Published in final edited form as:

Biometrics. 2011 December ; 67(4): 1379–1388. doi:10.1111/j.1541-0420.2011.01587.x.

On Path Restoration for Censored Outcomes

Brent A. Johnson*,

Department of Biostatistics and Bioinformatics, Emory University, Atlanta GA 30322, USA

Qi Long**, and

Department of Biostatistics and Bioinformatics, Emory University, Atlanta GA 30322, USA

Matthias Chung*

Department of Mathematics, Texas State University, San Marcos TX 78666, USA

Summary

Dimension reduction, model and variable selection are ubiquitous concepts in modern statistical science and deriving new methods beyond the scope of current methodology is noteworthy. This note briefly reviews existing regularization methods for penalized least squares and likelihood for survival data and their extension to a certain class of penalized estimating function. We show that if one's goal is to estimate the entire regularized coefficient path using the observed survival data, then all current strategies fail for the Buckley-James estimating function. We propose a novel two-stage method to estimate and restore the entire Dantzig-regularized coefficient path for censored outcomes in a least-squares framework. We apply our methods to a microarray study of lung adenocarcinoma with sample size $n = 200$ and $p = 1036$ gene predictors and find 10 genes that are consistently selected across different criteria and an additional 14 genes that merit further investigation. In simulation studies, we found that the proposed path restoration and variable selection technique has the potential to perform as well as existing methods that begin with a proper convex loss function at the outset.

Keywords

Accelerated failure time; Boosting; Dantzig selector; Lasso; Penalized least squares; Rank regression; Survival analysis

1. Introduction

In regression analyses of gene expression data, one often observes a large number of candidate genes relative to the sample size and the goal is to summarize the data in a meaningful way. These issues arise in cancer microarray studies and, here, we deal with the problem where the sample size n is 200 or smaller and the number of candidate predictors p is on the order of several thousand but not in the millions. This paper discusses regularized estimation of coefficient paths as well as selecting optimal point estimates on such paths, topics which have received considerable attention over the last decade (cf. Hastie et al., 2001; Efron et al., 2004; Candes and Tao, 2007). Here, we describe a popular semi-parametric regression estimator for censored outcomes (Buckley and James, 1979) in an accelerated lifetime model whose regularized coefficient path cannot be estimated using standard extensions of homotopy (Osborne et al., 2000), least angle regression (Efron et al.,

* bajohn3@emory.edu

** qlong@emory.edu

* mc85@txstate.edu

2004), corrector-predictor methods (e.g. Nocedal and Wright, 2006; Park and Hastie, 2007), coordinate-wise regression (Friedman et al., 2007; Friedman et al., 2010) or any other technique for linear programming (e.g. Candes and Tao, 2007), quadratic programming (Tibshirani, 1996, 1997) or convex optimization (Nocedal and Wright, 2006). In this paper, we favor a two-stage estimation approach to restoring the regularized coefficient path. The caveats in regularized estimation for this particular high-dimensional censored data problem apply to other missing data problems as well.

Assume the logarithm of the survival endpoint Y_i for the i -th subject is linearly related to p predictors $\{x_{ij}\}_{j=1}^p$ through the accelerated failure time (AFT) model,

$$Y_i = \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i, \quad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a p -vector of unknown regression coefficients and $\varepsilon_1, \dots, \varepsilon_n$ are independent errors from a common distribution function F_ε . The complete data are $\mathbf{Z}_{\text{comp}} = \{(Y_i, \mathbf{x}_i), i = 1, \dots, n\}$, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, whereas the observed data are $\mathbf{Z}_{\text{obs}} = \{(U_i, \delta_i, \mathbf{x}_i), i = 1, \dots, n\}$, with $U_i = \min(Y_i, C_i)$, $\delta_i = I(Y_i \leq C_i)$ and $I(\cdot)$ denotes the indicator function. Without loss of generality, we assume that the columns of the design matrix $\mathbf{X}$ have mean zero and unit L_2 -norm. One method of high-dimensional regression is through a penalized least squares framework. If $l_n(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ with $\|\cdot\|_2$ denoting L_2 -norm, then the regularized coefficient estimate minimizes the penalized L_2 -loss,

$$l_n(\boldsymbol{\beta}, \lambda) = l_n(\boldsymbol{\beta}) + J_\lambda(\boldsymbol{\beta}), \quad (2)$$

where $J_\lambda(\boldsymbol{\beta})$ a penalty function and λ a regularization parameter. Least absolute shrinkage and selection operator (lasso; Tibshirani, 1996) imposes L_1 constraints on the regression coefficients and its Lagrangian equivalent leads to (2) with $J_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$. For fixed regularization parameter λ , numerically solving the lasso problem is a quadratic programming problem. For λ sufficiently large, say $\lambda_{\max}$, all coefficient estimates (2) are set exactly to zero and λ sufficiently small, say $\lambda_{\min}$, results in a saturated fit (Efron et al., 2004; Park and Hastie, 2007; Friedman et al., 2007; Friedman et al., 2010). In the case where $n \gg p$, $\lambda_{\min} = 0$ and the saturated fit is the ordinary least squares fit with all p predictors. The entire coefficient path is the trace of regularized coefficient estimates over the index set $\Lambda_1 = \{\lambda \mid \lambda_{\min} \leq \lambda \leq \lambda_{\max}\}$,

$$\mathcal{P}_1(\mathbf{Z}_{\text{comp}}) = \{\widehat{\boldsymbol{\beta}}(\lambda), \lambda \in \Lambda_1 \mid \widehat{\boldsymbol{\beta}}(\lambda) = \arg \min l_n(\boldsymbol{\beta}, \lambda)\}. \quad (3)$$

The computational aspects of path estimation are themselves problems of independent interest, especially in statistical learning and computation.

Osborne et al. (2000) showed that entire L_1 -regularized coefficient path was piece-wise linear. In this case, the coefficient path is an organized collection of line segments and a description of how and where those line segments are joined. Efron et al. (2004) showed how to estimate the entire L_1 -regularized coefficient path efficiently. Friedman et al. (2007) reviewed how one can estimate the entire L_1 -regularized coefficient path through

coordinate-wise optimization. Path estimation is advantageous in practice because it simplifies parameter tuning and can mitigate user-induced round-off errors when an exact solution exists. The basic techniques are particularly useful in high dimensions because coefficient estimation is still possible even when estimation in the full model is not. When least squares loss $l_n(\beta)$ in (2) is replaced with other convex loss functions, several authors (Rossett and Zhu, 2004; Park and Hastie, 2005; Friedman et al., 2010) have noted that ordinary convex optimization techniques (cf. Nocedal and Wright, 1999) may be used to estimate the entire coefficient path even when the path is not piece-wise linear.

For semi-parametric modelers, an intriguing question is whether basic path estimation techniques extend to M - and Z -estimators, say, the solution to $0 = \Psi_n(\beta)$. Fu (2003) considered L_1 -regularized generalized estimating equations (GEE) and showed that solutions to smooth estimating equations could be well-defined and coefficient path construction follows straightforwardly. Now, consider the Buckley-James (1979) estimating equations, given by

$$0 = \Psi_n(\beta) = \sum_{i=1}^n \mathbf{x}_i^T \{\widehat{Y}_i(\beta) - \mathbf{x}_i^T \beta\}, \quad (4)$$

with synthetic response,

$$\widehat{Y}_i(\beta) = \delta_i Y_i + (1 - \delta_i) \left[\mathbf{x}_i^T \beta + \left\{ \widehat{S}_\varepsilon(e_i) \right\}^{-1} \int_{e_i}^{\infty} \widehat{S}_\varepsilon(t) dt \right], \quad (5)$$

$e_i = U_i - \mathbf{x}_i^T \beta$ and $\widehat{S}_\varepsilon(t)$ is the Kaplan-Meier estimator of $S_\varepsilon(t) = 1 - F_\varepsilon(t)$ based on the residuals $\{(e_i, \delta_i)\}_{i=1}^n$. From (4), we see that the Buckley-James estimator is the solution to the normal equations with imputed responses (5) for the censored observations. However, unlike Fu's (2003) setup, the estimating equations in (4) are discrete and his methods do not apply. Moreover, no simple extension of ordinary path-based estimation will recover the complete-data coefficient path because the expression in square brackets in (5) is only an unbiased estimator for $E(Y_i | Y_i > C_i, \beta)$ under the true model (1) with $\beta \equiv \beta_0$. Most estimators of the entire regularized coefficient path (Efron et al., 2004; Park and Hastie, 2007; Friedman et al., 2007; Friedman et al., 2010) initialize with $\beta = 0$ which leads to the synthetic response for censored observations, $E(Y_i | Y_i > C_i)$, and ignores information in the predictors. Sequential estimators of the coefficient path (cf. Friedman et al., 2007; Friedman et al., 2010) update the synthetic response with the current regularized iterate but what we really desire is a consistent estimate of the true coefficient β_0 at the initial step.

Restoring the entire regularized coefficient path $\mathcal{P}(\mathbf{Z}_{\text{comp}})$ using the the observed data $\mathbf{Z}_{\text{obs}}$ for missing data problems has received scant attention in the literature. In high-dimensional Buckley-James regression, our idea is to view the imputation and coefficient path estimation as completely separate problems. Separating the “imputer” from the “analyst” is an old idea advocated by missing data experts like Professor Donald Rubin (cf. Rubin, 1976). When the number of predictors is small and sample size large, Johnson (2009a) investigated a related lasso-type estimator where imputation is based on a consistent estimator from the full model. Here, the challenge is to find the best strategy when the number of predictors exceeds the sample size and the coefficients in the full model can no longer be estimated easily. We propose a new strategy of restoring the entire regularized coefficient path for

censored outcomes in Section 2 and variable selection procedure in Section 3. An analysis of microarray data is presented in Section 4 and simulation studies in Section 5.

2. Methods

2.1 Dantzig estimation

In this paper, we focus on regularized extensions of semiparametric estimators defined as solutions to estimating equations, $0 = \Psi_n(\boldsymbol{\beta})$, as in (4). Here, a rather direct definition of regularized estimator follows from the Dantzig estimation (Candes and Tao, 2007). In the least squares case where $\Psi_n(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} - \mathbf{X}^T \mathbf{Y}$, the Dantzig selector (DS) is the solution to the following linear programming problem,

$$\min_{\boldsymbol{\beta}} \|\mathbf{x}^*(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\|_{\infty} \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq \lambda, \quad (6)$$

where $\mathbf{x}^*$ is generic notation to indicate that the supremum is taken over each element of the linear system. The entire Dantzig coefficient path is denoted,

$$\mathcal{P}_{\text{DS}}(\mathbf{Z}_{\text{comp}}) = \{\widehat{\boldsymbol{\beta}}(\lambda), \lambda \in \Lambda_{\text{DS}}\} \quad \widehat{\boldsymbol{\beta}}(\lambda) = \arg \min \text{ of (6)},$$

where Λ_{DS} is the index set for the Dantzig path. In high-dimensional sparse linear models, i.e. models where most coefficients are truly zero, Candes and Tao (2007) provide non-asymptotic sharp bounds for DS coefficient estimates and show they are within a factor of $\log(p)$. At the time of this writing, Dantzig estimation has been extended to generalized linear models (James and Radchenko, 2009) and Cox's partial likelihood estimator (Antoniadis et al., 2009). The following idea for high-dimensional regularized Buckley-James estimation was initially discussed in an Emory University Technical Report by the first author (Johnson, 2009c). During the review process, an anonymous referee kindly informed us of parallel related work by Yi Li and colleagues.

2.2 Frequentist Restored Dantzig Selector (FRDS)

If $\boldsymbol{\beta}_0$ were known, then $\hat{Y}_i(\boldsymbol{\beta}_0)$ is the best possible imputed response in the sense that it minimizes the expected L_2 -distance between the observed and imputed response for any synthetic response in the class of censoring unbiased transformations. For this reason, the Buckley-James transformation in (5) with $\boldsymbol{\beta} \equiv \boldsymbol{\beta}_0$ is called the “best restoration.” Because the true value is unknown, a natural suggestion when $n \gg p$ is to replace it with a root- n consistent initial value fit to all p predictors (Johnson, 2009a). In high-dimensional regression, no ordinary root- n consistent estimator exists and a new strategy is needed. Here, we consider two different strategies for initializing the Buckley-James estimator.

- **“A” Initialization.** Initialize with a *regularized* estimate $\hat{\boldsymbol{\beta}}^\gamma$ in the AFT model (1) fit to all p predictors, where γ is a tuning parameter. Then, impute the censored observations $\hat{Y}(\hat{\boldsymbol{\beta}}^\gamma)$ via (5) and an optimal value of the tuning parameter $\hat{\gamma}$.
- **“B” Initialization.** Select a subset $\mathcal{S}_n$ of important variables using an auxiliary procedure; compute an *unregularized* estimator, say $\hat{\boldsymbol{\beta}}_{\mathcal{S}_n}^{\text{Un}}$, on this selected subset, then compute the imputed censored observations $\hat{Y}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_n}^{\text{Un}})$ via (5).

Our A initialization uses threshold gradient descent (TGD; Friedman and Popescu, 2004) on the Gehan loss function,

$$l_n(\beta) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \delta_i(e_j - e_i) I(e_j \geq e_i), \quad (7)$$

and the tuning parameter γ pertains to a target number of updates and $\hat{\gamma}$ is chosen via 5-fold cross-validation using (7). This TGD technique approximates the L_1 -regularized Gehan estimate (Friedman and Popescu, 2004), which can also be found using linear programming (cf. Johnson, 2008, 2009b; Cai et al., 2009). It is known that the regularized estimator $\hat{\beta}^\gamma$ shrinks closer to zero as the number of predictors diverges to infinity. Thus, A initialization may not perform well in high dimensions but retain it for consistency with earlier drafts of this manuscript.

The B initialization is based on the heuristic that a good imputation procedure does not need to be an optimal model selection procedure. The only requirement of the auxiliary imputation procedure is that it provides reliable and reproducible prediction, and this goal can often be achieved with only a small subset $\mathcal{S}_n$ of all p predictors. In this paper, we investigate two strategies for defining the subset $\mathcal{S}_n$:

- “B₁” initialization: $\mathcal{S}_n$ is the active set, $\{j: |\hat{\beta}_j^\gamma| > 0\}$, where $\hat{\beta}^\gamma$ is a high-dimensional regression fit in the AFT model from “A” initialization;
- “B₂” initialization: $\mathcal{S}_n$ is the active set after high-dimensional regression fit in Cox’s proportional hazards model.

In both B₁ and B₂ initialization, the unregularized estimates are always computed by minimizing Gehan loss in (7) on the subset of predictors $\mathcal{S}_n$. The only difference between B₁ and B₂ initialization is the underlying model assumption: B₁ operates under the semiparametric AFT model in (1) while B₂ operates under a possibly misspecified model. A referee suggested that subset choice could be constructed from parametric models if the actual coefficient estimates were ultimately ignored. Based on similar reasoning, we could easily fit a high-dimensional proportional hazards regression model using existing software and select its active set. Even if the proportional hazards model were incorrect, the subset $\mathcal{S}_n$ may be good enough as long as coefficient estimates for active variables were non-zero. The proposed B₂ initialization coincides with the referee’s suggestion when the errors in (1) are extreme-value distributed.

3. Operating Characteristics

3.1 Outline

In this section, we outline the rationale of our approach to high-dimensional path restoration. Let $\mathcal{P}_\bullet(\mathbf{Z}_{\text{comp}})$ be general notation for a regularized coefficient path constructed from complete data, using constrained and unconstrained loss functions, $l_n(\beta, \lambda)$ and $l_n(\beta)$, respectively. Let $\mathcal{P}_\bullet^R(\mathbf{Z}_{\text{obs}})$ be the reconstructed coefficient path from the observed data $\mathbf{Z}_{\text{obs}}$ using the reconstructed constrained and unconstrained loss functions, $l_n^R(\beta, \lambda)$ and $l_n^R(\beta)$, respectively. We define $\hat{\beta}^R(\lambda) = \arg \min \{l_n^R(\beta, \lambda)\}$ and $\hat{\beta}(\lambda)$ was defined in (3). Let $\mathcal{B}$ be a (compact) parameter space for the regression coefficients.

Under suitable regularity conditions, there exists an $\varepsilon_n > 0$ such that, for every $\beta \in \mathcal{B}$, $|r_n^R(\beta, \lambda) - l_n(\beta, \lambda)| < \varepsilon_n$ implies $\|\hat{\beta}^R(\lambda) - \hat{\beta}(\lambda)\| < \eta(\varepsilon_n)$ in a stochastic sense. The most direct method of minimizing the distance in constrained loss functions is to minimize the distance in unconstrained loss functions. In the Buckley-James problem with Dantzig-type constraints, we want the distance between the maximal normal equation $\|\mathbf{x}^*(\mathbf{Y} - \mathbf{X}\beta)\|$ and $\|\mathbf{x}^*\{\hat{\mathbf{Y}}(\hat{\beta}^{\text{ini}}) - \mathbf{X}\beta\}\|$ to be minimal for every $\beta \in \mathcal{B}$, where $\hat{\mathbf{Y}}(\hat{\beta}^{\text{ini}})$ is the imputed response vector (5) using initial estimator $\hat{\beta}^{\text{ini}}$. Without loss of generality, the imputation error (IE) for $\hat{\beta}^{\text{ini}}$ is,

$$\text{IE}(\hat{\beta}^{\text{ini}}) = \|\hat{\mathbf{Y}}(\hat{\beta}^{\text{ini}}) - \mathbf{Y}\|_1 \leq \|\hat{\mathbf{Y}}(\hat{\beta}^{\text{ini}}) - \hat{\mathbf{Y}}(\beta_0)\|_1 + \|\hat{\mathbf{Y}}(\beta_0) - \mathbf{Y}\|_1. \quad (8)$$

The second expression on the right-hand side of (8) is an error term that depends on the validity of (5) and the true data generating mechanism. The first expression on the right-hand side of (8) is controlled by the imputer and suggests that desirable synthetic responses are as close as possible to the “true-value” imputations $\hat{\mathbf{Y}}(\beta_0)$.

Let $\mathcal{V}(\mathbf{Z}_{\text{obs}})$ be the auxiliary variable selection procedure used to select the subset $\mathcal{S}_n$. We assume that the stochastic subset $\mathcal{S}_n$ converges to a limit, i.e. $P(\mathcal{S}_n = \mathcal{S}^*) \rightarrow 1$ as $n \rightarrow \infty$. If the limiting subset $\mathcal{S}^*$ is, in fact, the true active subset $\mathcal{S}_0 = \{j: |\beta_{0j}| > 0\}$, then the procedure $\mathcal{V}(\mathbf{Z}_{\text{obs}})$ is said to be consistent in model selection. If the auxiliary procedure $\mathcal{V}(\mathbf{Z}_{\text{obs}})$ is consistent in model selection, then our unregularized estimator on the subset $\mathcal{S}_n$, i.e. $\hat{\beta}_{\mathcal{S}_n}^{\text{Un}}$, has an oracle property (Fan and Li, 2001). It is known that estimators with an oracle property have smaller prediction error than, say, lasso or Dantzig estimation. Because our initial estimator, $\hat{\beta}_{\mathcal{S}_n}^{\text{Un}}$, is not regularized, it can still provide good imputation even when the number of potential predictors is ultra-high dimensional as long as the auxiliary procedure $\mathcal{V}(\mathbf{Z}_{\text{obs}})$ is consistent in model selection. Auxiliary procedures $\mathcal{V}(\mathbf{Z}_{\text{obs}})$ that are not consistent in model selection can still be excellent variable selection procedures (cf. Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006) and have imputation error on the same order as methods that are consistent in model selection. This topic is beyond the scope of the current work and the subject of future investigation.

3.2 Algorithm

Our goal is to solve the following optimization problem,

$$\min_{\beta} \|\mathbf{x}^*(\hat{\mathbf{Y}}(\hat{\beta}^{\text{ini}}) - \mathbf{X}\beta)\|_{\infty} \quad \text{subject to } \|\beta\|_1 \leq \lambda, \quad (9)$$

for values of the regularization parameter such that coefficient estimates are zero on one boundary and saturated on the other boundary. Along the lines of Friedman et al. (2007), our algorithm approximates the entire FRDS coefficient path for a grid of regularization parameters of length L ranging from $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min}$ in (9) results in setting all coefficients to zero and $\lambda_{\max}$ results in a saturated fit. The following lemma is analogous to one given in Park and Hastie (2005, Lemma 1) and provides bounds on the grid of values.

Lemma 1—There exists a minimal regularization parameter $\lambda_{\min}$ such that for every $\lambda \leq \lambda_{\min}$, the solution to (9) is exactly zero.

Proof: Rewrite the optimization in (9) as

$$\min_{\beta} \|\beta\|_1 \text{ subject to } |\mathbf{x}^{(j)}\{\widehat{\mathbf{Y}}(\widehat{\beta}^{\text{ini}}) - \mathbf{X}\beta\}| \leq \tau, j=1, \dots, p, \quad (10)$$

where $\mathbf{x}^{(j)}$ is the j -th column of $\mathbf{X}$. Because there is a 1-1 correspondence between $\tau_{\max}$ and $\lambda_{\min}$, we are done if we can show there exists a $\tau_{\max}$ such that the solution to (10) is exactly zero for $\tau \geq \tau_{\max}$. The conclusion is satisfied with $\tau_{\max} = \max_{1 \leq j \leq p} |\mathbf{x}^{(j)}\{\widehat{\mathbf{Y}}(\widehat{\beta}^{\text{ini}}) - \mathbf{X}\beta\}|$.

Since the initial estimator $\widehat{\beta}^{\text{ini}}$ is arbitrary, Lemma 1 can be applied to any initialization strategy. In our numerical work, we approximate the Dantzig path by using a grid of length $L = 100$ on a logarithmic scale. Furthermore, when $n \gg p$, the range of the regularization parameter is $[\lambda_{\min}, \infty]$ in (9) or equivalently $[0, \tau_{\max}]$ in (10). When $p \gg n$, we cannot run the regularization parameter all the way to zero. Here, we use $\tau_{\min} = 0.001 \times \tau_{\max}$ (cf. Friedman et al., 2007; Friedman et al., 2010).

3.3 Parameter Tuning

Because of imputation and our estimating equation approach to parameter estimation, it is clear that naïvely using ordinary statistics for parameter tuning will not suffice. For general semi-parametric estimators built on estimation equations, $0 = \Psi_n(\beta) = \sum_i \psi(\mathbf{Z}_i, \beta)$, we tune the L_1 -norm of the estimating function through V -fold cross-validation (CV). Let $\mathbb{J} = \{1, \dots, n\}$ be the full data set, $\mathbb{J}^v$ and $\mathbb{J}^{(-v)}$ are the validation and training data, respectively. Then, the idea we propose is to minimize

$$\text{CV}(\lambda) = \sum_{v=1}^V \left\| \frac{1}{|\mathbb{J}^v|} \sum_{k \in \mathbb{J}^v} \psi(\mathbf{Z}_k, \widehat{\beta}_{(-v)}(\lambda)) \right\|_1.$$

The specifics of CV tuning in a Buckley-James framework require careful consideration of the imputation step. Without loss of generality, let $\widehat{\beta}^{\text{ini}}$ be any regularized or unregularized estimate used in initialization. For tuning regularized Buckley-James estimators, we propose the CV criterion,

$$\text{CV}(\lambda) = \sum_{v=1}^V \left\| \frac{1}{|\mathbb{J}^v|} \sum_{k \in \mathbb{J}^v} \mathbf{x}_k^T \left\{ \widehat{\mathbf{Y}}_k^{(-v)}(\widehat{\beta}_{(-v)}^{\text{ini}}) - \mathbf{x}_k^T \widehat{\beta}_{(-v)}(\lambda) \right\} \right\|_1,$$

where $\widehat{\beta}_{(-v)}^{\text{ini}}$ and $\widehat{\beta}_{(-v)}^{\lambda}$ are the initial and regularized Dantzig estimate, respectively, from the training sample, $|\mathbb{J}^v|$ is the cardinality of the validation sample and $\widehat{\mathbf{Y}}_k^{(-v)}(\beta^*)$ are the imputed responses using β^* and the Kaplan-Meier $\widehat{S}_e(t)$ estimate from the training sample, not the validation sample. The initial estimators that we investigate in this paper, $\widehat{\beta}^{\text{ini}}$ and $\widehat{\beta}_{s_n}^{\text{Un}}$ discussed in Section 2, also require tuning and adds to the computational complexity of the overall procedure.

Tuning regularized Buckley-James estimators is challenging and has been attempted previously. For comparison purposes, we also consider CV tuning using the Gehan loss (Johnson, 2008) and using IPW loss (Johnson et al., 2008). Let $K(t) = P(C > t)$ be the

survivor distribution for the censoring random variable and define the inverse-weighted estimator squared-error loss function, $\sum_{i=1}^n \widehat{w}_i (U_i - \mathbf{x}_i^T \beta)^2$, $\widehat{w}_i = \delta_i / \widehat{K}(U_i)$, where $\widehat{K}(t)$ is the Kaplan-Meier estimator for $K(t)$. In cross-validation, the weights in the validation data set are computed using Kaplan-Meier estimator $\widehat{K}(t)$ computed from the training sample. Deriving rigorous proofs to justify CV tuning in general estimating equations is still ongoing work.

4. Application to Lung Andenocarcinoma Microarrays

Our application uses data from a pooled sample of microarrays from two different studies. Both microarray studies, one conducted at Harvard and another at Michigan, used Affymetrix oligonucleotide arrays but used different versions of Affymetrix chips. Michigan used the HuGeneFL chip while Harvard used the U95Av2 chip. Different versions of microarray chips make identifying prognostic genes in the pooled sample difficult. In addition to additional noise which may be present due to the fact that data were analyzed at different institutions with different technicians, the two chip types do not contain the same probesets. A natural way of pooling information is to focus on the subset of probes that match across chip types. This “partial probeset” approach was proposed by Morris et al. (2005) and the data we analyze was pre-processed using their method. The total combined sample includes data on $n = 200$ and $p = 1036$ probes, the latter of which we simply refer to as “genes.” The survival endpoint is interpreted as time-to-death (in months) due to lung andenocarcinoma and a total of 103 observations (53.5%) were censored. These data were originally part of Critical Assessment of Microarray Data Analysis (CAMDA) 2003 program and additional details may still be found on their website (<http://www.camda.duke.edu/camda03/datasets/index.html>).

In our data analysis and simulation studies in Section 5, the following estimators are considered.

- **Rank-based Threshold Gradient Descent (TGD).** TGD (Friedman and Popescu, 2004) applied to the rank-based Gehan loss function and observed data. This method approximates the lasso coefficient paths for the Gehan loss (Cai et al., 2009).
 - **Best Restoration Dantzig Selector (Best).** Impute the censored outcomes based on the true value β_0 followed by DS. BRDS is the gold standard among DS-based methods for variable selection based on the Buckley-James estimating function.
 - **Self-Restored Dantzig Selector (Self).** Iteratively impute the censored observations and simultaneously estimate the coefficient path $\mathcal{P}_{DS}(\mathbf{Z}_{obs})$. Regularized coefficient estimates $\hat{\beta}(\lambda)$ at the k -th stage are used as starting values at $(k + 1)$ -st stage.
 - **FRDS-A.** Frequentist Restoration Dantzig Selector (FRDS) initializing with the final regularized coefficient from TGD, $\hat{\beta}'$.
 - **FRDS-B₁.** FRDS using the active set defined by the optimal regularized TGD estimator to define the set $\mathcal{S}_n$. Then, the initial value, $\widehat{\beta}_{\mathcal{S}_n}^{Un}$, is the unregularized Gehan estimator defined on this subset of predictors.
 - **FRDS-B₂.** Same as FRDS-B₁ except that $\mathcal{S}_n$ is defined the active set from L_1 -regularized Cox regression (Goeman, 2009).
- 5-fold CV Tuning.** For FRDS-B₁ and FRDS-B₂, we tune the procedures using 5-fold cross-validation through the following loss functions: L_q -norm of the

estimating function, $q = 1, 2$ (FRDS-B₁-L₁, FRDS-B₁-L₂), the Gehan loss function in (7) (FRDS-B₁-G), and the IPW loss (FRDS-B₁-W). We have analogous abbreviations for FRDS-B₂ procedures.

We found that the L_1 -regularized Cox (L1Cox) model selected 12 genes while rank-based TGD selected 22 genes, 10 genes are common to both sets. Initializing the FRDS procedure with TGD regularized estimates (i.e. FRDS-A) led to a model with 25 active genes, 16 genes are common to TGD while 8 are common to L1Cox. Finally, we consider the FRDS-B procedures. We first examined the differences in the imputed responses from the L1Cox 12-gene model as opposed to the 22-gene model from TGD. Recall the imputation in (5) relies on an estimate of the error distribution function F_ϵ which is computed via Kaplan-Meier estimator. The two competing empirical survivor curves are displayed in Figure 1 and we note substantial agreement. Pearson's correlation coefficient between the two vectors of fitted residuals is $r = 0.87$. Similarly, the synthetic responses using L1Cox initialization are highly correlated with the imputed responses using TGD initialization: $r = 0.98$ for all 200 observations (including the identical uncensored responses) and $r = 0.86$ using the 103 imputed responses only. The maximum absolute deviation between the two synthetic response vectors was 1.16 months.

Using 50-fold CV and L_1 -norm tuning, FRDS-B₁ chooses a model with 42 active genes while FRDS-B₂ model has 22 active genes. We note that both procedures result in final models with about twice as many active genes as the initial active set S_n that generated the synthetic response vector. For the 20 genes in the FRDS-B₁ model but not included in the FRDS-B₂ model, many coefficient estimates were modest but not exactly zero; the largest absolute coefficient estimate among these 20 genes was 0.11, about 8% of one standard deviation of the synthetic response vector. The FRDS-B₁ final model shares 19, 10, and 16 genes in common with TGD, L1Cox, and FRDS-A, respectively. If we use FRDS-B₂ with L_1 -norm tuning, the final model intersects on 15, 9, and 17 genes with TGD, L1Cox, and FRDS-A, respectively.

We also considered four different loss functions for tuning the regularization parameter: L_1 - and L_2 -norm of the Buckley-James estimating function, Gehan and IPW loss. For the FRDS-B₁ method, tuning with L_2 -norm selected 10 active variables, tuning with Gehan loss had 19 active variables and tuning with IPW loss had no active variable. As for the FRDS-B₂ method, the cardinality of the active set was 28, 22, and zero variables for L_2 -norm, Gehan, and IPW tuning, respectively. We were curious why tuning with IPW loss consistently chose the null model and printed all CV curves for the two FRDS methods with B initialization. Four CV curves are displayed in Figure 2 with solid line for FRDS-B₁ and dashed line for FRDS-B₂. The abscissa is scaled as a ratio of $\lambda_{\max}$ (cf. Friedman et al., 2007; Friedman et al., 2010), the data-dependent value of the regularization parameter that sets all coefficients to zero.

For the microarray data, tuning with L_1 -, L_2 -norm or Gehan loss all have minima on the interior while tuning with IPW loss does not. Interestingly, CV tuning FRDS-B₁ with L_1 -norm seems to have two minima, one near 0.2 and another near 0.6, but the CV curve is smallest at 0.2. If we tune FRDS-B₁ with L_2 -norm of the estimating function, then the larger ratio is chosen and a much simpler model results. This 10-gene model has 5 genes in common with L1Cox, 9 genes in common with TGD, and is a proper subset of FRDS-B₂. Figure 2 also demonstrates the cardinality of the FRDS-B₂ models was rather similar across three different CV tuning criteria.

5. Simulation Studies

We performed numerous simulation studies to assess the finite sample performance of the proposed methods. The referees noted a number of competing high-dimensional survival techniques have been proposed in the literature and we agree that this is an active research area. We limit our attention to 12 estimators listed in Section 4 that highlight nuances of regularized Buckley-James estimation.

5.1 Path Restoration

We begin by demonstrating that iterative regularized Buckley-James methods fails to restore the entire regularized coefficient paths. We simulated true data according to a normal-theory linear model,

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \sigma Z_i, \quad (i = 1, \dots, n),$$

with (x_{i1}, x_{i2}, Z_i) distributed trivariate normal with identity covariance matrix and $\sigma = 3$. Censoring random variables simulated according to the model, $C_i = \beta_1 x_{i1} + \beta_2 x_{i2} + W_i$, where $W_i \sim \text{Un}(-3, 3)$. The observed random pair (U_i, δ_i) are computed accordingly. We considered two scenarios: one where the coefficients were equal to one standard deviation, $\beta = (3, -3)^T$, and a second empty or null scenario, $\beta = (0, 0)^T$. The observed data contains roughly 50% censoring in both scenarios. We considered two sample sizes, $n = 40$ and $n = 60$, and the results are displayed in Table 1.

We compute the integrated squared error, $\int \{\beta(\lambda) - \hat{\beta}(\lambda)\}^2 d\lambda$, over a fine grid of the regularization parameter λ , where $\hat{\beta}(\lambda)$ is estimated using the complete data DS and several different reconstructed paths $\hat{\beta}(\lambda)$ are considered. The Best FRDS imputes the censored observations using the true coefficient vector β_0 while the “proposed” FRDS computes

unregularized Gehan estimates of the coefficients $\hat{\beta}_{S_n}^{\text{Un}}$ on the subset S_n (here, the subset S_n is always $\{1, 2\}$). The self-restored coefficient path is computed using iterative Buckley-James methods at each λ_j in the grid spanning the index set Λ_{DS} . In this simple example with $p = 2$ and $\beta_j = \pm 3$, the cumulative error in path estimation using iterative Buckley-James methods (“Self”) is about twice as large as the proposed frequentist restoration and this remains true even as the sample size increases. However, when no coefficients are active, then iterative Buckley-James (“Self”) methods will generally have smaller cumulative error than FRDS. This is primarily due to the fact that initial imputation at the first step (i.e. $\beta = 0$) starts with the true value. But even in this extreme case, the relative advantage of iterative Buckley-James over FRDS is already less than 10% when $n = 60$. Thus,

5.2 Finite Sample Properties of FRDS

In our first simulation study, we generated data according the linear model (1) with $n = 60$, $p = 8$, $\beta = (3/2, 3/4, 0, 0, 1, 0, 0) \times \Delta$ and the errors ε_i are independent and identically distributed as mean-zero normal random variables with standard deviation 2. The Greek letter Δ represents the effect size. The covariate vector $\mathbf{x}_i$ is generated from a normal distribution with mean zero and covariance Ω , where Ω assumes an AR(1) structure parameterized by ρ . Censoring was uniformly distributed $\text{Un}(0, \tau)$, with τ chosen to yield approximately 20% censoring.

Table 2 summarizes the simulation results for small p settings ($p = 8$). First, tuning regularized Buckley-James estimators by IPW loss is uniformly and substantially worse than by the Gehan loss (cf. Johnson, 2008) or the L_q -norm of the estimating function, of which

tuning by the Gehan loss appears to perform slightly better with larger P_0 . Second, when using the same tuning method, the performance of FRDS-B₂ which identifies the initialization set $\mathcal{S}$ by simply sorting test statistics from a univariate fit is comparable to that of FRDS-B₁ which defines $\mathcal{S}$ as the active set from the optimal TGD coefficient estimate. Third, excluding FRDS-B-W, all other FRDS-B methods outperform self-restoration DS and FRDS-A in identifying the correct submodel and minimizing risk. Fourth, the best FRDS-B method and the rank-based gradient descent achieve fairly similar overall performance with each slightly better in some settings. Compared to the “best” regularized Buckley-James estimator, both the best FRDS-B method and TGD tend to have slightly higher probability of identifying true active/inactive variables ($P_0 + P_1$) but also have higher MSE. Finally, the mean squared error and ability to correctly identify true zeros suffers as the signal-to-noise ratio increases, which is a well-known with lasso and not surprisingly is true for DS as well; in addition, the performance suffers for all methods as the correlation strengthens.

Our simulation setup with large p ($p = 100$, $n = 60$) is similar to the one above with the exception that $(\beta_1, \dots, \beta_8) = (1, 1, 0, 0, 1, 0, 0, 1) \times \Delta$ and $\beta_9 = \dots = \beta_{100} = 0$. Most observations seen in the small p settings in Table 2 are still true in the large p settings. For example, tuning regularized Buckley-James estimators with IPW loss is still uniformly and substantially worse than tuning with either L_q -norm of the estimating function or Gehan loss. However, here we note that tuning with L_q -norm of the estimating function is somewhat better than Gehan loss in terms of MSE and variable selection but the improvement lessens as the correlation among predictors increases. It is also noteworthy that FRDS-B₂ methods that use L_q -norm of the estimating function or Gehan loss for tuning perform as well or better than TGD when $\rho = 0.5$ or less. Finally, in the large p settings, all methods achieve their best performance when $\rho = 0.5$, and their MSEs are substantially larger when $\rho = 0.5$ or less compared to the small p settings. A referee noted that TGD performed better than the Best FRDS when $\rho = 0.9$. This may be due, in part, to better tuning procedures for TGD than FRDS or because Best FRDS still relies on imputation whereas TGD does not. Recall, Best FRDS is the best restoration among other FRDS procedures but not necessarily among all estimators in the AFT model.

6. Remarks

This paper attempts to tell two stories. First, although penalized likelihood and penalized least squares (PLS) is a powerful tool, it is by no means a panacea for variable selection in complex data. Biometricians working with composite models, missing data, or discrete estimating functions will find the current PLS literature limiting. Fu’s (2003) extension of lasso to generalized estimating equations (GEE) does not pertain to discrete estimating functions, such as the Buckley-James estimators in (4). In high-dimensional missing data problems, imputing the missing data from the observed data, say $E(\mathbf{Z}_{\text{comp}}|\mathbf{Z}_{\text{obs}}, \boldsymbol{\beta})$, often depends on consistent estimation of the unknown regression coefficients. Imputation based on an ordinary regularized estimator (e.g. lasso) may work in some special cases but not in general (e.g. the ultra high-dimensional case). Similar to the missing data problem where there is an imputation model and an analytic model, joint and composite modeling problems have multiple models with potentially different objectives. In the current paper, the analyst’s objective is to estimate the entire regularized coefficient path while the imputer’s objective is to minimize imputation or prediction error. In practice, these two objectives may require rather different techniques and the same will be true of other high-dimensional complex data problems.

The second point of our paper is an attempt to address the high-dimensional Buckley-James problem. In the high-dimensional Buckley-James problem, we argued that no ordinary extension of existing path computation methods apply due the imputation of synthetic

responses for censored observations, $E(Y|Y > C, \beta)$. We imputed synthetic responses based on an unregularized estimator on a user-defined subset S_n from an auxiliary high-dimensional variable selection procedure applied to the observed data $\mathbf{Z}_{\text{obs}}$. After the synthetic responses are computed, the entire regularized coefficient path can be estimated easily using any path estimation technique developed for a least-squares regression problem. Here, the focus was Dantzig estimation (Candes and Tao, 2007) but the same ideas apply to other regularized estimators such as lasso (Tibshirani, 1996), lars (Efron et al., 2004), boosting, and gradient descent (Friedman and Popescu, 2004). We argued that if the cardinality of true active set S_0 is less than the sample size and the auxiliary variable selection procedure was consistent in model selection, then our imputation would have desirable operating characteristics and the resulting estimated coefficient path would be close to the original coefficient path based on complete data. Simulation studies appeared to support our intuition.

In Tables 2–3, a referee noted that the rank-based gradient descent could even beat the “Best” FRDS procedure. If the goal is high-dimensional survival analysis in the AFT model, then rank-based variable selection (Johnson, 2008, 2009b; Cai et al., 2009) are attractive due to their simplicity and a well-defined loss function in (7). Any high-dimensional statistical learner can be extended to Gehan loss although the estimator is not efficient in large samples and few predictors. When the goal is to reconstruct the complete-data regularized coefficient paths, then imputation and unbiased transformations are better tools. Nevertheless, to reconcile the high-dimensional Buckley-James procedure, we need to develop parameter tuning in the absence of a well-defined convex loss. Here, we considered parameter tuning using 5-fold cross-validation for each of four different loss functions: inverse-weighting (IPW), Gehan loss, and the L_1 - and L_2 -norm of the Buckley-James estimating function. We found that IPW tuning consistently produced estimates with the highest mean squared-error while the other three loss functions gave similar performance. In our analysis of the microarray data, the cross-validation curve using IPW loss was not convex for two regularized Buckley-James estimators as shown in Figure 2. Parameter tuning using Gehan loss or L_2 -norm of the estimating function gave the most reliable performance in low and high dimensions. Parameter tuning for general estimating functions is still an open problem.

Acknowledgments

We thank the Editor Zucker, an anonymous associate editor and referees for their suggestions which greatly improved an earlier draft of this manuscript. We acknowledge Editor Zucker for suggesting that we investigate initialization from a possibly misspecified model. The research of the author was supported in part by grants from the National Institutes of Allergies and Infectious Diseases (R03AI068484), and Emory’s Center for AIDS Research (P30AI050409). We thank Jeff Morris and Danyu Lin for sharing the lung cancer data and acknowledge Eugene Huang for suggesting the L_1 -norm as a surrogate loss in cross-validation.

References

- Antoniadis, A.; Fryzlewicz, P.; Letué, F. Technical report. University of Bristol; UK: 2009. The Dantzig selector in Cox’s proportional hazards model.
- Buckley J, James I. Linear regression with censored data. *Biometrika*. 1979; 66:429–436.
- Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n (with Discussion). *Ann Statist*. 2007; 35:2313–2404.
- Cox DR. Regression models and life-tables (with Discussion). *J Roy Statist Soc Ser B*. 1972; 34:187–202.
- Cox, DR.; Oakes, D. *Analysis of Survival Data*. London: Chapman and Hall; 1984.
- Datta S, Le-Rademacher J, Datta S. Predicting survival from microarray data by accelerated failure time modeling using partial least squares and lasso. *Biometrics*. 2007; 63:259–271. [PubMed: 17447952]

- Engler, D.; Li, Y. Technical Report. Harvard University and Dana Farber Cancer Institute; 2009. Survival analysis with high-dimensional covariates: an application in microarray studies.
- Efron B, Hastie T, Johnstone IM, Tibshirani R. Least angle regression (with Discussion). *Ann Statist.* 2004; 32:407–499.
- Friedman, J.; Popescu, BE. Technical report. Stanford University, Department of Statistics; 2004. Gradient directed regularization.
- Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. *Ann Appl Statist.* 2007; 1:302–332.
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010; 33:1–22. [PubMed: 20808728]
- Gehan EA. A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika.* 1965; 52:203–23. [PubMed: 14341275]
- Goeman J. L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal.* 2009; 52:70–84. [PubMed: 19937997]
- Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low- sample size settings, with application to microarray gene expression data. *Bioinformatics.* 2005; 21:3001–3008. [PubMed: 15814556]
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer; New York: 2000.
- Huang J, Ma S, Xie H. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics.* 2006; 62:813–820. [PubMed: 16984324]
- James G, Radchecko P. A generalized Dantzig selector with shrinkage tuning. *Biometrika.* 2009a; 96:323–338.
- Jin Z, Lin DY, Ying Z. On least-squares regression with censored data. *Biometrika.* 2006; 93:147–161.
- Johnson BA. Variable selection in semiparametric linear regression with censored data. *J R Statist Soc Ser B.* 2008; 70:351–370.
- Johnson BA. On lasso for censored data. *Electronic Journal of Statistics.* 2009a; 3:485–506.
- Johnson BA. Rank-based estimation in the ℓ_1 -regularized partly linear model with applications to integrated analyses of clinical predictors and gene expression data. *Biostatistics.* 2009b; 10:640–658.
- Johnson, BA. Emory University Technical Report 09-01. Department of Biostatistics; 2009c June. Fast restoration Dantzig selection for censored data.
- Johnson BA, Lin DY, Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. *J Amer Statist Assoc.* 2008; 103:672–680.
- Lange K, Hunter DR, Yang I. Optimization transfer using surrogate objective functions (with discussion). *J Comput Graph Statist.* 2000; 9:1–59.
- Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. *Ann Statist.* 2006; 34:1436–1462.
- Meinshausen N, Rocha G, Yu B. Discussion (of Candès and Tao, 2007): A tale of three cousins: lasso, l2boosting, and Dantzig. *Ann Statist.* 2009; 35:2373–2384.
- Morris, JS.; Yin, G.; Baggerly, KA.; Wu, C.; Zhang, L. Pooling information across different studies and oligonucleotide microarray chip types to identify prognostic genes for lung cancer. In: Shoemaker, JS.; Lin, SM., editors. *Methods of Microarray Data Analysis IV.* Springer-Verlag; New York: 2005.
- Nocedal, J.; Wright, SJ. *Numerical Optimization.* Springer; Berlin: 2006.
- Osborne MR, Presnell B, Turlach BA. On the lasso and its dual. *J Comput Graph Statist.* 2000; 9:319–337.
- Ritov Y. Estimation in a linear regression model with censored data. *Ann Statist.* 1990; 18:303–328.
- Rosset S, Zhu J, Hastie T. Boosting as a regularized path to a maximum margin classifier. *J Machine Learning Research.* 2004; 5:941–973.
- Tibshirani RJ. Regression shrinkage and selection via the lasso. *J Roy Statist Soc Ser B.* 1996; 58:267–288.

- Tibshirani RJ. The lasso method for variable selection in the Cox model. *Statist Med.* 1997; 16:385–395.
- Wang S, Nan B, Zhu J, Beer DG. Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics.* 2008; 64:132–140. [PubMed: 17680828]
- Zhao P, Yu B. On Model Selection Consistency of Lasso. *J Mach Learning Res.* 2006; 7:2541–2563.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Roy Statist Soc Ser B.* 2005; 67:301–320.

Figure 1.

Empirical survivor distribution of the unregularized fitted residuals using the active set $\mathcal{S}_n$ from either L_1 -regularized Cox regression (L1Cox, 12 genes) or rank-based threshold gradient descent (TGD, 22 genes) to initialize the FRDS procedure. This figure appears in color in the electronic version of this article.

Figure 2. Display of CV curves in the microarray data for FRDS-B₁ (solid) and FRDS-B₂ (dashed). This figure appears in color in the electronic version of this article.

Table 1
Cumulative squared error of coefficient paths to complete-data coefficient paths

$\beta =$	$(3, -3)$		$(0, 0)$	
	$n =$	40	60	60
Best		0.066	0.046	0.055
Self		0.165	0.116	0.063
FRDS		0.081	0.055	0.074

Table 2

Simulation results for $n = 60$ and $p = 8$ based on 500 Monte Carlo datasets. Table entries are mean squared error (MSE) computed as $(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)$ with true regression coefficients $\beta = (3/2, 3/4, 0, 0, 1, 0, 0) \times \Delta$; P_0 , the proportion of zero coefficients correctly set to zero; P_1 , the proportion of nonzero coefficients correctly set to nonzero.

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
	MSE	P_0	P_1	MSE	P_0	P_1	MSE	P_0	P_1
$\Delta = 1$									
Best	0.520	0.433	0.985	0.630	0.460	0.989	3.091	0.464	0.833
Self	0.579	0.425	0.981	0.748	0.465	0.986	3.575	0.464	0.817
TGD	0.603	0.482	0.981	0.728	0.503	0.977	2.060	0.525	0.876
FRDS-A	0.597	0.423	0.975	0.749	0.444	0.985	3.373	0.461	0.820
FRDS-B ₁ -L ₁	0.599	0.438	0.977	0.756	0.464	0.982	3.467	0.473	0.814
FRDS-B ₁ -G	0.563	0.438	0.986	0.743	0.495	0.983	3.165	0.531	0.799
FRDS-B ₁ -L ₂	0.584	0.434	0.981	0.758	0.451	0.982	3.456	0.462	0.813
FRDS-B ₁ -W	0.703	0.197	0.969	0.936	0.263	0.973	3.712	0.391	0.829
FRDS-B ₂ -L ₁	0.601	0.444	0.973	0.743	0.463	0.983	3.320	0.480	0.815
FRDS-B ₂ -G	0.564	0.439	0.984	0.738	0.494	0.982	3.068	0.534	0.800
FRDS-B ₂ -L ₂	0.578	0.439	0.982	0.745	0.460	0.983	3.317	0.473	0.817
FRDS-B ₂ -W	0.702	0.198	0.967	0.928	0.262	0.974	3.537	0.390	0.823
$\Delta = 2$									
Best	0.514	0.413	1.000	0.609	0.422	1.000	3.728	0.463	0.968
Self	0.743	0.390	1.000	0.932	0.433	1.000	5.659	0.466	0.939
TGD	0.749	0.426	1.000	0.919	0.446	1.000	3.112	0.448	0.978
FRDS-A	0.741	0.388	1.000	0.917	0.409	1.000	4.838	0.426	0.953
FRDS-B ₁ -L ₁	0.727	0.413	1.000	0.945	0.428	0.999	5.470	0.440	0.941
FRDS-B ₁ -G	0.706	0.425	1.000	0.923	0.479	0.999	4.342	0.458	0.955
FRDS-B ₁ -L ₂	0.713	0.399	1.000	0.939	0.428	0.999	5.369	0.434	0.950
FRDS-B ₁ -W	0.942	0.328	0.997	1.118	0.434	0.997	5.713	0.517	0.916
FRDS-B ₂ -L ₁	0.713	0.409	1.000	0.091	0.438	1.000	5.130	0.466	0.937
FRDS-B ₂ -G	0.696	0.424	1.000	0.892	0.480	1.000	4.101	0.457	0.955

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
	MSE	P_0	P_1	MSE	P_0	P_1	MSE	P_0	P_1
FRDS-B ₂ -L ₂	0.703	0.401	1.000	0.894	0.438	1.000	5.073	0.453	0.948
FRDS-B ₂ -W	0.930	0.321	0.997	1.087	0.448	0.998	5.579	0.522	0.913

Table 3

Simulation results for $n = 60$ and $p = 100$ based on 500 Monte Carlo datasets. Table entries and simulation summaries analogous to Table 2.

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
	MSE	P_0	P_1	MSE	P_0	P_1	MSE	P_0	P_1
$\Delta = 1$									
Best	1.425	0.853	0.925	1.164	0.868	0.963	2.950	0.916	0.754
Self	1.584	0.846	0.904	1.338	0.863	0.947	3.239	0.917	0.705
TGD	1.557	0.892	0.863	1.246	0.896	0.925	2.027	0.920	0.827
FRDS-A	1.605	0.849	0.899	1.345	0.863	0.947	3.063	0.911	0.731
FRDS-B ₁ -L ₁	1.614	0.848	0.897	1.348	0.865	0.944	3.288	0.914	0.732
FRDS-B ₁ -G	1.652	0.837	0.901	1.433	0.856	0.943	3.271	0.907	0.735
FRDS-B ₁ -L ₂	1.624	0.842	0.897	1.365	0.862	0.941	3.208	0.913	0.733
FRDS-B ₁ -W	3.760	0.618	0.928	3.883	0.650	0.937	8.824	0.767	0.770
FRDS-B ₂ -L ₁	1.577	0.853	0.903	1.329	0.867	0.943	3.112	0.918	0.719
FRDS-B ₂ -G	1.607	0.844	0.911	1.365	0.865	0.942	3.078	0.913	0.739
FRDS-B ₂ -L ₂	1.572	0.848	0.906	1.333	0.866	0.945	3.094	0.917	0.726
FRDS-B ₂ -W	3.730	0.625	0.931	3.800	0.661	0.940	8.436	0.774	0.765
$\Delta = 2$									
Best	1.450	0.833	0.999	1.131	0.871	0.999	4.394	0.926	0.943
Self	2.236	0.830	0.991	1.877	0.858	0.997	6.070	0.911	0.881
TGD	2.119	0.869	0.993	1.770	0.877	0.998	3.722	0.911	0.963
FRDS-A	2.199	0.823	0.995	1.825	0.858	0.997	5.419	0.898	0.921
FRDS-B ₁ -L ₁	2.052	0.822	0.997	1.780	0.856	0.998	5.964	0.904	0.916
FRDS-B ₁ -G	2.119	0.804	0.998	1.796	0.849	0.998	5.613	0.894	0.909
FRDS-B ₁ -L ₂	2.055	0.819	0.997	1.752	0.856	0.998	5.947	0.904	0.908
FRDS-B ₁ -W	3.177	0.701	0.993	2.684	0.782	0.992	7.795	0.861	0.899
FRDS-B ₂ -L ₁	2.068	0.817	0.998	1.846	0.845	0.997	5.997	0.900	0.891
FRDS-B ₂ -G	2.116	0.804	0.997	1.872	0.843	0.998	5.945	0.895	0.901
FRDS-B ₂ -L ₂	2.076	0.813	0.998	1.844	0.844	0.997	5.808	0.905	0.895
FRDS-B ₂ -W	3.151	0.700	0.994	2.686	0.779	0.993	7.828	0.862	0.885