Published in final edited form as:
Biometrics. 2011 June ; 67(2): 504–512. doi:10.1111/j.1541-0420.2010.01466.x.

High-Dimensional Variable Selection in Meta-Analysis for Censored Data

Fei Liu[1,*], **David Dunson**[2,**], and **Fei Zou**[3,***]

[1]IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, U.S.A

[2]Department of Statistics, Duke University, Durham, North Carolina 27708, U.S.A

[3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

Summary

This article considers the problem of selecting predictors of time to an event from a high-dimensional set of candidate predictors using data from multiple studies. As an alternative to the current multistage testing approaches, we propose to model the study-to-study heterogeneity explicitly using a hierarchical model to borrow strength. Our method incorporates censored data through an accelerated failure time model. Using a carefully formulated prior specification, we develop a fast approach to predictor selection and shrinkage estimation for high-dimensional predictors. For model fitting, we develop a Monte Carlo expectation maximization (MC-EM) algorithm to accommodate censored data. The proposed approach, which is related to the relevance vector machine (RVM), relies on maximum a posteriori estimation to rapidly obtain a sparse estimate. As for the typical RVM, there is an intrinsic thresholding property in which unimportant predictors tend to have their coefficients shrunk to zero. We compare our method with some commonly used procedures through simulation studies. We also illustrate the method using the gene expression barcode data from three breast cancer studies.

Keywords

Accelerated failure time; Expectation maximization (EM) algorithm; Lasso; Maximum a posteriori (MAP) estimation; Meta-analysis; Relevance vector machine; Shrinkage

1. Introduction

In modern biomedical research, it has become routine to encounter problems involving massive numbers of predictors, with gene expression data as one example. Often, interest focuses on identifying important predictors of an event time, such as patient survival following cancer treatment. Because the sample size from any single study is typically insufficient to allow accurate selection of important predictors, there has been increased emphasis in recent years on borrowing of strength across data from multiple studies. Different studies are often conducted by different labs and may involve varying platforms and event definitions. These differences lead to study-to-study heterogeneity, which must be accommodated in statistical analysis. This article focuses on the problem of flexibly

*feiliu@us.ibm.com
**dunson@stat.duke.edu
***fzou@bios.unc.edu

borrowing strength across studies in selecting predictors of an event time from a massive number of candidates.

Variable selection is typically of interest, even when prediction is the focus, since one may get better insight into the biological mechanisms by reducing the dimensionality of the predictive model. For variable selection using data from a single study, a broad variety of methods has been developed for prediction based on large numbers of predictors. Bovelstad et al. (2007) provided a recent review of the literature in this area, while also comparing predictive performance for different methods. They concluded that ridge regression had the best performance in terms of prediction for their data sets, with shrinkage outperforming simple variable selection methods, such as univariate selection or forward selection. Unlike ridge regression, Lasso (Tibshirani, 1996, 1997; Zhang and Lu, 2007) results in simultaneous shrinkage and variable selection, as many of the coefficients will be estimated to be zero. An alternative to Lasso, which also has this property and is widely used in the machine learning community, is the relevance vector machine (RVM; Tipping, 2001).

In considering generalizations of these approaches to accommodate data from multiple studies, an important factor is the computational speed. When data are available from several studies and for thousands of genes, standard Bayes methods of posterior computation that rely on Markov chain Monte Carlo (MCMC) algorithms will be very time consuming to implement. In fact, for large numbers of predictors, the size of the model space is too enormous to run stochastic search variable selection algorithms (George and McCulloch, 1997; Hans, Dobra, and West, 2007) to make it converge. Hence, as a pragmatic approach, it is useful to consider fast alternatives to MCMC based on the maximum a posteriori (MAP) estimation. The Lasso and RVM procedures both have a Bayesian interpretation as MAP estimates, with the Lasso placing a double exponential prior on the coefficients, while the RVM uses an improper t prior with zero degrees of freedom. In addition, computational speed can be improved by focusing on normal linear regression models.

In the presence of multiple studies, a common approach for gene selection is to conduct independent analysis of each dataset, and then examine the intersection of the genes selected (see, e.g., Chan et al., 2008). An alternative is to pool the data, and conduct a single analysis ignoring heterogeneity among the studies. Recently, there has been increased emphasis on multistage designs, which identify a subset of candidate genes in an initial study and then validate these genes in the subsequent studies. Refer to Beckly et al. (2008) for a recent example of this approach. Both the independent analyses and multistage approaches fail to borrow strength across the studies. The multistage method fails to detect genes that are very highly significant in the second stage study, but were not significant enough at the first stage to be maintained after false discovery rate control. The same problem arises in the independent analyses approach, which requires genes to be significant in each study. If one uses a union of genes from the different studies instead of an intersection, the false discovery rate will be increased and important genes may still be missed.

To address these problems, one would ideally simultaneously analyze the data from the different studies, while accommodating heterogeneity. Motivated by the very different problem of borrowing information across related signals in performing signal reconstruction from compressive sensing measurements, Ji, Dunson, and Carin (2009) proposed a multitask RVM (MT-RVM). The MT-RVM approach incorporates dependence in the selection of basis functions for related signals, and can potentially be used directly to incorporate dependence in variable selection across studies. However, the method proposed by Ji et al. (2009) does not account for censoring.

To adapt the variable selection models from normal linear regression settings to the analysis of censored data, one can use an accelerated failure time (AFT) model (Buckley and James, 1979; Kalbfleisch and Prentice, 1980; Koul, Susarla, and Van Ryzin, 1981). For example, Datta, Le-Rademacher, and Datta (2007) used an AFT model to predict patient survival from microarray data for a single study, with partial least squares (PLS) and Lasso used for estimation and imputation used to account for censoring. They concluded that Lasso had better performance. Wang et al. (2008) instead related high-dimensional genomic data to survival outcomes using a semi-parametric AFT model, with a doubly penalized Buckley–James method used for estimation. The approach utilized an elastic net penalty (Zou and Hastie, 2005), which is a hybrid of ridge regression and Lasso. To our knowledge, there are no methods currently available for formally combining data from multiple studies in conducting fast high-dimensional variable selection for survival outcomes.

In this article, we propose a multistudy AFT model, which accounts for heterogeneity among studies. The study-specific coefficients for the different genes are assigned carefully chosen hierarchical t priors. Expressing the t prior as a scale mixture of normals following West (1987) leads to a Gamma prior for gene-specific precision parameters. Taking an approach related to that of Ji et al. (2009), we propose to utilize the same gene-specific precision parameters for the different studies in order to borrow information. This specification allows the gene-specific coefficients to vary across studies, while including dependence in the degree of shrinkage toward zero. To allow censoring, a Monte Carlo expectation maximization (MC-EM) algorithm is developed for simultaneous variable selection and coefficient estimation. The proposed approach, which we refer to as hierarchical RVM with censoring (HRVM-C), produces sparse estimates of the gene-specific coefficients, with many of the coefficients set to zero.

The remainder of this article is organized as follows. We first give a brief review of MT-RVM in Section 2 and then introduce HRVM-C in Section 3. Section 3 also discusses the computational details. In Section 4, we present results from simulation studies. Section 5 demonstrates our method with gene expression barcode data from three breast cancer studies in Zilliox and Irizarray (2007). We give our final conclusion and the discussions in Section 6.

2. MT-RVM

Consider S related studies. Let n_i be the number of samples in the ith study ($i = 1, \ldots, S$), t_{ij} be the response of the jth subject in study i, ($j = 1, \ldots, n_i$), and x_{ijk} ($k = 1, \ldots, p$) be the corresponding kth predictor variable. For simplicity, we set $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})'$, $\boldsymbol{t}=(\boldsymbol{t}_1', \ldots, \boldsymbol{t}_S')'$, $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$, and $\boldsymbol{x}_i=(\boldsymbol{x}_{i1}', \ldots, \boldsymbol{x}_{in_i}')'$. The MT-RVM model in Ji et al. (2009) can be represented as

$$\boldsymbol{t}_i=\boldsymbol{x}_i\beta_i+\boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i\sim \mathrm{N}(\boldsymbol{0}, \alpha_{0i}^{-1}\boldsymbol{I}_{n_i}), \quad \beta_{ik}\sim \mathrm{N}(0, \alpha_{0i}^{-1}\alpha_k^{-1}), \quad (1)$$

where α_{0i} is the precision of the errors and $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})'$ is the coefficient vector for study i. To encourage sparsity and information sharing across studies, the MT-RVM further places independent Gamma priors on $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$.

$$p(\alpha|c,d)=\prod_{k=1}^{p}\mathrm{Ga}(\alpha_k|c,d)=\prod_{k=1}^{p}\frac{d^c}{\Gamma(c)}\alpha_k^{c-1}\exp(-d\alpha_k). \quad (2)$$

Similarly, Gamma priors are specified, independently, for α_{0i}, $p(\alpha_{0i} | a, b) = \mathrm{Ga}(\alpha_{0i} | a, b)$.

The MAP estimate for $\boldsymbol{\alpha}$ is defined as $\hat{\boldsymbol{\alpha}}^{MAP} = \arg\max_{\boldsymbol{\alpha}} \Sigma_k \log p(\alpha_k | c, d) + \log \mathrm{L}(\boldsymbol{t}; \boldsymbol{\alpha})$, where $\log \mathrm{L}(\boldsymbol{t}; \boldsymbol{\alpha})$ is the log-likelihood in (1) after integrating out $\boldsymbol{\beta}_i$ and α_0 with respect to their prior distributions. We have

$$\log \mathrm{L}(\boldsymbol{t};\alpha) = -\frac{1}{2}\sum_{i=1}^{S} \times \{(n_i + 2a)\log(\boldsymbol{t}_i' \boldsymbol{B}_i^{-1} \boldsymbol{t}_i + 2b) + \log|\boldsymbol{B}_i|\}, \quad (3)$$

with $\boldsymbol{B}_i = \boldsymbol{I} + \boldsymbol{x}_i' \boldsymbol{A}^{-1} \boldsymbol{x}_i$, and $\boldsymbol{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_p)$. A recommended default choice, given in Ji et al. (2009), is $a = b = c = d = 0$. Under the default choice of the hyperparameters, the MAP estimate for $\boldsymbol{\alpha}$ is equal to the maximum likelihood estimate (MLE).

The values of the hyperparameters *a, b, c,* and *d* control the shape of the priors, where small values refer to distributions with a large spike at 0 and heavy right tails. Although the default choice of the hyperparameters will result in an improper posterior distribution, the MAP estimates exist and have a sparseness-favoring property in which many of the regression coefficients will be exactly zero, with such elements shared across the different studies. We focus on the standard default noninformative prior for the error variances, which lets $a, b \to 0$. When prior information is available, this prior can be easily modified. Under the recommended choice, one obtains simultaneous variable selection across studies, while allowing heterogeneity. The value of α_k^{-1} reflects the importance of predictor *k*. In particular, for certain *k*, one obtains $\alpha \cap_k = \infty$, which implies that $\beta \cap_{i\,k} = 0$ for all *i*. As another extreme, for values of $\alpha \cap_k$ close to zero, substantial heterogeneity is allowed, with $\beta_{i\,k}$ and $\beta_{i'\,k}$ potentially very different for $i \neq i'$.

Note that α_j is closely related to the shrinkage factor. In fact, the conditional posterior distribution of $\boldsymbol{\beta}_i$ can be written as $p(\boldsymbol{\beta}_i | \boldsymbol{t}_i, \alpha_{0i}, \boldsymbol{\alpha}) \sim \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\mu_i = \alpha_{0i} \sum_i \boldsymbol{x}_i' \boldsymbol{t}_i$, $\sum_i = (\alpha_{0i} \boldsymbol{x}_i' \boldsymbol{x}_i + \boldsymbol{A})^{-1}$, and $\boldsymbol{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_p)$. When the value of α_k is close to 0, the conditional posterior distribution of $\beta_{i\,j}$ is centered close to the MLE instead of being shrunk toward zero; this suggests the importance of the *j*th variable. On the other hand, when the value of α_k is ∞, the $\beta_{i\,j}$ s will be shrunk to 0, and the *j*th variable will be excluded. In this sense, the value of α_j controls the importance of the *j*th predictor, with the predictors associated with smaller values being more important.

The key in borrowing information is to use common hyperprior variances, which occurs in the second hierarchy of MT-RVM. To further elaborate on this point, we consider the following simplified case, $t_{i\,j} \sim \mathrm{N}(\mu_i, 1)$, $\mu_i \sim \mathrm{N}(0, \alpha^{-1})$, $i = 1, \ldots, S$, $j = 1, \ldots, n$. We can write the log-likelihood for α terms sufficient statistics, $\ell(\alpha) = -\frac{1}{2}\sum_{i=1}^{S} (\log(\alpha^{-1} + 1/n) + \frac{\bar{t}_{i\cdot}^2}{\alpha^{-1} + 1/n})$, where $\bar{t}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} t_{ij}$. Differentiating $\ell(\alpha)$ with respect to α and setting the result to 0, we get the MLE for α: $\widehat{\alpha} = \frac{S}{\sum_{i=1}^{S} (\bar{t}_{i\cdot}^2 - 1/n)}$, if $\sum_{i=1}^{S} (\bar{t}_{i\cdot}^2 - 1/n) > 0$; and $\hat{\alpha} = \infty$, otherwise.

The estimation equation for α involves the sufficient statistics $\bar{t}_{i\cdot}$ from all studies. Further, if $\alpha \cap = \infty$, it allows simultaneous variable selection by setting all μ_i to 0. Borrowing information also occurs in estimation of the coefficients. In our simple example, the posterior mean of μ_i is $\frac{\alpha^{-1}}{\alpha^{-1} + 1/n}\bar{t}_{i\cdot}$, which has been shrunk toward the prior mean of zero.

In many applications, the goal of the selection process is to identify not only the predictors that are consistently important in all of the studies, but also those that are very significant in some of the studies. The example suggests that the MT-RVM is sensitive to both types of signals. A predictor will be selected whenever $\sum_{i=1}^{S}(\bar{t}_{i\cdot}^2 - 1/n) > 0$, which include the following two cases: (a) $\bar{t}_{i\cdot}^2 > 1/n$ for all i. This corresponds to the cases when μ_i is included in all the studies. It is obvious to see that $a \cap < \infty$, and thus μ_i will be selected; (b) $\bar{t}_{i\cdot}^2 \gg 1/n$ for some i. This corresponds to the cases when the signal is very strong in some studies. Again, it is clear that $a \cap < \infty$, and thus μ_i will be selected.

3. HRVM-C

3.1 Formulation

Building on the MT-RVM approach, we propose an HRVM-C method for high-dimensional variable selection in meta-analysis of survival data. We first extend the AFT model (Wei, 1992) to a multistudy AFT model as follows. Denoting the log-failure time (survival time) for subject j in study i by t_{ij}, we first model the log-failure time for each individual study by the AFT model as in (1), and then combine data from multiple studies by placing multivariate Student's t distributions as the priors for the study-specific coefficients as $p(\beta_i) \propto (2a + ab^{-1}\beta_i' \boldsymbol{A} \beta_i)^{-\frac{p+2a}{2}}$, where $\boldsymbol{A} = \mathrm{diag}(a_1, \ldots, a_p)$. For the precision parameters $\boldsymbol{a}$, we specify Gamma priors as in (2). Following West (1987), we can express the multivariate Student's t distribution as a scale mixture of normals, which leads to the MT-RVM model as in (1) and (2). The hyperparameters are set as the default values in the MT-RVM.

For censored data, the log-likelihood for $\boldsymbol{a}$ in (3) no longer holds. Here, we focus on the case of right censoring. Accounting for interval censoring will be straightforward using the same type of strategy. Denote the censored observation by y_{ij} and the censoring indicator by δ_{ij}, that is, $y_{ij} = t_{ij}$ if $\delta_{ij} = 1$, and $t_{ij} > y_{ij}$ if $\delta_{ij} = 0$. We thus observe $(\boldsymbol{y}_i, \boldsymbol{\delta}_i)$, for $i = 1, \ldots, S$, where $\boldsymbol{y}_i' = (y_{i1}, \ldots, y_{in_i})$ and $\delta_i' = (\delta_{i1}, \ldots, \delta_{in_i})$. In the rest of this article, we use $\boldsymbol{y}_i^{nc}$ and $\boldsymbol{y}_i^{c}$ to denote the vector of the noncensored observations and the censored observations, respectively, for study i. The corresponding matrices of the predictor variables are set $\boldsymbol{X}_i^{nc}$ and $\boldsymbol{X}_i^{c}$. Setting $\boldsymbol{y}^{nc} = (\boldsymbol{y}_1^{nc}, \ldots, \boldsymbol{y}_S^{nc})$, $\boldsymbol{y}^{c} = (\boldsymbol{y}_1^{c}, \ldots, \boldsymbol{y}_S^{c})$, and $\boldsymbol{y} = (\boldsymbol{y}^{nc}, \boldsymbol{y}^{c})$, the log-likelihood can be written as $\log \mathrm{L}(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{a}) = \log \mathrm{L}(\boldsymbol{y}^{nc}; \boldsymbol{a}) + \Sigma_{i,j:\delta_{ij}=0} \log \int_{t_{ij} > y_{ij}} p(t_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{a}) \mathrm{d}t_{ij}$, where $\log \mathrm{L}(\boldsymbol{y}^{nc}; \boldsymbol{a})$ is defined in (3) and $p(t_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{a})$ is defined according to (1), $(t_{ij} \mid \boldsymbol{x}_{ij}, \alpha) \sim \mathrm{N}(\boldsymbol{x}_{ij}'\beta_i, \alpha_{0i}^{-1})$. The MAP estimate for $\boldsymbol{a}$ under censoring is then defined as

$$\widehat{\alpha}^{MAP} = \arg\max_{\alpha} \left\{ \sum_k \log p(\alpha_k \mid c, d) + \log \mathrm{L}(\boldsymbol{y}, \delta; \alpha) \right\}. \quad (4)$$

After obtaining $\boldsymbol{a} \cap^{MAP}$, we keep variable k in the model as long as $\widehat{\alpha}_k^{MAP} < \infty$. In practice, some elements of $\boldsymbol{a} \cap^{MAP}$ may be large but not infinite, implying that the corresponding coefficients are very small but not exactly zero. However, the number of such large finite values is typically very small, so that the procedure tends to exhibit a thresholding behavior in which coefficients close to zero are shrunk exactly to zero.

3.2 MC-EM Algorithm for Censored Data

The optimization problem in (4) is a critical step of our method and it is challenging due to the high dimensionality of $\boldsymbol{a}$. We develop a MC-EM algorithm to solve this problem. This algorithm follows Wei and Tanner (1990) in implementing the intractable E-step using

Monte Carlo integration. In addition, similar to Tipping and Faul (2003), we develop an algorithm that breaks up the M-step into a series of alternating conditional maximization steps. Due to the dimensionality of the maximization problem, direction maximization is infeasible, but solving the high-dimensional maximization through a sequence of one-dimensional maximizations makes the computation tractable.

3.2.1 E-step—Let $\boldsymbol{t}^c$ be the complete data associated with the censored observations $\boldsymbol{y}^c$ and set $\boldsymbol{t} = (\boldsymbol{y}^{nc}, \boldsymbol{t}^c)$. In the E-step, we treat the censored observations as missing data, and define

$$Q(\alpha;\alpha^{(h-1)})=\int \log \mathrm{L}(\boldsymbol{t};\alpha)\, p(\boldsymbol{t}^c|\boldsymbol{y}^{nc}, \boldsymbol{y}^c, \alpha^{(h-1)})\mathrm{d}\boldsymbol{t}^c, \quad (5)$$

where $\mathrm{L}(\boldsymbol{t}; \boldsymbol{\alpha})$ is the likelihood for the complete data defined in (3) and $p(\boldsymbol{t}^c \mid \boldsymbol{y}^{nc}, \boldsymbol{y}^c, \boldsymbol{\alpha}^{(h-1)})$ is the posterior predictive distribution of the censored data given the MAP estimate of $\boldsymbol{\alpha}$ in the previous step (details are provided in the Appendix). The posterior predictive distribution of the censored data is the product of conditionally independent truncated univariate t distributions, which is straightforward to sample from. Let $\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(M)}$ be M independent draws of complete data generated by setting the uncensored log-survival times equal to the observed values and imputing the censored log-survival times from their conditional predictive distribution given $\boldsymbol{\alpha}^{(h-1)}$. We approximate the integral in $Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(h-1)})$ by Monte Carlo integration,

$$\begin{aligned} Q(\alpha;\alpha^{(h-1)}) &\approx \tfrac{1}{M}\sum_{m=1}^{M} \log \mathrm{L}(\boldsymbol{t}^{(m)};\alpha), \\ &= -\tfrac{1}{2M}\sum_{m=1}^{M}\sum_{i=1}^{S} \times \{(n_i+2a)\log(\boldsymbol{t}_i^{(m)'}\boldsymbol{B}_i^{-1}\boldsymbol{t}_i^{(m)}+2b)+\log|\boldsymbol{B}_i|\}, \end{aligned} \quad (6)$$

where $\boldsymbol{t}_i^{(m)}$ is the completed observations for the ith study in the mth imputed data.

3.2.2 M-step—Our goal in the M-step is to update $\boldsymbol{\alpha}$ by maximizing $Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(h-1)})$, that is, $\boldsymbol{\alpha}^{(h)} = \arg\max_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(h-1)})$. Note that this is a very challenging maximization problem due to the high dimensionality. In order to simplify this high-dimensional maximization task, we propose to use an alternating conditional maximization approach, which only requires a sequence of one-dimensional conditional optimizations. In our experience, the steps are all simple and efficient to implement, and convergence has occurred rapidly in each of the cases we have considered.

The key is to consider the dependence of the target function on a single hyperparameter, say α_k. Using results from linear algebra, we first write $|\boldsymbol{B}_i|$ and $\boldsymbol{B}_i^{-1}$ as $|\boldsymbol{B}_i|=|\boldsymbol{B}_{i,-k}||1+\alpha_k^{-1}\boldsymbol{x}_{i,k}'\boldsymbol{B}_{i,-k}\boldsymbol{x}_{i,k}|$ and $\boldsymbol{B}_i^{-1}=\boldsymbol{B}_{i,-k}^{-1}-(\boldsymbol{B}_{i,-k}^{-1}\boldsymbol{x}_{i,k}\boldsymbol{x}_{i,k}'\boldsymbol{B}_{i,-k}^{-1})/(\alpha_k+\boldsymbol{x}_{i,k}'\boldsymbol{B}_{i,-k}^{-1}\boldsymbol{x}_{i,k})$. Here, $\boldsymbol{B}_{i,k}$ denotes the matrix after removing the contribution of $\boldsymbol{x}_{i,k}$ from $\boldsymbol{B}_i$ Plugging the above facts into (6), we have

$$\begin{aligned} & Q(\alpha;\alpha^{(h-1)}) \\ &\approx -\tfrac{1}{2M}\sum_{m=1}^{M}\sum_{i=1}^{S} \times ((n_i+2a)\log(\boldsymbol{t}_i^{(m)'}\boldsymbol{B}_{i,-k}^{-1}\boldsymbol{t}_i^{(m)}+2b)+\log(|\boldsymbol{B}_{i,-k}|)) \\ &+\text{const}-\tfrac{1}{2M}\sum_{m=1}^{M}\sum_{i=1}^{S}\left(\log(1+\alpha_k^{-1}s_{i,k})+(n_i+2a)\times\log\left(1-\frac{q_{i,k,m}^2/g_{i,k,m}}{\alpha_k+s_{i,k}}\right)\right) \\ &=\ell(\alpha_{-k})+\ell(\alpha_k), \end{aligned} \quad (7)$$

where $s_{i,k}=\boldsymbol{x}_{i,k}^{'}\boldsymbol{B}_{i,k}^{-1}\boldsymbol{x}_{i,k}$, $q_{i,j,m}=\boldsymbol{x}_{i,j}^{'}\boldsymbol{B}_{i,-k}^{-1}\boldsymbol{t}_i^{(m)}$, $g_{i,k,m}=\boldsymbol{t}_i^{(m)'}\boldsymbol{B}_{i,-k}^{-1}\boldsymbol{t}_i^{(m)}$, and $\boldsymbol{\alpha}_{-k}$ is the resulting vector of $\boldsymbol{\alpha}$ after removing the kth component.

Differentiating $Q(\boldsymbol{\alpha};\boldsymbol{\alpha}^{(h-1)})$ with respect to α_k and setting the result to zero, we have,

$$\frac{\partial Q(\alpha;\alpha^{(h-1)})}{\alpha_k}\approx\frac{\partial\ell(\alpha_k)}{\alpha_k}$$
$$=-\frac{1}{2M}\sum_{m=1}^{M}\sum_{i=1}^{S}\times\frac{(n_i+2a)q_{i,k,m}^2/g_{i,k,m}-s_{i,k}-s_{i,k}\,(s_{i,k}-q_{i,k,m}^2/g_{i,k,m})/\alpha_k}{(\alpha_k+s_{i,k})\,(\alpha_k+s_{i,k}-q_{i,k,m}^2/g_{i,k,m})}$$
$$=0.$$

Assuming $\alpha_k \ll s_{i,k}$, we obtain an estimate of α_k that maximizes the conditional log-likelihood with respect to the kth dimension,

$$\widehat{\alpha}_k\approx\frac{MS}{\sum_{m=1}^{M}\sum_{i=1}^{S}\frac{(n_i+2a)q_{i,k,m}^2/g_{i,k,m}-s_{i,k}}{s_{i,k}(s_{i,k}-q_{i,k,m}^2/g_{i,k,m})}},\quad \text{if}\ \sum_{m=1}^{M}\sum_{i=1}^{S}\frac{(n_i+2a)q_{i,k,m}^2/g_{i,k,m}-s_{i,k}}{s_{i,k}(s_{i,k}-q_{i,k,m}^2/g_{i,k,m})}>0,$$
$$\widehat{\alpha}_k=\infty,\quad \text{otherwise.} \tag{8}$$

In each iteration of the M-step, we first calculate $\alpha\cap_k$ and the conditional likelihood $\ell(\alpha\cap_k)$ for all k, and then update the kth element in $\boldsymbol{\alpha}$ which has the maximum $\ell(\alpha\cap_k)$ among all k. Performing the above local maximization iteratively for varying k until convergence, we obtain a simple and seemingly (in the cases we have considered) efficient algorithm for the M-step. In practice, we monitor convergence by specifying a threshold η_2, and stop the M-step when the change in $\max_k \ell(\alpha\cap_k)$ is less than η_2.

The proposed optimization strategy is a version of alternating conditional maximization, which is well known to converge to a local mode in the likelihood surface. In each iteration, the algorithm leads to one of the three operations: (a) If $\alpha_k^{(h-1)}<\infty$ and the algorithm sets $\alpha_k^{(h)}=\infty$, we have $\beta_{i\,k}=0$ for all i, thereby removing the kth predictor from the model in all studies; (b) If $\alpha_k^{(h-1)}=\infty$ and the algorithm sets $\alpha_k^{(h)}<\infty$, we have $\beta_{i\,k}\neq 0$ for all i, thereby adding the kth predictor to the model in all studies; (c) If $\alpha_k^{(h-1)}<\infty$ and the algorithm sets $\alpha_k^{(h)}<\infty$, we have simply re-estimated the hyperparameter value and the kth predictor remains in the model. From expression (8), it is clear that $\alpha_k^{(h)}$ can take a value of exactly ∞ and each of the operations (a)–(c) is possible prior to convergence.

We iterate between the E-step and the M-step until the MC-EM algorithm converges, which is judged to have occurred when the change in the maximized log-likelihood between iterations is less than the threshold η_1. After convergence, the kth predictor will be excluded from all studies if $\alpha\cap_k=\infty$. If $\alpha\cap_k$ is finite but large, then the coefficients for the kth predictor will greatly shrink toward zero, and thus tends to be small in all the studies. If $\alpha\cap_k$ is small, the coefficients for the kth predictor will be shrunk less and the model allows for substantial heterogeneity in the kth predictor across different studies.

4. Simulation Study

4.1 Comparison with Existing Methods

In this section, we assess the performance of HRVM-C and compare it with popular existing methods for variable selection in meta-analysis. Because HRVM-C is the only method that accounts for censoring, the comparison with the other methods is carried out in the setting of complete data. We then consider separately how HRVM-C performs with censored data.

We first consider the Group Lasso method (Grp-Lasso) of Yuan and Lin (2006), which is applied by augmenting the model as $\mathrm{E}(\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_S')' = \mathrm{diag}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_S)(\beta_1', \ldots, \beta_S')'$, where for a given k, we define β_{ik} ($i = 1, \ldots, S$) as a group, that is, the regression coefficients in different studies for a given gene. In addition, the following generic methods are considered for combining multiple studies in high-dimensional variable selection problems:

- Fitting each study independently and then reporting the union of the selected predictor variables for each study (Ind).
- Reducing the false positive rate by a multistage analysis (MSA), that is, only considering the selected predictor variables in the analysis of the follow-up data sets.
- Fitting with pooled data (Pool).

For each individual study, we consider the following variable selection techniques: RVM, Lasso, and using p-values from simple linear regression models (Pvals). For the Pvals method, predictors are ranked according to their p-values each of which is obtained from a simple linear regression model only including that predictor. We thus consider 11 procedures: HRVM-C, Grp-Lasso, Ind-RVM, Ind-Lasso, Ind-Pvals, MSA-RVM, MSA-Lasso, MSA-Pvals, Pool-RVM, Pool-Lasso, and Pool-Pvals.

The simulation is set up to mimic the sample size of the real data in Section 5. In particular, we simulate three related studies with 226 subjects in study 1, 156 subjects in study 2, and 101 subjects in study 3. We fix p, the total number of predictors, at 1000 and then randomly choose $p_0 = 20$ of them to be related to the survival time, with the regression coefficients simulated, independently, from a uniform $\mathrm{U}([-1, -0.1] \cup [0.1, 1])$ distribution.

In simulating the predictor variables for microarray data, we use the strategy of Gui and Li (2005) and Sha, Tadesse, and Vannucci (2006), which runs as follows. For a study with sample size n, we first draw an $n \times n$ matrix A from a uniform $\mathrm{U}(-1.5, 1.5)$ distribution and randomly choose p_0 columns of $\boldsymbol{A}$ to be relevant to the survival time. The orthonormal basis of $\boldsymbol{A}$, constructed by Gram–Schmidt orthonormalization, is then obtained as $\{\xi_1, \ldots, \xi_{p0}, \zeta_1, \ldots, \zeta_{p-p0}\}$, where $\{\xi_1, \ldots, \xi_{p0}\}$ is an orthonormal basis for the p_0 columns that are relevant to the survival time. Let $\boldsymbol{T}$ be a $p_0 \times (n - p_0)$ matrix such that the largest eigenvalue of $\boldsymbol{T}'\boldsymbol{T}$ is ρ^2. By Cauchy's inequality, for any vector in the linear space spanned by $\{\xi_1, \ldots, \xi_{p0}\}$ and any vector in the linear space spanned by $\zeta + \xi T$, the maximum correlation between them is less than or equal to $\rho/\sqrt{1+\rho^2}$. We thus generate the remaining $p - p_0$ variables not relevant to the survival time from the linear space $\zeta + \xi T$. The log-survival time is then generated using the AFT model according to (1), where $\varepsilon_{ij} \sim \mathrm{N}(0, 1)$.

We choose the value of ρ such that the maximum correlation is controlled at 0 (low level), 0.5 (medium level), and 0.9 (high level). For each chosen ρ, we generate 100 data sets, and select variables by the 11 considered methods, respectively. In fitting HRVM-C and RVM, we set the hyperparameters $a, b, c,$ and d at their recommended values. The tuning parameter in Grp-Lasso is chosen as the maximal value of the penalty parameter in Group Lasso. For

the thresholds required in the Lasso and Pvals methods, we utilize sliding thresholds that keep the most influential 5, 10, 15, 20, 25, 30, 35, 40, 45, or 50 predictors in each study. For a fair comparison, we also apply the same thresholds to HRVM-C and RVM methods. The final selected predictors in Ind-RVM, Ind-Lasso, and Ind-Pvals are then the union set of the selected predictors after thresholding in all studies, while the predictors reported in MSA-RVM, MSA-Lasso, MSA-Pval are obtained by keeping the predictors after thresholding in the first stage and then restricting attention to those kept predictors in the subsequent analysis. Finally, in Pool-RVM, Pool-Lasso, and Pool-Pvals, we pool the data together and select the most influential predictors after thresholding.

To evaluate the performance of each procedure, we consider the true positive rate (TPR) and the false discovery rate (FDR), which are defined as the ratio of the number of correctly identified predictors and the total number of truly active predictors, and the ratio of the number of the falsely identified predictors and the total number of selected predictors, respectively.

The FDR is the error rate in variable selection and is typically controlled at a prespecified level. We report, in Table 1, the TPRs for the 11 considered approaches, based on the average of 100 simulations, while controlling the FDRs at 0.05 in all methods. HRVM-C, with TPRs about 0.97, clearly outperforms all the other methods. Table 1 also reflects the drawbacks of the other methods. First, we note that in the Grp-Lasso, the model tends to put a large penalty in the nonzero coefficients due to the large number of groups, leading to a final model that is too sparse. Indeed, from our experience, Grp-Lasso can only identify one or two predictors each time, while missing most predictors. Therefore, the TPRs for Grp-Lasso is very small, as shown in the table. Second, in the independent methods, after FDR control, few predictors are selected and thereby the methods still miss important predictors. Third, the multistage approaches, which control the FDR in the first-stage study, miss important predictors that are significant only after the second-stage study. Finally, by pooling the data together, we may miss the predictors that are positively correlated with the response in one study, but are negatively correlated with the response in another study. Apparently, the advantage of HRVM-C over the other methods is consistent under all correlation levels being considered.

4.2 Censored Case

In this section, we conduct simulation studies to investigate the performance of HRVM-C in the presence of censoring. We first simulate the complete log-survival time t_{ij} as in Section 4.1. We then generate the censored observations using the strategy described in Sha et al. (2006), as follows. Given the censoring rate λ, we first set the censoring indicators δ_{ij} for the first 100 * λ percentage of subjects in each study to 0. If $\delta_{ij} = 0$, we observe a censored data y_{ij} from the distribution $\exp(y_{ij}) \sim \text{Uniform}(0, \exp(t_{ij}))$. Otherwise, we observe the noncensored data t_{ij}. The simulation leads to a right-censoring data set. In this simulation, the total number of predictor variables p are fixed at 1000, 20 of which are related to the survival time (these are the same as in Section 4.1). For the censoring rate λ, we consider low censoring rate $\lambda = 0.1$, medium censoring rate $\lambda = 0.5$, and high censoring rate $\lambda = 0.9$.

We apply HRVM-C to 100 simulated data sets. To avoid numerical problems, we choose $a = 10^{-4}$ and $b = 10^{-4}$. In the EM algorithm, we impute 1000 complete data sets within each E-step, and the two thresholds are chosen as $\eta_1 = 0.01$ and $\eta_2 = 0.01$. For each data set, we start the EM algorithm from the empty set, that is, no predictors are included in the model. After convergence, we rank the selected predictors based on the values of the shrinkage factors α_k and apply the same sliding thresholds as in Section 4.1. In Table 2, we report the TPRs based on the average of 100 replications while controlling the FDRs at 0.05. For low censoring rate, the performance of HRVM-C, with TPRs about 0.95, is very good under all

correlation levels. As expected, as the censoring rat increases, the performance gets worse. Nevertheless, with TPRs around 0.75 for the medium censoring rate and around 0.5 for the high censoring rate, HRVM-C performs reasonably well when data are censored. Finally, in the last column of Table 2, we report the CPU time for one replicate under the varying censoring rates and correlation levels. The computation is inexpensive. We also note that the computational time decreases as the censoring rate gets higher. From our experience, this is due to the fact that the M-step needs fewer iterations to converge when the censoring rate increases.

In Table 3, we summarize the frequencies with which the 20 predictors that related to the response are selected by HRVM-C. As expected, the predictors with larger coefficients are more likely to be selected. We also note that the result does not vary much as the maximum correlation between related predictors and the unrelated predictors increases. This is an advantage of borrowing strength from all studies. Two predictors that are highly correlated in one specific study are not necessarily correlated in other studies. Therefore, by incorporating all information by hierarchical modeling, we are able to avoid the collinearity issues in any single study. Finally, as the censoring rate increases, predictors with smaller coefficients have less chances to be selected.

5. Analysis of the Gene Expression Barcode Data

We demonstrate our method with the gene expression bar-code data in Zilliox and Irizarray (2007). The data consist of three breast cancer studies (A3ymetrix HGU133A array) in Miller et al. (2005), Pawitan et al. (2005), and Sotiriou et al. (2006), that include patient survival data. There are 243 subjects in the first study, 156 in the second study, and 101 in the third study. In the first study, 52 patients are censored and 15 have missing data in the survival status. All observations in the second study are censored and in the third study, 61 observations are censored. Here, we focus on gene selection by HRVM-C, and hence remove the patients with missing data from consideration.

The gene expression profile consists of 22, 215 genes. To remove the variability in the gene expression profile between different studies, we use the gene barcode of the microarray data as our predictor variables. Many genes are in the same status (barcode is 1 or 0) among all the subjects. To avoid identifiability issues arising from including those genes in the model, we eliminate them from consideration. This reduces the total number of genes to 11, 879. Our goal is to select the genes that affect the patient survival time.

We perform the gene selection by the HRVM-C for the gene barcode data, under the choice of $a = 10^{-4}$, $b = 10^{-4}$, $N = 1000$, $\eta_1 = 0.001$, and $\eta_2 = 0.001$. From our experience, the result does not appear to be sensitive to the choices of a, b, and N. On the other hand, the choice of η_1 and η_2 are critical. These thresholds control how long the algorithm would run. In general, one would want to avoid values that are too small, which lead to bad convergence of the algorithm. Larger values, on the other hand, could make the algorithm run longer. We have tried a variety of values, and the ones we chose seem to provide a reasonable balance between convergence and running time. At the end of the algorithm, 34 genes are selected to be related to the survival time.

Finally, we check the biological meanings of the selected genes. Most of the selected genes are known to be cancer related. In Table 4, we list the information of those genes that are selected by our method.

6. Discussion

In this article, we develop the HRVM-C to combine multiple studies in high-dimensional variable selection. In contrast with the commonly used approaches, our method systematically borrows information across the studies using an explicit overall statistical model. For model fitting with censored data, we develop an MC-EM algorithm that can be implemented quickly even in high dimensions. In simulations studies, our method is found to outperform existing approaches in the setting of complete data, and to perform well with censored data. We demonstrate the usefulness of our method in a meta-analysis of multiple breast cancer studies.

HRVM-C provides a useful tool for dealing with censored data across multiple studies in high-dimensional variable selection problems. The MC-EM algorithm developed here can be easily extended to a Monte Carlo Expectation Conditional Maximization (MC-ECM) algorithm where the M-step is replaced by the Conditional Maximization (CM) step (Meng and Rubin, 1993). According to (8), a_k can be (approximately) maximized conditional on the rest of parameters, and therefore MC-ECM algorithm has the potential to further reduce the computational expense. HRVM-C requires at least one noncensored observation in each study in order to impute the complete data sets (see the discussion in Appendix). To overcome this difficulty, one may consider to utilize a similar structure in the Cox proportional hazards model. We will explore this extension in our future work.

7. Supplementary Material

Further details and complete information needed to recapitulate the analyses reported are available at the *Biometrics* website http://www.biometrics.tibs.org. This includes the Matlab code to conduct the analysis and a brief readme on use of the code.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by the Statistical and Applied Mathematical Sciences Institute (SAMSI) Summer 2008 research program on *Meta-analysis: Synthesis and Appraisal of Multiple Sources of Empirical Evidence*. The gene barcode data used in this article were kindly provided by Dr Rafael Irizarry and Dr Michael Zilliox. The authors gratefully acknowledge the many helpful comments received from the editor, the associate editor, and the two anonymous referees. Research of Fei Liu was partially supported by the University of Missouri-Columbia research board award. Research of Fei Zou was partially supported by NIH (R01GM074175). Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NIH.

References

Beckly J, Hancock L, Geremia A, Cummings J, Morris A, Cooney R, Pathan S, Guo C, Jewell D. Two-stage candidate gene study of chromosome 3p demonstrates an association between nonsynomous variants in the mst1r gene and Crohn's disease. Inflammatory Bowel Diseases. 2008; 14:500–507. [PubMed: 18200509]

Bovelstad H, Nygard S, Storvold H, Aldrin M, Borgan O, Frigessi A, Lingjaerde O. Predicting survival from microarray data—a comparative study. Bioinformatics. 2007; 23:2080–2087. [PubMed: 17553857]

Buckley J, James I. Linear regression with censored data. Biometrika. 1979; 66:429–436.

Chan S, Griffith O, Tai I, Jones S. Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. Cancer Epidemiology Biomarkers and Prevention. 2008; 17:543–552.

Datta S, Le-Rademacher J, Datta S. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and lasso. Biometrics. 2007; 63:259–271. [PubMed: 17447952]

George E, McCulloch R. Approaches for Bayesian variable selection. Statistica Sinica. 1997; 7:339–373.

Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics. 2005; 21:3001–3008. [PubMed: 15814556]

Hans C, Dobra A, West M. Shotgun stochastic search for "large p" regression. Journal of the American Statistical Association. 2007; 102:507–516.

Ji S, Dunson D, Carin L. Multitask compressive sensing. IEEE Transactions on Signal Processing. 2009; 57:92–106.

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. New York: Wiley; 1980.

Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data. Annals of Statistics. 1981; 9:1276–1288.

Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. 1993; 80:267–278.

Miller M, Wang C, Parisini E, Coletta R, Goto R, Lee S, Barral D, Townes M, Roura-Mir C, Ford H, Brenner M, Dascher CC. Characterization of two avian MHC-like genes reveals an ancient origin of the cd1 family. Proceedings of National Academy of Science, USA. 2005; 102:8674–8679.

Pawitan Y, Bjohle J, Amler L, Borg A, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu E, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw P, Smeds J, Skoog L, Wedren S, Bergh J. Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. Breast Cancer Research. 2005; 7:R953–R964. [PubMed: 16280042]

Sha N, Tadesse MG, Vannucci M. Bayesian variable selection for the analysis of microarray data with censored outcomes. Bioinformatics. 2006; 22:2262–2268. [PubMed: 16845144]

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh, Piccart M, Delorenzi M. Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute. 2006; 98:262–272. [PubMed: 16478745]

Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B. 1996; 58:267–288.

Tibshirani R. The Lasso method for variable selection in the Cox model. Statistics in Medicine. 1997; 16:385–395. [PubMed: 9044528]

Tipping ME. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research. 2001; 1:211–244.

Tipping, ME.; Faul, AC. Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics. Key West, FL: 2003. Fast marginal likelihood maximisation for sparse Bayesian models.

Wang S, Nan B, Zhu J, Beer D. Doubly penalized Buckley-James method for survival data with high-dimensional covariates. Biometrics. 2008; 64:132–140. [PubMed: 17680828]

Wei GCG, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. Journal of the American Statistical Association. 1990; 85:699–704.

Wei LJ. The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. Statistics in Medicine. 1992; 11:1871–1879. [PubMed: 1480879]

West M. On scale mixtures of normal distributions. Biometrika. 1987; 74:646–648.

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B. 2006; 68:49–67.

Zhang H, Lu W. Adaptive Lasso for Cox's proportional hazards model. Biometrika. 2007; 94:691–703.

Zilliox M, Irizarray R. A gene expression bar code for microarray data. Nature Methods. 2007; 4:911–913. [PubMed: 17906632]

Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B. 2005; 67:301–320.

Appendix

Imputing Missing Data

The conditional distribution for the censored observation can be written as

$$p(t_{ij}|\delta_{ij}=0, y_{ij}, \alpha_{0i}, \alpha, \boldsymbol{y}^{nc})=I(t_{ij}>y_{ij})\times\int p(t_{ij}|\boldsymbol{y}_i^{nc}, \alpha_{0i}, \beta_i)p(\beta_i|\boldsymbol{y}_i^{nc}, \alpha_{0i}, \alpha)\times p(\alpha_{0i}|\boldsymbol{y}_i^{nc}, \alpha)\mathrm{d}\beta_i\mathrm{d}\alpha_{0i}, \quad (A.1)$$

where $I(\cdot)$ is the indicator function of the set $t_{ij} > y_{ij}$, $(\beta_i|\boldsymbol{y}_i^{nc}, \alpha_{0i}, \alpha)\sim \mathrm{N}(\mu_i, \alpha_{0i}^{-1}\sum_i)$, with $\mu_i=(\boldsymbol{X}_i^{nc'}\boldsymbol{X}_i^{nc}+\boldsymbol{A})^{-1}\boldsymbol{X}_i^{nc'}\boldsymbol{y}_i^{nc}$, and $\sum_i^{-1}=\boldsymbol{X}_i^{nc'}\boldsymbol{X}_i^{nc}+\boldsymbol{A}$. Integrating out $\boldsymbol{\beta}_i$ and α_{0i}, we have

$p(t_{ij}|\delta_{ij}=0, y_{ij}, \boldsymbol{y}_i^{nc}, \alpha) \propto I(t_{ij}>y_{ij})\{2b+\boldsymbol{y}_i^{nc'}(\boldsymbol{I}_{n_i}+\boldsymbol{X}_i^{nc}\boldsymbol{A}^{-1}\boldsymbol{X}_i^{nc'})^{-1}\boldsymbol{y}_i^{nc}+(t_{ij}-\boldsymbol{x}_{ij}\mu_i)^2/(1+\boldsymbol{x}_{ij}\sum_i^{-1}\boldsymbol{x}_{ij}')\}^{-\frac{n_i-\sum_{j=1}^{n_i}\delta_{ij}+2a}{2}}$. This is a truncated noncentral Student's t distribution, with the degree of freedom $n_i-\sum_{j=1}^{n_i}\delta_{ij}+2a-1$, the location parameter $\boldsymbol{x}_{ij}\boldsymbol{\mu}_i$, and the scale parameter $[(1+\boldsymbol{x}_{ij}\sum_i^{-1}\boldsymbol{x}_{ij}')\{2b+\boldsymbol{y}_i^{nc'}(\boldsymbol{I}_{n_i}+\boldsymbol{X}_i^{nc'}\boldsymbol{A}^{-1}\boldsymbol{X}_i^{nc})^{-1}\boldsymbol{y}_i^{nc}\}/(n_i-\sum_{j=1}^{n_i}\delta_{ij}+2a-1)]^{-1/2}$. The distribution of $\boldsymbol{t}^c$ is thus given as $p(\boldsymbol{t}^c|\boldsymbol{y}^{nc}, \boldsymbol{y}^c, \alpha^{(h-1)})=\Pi_{i,j:\delta_{ij}=0}p(t_{ij}|\delta_{ij}=0, y_{ij}, \boldsymbol{y}_i^{nc}, \alpha)$. At each E-step, given the current value of $\boldsymbol{\alpha}$, we obtain a complete data set by sampling the censored observations from this distribution.

Note that the above distribution is well defined only when the degree of freedom $n_i-\sum_{j=1}^{n_i}\delta_{ij}+2a-1>0$. This implies that there must be at least one noncensored observation in any single study. For this reason, HRVM-C cannot incorporate studies where all the observations are censored.

Table 1

True positive rate (TPR) for the noncensored simulation studies. The false discovery rate (FDR) is controlled at 0.05. The maximum correlation is 0 (Low), 0.5 (Medium), and 0.9 (High). The results are based on 100 replicated simulations.

	Low	Medium	High
HRVM-C	0.9745	0.9685	0.9725
Grp-Lasso	0.0205	0.0195	0.026
Ind-RVM	0.7675	0.7430	0.7580
Ind-Lasso	0.8515	0.8445	0.8495
Ind-Pvals	0.8540	0.8485	0.8545
MSA-RVM	0.7825	0.7435	0.7450
MSA-Lasso	0.6495	0.6490	0.6495
MSA-Pvals	0.6165	0.6125	0.6125
Pool-RVM	0.7395	0.7150	0.7215
Pool-Lasso	0.7400	0.7375	0.7380
Pool-Pvals	0.7415	0.7360	0.7370

Table 2

TPRs and CPU running time (in seconds) in the censored simulation studies. The FDRs are controlled at 0.05. The first column is the level of the censoring rate: 0.1 (Low), 0.5 (Medium), 0.9 (High). The second column is the level of the maximum correlation between variables: 0 (Low), 0.5 (Medium), 0.8 (High). The results are based on the average of 100 simulated data sets.

Censoring rate	Correlation	TPR	CPU time
Low	Low	0.9555	183.2900
	Medium	0.9595	190.0400
	High	0.9510	188.7900
Medium	Low	0.7490	98.2700
	Medium	0.7475	95.6500
	High	0.7495	89.4200
High	Low	0.4895	25.8700
	Medium	0.4955	24.9900
	High	0.4930	31.5000

Table 3

Percentage of times that the predictors have been selected in the 100 replications in the censored simulation studies, when the censoring rate λ is 0.1, 0.5, and 0.9 and ρ is 0 and 4/3 (this corresponding to the maximum correlation between active predictors and inactive predictors are 0 and 0.8.) The first column shows the true coefficients of the 20 predictors in the three studies, which are related to the survival time.

	(λ, ρ)					
Coefficients	**(0.1, 0)**	**(0.1, 4/3)**	**(0.5, 0)**	**(0.5, 4/3)**	**(0.9, 0)**	**(0.9, 4/3)**
(−0.97, −0.36, −0.30)	0.96	0.94	0.83	0.87	0.34	0.41
(−0.35, −0.23, −0.81)	1.00	1.00	1.00	0.99	0.50	0.44
(0.40, 0.50, 0.69)	1.00	1.00	1.00	1.00	0.69	0.78
(0.61, −0.68, 0.39)	1.00	0.99	0.86	0.91	0.56	0.40
(−0.59, −0.43, 0.78)	1.00	1.00	1.00	1.00	0.83	0.84
(0.26, 0.23, 0.44)	1.00	1.00	0.88	0.87	0.21	0.16
(−0.21, −0.33, −0.46)	1.00	0.99	0.93	0.92	0.19	0.25
(0.35, −0.43, −0.11)	0.76	0.68	0.55	0.51	0.11	0.12
(−0.48, −0.67, −0.90)	1.00	1.00	1.00	1.00	0.89	0.90
(0.51, 0.16, 0.77)	1.00	1.00	0.98	1.00	0.68	0.72
(−0.25, 0.64, 0.66)	1.00	1.00	1.00	0.99	0.67	0.65
(−0.28, 0.25, 0.97)	1.00	1.00	1.00	1.00	0.46	0.38
(0.95, −0.94, 0.81)	1.00	1.00	1.00	1.00	0.90	0.88
(0.89, 0.32, −0.94)	1.00	1.00	1.00	1.00	0.98	0.96
(−0.66, −0.15, −0.25)	0.90	0.92	0.63	0.80	0.26	0.29
(−0.59, 0.36, 0.74)	1.00	1.00	1.00	1.00	0.73	0.84
(0.80, −0.65, 0.83)	1.00	1.00	1.00	1.00	0.89	0.86
(0.64, 0.98, −0.96)	1.00	1.00	1.00	1.00	0.95	0.97
(−0.75, −0.16, −0.25)	0.93	0.97	0.76	0.79	0.39	0.25
(−0.85, −0.48, −0.83)	1.00	1.00	1.00	1.00	0.88	0.90

Table 4

Information about the selected genes

A3yID	Gene symbol	Description
201167 x at	ARHGDIA	Rho GDP dissociation inhibitor (GDI) alpha
200969 at	SERP1	Stress-associated endoplasmic reticulum protein 1
201280 s at	DAB2	Disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)
200008 s at	GDI2	GDP dissociation inhibitor 2
200958 s at	SDCBP	Syndecan binding protein (syntenin)
201341 at	ENC1	Ectodermal-neural cortex (with BTB-like domain)
201384 s at	NBR1	Neighbor of BRCA1 gene 1
201399 s at	TRAM1	Translocation associated membrane protein 1
160020 at	MMP14	Matrix metallopeptidase 14 (membrane-inserted)
200957 s at	SSRP1	Structure specific recognition protein 1
201275 at	FDPS	Farnesyl diphosphate synthase
200994 at	IPO	Importin 7
200835 s at	MAP4	Microtubule-associated protein 4
200902 at	SEP15	15 kDa selenoprotein
201404 x at	PSMB2	Proteasome (prosome, macropain) subunit, beta type, 2
200626 s at	MATR3	Matrin 3
200923 at	LGALS3BP	Lectin, galactoside-binding, soluble, 3 binding protein
201087 at	PXN	Paxillin
201040 at	GNAI2	Guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2
201264 at	COPE	Coatomer protein complex, subunit epsilon
200744 s at	GNB1	Guanine nucleotide binding protein (G protein), beta polypeptide 1
200672 x at	SPTBN1	Spectrin, beta, non-erythrocytic 1
200914 x at	KTN1	Kinectin 1 (kinesin receptor)
200607 s at	RAD21	RAD21 homolog (S. pombe)
201091 s at	CBX3	Chromobox homolog 3 (HP1 gamma homolog, Drosophila)
200749 at	RAN	Member RAS oncogene family
201316 at	PSMA2	Proteasome (prosome, macropain) subunit, alpha type, 2
201343 at	Hs.693967	Transcribed locus
201041 s at	DUSP1	Dual specificity phosphatase 1
201069 at	MMP2	Matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
200962 at	RPL31	Ribosomal protein L31
201129 at	SFRS7	Splicing factor, arginine/serine-rich 7, 35kDa
201291 s at	TOP2A	Topoisomerase (DNA) II alpha 170kDa
200920 s at	BTG1	B-cell translocation gene 1, anti-proliferative