SCIENTIFIC REPORTS
nature research

OPEN

Analysis of microarray right-censored data through fused sliced inverse regression

Jae Keun Yoo*

Sufficient dimension reduction (SDR) for a regression pursue a replacement of the original *p*-dimensional predictors with its lower-dimensional linear projection. The so-called sliced inverse regression (SIR; [5]) arguably has the longest history in SDR methodologies, but it is still one of the most popular one. The SIR is known to be easily affected by the number of slices, which is one of its critical deficits. Recently, a fused approach for SIR is proposed to relieve this weakness, which fuses the kernel matrices computed by the SIR application from various numbers of slices. In the paper, the fused SIR is applied to a large-*p*-small *n* regression of a high-dimensional microarray right-censored data to show its practical advantage over usual SIR application. Through model validation, it is confirmed that the fused SIR outperforms the SIR with any number of slices under consideration.

Sufficient dimension reduction (SDR) in regression of $Y|\mathbf{X} \in R^p = (X_1, \ldots, X_p)^{\mathrm{T}}$ pursue a replacement of the original p-dimensional predictors $\mathbf{X}$ with its lower-dimensional linear projection without loss of information about the conditional distribution of $Y|\mathbf{X}$. Equivalently, SDR seeks for finding $\mathbf{M} \in R^{p\times q}$ such that

$$Y \coprod \mathbf{X}|\mathbf{M}^{\mathrm{T}}\mathbf{X}, \tag{1}$$

where a notation of $\coprod$ represents a statistical independence and $q \leq p$.

The conditional independence statement (1) indicates that the two conditional distributions of $Y|\mathbf{X}$ and $Y|\mathbf{M}^{\mathrm{T}}\mathbf{X}$ are equivalent, so $\mathbf{X}$ is replaced by $\mathbf{M}^{\mathrm{T}}\mathbf{X}$ with preventing loss of the information about $Y|\mathbf{X}$. A subspace spanned by the columns of $\mathbf{M}$ satisfying (1) is called a dimension reduction subspace. If the subspace acquired by intersecting all possible dimension reduction subspaces is still a dimension reduction subspace, the intersection subspace is defined as the *central subspace* $S_{Y|\mathbf{X}}$[1]. The central subspace is minimal and unique, and its restoration is the main purpose of SDR literature. Hereafter, notations of d and $\boldsymbol{\eta} \in R^{p\times q}$ represent the true dimension and orthonormal basis matrix of $S_{Y|\mathbf{X}}$, respectively. The dimension-reduced predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ is called sufficient predictors.

Data, whose sample size n is smaller than p, such as microarray data, high-throughput data, etc., are quite popular these days. In such data, so-called curse of dimensionality usually occurs, so a proper model-building are often problematic in practice. Then the SDR of $\mathbf{X}$ through $S_{Y|\mathbf{X}}$ can facilitate a model specification, so it turns out to be practically useful in such data.

One of the most popular SDR methods should be sliced inverse regression (SIR[2]). Implementation of SIR requires a categorization of a response variable Y, called *slicing*, and the selection of the appropriate number of slices are often critical in the application results. So far, any ideal or recommended selection guidelines to choose the number of slices are not yet known. To overcome this, a fused approach is proposed in[3] by combining sample kernel matrices of SIR constructed by varying the numbers of slices. The combining approach in[3] is called *fused sliced inverse regression* (FSIR). According to[3], FSIR results in robust basis estimates of $S_{Y|\mathbf{X}}$ to the numbers of slices.

The purpose of this paper is to analyze a micro array right-censored survival data by implementing fused sliced inverse regression (FSIR) by[3]. The performances of FSIR will be compared with the usual SIR applications with different numbers of slices. The organization of the paper is as follows. The SIR and FSIR along with the applicability to survival regression is discussed in section 2. In the same section, the permutation dimension test is discussed. Diffuse large-B-cell lymphoma data is analyzed through SIR and FSIR, and their results are compared in section 3. We summarize our work in section 4.

Department of Statistics, Ewha Womans University, Seoul, 03760, Republic of Korea. *email: peter.yoo@ewha.ac.kr

We will define the following notations, which will be used frequently throughout the rest of the paper. A subspace $S(\mathbf{B})$ stands for a subspace spanned by the columns of $\mathbf{B}$. And, we define that $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{X})$.

Material and Methods

Sliced inverse regression and fused sliced inverse regression. Before explaining sliced inverse regression[2], the predictor $\mathbf{X}$ is normalized to $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{X}–E(\mathbf{X}))$. Letting $S_{Y|\mathbf{Z}}$ be the central subspace for a regression of $Y|\mathbf{Z}$, then the relationship that $S_{Y|\mathbf{X}} = \mathbf{\Sigma}^{-1/2}S_{Y|\mathbf{Z}}$ holds. Define η_z be $p \times d$ orthonormal basis matrix for $S_{Y|\mathbf{Z}}$. Consider the so-called linearity condition: (C1) $E(\mathbf{Z}|\boldsymbol{\eta}_z{}^{\mathrm{T}}\mathbf{Z})$ is linear in $\boldsymbol{\eta}_z{}^{\mathrm{T}}\mathbf{Z}$. According to[2], a proper subspace of $S_{Y|\mathbf{Z}}$ can be constructed under linearity condition:

$$S(E(\mathbf{Z}|Y)) \subseteq S_{Y|\mathbf{Z}} \Leftrightarrow S(\mathbf{\Sigma}^{-1}E(\mathbf{X}|Y)) \subseteq S_{Y|\mathbf{X}}.$$

For estimating of $S_{Y|\mathbf{X}}$ completely, it is typically assumed that $S(\mathbf{\Sigma}^{-1}E(\mathbf{X}|Y)) = S_{Y|\mathbf{X}}$. The so-called sliced inverse regression is a method to recover $S_{Y|\mathbf{X}}$ by computing $E(\mathbf{X}|Y)$.

In population, the quantity $E(\mathbf{Z}|Y)$ should be computed without any specific assumptions on $Y|\mathbf{Z}$. If Y is discrete with h levels, $E(\mathbf{Z}|Y = s)$ is the average of $\mathbf{Z}$ within the sth category of Y. Following this idea, if Y is continuous or many-valued, Y is transformed to a categorized response $\widetilde{Y}$ with h levels. Then $E(\mathbf{Z}|\widetilde{Y} = s)$ becomes the average of $\mathbf{Z}$ within the sth category of $\widetilde{Y}$ for $s = 1, \ldots, h$. This categorization of Y is called *slicing*, which is done for each category to have equal numbers of observations. The SIR constructs:

$$\mathrm{M}_{\mathrm{SIR}} = \mathrm{cov}(E(\mathbf{Z}|Y)) \text{ or } \mathrm{M}_{\mathrm{SIR}} = \mathrm{cov}(E(\boldsymbol{Z}|\widetilde{Y})).$$

In sample structure, the algorithm of SIR is as follows:

1. Construct $\widetilde{Y}$ by dividing the range of Y into h non-overlapping intervals. Let n_s be the number of observations for the sth category of $\widetilde{Y}$ for $s = 1, \ldots, h$.
2. Compute $\hat{E}(\mathbf{Z}|\widetilde{\mathrm{Y}} = \mathrm{S}) = \sum_{\hat{i}\in Y=s}(\hat{\mathbf{Z}}_i/n_s)$, for $s = 1, \ldots, h$, where $\hat{\mathbf{Z}}_i = \hat{\sum}^{-1/2}(\mathbf{X}_{\mathrm{i}} - \overline{\mathbf{X}})$.
3. Construct $\hat{\mathbf{M}}_{\mathrm{SIR}}$ as follows: $\hat{\mathbf{M}}_{\mathrm{SIR}}\hat{\mathrm{c}}\mathrm{ov}(E(\hat{\mathbf{Z}}|\widetilde{Y})) = \sum_{s=1}^{n}(n_s/n)\ \hat{\mathrm{E}}(\mathrm{Z}|\widetilde{Y} = s)\ \ \hat{\mathrm{E}}(\mathbf{Z}|\widetilde{Y} = s)^{\mathrm{T}}$
4. Spectral-decompose $\hat{\mathbf{M}}_{\mathrm{SIR}}$: $\hat{\mathbf{M}}_{\mathrm{SIR}} = \sum_{j=1}^{p}\hat{\lambda}_j\hat{\gamma}_j\hat{\gamma}_j^{\mathrm{T}}$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.
5. Determine the structural dimension d. Let $\hat{d}$ denote an estimate of d.
6. A set of eigenvectors $(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ corresponding to first $\hat{d}$ largest eigenvalues are the estimate of an orthonormal basis for $S_{Y|\mathbf{Z}}$.
7. Back-transform $\hat{\sum}^{-1/2}(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ to have the estimate of an orthonormal basis of $S_{Y|\mathbf{X}}$.

As we can see the implementation of SIR in practice, the results may critically vary depending on the selection of h. This is discussed in[3]. Define that $\mathbf{M}_{\mathrm{FSIR(h)}} = (\mathbf{M}_{\mathrm{SIR}(1)}, \ldots, \mathbf{M}_{\mathrm{SIR}(h)})$, where $\mathbf{M}_{\mathrm{SIR}(h)}$ stands for the kernel matrix of SIR with h slices. Since $S(\mathbf{M}_{\mathrm{SIR}(k)}) = S_{Y|\mathbf{Z}}$ for $k = 2, 3, .., h$, we have.

$$S(\mathbf{M}_{\mathrm{SIR}(k)}) \subseteq S(\mathbf{M}_{\mathrm{FSIR}(h)}) = S_{Y|\mathbf{Z}}, \quad k = 2, 3, \ldots, h.$$

In[3], the matrix $\mathbf{M}_{\mathrm{FSIR}(h)}$ is proposed as another kernel matrix to estimate $S_{Y|\mathbf{X}}$, and this approach is called *fused sliced inverse regression* (FSIR). In[3], it is confirmed that $\mathbf{M}_{\mathrm{FSIR}(h)}$ is robust to the choices of h through various numerical studies.

Inference on $S_{Y|\mathbf{Z}}$ is done by the spectral decomposition of $\hat{\mathbf{M}}_{\mathrm{FSIR}(k)}$. The eigenvectors of $\hat{\mathbf{M}}_{\mathrm{FSIR}(k)}$ corresponding to its non-zero eigenvalues form an estimate of an orthonormal basis of $S_{Y|\mathbf{Z}}$.

Permutation dimension test. The true structural dimension d is determined by a sequence of hypothesis tests[4]. Starting with $m = 0$, test H_0: $d = m$ versus H_1: $d = m + 1$. If H_0: $d = m$ is rejected, increment m by 1 and redo the test, stopping the first time H_0 is not rejected and setting $\hat{d} = m$. This dimension test is equivalent to testing the rank of $\mathbf{M}_{\mathrm{FSIR}(h)}$. So, a proposed test statistics is as follows:

$$\hat{\Lambda}_m = n\sum_{i=m+1}^{p}\hat{\lambda}_i,$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.

Here a permutation approach is adopted to implement the dimension estimation. An advantage of the permutation test is no requirement of the asymptotics of $\hat{\Lambda}_m$. The permutation test algorithm is as follows:

1. Construct $\hat{\mathbf{M}}_{\mathrm{FSIR}(h)}$. Under H_0: $d = m$, compute $\hat{\Lambda}_m$ and partition the eigenvectors: $\hat{\mathbf{\Gamma}}_1 = (\hat{\gamma}_1, \ldots, \hat{\gamma}_m)$ and $\hat{\mathbf{\Gamma}}_2 = \left(\hat{\gamma}_{m+1}, \ldots, \hat{\gamma}_p\right)$.
2. Construct two sets of vectors: $\hat{V}_i \in R^{m\times 1} = \hat{\Gamma}_1{}^{T}\hat{\mathrm{Z}}_i$ and $\hat{U}_i \in R^{(p-m)\times 1} = \hat{\Gamma}_2{}^{\mathrm{T}}\hat{\mathrm{Z}}_i, \quad i = 1, \ldots, n$
3. Randomly permute index i of the $\hat{U}_i$ with the permuted set $\hat{U}_i^*$.
4. Construct the test statistics $\hat{\Lambda}_m^*$ based on a regression of $Y_i|(\hat{V}_i\ \hat{U}_i^*)$.
5. Repeat steps (3–4) N times, where N is the total number of permutations. The p-value of the hypothesis testing is the fraction of $\hat{\Lambda}_m^*$ that exceed $\hat{\Lambda}_m$.

The setting $N = 1000$ is a widely-used choice.

Application to survival regression. Survival regression is a study of the conditional distribution of survival time T given a set of predictors $\mathbf{X}$. Naturally, SDR in the survival regression should seek for recovering the central subspace $S_{T|\mathbf{X}}$:

$$T \coprod \mathbf{X} | \eta^T \mathbf{X}. \tag{2}$$

However, since the true survival time T cannot be completely observed due to censoring, the direct study of $T|\mathbf{X}$ cannot be usually done.

Instead, the data $(Y_i, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, are collected as n independent and identically distributed realizations of $(T, C, \mathbf{X})$, where $Y = T\,\delta + C(1-\delta)$, $\delta = 0, 1$ is an indicator variable whose value is equal to 1, if $\delta(C > T) = 1$ and 0, otherwise, and C stands for a censoring time. This type of censoring is called right-censoring. Using $(Y_i, \delta_i, \mathbf{X}_i)$, the regression of $T|\mathbf{X}$ is replaced as follows. The first step is a consideration of a regression of $(T, C)|\mathbf{X}$. The construction of $(T, C)|\mathbf{X}$ directly implies that $S_{T|\mathbf{X}} \subseteq S_{(T,C)|\mathbf{X}}$. According to[5], the central subspace $S_{(Y,\delta)|\mathbf{X}}$ from a bivariate regression of $(Y, \delta)|\mathbf{X}$ is informative to $S_{(T,C)|\mathbf{X}}$, because $S_{(Y,\delta)|\mathbf{X}} \subseteq S_{(T,C)|\mathbf{X}}$. Since (Y, δ, X) are collected for survival analysis, the estimation of $S_{(Y,\delta)|\mathbf{X}}$ can be done. The two regressions of $T|\mathbf{X}$ and $(Y, \delta)|\mathbf{X}$ are connected in[3] under condition: (C2) $C \coprod \mathbf{X} | (\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}, T)$. Conditionc2 is weaker than $C \coprod (T, \mathrm{X})$, which is normally assumed in in survival analysis. Then, condition C2 guarantees that statement (2) is equivalent to $(T,C) \coprod \mathbf{X} | \eta^{\mathrm{T}}\mathbf{X}$, so we have $S_{(T,C)|\mathbf{X}} = S_{T|\mathbf{X}}$. Therefore, the following relation directly implied:

$$S_{(Y,\delta)|\mathbf{X}} \subseteq S_{(T,C)|\mathbf{X}} = S_{T|\mathbf{X}}.$$

According to[5,6], the equality would normally hold, because proper containment requires carefully balanced conditions. Then, SIR and FSIR are directly applicable with bivariate slicing of Y and δ to recover $S_{T|\mathbf{X}}$. Similar discussion about this can be found in section 4.2 of[6].

Results

Analysis of diffuse large-B-cell lymphoma data. The diffuse large-B-cell lymphoma dataset (DLBCL[7]) contains measurements of 7399 genes from 240 patients obtained from customized cDNA microarrays. For each patient, his/her survival time was recorded and varied from 0 to 21.8 years. The total uncensored cases (deceased) are 138 among 240 patients. More detailed description on the data is founded in[6–8].

We follow the approach in[9] to analyze the DLBCL. The DLBCL is randomly divided into the training set of 160 and the test set of 80. As usual, the training set is used for model-building, and the test set is utilized for model-validation. First, the 7399 genes in the training set, which are denoted as $\mathbf{X}_{\mathrm{tr}}$, are initially reduced to their 40 principal components through principal component analysis. Letting $\hat{\boldsymbol{\Omega}} \in R^{7399\times 40}$ be the rotation matrix, the 40 principal components are $\hat{\boldsymbol{\Omega}}^{\mathrm{T}}\mathbf{X}_{\mathrm{tr}}$.

Second, the SIR is employed for the additional dimension reduction of $\hat{\boldsymbol{\Omega}}^{\mathrm{T}}\mathbf{X}_{\mathrm{tr}}$ with observed survival time and censoring status as bivariate responses. Let $\hat{\mathbf{B}} \in R^{40\times d}$ stand for the estimated matrix. According to[9], the dimension d is estimated to be one. The finalized estimated sufficient predictors through this two-step dimension reduction are denoted as $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}_{\mathrm{tr}}$ with $\hat{\boldsymbol{\eta}} \in R^{7399\times 1} = \hat{\boldsymbol{\Omega}}\hat{\mathbf{B}}$.

For model-building, the Cox-proportional hazards model was fitted with $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}_{\mathrm{tr}}$. For model-validation, the predicted scores and the corresponding area under ROC curves for prediction of survival time from 1 to 10 years for both the training and test sets were computed. For the test set, the dimension- reduced predictors are defined as $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}_{\mathrm{te}}$, where $\hat{\boldsymbol{\eta}}$ is obtained from the training set and $\mathbf{X}_{\mathrm{te}}$ stands for the predictors in the test set. The area closer to one indicates better estimation.

One potentially arguable issue in the analysis in the context should arise on the selection of the number of slices h in the SIR application. As discussed in the previous section, its performance inevitably depends on h. To investigate how serious they impact on the model-validation, we consider $h = 4$, 6, 8 and 10 for SIR along with FSIR. Following the guidance of[3], 10 slices are used in FSIR. The area under ROC curve for the training and test sets are reported in Fig. 1.

First, we see the areas under ROC curves for the training set in Fig. 1(a). Larger areas indicate better prediction performances. For the SIR application, the smaller numbers of slices show the better performances. The FSIR is not best among the all application of SIR considered here, but there are no notable differences to the best results, which is with $h = 2$, among all the SIR applications. Therefore, for the training set, the FSIR is not cause of concern at all. In the case of the test set in Fig. 1(b), the FSIR shows better prediction performances than any of the SIR applications. The prediction results by the FSIR is consistent in both the training and test sets, while the usual SIR applications are very sensitive to the choices of h, as expected. The application of the FSIR to the data is concluded to be successful.

Discussion

According to Fig. 1(a,b), the areas under ROC curves for the training and test sets are reversed against h in the SIR applications. In the training set, smaller numbers of slices have larger areas, while the areas with smaller numbers of h become smaller in the test sets, which is even below 0.5. The area equal to 0.5 is often used as the cut-off. Therefore, for SIR, the application with $h = 10$ alone is above 0.5 in both train and test sets, although its performance is worst among the others in the train set. The FSIR, however, shows reliable and consistently good performances in both training and test sets.

The best selections of h in the training set and the test set are different, and this selection bias in h can cause the ironic results in SIR. This bias also affects the estimation of h in the analysis. With level 5%, the SIR application with $h = 4$ and 8 determines that $\hat{d} = 0$ with the corresponding p-values of 0.139 and 0.244 for H_0: $d = 0$, respec-

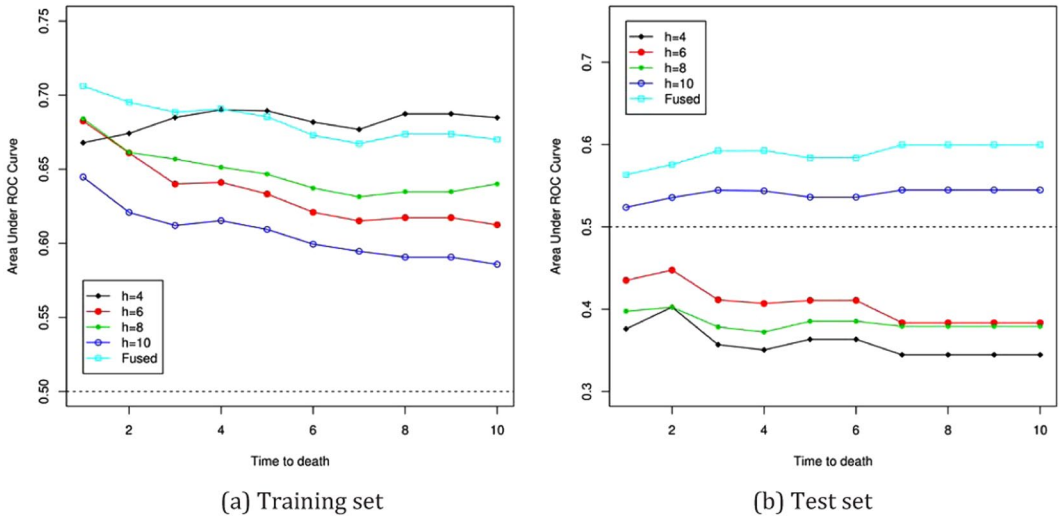

Figure 1. Area under ROC curves at time 1 to 10 years for DLBCL data in Section 3: $h = 4, 6, 8, 10$, sliced inverse regression with the according number of slices; Fused, fused sliced inverse regression with $h = 10$.

tively. However, the SIR with $h = 6$ and 10 determines that $\hat{d} = 1$ ($h = 6$: 0.009 for H_0: $d = 0$ and 0.097 for H_0: $d = 1$ & $h = 10$: 0.007 for H_0: $d = 0$ and 0.10 for H_0: $d = 1$). This confirms the severe sensitivity of the SIR to the selection of h in the high-dimensional data analysis. The FSIR determines that $\hat{d} = 1$ with the p-values of 0.014 for H_0: $d = 0$ and of 0.115 for H_0: $d = 1$. This shows that the FSIR has potential advantages over the SIR in high-dimensional data analysis in practice.

Conclusion

Fused sliced inverse regression (FSIR) proposed by[3] solves the sensitiveness of slice inverse regression (SIR[2]) to the number of slices by combining SIR kernel matrices. In this paper, the fused sliced inverse regression is applied to high-dimensional microarray right-censored data to show the potential advantage to large p-small n data over the usual SIR application. The predictors are initially reduced through principal components analysis, and then SIR and FSIR are implemented with 40 principal components. According to model-validation, the SIR reveals its sensitiveness to the number of slices. Moreover, ironic validation results are observed in the training and test sets. For SIR, the numbers of slices to have better performances in the training set show worse performances in the test set. This may be because good slicing schemes in the training set do not coincide with that in the test set. This is confirmed again through the estimation of the true structural dimension. However, FSIR shows better performances in the training and test sets than all SIR-application under consideration. This proves a practical advantage of FSIR over SIR.

The usage of FSIR can improve the accuracy in high-dimensional data analysis, which often arise in many scientific fields including biological sciences, so it can contribute to discover new founding in the many science areas.

Data availability

The dataset of the diffuse large-B-cell lymphoma dataset (DLBCL[7]) is available at the following two web locations: http://llmpp.nih.gov/DLBCL; http://statweb.stanford.edu/~tibs/superpc/staudt.html.

Received: 28 February 2019; Accepted: 1 October 2019;
Published online: 22 October 2019

References

1. Cook, R. D. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley: New York (1998).
2. Li, K. C. Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association.* **86**, 316–327 (1991).
3. Cook, R. D. & Zhang, X. Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association.* **109**, 815–827 (2014).
4. Rao, C. R. *Linear Statistical Inference and Its Application*. Wiley: New York (1965).
5. Cook, R. D. Dimension reduction and graphical exploration in regression including survival analysis. *Statistics in Medicine.* **22**, 1399–1413 (2003).
6. Yoo, J. K. Advances in seeded dimension reduction: Bootstrap criteria and extension. *Computational Statistics and Data Analysis.* **60**, 70–79 (2013).
7. Rosenwald, A. *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine.* **346**, 1937–1947 (2002).
8. Yoo, J. K. Multivariate seeded dimension reduction. *Journal of the Korean Statistical Society.* **43**, 559–566 (2014).
9. Li, L. Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics.* **20**, 3406–3412 (2004).

Acknowledgements
This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (NRF-NRF-2019R1F1A1050715/2019R1A6A1A11051177).

Competing interests
The author declares no competing interests.

Additional information
Correspondence and requests for materials should be addressed to J.K.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.