Published in final edited form as:
J Stat Comput Simul. 2017 ; 87(7): 1363–1378. doi:10.1080/00949655.2016.1263992.

Temporal Prediction of Future State Occupation in a Multistate Model from High-Dimensional Baseline Covariates via Pseudo-Value Regression

Sandipan Dutta[a], **Susmita Datta**[b], and **Somnath Datta**[b]

[a]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA

[b]Department of Biostatistics, University of Florida, Gainesville, FL, USA

Abstract

In many complex diseases such as cancer, a patient undergoes various disease stages before reaching a terminal state (say disease free or death). This fits a multistate model framework where a prognosis may be equivalent to predicting the state occupation at a future time *t*. With the advent of high throughput genomic and proteomic assays, a clinician may intent to use such high dimensional covariates in making better prediction of state occupation.

In this article, we offer a practical solution to this problem by combining a useful technique, called pseudo value regression, with a latent factor or a penalized regression method such as the partial least squares (PLS) or the least absolute shrinkage and selection operator (LASSO), or their variants. We explore the predictive performances of these combinations in various high dimensional settings via extensive simulation studies. Overall, this strategy works fairly well provided the models are tuned properly. Overall, the PLS turns out to be slightly better than LASSO in most settings investigated by us, for the purpose of temporal prediction of future state occupation. We illustrate the utility of these pseudo-value based high dimensional regression methods using a lung cancer data set where we use the patients' baseline gene expression values.

Keywords

Censoring; Covariate; Gene expression; LASSO; PLS; Survival

1. Introduction

Multistate models are typically used to describe the progression of a set of subjects through a succession of stages until they reach a certain endpoint. This endpoint is called an absorbing state as no more transitions are possible from this state. A simple example of such a model is the setting of a survival analysis where there are only two states, viz., alive or the initial state, and dead or the final state. In disease studies, like cancer, prognosis of patients is of much importance. This includes predicting how complicated the stage of the disease will be for a patient, e.g. whether the patient will still be at an intermediate stage II or at the

---

Appendix. Supplementary Material: Web Appendices, Web-Tables, and Web-Figures referenced throughout the article can be found in the online Supplementary Material available with this article.

final and more severe stage IV of lung cancer, after t (say) months from the point of study, or, whether a patient can really survive till t months after a follow-up study. This requires estimation of state occupation probability, which is the probability that an individual would be occupying a particular stage of the disease process at a given time. For survival models the survival probability at a given time can be interpreted as one of the two state occupation probabilities. Estimation of these state occupation probabilities become difficult in the presence of censored data. In such cases one estimates the state occupation probability at given time in presence of censoring using the Aalen-Johansen estimator [1]. However, often we have baseline covariates on the patients during a disease study; one has to assimilate these additional covariate information for better prognosis of the disease pattern in a given individual. This can be done through regression modeling of state occupation probabilities at a given time incorporating the covariate information of the subjects under study and using the resultant model for prediction purposes.

Andersen, Klein, and Rosthoj [2] invented a simple yet effective technique for directly modeling state occupation probability in a multistate process based on a given set of covariates. They proposed the overall marginal estimation of a state occupation probability using Aalen-Johansen estimator, and then using the ‘leave-one-out’ jackknife based ‘pseudo-values’ [3] of the marginal estimate as responses in regression modeling based on covariates. The leave-one-out jackknife method for a summary statistic (e.g. Aalen-Johansen estimator of marginal state occupation probabilities) makes it possible to attach a separate estimate for each individual under study. This leads to the formation of the pseudo-value (PV). A pseudo-value corresponding to an individual is constructed in such a way that it reflects the extent to which the overall marginal estimator is affected by the presence or absence of that individual in the study. So these pseudo-values can be, intuitively, related to the covariates at the individual level. In that case PV corresponding to an individual can be thought of containing information on how the covariates of that individual affect the overall marginal estimator. The usefulness of this pseudo-value based regression is largely due to the fact that, under suitable regularity conditions, the pseudo-values computed from an asymptotically linear and unbiased estimator will be approximately i.i.d with the same conditional expectation (regression function) that we are trying to estimate [4]. The pseudo-value based regression technique has since then been applied to other time to event data problems; see, [5-8] among others. Although originally developed for testing the effects of covariates in censored data settings, the pseudo-value based regression technique can also be used for prediction of future state occupation. However, most of the existing works based on the pseudo value technique have been carried out under the generalized linear regression framework.

With the advent of high throughput genomic and proteomic assays, a clinician may want to use these information collected on a patient at baseline in making better prediction of future state occupation. In such situations, the standard linear or generalized linear models will not be applicable as the covariate dimension (e.g., number of genes in microarrays or next generation sequencing arrays, number of proteins in protein arrays, or mass over charge ratios in mass spectrometry based proteomic profiles) is typically very large compared to the number of individuals under study. A recent work involving the pseudo-value technique in high dimensional settings was pursued by Mogensen and Gerds [9] for a classification problem through the random forest approach, but it was limited only to competing risk models. In this article, we consider estimating the probability that an individual would be in a certain state of a general multistate disease process at a given time based on his or her covariate profile. Generally speaking, either a latent factor regression, or a penalized regression, or a combination of the two, have been used in literature to handle high dimensional covariates. Thus we consider the pseudo-value based regression approach in combination with a latent factor or a penalized regression technique. We explore the predictive performances of latent factor regressions such as Partial Least Squares [10, 11] as well as, penalized regressions such as Least Absolute Shrinkage and Selection Operator [12] all using the pseudo-value approach in cases where the covariate dimension exceeds the sample size.

The rest of the article is organized as follows. In Section 2, we give an overview of the PV approach in regression modeling with special emphasis on the regression of state occupation probability in multistate models. Section 3 contains simulation studies involving an irreversible illness-death multistate model in which we compare the predictive performances of different PV based high dimensional regressions, namely, Partial Least Squares (PLS), Sparse PartialLeast Squares (SPLS) [13], Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net [14], and Adaptive LASSO (AdLASSO) [15]. This section also has a simulation study representing a two-state survival framework, where we compare the performances of Cox model based LASSO regression [16] and two competing PV based high dimensional regression methods. In Section 4, we demonstrate the use of the PV based high dimensional methods in predicting patient survival through a lung cancer dataset that contain censored observations. The article ends with a discussion in Section 5.

2. Background of the Methods

Let us briefly discuss the notations used in this article for developing the methods. Suppose we have a multistate model framework involving n individuals. We are interested in inferring the probability of occupation of a given state h of the multistate system at a given time t. We denote this probability by $P_h(t)$. If $U(t)$ denote the state at time t, then we have Ph(t) = $E(I(U(t) = h))$. We denote a marginal estimator of $P_h(t)$ as $\hat{P}_h(t)$. Also, we have information on q covariates related to each of the n subjects under study.

2.1 Pseudo-values and their Application in Regression Modeling

The pseudo-value approach was first obtained for the 'leave-one-out' jackknife resampling technique, with the initial purpose being studying the bias and standard error of an estimator. The idea behind the construction of pseudo-values is easily comprehensible when the estimator is linear. If $\hat{\theta}$ is an estimator of a parameter of interest θ based on a random sample of size n, and if $\hat{\theta}^{(-i)}$ is the estimate of θ obtained by deleting the i^{th} observation from the original sample, then the i^{th} pseudo-value is defined as $\eta_i = n\hat{\theta} - (n-1)\hat{\theta}^{(-i)}$, where $i \in \{1, 2, \ldots, n\}$. Intuitively, η_i can be regarded as the contribution of the individual i on the marginal estimator $\hat{\theta}$. Andersen et al. [2] proposed the use of these pseudo-values in the context of regression modeling of state occupation probabilities via generalized linear models and showed that this pseudo-value approach efficiently estimates the regression parameters. Suppose data consists of n pairs of independent and identically distributed pairs (X_i, Z_i), $1 \leq i \leq n$, of response X and covariates Z, and we are interested in estimating $\theta(Z) = E(f(X)|Z)$, for some known function f. Starting with an asymptotically linear and unbiased estimator $\hat{\theta}$ of the corresponding marginal parameter $\theta = E(f(X))$, Andersen et al. [2] proposed that one can regress the corresponding pseudo-values η_i on Z_i to obtain an estimator of the regression function $\theta(Z)$. In this article, we let f be the indicator function $I(U(t) = h)$ which denotes whether an individual is at state h at time t. With this choice, $P_h(t)$, the occupation probability of a certain state h at a given time t, becomes the parameter θ of interest.

2.2 Estimation and Regression of State Occupation Probability in Multistate Models

Aalen and Johansen [1] proposed a non-parametric estimator of the state occupation probability in multistate models with censored outcomes. They showed that, under independent censoring, $\hat{P}_h(t)$, the Aalen-Johansen estimate of occupation probability of state h at time t, is consistent for estimating the true occupation probability $P_h(t)$ if the underlying multistate process is Markov. Later on, Datta and Satten [17] showed that even if the underlying process is non Markov, the Aalen-Johansen estimator of state occupation probability remains consistent. The Aalen-Johansen estimator can be thought of as a generalization of the Kaplan-Meier estimator of survival probability in a two-state survival framework. Steps for computing the Aalen-Johansen estimator and related details are discussed in the Web Appendix A of online Supplementary Material (See Appendix).

Two existing approaches for regression and prediction of state occupation are the pseudo-value regression approach and the binomial regression approach [18]. The binomial regression approach only uses the binary indicator of an event status (e.g., occupation or non-occupation of the given state), at a given time point, as the outcome, unlike the pseudo-value approach that considers a marginal estimator like Aalen-Johansen estimator. For prediction of state occupation in a system with multiple states, the pseudo-value approach based on the Aalen-Johansen estimator is expected to be more efficient than the binomial regression approach. This is so because the Aalen-Johansen estimator uses information on full event history up to the time point of interest rather than relying just on the binary event status at the time point of interest. Therefore, we focus on extending the pseudo-value approach for regressing the state occupation probabilities in multistate models, when the covariate dimension is high.

Aalen-Johansen estimator is a very suitable marginal estimator of state occupation probability when the aim is to regress the state occupation probabilities based on the covariates. If one wants to predict the occupation probability of a typical state h at some future time t through the pseudo-value based regression approach as discussed earlier then the pseudo-values of state occupation probability of state h at time t can be generated as

$$\hat{P}_h^i(t)=n\hat{P}_h(t)-(n-1)\hat{P}_h^{(-i)}(t), i=1,2,\ldots,n,$$

where $\hat{P}_h^{(-i)}(t)$ is the Aalen-Johansen estimate of occupation probability of state h at time t calculated after removing the individual i from the data. Now, we can regress these pseudo-values on available covariates through a linear model or a generalized linear model as discussed before.

2.3 High Dimensional Regression

In case the number of covariates (q) available for each individual exceed the total number of individuals (n) under study, i.e. $n \ll q$, the standard linear or generalized linear models fail. Then we have to resort to one of the high dimensional regression techniques. Out of the different high dimensional regression techniques, we consider latent factor regression such as Partial Least Squares (PLS), Sparse Partial Least Squares (SPLS), and penalized regression methods such as LASSO, Elastic Net (ENET), and Adaptive LASSO (AdLASSO) in this article.

PLS is a latent factor regression technique which constructs a handful of latent variables from a collection of a large set of explanatory variables in such a way that most of the predictive power can be accomplished by these extracted latent variables. SPLS may lead to an improvement upon PLS as it performs latent factor extraction and variable selection simultaneously. SPLS imposes an additional L_1 constraint during the formation of PLS components that leads to the variable selection feature. Penalized regression techniques are widely used in case the covariate dimension is too large compared to the sample size. LASSO is one of the most popular penalized regression methods that introduce an L_1 penalty to the regression coefficients so that many coefficients shrink to zero. This is a very useful feature when there is an enormous list of covariates with most of them not contributing to the outcome of interest. However, LASSO has the tendency of shrinking most coefficients to zero in a set of correlated regressors which can lead to the elimination of many important covariates. ENET regression is an extension of LASSO that addresses this issue of over-shrinking. ENET uses a mixture of L_1 and L_2 penalty and is robust to the presence of highly correlated variable. AdLASSO regression is another extension of LASSO where different adaptive weights are used in penalizing different coefficients in the L_1 penalty. The advantage of the AdLASSO method is that the estimated regression coefficients have the oracle property which means that the penalized estimator of the coefficients is consistent in the parameter estimation and variable selection. This feature is lacked by the regular LASSO. Details on these high dimensional methods, including their computational steps and important features, can be found in the Web Appendix B of online Supplementary Material.

The development of the pseudo-value based high dimensional regression of state occupation probabilities can be regarded as the unification of all the different concepts discussed in this section, namely, estimation and regression of state occupation probabilities, and the high dimensional regression methods.

3. Simulation Studies

We now describe a number of simulation settings. Through the study of these varied simulation settings involving sparse and non-sparse scenarios, different censoring rates and noise-to signal ratios, we aim to find out which of the aforementioned pseudo-value based high dimensional regression methods is the most powerful for predicting future state occupation and survival. In the first setting, we have a multistate model framework where we compare the performances of different high dimensional regression methods such as PLS, SPLS, LASSO, Adaptive LASSO (AdLasso), and Elastic Net (ENET), based on pseudo-values. In the second setting, we have a survival (two-state) model where we compare the performance of the Cox model based LASSO regression with that of the PV based high dimensional regression techniques when the underlying true model is non-Cox type.

3.1 Simulation designs for a multistate model

We generate an irreversible three-stage illness-death model with censored outcomes where all the state-to-state transition times of a typical individual are generated from an accelerated failure time (AFT) models based on the covariate information of that individual. The three states in this illness-death model are the ‘disease-free’ state, ‘ill’ or ‘disease’ state, and the ‘death’ state, which are indexed as states 1, 2, and 3, respectively. Every individual starts from state 1 and can move into either state 2 or state 3. Once an individual leaves a state it cannot return to it. Also, no transition is possible from state 3 (absorbing state).

For a typical individual i, the transition time from the state h to the state k, T_{ihk} (say), is such that the $(\log(T)_{ihk})$ is generated from a linear model based on the available covariates Z_{i1}, Z_{i2}, …, Z_{iq}. For our illness death model, we generate the transition times as follows:

$$\log(T_{i12})=\sum_{j=1}^{q}\beta_j Z_{ij}+\in_{i12}=Z_i^T\beta+\in_{i12},\ \log(T_{i12})=\sum_{j=1}^{q}\gamma_j Z_{ij}+\in_{i13}=Z_i^T\gamma+\in_{i13}$$

and

$$\log(T_{i23})=\sum_{j=1}^{q}\gamma_j Z_{ij}+\in_{i13}=Z_i^T\gamma+\in_{i13},\ \text{provided} T_{i23}\geq Ti12,$$

where ϵ_{12} and ϵ_{13} are the error components, q is the number of available covariates,

$$Z_i^T=(Z_{i1},Z_{i2},\dots,Z_{iq}),\ \beta=(\beta_1,\beta_2,\dots,\beta_q)^T,\gamma=(\gamma_1,\gamma_2,\dots,\gamma_q)^T,$$

and $i = 1, 2, \dots, n$. For a typical i, if $T_{i13} < T_{i12}$, we ignore all other transition times as the individual has moved to the absorbing state at the very first transition. Otherwise, we repeatedly simulate T_{i23} till the condition $T_{i23} \geq T_{i12}$ is fulfilled. The transition times are generated so that the hazard of the second possible transition of an individual from state 2 to state 3 does not depend on the time of the first transition, and the resulting process is

Markov. Here we choose $n = 200$, $q = 10{,}000$. The regression coefficient parameters are chosen as one of the two following combinations:

$$
\begin{array}{lll}
(a) \qquad \beta_j = \begin{cases} 1, & \text{if} 1 \le j \le 50 \\ 0, & \text{otherwise} \end{cases}, & \gamma_j = \frac{1}{j}, & 1 \le j \le q \\
(b) \beta_j = 1, & \gamma_j = \frac{1}{j}, & 1 \le j \le q
\end{array}
$$

Case (a) corresponds to the situation where only a few (0.5%) of the total covariates actually contribute to the time of transition of an individual from state 1 to state 2. So this can be thought of as a sparse regression model for transition into the disease state. In case (b), all the available covariates contribute to the transition time from state 1 to state 2, which implies a non-sparse (dense) regression model for transition to the disease state. In both the cases, the number of covariates contributing to the transition to state 3 is neither too large nor too small. Let Z be the design matrix such that covariate vector for the i^{th} individual, namely Z_i, defines the i^{th} row of Z. We generate Z_i from a multivariate normal distribution with zero mean vector and variance covariance matrix Σ_Z. For our simulation we choose Σ_Z as an identity matrix. We generate both the errors ϵ_{12} and ϵ_{13} from a normal distribution with mean 0 and variance $r\sigma^2$. Here $\sigma^2 = max(\beta^T \Sigma \beta, \gamma^T \Sigma \gamma)$, where β and γ are normalized versions of regression coefficient vectors β and γ respectively, and r is a constant factor controlling the noise-to-signal ratio (NSR) of the simulated regression model. Two choices of NSR were considered, namely, 0.01 and 1.0. The censoring time, C_i for the i^{th} individual at each of the states 1 and 2, is generated from a lognormal distribution such that $\log(C_i) \sim N(c_0, \sigma_c^2)$ independently for $i = 1, 2, \ldots, n$. Here c_0 is determined by the overall censoring rate. We consider three different choices for the censoring rate, namely, 0% (no censoring), 35% (moderate censoring), and 80% (heavy censoring).

In order to directly predict the future state occupation based on the huge number of available covariates, we start with the Aalen-Johansen estimator as the marginal estimator for state occupation probability as discussed in Section 2.2. Here we focus on state 2 (illness state). So, in our simulation study we have directly modeled the state occupation probability of state 2 at a specific time t using the pseudo-value regression as outlined in section 2.2. We use different high dimensional regression methods discussed in Section 2.3 such as PLS, SPLS, LASSO, Adaptive LASSO, and Elastic net for this purpose. To get a complete picture on the performances of these high dimensional regression methods we vary the number of PLS or SPLS terms (latent factors) as well as the index of regularization (penalty) parameter in LASSO, AdLasso, and ENET. We obtain predicted values using different number of latent factors in PLS and SPLS regression, where the threshold tuning parameter (see Web Appendix B) for a fixed number of latent components in SPLS regression is obtained through cross-validation. Similarly for LASSO, ENET, and AdLasso, we compute the predicted values for the extensive solution path of the regularization parameter corresponding to the L_1 penalty (refer Web Appendix B). In addition, for ENET regression we consider four choices for the elastic net mixing parameter α (described in Web Appendix B), namely 0.2, 0.4, 0.6, and 0.8. For AdLasso we take the ridge regression estimate of the regression coefficient corresponding to minimum cross-validated error as the initial consistent estimator along with the three different choices of the weight tuning parameter γ, namely 0.5, 1 and 2 (as mentioned in Web Appendix B).

Performance measure—To evaluate the predictive performances of the PV based high dimensional regression methods, we can derive the theoretical (true) state occupation probability at a given time conditional on the covariates $Z_1, Z_2, \ldots, Z_q$, for the irreversible illness-death model described in our simulation settings. In this setting, $P_2(t; Z_i)$ denotes the true occupation probability of the i^{th} individual (conditional oncovariate vector Z_i) at state 2 at time t, and the detailed steps of deriving $P_2(t; Z_i)$ can be found in Web Appendix C. Let $\hat{P}_{2;i}(t; Z_i)$ denote the estimated value of the state 2 occupation probability at time t for the i^{th} individual using a PV based regression method. Then a measure of the predictive power of that PV based regression method can be given by the mean relative error of estimation

$$\text{MREE}=\frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{P}_{2;i}(t;Z_i)-P_2(t;Z_i)|}{P_2(t;Z_i)}\quad i=1,2,\ldots,n$$

Here lower values of *MREE* indicate better prediction power of the corresponding regression method. In addition, to compare the performances of high dimensional regression methods with that of a ‘no-covariate’ model, i.e., a model based on the marginal probabilities without any covariate information, we calculate the above measures for a no covariate model, where $\hat{P}_{2;i}(t; Z_i)$ is replaced by the marginal Aalen-Johansen estimate $\hat{P}_2(t)$ (ignoring the covariate information Z_i) for state 2 for all values of $i = 1, 2, \ldots, n$. For most parts of our simulation study, we choose the time point t as the median of all the first transition times obtained from the complete data on n individuals. We calculated all the *MREE* values of the pseudo-value based PLS, SPLS, LASSO, AdLasso, and ENET regression methods as well as that of the ‘no-covariate’ model by averaging over 50 independent Monte-Carlo runs of the previously described data set.

Results—First we consider the sparse setting, i.e., case (*a*). Table 1 presents the optimal *MREE* values for the different regression methods based on pseudo-values as well as that of a ‘no-covariate’ model under different censoring rates at NSR value 0.01. From Table 1 we find that the optimal (minimum) *MREE* value for each of PLS, SPLS, LASSO, AdLasso, and ENET is substantially less than the *MREE* value of the ‘no-covariate’ model. This implies that indeed the PV based high dimensional regression methods are much more effective compared to a marginal model in predicting state occupation when covariates are involved. Also from Table 1, we see that among the penalized regression methods, the LASSO performs the best for all types of censoring rates considered. But even then, the optimal values of PLS and SPLS regression is less than that of LASSO, with PLS regression emerging out to be the best in terms of having minimum *MREE* values overall. Also, we see that with the increase in the censoring rate in the data, the *MREE* values tend to increase, albeit not by a large margin. Figure 1 compares the performances of PLS, SPLS, LASSO, and ENET for a range of PLS/SPLS components and varying index of regularization parameter of LASSO/ENET/AdLASSO under the two different NSR values of 0.01 and 1.0 and a censoring rate of 80%, while Table 2 shows the optimal values of different regression methods based *MREE* on pseudo-values as wellas that of the no-covariate model under the high NSR of 1.0 and 80% censoring. Interestingly, with the increase in the NSR the performances of the PV based regression methods tend to deteriorate to such an extent that it is difficult to distinguish them from a no-covariate model, although the optimal values of the PV based methods are marginally better (lower) than that of the no-covariate model. Also, we see that the wide difference between the PLS-type and the LASSO-type regression methods disappears under high NSR. Web-Figure 1 displays the *MREE* values of different regression methods based on pseudo-values for all the components or values of the index of regularization parameter considered and varying rates of censoring present in the data. Since the marginal estimator of state occupation probability is a function of time, one may be interested in observing how the *MREE* for the PV based regression methods behave as a function of time. Web-Figure 2 shows the optimal *MREE* values of the PLS and LASSO regression based on pseudo-values as a function of the time at which the underlying Aalen-Johansen estimates are calculated.

Next we consider the case (*b*), the non-sparse regression scenario. Table 1 summarizes the optimal *MREE* values of these PV based regression techniques as well as that of a no-covariate model for different censoring rates at a NSR value of 0.01. Table 1 shows that all the PV based regression methods perform substantially better than the no-covariate model, and PLS regression has the minimum overall *MREE* values among all the PV based regression techniques, similar to the results found in case (*a*). But the most striking difference in case (*b*) is that the difference between optimal *MREE* values of PLS and LASSO increase to such an extent that the minimum *MREE* of PLS is around six times smaller than that of LASSO, whereas in case (*a*) the minimum *MREE* of PLS was only 1.7 times lower than that of LASSO. This signifies the fact in case of non-sparse (dense) regression scenario the PLS method is vastly superior to the penalized regression techniques in predicting state occupation. This is mainly because PLS regression, unlike the LASSO-type penalized regression, does not remove covariates that have some predictive powers. Next, we investigate the predictive performances of the PV based methods in this simulation setting when we increase the NSR to 1.0. Figure 2 and Table 2 show that as the NSR increases the performances of the PV based regression methods do get worse and the wide difference between the optimal *MREE* values of PLS and LASSO vanishes. However, the optimal *MREE* values are still better than the *MREE* of the no-covariate model. Web-Figure 3 displays the *MREE* values for all the components of PLS, SPLS and the values of the index of regularization parameter of LASSO, AdLasso, and ENET regression under different censoring rates when the NSR is 0.01.

3.2 Simulation designs for a survival model

In case of survival model, $S(t)$, the survival probability of an individual at time t, can be interpreted as $S(t) = P_1(t) = 1 - P_2(t)$, where $P_2(t)$ is the probability that individual is occupying state 2 at that time t. So the PV based regression method can be applied to directly model the survival probabilities based on large number of covariates and small sample sizes in presence of potential censoring. There exists a LASSO regression method based on Cox regression model in survival framework [16]. Note that, PLS regression can also be carried out under the assumption of an underlying Cox model. But, on trying to implement PLS with Cox model in our simulated data setting, the estimation of the model parameters failed due to the non-convergence of the algorithm. While Cox regression may be a natural choice for analyzing data that are known to follow proportional hazard assumption, it might be interesting to see how the PV based regression methods fare compared to the Cox model based LASSO method in survival prediction when the underlying survival model is not a Cox type model (i.e., the hazards are not proportional). For the purpose of comparison, we engage PLS as well as LASSO, based on pseudo-values, as these two high dimensional methods performed near the top in our simulation studies in Section 3.1. We choose the Kaplan-Meier estimator of survival probability as the marginal estimator and compute the pseudo-values based on this estimator in the same way as we do it for the Aalen-Johansen estimator of state occupation probability.

For carrying out this comparison we simulate a survival data with right censoring. The true event (transition from state 1 to state 2) times are generated via an AFT model based on the available covariates $Z_1, Z_2, \ldots, Z_q$. This results in a model that is not Cox-type. If T_i denote the true event time for the i^{th} individual, then we have $\log(T_i)=\sum_{j=1}^{q}\beta_j Z_{ij}+\in_i$, where ϵ_i is the error component, q is the covariate dimension, and $i = 1, 2, \ldots, n$. Here we choose n as 200, and q as 10,000. The regression coefficient vector β in the above mentioned regression model is chosen in one of the two following combinations:

$$
\begin{array}{lll}
(a) & \beta_j=\begin{cases} j\text{mod}5, & \text{if} 1\le j\le 100 \\ 0, & \text{otherwise} \end{cases}, & 1\le j\le q \\
(b)\beta_j=\frac{1}{j}, & & 1\le j\le q
\end{array}
$$

Case (a) corresponds to a highly sparse regression scenario where only a few (0.8%) of the total covariates actually contribute to the true event time. In case (b) the number of covariates contributing to the true event time is neither too large nor too small. We generate Z_i, the i^{th} row of the design matrix Z, from a multivariate normal distribution with zero mean vector and variance covariance matrix Σ_Z. For our simulation we choose Σ_Z as an identity matrix. We generate the errors ϵ_{12} from a normal distribution with mean 0 and variance $10\sigma^2$, where $\sigma^2 = \beta^T\Sigma_Z\beta$, where β is the normalized version of regression coefficient vectors β. Also, C_i, the right censoring time for the i^{th} individual, is generated such that $(\log(C)_i) \sim N(c_0, \sigma^2)$ independently for $i = 1, 2, \ldots, n$. Here c_0 is determined by the censoring rate. For this simulation we choose the censoring rate to be around 50%.

With the right censored data generated from the above simulation setting, we estimate the marginal survival probability at time t through the Kaplan-Meier estimator, and then calculate the pseudo-values of the survival probability for each of the n individuals. Now, with these pseudo values as responses we fit a regression model based on either LASSO or PLS regression technique, and predict the survival probabilities of the individuals from the fitted model. In addition, we separately fit a Cox proportional hazard model with LASSO-type (L_1) penalization on the simulated right censored data and again estimate the survival probabilities of all the individuals at time t using the fitted Cox-LASSO regression model with Breslow estimate of baseline survival.

Performance measure—Similar to the simulation study in Section 3.1, we derive the theoretical survival probabilities (assuming no censoring) at a given time conditional on the covariates $Z_1, Z_2, \ldots, Z_q$. If $S(t; Z_i)$ denote the theoretical survival probability at time t of the i^{th} individual with covariate information Z_i, then for the above mentioned simulation setting, we have $S(t;Z_i)=\left(1-\Phi\left(\frac{\log(t)-Z_i^T\beta}{\sigma}\right)\right)=S_i(t)$ (say). In addition, if $\hat{S}_i(t)$ denote the estimated survival probability of the i^{th} individual at time t, either from the pseudo-value based high dimensional regression or a Cox model based LASSO regression, then

$$\text{MREE}=\frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{S}_i(t;Z_i)-S_i(t;Z_i)|}{S_i(t;Z_i)}$$

Results—The performances of the different methods in simulation scenarios (*a*) and (*b*) are displayed in Figure 3. In regression scenario (*a*), the optimal (minimum) *MREE* values of Cox-LASSO, pseudo-value based LASSO, and pseudo-value based PLS regression are obtained as 0.5117, 0.4633, and 0.1241 respectively. For regression scenario (*b*), the optimal *MREE* of Cox-LASSO, pseudo-value based LASSO, and pseudo-value based PLS regression are 0.8826, 0.9577, and 0.7305 respectively. In case of the highly sparse regression scenario (*a*), the minimum *MREE* value for the pseudo-value based LASSO is lower than that of a Cox model based LASSO. It is the other way around in case (*b*) when the underlying model is less sparse. In both cases, however, the pseudo-value based PLS regression has the least minimum *MREE* value amongst the three.

4. Applications

4.1 Michigan Lung Cancer Data

We demonstrate the use of pseudo-value based PLS and LASSO regression methods in predicting patient survival using a Michigan lung adenocarcinoma data set which was originally analyzed by Beer et al.[19]. The original data set had 7129 gene expressions for 86 lung tumor samples and 10 normal tissue samples. Genes with extremely low levels of expressions were excluded from the final data set. The remaining 4966 genes were used for dividing the 86 lung cancer patients into three clusters by hierarchical clustering. In the original study, Beer et al. [19] found that these three clusters showed significant differences based on tumor stage and tumor differentiation. That study intended to investigate the relationships between clusters, cancer stages, gene differentiation and overall survival, details of which can be found in [19].

In this article we use the 4966 gene expressions obtained from the original study along with the survival times, survival indicator, information on tumor status (either stage 1 tumor or stage 3 tumor) and gender information of the 86 lung cancer patients to demonstrate the pseudo-value based prediction of patient survival in presence of a high dimensional covariate. There are 67 patients with stage 1 tumor whereas there are 19 patients with stage 2 tumor. From the full data of 86 lung cancer patients, we estimate the overall survival probability at a given time t (in months) by the Kaplan-Meier estimator. We predict the survival at time t of each of the 86 patients based on his or her gene expression profile, irrespective of the censoring status. For this we use the PV based PLS and LASSO regression where the pseudo-values are based on the Kaplan-Meier estimator as described in details in Section 3. The data under study has 70% of censored observations.

Unlike in simulation studies, we do not have the true (theoretical) survival probability at any given time for any of the patients. At a given time t, the only information we have is that whether a patient is alive and under study at that time, or is dead at some time before t, or is censored at some time before t. We use the survival status of the set of individuals who are alive at time t to tune the regression model parameters, while we use the survival status of the set of patients who are known to be dead by time t to check the predictive power of the optimally tuned regression models at time t. Thus, for choosing the optimal number of PLS components or optimal index of regularization parameter in LASSO, we calculate the following data-based measure of mean absolute error of prediction

$$\mathrm{MAEP}=\frac{1}{n_R(t)}\sum_{i=1}^{n}\delta_i(t)|\hat{S}_i(t;Z_i)-1|$$

where $\delta_i(t) = 1$ if the i^{th} patient is alive and under study at time t, and $\delta_i(t) = 0$ otherwise, $n_R(t)=\sum_{i=1}^{n}\delta_i(t)$, and $\hat{S}_i(t;Z_i)$ is the estimated survival probability at time t for the i^{th} patient using PV based regression. Note that *MAEP* measures average absolute error of fit for the state prediction amongst subjects who are still known to be alive at the time point under consideration. The regression model having the minimum *MAEP* value at a given time t is chosen as the optimal model.

Next we test the classification ability of our optimally tuned model when applied to the individuals who are dead by time t. In other words, we check for individuals who are already dead by the time t, how well this could have been predicted from their baseline covariates by using the optimal PV based regression model. For better interpretation of the estimated survival probability as an indicator of the survival status, we classify a typical individual i as 0 or 1 based on whether the estimated survival probability $\hat{S}_i(t;Z_i)$ is less than 0.5 or not, respectively. We define $D_i(t) = 0$ if $\hat{S}_i(t;Z_i) < 0.5$ and $D_i(t) = 1$ if $\hat{S}_i(t;Z_i) \geq 0.5$, implying that the patient i is predicted to be more likely to be dead than alive at time t if $D_i(t) = 0$. In that case a measure of the misclassification rate for the set of patients already known to be dead at time t can be obtained as

$$\mathrm{MR}=\frac{1}{n_D(t)}\sum_{i=1}^{n}\alpha_i(t)D_i(t)$$

where $\alpha_i(t) = 1$ if the i^{th} patient is known to be dead by time t, and $\alpha_i(t) = 0$ if the i^{th} patient is not known to be dead by time t, $n_D(t)=\sum_{i=1}^{n}\alpha_i(t)$. It is easy to see that $0 \leq MR \leq 1$, where 0 denotes the case of no misclassification while 1 represents the case of maximum misclassification.

We have calculated the $MAEP$ values for different values of the index of regularization parameter of LASSO and different PLS components for a wide range of the values of time t. For the choice of t as 30 months, the minimum value of $MAEP$ using LASSO is 0.0315, while the minimum value of $MAEP$ using PLS is 0.0477 corresponding to a PLS regression with 7 components. Details of the performances of LASSO and PLS regression at 30 months can be obtained from Web-Figure 4. Next, we check how well the optimal PV based regression model at a given time point t perform in terms of the misclassification rate (MR) for the individuals already dead by time t. Interestingly, for the optimal PV based regression model at a given time t obtained by minimizing $MAEP$, the MR value turns out to be 0 implying perfect classification, and this is true for all the choices of t considered in our analyses. This indicates that, indeed, the optimal regression model having the minimum $MAEP$ value at a given time t, perfectly identifies the individuals who have died before time t.

One interesting question is whether there is any difference in the survival chances based on tumor status. For this we predict the survival at time t for each of the 86 patients through both the PLS and LASSO regression based on pseudo-values where the optimal number of PLS components and the optimal value of the index of regularization of LASSO are obtained by the procedure described in the last paragraph. Then we classify the patients according to their tumor status, namely stage 1 and stage 3, and then take the average of the estimated probabilities in each of the two groups. We repeat this for different choices of t and plot these average survival probabilities of stage 1 and stage 3 tumors as a function of time as shown in Web-Figure 5. It can be seen that at any time point t the average survival probability of a stage 3 tumor patient ismuch less than that of a stage 1 tumor patient. So tumor status does play a differentiating role in overall patient survival. Web-Figure 6 compares the average predicted survival probability of the male population with that of female population at different points of time. At initial time points there appear to be no significant differences between the average survival probabilities of the male and the female population, but the male and female survival do differ at later time points. Due to the censoring present in the data the main challenge is to get survival information of the patients who have already been censored before the time point of interest. We demonstrate the use of pseudo-value regression in estimating the survival probability of a patient at time point later than its censoring time. Figure 4 shows the temporal estimated survival probability of a lung cancer patient who was censored at 28.3 months. Survival probabilities are estimated using both LASSO and PLS regression methods based on pseudo-values where the optimal value of the regularization index of LASSO or the optimal number of PLS components is chosen in the same way of minimizing $MAEP$. We can see that the estimated survival probability of this patient goes on decreasing as time increases which is consistent with the general nature of survival function. Due to the lack of adequate methods for obtaining exact confidence intervals for LASSO based estimates, we rely on bootstrap percentile based confidence interval (CI) for the estimated survival of a patient at a given time. We calculate the 95% confidence intervals based on bootstrap percentile method (considering 1000 bootstrap resamples). We compute these confidence intervals of the estimated survival probabilities at multiple time points for the patient censored at 28.3 months. Using the LASSO based PV regression we obtain the bootstrap based 95% CI for survival at time points 12, 36, and 48 as (0.619, 1.000), (0.516, 1.000) and (0.502, 1.000), respectively. Using PLS based PV regression the bootstrap based 95% CIs for survival at time points 12, 36, and 48 turn out to be (0.618, 1.000), (0.445, 1.000) and (0.405, 1.000), respectively. Although the bootstrap based confidence intervals appear to be wide in this case, these are the only CI that can give some idea on the variability of the LASSO based estimates of survival probabilities.

5. Discussion

The pseudo-value method allows direct prediction of future state occupation instead of indirect modeling through state-to-state transition hazards even when censoring is present. This is particularly useful when the main objective is to interpret the estimated probability that an individual is in a particular stage of a multistate disease process at a given time in terms of the covariate profile of that individual. When the dimension of the covariate profile (e.g. gene expression profile) exceeds the underlying sample size, one can use latent factor regressions such as PLS regression or penalized regression such as LASSO regression in conjunction with the pseudo-value approach. Through extensive simulation studies, we have seen that, among the various high dimensional regression techniques that we considered, overall PLS works the best with the pseudo-value based responses for predicting future state occupation or survival. In cases of underlying sparsity, where a majority of the available covariates are noise variables not contributing to the state occupation or survival probabilities, the pseudo-value based LASSO regression is a powerful alternative to PLS regression for prediction purposes. Even in case of simple survival (two-state) model with a huge covariate dimension, the pseudo-value based regression methods seem to work better than the Cox model based penalized regression for predicting survival when the proportional hazards assumption is violated.

We have demonstrated the use of pseudo-value based high dimensional regression using a lung cancer data set which had a high proportion of censored samples. We employed pseudo-value regression using PLS as well as LASSO regression in predicting patient survival at a given time. Overall, meaningful and consistent results were obtained on patient survival, e.g., differentiation based on tumor stages.

This article is mainly motivated by the question of forecasting the state occupation or survival probability of a typical patient at some future time point based on its high dimensional covariate profile. Another interesting task can be finding out which of the available covariates (genes in case of gene expression profiles) are most significant in predicting state occupation or survival. But as state occupation probabilities are functions of time, one may have different optimal pseudo-value based regression models at different time points leading to the possibility of non-uniformity in the list of significant covariates over time. For example, some genes may turn out to be significant at two widely different time points but insignificant in the intermediate time points. Such results may be difficult to interpret from a biological perspective, and we plan to focus on this need of using the pseudo-value based high dimensional regression techniques for variable selection in future studies.

In our current work, as well as most of the past works based on pseudo-value regression, it has been assumed that the censoring mechanism is independent of the covariates under study. This assumption can be relaxed and a recent work in this direction [20] suggests using a correctly specified regression model for the censoring time in a competing risk framework where the covariate dimension is smaller than the sample size. However, extension of this approach to high dimensional settings, especially in the presence of huge covariate dimensions consisting of omic expression profiles, is not straightforward as it would be challenging to specify the correct model for the censoring time based on the high dimensional genomic or proteomic covariates. So, the idea of including covariate dependent censoring in the temporal prediction of state occupation based on high dimensional baseline covariates needs separate attention in future studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the NCI/NIH under Grant CA170091-01A1. We thank the reviewers for their constructive comments that helped to improve the article.

References

1. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scand J Statist. 1978; 5:141–150.
2. Andersen PK, Klein JP, Rosthøj S. Generalised linear models for correlated pseudoobservations, with applications to multi-state models. Biometrika. 2003; 90:15–27.
3. Miller RG. The jackknife-a review. Biometrika. 1974; 61:1–15.
4. Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. Lifetime Data Anal. 2009; 15:241–255. [PubMed: 19051013]
5. Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. Lifetime Data Anal. 2004; 10:335–350. [PubMed: 15690989]
6. Andersen PK, Klein JP. Regression Analysis for Multistate Models Based on a Pseudo-value Approach, with Applications to Bone Marrow Transplantation Studies. Scand J Statist. 2007; 34:3–16.
7. Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics. 2005; 61:223–229. [PubMed: 15737097]
8. Klein JP, Logan B, Harhoff M, et al. Analyzing survival curves at a fixed point in time. Stat Med. 2007; 26:4505–4519. [PubMed: 17348080]
9. Mogensen UB, Gerds TA. A random forest approach for competing risks based on pseudo-values. Stat Med. 2013; 32:3102–3114. [PubMed: 23508720]
10. Wold, H. Estimation of principal components and related models by iterative least squares. In: Krishnaiaah, PR., editor. Multivariate Analysis. New York: Academic Press; 1966. p. 391-420.
11. Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. Technometrics. 1993; 35:109–135.
12. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996; 58:267–288.
13. Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Series B Stat Methodol. 2010; 72:3–25. [PubMed: 20107611]
14. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005; 67:301–320.

15. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006; 101:1418–1429.

16. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997; 16:385–395. [PubMed: 9044528]

17. Datta S, Satten GA. Validity of the Aalen–Johansen estimators of stage occupation probabilities and Nelson–Aalen estimators of integrated transition hazards for non- Markov models. Statist Probab Lett. 2001; 55:403–411.

18. Scheike TH, Zhang MJ. Direct modelling of regression effects for transition probabilities in multistate models. Scand J Statist. 2007; 34:17–32.

19. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med. 2002; 8:816–824. [PubMed: 12118244]

20. Binder N, Gerds TA, Andersen PK. Pseudo-observations for competing risks with covariate dependent censoring. Lifetime Data Anal. 2014; 20:303–315. [PubMed: 23430270]

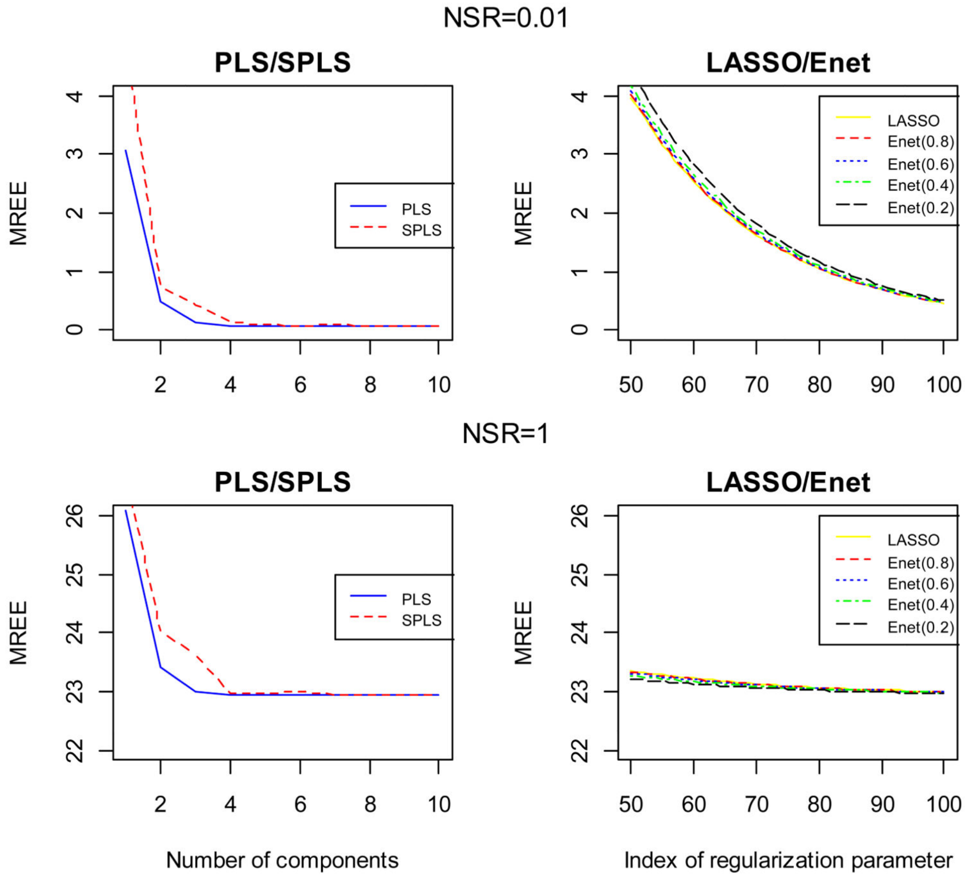

Figure 1.

The values of different pseudo-value based *MREE* regression methods under low and high NSR values and 80% censoring for sparse regression scenario (*a*) of illness-death model from Section 3.1

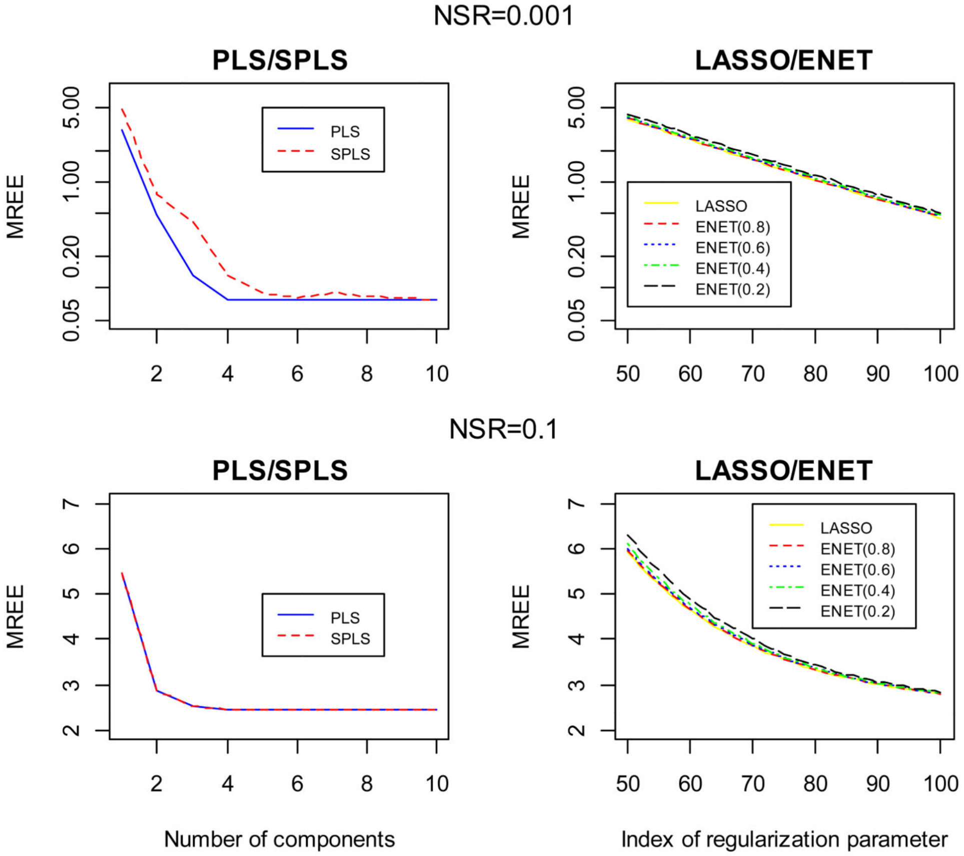

Figure 2.

The values of different pseudo-value based *MREE* regression methods under low and high NSR values and 80% censoring for non-sparse regression scenario (*b*) of illness-death model from Section 3.1

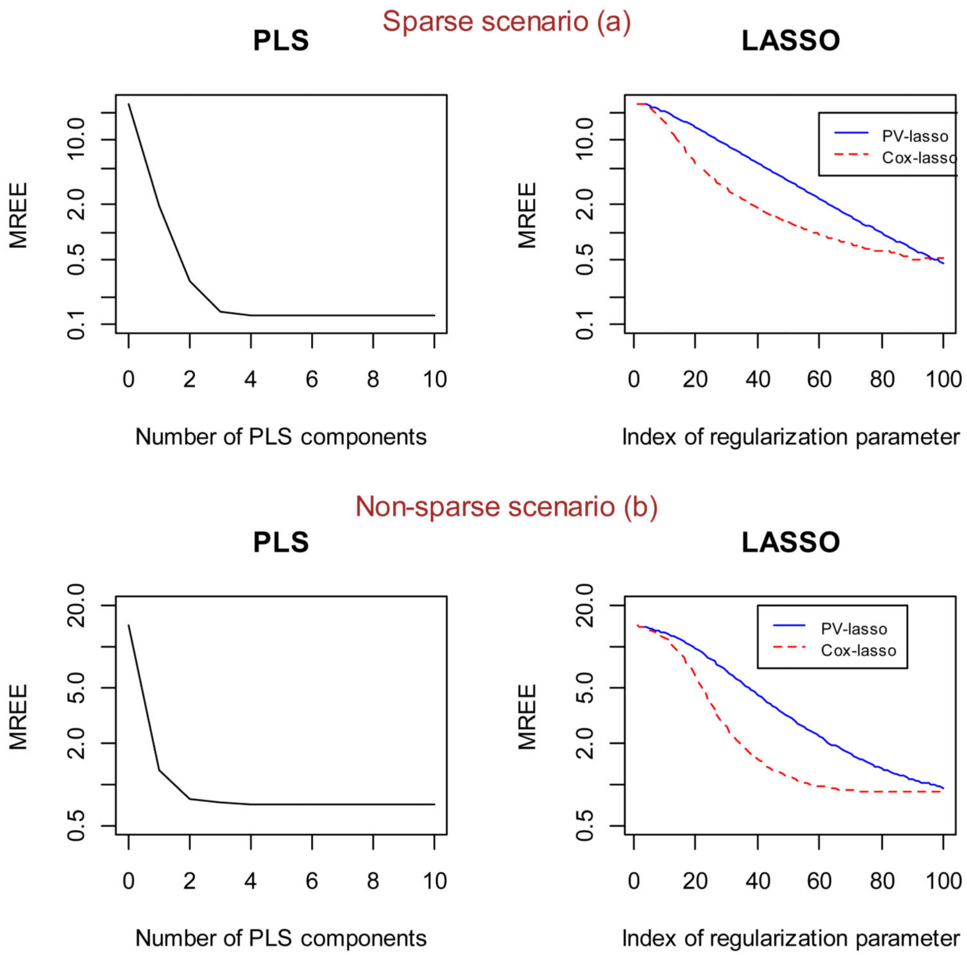

Figure 3.

The *MREE* values for Cox-LASSO, Pseudo-value (PV) LASSO and pseudo-value PLS for survival models with sparse scenario (*a*) and the non-sparse scenario (*b*) from Section 3.2.

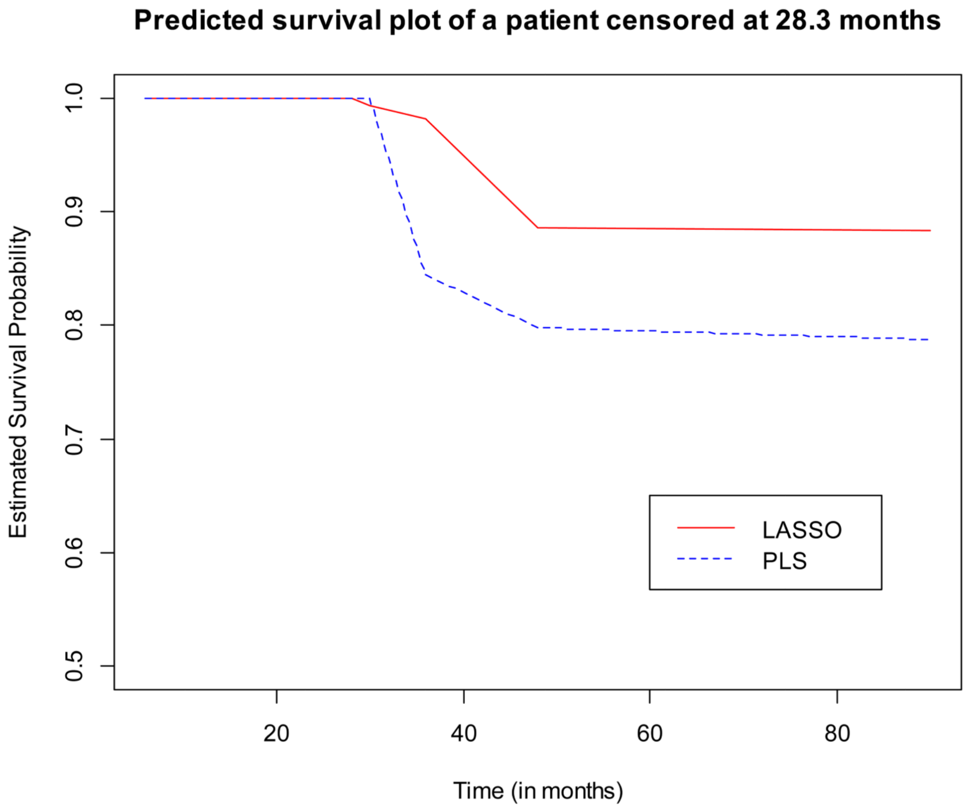

Figure 4.

Predicted survival (using both LASSO and PLS methods) of a patient in the Michigan Lung cancer study who was actually censored at 28.3 months of the study.

Table 1

Minimum values for different pseudo-value *MREE* based regression as well as the *MREE* of a no-covariate model under different censoring rates and 0.01 NSR in the sparse regression scenario (*a*) and the non-sparse scenario (*b*) from the illness-death model in Section 3.1

Type of regression	**Sparse scenario(*a*)**			**Non-sparse scenario(*b*)**		
	Censoring rate			**Censoring rate**		
	0%	**35%**	**80%**	**0%**	**35%**	**80%**
PLS	0.4070	0.4116	0.5085	0.0043	0.0178	0.0765
LASSO	0.7565	0.7768	0.8524	0.3548	0.3969	0.4630
ENET (0.8)	0.7582	0.7779	0.8539	0.3603	0.3995	0.4661
ENET (0.6)	0.7596	0.7782	0.8560	0.3760	0.4049	0.4717
ENET (0.4)	0.7636	0.7825	0.8603	0.3899	0.4146	0.4821
ENET (0.2)	0.7789	0.7966	0.8767	0.4187	0.4406	0.5080
AdLasso (0.5)	0.9289	0.9409	0.9623	0.5047	0.5329	0.6022
AdLasso (1.0)	1.2131	1.2343	1.2988	0.7094	0.7490	0.8230
AdLasso (2.0)	2.0903	2.1131	2.1481	1.3869	1.4426	1.5402
No-covariate	21.3256	21.3386	21.9483	24.3330	24.6635	25.0461

Table 2

Minimum values for different pseudo-value *MREE* based regression as well as MREE of a no-covariate model under a high NSR of 1.0 and 80% censoring rate for the sparse regression scenario (*a*) and the non-sparse scenario (*b*) from the illness-death model in Section 3.1

Type of regression	**Sparse scenario(*a*)**	**Non-sparse scenario(*b*)**
PLS	22.9332	2.4591
LASSO	22.9922	2.8076
ENET (0.8)	22.9886	2.8096
ENET (0.6)	22.9839	2.8138
ENET (0.4)	22.9780	2.8213
ENET (0.2)	22.9686	2.8428
No-covariate	24.0535	24.1069