Published in final edited form as:

J Comput Graph Stat. 2012 ; 21(1): 110–133. doi:10.1198/jcgs.2011.10102.

A Sparse Structured Shrinkage Estimator for Nonparametric Varying-Coefficient Model with an Application in Genomics

Z. John Daye, Jichun Xie, and Hongzhe Li*

University of Pennsylvania, School of Medicine

Abstract

Many problems in genomics are related to variable selection where high-dimensional genomic data are treated as covariates. Such genomic covariates often have certain structures and can be represented as vertices of an undirected graph. Biological processes also vary as functions depending upon some biological state, such as time. High-dimensional variable selection where covariates are graph-structured and underlying model is nonparametric presents an important but largely unaddressed statistical challenge. Motivated by the problem of regression-based motif discovery, we consider the problem of variable selection for high-dimensional nonparametric varying-coefficient models and introduce a sparse structured shrinkage (SSS) estimator based on basis function expansions and a novel smoothed penalty function. We present an efficient algorithm for computing the SSS estimator. Results on model selection consistency and estimation bounds are derived. Moreover, finite-sample performances are studied via simulations, and the effects of high-dimensionality and structural information of the covariates are especially highlighted. We apply our method to motif finding problem using a yeast cell-cycle gene expression dataset and word counts in genes' promoter sequences. Our results demonstrate that the proposed method can result in better variable selection and prediction for high-dimensional regression when the underlying model is nonparametric and covariates are structured. Supplemental materials for the article are available online.

Keywords

High-dimensional data; Model selection; Motif analysis; Nonparametric regression; Sparsity; Structured covariates

1 Introduction

Research in genomics has motivated many important developments in high-dimensional regression and variable selection, where high-dimensional genomic data are treated as covariates. Such genomic data often have certain structures and can be represented as vertices of an undirected graph. For example, genes in the same metabolic pathways or proteins that share similar functionalities often serve as covariates in genomic studies. These covariates can be considered as groups and can be represented as nodes in a biological network. Statistical procedures that incorporate covariate structures can often lead to models

*Z. John Daye is a Postdoctoral Researcher, Jichun Xie is a Ph.D. student, and Hongzhe Li is Professor, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021. (zdaye@upenn.edu)

SUPPLEMENTAL MATERIALS

Appendix: The Appendix contains detailed proofs for Theorem 3.1, 3.2, and 3.3. (vc appendix.pdf)

Fortran & R Code: Fortran code and its R-language wrapper for computing the SSS estimator. vc.R contains the main program. vcSim.R contains simulation examples used in the paper. Fortran 90 codes need to be pre-compiled, see README.txt. (vc.zip)

that are scientifically more interpretable than those that ignore such structures. Furthermore, in many interesting applications, the dimension of covariates is often very large and the effects of predictors on the response vary nonparametrically depending upon some intrinsic state, such as time. For instance, gene expression level varies depending upon the stage in the cell cycle at which they are measured and can relate to a large number of genetic factors.

Motivated by applications in genomics, we consider the problem of high-dimensional and non-parametric variable selection with structured covariates. We assume that the response $Y(t)$ is a random process, and the predictor $\mathbf{X}(t) = (X_1(t), \dots, X_p(t))^T$ is a p -dimensional random process. In applications, observations for n randomly selected subjects are obtained as $(y_i(t_{ik}), x_i(t_{ik}))$ for the i th subject at time point t_{ik} , for $i = 1, \dots, n$ and $k = 1, \dots, K$. We consider the variable selection problem for the linear varying-coefficient (LVC) model

$$y_i(t_{ik}) = \sum_{g=1}^p x_{ig}(t_{ik}) \beta_g^0(t_{ik}) + \varepsilon_{ik}, \quad i=1, \dots, n, k=1, \dots, K, \quad (1)$$

where $\beta_g^0(t_{ik})$ are true coefficient functions and ε_{ik} are *i.i.d* random errors with zero means. For simplicity, we assume that the discrete time points t_{ik} sampled are the same across observations and denote them as t_k for $k = 1, \dots, K$. Our results in this paper can be easily extended to the case where different time points are sampled across observations. The LVC model (1) presents an important nonparametric generalization of linear regression. It allows the effects of covariates to depend upon an intrinsic quantity, such as time, while retaining the simplicity of having a linear model. Because of its flexibility, the LVC model has received much attention (Hastie and Tibshirani, 1993; Hoover et al., 1998; Cai et al., 2000; Fan and Zhang, 2008). It is particularly useful in genomic studies, where biological processes often vary depending on meaningful states, such as time or cell-cycle stage (Wang et al., 2007, 2008).

When the number of the variables in the LVC model (1) is large, it is often important to select the relevant variables. High-dimensional variable selection for parametric linear or generalized linear regressions has been a topic of intensive research in recent years. Among many recent developments, the lasso (Tibshirani, 1996) provides an important strategy for variable selection. It can be easily computed in high-dimensional regression using available algorithms, such as the LARS (Efron et al., 2004; Rosset and Zhu, 2007) or coordinate descent (Tseng, 1988, 2001; Friedman et al., 2007). Consistency properties of model selection are also well understood for the lasso (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006). Moreover, several methods have been developed under linear regression to incorporate various covariate structures. The group lasso (Yuan and Lin, 2006) incorporates group or hierarchical structure of variables, whereas the fused lasso (Land and Friedman, 1996; Tibshirani et al., 2005) promotes smoothness over ordered variables. In genomic applications, the network-constrained regression (Li and Li, 2008) and graph-constrained estimation procedure (Li and Li, 2010) are introduced to utilize *a priori* network or graphical information.

Extensions of penalized estimation to nonparametric regression in high-dimensional settings have also been developed, including variable selection for the additive models (Meier et al., 2009; Ravikumar et al., 2009; Huang et al., 2010) and varying-coefficient models (Wang et al., 2008; Wang and Xia, 2009). However, these methods do not explicitly incorporate covariate structures for high-dimensional and nonparametric modeling. The structures among the covariates often imply certain structure on the regression coefficient functions

that can lead to better interpretation and variable selection in nonparametric regression modeling.

In this paper, we consider the setting when the covariates can be represented as vertices of an undirected graph. Such a graph can be defined either by the correlation matrix of the covariates or defined based on some prior knowledge. For many applications with structured covariates, certain smoothness of the regression coefficients can be imposed. We propose a sparse structured shrinkage (SSS) estimator to impose both sparsity and smoothness of the coefficient functions in the LVC model (1). Our procedure utilizes a nonparametric L_1 -norm penalty for variable selection and a nonparametric pairwise L_2 -norm fusion penalty for incorporating *a priori* covariate structures. We estimate our model using B-splines and present an efficient algorithm for computing it under high-dimensionality. Model selection consistency and estimation bounds are established under the fixed-knots asymptotics.

The rest of the paper is organized as follows. In Section 2, we first introduce the SSS estimator for the LVC model. Section 2.2 details an efficient algorithm for computing our estimator, whereas Section 3 provides theoretical results on model selection consistency and estimation bounds. We demonstrate the applicability of our method via simulation examples in Section 4 and real-data example in Section 5. We conclude in Section 6 with a brief discussion. All proofs are given in the Appendix, available as online supplemental materials.

2 A Sparse Structured Shrinkage Estimator for the LVC Model

2.1 A sparse structured shrinkage estimator

We consider structured covariates in the LVC model (1) where the structures can be summarized using an undirected covariate graph. One example of such a graph is the co-expression network often defined for gene expression data, when links between two nodes represent two highly correlated genes. Another example is the motif finding problem where the covariates are the frequencies of the words of certain length in genes' promoter sequences and many words differ only by one or two bases. In such an application, words can form a hypercube based on their similarity (See Section 5 for details). In general, for a given graph G that represents the covariate structure, let $w_{g_1 g_2}$ denote the weights for the edges between vertices g_1 and g_2 . We propose the following SSS estimator,

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K \left[y_i(t_k) - \sum_{g=1}^p x_{ig}(t_k) \beta_g(t_k) \right]^2 + \lambda_1 \sum_{g=1}^p \|\beta_g\| + \frac{\lambda_2}{2} \sum_{g_1=1}^{p-1} \sum_{g_2=g_1+1}^p w_{g_1 g_2} \|\beta_{g_1} - \beta_{g_2}\|_2^2 \right\},$$

where λ_1 and λ_2 are two tuning parameters, $\beta = \{\beta_1(t), \dots, \beta_p(t)\}$ is the vector of coefficient functions, and the functional norm $\|f\| = \sqrt{\int f^2(t) dt}$. This functional norm has also been used for variable selection for additive models (Meier et al., 2009; Ravikumar et al., 2009; Huang et al., 2010). We further define a re-scaled SSS estimator as

$$\widehat{\beta}^0(c, \lambda_1, \lambda_2) = c \widehat{\beta}, \quad (3)$$

where c is a re-scaling parameter, which is introduced to reduce the bias of the estimate due to regularization. As we show in our simulation, this re-scaling parameter often leads to better parameter estimates than the original estimates $\widehat{\beta}$. A similar re-scaling procedure was also employed in Witten and Tibshirani (2009) for the simple regression setting and justified through numerical studies. In this paper, we assume non-intercept predictors $x_{ig}(t_k)$ in (2) to be standardized to have zero mean and equal variances across observations. An intercept

term may be included by setting $x_{j1}(t_k) = 1/\delta$ for δ very small and the resulting coefficient $\hat{\beta}_1(t_k)$ divided by δ , which prevents the intercept from being penalized.

The SSS estimator penalizes the sum of functional norms and norms of weighted pairwise differences of β_g , which do not depend upon time t_k . The first penalty term in (2) allows $\hat{\beta}$ to be sparse, whereas the second penalty term incorporates structural information of covariates to encourage the smoothness of the coefficients for covariates that are linked on the graph. We call them *sparsity* and *structural* penalties, respectively. These penalties on coefficient functions β are direct extensions of the Lasso (Tibshirani, 1996) and variable fusion (Land and Friedman, 1996; Tibshirani et al., 2005) in the ordinary linear regression setting.

2.2 Basis function expansion and computational algorithm

In this section, we describe an efficient algorithm for computing the SSS estimate. We first reformulate our optimization problem (2) into that of an ordinary penalized regression by approximating each coefficient function $\beta_g(t)$ using linear combinations of the B-spline bases. We then compute the resulting optimization problem using an efficient coordinate descent method, proposed by Tseng (1988, 2001) and used previously by Friedman et al. (2010) for computing the lasso. Our algorithm can compute the SSS estimator exactly, even when p is large.

Consider the B-spline expansion $\beta_g(t) \approx \sum_{l=1}^{L_g} \gamma_{gl} B_l(t)$, where the number of basis functions $L_g = \text{no. interior knots} + \text{degree} + 1$. The degree is set equal to 3 throughout this paper. We

approximate the norm of a coefficient function $\|\beta_g\|$ as $\|\gamma_g\|_{\mathbf{R}_{gg}} = \sqrt{\gamma_g^T \mathbf{R}_{gg} \gamma_g}$, where $\gamma_g = (\gamma_{g1}, \dots, \gamma_{gL_g})^T$ and $\mathbf{R}_{g1g2} = (r_{ij})_{L_{g1} \times L_{g2}}$, $r_{ij} = \int B_i(t) B_j(t) dt$. Furthermore, the squared norm of the difference of a pair of coefficient functions $\|\beta_{g1} - \beta_{g2}\|^2$ can be written as

$$(\gamma_{g1}^T, \gamma_{g2}^T) \begin{pmatrix} \mathbf{R}_{g1g1} & -\mathbf{R}_{g1g2} \\ -\mathbf{R}_{g2g1} & \mathbf{R}_{g2g2} \end{pmatrix} \begin{pmatrix} \gamma_{g1} \\ \gamma_{g2} \end{pmatrix}$$

and the structural penalty $\sum_{g1=1}^{p-1} \sum_{g2=g1+1}^p w_{g1,g2} \|\beta_{g1} - \beta_{g2}\|^2$ as $\|\gamma\|_{\Omega}^2 = \gamma^T \Omega \gamma$, where

$$\gamma = (\gamma_1^T, \dots, \gamma_p^T)^T$$

and

$$\Omega = \begin{pmatrix} (\sum_{g \neq 1} w_{1g}) \mathbf{R}_{11} & -w_{12} \mathbf{R}_{12} & \cdots & -w_{1,p} \mathbf{R}_{1,p} \\ -w_{12} \mathbf{R}_{21} & (\sum_{g \neq 2} w_{2g}) \mathbf{R}_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & -w_{p-1,p} \mathbf{R}_{p-1,p} \\ -w_{1p} \mathbf{R}_{p1} & \cdots & -w_{p-1,p} \mathbf{R}_{p,p-1} & (\sum_{g \neq p} w_{pg}) \mathbf{R}_{pp} \end{pmatrix}.$$

Let

$$\mathbf{B}(t) = \begin{pmatrix} B_1(t) & \cdots & B_{L_1}(t) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ & \vdots & & & \ddots & & & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & B_1(t) & \cdots & B_{L_p}(t) \end{pmatrix},$$

$$\mathbf{U}_i(t_k) = (x_{i1}(t_k), \dots, x_{ip}(t_k)) \mathbf{B}(t_k), \mathbf{U} = ((\mathbf{U}_1(t_1))^T, \dots, \mathbf{U}_1(t_K)^T), \dots, (\mathbf{U}_n(t_1))^T, \dots, \mathbf{U}_n(t_K)^T)^T,$$

$$\mathbf{y}_i = (y_i(t_1), \dots, y_i(t_K))^T, \text{ and } \mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T.$$

Then, the optimization problem (2) is equivalent to

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{y} - \mathbf{U}\gamma\|^2 + \lambda_1 \sum_{g=1}^P \|\gamma_g\|_{\mathbf{R}_{gg}} + \frac{\lambda_2}{2} \gamma^T \mathbf{\Omega} \gamma. \quad (4)$$

We further employ the Cholesky decomposition of $\mathbf{R}_{gg} = \mathbf{V}_g^T \mathbf{V}_g$ to obtain a reparameterization of (4) as

$$\hat{\gamma}^* = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{y} - \mathbf{U}^* \gamma^*\|^2 + \lambda_1 \sum_{g=1}^P \|\gamma_g^*\| + \frac{\lambda_2}{2} (\gamma^*)^T \mathbf{\Omega}^* \gamma^*, \quad (5)$$

where $\gamma_g^* = \mathbf{V}_g \gamma_g$, $\mathbf{\Omega}^* = (\mathbf{V}^{-1})^T \mathbf{\Omega} \mathbf{V}^{-1} = \operatorname{diag}(\mathbf{V}_1^{-1}, \dots, \mathbf{V}_p^{-1})^T \mathbf{\Omega} \operatorname{diag}(\mathbf{V}_1^{-1}, \dots, \mathbf{V}_p^{-1})$, and the columns of $\mathbf{U}^*$ that correspond to the g th covariate is defined as $\mathbf{U}_g^* = \mathbf{U}_g \mathbf{V}_g^{-1}$. This allows us to reparameterize $\|\gamma_g\|_{\mathbf{R}_{gg}}$ in (4) as a Euclidean norm $\|\gamma_g^*\|$, which relates our estimator to the group lasso of Yuan and Lin (2006) and provides a computationally simpler formulation.

We now compute the solution to (5) using the coordinate descent algorithm. Given the current coefficients γ^* , we update γ^* at each step by setting

$$\gamma_g^* \leftarrow 0, \text{ if } \left\| -\frac{1}{n} (\mathbf{U}_g^*)^T \mathbf{y} + \frac{1}{n} (\mathbf{U}_g^*)^T \mathbf{U}_{-g}^* \gamma_{-g}^* + \lambda_2 \mathbf{\Omega}_{g,-g}^* \gamma_{-g}^* \right\| \leq \lambda_1, \quad (6)$$

where the indices $-g = \{1, \dots, \sum_{g'=1}^{(g-1)} L_{g'}, (\sum_{g'=1}^g L_{g'}) + 1, \dots, \sum_{g'=1}^p L_{g'}\}$, and

$$\gamma_g^* \leftarrow \underset{\gamma_g^*}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{y} - \mathbf{U}^* \gamma^*\|^2 + \lambda_1 \sum_{g=1}^P \|\gamma_g^*\| + \frac{\lambda_2}{2} (\gamma^*)^T \mathbf{\Omega}^* \gamma^*, \text{ otherwise.} \quad (7)$$

The gradient and Jacobian of the objective function in (7) are, respectively,

$$\mathbf{Gr}_g = -\frac{1}{n} (\mathbf{U}_g^*)^T \mathbf{y} + \frac{1}{n} (\mathbf{U}_g^*)^T \mathbf{U}_{-g}^* \gamma_{-g}^* + \lambda_2 \mathbf{\Omega}_{g,-g}^* \gamma_{-g}^* + \frac{\lambda_1}{\|\gamma_g^*\|} \gamma_g^* + \left(\frac{1}{n} (\mathbf{U}_g^*)^T \mathbf{U}_g^* + \lambda_2 \mathbf{\Omega}_{gg}^* \right) \gamma_g^* \quad \text{and} \quad (8)$$

$$\mathbf{J}_g = \frac{\lambda_1}{\|\gamma_g^*\|} \left(\mathbf{I}_{L_g} - \frac{1}{\|\gamma_g^*\|^2} \gamma_g^* \gamma_g^{*T} \right) + \frac{1}{n} (\mathbf{U}_g^*)^T \mathbf{U}_g^* + \lambda_2 \mathbf{\Omega}_{gg}^*. \quad (9)$$

We note that updating the objective in (7) can be computationally expensive, especially when the dimension of the response $\mathbf{y}_{n \times K}$ is large, which is often the case for data of nonparametric models. To make our algorithm scalable, we compute (7) as a system of nonlinear equations of gradient $\mathbf{Gr}_g$, where the objective $\|\mathbf{Gr}_g\|^2$ is used instead of the one in (7). The gradient $\mathbf{Gr}_g$ has a dimension of L_g , which is usually chosen to be smaller than K . As the optimization problem is convex, the solution to (7) can be readily computed using the system of nonlinear equations of $\mathbf{Gr}_g$ and the Jacobian $\mathbf{J}_g$, of which only the last two terms and the first term, respectively, need to be recomputed at each iteration. We note that the rate of convergence in terms of number of iterations for nonlinear equations is typically slower than that for direct optimization, as the objective function is misspecified (Nocedal and Wright, 1999). However, in this situation, the benefit of computing an objective function with a much smaller dimension exceeds the effect of having slower convergence rate.

We update γ^* as the following. At each step, we first compute (7) for all γ_g^* such that $\gamma_g^* \neq 0$. This is repeated at most a few times. Then, we update γ_g^* for all $g = 1, \dots, p$ using both (6) and (7). The above procedures are typically repeated until the norm of the difference of γ^* with its updated values after a complete cycle do not exceed some tolerance point, the maximum number of iterations is reached, or a prescribed number of variables is exceeded.

Finally, the SSS estimates (2) are obtained by setting $\widehat{\beta}_g(t) = \sum_{l=1}^{L_g} \widehat{\gamma}_{gl} B_l(t)$ where $\widehat{\gamma}_g = \mathbf{V}_g^{-1} \widehat{\gamma}_g^*$.

The tuning parameters $(\lambda_1, \lambda_2, c)$ are chosen using the standard cross-validation. The programming code, written in Fortran, and its R-language wrapper for computing the SSS estimator using the algorithm described in this section are available online as supplemental materials.

3 Theoretical Properties

In this section, we present theoretical properties for the SSS estimate of $\beta_g(t)$. We derive fixed-knots asymptotic results, in which we assume that the true coefficient function $\beta_g^0(t)$ itself is a spline function with known number of knots L_g . Fixed-knots asymptotics are often employed in the nonparametric literature, such as Yu and Ruppert (2002), for obtaining results that are more easily related to practical methodology. Knots-increasing asymptotics, on the other hand, often provide only rate of convergence results but little further insights. A setback of fixed-knots asymptotics is that it does not incorporate possible bias of approximating $\beta_g^0(t)$ with a fixed number of knots. However, we note that the asymptotic bias when β_g^0 itself is not a spline function is often very small and may usually be ignored in practice (Ruppert, 2002). Furthermore, due to computational expenses and limited number of time points, using a very large number of bases is not practical in high-dimensional applications.

We employ the B-spline formulation of the SSS estimates presented in Section 2.2. Let $\beta_g^0(t) = \sum_{l=1}^{L_g} \gamma_{gl}^0 B_l(t)$ for fixed number of knots L_g . The original model (1) can then be rewritten as

$$y_i(t_k) = \sum_{g=1}^p \sum_{l=1}^{L_g} x_{ig}(t_k) B_l(t_k) \gamma_{gl}^0 + \varepsilon_{ik}, \quad (10)$$

with the corresponding population version of the model as

$$Y = U\gamma^0 + \varepsilon.$$

Due to one-to-one correspondence between $\hat{\beta}(t)$ and $\hat{\gamma}$, we only need to consider the theoretical properties of the estimate $\hat{\gamma}$ as defined in equation (4).

3.1 Model selection consistency when p is fixed

We first present the results on model selection consistency in the standard setting when p is fixed and $n \rightarrow \infty$. Without loss of generality, we assume that the true coefficient functions $\beta^0 = \{(\beta_{(1)}^0)^T, (\beta_{(2)}^0)^T\}^T$, where $\beta_{(1)}^0 = \{\beta_g^0 : \beta_g^0 \neq 0\}$ and $\beta_{(2)}^0 = \{\beta_g^0 : \beta_g^0 = 0\}$. Coefficients $\gamma_{(1)}^0$ and $\gamma_{(2)}^0$ are defined as concatenations of γ_g^0 defining $\beta_{(1)}^0$ and $\beta_{(2)}^0$, respectively. Note that $\beta_{(2)}^0 = \gamma_{(2)}^0 = 0$. With the same partition, let

$$U = (U_{(1)}, U_{(2)})$$

and

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}.$$

We assume the following regularity conditions on the distributions of the response Y and covariates U .

- A1** $E\|Y\|^4 < \infty$ and $E\|U\|^4 < \infty$.
- A2** $\Sigma_{UU} = E(U(U)^T) - E(U)E(U)^T$ is positive definite.
- A3** Assume $E((Y - \gamma^T U)^2 | U)$ is greater than $\sigma_0^2 > 0$ almost surely, where $\gamma = \arg \min_{\gamma} E(Y - \gamma^T U)^2$.

The SSS estimator $\hat{\gamma}$ is model selection consistent if

$$\lim_{n \rightarrow \infty} P(\{g : \hat{\gamma}_g \neq 0\} = \{g : \gamma_g^0 \neq 0\}) = 1. \quad (11)$$

Let the norm of a matrix $\mathbf{A}$ be defined as $\|\mathbf{A}\| = \sqrt{\text{tr}(\mathbf{A}^T \mathbf{A})}$ and $\mathbf{D}_{(1)} = \text{Diag}(\|\gamma_g^0\|_{R_{gg}}^{-1})$ as the block-diagonal matrix of $\|\gamma_g^0\|_{R_{gg}}^{-1} \mathbf{I}_{L_g}$ for all $g \in \{g : \beta_g^0 \neq 0\}$, where $\mathbf{I}_{L_g}$ denotes the identity matrix of dimension L_g . We provide the following necessary and sufficient theorems for $\hat{\gamma}$ to be model selection consistent.

Theorem 3.1—Assume the regularity conditions (A1) and (A2). The SSS estimator $\hat{\gamma}$ is model selection consistent if the condition

$$\|\mathbf{V}_g^{-1} \left[\sum_{u_g u_{(1)}} \sum_{u_{(1)} u_{(1)}}^{-1} (\mathbf{D}_{(1)} + \alpha \Omega_{11}) - \alpha \Omega_{21} \right] \gamma_{(1)}^0\| < 1 \quad (12)$$

for all $g \in \{g: \gamma_g^0=0\}$ is satisfied for λ_1 and λ_2 such that $\lambda_1 \rightarrow 0$, $\lambda_2 \rightarrow 0$, $\lambda_1 \sqrt{n} \rightarrow \infty$, and $\lambda_2/\lambda_1 \rightarrow \alpha$.

Theorem 3.2—Assume the regularity conditions (A1)–(A3). The SSS estimator $\hat{\beta}$ is model selection consistent only if the condition

$$\|\mathbf{V}_g^{-1} \left[\sum_{\mathbf{u}_g \mathbf{u}_{(1)}} \sum_{\mathbf{u}_{(1)} \mathbf{u}_{(1)}}^{-1} (\mathbf{D}_{(1)} + \alpha \mathbf{\Omega}_{11}) - \alpha \mathbf{\Omega}_{21} \right] \gamma_{(1)}^0\| \leq 1 \quad (13)$$

for all $g \in \{g: \gamma_g^0=0\}$ is satisfied for λ_1 and λ_2 such that $\lambda_1 \rightarrow 0$, $\lambda_2 \rightarrow 0$, and $\lambda_2/\lambda_1 \rightarrow \alpha$.

The proofs for Theorem 3.1 and 3.2, which extend those of Bach (2008) for the group lasso, are given in Appendix, available as online supplemental materials. Theorem 3.1 and 3.2 suggest that the SSS estimator is model selection consistent under necessary and sufficient conditions (13) and (12) and restrictions on tuning parameters λ_1 , λ_2 . These conditions are similar to the irrerepresentable conditions for the lasso (Zhao and Yu, 2006) and group lasso (Bach, 2008), which involve both covariances among relevant covariates and covariances between relevant and irrelevant covariates. In addition, the conditions also involve the connection matrix in the covariate graph through $\mathbf{\Omega}$. When $\alpha \rightarrow 0$ and $\mathbf{V}_g = \mathbf{I}_{L_g}$, conditions (12) and (13) become those for the ordinary group lasso with $\mathbf{U}$ as predictors (Bach, 2008). When the group size is one and $\lambda_2 = 0$, the conditions reduce to the irrerepresentable conditions of Zhao and Yu (2006) for standard lasso.

3.2 Estimation bounds when p is larger than n

In this section, we study the performance of the estimator $\hat{\gamma}$ defined as in equation (4) in the high-dimensional case when p can be larger than n . In particular, we provide non-asymptotic bounds on the squared risk and the l_2 estimation error of $\hat{\gamma}$. In this asymptotic analysis, the tuning parameters λ_1 and λ_2 are chosen depending on the sample size n . We emphasize this dependency by adding a subscript to these two parameters. In order to facilitate the

presentation of the theorem, we introduce the following notation. Let $q = \sum_{g=1}^p L_g$ be the dimension of γ and $\iota = \max_g L_g$ be the maximum group size. In addition, we define

$\sigma_j^2 = \mathbf{U}_j^{*T} \mathbf{U}_j^*$, where $\mathbf{U}_j^*$ is the j th column of the matrix $\mathbf{U}^*$. We assume σ_j^2 is bounded away from 0 and ∞ , i.e. there exist $\sigma_{\min}$ and $\sigma_{\max}$ such that

$$0 < \sigma_{\min} \leq \sigma_i \leq \sigma_{\max} < \infty, \quad \forall i=1, \dots, q.$$

In addition, let μ_g be the smallest eigenvalue of the positive-definite matrix $\mathbf{R}_{gg}$ and is bounded away from zero, $\min_g \mu_g > \mu_{\min} > 0$.

In order to obtain the estimation bounds for $\hat{\gamma}$, we require an additional condition on the matrix

$$\mathbf{D}_n = \frac{1}{2n} (\mathbf{V}^{-1})^T (\mathbf{U}^T \mathbf{U} + n \lambda_{2n} \mathbf{\Omega}) \mathbf{V}^{-1}.$$

Condition $B(\Theta, \tau)$ —There exists a constant $\varphi > 0$, and $0 < \tau < 1/2$ such that, for any $\eta \in \mathbb{R}^q$ that satisfies $\sum_g \|\eta_g\| \leq \frac{2}{1-2\tau} \sum_{g \in \Theta} \|\eta_g\|$, we have

$$\eta^T \mathbf{D}_n \eta \geq \varphi \sum_{g \in \Theta} \|\eta_g\|^2,$$

where η_g is the subset of η that corresponds to the coefficients of the $\beta_g(t)$ in B-spline basis expansion. This assumption is similar to the restricted eigenvalue assumption introduced in Bickel et al. (2009).

Theorem 3.3—Consider model (10). Let $\mathcal{A}^0 = \{g: \gamma_g^0 \neq 0\}$ be the sparsity set, and let the tuning parameters λ_{1n} and λ_{2n} be defined as

$$\lambda_{1n} = \kappa \sigma_{\min} \sqrt{(nK)^{-1} \log(q)}, \text{ and } \lambda_{2n} = \frac{\tau \lambda_{1n}}{r^*},$$

where $r^* = \|\Omega \gamma\|_{\infty}$. If Condition $B(\mathcal{A}^0, \tau)$ holds, then with probability greater than $1 - u_{n,p}$ we have

$$\begin{aligned} \frac{1}{n} \|\mathbf{U} \gamma^0 - \mathbf{U} \hat{\gamma}\|^2 &\leq \frac{8\lambda_{1n}^2}{\varphi} |\mathcal{A}^0|, \\ (\gamma^0 - \hat{\gamma})^T \Omega (\gamma^0 - \hat{\gamma}) &\leq \frac{8r^* \lambda_{1n}}{\varphi \tau} |\mathcal{A}^0|, \\ \text{and } \sum_g \|\gamma_g^0 - \hat{\gamma}_g\| &\leq \frac{4\lambda_{1n}}{(1-2\tau)\varphi \mu_{\min}} |\mathcal{A}^0|, \end{aligned}$$

where $u_{n,p} = p^{1 - \kappa^2 \tau^2 \sigma_{\min}^2 / (2K^2 \iota^2 \sigma_{\max}^2)}$.

The proof of this theorem is based on the ‘argmin’ definition of the estimation $\hat{\gamma}$ and certain concentration inequalities, extending that of Hebiri and van de Geer (2010) for $L_1 + L_2$ penalized estimation methods. Details are given in Appendix, available as online supplemental materials. It provides bounds on the prediction and estimation errors with high probability. These prediction and estimation errors are bounded by $C_0 |\mathcal{A}^0| \sqrt{\log(p/n)}$, where C_0 is a constant. If the number of relevant variables $r_n = |\mathcal{A}^0|$ is fixed, these bounds go to zero when $\log(p)/n \rightarrow 0$. The theorem also holds when r_n diverges, in which case we require that $r_n \sqrt{\log(p)/n} \rightarrow 0$ in order for these bounds to go to zero.

4 Simulation Studies

In this section, we examine the finite-sample performances of the SSS estimator via simulation studies. The effects of incorporating the structural information of the covariates and high-dimensionality on variable selection are illustrated. Further, a study is conducted to corroborate the re-scaling scheme introduced in Section 2 for estimating the model (3).

We assume that the linear varying-coefficient model

$$y_i(t_k) = \beta_0(t) + \sum_{g=1}^{p-1} x_{ig}(t_k) \beta_g(t_k) + \varepsilon_{ik}, \quad (14)$$

has random error $\varepsilon_{ik} \sim N(0, \sigma)$ and covariates $\mathbf{X}(t_k) \sim N(\mathbf{0}, \Sigma)$. We repeatedly generate the data from the model (14) 200 times with $n = 50$ number of observations, and we apply a five-fold cross-validation for each simulation to estimate the tuning parameters λ_1 , λ_2 , and c (Huang et al., 2004). Let p_0 be the number of true non-zero coefficients. We examine variable selection performances using sensitivity, specificity, and G -measure. Sensitivity is defined as $(\text{no. of true positives})/p_0$ and describes the proportion of selecting the true variables correctly, whereas specificity, $1 - (\text{no. of false positives})/(p - p_0)$, is the proportion of discarding the irrelevant variables correctly. G -measure is defined as the geometric mean between sensitivity and specificity, $\sqrt{\text{sensitivity} * \text{specificity}}$. G -measure close to 1 indicates accurate selection, and G -measure close to 0 implies that few true variables or too many irrelevant variables are selected, or both. We report median values and bootstrapped standard deviations of medians out of 500 re-samplings, in parentheses, for each measurement based upon 200 repetitions.

We consider the following true coefficient functions over $t \in [0, 1]$

- Piecewise Constant: $f_{pc}(t) = -I\{t < 0.5\} + I\{t \geq 0.5\}$
- Polynomial: $f_{pn1}(t) = 2t - 1$, $f_{pn2}(t) = (1 - 2t)^3$
- Low-Frequency: $f_{l1}(t) = \sin(2\pi t)$, $f_{l2}(t) = \cos(2\pi t)$
- High-Frequency: $f_{h1}(t) = \sin(8\pi t)$, $f_{h2}(t) = \cos(8\pi t)$
- Compound: $f_{cp}(t) = (1 - 2t)^3 - \cos(4\pi t)$.

In each example, we use $K = 20$ time points sampled equidistantly from $[0, 1]$. We note that, in many applications such as motif finding, data are often sampled over well-defined biological cycles and important effects often vary over time via simple functions, which may be well approximated using a small number of bases. Furthermore, the number of bases that can be applied in high-dimensional applications may be limited by computational feasibility. In the following analyses, we employ only $L = 4$ bases, unless otherwise stated, to construct cubic B-splines for each coefficient function $\beta_g(t)$. The results presented in this section appear to suggest that good variable selection performances may be achieved even though the number of bases is limited.

4.1 Structured covariates

We examine the effect of having structured covariates on variable selection for our SSS estimator. We present results for the SSS estimator (3) with potentially non-zero λ_2 and with $\lambda_2 = 0$ in order to demonstrate the advantage of incorporating the covariate structures. This example has $p = 101$ predictors with true coefficient functions $\beta_0 = 2f_{cp}$ and

$$\begin{aligned} \beta_1 &= (1 - \zeta)f_{lf2} + \zeta f_{pn1}, & \beta_5 &= (1 - \zeta)f_{pn2} + \zeta f_{pn1}, & \beta_9 &= 3f_{hf1}, \\ \beta_2 &= (1 - \zeta)f_{lf2} + \zeta f_{hf1}, & \beta_6 &= (1 - \zeta)f_{pn2} + \zeta f_{hf1}, & \beta_{10} &= 1.5f_{pn2}, \\ \beta_3 &= (1 - \zeta)f_{lf2} + \zeta f_{lf1}, & \beta_7 &= (1 - \zeta)f_{pn2} + \zeta f_{lf1}, & \beta_{11} &= 2f_{lf1}, \\ \beta_4 &= (1 - \zeta)f_{lf2} + \zeta f_{hf2}, & \beta_8 &= (1 - \zeta)f_{pn2} + \zeta f_{hf2}, & \beta_{12} &= 2f_{pc}, \end{aligned}$$

and $\beta_g = 0$ for $g = 13, \dots, 100$. The coefficient functions are identical within the groups $I_1 = \{1, 2, 3, 4\}$ and $I_2 = \{5, 6, 7, 8\}$ when $\zeta = 0$ and distinct when $\zeta = 1$. The covariates $\mathbf{X}$ are

generated with $\Sigma_{ij} = \rho$ for $i, j \in I_1$ and $i, j \in I_2$ and $\Sigma_{ij} = 0$ for $i \neq j$, otherwise. Thus, covariates are correlated within groups I_1 and I_2 . Further, we use $\sigma = 0.5$ and sample size $n = 50$. We set the weights such that $w_{ij} = 1$ for $i, j \in I_1$ and $i, j \in I_2$ and $w_{ij} = 0$, otherwise.

Denote the sparse structured estimator with possibly nonzero λ_2 as SSS and with zero-valued λ_2 as $\text{SSS}(\lambda_2 = 0)$. Table 1 presents variable selection results for both estimators. We note that SSS with larger values of G -measures tends to dominate $\text{SSS}(\lambda_2 = 0)$ in general for variable selection. The gains in performances for SSS tend to come from sensitivity, or the percentage of detecting relevant variables correctly. For example, SSS selects 92.3% of the true variables correctly, whereas $\text{SSS}(\lambda_2 = 0)$ selects only 76.9% when $\zeta = 0.25$ and $\rho = 0.95$. The differences between SSS and $\text{SSS}(\lambda_2 = 0)$ are most pronounced when ζ is relatively small or when the within-group effects are similar. Nonetheless, SSS still shows some advantages over $\text{SSS}(\lambda_2 = 0)$ even at $\zeta = 0.5$ when within-group effects are quite different. This suggests that the SSS estimator can sometimes be advantageous even when the within-group effects are not necessarily identical, allowing it to be useful when prior knowledge is only approximate. Further, the performances of $\text{SSS}(\lambda_2 = 0)$ tend to deteriorate when ρ is large or multicollinearity is strong. This may be due to difficulty for linear models to distinguish highly collinear variables during selection. We note that SSS tends to perform similarly across different values of ρ , which suggests that the SSS estimator can be robust against within-group collinearity in terms of variable selection.

Table 2 presents the percentage of times each individual relevant variable is selected and bias squared and variance for its coefficient estimate when $\zeta = 0.25$ and $\rho = 0.95$. Bias and variance for each coefficient is estimated by averaging over the K time points sampled. We see that SSS dominates $\text{SSS}(\lambda_2 = 0)$ by a large margin in terms of both variable selection and estimation for variables in groups I_1 and I_2 . On the other hand, variable selection and estimation results are quite similar between SSS and $\text{SSS}(\lambda_2 = 0)$ for variables with indices in $\{0, 9, 10, 11, 12\}$. This seems to suggest that the advantages of the SSS estimator come from utilizing inherent covariate structures.

Finally, we plot in Figure 1 the true and estimated relevant coefficient functions from one repetition using $L = 12$ for the model with $\zeta = 0.25$, $\rho = 0.95$ and $\sigma = 0.1$. We see that the estimated coefficient functions approximate the true ones pretty well. Further, we observe that estimates of $\beta_1(t), \dots, \beta_4(t)$ and $\beta_5(t), \dots, \beta_8(t)$ are smoothed amongst themselves to effect in preserving covariate structures in groups $I_1 = \{1, \dots, 4\}$ and $I_2 = \{5, \dots, 8\}$, respectively.

4.2 High-dimensionality

In this example, we study the effect of high-dimensionality on variable selection for the SSS estimator under finite-samples. This example uses the same setup as in Section 4.1, except that we fix $\zeta = 0.25$ and $\rho = 0.95$. Furthermore, we vary $p - 1$ over $\{50, 100, 200, 400\}$ and σ over $\{0.1, 0.25, 0.5\}$ in order to demonstrate the effects of having different numbers of variables and signal-to-noise ratios on variable selection.

Table 3 presents variable selection results for our estimator. Median numbers of true positives (TP) and false positives (FP) are also displayed, providing the exact counts of correctly selected and incorrectly selected variables, respectively. We observe that the numbers of true positives and, in complement, sensitivity decrease with increasing p . Specificity also tends to increase with increasing p . The effect is most significant when σ is large or signal-to-noise ratio is small. As the underlying model is nonparametric, the requirement on estimation accuracy of selected coefficients can be stringent, which may cause variable selection to be conservative when p is large. In practice, one may want to

relax the value of λ_1 estimated by cross-validation in order to increase the number of variables selected if the data at hand may have a very low signal-to-noise ratio.

Table 4 provides details on estimation and variable selection for $p - 1 = 50$ and $p - 1 = 400$ when $\sigma = 0.25$. We see that, when p is large, bias tends to increase but variances can sometimes be smaller. Furthermore, percentages of selection tend to be smaller for these relevant predictors when p is large. Variables with simple coefficient functions $\beta_{10} = 1.5f_{pn2}$, $\beta_{11} = 2f_{fl}$, $\beta_{12} = 2f_{pc}$ are selected well at high-dimensions, whereas those with complex coefficient functions, such as $\beta_9 = 3f_{fl}$, are selected poorly. Variable selection for predictors with complex coefficient functions can be improved by using larger number of bases. However, we note that, in applications such as motif finding, important biological effects are expected to vary over time via simple functions and can be readily detected even when L is small.

4.3 Coefficient re-scaling and effects of sample size

In this example, we verify the advantage of using the re-scaling scheme proposed in Section 2 for estimating the structured varying-coefficient model (3). We summarize results for the SSS estimator when its estimates are re-scaled and when they are not ($c = 1$). We employ the setup as in Section 4.1, except that we fix $\zeta = 0.25$ and $\rho = 0.95$ and vary σ over $\{0.1, 0.25, 0.5\}$ and n over $\{50, 200, 2000\}$.

We denote the SSS estimator with re-scaling as SSS and without re-scaling as $SSS(c = 1)$. Variable selection results are presented in Table 5. We see that SSS, with larger values of G -measures, significantly dominates $SSS(c = 1)$ in terms of variable selection. In particular, SSS dominates $SSS(c = 1)$ mainly in terms of specificity while sensitivity results are similar. This suggests that using the re-scaling scheme may correct for over-selection caused by bias from shrinkage of $\|\beta_g\|$ towards 0. Moreover, as sample size increases, we observe that the selection performances for the SSS with re-scaling improve quickly and can approach that of an optimal selection when n is large.

5 Application to DNA Motif Finding

In this section, we illustrate the SSS estimator in an application to DNA motif finding. Gene transcription is regulated by proteins, called transcription factors, which control the expression of genes by binding to recognition sites in promoter regions of the genes. Proteins often bind to the gene promoter regions by recognizing the low-entropy patterns of short sequences, called motifs. The identification of transcription factor binding motifs presents an important problem in genomics, which plays an integral role in the discovery of gene regulatory circuitries and understanding gene expression. One approach of identifying the motifs is through regression analysis of gene expression data using word frequencies as covariates. Previous works on DNA motif finding in such regression framework include Bussemaker et al. (2001), Conlon et al. (2003), Zhong et al. (2005) and Zhang et al. (2007). These papers, however, only assume a simple linear regression model, which may lose efficiency when the gene expression data are measured over time.

Transcription factor binding sites with similar sequences or words often serve as alternative binding sites for the same transcription factor. In particular, binding sites for the same transcription factor usually share a common core sequence but may have flanking sequences that are different (Mirny and Gelfand, 2002; Moses et al., 2003). As transcription factors usually contact the DNA at its major or minor grooves, core sequences are typically very short. For example, the transcription factor MCB, that controls gene expression at G1/S transition during the cell cycle, binds with the DNA at a 4-base core sequence of CGCG with various flanking bases. That is, the DNA binding sequences for MCB can be

ACGCGA, ACGCGT, CCGCGT, or TCGCGA. In de novo discovery of motifs, counts of occurrences of different word patterns in promoter sequences often serve as covariates, and their effects on gene expression are modeled. By considering words with the same core sequences to have similar functionalities, we see that counts of word patterns exemplify covariates that are structured, e.g., these words can be summarized as nodes of a hypercube (Li and Zhang, 2010). Utilizing structures in DNA motif finding has been shown to improve sensitivity in several recent papers (Kechris et al., 2004; Li and Zhang, 2010).

We consider the problem of identifying the motifs that are associated with yeast cell cycle process using the microarray time-course gene expression dataset from Spellman et al. (1998), which identified about 800 cell cycle-related genes. For each gene, expression data is sampled over two cell cycles at $K = 18$ time points. We use the counts of occurrences of 7-base word patterns as predictors, where a word and its complement in reverse are considered identical. Hence, there are a total of $4^7/2 = 8192$ different words. We consider the following LVC model,

$$y_i(t) = \beta_0(t) + \sum_{g=1}^{8192} x_{ig} \beta_g^0(t) + \varepsilon_{it}, \quad (15)$$

where $y_i(t)$ is the expression level of gene i at time t , x_{ig} is the number of occurrences of word g in the promoter sequence of gene i , and $\beta_g^0(t)$ measures the effect of the g th word on gene expression at time t .

Our goal is to identify these words in model (15) that have non-zero coefficients. To incorporate covariate structures in our model, we assign weights between effects of counts of 7-base word pairs, $\{b_{1g_1}, \dots, b_{7g_1}\}$ and $\{b_{1g_2}, \dots, b_{7g_2}\}$, as

$$w_{g_1, g_2} = \begin{cases} 1, & \text{if } b_{ig_1} \neq b_{ig_2} \text{ for exactly one } i \in \{1, 2, 6, 7\} \\ 0, & \text{otherwise,} \end{cases} \quad (16)$$

where we conservatively assume 3-base core sequences $\{b_{3g}, b_{4g}, b_{5g}\}$ and enforce affinity between words that differ in the flanking region $\{b_{1g}, b_{2g}, b_{6g}, b_{7g}\}$ by only one base. In order to discern the effect of utilizing covariate structures, we estimate the structured varying-coefficient model with possibly nonzero λ_2 and zero-valued λ_2 . As in Example 4.1, we denote these procedures as SSS and SSS($\lambda_2 = 0$), respectively.

Previous research has shown that B-splines with $L = 4$ basis functions can model the yeast cell cycle time course expression data well (Wang et al., 2008). We use simple cubic B-splines to approximate the coefficient functions $\beta_g(t)$, which results in 8,192 groups of covariates with each group containing four covariates, for a total of 32,768 covariates. In order to facilitate the computation, we employ two strategies to pre-select the words that are associated with gene expression. We first perform single-word pre-screening, in the same spirit as the sure independent screening (SIS) procedure of Fan and Lv (2008). Specifically, we perform a simple linear regression analysis to select the words that are associated with gene expression level at each time point. Following Conlon et al. (2003), words are only kept if they are associated with gene expression at at least one time point at p -value < 0.01 . This results in a total of 1,896 words as predictors.

We next employ the tournament screening strategy of Chen and Chen (2009) to consider groups of words using a divide-and-conquer scheme, where we randomly partition the 1896 predictors into 4 groups of 474 predictors each. Variables are selected using five-fold cross-

validation within each group, and the selected variables are combined. Since counts of related words may be assigned to different groups, to preserve covariate structures, we include discarded variables with words that differ from those selected by only one base in the flanking region. Over 25 repetitions, a median of 426 and 279 variables and a maximum of 952 and 553 variables are retained for SSS and $\text{SSS}(\lambda_2 = 0)$, respectively, at this stage. Finally, a five-fold cross-validation is performed upon the retained variables to yield the final result. The same procedure, SSS or $\text{SSS}(\lambda_2 = 0)$, is applied at each step. A median of 327 (31.4) and 101 (10.2) variables are selected, using the above procedure, for SSS and $\text{SSS}(\lambda_2 = 0)$, respectively, over 25 repetitions, where the values in parentheses refer to bootstrapped standard deviations of medians out of 500 re-samplings. To provide a succinct summary of results, we examine words selected 85% or more times by SSS among 25 repetitions. This amounts to 112 words that we depict in Figure 2, where an edge is drawn if a pair of words differ by only one base in the flanking region.

To further study the effect of utilizing covariate structures, we examine linked graphs with 3 or more words in Figure 2. Table 6 presents percentages of selection for both SSS and $\text{SSS}(\lambda_2 = 0)$ and the associated transcription factors for these words. The database, <http://rulai.cshl.org/SCPD/searchmotif.html>, is used to determine which transcription factor a word is associated with on its mapped binding site. First of all, we note that Table 6 includes many known cell cycle-regulating transcription factors, including ACE2, MCB, MCM1, ROX1, STE12, and SWI5. Different words for the same transcription factor often belong to the same linked graph. Most of these words correspond to known consensus sequences, indicating that the SSS procedure can indeed recover the motifs related to the cell cycle-related transcription factors. By imposing the smoothness constraints on the coefficient functions, the SSS identified more relevant words than the $\text{SSS}(\lambda_2 = 0)$. The comparisons are apparent between SSS and $\text{SSS}(\lambda_2 = 0)$ in linked graph #2 for ROX1, #2 and #5 for ACE2, #4 for MCB, and #7 for SWI5 (see Table 6).

It is interesting to note that MCM1, though an important factor in cell cycle-regulation, is often missed by existing motif finding procedures due to its high degeneracy (Bussemaker et al., 2001). The consensus of the binding motifs of MCM1 is CCNNNWWRGG, where letter N can be any of the bases A, T, G, C, letter W represents letter A or T, and letter R represents letter A or G. The SSS discovered 5 alternative words for MCM1, 2 in linked graph #1 with core sequence of TTT and 3 in #16. The linked graph #16 contains 3 words (TTTCCGA, TTTCCGT, TTTCCTA), all of which are variants for the MCM1 consensus motifs. The number of MCM1 associated words is also more than the number identified in Li and Zhang (2010), which discovered 3 variants for MCM1 using a Bayesian procedure to incorporate covariate structures. Moreover, variants of MCM1 are selected more than 90% of the time by SSS, whereas one variant in linked graph #16 and both variants in #1 are never selected by $\text{SSS}(\lambda_2 = 0)$.

6 Discussions and Extensions

In this paper, we have proposed a sparse structured shrinkage estimator for LVC models, a new nonparametric method that can incorporate covariate structures for high-dimensional variable selection. We presented simulation results that demonstrate the efficacy of our method under both high-dimensionality and structured covariates. Furthermore, we applied our method to DNA motif finding and demonstrated the applicability and advantage of our procedure.

Our method is unique in incorporating structural information of covariates under both high-dimensionality and nonparametric modeling. Our procedure specifically encourages smoothness of coefficient functions with related covariates. However, other types of

covariate structures can be incorporated, such as group or hierarchical structure (Yuan and Lin, 2006) and heredity or marginality principle (Hamada and Wu, 1992). Furthermore, we demonstrated the advantage of incorporating covariate structures for the varying-coefficient models. Other nonparametric procedures may also benefit from incorporating structural information of covariates, such as the generalized additive models (Hastie and Tibshirani, 1990) and partially linear models (Carroll et al., 1997). When *a priori* knowledge of covariate structures is not known, we may also utilize weights determined by similarity measures of covariates in (2). For example, Daye and Jeng (2008) have shown that, in the linear regression, correlation-driven weights with pairwise fusion can sometimes be useful under multicollinearity. Potential applicability of similarity-determined weights for the varying-coefficient model is suggested by Table 1, where the advantages of incorporating covariate structures are most apparent when covariates are similar.

The re-scaling parameter c is applied in (3) in order to alleviate bias due to shrinkage of coefficient estimations by the sparsity penalty. Alternatively, one can use a non-convex penalty such as the SCAD (Fan and Li, 2001) in place of the L_1 penalty to achieve sparse solutions as in Wang et al. (2008) and to alleviate biases. However, under high-dimensionality, non-convex optimization can be difficult to compute for a nonparametric model. The technique in (3) allows us to approximate for unbiased estimates while using an efficient convex optimization. In Section 4, we show that this strategy works well in improving variable selection. However, theoretical work is needed to better understand the effect of re-scaling on parameter estimate.

Finally, in our application to the motif finding problem, we only consider the main effects of the motifs on gene expressions, as in most of the regression-based motif finding methods (Bussemaker et al., 2001; Conlon et al., 2003; Li and Zhang, 2010). We can in principle include the interaction effects in the model; however, this creates additional computational challenges to our problem given the large sample size and the large number of words considered. The fact that our procedure can recover most of the known cell-cycle related motifs indicates that such interaction effects might not be strong or that they can also be reflected in the main effects.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by NIH grants ES009911 and CA127334. We thank Nancy R. Zhang for kindly providing us the data set they compiled for Li and Zhang (2010) and three reviewers for their very helpful comments.

References

- Bach FR. Consistency of the Group Lasso and Multiple Kernel Learning. *Journal of Machine Learning Research*. 2008; 9:1179–1225.
- Bickel PJ, Ritov Y, Tsybakov A. Simultaneous Analysis of Lasso and Dantzig Selector. *The Annals of Statistics*. 2009; 37:1705–1732.
- Bussemaker HJ, Li H, Siggia ED. Regulatory Element Detection Using Correlation with Expression. *Nature Genetics*. 2001; 27:167–171. [PubMed: 11175784]
- Cai ZW, Fan JQ, Li RZ. Efficient Estimation and Inferences for Varying-Coefficient Models. *Journal of the Royal Statistical Society, Ser B*. 2000; 95:888–902.
- Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized Partially Linear Single-Index Models. *Journal of the American Statistical Association*. 1997; 92:477–489.

- Chen Z, Chen J. Tournament Screening cum EBIC for Feature Selection with High Dimensional Feature Spaces. *Science in China, Series A: Mathematics*. 2009; 52:1327–1341.
- Conlon EM, Liu XS, Lieb JD, Liu JS. Integrating Regulatory Motif Discovery and Genome-Wide Expression Analysis. *Proceedings of National Academy of Science USA*. 2003; 100:3339–3344.
- Daye ZJ, Jeng XJ. Shrinkage and Model Selection with Correlated Variables via Weighted Fusion. *Computational Statistics and Data Analysis*. 2008; 53:1284–1298.
- Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. *The Annals of Statistics*. 2004; 32:407–499.
- Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Fan J, Lv J. Sure Independence Screening for Ultra-High Dimensional Feature Space. *Journal of the Royal Statistical Society, Ser B*. 2008; 70:849–911.
- Fan J, Zhang W. Statistical Methods with Varying Coefficient Models. *Statistics and its Interface*. 2008; 1:179–195. [PubMed: 18978950]
- Friedman J, Hastie T, Hoefling H, Tibshirani R. Pathwise Coordinate Optimization. *The Annals of Applied Statistics*. 2007; 1:302–332.
- Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*. 2010; 33
- Hamada M, Wu C. Analysis of Designed Experiments with Complex Aliasing. *Journal of Quality Technology*. 1992; 24:130–137.
- Hastie, T.; Tibshirani, R. *Generalized Additive Models*. Chapman & Hall/CRC; 1990.
- Hastie T, Tibshirani R. Varying-Coefficient Models. *Journal of the Royal Statistical Society, Ser B*. 1993; 55:757–796.
- Hebiri M, van de Geer S. The Smooth-Lasso and Other $l_1 + l_2$ -Penalized Methods. 2010:1003.4885v1. arXiv.
- Hoover DR, Rice JA, Wu CO, Yang L. Nonparametric Smoothing Estimates of Time-Varying Coefficient Models with Longitudinal Data. *Biometrika*. 1998; 85:809–822.
- Huang J, Horowitz JL, Wei F. Variable Selection in Nonparametric Additive Models. *The Annals of Statistics*. 2010; 38:2282–2313.
- Huang JZ, Wu CO, Zhou L. Polynomial Spline Estimation and Inference for Varying Coefficient Models with Longitudinal Data. *Statistica Sinica*. 2004; 14:763–788.
- Kechris K, van Zwet E, Bickel P, Eisen MB. Detecting DNA Regulatory Motifs by Incorporating Positional Trends in Information Content. *Genome Biology*. 2004; 5:R50. [PubMed: 15239835]
- Land, S.; Friedman, J. Tech rep. Department of Statistics, Stanford University; 1996. Variable Fusion: a New Method of Adaptive Signal Regression.
- Li C, Li H. Network-Constrained Regularization and Variable Selection for Analysis of Genomic Data. *Bioinformatics*. 2008; 24:1175–1182. [PubMed: 18310618]
- Li C, Li H. Variable Selection and Regression Analysis for Graph-Structured Covariates with an Application to Genomics. *The Annals of Applied Statistics*. 2010; 4:1498–1516.
- Li F, Zhang NR. Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces with Applications in Genomics. *Journal of the American Statistical Association*. 2010; 105:1202–1214.
- Meier L, van de Geer S, Bühlmann P. High-Dimensional Additive Modeling. *The Annals of Statistics*. 2009; 37:3779–3821.
- Meinshausen N, Bühlmann P. High Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*. 2006; 34:1436–1462.
- Mirny L, Gelfand M. Structural Analysis of Conserved Base Pairs in Protein-DNA Complexes. *Nucleic Acids Research*. 2002; 30:1704–1711. [PubMed: 11917033]
- Moses A, Chiang D, Kellis M, Lander E, Eisen M. Position Specific Variation in the Rate of Evolution in Transcription Factor Binding Sites. *BMC Evolutionary Biology*. 2003; 3:19. [PubMed: 12946282]
- Nocedal, J.; Wright, S. *Numerical Optimization*. Springer; 1999.
- Ravikumar P, Lafferty J, Liu H, Wasserman L. Sparse Additive Models. *Journal of the Royal Statistical Society, Ser B*. 2009; 71:1009–1030.

- Rosset S, Zhu J. Piecewise Linear Regularized Solution. *The Annals of Statistics*. 2007; 35:1012–1030.
- Ruppert D. Selecting the Number of Knots for Penalized Splines. *Journal of Computational and Graphical Statistics*. 2002; 11:735–757.
- Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell*. 1998; 9:3273–3297. [PubMed: 9843569]
- Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Ser B*. 1996; 58:267–288.
- Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society, Ser B*. 2005; 67:91–108.
- Tseng, P. Tech Rep. Massachusetts Institute of Technology, Laboratory for Information and Decision Systems; 1988. Coordinate Ascent for Maximizing Nondifferentiable Concave Functions. LIDS-P 1840
- Tseng P. Convergence of Block Coordinate Descent Method for Nondifferentiable Maximization. *Journal of Optimization Theory and Applications*. 2001; 109:474–494.
- Wang H, Xia Y. Shrinkage Estimation of the Varying Coefficient Model. *Journal of the American Statistical Association*. 2009; 104:747–757.
- Wang L, Chen G, Li H. Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data. *Bioinformatics*. 2007; 23:1486–1544. [PubMed: 17463025]
- Wang L, Li H, Huang J. Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements. *Journal of the Royal Statistical Society, Ser B*. 2008; 103:1556–1569.
- Witten DM, Tibshirani R. Covariance-Regularized Regression and Classification for High-Dimensional Problems. *Journal of the Royal Statistical Society, Ser B*. 2009; 71:615–636.
- Yu Y, Ruppert D. Penalized Spline Estimation for Partially Linear Single-Index Models. *Journal of the American Statistical Association*. 2002; 97:1042–1054.
- Yuan M, Lin Y. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, Ser B*. 2006; 68:49–67.
- Zhang NR, Wildermuth MC, Speed TP. Transcription Factor Binding Site Prediction with Multivariate Gene Expression Data. *The Annals of Applied Statistics*. 2007; 2:332–365.
- Zhao P, Yu B. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*. 2006; 7:2541–2567.
- Zhong W, Zeng P, Ma P, Liu JS, Zhu Y. RSIR: Regularized Sliced Inverse Regression for Motif Discovery. *Bioinformatics*. 2005; 21:4169–4175. [PubMed: 16166098]

Figure 1.

Estimated (dashed lines) and true (solid lines) coefficient functions $\beta_1(t) - \beta_{12}(t)$ for one replication for the model with $\zeta = 0.25$, $\rho = 0.95$ and $\sigma = 0.1$.

Figure 2. DNA motif finding. Words selected 85% or more times by the structured varying-coefficient model among 25 repetitions. Words that are linked differ only by one letter in the flanking regions.

Table 1

Example 4.1. Variable selection performance results based upon 200 replications with structured covariates, $p = 101$, $n = 50$. SSS: the sparse structured shrinkage estimator with the tuning parameters selected by 5-fold CVs; SSS($\lambda_2 = 0$): the sparse structured shrinkage estimator with $\lambda_2 = 0$ and other tuning parameters selected by 5-fold CVs.

ζ	ρ	SSS		G-measure		SSS ($\lambda_2 = 0$)		G-measure	
		sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
0	0.95	0.923 (0.002)	0.920 (0.006)	0.916 (0.006)		0.769 (0.000)	0.920 (0.006)	0.834 (0.007)	
	0.85	1.000 (0.037)	0.886 (0.007)	0.916 (0.003)		0.923 (0.031)	0.886 (0.006)	0.877 (0.003)	
	0.65	1.000 (0.000)	0.818 (0.009)	0.896 (0.004)		0.923 (0.010)	0.818 (0.010)	0.885 (0.004)	
	0.25	0.923 (0.024)	0.920 (0.007)	0.899 (0.008)		0.923 (0.000)	0.920 (0.007)	0.893 (0.007)	
0.25	0.95	0.923 (0.004)	0.830 (0.014)	0.853 (0.007)		0.769 (0.032)	0.835 (0.015)	0.784 (0.003)	
	0.85	0.923 (0.002)	0.830 (0.020)	0.862 (0.006)		0.846 (0.036)	0.830 (0.023)	0.812 (0.006)	
	0.65	0.923 (0.022)	0.932 (0.011)	0.872 (0.007)		0.846 (0.012)	0.938 (0.010)	0.857 (0.005)	
	0.25	0.846 (0.014)	0.915 (0.009)	0.853 (0.004)		0.846 (0.000)	0.920 (0.009)	0.848 (0.005)	
0.5	0.95	0.692 (0.036)	0.943 (0.006)	0.793 (0.012)		0.615 (0.031)	0.943 (0.006)	0.758 (0.009)	
	0.85	0.769 (0.000)	0.943 (0.005)	0.819 (0.009)		0.692 (0.000)	0.943 (0.005)	0.784 (0.007)	
	0.65	0.769 (0.002)	0.932 (0.005)	0.818 (0.008)		0.692 (0.037)	0.932 (0.005)	0.803 (0.006)	
	0.25	0.692 (0.028)	0.920 (0.007)	0.784 (0.007)		0.692 (0.024)	0.920 (0.007)	0.782 (0.008)	

Table 2

Example 4.1. Percentage selected, bias², and variance when $\zeta = 0.25$ and $\rho = 0.95$ based upon 200 replications with structured covariates, $p = 101$, $n = 50$. SSS: the sparse structured shrinkage estimator with the tuning parameters selected by 5-fold CVs; SSS($\lambda_2 = 0$): the sparse structured shrinkage estimator with $\lambda_2 = 0$ and other tuning parameters selected by 5-fold CVs.

	SSS			SSS ($\lambda_2 = 0$)		
	% Selected	Bias ²	Var	% Selected	Bias ²	Var
β_0	1.000	2.125 (0.013)	1.263 (0.045)	1.000	2.118 (0.015)	1.244 (0.046)
β_1	0.870	0.141 (0.005)	0.269 (0.015)	0.665	0.245 (0.032)	0.363 (0.051)
β_2	0.870	0.152 (0.003)	0.292 (0.010)	0.725	0.262 (0.022)	0.398 (0.041)
β_3	0.880	0.139 (0.005)	0.286 (0.017)	0.710	0.269 (0.031)	0.446 (0.055)
β_4	0.900	0.145 (0.003)	0.305 (0.010)	0.700	0.257 (0.038)	0.390 (0.037)
β_5	0.735	0.036 (0.003)	0.058 (0.004)	0.435	0.067 (0.000)	0.085 (0.000)
β_6	0.835	0.083 (0.004)	0.088 (0.004)	0.575	0.130 (0.003)	0.139 (0.009)
β_7	0.795	0.087 (0.006)	0.080 (0.005)	0.545	0.133 (0.006)	0.127 (0.004)
β_8	0.795	0.089 (0.004)	0.098 (0.007)	0.535	0.136 (0.000)	0.148 (0.005)
β_9	0.680	4.363 (0.019)	0.038 (0.004)	0.680	4.364 (0.018)	0.037 (0.004)
β_{10}	0.900	0.150 (0.008)	0.097 (0.011)	0.895	0.151 (0.009)	0.096 (0.011)
β_{11}	0.995	0.267 (0.012)	1.138 (0.036)	0.995	0.272 (0.011)	1.121 (0.040)
β_{12}	1.000	0.762 (0.009)	2.647 (0.076)	1.000	0.768 (0.008)	2.611 (0.080)

Table 3

Example 4.2. Results of variable selection performance of the SSS estimator based upon 200 replications for different numbers of predictors p and σ , where the sample size is fixed as $n = 50$.

σ	$p - 1$	TP	FP	sensitivity	specificity	G-measure
0.1	50	13.0 (0.0)	4.0 (0.3)	1.000 (0.000)	0.895 (0.006)	0.934 (0.006)
	100	13.0 (0.0)	10.0 (0.5)	1.000 (0.000)	0.886 (0.006)	0.941 (0.003)
	200	13.0 (0.0)	11.0 (4.4)	1.000 (0.000)	0.941 (0.022)	0.959 (0.004)
	400	12.0 (0.2)	3.0 (0.4)	0.923 (0.018)	0.992 (0.001)	0.960 (0.003)
0.25	50	13.0 (0.0)	9.0 (0.5)	1.000 (0.000)	0.763 (0.014)	0.871 (0.008)
	100	13.0 (0.0)	3.5 (0.5)	1.000 (0.000)	0.960 (0.007)	0.950 (0.009)
	200	12.0 (0.4)	13.0 (0.7)	0.923 (0.031)	0.931 (0.004)	0.935 (0.003)
	400	10.0 (0.2)	3.5 (0.6)	0.769 (0.020)	0.991 (0.002)	0.832 (0.011)
0.5	50	13.0 (0.0)	4.0 (0.6)	1.000 (0.003)	0.895 (0.015)	0.918 (0.007)
	100	12.0 (0.1)	15.0 (1.2)	0.923 (0.004)	0.830 (0.014)	0.853 (0.007)
	200	9.0 (0.3)	11.0 (0.6)	0.692 (0.020)	0.941 (0.003)	0.807 (0.008)
	400	6.0 (0.1)	6.0 (1.3)	0.462 (0.011)	0.985 (0.004)	0.676 (0.003)

Table 4

Example 4.2. Percentage selected, bias², and variance of the SSS estimator when $\sigma = 0.25$ based upon 200 replications, where the sample size is fixed as $n = 50$.

	$p - 1 = 50$			$p - 1 = 400$		
	% Selected	Bias ²	Var	% Selected	Bias ²	Var
β_0	1.000	1.946 (0.004)	0.974 (0.020)	1.000	2.216 (0.028)	1.532 (0.030)
β_1	0.980	0.065 (0.001)	0.246 (0.005)	0.785	0.221 (0.020)	0.409 (0.047)
β_2	0.965	0.085 (0.001)	0.261 (0.005)	0.805	0.223 (0.015)	0.359 (0.021)
β_3	0.985	0.077 (0.001)	0.268 (0.005)	0.725	0.222 (0.014)	0.308 (0.029)
β_4	0.975	0.074 (0.001)	0.253 (0.005)	0.755	0.266 (0.021)	0.414 (0.041)
β_5	0.895	0.019 (0.001)	0.058 (0.001)	0.585	0.072 (0.000)	0.105 (0.006)
β_6	0.955	0.052 (0.001)	0.092 (0.002)	0.665	0.103 (0.013)	0.099 (0.007)
β_7	0.970	0.050 (0.002)	0.102 (0.003)	0.570	0.118 (0.015)	0.085 (0.004)
β_8	0.925	0.048 (0.001)	0.074 (0.002)	0.505	0.104 (0.000)	0.085 (0.003)
β_9	0.985	4.220 (0.009)	0.085 (0.005)	0.410	4.427 (0.000)	0.017 (0.000)
β_{10}	1.000	0.053 (0.003)	0.179 (0.007)	0.705	0.212 (0.023)	0.052 (0.006)
β_{11}	1.000	0.071 (0.003)	1.659 (0.023)	1.000	0.381 (0.022)	0.889 (0.037)
β_{12}	1.000	0.614 (0.003)	3.323 (0.045)	1.000	0.817 (0.012)	2.433 (0.071)

Example 4.3. Results of variable selection performance of the SSS estimator based upon 200 replications for examining re-scaling of estimates, where $p = 101$. SSS: the sparse structured shrinkage estimator with the tuning parameters selected by 5-fold CVs; SSS($c = 1$): the sparse structured shrinkage estimator with $c = 1$ and other tuning parameters selected by 5-fold CVs.

Table 5

σ	n	SSS			SSS ($c = 1$)		
		sensitivity	specificity	G-measure	sensitivity	specificity	G-measure
0.1	50	1.000 (0.000)	0.886 (0.006)	0.941 (0.003)	1.000 (0.000)	0.875 (0.006)	0.926 (0.006)
	200	1.000 (0.000)	0.920 (0.003)	0.959 (0.002)	1.000 (0.000)	0.920 (0.004)	0.959 (0.002)
	2000	1.000 (0.000)	0.920 (0.003)	0.959 (0.002)	1.000 (0.000)	0.920 (0.003)	0.959 (0.002)
0.25	50	1.000 (0.000)	0.960 (0.007)	0.950 (0.009)	1.000 (0.002)	0.705 (0.008)	0.826 (0.009)
	200	1.000 (0.000)	0.989 (0.002)	0.994 (0.003)	1.000 (0.000)	0.705 (0.007)	0.839 (0.005)
	2000	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.670 (0.006)	0.819 (0.003)
0.5	50	0.923 (0.004)	0.830 (0.014)	0.853 (0.007)	0.923 (0.015)	0.773 (0.006)	0.826 (0.006)
	200	1.000 (0.000)	0.830 (0.006)	0.905 (0.003)	1.000 (0.000)	0.795 (0.006)	0.885 (0.005)
	2000	1.000 (0.000)	0.830 (0.014)	0.911 (0.008)	1.000 (0.000)	0.784 (0.007)	0.883 (0.004)

DNA motif finding. Words selected 85% or more times by the structured varying-coefficient model among 25 repetitions and belonging to linked graphs of 3 or more elements. (Underlined indicates transcription factors known to be cell cycle-regulating.)

Table 6

Word pattern	% select SSS	% select SSS ($\lambda_2 = 0$)	Transcription factors
# 1 (12 words)			
TATTTGT	0.92	0.00	CUP2
TCTTTTA	0.88	0.60	
TCTTTTG	0.92	0.20	
TGTTTAC	0.88	0.64	
TGTTTAG	0.92	0.00	<u>MCM1</u> , <u>ROX1</u> , <u>MIG1</u> , <u>MAL63</u>
TGTTTAT	0.92	0.76	STE12, PRE URSSGA
TGTTTCA	1.00	1.00	
TGTTTCC	1.00	0.00	
TGTTTGT	0.92	0.84	
TGTTTIA	1.00	0.80	<u>MCM1</u> , <u>PHO2</u> , <u>SWI5</u>
TGTTTIT	0.92	0.00	
TTTTTTA	0.92	0.00	
# 2 (10 words)			
TAGTTAT	0.88	0.00	URS1ERG11 URS1ERG11, TAF
TCGTTTA	0.92	0.32	
TCGTTIT	0.92	0.00	<u>ROX1</u> <u>SFE</u> , MOT3 <u>ROX1</u> , <u>SFE</u>
TTGTTAT	0.92	0.92	
TTGTTCT	0.92	0.00	<u>ABF1</u> , <u>BAF1</u> <u>ROX1</u>
TTGTTGA	1.00	0.80	
TTGTTTA	1.00	1.00	0.00 0.76
TTGTTTC	1.00	0.76	
TTGTTTG	1.00	0.76	1.00 0.76
TTGTTTT	1.00	1.00	
# 3 (9 words)			
CCTGGTT	0.96	0.00	0.00 0.76
GATGGTT	0.96	0.76	

Word pattern	% select SSS	% select SSS ($\lambda_2 = 0$)	Transcription factors
GCTGGAT	1.00	0.00	<u>ACE2</u> <u>ACE2, SWI5, PHO2, SWI5</u>
GCTGGCT	1.00	0.76	
GCTGGGC	1.00	0.04	
GCTGGGT	1.00	0.92	
GCTGGTG	0.96	0.00	
GCTGGTT	1.00	1.00	<u>MCB</u> <u>MCB</u> <u>MCB</u> PHO4
GTTGGGC	0.92	0.64	
<u># 4 (6 words)</u>			
CACGCGT	0.88	0.00	
GACGCGT	0.88	0.40	
TACGCAT	0.96	0.04	<u>ACE2</u> CUP2 <u>ACE2, SWI5</u> <u>ROX1</u> BAS1
TACGCGT	0.88	0.40	
TTCGCAT	0.92	0.28	
TTCGCGT	0.96	0.12	
<u># 5 (6 words)</u>			
GGCTGGG	1.00	0.72	<u>ACE2</u> CUP2 <u>ACE2, SWI5</u> <u>ROX1</u> BAS1
GGCTGGT	0.92	0.00	
GGCTGTG	1.00	0.88	
TGCTGGC	1.00	0.92	
TGCTGGT	1.00	0.92	
TTCCTGT	0.92	0.04	UASH, CUP2, ACE1, UASPHR
<u># 6 (5 words)</u>			
CGTTGTT	0.96	0.00	
TCTTGTT	0.96	0.00	
TGTTGTT	0.96	0.92	
TTTGTCT	0.88	0.76	<u>SWI5</u>
TTTGTGT	0.96	0.00	
<u># 7 (4 words)</u>			
CTGCTGG	0.92	0.00	
TAGCTGG	0.96	0.00	
TGGCTGG	0.96	0.20	

Word pattern	% select SSS	% select SSS ($\lambda_2 = 0$)	Transcription factors
TTGCTGG # 8 (4 words)	0.96	0.92	<u>SWI5</u> , ACE1, CUP2
CTGGCTT	0.88	0.04	
ATGGCTT	0.88	0.56	URS1ERG11
GTGGCTT	0.88	0.04	
TTGGCTT	0.88	0.00	PUT3
# 9 (3 words)			
ACGCGGT	0.96	0.00	
ACGCGTT	1.00	0.96	<u>MCB</u>
TCGCGTT	1.00	0.00	
# 10 (3 words)			
GAGATCT	0.96	0.80	
GAGATGT	0.92	0.04	
GGGATCT	0.92	0.00	
# 11 (3 words)			
GCGGTCT	0.92	0.60	<u>RAP1</u> , <u>GAL4</u>
GGGGTCT	0.92	0.00	BUF
TCGGTCT	0.92	0.00	
# 12 (3 words)			
GTTTCAG	1.00	0.00	
GTTTCAT	1.00	0.76	<u>STE12</u> , UASPHR
GTTTCCT	0.92	0.00	URSSGA
# 13 (3 words)			
TCAGAAC	0.88	0.00	
TGAGAAC	1.00	1.00	
TGAGATC	1.00	0.88	
# 14 (3 words)			
TCCTACT	0.88	0.0	
TTCTACG	0.88	0.12	
TTCTACT	0.88	0.52	
# 15 (3 words)			

Word pattern	% select SSS	% select SSS ($\lambda_2 = 0$)	Transcription factors
TGTGCGA	0.96	0.00	UASPHR
TGTGCGT	0.96	0.76	
TTTGCGT	0.96	0.00	
<u># 16 (3 words)</u>			
TTTCCGA	0.92	0.84	<u>MCM1</u>
TTTCCGT	0.92	0.00	<u>MCM1</u> , IRE
TTTCCTA	1.00	0.76	<u>MCM1</u>
<u># 17 (3 words)</u>			
TTTGCTA	0.88	0.56	<u>CuRE;MAC1</u> ACE1, CUP2, UASH
TTTGCTC	0.88	0.00	
TTTGCTG	0.88	0.04	