of generalized ridge regression. An algorithm for MLGEVD is given below:

Algorithm: MLGEVD

1) Input data matrices $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{q\times n}$.
2) Initial approximation: we have $D = \tilde{A}(B^TB)^{-1}U$ (see Equation 6). Compute a rank p truncated SVD approximation $D^{(0)} = \tilde{A}(B^TB)^{-1}U_{a_p}$ of D. And error covariance matrix $W_i \in \mathbb{R}^{m\times m}$ (i=1,. ., n).
3) k=0;
4) repeat
5) Compute the solution of (8) $r_i = (A^TW_i^{-1}A + B^TB)^{-1}A^TW_i^{-1}D_i$, $i = 1\ldots$,n, are the GEVs (columns of $R^{(k)}$).
6) Compute $D^{(k+1)}$ using Equation (6).
7) k=k+1
8) Until$||D^{(k)} - D^{(k-1)}||_F/||D^{(k)}||_F \leq \varepsilon$, where ε is the convergence parameter.
9) Output: $\tilde{D} = D^{(k)}, \tilde{R} = R^{(k)}$

The Matlab implementation of the algorithm is given in Appendix A. For numerical reasons [31] however, the explicit formation of matrix product of the form A^TA, B^TB should be avoided. Hence we will work with SVD and GSVD, instead of EVD and GEVD, for which explicit commands are provided in Matlab.

3 Classification Models

The constrained optimization problem for an Least Squares-Support Vector Machine (LS-SVM) [32], [33] for classification has the following form:

$$\min_{w,b,e}(\frac{1}{2}w^Tw + \gamma\frac{1}{2}\Sigma_{k=1}^{N}e_k^2)$$

subject to:

$$y_k[w^T\varphi(x_k) + b] = 1 - e_k, \qquad k = 1,\ldots,N$$

where $\varphi(.)$: $\mathbb{R}^d \to \mathbb{R}^{d_h}$ is a nonlinear function which maps the d-dimensional input vector x from the input space to the d_h-dimensional feature space, possibly infinite. In the dual space the solution is given by

$$\begin{bmatrix} 0 & y^T \\ y & \Omega + \frac{I}{\gamma} \end{bmatrix}\begin{bmatrix} b \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 1_v \end{bmatrix}$$

with $y = [y_1,\ldots,y_N]^T$, $1_N = [1,\ldots,1]^T$, $e = [e_1,\ldots,e_N]^T$, $\beta = [\beta_1,\ldots,\beta_N]^T$ and $\Omega_{i,j} = y_iy_jK(x_i,x_j)$ where $K(x_i,x_j)$ is the kernel function. The functions that are most frequently employed in classification problems are the linear kernel ${x_i}^T x_j$, the polynomial kernel $({x_i}^T x_j + t)^d$ with the intercept constant $t \in \mathbb{R}^+$ and degree $d \in \mathbb{N}$, and the radial basis function (RBF) $\exp(-||x_i - x_j||_2^2)/\sigma^2$.

The classifier in the dual space takes the form

$$y(x) = \text{sign}[\sum_{k=1}^{N}\beta_k y_k K(x,x_k) + b]$$

where β_k are Lagrange multipliers.

4 Results

For all case studies, 2/3rd of the samples were randomly assigned to the training set and 1/3rd to the test set. The split was performed stratified to outcome, to ensure that the relative proportion of outcomes sampled in both training and test set was similar to the original proportion in the full data set. To allow proper comparison of all methods, the randomizations are the same for all numerical experiments per case study.

4.1 Classification of Breast Cancer Patients

MLGEVD approximates GEVs by incorporating microarray data as prior information into the model.

In the first step, GEVs are obtained from training - both clinical and microarray - data sets by MLGEVD and GEVD, separately. Projected variables corresponding to training and test clinical data sets are obtained by projecting these data onto the direction of GEVs. Next, the LS-SVM classifier is trained using the projected-training data, followed by classification of the projected-test data. Classification accuracy is given in terms of test set Area Under the ROC Curve (AUC) and F-score [34]. In this section all the steps are implemented using Matlab R2012b and LS-SVMlab v1.8 toolbox [35] with the default parameter settings.

Table 4 summarizes the average test results over 30 iterations of the MLGEVD and GEVD methods for all three breast cancer data sets. For the sake of comparison, LS-SVM classifiers with the linear, RBF

TABLE 4
Comparison of LS-SVM classification performance on MLGEVD and GEVD for three breast cancer studies: average test AUC (std) and average F-score (std) over 30 iterations

	kernel function	linear	RBF	polynomial
Case I				
test AUC	MLGEVD	**0.80**(0.09)	**0.79**(0.01)	**0.77**(0.07)
	GEVD	0.77(0.08)	0.74(0.09)	0.63(0.02)
	p-value	0.03	0.01	2.72E-10
test F-score	MLGEVD	**0.64**(0.01)	**0.57**(0.02)	**0.51** (0.13)
	GEVD	0.55(0.01)	0.49 (0.01)	0.26(0.03)
	p-value	0.17	0.23	0.03
Case II				
test AUC	MLGEVD	**0.80**(0.05)	0.75(0.09)	**0.70**(0.10)
	GEVD	0.79(0.06)	**0.78**(0.08)	0.61(0.06)
	p-value	0.01	0.04	0.01
test F-score	MLGEVD	**0.60**(0.03)	**0.46** (0.06)	0.47(0.01)
	GEVD	0.56(0.02)	0.51(0.07)	**0.50**(0.05)
	p-value	0.02	0.06	0.11
Case III				
test AUC	MLGEVD	**0.67**(0.07)	**0.60**(0.08)	**0.60**(0.04)
	GEVD	0.66(0.08)	0.57(0.05)	0.54(0.06)
	p-value	0.02	0.26	0.86
test F-score	MLGEVD	**0.44**(0.04)	**0.29**(0.46)	**0.28**(0.03)
	GEVD	0.32(0.08)	0.23 (0.12)	0.24 (0.08)
	p-value	0.06	0.18	0.04

p-value: two-sided sign test ; RBF: radial basis function

and polynomial kernel function were applied to the MLGEVD and GEVD models. The LS-SVM classifier with the linear kernel function resulted in the best test AUC for both GEVD and MLGEVD. With this kernel function, MLGEVD significantly outperformed GEVD for all three breast cancer case studies. The F-scores for these classifiers shown in Table 4 are indicative of precision and recall, with scores ranging from 0 to 1.

High-throughput data, such as microarray data are in general much more difficult and expensive to collect while clinical parameters are routinely measured by clinicians. The main advantage of MLGEVD/GEVD for prediction is that high-throughput data are only used as prior information during model development, while the final clinical decision only depends on clinical variables.

4.2 Identification of differentially expressed genes in colon cancer

In the colon cancer case study, GEVs are obtained from training data using MLGEVD and GEVD. Then the training data is divided into two groups: normal and cancerous samples. Each of these sets of data are projected onto the direction of GEVs, resulting into two sets of scores Z^1 and Z^2. Let $g_i = Z_i^1 - Z_i^2$ be the difference in score for gene i between normal and cancerous samples.

Each gene can graphically be represented as a point in the k-dimensional space (with k the number of GEVs selected for projection). Each gene i with similar expression levels in both sets of scores has approximately the same scores $Z_i^1 \approx Z_i^2$ and form a cloud of points around the origin. Differentially expressed genes have significantly different scores and are located away from the origin. To identify the outliers in this k-dimensional space, the Mahalanobis distance is calculated for each gene $MD_i^2 = (g_i - c)\Sigma^{-1}(g_i - c)^T$, with c the multivariate arithmetic mean and Σ^{-1} the inverse of the covariance matrix of the differences in scores [8]. Genes with the largest Mahalanobis distances are defined as the most differentially expressed genes.

Table 5 shows the top 50 differentially expressed genes obtained with MLGEVD and GEVD. Among these genes, relevance for colon cancer has been shown for 44 and 38 genes with MLGEVD and GEVD, respectively. An LS-SVM model with an RBF kernel for the prediction of tumour vs. non-tumour samples was built. In Table 6, we compared the prediction performances of full data sets and GEVD with MLGEVD. We show that these 50 genes obtained from MLGEVD can be used to form an colon cancer signature, to distinguish normal from colon cancer subjects and can be used to classify good and poor prognostic tumors (see Table 6).

Several genes selected by this approach are known to be involved in, and important for, colon cancer. The ribosome, the essential cellular organelle for protein synthesis in all cells, consists of ribosomal RNAs (rRNAs) and ribosomal proteins (RPs). Ribosomal protein L41 (RPL41) is a microtubule-associated protein essential for functional spindles and for the integrity of centrosome. Abnormal mitosis and a disrupted centrosome associated with RPL41 downregulation may be related to malignant transformation [45]. In our analysis, RPL41 ranked as one of the differentially expressed gene. Studies in [37] and [46] already reported on the importance of RPL41 in colon cancer.

Ectopic expression of tumor rejection antigen 1 (Tra1) was detected in the ulcerative colitis (UC) affected colonic mucosa [47]. Tra1 is reported as a differentially expressed gene in colon cancer [37]. Calcyclin

TABLE 5
The 50 top ranked genes for relevance in colon cancer diagnosis identified by MLGEVD and GEVD.

MLGEVD		GEVD	
Gene Symbol	Ref	Gene Symbol	Ref
IGLC1	[36]	EIF4A1	[37]
RPLP1	[37]	N2b5HR	[37]
TMSB4X	[37]	ITIH1	[37]
FTL	[37]	IGHG3	[38]
IGKC	[39]	IGLC1	[36]
TCTP	[37]	RPLP2	[37]
EIF4A1	[37]	MYL6	[37]
S100A6	[37]	TCTP	[37]
RPS	[37]	RPL41	[37]
SELENBP1	–	ACTB	[37]
RPS29	[37]	RPS9	[37]
RPSA	[37]	HSP90B1	[37]
RPL30	[37]	RPL37A	[37]
RPL37A	[37]	BBC1	[37]
RPL32	[37]	RPLP1	[37]
IGHG3	[38]	YBX1	[37]
YBX1	[37]	UBB	[37]
CPSF1	–	RPSA	[37]
RPL37A	–	GAPDH	[37]
LGALS3	–	SRF	[37]
RPS18	[37]	IGF2	–
UBB	[37]	RPL37A	–
PFN1	[37]	RPL1	–
RPS6	[37]	HSPB1	–
GAPDH	[37]	RPS29	[37]
HSP90B1	[37]	RPS18	[37]
RPS24	[37]	FTL	[37]
BBC1	[37]	IGKC	[39]
RPS28	[37]	RPL30	[37]
RPL38	[40]	RPS	[37]
MUC2	[36]	RPS28	[37]
IGHG3	[41]	HLA-B	[37]
ITIH1	[37]	S100A6	[37]
RPLP2	[37]	EEF1A2	[37]
RPL41	[37]	IFI27	–
ALDOA	[37]	EEF1B2	–
ACTB	[37]	JUND	–
RPS9	[37]	MT1G	–
OAZ	[37]	SELENBP1	–
HSP90AB1	[42]	RPS8	–
RPS24	[40]	ARNT	–
B2M	[37]	TSPAN8	–
MAMDC2	[37]	OAZ	[37]
SRF	[37]	RPS11	[37]
DESMIN	[43]	RPS24	[37]
LYZ	[44]	MUC2	[36]
N2b5HR	[37]	TPM2	[43]
MYL6	[37]	RPS19	[40]
FCGRT	–	RPL32	[37]
RPL37	–	LYZ	[44]

TABLE 6
LS-SVM model for prediction of tumour and non-tumour samples of colon cancer. Average classification performance (std) on test sets for all genes and the subsets of 50 genes selected by GEVD and MLGEVD.

Genes selected by	kernel function	LS-SVM	p-value[a]
full data set	RBF	0.821(0.147)	0.019
GEVD	RBF	0.841(0.087)	0.072
MLGEVD	RBF	**0.895(0.060)**	

[a] two-sided sign test for the comparison of full data sets and GEVD with MLGEVD.

binding protein (CacyBP) was a promising candidate biomarker for colorectal cancer (CRC) metastasis and also sheds light on the underlying molecular mechanism by which CacyBP promotes CRC metastasis [48]. 60S acidic ribosomal protein P1 was reported as a top-ranked gene in colon cancer [37], [49].

Translationally controlled tumor protein (TCTP) is a highly conserved and ubiquitously expressed protein in all eukaryotes highlighting its important functions in the cell. Previous studies revealed that TCTP is implicated in many biological processes, including cell growth, tumor reversion, and induction of pluripotent stem cells. In human colon cancer, the level of TCTP mRNA was detected in three human colon carcinoma cell lines (SNU-C2A, SNU-C4, and SNU-C5) [50]. Ornithine decarboxylase (OAZ1) catalyzes the conversion of ornithine to putrescine in the first and apparently rate-limiting step in polyamine biosynthesis. The ornithine decarboxylase antizymes play a role in the regulation of polyamine synthesis by binding to and inhibiting ornithine decarboxylase. OAZ was reported as a top ranked gene in colon cancer [37], [40].

Alterations in the distribution and/or adhesiveness of laminin receptors in colon cancer cell lines were suggested to be associated with increased tumorigenicity [51]. A study of cultured colon cancer cells suggests that laminin may play an important role in hematogeneous metastasis by mediating tethering and spreading of colon cancer cells under blood flow [52]. In general, the markers are involved in cell signaling, adhesion and communication, immune response, heat shock, and DNA repair [37].

In short, out of 50 genes identified as differentially expressed in colon cancer, the majority of these genes are reported as top ranked genes in various studies. In addition, the improved prediction performance obtained with the selected genes clearly indicate that these genes distinguish cancerous from non-cancerous samples.

The experiments show that the convergence time will depend on the dimensionality of the problem. These times were generally reasonable, less than a minute for the breast cancer cases and a few minutes for colon cancer on Windows XP operating system with 2.40 GHz processor.

5 Discussion

Generalized Eigenvalue Decomposition (GEVD) is a method for comparing two data sets, in which generalized eigenvectors capture common information between these two data sets. In this study, MLGEVD

shows that GEVD is strongly related to generalized regression, if the second matrix is of full rank.

MLGEVD approach has been applied to four case studies for which gene expression with corresponding clinical/literature information were available. Microarray and clinical parameters were gathered from patients with breast cancer. Literature information from Pubmed were collected for colon cancer. The main aim of the work is to interpret GEVD in the framework of maximum likelihood estimation. To validate the merit of MLGEVD over GEVD/GSVD, models were built for classifying patients. In this study we performed MLGEVD, on clinical data sets with microarray as prior information and on microarray data with literature as prior information. In both cases, the model parameters (generalized eigenvectors) are obtained with MLGEVD. Subsequently clinical parameters/microarray were projected onto the generalized eigenvectors, referred to as the projected clinical space/gene space. Similarly generalized eigenvectors are obtained for GEVD. The advantages of MLGEVD over GEVD were that MLGEVD reduce the projection error. Finally LS-SVM was built on these clinical projected spaces or the sub sets of genes identified by MLGEVD, and validated on test samples for prediction.

For all the data sets on breast cancer, binary outcomes (cancerous vs non-cancerous) could more accurately be predicted with MLGEVD than with GEVD. In addition, incorporation of external knowledge into the analysis of microarray improves the identification of disease related genes. In general, MLGEVD can be applied to any two data sets that satisfy the data properties of GEVD, that is, one of the data sets is invertible, and the number of rows or columns of both data sets are the same.

The proposed model is very cost effective. The high throughput technologies which are difficult and expensive to collect are used only for the model development. The clinical parameters which are routinely measured by clinicians are used for prediction. In real data examples, we have shown how to incorporate external knowledge, extracted from microarray data/literature information into medical diagnosis. The proposed method provides a general way to incorporate such ever-increasing amounts of prior knowledge into the analysis and to further improve the predictive performance.

6 Conclusion

In this paper, we developed an algorithm for MLGEVD and compared its performance with GEVD. We show that prediction performances improve with the incorporation of prior data. Both GEVD and MLGEVD can use the high-throughput data, which are difficult and expensive to collect, as prior information. The MLGEVD obtained the best approximation of model parameters, GEVs, which minimize the projection error. In our analysis, we have shown that MLGEVD can be used as an alternative to GEVD with more accurate classification/prediction. Overall, the proposed model can be used as a noise reduction technique in medical prognosis. In the near future, we will investigate the applicability of MLGEVD to more than two matrices and interpret these matrix results in a Bayesian context.

Acknowledgments

BDM is full professor at the Katholieke Universiteit Leuven, Belgium. Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, START 1, OT 09/052 Biomarker, several PhD/postdoc & fellow grants; Industrial Research fund (IOF): IOF/HB/13/027 Logic Insulin, IOF: HB/12/022 Endometriosis; Flemish Government: FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits), research community MLDM; IWT: PhD Grants; TBM-Logic Insulin, TBM Haplotyping, TBM Rectal Cancer, TBM IETA; Hercules Stichting: Hercules III PacBio RS; iMinds: SBO 2013; Art&D Instance;IMEC: phd grant;VLK van der Schueren: rectal cancer;VSC Tier 1: exome sequencing; Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate); COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network. The scientific responsibility is assumed by its authors.

References

[1] V. Roth and T. Lange, "Bayesian class discovery in microarray data," *IEEE Transactions on Biomedical Engineering*, vol. 51, 2004.

[2] P. Qiu and S. K. Plevritis, "Simultaneous class discovery and classification of microarray data using spectral analysis," *Journal of Computational Biology*, vol. 16, pp. 935–944, 2009.

[3] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19, pp. 1484–1491, 2003.

[4] L. Conde, A. Mateos, J. Herrero, and D. J, "Improved class prediction in dna microarray gene expression data by unsupervised reduction of the dimensionality followed by supervised learning with a perceptron," *Journal of VLSI Signal Processing*, vol. 35, pp. 245–253, 2003.

[5] C. Eden, P andnRitz, C. Rose, M. Ferno, and C. Peterson, ""good old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers," *Eur J Can.*, vol. 40, pp. 1837–41, 2004.

[6] A. Daemen, D. Timmerman, T. Van den Bosch, C. Bottomley, E. Kirk, C. Van Holsbeke, L. Valentin, T. Bourne, and B. De Moor, "Improved modeling of clinical data with kernel methods," *Artificial Intelligence in Medicine*, vol. 54, pp. 103–114, 2012.

[7] W. Harriet, K. Eeva, K. S. Jouni, K. Antti, H. Jaakko, A. Sisko, and K. Sakari, "Identification of differentially expressed genes in pulmonary adenocarcinoma by using cdna array," *Oncogene*, vol. 21, pp. 5804–5813, 2002.

[8] S. Jonnalagadda and R. Srinivasan, "A pca based approach for gene target selection to improve industrial strains," *17th European Symposium on Computer Aided Process Engineering ESCAPE17*, 2007.

[9] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman, "Nonparametric methods for identifying differentially expressed genes in microarray data," *Bioinformatics*, vol. 18, pp. 1454–1461, 2002.

[10] L. Shao, Q. Huang, M. Luo, and M. Lai, "Detection of the differentially expressed gene igf-binding protein-related protein-1 and analysis of its relationship to fasting glucose in chinese colorectal cancer patients," *Endocrine-Related Cancer*, vol. 11, pp. 141–148, 2004.

[11] C.-L. Hsu and W.-C. Lee, "Detecting differentially expressed genes in heterogeneous disease using half student's t-test," *Int. I. Epidemiol.*, vol. 10, pp. 1–8, 2010.

[12] S. Sedeh, R, M. Bathe, and B. K-J, "The subspace iteration method in protein normal mode analysis," *J Comput Chem*, vol. 31, pp. 66–74, 2010.

[13] O. Alter, P. O. Brown, and D. Botstein, "Generalized singular value decomposition for comparative analysis of genomescale expression data sets of two different organisms," *PNAS*, vol. 100, pp. 3351–3356, 2003.

[14] P. D. Wentzell, D. T. Andrews, and B. R. Kowalski, "Maximum likelihood multivariate calibration," *Anal. Chem.*, vol. 69, pp. 2299–2311, 1997.

[15] P. D. Wentzell, D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski, "Maximum likelihood principal component analysis," *J. Chemomet*, vol. 11, pp. 339–366, 1997.

[16] P. D. Wentzell and M. T. Lohnes, "Maximum likelihood principal componnet analysis with correlated measuremnet errors: theoretical and practical considerations," *Chemomet. Intell. Lab. Syst*, vol. 45, pp. 65–85, 1999.

[17] M. Schuermans, I. Markovsky, P. D. Wentzell, and S. V. Huffel, "On the equivalence between total least squares and maximum likelihood pca," *Analytica Chimica Acta*, vol. 544, pp. 254–267, 2005.

[18] A. Faro, D. Giordano, and C. Spampinato, "Combining literature textmining with microarray data: advances for system biology modeling," *Briefings in bioinformatics*, vol. 13, pp. 61–82, 2011.

[19] O. Gevaert, S. Van Vooren, and B. De Moor, "Integration of microarray and textual data improves the prognosis prediction of breast, lung and ovarian cancer patients," *PSB08 proceedings*, pp. 13–16, 2008.

[20] K. Chin, S. De Vries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W.-L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, T. Ryder, F. Chen, H. Feiler, T. Tokuyasu, C. Kingsley, S. Dairkee, Z. Meng, K. Chew, D. Pinkel, A. Jain, B. M. Ljung, L. Esserman, D. G. Albertson, F. M. Waldman, and J. W. Gray, "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies," *Cancer Cell.*, vol. 10, pp. 529–541, 2006.

[21] K. R. Hess, K. Anderson, W. F. Symmans, V. Valero, N. Ibrahim, J. A. Mejia, D. Booser, R. L. Theriault, A. U. Buzdar, P. J. Dempsey, R. Rouzier, N. Sneige, J. S. Ross, T. Vidaurre, H. L. Gómez, G. N. Hortobagyi, and L. Pusztai, "Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer," *J Clin Oncol*, vol. 24, pp. 4236–4244, 2006.

[22] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi, "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis," *J Natl Cancer Inst*, vol. 98, pp. 262–272, 2006.

[23] M. Dai, A. D. Wang, Pand Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, and F. Watson, S Jand Meng, "Evolving gene/transcript definitions significantly alter the interpretation of genechip data," *Nucleic Acids Res.*, vol. 33, p. e175, 2005.

[24] D. Mishra and B. Sahu, "Feature selection for cancer classification: A signal-to-noise ratio approach," *International Journal of Scientific & Engineering Research*, vol. 2, pp. 1–7, 2011.

[25] A. Daemen, O. Gevaert, F. Ojeda, A. Debucquoy, J. A. K. Suykens, C. Sempoux, J.-P. Machiels, K. Haustermans, and B. De Moor, "Kernel-based integration of genome-wide data for clinical decision support," *Genome Medicine*, vol. I, pp. 39.1–39.17, 2009.

[26] Bioinformatics research group. [Online]. Available: http://www.upo.es/eps/bigs/datasets.html

[27] National cancer institute dictionary of cancer terms. [Online]. Available: http://www.cancer.gov/dictionary

[28] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine 2*, vol. 11, pp. 559–572, 1901.

[29] G. H. Golub and C. F. Van Loan, *Matrix Computations*. 2nd ed. (Baltimore: Johns Hopkins University Press), 1989.

[30] N. J. Higham, "Newton's method for the matrix square root," *Mathematics of Computation*, vol. 46, pp. 537–549, 1986.

[31] D. S. Watkins, "Product eigenvalue problems," *Society for Industrial and Applied Mathematics*, vol. 47, pp. 3–40, 2005.

[32] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, pp. 293–300, 1999.

[33] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[34] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminnat measures for performance evaluation," *American Association for Artificial Intelligence*, 2006.

[35] K. De Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, and J. A. K. Suykens, "Ls-svmlab toolbox user's guide version 1.8," 2010.

[36] S. Hu and R. J. Sunil, "Statistical redundancy testing for improved gene selection in cancer classification using microarray data," *Cancer Inform*, vol. 3, pp. 29–41, 2007.

[37] E. J Moler, M. L Chow, and I. S Mian, "Analysis of molecular profile data using generative and discriminative methods," *Physiol Genomics*, vol. 4, pp. 109–126, 2000.

[38] B. Fogel, G and W. Corne, D, *Computational intelligence in bioinformatics*. IEEE Press Series on Computstional Intelligence.

[39] D. Venet, C. Maenhaut, and H. Bersini, "Separation of samples into their constituents using gene expression data," *Bioinformatics*, vol. 17, pp. S279–S287, 2001.

[40] U. Alon, N. Barkai, D. A Notterman, K. Gish, S. Ybarra, D. Mack, and A. J Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *PNAS*, vol. 96, pp. 6745–6750, 1999.

[41] M. Filippone, F. Masulli, and S. Rovetta, "Simulated annealing for supervised gene selection," *Springer-Verlag*, 2010.

[42] O. H. Fang, N. Mustapha, and S. M. N, "Integrative gene selection for classification of microarray data," *Computer and Information Science*, vol. 4, pp. 55–63, 2011.

[43] H. Zhang, X. Song, and X. Zhang, "Miclique: An algorithm to identify differentially coexpressed disease gene subset from microarray data," *Journal of Biomedicine and Biotechnology*, 2009.

[44] X. Yana, M. Denga, W. K Fungb, and M. Qiana, "Detecting differentially expressed genes by relative entropy," *Journal of Theoretical Biology*, vol. 234, pp. 395–402, 2005.

[45] S. Wang, J. Huang, J. He, A. Wang, S. Xu, S. F. Huang, and S. Xiao, "Rpl41, a small ribosomal peptide deregulated in tumors, is essential for mitosis and centrosome integrity," *Neoplasia*, vol. 3, pp. 284–93, 2010.

[46] X. Li, S. Rao, Y. Wang, and B. Gong, "Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling," *Nucleic Acids Res.*, vol. 9, pp. 2685–2694, 2004.

[47] M. S. Kurokawa, M. Hatsugai, Y. Noguchi, T. Yoshioka, H. Mitsui, H. Yasuda, and T. Kato, *Proteomic Approaches for Biomarker Discovery in Ulcerative Colitis*, 2011.

[48] D. Ghosh, Z. Li, X. F. Tan, T. K. Lim, Y. Mao, and Q. Lin, "itraq based quantitative proteomics approach validated the role of calcyclin binding protein (cacybp) in promoting colorectal cancer metastasis. mol cell proteomics," *Mol Cell Proteomics*, vol. 7, pp. 1865–80, 2013.

[49] Z. Wang and V. Palade, *Fuzzy gene mining: A fuzzy-based framework for cancer microarray data analysis*. A john wiley & Sons, 2008.

[50] T. H. M. Chan, L. Chen, and X.-Y. Guan, "Role of translationally controlled tumor protein in cancer progression," *Biochem Res Int.*, vol. 7, pp. 1–5, 2012.

[51] W. H. Kim, B. L. Lee, S. H. Jun, and K. H. K. Song, S Y and, "Expression of 32/67-kda laminin receptor in laminin adhesionselected human colon cancer cell lines," *Br J Cancer*, vol. 77, pp. 15–20, 1998.

[52] J. Kitayama, H. Nagawa, N. Tsuno, T. Osada, K. Hatano, E. Sunami, H. Saito, and T. Muto, "Laminin mediates tethering

and spreading of colon cancer cells in physiological shear flow," *Br J Cancer*, vol. 80, pp. 1927–1934, 1999.

Minta Thomas was born in Kottyam, India on 1983. She received Masters degree in computer science in 2005 at Mahatma University, Kerala, India and M.Phil in bioinformatics in 2007 at Kerala University, Kerala, India. Presently, she is pursing Ph.D in informatics at K U Leuven, Belgium. The subject of her research is the development of new computational techniques for biological data analysis and classification problems.

Anneleen Daemen obtained a masters degree in Electrical Engineering and a doctoral degree in Bioinformatics from the Katholieke Universiteit Leuven (Belgium) in 2010. She contributed to the field of breast cancer during a postdoc at Lawrence Berkeley National Laboratory in the labs of Joe Gray and Paul Spellman, and at the University of California, San Francisco in the lab of Laura van't Veer. In 2011, she was appointed Visiting Professor in Statistics at the Department of Obstetrics and Gynecology (Leuven, Belgium). She now works as scientist at Genentech, Inc (South San Francisco, CA), where she focuses on high-dimensional biological data analysis to obtain as complete a picture of what is happening within a tumor cell as possible for drug development and diagnostics. Dr. Daemen has a distinguished publications record of over 50 full-length scientific articles in the field of bioinformatics, and holds an international patent for a kit to evaluate the biological stage of HCC tumors.

Bart De Moor was born in Halle, Brabant, Belgium, on July 12, 1960. He received the doctoral degree in applied sciences in 1988 from the Katholieke Universiteit, Leuven, Belgium. He was a Visiting Research Associate from 1988 to 1989 in the Department of Computer Science and Electrical Engineering of Stanford University, Stanford, CA. He is a full Professor at the Katholieke Universiteit, Leuven. His research interests include numerical linear algebra, system identification, advanced process control, data mining, and bio-informatics. He is the (co-)author of several books and several hundreds of papers, some of which have been awarded. Dr.De Moor received the Leybold-Heraeus Prize in 1986, the Leslie Fox Prize in 1989, the Guillemin-Cauer Best Paper Award, of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS in 1990, the biannual Siemens prize in 1994, and became a Laureate of the Belgian Royal Academy of Sciences in 1992.