Biometrika (2012), **99**, 1, *pp.* 1–15

Printed in Great Britain

Advance Access publication on 31 July 2012

Sparse Envelope Model: Efficient Estimation and Response Variable Selection in Multivariate Linear Regression

BY Z. SU, G. ZHU

Department of Statistics, University of Florida, 102 Griffin-Floyd Hall, Gainesville, Florida, 32611, U.S.A.

zhihuasu@stat.ufl.edu gzhu22@ufl.edu

X. CHEN

Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, 117546, Singapore

stacx@nus.edu.sg

AND Y. YANG

Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, Quebec, H3A 0B9, Canada

yi.yang6@mcgill.ca

SUMMARY

The envelope model is a method for efficient estimation in multivariate linear regression. In this article, we propose the sparse envelope model, which is motivated by applications where some response variables are invariant to changes of the predictors and have zero regression coefficients. The envelope estimator is consistent but not sparse, and in many situations it is important to identify the response variables for which the regression coefficients are zero. The sparse envelope model performs variable selection on the responses and preserves the efficiency gains offered by the envelope model. Response variable selection arises naturally in many applications, but has not been studied as thoroughly as predictor variable selection. In this article, we discuss response variable selection in both the standard multivariate linear regression and the envelope contexts. In response variable selection, even if a response has zero coefficients, it still should be retained to improve the estimation efficiency of the nonzero coefficients. This is different from the practice in predictor variable selection. We establish consistency, the oracle property and obtain the asymptotic distribution of the sparse envelope estimator.

Some key words: Canonical correlation, Dimension reduction, Envelope model, Grassmann manifold, Oracle property

1. INTRODUCTION

1·1. *Background*

Throughout the article, we consider multivariate linear regression

$$Y = \alpha + \beta(X - \mu_X) + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}^r$ is a multivariate response vector, $X \in \mathbb{R}^p$ denotes the vector of random predictors with mean $\mu_X \in \mathbb{R}^p$ and covariance matrix $\Sigma_X \in \mathbb{R}^{p\times p}$. The error vector $\varepsilon \in \mathbb{R}^r$ has mean 0 and

positive definite covariance matrix $\Sigma \in \mathbb{R}^{r\times r}$, and is independent of the predictor vector X. The intercept $\alpha \in \mathbb{R}^r$ and regression coefficients $\beta \in \mathbb{R}^{r\times p}$ are unknown parameters.

The standard approach estimates each row of β separately by regressing the corresponding element of Y on X, and relationships among the elements of Y are not used. The envelope model (Cook et al., 2010) makes use of the stochastic relationships among the elements of Y, and identifies a part of the response that is immaterial to changes in X. Excluding this immaterial part in the estimation of β leads to gains in efficiency. Building on the development in Cook et al. (2010), several papers have applied the idea of enveloping to more general contexts, and have proposed new models to achieve even greater gains in efficiency; see, e.g. Su & Cook (2011), Cook & Su (2013), and Cook & Zhang (2015). Moreover, a connection between the envelope model and partial least squares that has allowed for a new understanding of the working mechanism of partial least squares was established by Cook et al. (2013).

Compared to predictor variable selection, the literature on response variable selection is limited. Response variable selection is motivated by applications in which some response variables do not depend on any of the predictors and have zero regression coefficients. For example, the expression levels for some genes of the fission yeast *Schizosaccharomyces pombe* show little variation in a cell cycle while the expression levels for other genes have large variation, see Section 3·2. Finding inactive response variables can lead to more interpretable results and also improve estimation efficiency; see Section 2·5. The standard procedure for identifying inactive responses is to evaluate, for $i = 1, \ldots, r$, whether Y_i depends on X via the F test, adjusting for multiple testing (see, e.g. Benjamini & Yekutieli 2001). However, since the relationship between the response variables is not used, this procedure is not efficient, as is demonstrated in the simulations in Section 3·1.

In this article, we develop a sparse envelope model that performs response variable selection efficiently under the envelope model. We also discuss issues in response variable selection, especially how to use the inactive responses to improve estimation efficiency for nonzero regression coefficients. Our theoretical discussion addresses both large-sample and high-dimensional scenarios. Throughout the article, we assume that the number of predictors p is fixed and smaller than the sample size n. If p is large, we can apply a standard approach like the lasso to reduce p before applying our method.

We use P_A to indicate the projection matrix onto A or $\mathrm{span}(A)$ if A is a subspace or a matrix, and $Q_A = I - P_A$. The symbol $\sim$ stands for equality in distribution. If V_1 and V_2 are random variables, $V_1 \perp\!\!\!\perp V_2$ indicates that they are independent. The L_2 norm of a vector v is denoted by $\|v\|_2$. For a matrix M, we use $\|M\|$ for its spectral norm and $\|M\|_F$ for its Frobenius norm. The operator vec stacks a matrix into a vector column-wise. The Kronecker product for matrices A and B is indicated by $A \otimes B$. A notation table is in the Supplement.

1·2. *Envelopes*

Let $(\Gamma, \Gamma_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix. Then Y can be decomposed into two parts, $P_\Gamma Y$ and $Q_\Gamma Y$. We assume that these satisfy the conditions: (i) $Q_\Gamma Y \mid X \sim Q_\Gamma Y$ and (ii) $\mathrm{cov}(P_\Gamma Y, Q_\Gamma Y \mid X) = 0$. Condition (i) implies that the distribution of $Q_\Gamma Y$ does not depend on X. So $Q_\Gamma Y$ does not carry any information about β. Condition (ii) implies that $Q_\Gamma Y$ does not carry any information about β through its conditional correlation with $P_\Gamma Y$. Together these conditions imply that $Q_\Gamma Y$ does not carry any information about β directly or indirectly, and therefore $Q_\Gamma Y$ is immaterial to the regression. Thus we call $P_\Gamma Y$ and $Q_\Gamma Y$ the material part and immaterial part, respectively. Cook et al. (2010) showed that (i) and (ii) are equivalent to the following conditions: (a) $\mathcal{B} \subseteq \mathrm{span}(\Gamma)$, where $\mathcal{B} = \mathrm{span}(\beta)$, and (b) $\Sigma = \Sigma_1 + \Sigma_2 = P_\Gamma \Sigma P_\Gamma + Q_\Gamma \Sigma Q_\Gamma$. When (b) holds, $\mathrm{span}(\Gamma)$ is a reducing subspace of Σ (Conway, 2013, Sec-

tion 2·3). The Σ-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_\Sigma(\mathcal{B})$, is defined as the smallest reducing subspace of Σ that contains $\mathcal{B}$ (Cook et al., 2010). Consequently, $\mathcal{E}_\Sigma(\mathcal{B})$ decomposes Σ into variation related to the material and immaterial parts of Y: $\Sigma_1 = \mathrm{var}(P_\Gamma Y \mid X)$ and $\Sigma_2 = \mathrm{var}(Q_\Gamma Y)$. We call (1) an envelope model when conditions (a) and (b) are imposed. Because β is related only to the material variation, the decomposition of Σ suggests that excluding the immaterial information makes estimation of β more efficient. In particular, massive efficiency gains can be obtained when $\|\Sigma_2\| \gg \|\Sigma_1\|$. Based on (a) and (b), the coordinate form of the envelope model is

$$Y = \alpha + \Gamma\eta(X - \mu_X) + \varepsilon, \quad \Sigma = \Sigma_1 + \Sigma_2 = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T, \tag{2}$$

where $\beta = \Gamma\eta$, $\Gamma \in \mathbb{R}^{r\times u}$ is an orthogonal basis for $\mathcal{E}_\Sigma(\mathcal{B})$, Γ_0 is a completion of Γ, and u is the dimension of $\mathcal{E}_\Sigma(\mathcal{B})$. The matrix $\eta \in \mathbb{R}^{u\times p}$ holds the coordinates of β relative to Γ, and $\Omega \in \mathbb{R}^{u\times u}$ and $\Omega_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are positive definite. If $u = r$, then $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^r$, which implies that there is no immaterial information and the envelope model reduces to the standard model.

To estimate the envelope $\mathcal{E}_\Sigma(\mathcal{B})$, Cook et al. (2010) solved the manifold optimization problem

$$\widehat{\mathcal{E}}_\Sigma(\mathcal{B}) = \underset{\mathrm{span}(\Gamma)\in\mathcal{G}(r,u)}{\arg\min} \{\log|\Gamma^T\widehat{\Sigma}_{\mathrm{res}}\Gamma| + \log|\Gamma^T\widehat{\Sigma}_Y^{-1}\Gamma|\} \tag{3}$$

where $|\cdot|$ denotes determinant, $\mathcal{G}(r,u)$ denotes an $r\times u$ Grassmann manifold, which is the set of all u-dimensional subspaces in an r-dimensional space. The matrix $\widehat{\Sigma}_Y$ is the sample covariance matrix of Y and $\widehat{\Sigma}_{\mathrm{res}}$ denotes the sample covariance matrix of the residuals from the regression of Y on X. As the search of $\mathcal{E}_\Sigma(\mathcal{B})$ is on $\mathcal{G}(r,u)$, (3) is a Grassmann manifold optimization problem. The objective function is non-convex. Tools for solving non-convex optimization problems on manifolds, especially when r is large, are quite limited. Cook et al. (2016) addressed this issue by converting (3) to a non-Grassmann manifold optimization, which is faster and more reliable in such cases. Without loss of generality, we assume that Γ_1, the submatrix that consists of the first u rows of Γ, is non-singular. Then

$$\Gamma = \begin{pmatrix}\Gamma_1\\ \Gamma_2\end{pmatrix} = \begin{pmatrix}I_u\\ A\end{pmatrix}\Gamma_1 \equiv G_A\Gamma_1,$$

where $A = \Gamma_2\Gamma_1^{-1}$. Notice that A depends on Γ only through $\mathrm{span}(\Gamma)$: for an orthogonal matrix $O \in \mathbb{R}^{u\times u}$, if $\Gamma^* = \Gamma O$, then $\Gamma_1^* = \Gamma_1 O$, $\Gamma_2^* = \Gamma_2 O$, and $A^* = \Gamma_2 O O^{-1}\Gamma_1^{-1} = A$. Because A is unconstrained, (3) can be converted to the non-Grassmann optimization

$$\widehat{A} = \underset{A\in\mathbb{R}^{(r-u)\times u}}{\arg\min} \{-2\log|G_A^TG_A| + \log|G_A^T\widehat{\Sigma}_{\mathrm{res}}G_A| + \log|G_A^T\widehat{\Sigma}_Y^{-1}G_A|\}. \tag{4}$$

Cook et al. (2015) developed an effective algorithm and a good starting value for solving (4).

Once we have $\widehat{A}$, $\widehat{\mathcal{E}}_\Sigma(\mathcal{B}) = \mathrm{span}(\widehat{G}_A)$, and the envelope estimator of β is $\widehat{\beta}_{\mathrm{env}} = P_{\widehat{\mathcal{E}}}\widehat{\beta}_{\mathrm{ols}}$, where $\widehat{\beta}_{\mathrm{ols}}$ is the ordinary least squares estimator of β and $\mathcal{E}_\Sigma(\mathcal{B})$ is abbreviated as $\mathcal{E}$ if it appears in subscripts. Cook et al. (2010) showed that $\widehat{\beta}_{\mathrm{env}}$ is asymptotically at least as efficient as $\widehat{\beta}_{\mathrm{ols}}$. A more detailed review about the envelope models can be found in (Cook & Su, 2013, Section 2).

2. Sparse Envelope Model

2·1. *Response Variable Selection*

In some cases, certain response variables are immaterial to X, i.e., the corresponding rows of Γ consist of zeros. We call such response variables inactive. We call a response variable active if its corresponding row in Γ is nonzero. Since different orthogonal bases of a subspace have

the same row-wise sparsity pattern, the active and inactive responses are invariant under column transformation of Γ. Because $\beta = \Gamma\eta$, the regression coefficients of the inactive responses are zero. However, an active response may also have zero regression coefficients. Proposition 1 characterizes the active responses, and shows their relationship with responses that have non-zero regression coefficients.

In preparation, we use the covariance graph model (Cox & Wermuth, 1993) to represent the structure of Σ. The covariance graph model was recently used in Chen et al. (2012) to construct a graph-guided fused lasso penalty for predictor variable selection. Let $G = (V, E)$ be an undirected graph with vertices $V = \{1, \ldots, r\}$ and an edge set E consisting of all pairs (i, j) for which the (i, j)th element in Σ is nonzero. The response variables Y_i and Y_j are said to be connected if there is a sequence of edges in the graph connecting vertices i and j.

PROPOSITION 1. *If the regression coefficients of an active response are all zero, then the response must be connected with a response that has non-zero regression coefficients.*

Proposition 1 indicates that if an active response has zero regression coefficients, it still offers information in estimating the non-zero regression coefficients. This is a new feature of response variable selection. In predictor variable selection, if a predictor has zero regression coefficients, it offers no information in estimating any non-zero regression coefficients. More discussion on Proposition 1 is in the Supplement.

In this article, we are not trying to identify the responses having zero regression coefficients and the responses having non-zero regression coefficients; rather we are interested in identifying the active and inactive responses, i.e., whether or not a response contributes in the material part.

2·2. *Formulation*

We use $Y_{\mathcal{A}}$ and $Y_{\mathcal{I}}$ to denote the active and inactive responses. The subscripts $\mathcal{A}$ and $\mathcal{I}$ are used if a quantity is associated with the active or inactive responses. Without loss of generality, let $Y = (Y_{\mathcal{A}}^T, Y_{\mathcal{I}}^T)^T$, and let q denote the dimension of $Y_{\mathcal{A}}$ ($q \le r$). Thus $Y_{\mathcal{A}} \in \mathbb{R}^q$ and $Y_{\mathcal{I}} \in \mathbb{R}^{r-q}$. Then Γ and Γ_0 should have the following structure:

$$\Gamma = \begin{pmatrix} \Gamma_{\mathcal{A}} \\ 0 \end{pmatrix}, \qquad \Gamma_0 = \begin{pmatrix} \Gamma_{\mathcal{A},0} & 0 \\ 0 & I_{r-q} \end{pmatrix} R \equiv \widetilde{\Gamma}_0 R, \tag{5}$$

where $\Gamma_{\mathcal{A}} \in \mathbb{R}^{q\times u}$ is a semi-orthogonal matrix, $\Gamma_{\mathcal{A},0} \in \mathbb{R}^{q\times(q-u)}$ is its completion, and $R \in \mathbb{R}^{(r-u)\times(r-u)}$ is an orthogonal matrix. Since $\Gamma^T Y = \Gamma_{\mathcal{A}}^T Y_{\mathcal{A}}$, the inactive responses do not appear in the material part. Because $\beta = \Gamma\eta$, we have $\beta = (\beta_{\mathcal{A}}^T, 0)^T$, where $\beta_{\mathcal{A}} = \Gamma_{\mathcal{A}}\eta \in \mathbb{R}^{q\times p}$ and the zero matrix has dimension $(r - q) \times p$. The completion of Γ has the general form $\Gamma_0 = \widetilde{\Gamma}_0 R$, where $\widetilde{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$ is a completion with a block diagonal structure, and R represents a rotation of the orthogonal basis. Because $\widetilde{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$ has a simple block diagonal structure, it will be convenient to use it in some of our later development. From the structure of $\widetilde{\Gamma}_0$, it is easy to see that the immaterial part $\widetilde{\Gamma}_0^T Y = \left((\Gamma_{\mathcal{A},0}^T Y_{\mathcal{A}})^T,\ Y_{\mathcal{I}}^T\right)^T$ has two parts, one from the immaterial information of the active responses $\Gamma_{\mathcal{A},0}^T Y_{\mathcal{A}}$, and the other from the inactive responses $Y_{\mathcal{I}}$.

We call (2) the sparse envelope model if Γ and Γ_0 have the structures given by (5). We require $u \le q$ because the dimension of $\Gamma_{\mathcal{A}}^T Y_{\mathcal{A}}$ should be at most the dimension of $Y_{\mathcal{A}}$. When $u = q$, there is no immaterial information in the active responses, and $\Gamma_{\mathcal{A}} = I_q$. Therefore, up to an orthogonal transformation, $\Gamma^T Y = Y_{\mathcal{A}}$ and $\Gamma_0^T Y = Y_{\mathcal{I}}$, and Σ has a block diagonal structure. If $q = r$, there are no inactive responses and all rows in Γ are non-zero. The sparse envelope model is then equivalent to the envelope model.

2·3. *Response Variable Selection via Penalized Likelihood*

Since $\Gamma = G_A\Gamma_1$, a row in Γ is zero if and only if the corresponding row in A is zero. To induce row-wise sparsity in A, we add a group lasso penalty (Yuan & Lin, 2006) to (4), so that the optimization problem becomes

$$\widehat{A} = \underset{A\in\mathbb{R}^{(r-u)\times u}}{\arg\min}\{-2\log|G_A^TG_A| + \log|G_A^T\widehat{\Sigma}_{\rm res}G_A| + \log|G_A^T\widehat{\Sigma}_Y^{-1}G_A| + \sum_{i=1}^{r-u}\lambda_i\|a_i\|_2\}, \quad (6)$$

where a_i^T denotes the ith row of A and the λ_i's are tuning parameters.

We choose this penalty function for the following reasons. First, it treats each row of Γ as a group, so the sparsity is row-wise instead of element-wise. This fits the response variable selection context: $\|a_i\|_2 = 0$ means the $(i+u)$th row of Γ is zero, so the $(i+u)$th response is inactive. Second, it is invariant to a change of basis. Since A depends on Γ only through its span, $\sum_{i=1}^{r-u}\lambda_i\|a_i\|_2$ is unchanged if a different orthogonal basis of $\mathcal{E}_\Sigma(\mathcal{B})$ is used. Third, the estimator (6) has the desirable features of $\sqrt{n}$-consistency, asymptotic normality, selection consistency, and has an optimal estimation rate; see Section 2·5. Finally, its numerical performance is substantially better than the performance of some alternatives, in particular the method that involves applying F tests to each row of $\widehat{\beta}_{\rm ols}$, or hard-thresholding the envelope estimator; see Section 3·1.

When r tends to infinity with n, we denote r by r_n. If $r_n > n$, both $\widehat{\Sigma}_Y$ and $\widehat{\Sigma}_{\rm res}$ are singular, which is problematic because the objective function in (6) depends on $\widehat{\Sigma}_Y^{-1}$ and the optimization algorithm used to solve (6) requires $\widehat{\Sigma}_{\rm res}^{-1}$; see Section 2·4. We can resolve these issues by obtaining estimators for Σ_Y^{-1} and Σ^{-1} directly using methods like sparse permutation invariant covariance estimation (Rothman et al., 2008), lasso penalized D-trace estimation (Zhang & Zou, 2014), or convex pseudo-likelihood based partial correlation graph estimation (Khare et al., 2015). Among these methods, sparse permutation invariant covariance estimation is the only one that does not require a sparsity structure for the target parameter in order to establish the consistency of its estimator. Cook et al. (2012) used this method to estimate a target parameter which is not necessarily sparse, and their numerical experiments showed that the estimator is very stable. In the sparse envelope model, Σ_Y^{-1} and Σ^{-1} may not contain zero elements. We then use sparse permutation invariant covariance estimators of Σ_Y^{-1} and Σ^{-1}, and denote them by $\widehat{\Sigma}_{Y,{\rm sp}}^{-1}$ and $\widehat{\Sigma}_{{\rm res,sp}}^{-1}$. Then $\widehat{\Sigma}_{Y,{\rm sp}}$ and $\widehat{\Sigma}_{{\rm res,sp}}$ are obtained by taking the inverses of $\widehat{\Sigma}_{Y,{\rm sp}}^{-1}$ and $\widehat{\Sigma}_{{\rm res,sp}}^{-1}$. Replacing $\widehat{\Sigma}_{\rm res}$ and $\widehat{\Sigma}_Y^{-1}$ by $\widehat{\Sigma}_{{\rm res,sp}}$ and $\widehat{\Sigma}_{Y,{\rm sp}}^{-1}$ in (6), the optimization problem is

$$\widehat{A} = \underset{A\in\mathbb{R}^{(r_n-u)\times u}}{\arg\min}\{-2\log|G_A^{\rm T}G_A| + \log|G_A^{\rm T}\widehat{\Sigma}_{{\rm res,sp}}G_A| + \log|G_A^{\rm T}\widehat{\Sigma}_{Y,{\rm sp}}^{-1}G_A| + \sum_{i=1}^{r_n-u}\lambda_i\|a_i\|_2\}. \quad (7)$$

Optimization of (6) and (7) is discussed in Section 2·4. After we have $\widehat{A}$, an orthogonal basis of $\mathrm{span}(\widehat{G}_A)$ is used to form $\widehat{\Gamma}$, and $\widehat{\Gamma}_0$ is taken as a completion of $\widehat{\Gamma}$. The sparse envelope estimators of β and Σ are

$$\widehat{\beta} = P_{\widehat{\Gamma}}\widehat{\beta}_{\rm ols}, \qquad \widehat{\Sigma} = P_{\widehat{\Gamma}}\widehat{\Sigma}_{\rm res}P_{\widehat{\Gamma}} + Q_{\widehat{\Gamma}}\widehat{\Sigma}_YQ_{\widehat{\Gamma}}.$$

The estimators for the constituent parameters are $\widehat{\eta} = \widehat{\Gamma}^{\rm T}\widehat{\beta}_{\rm ols}$, $\widehat{\Omega} = \widehat{\Gamma}^{\rm T}\widehat{\Sigma}_{\rm res}\widehat{\Gamma}$ and $\widehat{\Omega}_0 = \widehat{\Gamma}_0^{\rm T}\widehat{\Sigma}_Y\widehat{\Gamma}_0$. The sparse envelope estimators have the same form as the envelope estimators, except that $\widehat{\Gamma}$ and $\widehat{\Gamma}_0$ have the special structures specified in (5).

2·4. *Algorithm*

We first discuss the algorithm for solving (6). Since selection of $r-u$ tuning parameters can be computationally intensive, we use the idea of the adaptive lasso (Zou, 2006) and set $\lambda_i = \lambda\omega_i$, where the ω_i are adaptive weights. Then the optimization becomes

$$\widehat{A} = \underset{A\in\mathbb{R}^{(r-u)\times u}}{\arg\min}\ \{-2\log|G_A^{\mathrm{T}}G_A| + \log|G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A| + \log|G_A^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}G_A| + \lambda\sum_{i=1}^{r-u}\omega_i\|a_i\|_2\}. \tag{8}$$

The optimization problem in (8) is non-convex and the objective function is not differentiable due to the group lasso penalty. Blockwise coordinate descent algorithms have been very successful in solving a wide class of group lasso penalized high-dimensional learning problems (Friedman et al., 2008; Simon et al., 2013; Yang & Zou, 2015). Cook et al. (2015) used a blockwise coordinate descent algorithm to optimize the envelope objective function (4), and the method worked well. Here we develop a fast blockwise coordinate descent algorithm for efficiently solving (8). Our algorithm cyclically updates each row of A, such that after each operation the objective function (8) strictly decreases. Let $A_{-i}\in\mathbb{R}^{(r-u-1)\times u}$ be the submatrix of A with row a_i^{T} removed. Without loss of generality, we consider the case when a_i^{T} is the last row of A. Form the partitions

$$G_A = \begin{pmatrix} I_u \\ A \end{pmatrix} = \begin{pmatrix} G \\ a_i^{\mathrm{T}} \end{pmatrix}, \qquad \widehat{\Sigma}_{\mathrm{res}} = \begin{pmatrix} U_{11}\ U_{12} \\ U_{21}\ U_{22} \end{pmatrix}, \qquad \widehat{\Sigma}_Y^{-1} = \begin{pmatrix} V_{11}\ V_{12} \\ V_{21}\ V_{22} \end{pmatrix}.$$

Let $L(A) = -2\log|G_A^{\mathrm{T}}G_A| + \log|G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A| + \log|G_A^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}G_A|$. We can write $L(A)$ in terms of a_i up to a constant while holding all the other rows of A at their current value $\widetilde{A}_{-i}$: we have

$$\begin{aligned} L(a_i \mid \widetilde{A}_{-i}) = &-2\log(1+a_i^{\mathrm{T}}B_1a_i) + \log\{1+(a_i+v_2)^{\mathrm{T}}B_2(a_i+v_2)\} \\ &+ \log\{1+(a_i+v_3)^{\mathrm{T}}B_3(a_i+v_3)\} + \text{const}, \end{aligned} \tag{9}$$

where $v_2 = U_{22}^{-1}G^{\mathrm{T}}U_{12}$, $v_3 = V_{22}^{-1}G^{\mathrm{T}}V_{12}$, $B_1 = (I_u + A_{-i}^{\mathrm{T}}A_{-i})^{-1}$, $B_2 = U_{22}(G^{\mathrm{T}}U_{11}G - U_{22}^{-1}G^{\mathrm{T}}U_{12}U_{21}G)^{-1}$ and $B_3 = V_{22}(G^{\mathrm{T}}V_{11}G - V_{22}^{-1}G^{\mathrm{T}}V_{12}V_{21}G)^{-1}$. Within the blockwise coordinate descent loops, we need to solve the optimization problem

$$\widehat{a}_i = \underset{a_i}{\arg\min} L(a_i \mid \widetilde{A}_{-i}) + \lambda\omega_i\|a_i\|_2. \tag{10}$$

Unfortunately, there is no closed-form solution to (10), so we apply the majorization-minimization principle (Wu & Lange, 2010; Lange et al., 2000; Hunter & Lange, 2004; Zhou & Lange, 2010) within the blockwise coordinate descent loop by iteratively minimizing a function that majorizes the objective function in (9). The majorization function $Q(a_i)$ is equal to $L(a_i \mid \widetilde{A}_{-i})$ at the current value $\widetilde{a}_i$ and lies strictly above $L(a_i \mid \widetilde{A}_{-i})$ when $a_i \neq \widetilde{a}_i$. Specifically, the majorization function $Q(a_i)$ has the form

$$Q(a_i) = L(\widetilde{a}_i \mid \widetilde{A}_{-i}) + (a_i-\widetilde{a}_i)^{\mathrm{T}}\frac{dL(a_i|\widetilde{A}_{-i})}{da_i}\Big|_{a_i=\widetilde{a}_i} + 0.5\delta_i(a_i-\widetilde{a}_i)^{\mathrm{T}}(a_i-\widetilde{a}_i),$$

where

$$\frac{dL(a_i|\widetilde{A}_{-i})}{da_i}\Big|_{a_i=\widetilde{a}_i} = \frac{-4B_1\widetilde{a}_i}{1+\widetilde{a}_i^{\mathrm{T}}B_1\widetilde{a}_i} + \frac{2B_2(\widetilde{a}_i+v_2)}{1+(\widetilde{a}_i+v_2)^{\mathrm{T}}B_2(\widetilde{a}_i+v_2)} + \frac{2B_3(\widetilde{a}_i+v_3)}{1+(\widetilde{a}_i+v_3)^{\mathrm{T}}B_3(\widetilde{a}_i+v_3)},$$

$\delta_i = (1+\varepsilon^*)\{4\gamma_{\max}(B_1) + 2\gamma_{\max}(B_2) + 2\gamma_{\max}(B_3)\}$, and $\gamma_{\max}(\cdot)$ denotes the largest eigenvalue of the corresponding matrix. We must have $\varepsilon^* > 0$ such that $Q(a_i) > L(a_i \mid \widetilde{A}_{-i})$ holds

for any $a_i \neq \widetilde{a}_i$. In this article we set $\varepsilon^* = 10^{-6}$. Then instead of minimizing (10) we solve

$$\min_{a_i}\{Q(a_i) + \lambda\omega_i\|a_i\|_2\}. \tag{11}$$

The solution to (11) has a simple closed-form expression. Algorithm 1 summarizes our blockwise coordinate descent algorithm. It takes $O(u^3 + ru)$ flops to compute δ_i, and each update of $\widetilde{a}_i$ to $\widetilde{a}_{i,\text{new}}$ takes $O(u^2)$ flops. The starting value can be taken as the envelope estimator of A, which is the minimizer of (4).

Algorithm 1 The blockwise coordinate descent algorithm for solving (8).

Initialize $\widetilde{A}$

Repeat until convergence of $\widetilde{A}$

For $i = 1$ to $i = r - u$

$\delta_i \leftarrow (1 + \varepsilon^*)\{4\gamma_{\max}(B_1) + 2\gamma_{\max}(B_2) + 2\gamma_{\max}(B_3)\}$

Repeat until convergence of $\widetilde{a}_i$

$$\widetilde{a}_{i,\text{new}} \leftarrow \tfrac{1}{\delta_i}\left\{\delta_i\widetilde{a}_i - \tfrac{dL(a_i|\widetilde{A}_{-i})}{da_i}\Big|_{a_i=\widetilde{a}_i}\right\}\left\{1 - \frac{\lambda\omega_i}{\left\|\delta_i\widetilde{a}_i - \frac{dL(a_i|\widetilde{A}_{-i})}{da_i}\Big|_{a_i=\widetilde{a}_i}\right\|_2}\right\}_+$$

$\widetilde{a}_i \leftarrow \widetilde{a}_{i,\text{new}}$

Output $\widetilde{A}$

Theorem 1 shows that Algorithm 1 has a descent property and the updates converge to a stationary point of the objective function in (8). A figure that empirically confirms the convergence of Algorithm 1 is in the Supplement.

THEOREM 1. *After updating $\widetilde{a}_i$, if $\widetilde{a}_{i,\text{new}} \neq \widetilde{a}_i$, the objective function in (*10*) strictly decreases after updating the block:*

$$L(\widetilde{a}_{i,\text{new}} \mid \widetilde{A}_{-i}) + \lambda\omega_i\|\widetilde{a}_{i,\text{new}}\|_2 < L(\widetilde{a}_i) + \lambda\omega_i\|\widetilde{a}_i\|_2.$$

If the solution stays unchanged after each blockwise coordinate update, i.e., $\widetilde{a}_{i,\text{new}} = \widetilde{a}_i$ for all i, then this solution satisfies the Karush–Kuhn–Tucker conditions, and this indicates that the algorithm has converged to a stationary point.

We solve the adaptive group lasso problem (8) by applying Algorithm 1 in a two-stage procedure. In the first stage, we set all ω_i to be 1 in Algorithm 1 and obtain the group lasso estimator $\widehat{A}_{\text{stage1}}$. In the second stage, we set weights $\omega_i = \|\widehat{a}_{i,\text{stage1}}\|_2^{\nu}$ and obtain the weighted group lasso estimator $\widehat{A}$. If $\|\widehat{a}_{i,\text{stage1}}\| = 0$, we exclude a_i in the second stage and set $\hat{a}_i = 0$. The parameter ν can be selected by cross-validation. Based on the discussion in Zou (2006), it is sufficient to choose ν from a small candidate set like $\{0{\cdot}5, 1, 2, 4\}$. To choose the tuning parameter λ, we use the Bayesian information criterion. For a fixed λ, the criterion is defined as $-2l_\lambda + (q_\lambda - u)u\log n$, where l_λ is the log likelihood given λ and q_λ is the number of active responses given λ. We choose the λ that minimizes the criterion. This criterion is used in Chen et al. (2010) and its consistency is proved in Zou & Chen (2012). We use the warm-start trick of Friedman et al. (2010) to compute the solution paths along a sequence of K values of λ, with $\log\lambda$ equally spaced between $\log\lambda_{\max}$ and $\log\lambda_{\min}$. The solution $\widehat{A}^{(\lambda_k)}$ computed at λ_k is used as the initial value for computing the solution for λ_{k+1} in Algorithm 1. An expression for the smallest λ that yields the null model is given in the Supplement. Since the sparse envelope estimator is asymptotically equivalent to the maximum likelihood estimator of the oracle envelope

model, see Section 2·5, we can use likelihood-based procedures such as the Akaike information criterion, the Bayesian information criterion or likelihood ratio testing to select u. We compare the performance of these procedures in the Supplement.

Solving (7) follows the same procedure as solving (6). For choosing λ and u we prefer cross-validation over the Bayesian information criterion and other likelihood-based procedures because these require the sample size to be at least moderately large in order to give good performance.

2·5. *Theoretical Properties of the Sparse Envelope Estimator*

Theorems 2–4 gives results regarding consistency and oracle properties of the sparse envelope estimator in the large-sample case, i.e., when r is fixed and n tends to infinity. Theorems 5 and 6 address selection consistency and the convergence rate when both r_n and n tend to infinity.

If $\mathcal{S}$ is a subspace and $\widehat{\mathcal{S}}$ is an estimator of $\mathcal{S}$, we say that $\widehat{\mathcal{S}}$ is a $\sqrt{n}$-consistent estimator of $\mathcal{S}$ if $P_{\widehat{\mathcal{S}}}$ is a $\sqrt{n}$-consistent estimator of $P_{\mathcal{S}}$. Let $\lambda_{\max,n} = \max(\lambda_1, \ldots, \lambda_{q-u})$ and $\lambda_{\min,n} = \min(\lambda_{q-u+1}, \ldots, \lambda_{r-u})$ at sample size n.

THEOREM 2. *Assume that the sparse envelope model (2) and (5) holds, the errors ε are independent and have finite fourth moment, and $n^{1/2}\lambda_{\max,n} \to 0$ as n tends to infinity. Then there exists a local minimizer $\widehat{A}$ of (6), such that $P_{\widehat{\Gamma}}$ is a $\sqrt{n}$-consistent estimator of P_Γ, and $\widehat{\beta}$ is a $\sqrt{n}$-consistent estimator of β.*

Theorem 2 implies that although the objective function for the sparse envelope estimator is based on a normal likelihood, normality is not required to establish $\sqrt{n}$-consistency of $\widehat{\mathcal{E}}_\Sigma(\mathcal{B})$ and $\widehat{\beta}$. Theorem 3 regards selection consistency and states that the sparse envelope model identifies the inactive responses with probability tending to 1.

THEOREM 3. *Assume that the conditions in Theorem 2 hold, and that $n^{1/2}\lambda_{\min,n} \to \infty$. Then $\mathrm{pr}(\hat{a}_i = 0) \to 1$ for $i = q - u + 1, \ldots, r - u$.*

An oracle estimator must consistently select the active responses, and estimate them with an optimal rate. While the oracle property is well studied in predictor variable selection (Fan & Li, 2001; Zou, 2006), it has not been studied in response variable selection. Therefore we first discuss how to define the oracle model for response variable selection under the standard model (1) and then define the oracle envelope model.

Because the definitions of active and inactive responses rely on the envelope construction, we introduce some new definitions for the standard model. Under the standard model (1), we call a response variable dynamic if its regression coefficients are not zero. We call a response variable static if its regression coefficients are zero. Let d denote the number of dynamic responses, and let $Y_D \in \mathbb{R}^d$ and $Y_S \in \mathbb{R}^{r-d}$ denote the dynamic and static responses. The subscripts D or S are attached to a quantity if it is associated with the dynamic or static responses. Without loss of generality, let $Y = (Y_D^{\mathrm{T}}, Y_S^{\mathrm{T}})^{\mathrm{T}}$. Then $\beta \in \mathbb{R}^{r\times p}$ has the structure $\beta = (\beta_D^{\mathrm{T}}, 0)^{\mathrm{T}}$, where $\beta_D \in \mathbb{R}^{d\times p}$ contains the regression coefficients for the dynamic responses. The oracle model is defined by

$$\begin{pmatrix} Y_D \\ Y_S \end{pmatrix} = \alpha + \begin{pmatrix} \beta_D \\ 0 \end{pmatrix} (X - \mu_X) + \varepsilon, \qquad \mathrm{var}(\varepsilon) = \Sigma = \begin{pmatrix} \Sigma_D & \Sigma_{DS} \\ \Sigma_{DS}^{\mathrm{T}} & \Sigma_S \end{pmatrix}, \tag{12}$$

where $\alpha \in \mathbb{R}^r$, $\beta_D \in \mathbb{R}^{d\times p}$, d is now known, and the partition of Σ corresponds to the allocation of Y_D and Y_S. The oracle model includes the static responses Y_S. This is in contrast to the oracle model for predictor variable selection, where predictors which are inactive are not included in the model. Since Y_S may be correlated with Y_D, including this information can improve the

efficiency in estimating β_D. Excluding Y_S leads to the model

$$Y_D = \alpha_D + \beta_D(X - \mu_X) + \varepsilon_D, \tag{13}$$

where α_D and ε_D are the first d elements of α and ε in (12). We call (13) the dynamic model because it includes only the dynamic responses. It is tempting to view (13) rather than (12) as the target model for oracle estimation, but we do not do so because (13) ignores information available from Y_S which may be used to devise a more efficient estimator in the current context. To compare models (13) and (12), we assume normality of the error distributions in Propositions 2 and 3 in order to get an explicit form for the asymptotic variance. Let $\widehat{\beta}_{D,\mathrm{ols}}$ and $\widehat{\beta}_{S,\mathrm{ols}}$ be the ordinary least squares estimators of the coefficients from the regression of Y_D on X and the regression of Y_S on X respectively, and let R_D and R_S be the residuals from the regression of Y_D on X and the regression of Y_S on X respectively. Define $\Sigma_{D|S} = \Sigma_D - \Sigma_{DS}\Sigma_S^{-1}\Sigma_{SD}$.

PROPOSITION 2. *Assume that the oracle model (*12*) holds and that the errors are normally distributed. The maximum likelihood estimator of β_D under the oracle model is $\widehat{\beta}_{D,1} = \widehat{\beta}_{D,\mathrm{ols}} - \widehat{\beta}_{D|S}\widehat{\beta}_{S,\mathrm{ols}}$, where $\widehat{\beta}_{D|S}$ is the ordinary least squares estimator of the coefficients from the regression of R_D on R_S; and as $n \to \infty$, $\sqrt{n}\{\mathrm{vec}(\widehat{\beta}_{D,1}) - \mathrm{vec}(\beta_D)\}$ is asymptotically normally distributed with mean 0 and covariance matrix $V_1 = \Sigma_X^{-1} \otimes \Sigma_{D|S}$.*

PROPOSITION 3. *Under the conditions in Proposition* 2*, the maximum likelihood estimator of β_D under the dynamic model (*13*) is $\widehat{\beta}_{D,2} = \widehat{\beta}_{D,\mathrm{ols}}$; and as $n \to \infty$, $\sqrt{n}\{\mathrm{vec}(\widehat{\beta}_{D,2}) - \mathrm{vec}(\beta_D)\}$ is asymptotically normally distributed with mean 0 and covariance matrix $V_2 = \Sigma_X^{-1} \otimes \Sigma_D$.*

COROLLARY 1. *Under the conditions in Proposition* 2*,*

$$V_2 - V_1 = \Sigma_X^{-1} \otimes \Sigma_D^{1/2}\rho\Sigma_D^{1/2},$$

where $\rho = \Sigma_D^{-1/2}\Sigma_{DS}\Sigma_S^{-1}\Sigma_{SD}\Sigma_D^{-1/2}$. The eigenvalues of ρ are the squared canonical correlations between Y_D and Y_S.

Corollary 1 quantifies the efficiency gains obtained by including Y_S. The result states that the stronger the correlation between Y_D and Y_S, the greater is the variance reduction obtained by including Y_S. When Y_D and Y_S are uncorrelated, $\widehat{\beta}_{D,1}$ and $\widehat{\beta}_{D,2}$ have the same asymptotic variance. In that case, we can ignore Y_S, since it does not carry information on β_D through Y_D.

Under the envelope model, the inactive response contains information on $\beta_{\mathcal{A}}$ through its covariance with the active response. We then define the oracle envelope model as

$$\begin{pmatrix} Y_{\mathcal{A}} \\ Y_{\mathcal{I}} \end{pmatrix} = \alpha + \Gamma\eta(X - \mu_X) + \varepsilon, \qquad \Sigma = \Gamma\Omega\Gamma^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}, \qquad \Gamma = \begin{pmatrix} \Gamma_{\mathcal{A}} \\ 0 \end{pmatrix}. \tag{14}$$

The oracle envelope model (14) appears similar to the sparse envelope model (2) and (5), but in (14) we know q and which rows in Γ consist of only zeros. We attach a subscript O if an estimator is the oracle envelope estimator. Let $\widehat{\Sigma}_{Y_{\mathcal{A}}|X} \in \mathbb{R}^{q\times q}$ be the sample covariance matrix of the residuals from the regression of $Y_{\mathcal{A}}$ on X, and $(\widehat{\Sigma}_Y^{-1})_{\mathcal{A}} \in \mathbb{R}^{q\times q}$ be the $q \times q$ upper left block of $\widehat{\Sigma}_Y^{-1}$. Let $\widetilde{\Omega}_0 = \widetilde{\Gamma}_0^{\mathrm{T}}\Sigma\widetilde{\Gamma}_0$. Based on the structure of $\widetilde{\Gamma}_0$, we partition $\widetilde{\Omega}_0$ into

$$\widetilde{\Omega}_0 = \begin{pmatrix} \widetilde{\Omega}_{0,\mathcal{A}} & \widetilde{\Omega}_{0,\mathcal{A}\mathcal{I}} \\ \widetilde{\Omega}_{0,\mathcal{A}\mathcal{I}}^{\mathrm{T}} & \widetilde{\Omega}_{0,\mathcal{I}} \end{pmatrix}, \qquad \widetilde{\Omega}_{0,\mathcal{A}} \in \mathbb{R}^{(q-u)\times(q-u)}, \quad \widetilde{\Omega}_{0,\mathcal{I}} \in \mathbb{R}^{(r-q)\times(r-q)}.$$

Let $\widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}} = \widetilde{\Omega}_{0,\mathcal{A}} - \widetilde{\Omega}_{0,\mathcal{A}\mathcal{I}}\widetilde{\Omega}_{0,\mathcal{I}}^{-1}\widetilde{\Omega}_{0,\mathcal{I}\mathcal{A}}$. Proposition 4 gives the maximum likelihood estimator $\widehat{\beta}_{\mathcal{A},O}$ and its asymptotic distribution.

PROPOSITION 4. *Assume that the oracle envelope model (*14*) holds and the errors are normally distributed. Then the maximum likelihood estimator of $\beta_{\mathcal{A}}$ under the oracle model is $\widehat{\beta}_{\mathcal{A},O} = P_{\widehat{\Gamma}_{\mathcal{A},O}}\widehat{\beta}_{\mathcal{A},\mathrm{ols}}$, where*

$$\mathrm{span}(\widehat{\Gamma}_{\mathcal{A},O}) = \underset{\mathrm{span}(G)\in\mathcal{G}(q,u)}{\arg\min} \log|G^{\mathrm{T}}\widehat{\Sigma}_{Y_{\mathcal{A}}|X}G| + \log|G^{\mathrm{T}}(\widehat{\Sigma}_{Y}^{-1})_{\mathcal{A}}G|.$$

Additionally, as $n\to\infty$, $\sqrt{n}\{\mathrm{vec}(\widehat{\beta}_{\mathcal{A},O}) - \mathrm{vec}(\beta_{\mathcal{A}})\}$ is asymptotically normally distributed with mean 0 *and covariance matrix $V_O = \Sigma_X^{-1}\otimes\Gamma_{\mathcal{A}}\Omega\Gamma_{\mathcal{A}}^{\mathrm{T}} + (\eta^{\mathrm{T}}\otimes\Gamma_{\mathcal{A},0})T^{-1}(\eta\otimes\Gamma_{\mathcal{A},0}^{\mathrm{T}})$, where $T = \eta\Sigma_X\eta^{\mathrm{T}}\otimes\widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1} + \Omega\otimes\widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1} + \Omega^{-1}\otimes\widetilde{\Omega}_{0,\mathcal{A}} - 2I_u\otimes I_{q-u}$.*

From Proposition 4, we see that $Y_{\mathcal{I}}$ appears in the objective function for $\mathrm{span}(\widehat{\Gamma}_{\mathcal{A},O})$, and therefore affects $\widehat{\beta}_{\mathcal{A},O}$. We now define the active envelope model, which contains only the active responses:

$$Y_{\mathcal{A}} = \alpha_{\mathcal{A}} + \Gamma_{\mathcal{A}}\eta(X-\mu_X) + \varepsilon_{\mathcal{A}}, \qquad \Sigma_{\mathcal{A}} = \Gamma_{\mathcal{A}}\Omega\Gamma_{\mathcal{A}}^{\mathrm{T}} + \Gamma_{\mathcal{A},0}\widetilde{\Omega}_{0,\mathcal{A}}\Gamma_{\mathcal{A},0}^{\mathrm{T}}. \tag{15}$$

PROPOSITION 5. *Assume that the conditions in Proposition* 4 *hold. Then the maximum likelihood estimator of $\beta_{\mathcal{A}}$ under the active envelope model is $\widehat{\beta}_{\mathcal{A},2} = P_{\widehat{\Gamma}_{\mathcal{A},2}}\widehat{\beta}_{\mathcal{A},\mathrm{ols}}$, where*

$$\mathrm{span}(\widehat{\Gamma}_{\mathcal{A},2}) = \underset{\mathrm{span}(G)\in\mathcal{G}(q,u)}{\arg\min} \log|G^{\mathrm{T}}\widehat{\Sigma}_{Y_{\mathcal{A}}|X}G| + \log|G^{\mathrm{T}}\widehat{\Sigma}_{Y_{\mathcal{A}}}^{-1}G|.$$

Additionally, as $n\to\infty$, $\sqrt{n}\{\mathrm{vec}(\widehat{\beta}_{\mathcal{A},2}) - \mathrm{vec}(\beta_{\mathcal{A}})\}$ is asymptotically normally distributed with mean 0 *and covariance matrix $V_3 = \Sigma_X^{-1}\otimes\Gamma_{\mathcal{A}}\Omega\Gamma_{\mathcal{A}}^{\mathrm{T}} + (\eta^{\mathrm{T}}\otimes\Gamma_{\mathcal{A},0})T_2^{-1}(\eta\otimes\Gamma_{\mathcal{A},0}^{\mathrm{T}})$, where $T_2 = \eta\Sigma_X\eta^{\mathrm{T}}\otimes\widetilde{\Omega}_{0,\mathcal{A}}^{-1} + \Omega\otimes\widetilde{\Omega}_{0,\mathcal{A}}^{-1} + \Omega^{-1}\otimes\widetilde{\Omega}_{0,\mathcal{A}} - 2I_u\otimes I_{q-u}$.*

Comparing V_O and V_3, we see that because $\widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1} \geq \widetilde{\Omega}_{0,\mathcal{A}}^{-1}$, $T_2^{-1} \geq T^{-1}$, the oracle envelope model (14) is more efficient than the active envelope model (15) in estimating $\beta_{\mathcal{A}}$. Therefore in the envelope context, including $Y_{\mathcal{I}}$ also improves efficiency.

We now return to the discussion of the theoretical properties of the sparse envelope estimator.

THEOREM 4. *Assume that the conditions in Theorem* 3 *hold. Then as $n\to\infty$, $\sqrt{n}\{\mathrm{vec}(\widehat{\beta}_{\mathcal{A}}) - \mathrm{vec}(\beta_{\mathcal{A}})\}$ is asymptotically normally distributed with mean* 0 *and asymptotic variance equal to that of $\widehat{\beta}_{\mathcal{A},O}$. If we further assume that the errors are normally distributed, then the asymptotic variance V is given in closed form: $V = \Sigma_X^{-1}\otimes\Gamma_{\mathcal{A}}\Omega\Gamma_{\mathcal{A}}^{\mathrm{T}} + (\eta^{\mathrm{T}}\otimes\Gamma_{\mathcal{A},0})T^{-1}(\eta\otimes\Gamma_{\mathcal{A},0}^{\mathrm{T}})$, where $T = \eta\Sigma_X\eta^{\mathrm{T}}\otimes\widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1} + \Omega\otimes\widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1} + \Omega^{-1}\otimes\widetilde{\Omega}_{0,\mathcal{A}} - 2I_u\otimes I_{q-u}$.*

Theorem 4 indicates that the sparse envelope estimator is asymptotically normal, and has the asymptotic distribution we would have if we knew in advance which responses are active and which are inactive. The optimal estimation rate asserted in Theorem 4 combined with selection consistency shows that the sparse envelope estimator has the oracle property: the sparse envelope model selects the inactive responses with probability tending to 1 and estimates the coefficients for the active responses as efficiently as does the oracle envelope model.

Now we discuss the convergence rate and selection consistency of the sparse envelope estimator when r_n tends to infinity with n. We first make a few assumptions about the true model: (A1) There exist positive constants $\bar{k}$ and $\underline{k}$ such that $\gamma_{\max}(\Sigma) \leq \bar{k}$ and $\gamma_{\min}(\Sigma) \geq \underline{k}$, where $\gamma_{\max}(\Sigma)$

and $\gamma_{\min}(\Sigma)$ be the largest and smallest eigenvalue of Σ. (A2) The samples of ε are independent and identically sampled from a sub-gaussian distribution, i.e., $E\{\exp(t_1^{\mathrm{T}}\varepsilon)\} \leq \exp(c_1 t_1^{\mathrm{T}}\Sigma t_1)$ for some constant $c_1 > 0$ and every $t_1 \in \mathbb{R}^{r_n}$. Samples of X are independent and identically distributed, and $X - \mu_X$ follows a sub-gaussian distribution, i.e., $E[\exp\{t_2^{\mathrm{T}}(X-\mu_X)\}] \leq \exp(c_2 t_2^{\mathrm{T}}\Sigma_X t_2)$ for some constant $c_2 > 0$ and every $t_2 \in \mathbb{R}^p$.

Let s_1 and s_2 denote the number of nonzero elements in the lower triangle (not including the diagonal elements) of Σ^{-1} and Σ_Y^{-1} respectively, and $s = \max\{s_1, s_2\}$.

THEOREM 5. *Assume the sparse envelope model (2) and (5) holds. Under Assumptions A1 and A2, if* $\lambda_{\max,n} = o[\{(r_n+s)\log r_n/n\}^{1/2}]$, *then as* $n \to \infty$, *there exists a solution* $\widehat{A}$ *of the optimization problem (7) such that* $\|\widehat{A} - A\|_F = O_p[\{(r_n+s)\log r_n/n\}^{1/2}]$, *and the sparse envelope estimator* $\widehat{\beta}$ *satisfies that* $\|\widehat{\beta} - \beta\|_F = O_p[\{(r_n+s)\log r_n/n\}^{1/2}]$.

Inspection of the proof of Theorem 5 reveals that the convergence rate of the sparse envelope estimator is limited by the convergence rate of $\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1}$ and $\widehat{\Sigma}_{\mathrm{res,sp}}^{-1}$. If we have a different inverse covariance matrix estimator that converges at a faster rate, then the convergence rate of the sparse envelope estimator can be improved. Assumptions A1 and A2 are required for the consistency of $\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1}$ and $\widehat{\Sigma}_{\mathrm{res,sp}}^{-1}$. We relaxed the normality assumption in Rothman et al. (2008) to the sub-gaussian assumption based on the work in Ravikumar et al. (2011).

THEOREM 6. *Suppose the assumptions in Theorem* 5 *hold,* $\{(r_n+s)\log r_n/n\}^{1/2} \to 0$ *as* n *tends to infinity, and* $\{(r_n+s)\log r_n/n\}^{1/2} = o(\lambda_{\min,n})$. *Then* $\mathrm{pr}(\hat{a}_i \neq 0) \to 1$ *for* $i = 1, \ldots, q-u$, *and* $\mathrm{pr}(\hat{a}_i = 0) \to 1$ *for* $i = q-u+1, \ldots, r_n - u$.

Theorem 6 establishes selection consistency of the sparse envelope estimator. When r_n tends to infinity with n, the sparse envelope estimator still identifies active and inactive responses with probability tending to 1.

3. SIMULATIONS AND DATA ANALYSIS

3·1. *Simulations*

We report the results of two simulation studies, one in the large-sample setting and one in high-dimensional setting. In the first simulation, we fixed $p = 2$, $r = 10$, $q = 4$ and $u = 2$. The matrix $(\Gamma_{\mathcal{A}}, \Gamma_{\mathcal{A},0})$ was obtained by orthogonalizing a $q \times q$ matrix of independent uniform $(0,1)$ variates. Then we added 0's and 1's following the structure in (5) to get Γ and Γ_0. We took $\Omega = 9I_u$, and the eigenvalues of Ω_0 varied from 0·67 to 28·33. The canonical correlation between $\Gamma_0^{\mathrm{T}} Y_{\mathcal{A}}$ and $Y_{\mathcal{I}}$ was 0·9. The elements in X and η were generated from independent $N(0,4)$ random variables. We varied the sample size from 25 to 1000, and generated 200 replications for each sample size. For each replication, we fit the standard model (1), the sparse envelope model (2) and (5), the oracle envelope model (14), the active envelope model (15), and got their estimators of β. The estimation standard deviation for each element in β was calculated from the 200 estimators. For each sample size, the bootstrap standard deviation was obtained by computing the standard deviations from 200 bootstrap samples. The results for a randomly chosen element in β are plotted in Fig. 1. For better visibility, only the asymptotic standard deviation of the standard model is displayed. In all cases, the standard deviations are multiplied by $\sqrt{n}$.

Figure 1 shows that sparse envelope estimator is more efficient than the standard estimator and the active envelope estimator for all sample sizes. The ratio of the asymptotic standard deviation of the standard estimator to that of the sparse envelope estimator is 2·71, and for the active enve-

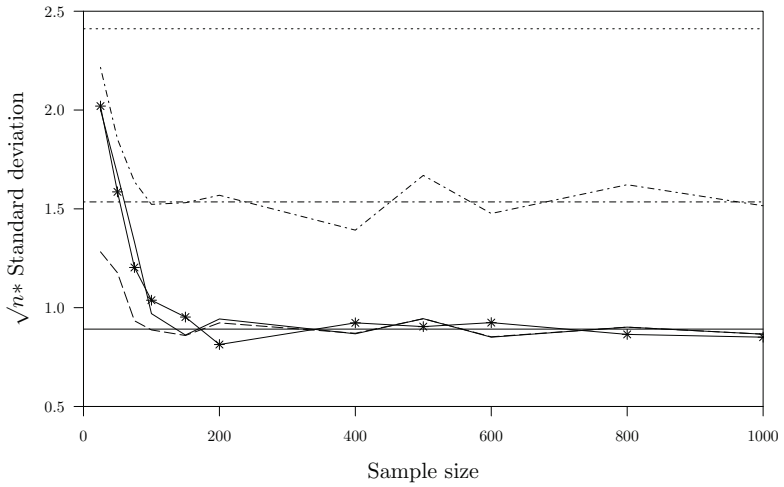

Fig. 1. Comparison of the standard deviations for sparse envelope estimator (solid), active envelope estimator (dash-dotted), oracle envelope estimator (dashed) and standard estimator (dotted). The horizontal lines mark the asymptotic standard deviation of the corresponding estimators. The solid line with asterisks marks the bootstrap standard deviations.

lope estimator versus the sparse envelope estimator comparison, the ratio is 1·73. The difference between the sparse envelope estimator and the oracle envelope estimator becomes quite small for sample sizes bigger than 100, which is consistent with the optimal estimation rate described in Theorem 4. We also notice that the bootstrap standard deviation is a good estimator of the actual standard deviation. In order to evaluate the variable selection performance of the sparse envelope model, we considered the true positive rate c_1/q, where c_1 is the number of active responses correctly chosen; true negative rate $c_2/(r-q)$, where c_2 is the number of inactive responses correctly chosen; and accuracy, which is an integer taking value 0 or 1, with 1 indicating that both the active and inactive responses are correctly chosen and 0 otherwise. The average of each quantity is given in Table 1. The accuracy tends to 1 as n increases, which confirms the selection consistency stated in Theorem 3. For comparison, we applied a hard-thresholding on the envelope estimator of Γ to select the active responses, with the threshold chosen by cross-validation. We also performed an F test on each row of $\widehat{\beta}_{\rm ols}$ with adjustments for multiple testing. The sparse envelope estimator dominants these two estimators for all sample sizes in this case.

Table 1. *Average true positive rate (%), true negative rate (%) and accuracy (%) of sparse envelope estimator, hard thresholding estimator and F test*

	sparse envelope			hard thresholding			F test		
n	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.
25	92·6	81·0	33·5	75·4	97·9	30·5	51·7	99·8	0·0
50	97·0	90·5	69·0	85·0	99·0	52·5	61·6	99·5	2·0
75	98·6	95·9	85·5	90·5	99·8	70·0	70·6	99·5	13·5
100	99·8	98·3	94·5	96·9	99·9	89·0	77·8	99·4	23·5
150	100·0	99·3	96·0	99·2	100·0	97·0	84·6	99·6	39·0
200	100·0	100·0	100·0	100·0	100·0	100·0	91·6	99·7	64·5

Table 2. *Average true positive rate (%), true negative rate (%) and accuracy (%) of sparse envelope estimator, hard thresholding estimator and F test in high dimensional setting*

	sparse envelope			hard thresholding			F test		
n	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.
50	78·5	99·1	6·5	53·4	100·0	0·0	35·2	100·0	0·0
100	91·6	99·9	54·5	62·6	100·0	0·0	55·9	100·0	0·0
150	98·0	100·0	91·5	81·0	100·0	2·0	71·2	100·0	0·0
200	99·8	100·0	98·0	86·6	100·0	12·0	85·2	100·0	10·0
250	99·8	100·0	98·5	89·6	100·0	19·0	95·1	100·0	48·0
300	100·0	100·0	100·0	91·8	100·0	28·0	98·4	100·0	79·0

Now we consider the high-dimensional scenario. We set $r = 1000$, $q = 10$, $p = 5$, $u = 2$ and varied n from 50 to 300. The first $q/2$ rows in $\Gamma_{\mathcal{A}}$ were $\{(2/q)^{1/2}, 0\}^{\mathrm{T}}$ and the remaining $q/2$ rows in $\Gamma_{\mathcal{A}}$ were $\{0, (2/q)^{1/2}\}^{\mathrm{T}}$. Then we used the structure in (5) to construct Γ and Γ_0. The elements in η were independent $N(0, 9)$ random variables, Ω =0·04I_u and Ω_0 was a block diagonal matrix with the upper left block being $25I_{q-u}$ and lower right block being $4I_{r-q}$. The elements in X were independent $N(0, 1)$ random variables. For each sample size, 200 replications were generated. Table 2 shows that performance of the sparse envelope estimator is better than that of the hard thresholding estimator and F test in this scenario as well. A figure that describes the convergence of $\|\widehat{\beta} - \beta\|_F$ is in the Supplementary Material.

Remark 1. The sparse envelope model also achieves efficiency gains when $r < p < n$, or with weak signals; see the Supplementary Material.

3·2. *Data analysis*

We illustrate the sparse envelope model using microarray time-course data on cell-cycle control in the fission yeast *Schizosaccharomyces pombe*. This dataset is analyzed in Gilks et al. (2005) using multivariate linear regression to study how gene expression levels change in a cell cycle. The response variables are expression levels of genes. Among the 407 genes measured, 11 genes have missing values. We only used the genes with complete data, and this gave 396 responses, which we log transformed to reduce skewness. The predictors are 10 equally spaced time points of the cell cycle and the sample size is 177. We fit the sparse envelope model to the data, with $u = 2$ suggested by cross-validation. The model identified 25 inactive responses. This indicates that the expression level of most genes varies in a cell cycle, but there are a few genes whose intensities do not change in a cell cycle. Among the 25 inactive responses, gene cdc20 was also identified by Gilks et al. (2005) to have "very little cell-cycle activity." We estimated $\|\widehat{\beta}_{\mathrm{ols}} - \beta\|_F$ and $\|\widehat{\beta} - \beta\|_F$ by the average of 200 bootstrap samples. The ratio of the estimated $\|\widehat{\beta}_{\mathrm{ols}} - \beta\|_F$ to $\|\widehat{\beta} - \beta\|_F$ is 1·52, which shows a clear efficiency gain due to the sparse envelope model.

4. DISCUSSION

In this paper, the sparse envelope model is developed by assuming row-wise sparsity in Γ under the envelope model. In ultra-high dimensional problems where $r_n \gg n$, we need to make additional assumptions such as sparsity of Σ or Σ^{-1} in order to establish the consistency of the sparse envelope model. The convergence rate of the sparse envelope estimator $\widehat{\beta}$ can be improved to $\|\widehat{\beta} - \beta\|_F = O_p\{(\log r_n/n)^{1/2}\}$ if we assume the number of nonzero off-diagonal elements

in Σ^{-1} is fixed as n tends to infinity. It may also be of interest to study prediction performance rather than estimation of parameters in ultra-high dimensional problems.

When the envelope structure does not hold, some preliminary numerical results show that the envelope estimator may still have a smaller mean square error than the standard estimator, as a result of the bias-variance tradeoff. The properties of the envelope estimator under this situation are open.

ACKNOWLEDGEMENT

We thank Professors Dennis Cook, Hani Doss and Bing Li for helpful discussions and Professor Adam Rothman for providing the codes for sparse permutation invariant covariance estimation. We are grateful to the editor, associate editor and three referees for comments that helped us greatly improve the paper. Research for this article was supported in part by the U.S. National Science Foundation, National University of Singapore and the Natural Sciences and Engineering Research Council of Canada.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes a notation table, proofs of theorems and propositions, and additional simulations.

REFERENCES

BENJAMINI, Y. & YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–1188.

CHEN, X., LIN, Q., KIM, S., CARBONELL, J. G. & XING, E. P. (2012). Smoothing proximal gradient method for general structured sparse regression. *Ann. Appl. Statist.* **6**, 719–752.

CHEN, X., ZOU, C. & COOK, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38**, 3696–3723.

CONWAY, J. B. (2013). *A Course in Functional Analysis*. New York: Springer.

COOK, R. D., FORZANI, L. & ROTHMAN, A. J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *Ann. Statist.* **40**, 353–384.

COOK, R. D., FORZANI, L. & SU, Z. (2016). A note on fast envelope estimation. *J. Multivar. Anal.* **150**, 42–54.

COOK, R. D., HELLAND, I. S. & SU, Z. (2013). Envelopes and partial least squares regression. *J. R. Statist. Soc. B* **75**, 851–877.

COOK, R. D., LI, B. & CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statist. Sinica* **20**, 927–1010.

COOK, R. D. & SU, Z. (2013). Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression. *Biometrika* **100**, 939–954.

COOK, R. D. & ZHANG, X. (2015). Foundations for envelope models and methods. *J. Am. Statist. Assoc.* **110**, 599–611.

COX, D. R. & WERMUTH, N. (1993). Linear dependencies represented by chain graphs. *Statist. Sci.* **8**, 204–218.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–1360.

FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**.

GILKS, W. R., TOM, B. D. M. & BRAZMA, A. (2005). Fusing microarray experiments with multivariate regression. *Bioinformatics* **21**, ii137–ii143.

HUNTER, D. & LANGE, K. (2004). A tutorial on MM algorithms. *Am. Statistician* **58**, 30–37.

KHARE, K., OH, S. Y. & RAJARATNAM, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. R. Statist. Soc. B* **77**, 803–825.

LANGE, K., HUNTER, D. & YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comp. Graph. Statist.* **9**, 1–20.

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. & YU, B. (2011). High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electron. J. Statist.* **5**, 935–980.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

SIMON, N., FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2013). A sparse-group lasso. *J. Comp. Graph. Statist.* **22**, 231–245.

SU, Z. & COOK, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika* **98**, 133–146.

WU, T. & LANGE, K. (2010). The MM alternative to EM. *Statist. Sci.* **4**, 492–505.

YANG, Y. & ZOU, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statist. Comp.* **25**, 1129–1141.

YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* **68**, 49–67.

ZHANG, T. & ZOU, H. (2014). Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika* **101**, 103–120.

ZHOU, H. & LANGE, K. (2010). MM algorithms for some discrete multivariate distributions. *J. Comp. Graph. Statist.* **19**, 645–665.

ZOU, C. & CHEN, X. (2012). On the consistency of coordinate-independent sparse estimation with BIC. *J. Multivar. Anal.* **112**, 248–255.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–1429.

[*Received April* 2012. *Revised September* 2012]