BIOMETRICS ??, 1–22 DOI: x.x/j.x-x.x.x
January 2022

Sparse Estimation in Semi-parametric Finite Mixture of Varying Coefficient Regression Models

Abbas Khalili[1,*], Farhad Shokoohi[2,], Masoud Asgharian[3,***], and Shili Lin[4,****]**

[1,3]Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada

[2]Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA

[4]Department of Statistics, Ohio State University, Columbus, OH 43210, USA

*_email:_ abbas.khalili@mcgill.ca

**_email:_ farhad.shokoohi@unlv.edu

***_email:_ masoud.asgharian2@mcgill.ca

****_email:_ shili@stat.osu.edu

SUMMARY: Finite mixture of regressions (FMR) are commonly used to model heterogeneous effects of covariates on a response variable in settings where there are unknown underlying subpopulations. FMRs, however, cannot accommodate situations where covariates' effects also vary according to an "index" variable–known as finite mixture of varying coefficient regression (FM-VCR). Although complex, this situation occurs in real data applications: the osteocalcin (OCN) data analyzed in this manuscript presents a heterogeneous relationship where the effect of a genetic variant on OCN in each hidden subpopulation varies over time. Oftentimes, the number of covariates with varying coefficients also presents a challenge: in the OCN study, genetic variants on the same chromosome are considered jointly. The relative proportions of hidden subpopulations may also change over time. Nevertheless, existing methods cannot provide suitable solutions for accommodating all these features in real data applications. To fill this gap, we develop statistical methodologies based on regularized local-kernel likelihood for simultaneous parameter estimation and variable selection in sparse FM-VCR models. We study large-sample properties of the proposed methods. We then carry out a simulation study to evaluate the performance of various penalties adopted for our regularized approach and ascertain the ability of a BIC-type criterion for estimating the number of subpopulations. Finally, we applied the FM-VCR model to analyze the OCN data and identified several covariates, including genetic variants, that have age-dependent effects on osteocalcin.

KEYWORDS: finite mixture of regressions, local-kernel likelihood, non-parametric models, penalized likelihood

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Finite mixture of regression (FMR) models (McLachlan and Peel, 2000) are commonly used to accommodate heterogeneous effects of covariates $\boldsymbol{X} = (X_1, \ldots, X_d)$ on a response variable Y when the population under study is believed to consist of multiple hidden subpopulations. While FMRs can successfully capture such heterogeneity, they fall short if the effects of X_j's on Y also vary as functions of an index variable, U, such as time or location. The mixing proportions, representing the sizes of hidden subpopulations, may also change as functions of U. It becomes even more challenging when only a few covariates among a large set have significant effects on Y. The data from the osteocalcin (OCN) study under our consideration presents such challenges on all fronts: there are many genetic and non-genetic covariates (X_j's); there may be heterogeneous relationships, also varying over age (U), between the response variable (OCN, Y) and only a subset of the covariates; and the relative proportions of hidden subpopulations may also change over age.

An extension of FMRs, semi-parametric finite mixture of varying coefficient regressions (FM-VCR), was further introduced to account for heterogeneous varying covariates' effects (Xiang et al., 2019). These models facilitate the use of varying coefficient regressions, as functions of U, in studying the relationship between Y and X_j's in a heterogeneous population. The functional forms of varying coefficients are seldom known and left unspecified. McLachlan and Peel (2000) and Xiang et al. (2019), discuss statistical inference and applications of various special cases of FM-VCRs in applied sciences and machine learning.

In the OCN study, the investigators collected data on environmental and genetic factors to study whether and how osteocalcin (the phenotype) is affected by a subset of these factors (Liao et al., 2014). Several genetic variants in Chromosome 7 (Chr7) were implicated (Zhang, 2017), and the gene harboring these variants was in fact linked to bone morphogenetic protein (BMP) (Harada et al., 2003). Population heterogeneity has not, however, been considered in

these studies. Inspired by recent studies demonstrating that relationships between OCN and some underlying factors are not the same across the population (Liu et al., 2015), we are interested in exploring whether there exist subpopulations, where the relationships between OCN and the factors of interest, especially the single-nucleotide polymorphisms (SNPs) in Chr7, are different across the subpopulations. Further, genetic effects on a phenotype may change over time (Zhang, 2017), which may well be the case for genes related to BMP. Finally, it is known that aging leads to bone density loss both in men and women (Demontiero et al., 2012); the mixing proportions of the components may well vary with age. Thus, a reanalysis of the data to accommodate population heterogeneity, varying subpopulation relative proportions, and varying genetic effects over time is warranted.

In this paper, motivated by the challenges in analyzing the OCN data, we study methodologies for sparse estimation in FM-VCR models. To the best of our knowledge, the existing literature (Xiang et al., 2019) mainly focuses on maximum likelihood estimation (MLE), while variable selection problems are largely understudied. In FM-VCRs, even with $d = 43$ in the OCN study, the dimension of the parameter space is large enough to render classical methods for variable selection almost impractical. Thus, we develop new results based on regularized local-kernel likelihood methods, demonstrating that regularized estimation in FM-VCRs consistently estimates the model parameters while recovering its sparse structure.

Since the seminal work of Hastie and Tibshirani (1993), varying coefficient regressions (VCR) and their extensions to mixture models (Xiang et al., 2019) have attracted much attention in the statistics literature. The parameters in these models can change as smooth non-parametric functions of an index variable such as time or location, which results in reducing modeling bias while avoiding the curse of dimensionality (Fan and Zhang, 1999). Due to the non-parametric nature of the models, parameter estimation requires careful consideration. Several estimation techniques are available in the literature, ranging from

the local kernel (Huang et al., 2018), to splines and basis functions (Hastie and Tibshirani, 1993) and local polynomial approximations (Fan and Gigbels, 1996). Another line of research particularly in VCRs has focused on variable selection, thanks to the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive LASSO (AdpLASSO) (Zou, 2006), MCP (Zhang, 2010), and group LASSO (Yuan and Lin, 2006). Among others, Wang and Xia (2009) studied variable selection using group LASSO/AdpLASSO and basis expansion or local-kernel methods. Wei et al. (2011) and others studied the problem in high dimensions using direct penalization or two-stage screening and penalization.

Despite the surge of research on variable selection in VCRs, the problem in FM-VCRs under our consideration has not been studied. The works reviewed by Xiang et al. (2019) focus on MLE in FM-VCRs and their special cases. One could use best-subset selection methods such as the AIC, BIC, or their variations to perform variable selection based on the MLE. Such techniques, however, require intensive computations as there are potentially $2^{C\times d} > 10^6$ submodels to be examined for selecting a sparse 2-components FM-VCR even with only $d = 10$ covariates. Variable selection based on regularization techniques such as LASSO is mainly studied in FMRs (Khalili and Chen, 2007; Städler et al., 2010; Shokoohi et al., 2019), without allowing for varying coefficients. The challenges of sparse estimation in FM-VCRs are the non-parametric nature of varying regression coefficients, mixing probabilities, and dispersion parameters, and often many covariates where only a handful are significant in the model.

In this paper, we develop computationally efficient penalized local-kernel likelihood methods for sparse estimation in FM-VCRs. We establish consistency in estimation and variable selection and oracle properties of the proposed estimators. We develop a modified EM algorithm (Dempster et al., 1977) for the numerical implementation of the methods. We evaluate the finite-sample performance of the methods via simulation and analyze the OCN data.

2. Sparse FM-VCR models

Let Y be a real-valued response variable, $\boldsymbol{X}$ be a d-dimensional vector of covariates, and U be an index variable. In a population with $C > 1$ hidden subpopulations, we are interested in the conditional distribution of $Y|(\boldsymbol{X} = \boldsymbol{x}, U = u)$, for $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ and $u \in \mathcal{U} \subset \mathbb{R}$.

Let $\mathcal{C}$ be a discrete latent variable taking values $1, \ldots, C$. For any $\boldsymbol{x} \in \mathcal{X}, u \in \mathcal{U}$, we denote

$$0 < \pi_j(u) = \Pr(\mathcal{C} = j|\boldsymbol{X} = \boldsymbol{x}, U = u) \ , \quad j = 1, \ldots, C, \tag{1}$$

as non-parametric functions of u, and $\sum_{j=1}^{C} \pi_j(u) = 1$. This setting differs from the mixture-of-experts (ME) and their variations (McLachlan and Peel, 2000) including finite mixtures with concomitant variables (Dayton and Macready, 1988; Frühwirth-Schnatter et al., 2018), wherein each π_j is modeled as a parametric function of u and possibly other covariates $\boldsymbol{x}$.

Suppose the conditional density (mass) function of $Y|(\boldsymbol{X} = \boldsymbol{x}, U = u, \mathcal{C} = j)$ is given by

$$f\big(y; \theta_j(\boldsymbol{x}, u), \phi_j(u)\big), \quad \text{for any} \quad y \in \mathcal{Y} \subset \mathbb{R}, \tag{2}$$

where f belongs to a known parametric family $\mathcal{G} = \big\{f(y; \theta, \phi) : (\theta, \phi) \in \Theta \times (0, \infty), \Theta \subset \mathbb{R}\big\}$ with respect to a σ-finite measure ν, and $\phi > 0$ is a dispersion parameter. We assume $\theta_j(\boldsymbol{x}, u) = g(\boldsymbol{x}^\top \boldsymbol{\beta}_j(u))$, for a known link function g and the regression coefficients $\boldsymbol{\beta}_j^\top(u) = \big(\beta_{j1}(u), \ldots, \beta_{jd}(u)\big)$, where $(\phi_j(u), \beta_{jl}(u))$ are also unknown non-parametric functions of u. A well-known example of f in (2) is the Gaussian density with the mean $\theta_j(\boldsymbol{x}, u) = E\{Y|(\boldsymbol{x}, u, j)\} = \boldsymbol{x}^\top \boldsymbol{\beta}_j(u)$ which corresponds to the identity link function $g(\eta) = \eta$, and variance $\mathrm{Var}\{Y|(\boldsymbol{x}, u, j)\} = \phi_j(u)$; see Web Appendix F for more details. Also note that each $\boldsymbol{\beta}_j(u)$ and $\phi_j(u)$ is a function of u rather than a constant as in the ME models.

By putting (1) and (2) together, the conditional density (mass) function of $Y|(\boldsymbol{X} = \boldsymbol{x}, U = u)$ in an FM-VCR model with order C is given by

$$f_C^*\big(y|\boldsymbol{\psi}(u), \boldsymbol{x}\big) = \sum\nolimits_{j=1}^{C} \pi_j(u) f\big(y; \theta_j(\boldsymbol{x}, u), \phi_j(u)\big), \ y \in \mathcal{Y} \tag{3}$$

where $\boldsymbol{\psi}(u) = \big(\boldsymbol{\pi}^\top(u), \boldsymbol{\phi}^\top(u), \boldsymbol{\beta}_1^\top(u), \ldots, \boldsymbol{\beta}_C^\top(u)\big)^\top \in \mathbb{R}^p$, with $p = Cd + 2C - 1$, and the sub-vectors $\boldsymbol{\phi}(u) = \big(\phi_1(u), \ldots, \phi_C(u)\big)^\top \in \mathbb{R}^C$, $\boldsymbol{\pi}(u) = \big(\pi_1(u), \ldots, \pi_{C-1}(u)\big)^\top \in \mathbb{R}^{(C-1)}$.

Identifiability is essential in mixture models; if for any $C_1 > 1, C_2 > 1$ and given $(u, \boldsymbol{x})$, we have $f^*_{C_1}\big(y|\boldsymbol{\psi}_1(u), \boldsymbol{x}\big) = f^*_{C_2}\big(y|\boldsymbol{\psi}_2(u), \boldsymbol{x}\big)$, for all $y \in \mathcal{Y}$, then we must have $C_1 = C_2$ and $\boldsymbol{\psi}_1(u) = \boldsymbol{\psi}_2(u)$ (up to a mixture component permutation). Sufficient conditions for identifiability of (3) are discussed in Theorem 1 of Huang et al. (2018), and the conditions are given in Web Appendix A. We assume that the models considered here are identifiable.

In an FM-VCR, even for moderate values of C and d, the number $p = Cd + (2C - 1)$ of non-parametric functions to be estimated becomes very large compared to a typical sample size n. In our real data application for Chr7, after an initial covariate screening, we have 36 SNPs plus 7 extra covariates ($d = 43$), with $1 \leqslant C \leqslant 5$ resulting in potentially $44 \leqslant p \leqslant 224$ non-parametric functions to be estimated based on a sample size of $n = 1704$. However, not all the covariates have significant effects on the response variable of interest, suggesting a sparse model for the data. Thus, we assume that the true FM-VCR model is *sparse*, that is, for every $j = 1, \ldots, C$, there exists an integer $1 \leqslant d^0_j < d$ such that

$$0 < E\big[\big\{\beta^0_{jl}(U)\big\}^2\big] < \infty\ , \quad \text{for all } l = 1, \ldots, d^0_j, \text{ and} \tag{4}$$

$$E\big[\big\{\beta^0_{jl}(U)\big\}^2\big] = 0\ , \quad \text{for all } l = d^0_j + 1, \ldots, d, \tag{5}$$

where E is with respect to U, and d^0_j is the number of non-zero $\beta^0_{jl}(\cdot)$'s in component j. By (4)-(5), the total number of non-parametric functions is $p_0 = \sum_{j=1}^{C} d^0_j + (2C - 1)$ and presumably much smaller than p. More discussion on Condition (5) is given in Remark 1 of Web Appendix E. Next, we discuss estimation and variable selection in sparse FM-VCRs.

3. Simultaneous estimation and variable selection

Let $(u_i, \boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, be the observations based on a random sample from the FM-VCR model (3). The (conditional) log-likelihood function is given by

$$\mathcal{L} = \sum\nolimits_{i=1}^{n} \log\{f^*_C(y_i|\boldsymbol{\psi}(u_i), \boldsymbol{x}_i)\} = \sum\nolimits_{i=1}^{n} \log\Big\{\sum\nolimits_{j=1}^{C} \pi_j(u_i) f\big(y_i; \theta_j(\boldsymbol{x}_i, u_i), \phi_j(u_i)\big)\Big\}, \tag{6}$$

where $\theta_j(\boldsymbol{x}_i, u_i) = g(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j(u_i))$. Since $(\pi_j(u), \phi_j(u), \boldsymbol{\beta}_j(u))$ are unknown non-parametric functions, the parameter space has infinite dimension and $\mathcal{L}$ is thus intractable. Several

techniques are available in the literature (Fan and Gigbels, 1996) for non-parametric function estimation, of which the most popular are local-kernel methods, splines and local polynomial approximations. Silverman (1984) shows that spline smoothing corresponds approximately to smoothing by a kernel method with bandwidth depending on the local density of design points. In the following, we use the local-kernel method where each functional parameter in the FM-VCR model is locally approximated by constants. As a result, the locally approximated model becomes a standard finite mixture of regression which in turn makes implementation of our penalization method via the EM algorithm easier (Huang et al., 2013).

Local log-likelihood. For any $u \in \mathcal{U}$, we consider the local-kernel log-likelihood

$$\ell_n\big(\boldsymbol{\psi}(u); h\big) = \sum\nolimits_{i=1}^{n} \log\Big\{ \sum\nolimits_{j=1}^{C} \pi_j f(y_i; \theta_j(\boldsymbol{x}_i), \phi_j)\Big\} K_h(u_i - u), \tag{7}$$

where $\theta_j(\boldsymbol{x}_i) = g(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$, and $K_h(t) = h^{-1}K(t/h)$ is a kernel function with a band-width h. The locally constant vector of parameters is $\boldsymbol{\psi}(u) = (\boldsymbol{\pi}^\top, \boldsymbol{\phi}^\top, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_C^\top) \in \mathbb{R}^p$, where $\boldsymbol{\pi}^\top = (\pi_1, \ldots, \pi_{C-1})$, $\boldsymbol{\phi}^\top = (\phi_1, \ldots, \phi_C)$, and $\boldsymbol{\beta}_j^\top = (\beta_{j1}, \ldots, \beta_{jd})$, $j = 1, \ldots, C$. The entries of $\boldsymbol{\psi}(u)$ are local-constant approximations of the functions $(\pi_j(u), \phi_j(u), \beta_{jl}(u))$, which depend on u, and for simplicity, we suppress u in the notation but keep it for $\boldsymbol{\psi}(u)$.

For any $u \in \mathcal{U}$, the maximum local-kernel log-likelihood estimator (MLLE) of $\boldsymbol{\psi}(u)$ is defined as the maximizer of $\ell_n\big(\boldsymbol{\psi}(u); h\big)$, and is denoted by $\breve{\boldsymbol{\psi}}_n(u)$. In practice, the estimation is usually done at the observed points $u_1, \ldots, u_n$; more discussion is given at the end of this section. In Proposition 1 of Web Appendix D, we establish estimation consistency of $\breve{\boldsymbol{\psi}}_n(u)$. It is, however, well-known that similar to the MLE in parametric regression, $\breve{\boldsymbol{\psi}}_n(u)$ does not provide a sparse FM-VCR model as postulated in (5), which is the main focus of our research. Therefore, we propose a regularization method that yields fitted sparse FM-VCR models.

Penalized local log-likelihood. For $j = 1, \ldots, C$ and $t = 1, \ldots, n$, let $\boldsymbol{\beta}_{j,t} = (\beta_{j1,t}, \ldots, \beta_{jd,t})^\top$ be the local constant approximation of the vector $\boldsymbol{\beta}_j(u)$ at u_t's. We denote the $(n \times d)$ matrices $\boldsymbol{B}_j = (\boldsymbol{\beta}_{j,1}, \ldots, \boldsymbol{\beta}_{j,n})^\top$. Similarly, let $\boldsymbol{\pi}_j^\top = (\pi_{j,1}, \ldots, \pi_{j,n})$ and $\boldsymbol{\phi}_j^\top = (\phi_{j,1}, \ldots, \phi_{j,n})$ be the

local constant approximations of $\pi_j(u)$ and $\phi_j(u)$ at u_t's, respectively. Also, denote the $(n \times p)$ matrix $\boldsymbol{\Psi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{C-1}, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_C, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_C)$, where $p = Cd + 2C - 1$.

Using (7), the corresponding (total) local-kernel log-likelihood is given by

$$L_n(\boldsymbol{\Psi}; h) = \sum_{t=1}^{n} \ell_n\big(\boldsymbol{\psi}(u_t); h\big) = \sum_{t,i=1}^{n} \log\Big\{\sum\nolimits_{j=1}^{C} \pi_{j,t}\; f\big(y_i; \theta_{j,t}(\boldsymbol{x}_i), \phi_{j,t}\big)\Big\}\, K_h(u_i - u_t), \tag{8}$$

and $\theta_{j,t}(\boldsymbol{x}_i) = g(\boldsymbol{x}_i^\top \boldsymbol{\beta}_{j,t})$. We estimate $\boldsymbol{\Psi}$ by maximizing the penalized local log-likelihood

$$\widetilde{L}_n(\boldsymbol{\Psi}; \boldsymbol{\lambda}, h) = L_n(\boldsymbol{\Psi}; h) - \mathbb{P}_n(\boldsymbol{\Psi}; \boldsymbol{\lambda}), \tag{9}$$

$$\mathbb{P}_n(\boldsymbol{\Psi}; \boldsymbol{\lambda}) = \sum\nolimits_{j=1}^{C} \sum\nolimits_{l=1}^{d} p_n(\|\boldsymbol{b}_{jl}\|/\sqrt{n}; \lambda_j), \tag{10}$$

where p_n is a penalty function with tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_C)^\top$ that control the level of penalization on $\|\boldsymbol{b}_{jl}\|/\sqrt{n}$. Here, $\boldsymbol{b}_{jl} = \big(\beta_{jl,1}, \ldots, \beta_{jl,n}\big)^\top = \big(\beta_{jl}(u_1), \ldots, \beta_{jl}(u_n)\big)^\top$ is the l^{th} column of the matrix $\boldsymbol{B}_j$, $j = 1, \ldots, C$, and $\|\boldsymbol{b}_{jl}\|^2 = \sum_{t=1}^{n} \beta_{jl}^2(u_t)$. Examples of p_n are the LASSO, AdpLASSO, SCAD, and MCP (Web Appendix C). The group penalization in (10) (Yuan and Lin, 2006) enforces zero estimates of some of $\|\boldsymbol{b}_{jl}\|$ and thus the vectors $\boldsymbol{b}_{jl}$. By Condition (RC.2) on the density $m(u)$ of U, $u_1, \ldots, u_n$ are dense in the support $\mathcal{U}$ (Janson, 1987), and thus it suffices to perform estimation at only u_i's rather than on the entire $\mathcal{U}$.

Given $(\boldsymbol{\lambda}, h)$, the maximum penalized local log-likelihood estimate (MPLLE) of $\boldsymbol{\Psi}$ is

$$\widehat{\boldsymbol{\Psi}}_n(\boldsymbol{\lambda}, h) \equiv \widehat{\boldsymbol{\Psi}}_n = \operatorname*{argmax}_{\boldsymbol{\Psi}} \widetilde{L}_n(\boldsymbol{\Psi}; \boldsymbol{\lambda}, h). \tag{11}$$

We induce zero estimates for some of the regression coefficients $\beta_{jl}(\cdot)$ by appropriate choices of the penalty p_n, and the tuning parameters $(\boldsymbol{\lambda}, h)$. Thus, the MPLLE performs simultaneous estimation and variable selection, resulting in a fitted sparse FM-VCR model.

4. Large-sample theory

To distinguish from earlier notation, while drawing the connection, for any $u \in \mathcal{U}$, let $\boldsymbol{\psi}^0(u)$ be the p-dimensional parameter vector in (3) corresponding to the true sparse FM-VCR model satisfying (4)-(5). In particular, for the observed points $u_1, \ldots, u_n$, the $(n \times p)$-dimensional parameter matrix in (8) corresponding to the true model is denoted by $\boldsymbol{\Psi}_n^0$ with

its t^{th} row as $[\boldsymbol{\psi}^0(u_t)]^\top, t = 1, \ldots, n$. Without loss of generality, we assume the partitioning $\boldsymbol{\psi}^0(u) = \big(\boldsymbol{\psi}_1^0(u), \boldsymbol{\psi}_2^0(u)\big)$ such that $\boldsymbol{\psi}_2^0(u)$ contains only those $\beta_{jl}^0(\cdot)$'s that satisfy the sparsity Condition (5), and $\boldsymbol{\psi}_1^0(u)$ consists of the non-zero functions $\beta_{jl}^0(\cdot)$ among other parameters $(\pi_j^0(u), \phi_j^0(u))$. By (5), we have $E\{\|\boldsymbol{\psi}_2^0(U)\|^2\} = 0$. Thus, Conditions (C1) and (RC.2) on $\beta_{jl}(\cdot)$'s and the density $m(u)$ of U (Web Appendix A and Web Appendix B) imply that $\boldsymbol{\psi}_2^0(u) = \mathbf{0}$, uniformly in $u \in \mathcal{U}$. This results in $\boldsymbol{\Psi}_n^0 = (\boldsymbol{\Psi}_{n1}^0, \boldsymbol{\Psi}_{n2}^0)$ such that $\boldsymbol{\Psi}_{n2}^0$ contains all the zero regression functions, and $\boldsymbol{\Psi}_{n1}^0 = \{[\boldsymbol{\psi}_1^0(u_t)]^\top : t = 1, \ldots, n\}$ contains the nonzero regression functions and other parameters. Denote $l(\boldsymbol{\psi}^0(u), \boldsymbol{x}, y) = \log\{f_C^*(y|\boldsymbol{\psi}^0(u), \boldsymbol{x})\}$, $y \in \mathcal{Y}$, $\boldsymbol{x} \in \mathcal{X}$. For $t = 1, \ldots, n$, define $l'(\boldsymbol{\psi}^0(u_t), \boldsymbol{x}, y)$ and $l''(\boldsymbol{\psi}^0(u_t), \boldsymbol{x}, y)$, respectively, as the gradient and Hessian of $l(\boldsymbol{\psi}, \boldsymbol{x}, y)$ with respect to $\boldsymbol{\psi}$ and evaluated at $\boldsymbol{\psi} = \boldsymbol{\psi}^0(u_t)$. Also, let

$$\boldsymbol{I}(u_t) = E_{(Y,\boldsymbol{X})|U=u_t}\left[l'(\boldsymbol{\psi}^0(U), \boldsymbol{X}, Y) l'^\top(\boldsymbol{\psi}^0(U), \boldsymbol{X}, Y) | U = u_t\right] \tag{12}$$

$$\Delta(u; u_t) = \int_{\mathcal{Y}} \int_{\mathcal{X}} l'(\boldsymbol{\psi}^0(u_t), \boldsymbol{x}, y) f_C^*(y|\boldsymbol{\psi}^0(u), \boldsymbol{x}) g(\boldsymbol{x}|u) d\boldsymbol{x} dy, \tag{13}$$

and $E_{(\boldsymbol{X},Y)|U=u_t}$ is the expectation with respect to the distribution of $(\mathbf{X}, Y)$ given $U = u_t$.

With regard to the penalty function p_n, we denote the quantity

$$r_{1n} = \max\left\{h_n^{1/2}|p_n'(\theta_{jl}^0; \lambda_n)|/n^{3/2} : \theta_{jl}^0 = \sqrt{E_0\{\beta_{jl}^0(U)\}^2} \neq 0, 1 \leqslant j \leqslant C, 1 \leqslant l \leqslant d_j^0\right\} \tag{14}$$

where $p_n'(\theta_{jl}^0; \lambda_n)$ is the first derivative of $p_n(\theta_{jl}; \lambda_n)$ evaluated at $\theta_{jl} = \theta_{jl}^0 \neq 0$. Also, denote

$$\mathcal{V}_0 = \int_{\mathcal{U}} K^2(t) dt \ , \ \mathcal{K}_2 = \int_{\mathcal{U}} t^2 K(t) dt. \tag{15}$$

Regularity Conditions (RC.1)-(RC.5) on the mixture density f_C^*, the conditional density $g(\boldsymbol{x}|u)$ of $\boldsymbol{X}|U = u$, the density $m(u)$ of U, and Conditions (KC.1)-(KC.2) and (PC.1)-(PC.3) on the kernel K and the penalty p_n, respectively, are all given in Web Appendix B. The proofs of the following lemma and theorems are given in Web Appendix D.

THEOREM 1: *(Point-wise estimation consistency) Suppose that Conditions (RC.1)-(RC.4) hold, the kernel K satisfies Conditions (KC.1)-(KC.2), and (λ_n, p_n, h_n) satisfy Conditions (PC.1)-(PC.2). Then, there exists a local maximizer $\widehat{\boldsymbol{\Psi}}_n$ of $\widetilde{L}_n$ in (9) that $n^{-1}\big\|\widehat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}_n^0\big\|_F^2 =$*

$n^{-1}\sum_{t=1}^{n}\big\|\widehat{\boldsymbol{\psi}}_n(U_t)-\boldsymbol{\psi}^0(U_t)\big\|^2=O_p\big\{(1+r_{1n})^2(nh_n)^{-1}\big\}$, *where* r_{1n} *is given in* (14), $\widehat{\boldsymbol{\psi}}_n(U_t)$ *and* $\boldsymbol{\psi}^0(U_t)$ *are the* t^{th} *rows of the matrices* $\widehat{\boldsymbol{\Psi}}_n$ *and* $\boldsymbol{\Psi}_n^0$, *and* $\|\cdot\|_F$ *is the Frobenius norm.*

By Theorem 1, if $r_{1n}=O(1)$, $\widehat{\boldsymbol{\Psi}}_n$ achieves the point-wise consistency rate $\{nh_n\}^{-1/2}$ in estimating $\boldsymbol{\Psi}_n^0$, a property shared by the MLLE (Proposition 1, Web Appendix D). This result clearly depends on the choice of (λ_n,p_n,h_n). For the LASSO, $\lambda_n=O(\{nh_n\}^{-1/2})$ suffices. For SCAD and MCP, as long as $(\lambda_n,h_n)\to 0$ as $n\to\infty$, the desired rate is achieved since $r_{1n}=0$. For AdpLASSO, with the (possibly random) weights w_{jl}'s, we require $\lambda_n\max_{1\leqslant j\leqslant C,1\leqslant l\leqslant d_j^0}w_{jl}=o_p(\{nh_n\}^{-1/2})$. The practical choice of w_{jl} is given in Remark 3 of Web Appendix E. Theorem 1, however, does not imply the sparsity of $\widehat{\boldsymbol{\Psi}}_n$. We strengthen this result by establishing the sparsity and oracle property of the MPLLE, beginning with a lemma.

LEMMA 1: *Assume that the conditions of Theorem 1 are met, and that* (λ_n,p_n,h_n) *also satisfy Condition (PC.3). Then, for any* $\sqrt{nh_n}$*-consistent* MPLLE $\widehat{\boldsymbol{\Psi}}_n$ *of* $\boldsymbol{\Psi}_n^0$, *as* $n\to\infty$, $P\big(\big\{\sum_{t=1}^{n}\widehat{\beta}_{jl}^2(U_t)/n\big\}^{1/2}=\big\|\widehat{\boldsymbol{b}}_{jl}/n\big\|=0\big)\to 1\ ,\ j=1,\ldots,C\ ,\ l=d_{j+1}^0,\ldots,d.$

By Lemma 1, with probability tending to one, the MPLLE of those $\beta_{jl}(u)$ that satisfy the sparsity property (5) are zero at the observed index variable: $\widehat{\beta}_{jl}(U_t)=0, t=1,\ldots,n$ (point-wise sparsity). It is more appealing to establish a uniform sparsity. Recall the partitioning $\boldsymbol{\psi}^0(u)=\big(\boldsymbol{\psi}_1^0(u),\boldsymbol{\psi}_2^0(u)\big)$, where $\boldsymbol{\psi}_2^0(u)=\mathbf{0}$ uniformly in $u\in\mathcal{U}$. We denote the oracle estimator $\widehat{\boldsymbol{\psi}}_{n1,orc}(u)$ as the MLLE of $\boldsymbol{\psi}_1^0(u)$, having known $\boldsymbol{\psi}_2^0(u)=\mathbf{0}$ a priori. For any $u\in\mathcal{U}$, let $\widehat{\boldsymbol{\psi}}_n(u)=\big(\widehat{\boldsymbol{\psi}}_{n1}(u),\widehat{\boldsymbol{\psi}}_{n2}(u)\big)$ be the MPLLE of $\boldsymbol{\psi}^0(u)$, where its partitioning corresponds to that of $\boldsymbol{\psi}^0(u)$. Next, we describe the behavior of $\widehat{\boldsymbol{\psi}}_n(u)$ with respect to the oracle estimator.

THEOREM 2: *Under the conditions of Lemma 1, and Condition (RC.5), as* $n\to\infty$, *(i) uniform sparsity:* $P\big\{\sup_{u\in\mathcal{U}}\big\|\widehat{\boldsymbol{\psi}}_{n2}(u)\big\|=0\big\}\to 1$; *(ii) oracle property:* $\sup_{u\in\mathcal{U}}\big\|\widehat{\boldsymbol{\psi}}_{n1}(u)-\widehat{\boldsymbol{\psi}}_{n1,orc}(u)\big\|=o_p\big\{(\log h_n^{-1}/nh_n)^{1/2}\big\}$; *(iii) asymptotic normality: for any* $u\in\mathcal{U}$,

$$\sqrt{nh_n}\big\{\big[\boldsymbol{I}(u)+\mathbb{P}_n''(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n)/n^2m(u)\big]\big(\widehat{\boldsymbol{\psi}}_{n1}(u)-\boldsymbol{\psi}_1^0(u)\big)-\mathcal{B}_n(u)\big\}\to N\big(\mathbf{0},\mathcal{V}_0\boldsymbol{I}(u)/m(u)\big),$$

where $\boldsymbol{I}(u)$ *in* (12) *is positive definite, the bias* $\mathcal{B}_n(u) = \frac{1}{2}\mathcal{K}_2\big\{\Delta''(u;u)+2\Delta'(u;u)m'(u)/m(u)\big\}h_n^2 - \mathbb{P}'_n(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n)/n^2m(u) + o_p(h_n^2)$, *and* $\boldsymbol{\theta}_1^0 = \big\{\theta_{jl}^0 = (E_0\{\beta_{jl}^0(U)\}^2)^{1/2} : 1 \leqslant j \leqslant C, 1 \leqslant l \leqslant d_j^0\big\}$.

Theorem 2(i) implies that the MPLLE of those $\beta_{jl}(\cdot)$'s that satisfy the sparsity property (5) are zero, that is $\widehat{\beta}_{jl}(u) = 0$ uniformly in $u \in \mathcal{U}$, for all $j = 1,\ldots,C$ and $l = d_{j+1}^0,\ldots,d$. Condition (PC.3) on (λ_n, p_n, h_n) guarantees sparsity of the MPLLE. For the LASSO, SCAD, and MCP, this condition requires $\sqrt{nh_n}\lambda_n \to \infty$, and for the AdpLASSO we need $\sqrt{nh_n}\lambda_n \min_{1\leqslant j\leqslant K, d_j^0+1\leqslant l\leqslant d} w_{jl} \to \infty$, as $n \to \infty$. Theorem 2(ii) and (iii) imply that the MPLLE asymptotically attains a similar information bound as the oracle estimator described above (Huang et al., 2018). Regarding (iii), for all the four penalties, we have $\mathbb{P}''_n(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n) = \mathbf{0}$. On the other hand, with respect to the bias $\mathcal{B}_n(u)$, for the SCAD and MCP, we have $\mathbb{P}'_n(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n)/n^2 = \mathbf{0}$, and for the AdpLASSO, $\mathbb{P}'_n(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n)/n^2 \sim \lambda_n \max_{1\leqslant j\leqslant C, 1\leqslant l\leqslant d_j^0} w_{jl}$, which tends to zero when scaled by $\sqrt{nh_n}$, as $n \to \infty$. For the LASSO, $\mathbb{P}'_n(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n)/n^2 = \lambda_n$, which tends to ∞ once scaled by $\sqrt{nh_n}$, as required by Condition (PC.3) for sparsity. This behavior is also known in parametric regression, as it gives sparse estimators but does not achieve the oracle property. Finally, the asymptotic normality in (iii) is obtained from the oracle perspective that the true sparse structure of the model is known in advance. In practice, a sparse model is typically fitted based on the MPLLE. As such there are two sources of uncertainty that are caused by variable selection and parameter estimation. The asymptotic normality does not take the uncertainty due to variable selection into account, which is a topic of post-selection inference (PoSI, Berk et al. (2013)). More discussion is given in Remark 2 of Web Appendix E.

5. Computational strategies

The goal is to estimate $\big(\pi_j(u), \phi_j(u), \boldsymbol{\beta}_j(u)\big), j = 1,\ldots,C$, at points $u_1,\ldots,u_n$, by obtaining an approximate solution for (11). To avoid mixture label switching (Huang et al., 2013), we use a modified EM algorithm for estimation of each function simultaneously over $u_1,\ldots,u_n$.

We view the observations $\{(u_i, \boldsymbol{x}_i, y_i) : i = 1,\ldots,n\}$ as incomplete data, and introduce

the unobserved Bernoulli variables Z_{ij} to represent the membership of the i^{th} observation to the j^{th} mixture component, $j = 1, \ldots, C$, such that $\Pr(Z_{ij} = 1 | u_i, \boldsymbol{x}_i, y_i) = \pi_j(u_i)$. The complete data are $\{(u_i, \boldsymbol{x}_i, y_i, \boldsymbol{z}_i) : i = 1, \ldots, n\}$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iC})^\top$. We define the complete (total) local-kernel log-likelihood function as

$$L_n^c(\boldsymbol{\Psi}; h) = \sum_{t=1}^{n} \sum_{i=1}^{n} \sum_{j=1}^{C} Z_{ij} \big\{ \log \pi_{j,t} + \log f(y_i; \theta_{j,t}(\boldsymbol{x}_i), \phi_{j,t}) \big\} K_h(u_i - u_t),$$

and the penalized complete local-kernel log-likelihood as $\widetilde{L}_n^c(\boldsymbol{\Psi}; h) = L_n^c(\boldsymbol{\Psi}; h) - \mathbb{P}_n(\boldsymbol{\Psi}; \boldsymbol{\lambda})$, where the penalty $\mathbb{P}_n$ is given in (10). Due to the non-differentiability of $p_n(\theta; \lambda)$ at $\theta = 0$, as suggested by Fan and Li (2001), we use the local quadratic approximation (LQA)

$$\bar{p}_n(\theta; \theta^{(0)}, \lambda) = p_n(\theta^{(0)}; \lambda) + p_n'(\theta^{(0)}; \lambda)(\theta^2 - \theta^{2(0)})/(2\theta^{(0)}), \tag{16}$$

for p_n, where $\theta^{(0)}$ is an initial value. The LQA is used in the modified EM algorithm. Given $\boldsymbol{\Psi}^{(m)}$, at the $(m+1)^{\text{th}}$ iteration, the algorithm proceeds as follows.

E-step. Since the Z_{ij}'s are unobservable, we compute the expectation of $\widetilde{L}_n^c(\boldsymbol{\Psi}; h)$ with respect to Z_{ij} conditional on $\{(u_i, \boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ and $\boldsymbol{\Psi}^{(m)}$. For all $i = 1, \ldots, n$, $j = 1, \ldots, C$, it boils down to the computation of the conditional expectations

$$E(Z_{ij} | \boldsymbol{\Psi}^{(m)}, u_i, \boldsymbol{x}_i, y_i) \equiv w_{ij}^{(m)} = \frac{\pi_j^{(m)}(u_i) f\big(y_i; \theta_j^{(m)}(\boldsymbol{x}_i, u_i), \phi_j^{(m)}(u_i)\big)}{\sum_{j=1}^{C} \pi_j^{(m)}(u_i) f\big(y_i; \theta_j^{(m)}(\boldsymbol{x}_i, u_i)), \phi_j^{(m)}(u_i)\big)}.$$

M-step. Using the LQA (16) evaluated at $\theta = \|\boldsymbol{b}_{jl}\|_2/\sqrt{n}$ and $\theta^{(m)} = \|\boldsymbol{b}_{jl}^{(m)}\|_2/\sqrt{n}$, we maximize the objective function (up to some constants)

$$\widetilde{Q}(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) = \sum_{j=1}^{C} \sum_{t=1}^{n} \Big\{ \sum_{i=1}^{n} w_{ij}^{(m)} \big\{ \log \pi_{j,t} + \log f(y_i; \theta_{j,t}(\boldsymbol{x}_i), \phi_{j,t}) \big\} K_h(u_i - u_t) - \boldsymbol{\beta}_{j,t}^\top \boldsymbol{\Sigma}_j^{(m)} \boldsymbol{\beta}_{j,t}/2 \Big\}$$

with respect to $\boldsymbol{\Psi}$, where $\theta_{j,t}(\boldsymbol{x}_i) = g(\boldsymbol{x}_i^\top \boldsymbol{\beta}_{j,t})$ with the vector $\boldsymbol{\beta}_{j,t} = \big(\beta_{j1}(u_t), \ldots, \beta_{jd}(u_t)\big)^\top$, the diagonal matrices $\boldsymbol{\Sigma}_j^{(m)} = \text{diag}\{\tau_{jl}^{(m)} : l = 1, \ldots, d\}$ and $\tau_{jl}^{(m)} = p_n'\big(\theta^{(m)}; \lambda_j\big)/[n\theta^{(m)}]$.

The maximization of $\widetilde{Q}$ with respect to $\boldsymbol{\Psi}$ results in the following updates. The probabilities $\pi_j(u_t)$ and the vectors $\boldsymbol{\beta}_{j,t} = \boldsymbol{\beta}_j(u_t)$, $t = 1, \ldots, n, j = 1, \ldots, C$, are updated by

$$\pi_{j,t}^{(m+1)} = \pi_j^{(m+1)}(u_t) = \sum_{i=1}^{n} w_{ij}^{(m)} K_h(u_i - u_t) \Big/ \sum_{i=1}^{n} K_h(u_i - u_t),$$

$$\boldsymbol{\beta}_j^{(m+1)}(u_t) = \arg\max_{\boldsymbol{\beta}_{j,t}\in\mathbb{R}^d} \sum\nolimits_{i=1}^{n} w_{ij}^{(m)} \big\{ \log f(y_i; \theta_{j,t}(\boldsymbol{x}_i), \phi_{j,t}) \big\} K_h(u_i - u_t) - \boldsymbol{\beta}_{j,t}^{\top} \boldsymbol{\Sigma}_j^{(m)} \boldsymbol{\beta}_{j,t}/2. \quad (17)$$

The dispersion parameters $\phi_{j,t} = \phi_j(u_t)$ are updated by solving the estimation equations

$$\sum\nolimits_{i=1}^{n} w_{ij}^{(m)} \frac{\partial}{\partial \phi_{j,t}} \big\{ \log f(y_i; \theta_{j,t}^{(m+1)}(\boldsymbol{x}_i), \phi_{j,t}) \big\} K_h(u_i - u_t) = 0. \quad (18)$$

To solve (17)-(18), depending on f, we may need to use the Newton-Raphson method. Details of the algorithm for Gaussian and t-distribution f are given in Web Appendix F.

Starting from an initial $\boldsymbol{\Psi}^{(0)}$ (discussed in Web Appendix F.4), the EM algorithm iterates until a convergence criterion is satisfied. We used the stopping rule $\|\boldsymbol{\Psi}^{(m+1)} - \boldsymbol{\Psi}^{(m)}\|_{\mathrm{F}} < \epsilon$. We set the estimates of the regression functions $\beta_{jl}(u)$ to zero if $\|\boldsymbol{b}_{jl}^{(m+1)}/n\| < \delta$. In our simulation and the real data analysis, we chose $\epsilon = \delta = 10^{-4}$. We describe data-adaptive strategies for the selection of the band-width h and the tuning parameters λ_j's, and also estimation of the mixture order C in Web Appendix G.

[Table 1 about here.]

6. Simulation study

We asses the finite-sample performance of the methods via simulations. We considered Gaussian FM-VCRs with $C = 2$ and 3 components. The parameter settings for the model with $C = 2$ are given in Table 1, and those for $C = 3$ are given in Web Table 1.

Let CEZ = # Correctly Estimated Zero, CEN = # Correctly Estimated Non-zero, IEZ = # Incorrectly Estimated Zero, and IEN = # Incorrectly Estimated Non-zero $\beta_{jl}(\cdot)$'s. We define

$$\text{Sensitivity} = \text{CEN}/(\text{CEN} + \text{IEZ}) \quad \text{and} \quad \text{Specificity} = \text{CEZ}/(\text{CEZ} + \text{IEN}),$$

which assess the performance of the penalization method for variable selection. We measure the estimation errors of the proposed estimators using

$$L_2^2(\widehat{\boldsymbol{\beta}}_j) = n^{-1} \sum\nolimits_{i=1}^{n} \sum\nolimits_{l=0}^{d} \big(\widehat{\beta}_{jl}(u_i) - \beta_{jl}(u_i)\big)^2, \quad L_2^2(\widehat{\boldsymbol{\sigma}}_j) = n^{-1} \sum\nolimits_{i=1}^{n} \big(\widehat{\sigma}_j(u_i) - \sigma_j(u_i)\big)^2,$$
$$L_2^2(\widehat{\boldsymbol{\pi}}_j) = n^{-1} \sum\nolimits_{i=1}^{n} \big(\widehat{\pi}_j(u_i) - \pi_j(u_i)\big)^2, \quad j = 1, \ldots, C.$$

EXAMPLE 1: We generated the vectors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^\top, i = 1, \ldots, n$, from a zero-mean multivariate normal with a covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{kl} = (0.5)^{|k-l|} : 1 \leqslant k, l \leqslant d\}$. The points $u_1, \ldots, u_n$ were generated from Uniform$[0, 1]$. Given each $(\boldsymbol{x}_i, u_i)$, we generated the response y_i from a Gaussian FM-VCR with $C = 2$ and the parameter setting given in Table 1, for the sample sizes $n = 200, 400$. We considered dimensions $d = 5, 10, 20, 50$, and we set non-zero coefficients $\big(\beta_{11}(u), \beta_{12}(u), \beta_{14}(u)\big)$ and $\big(\beta_{21}(u), \beta_{23}(u)\big)$ in the 1st and 2nd components of the mixture model. In Table 1, by increasing $d = 5$ to $10, 20, 50$, the non-zero $\beta_{jl}(\cdot)$'s remain the same and we add more zeros to each component.

The average sensitivity, specificity, and estimation errors (over $R = 300$ simulated samples) for the model with $C = 2$ and $d = 5, 10, 20$ are given in Table 2. An extended version of this table which includes the results for $d = 50$ is given in Web Table 2.

[Table 2 about here.]

From Table 2, the average sensitivity and specificity corresponding to $n = 200$ vary approximately between 47% to 95% and 80% to 99%, respectively, depending on the dimension $d = 5, 10, 20$, the mixture component, and the penalty function. It seems that the task of identifying the true non-zero regression coefficients $\beta_{jl}(\cdot)$ compared to the identification of true zero coefficients, for the smaller sample size $n = 200$, is more difficult. As the sample size increases to $n = 400$, all the penalties improve in terms of identifying both true non-zero and zero regression functions. In terms of the estimation error (L_2), the SCAD and MCP are closer to the oracle estimator, followed by the AdpLASSO, LASSO, and MLLE has the worst performance as expected. Overall, the AdpLASSO, SCAD, and MCP perform similarly in terms of the three performance measures, followed by the LASSO. The results for $d = 50$ given in Web Table 2 follow the same trend with better performance for larger n.

The results for the model with $C = 3$, $d = 5, 10, 20, 50$ and $n = 200, 400, 600, 800$ are given in Web Tables 3 and 4. This is clearly a more challenging setting and we can see that

when $n = 200$, the sensitivity in one of the mixture components is 35% with a specificity around 80%. For $n = 800$, the sensitivity reaches 75% for this component. For the other two components, the sensitivity and specificity values are 80%-96% and 75%-92%, respectively.

Figure 1 shows the mixture order selection results based on the BIC_2 in (A.79), for the model with $C = 2$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version. For a given dimension d, as n increases the performance of BIC_2 improves. As d increases from 5 to 20, the model becomes more complex and the model selection task becomes more difficult, as expected. Overall, the correct order is selected most often (and the majority is over 85%) even in the most difficult scenario.

[Figure 1 about here.]

EXAMPLE 2: Since the covariates in our real data are SNPs, we simulate x_j's to mimic such variables. We considered discrete covariates, each taking values in $\{0, 1, 2\}$, and each was randomly generated from the multinomial distribution $\text{MNomial}\big((0, 1, 2), [(1-p)^2, 2p(1-p), p^2]\big)$, where $p \in \{.05, .1, .2, .3, .4, .5\}$. We note that the smaller the value of p, the lesser the variation in the simulated covariates from this model. The parameter setting for the Gaussian FM-VCR is the same as those in Table 1. The results for $d = 5$ and $n = 200, 400$ are respectively given in Web Tables 5 and 6. The main purpose of this example is to show that for discrete covariates like SNPs, even for relatively small dimension $d = 5$, larger sample sizes are required to achieve a similar performance as those discussed in Example 1. From Web Table 5, we can see that for $n = 200$, the task of identifying non-zero regression coefficients (sensitivity) is more difficult for $p = .05$ and .10, which also results in higher estimation errors in one of the mixture components. As the sample size increases to $n = 400$ (Web Table 6), the overall performance of the method improves for all penalties. This result is encouraging for our real data analysis, since our sample size is even larger. In this setting, SCAD and MCP perform better than the LASSO and AdpLASSO.

7. Real data analysis

The *Fangchenggang Area Male Health and Examination Survey* was conducted to identify factors that might influence Osteocalcin (OCN) (Liao et al., 2014) and other phenotypes of interest. OCN is believed to play a role in metabolic regulation (Lee et al., 2007). It has also been observed that higher serum OCN levels are correlated with increases in bone mineral density during drug treatment for osteoporosis (Bharadwaj et al., 2009). The data include information on 2,200 unrelated healthy Chinese males, with their age, serving as our index variable (U), ranging from 20 to 69. To guard against potential confounders, individuals included in the study were free from a list of diseases, such as stroke, diabetes mellitus, primary hypertension, and hyperthyroidism. Following Zhang (2017), we focus on studying the relationship between OCN and SNPs in Chr7. We also include 7 other covariates: smoking status, physical activity, drinking, (log) body mass index, sex hormone-binding globulin, ferritin, and folic acid. After deleting individuals with missing data, SNPs with close to zero variability, and those contributing little additional information in the presence of others, we are left with $d = 43$ covariates (including 36 SNPs) and $n = 1704$ individuals.

In our analysis, we considered a Gaussian FM-VCR, where the effects of the covariates (X_j's) on OCN (Y) and the mixing proportions (π's) are all functions of age (U). We fitted models with $C = 1, \ldots, 5$, and BIC selected $C = 2$. Since our algorithm and Example 2 based on SCAD (compared to other penalties) provided more stable estimates for different initial values of the algorithm, our results below are based on the fitted model with SCAD.

[Figure 2 about here.]

The density plots of the observed y_i's classified into the two components of the fitted FM-VCR model are given in Figure 2(a). This figure appears in color in the electronic version of this article, and any mention of color refers to that version. Component 1 can

be interpreted as representing the high "OCN" subpopulation (hOCN), and Component 2 as the low "OCN" (lOCN) subpopulation. Figure 2(b) shows that as one ages, the probability of being in the hOCN decreases, which is consistent with biological theory since it has been found that osteocalcin is related to bone mineral density (Bharadwaj et al., 2009). From Figure 2(c,d), we observe that the estimated functional intercepts in both subpopulations are age-dependent, with decreasing OCN values as age increases; the decrease appears to slow down starting from age 35 for the hOCN, whereas the decrease is fairly linear for the lOCN. Furthermore, for the hOCN subpopulation, we found the effect of log(BMI) on OCN to be negative but non-linear (varying with age), with the negative effect generally decreasing with age (Figure 2(e)). Among the other non-genetic covariates considered, folic acid also has a non-linear negative effect on OCN (Web Figures 1(d)). Most significantly, rs7456421, an SNP linked to the OCN level (Zhang, 2017), is identified to have a significant effect on OCN in the hOCN subpopulation. The effect is non-constant over time; in particular, having the SNP would negatively impact the OCN at a younger age (Figure 2(f)). Further discussion is provided in Section 8. This SNP resides in gene HIPK2, which codes for homeodomain interacting protein kinase 2, and is a multi-functional signaling molecule, including being studied as a tumor suppressor recently (Feng et al., 2017) and as a bone morphogenetic protein (Harada et al., 2003). There are a number of other SNPs that show significant effects in the hOCN subpopulation (Web Figure 1(a)-(c)). The estimated variances in both subpopulations are given in Web Figure 1(e)-(f).

In contrast, in the lOCN subpopulation, beyond the finding that the OCN level decreases with age, it is rather interesting to see that none of the other factors investigated are selected by the penalization method. Taken together, our results not only corroborated the findings of Zhang (2017) but also strengthened the results by providing evidence of a dynamic relationship in one of the two subpopulations identified by the model. In the literature, OCN

levels have been used as a biomarker for measuring osteoporosis treatment effects (Bharadwaj et al., 2009). Therefore, our finding of the differential effects of SNPs and other covariates on OCN between the two subpopulations perhaps warrants further investigation into whether the conclusion on the use of OCN as a biomarker still holds within each of these subpopulations.

8. Discussion

We developed penalized local-kernel likelihood methods for FM-VCR models and established their consistency in estimation and variable selection. We examined the finite-sample performance of the methods via simulation, and they performed well for the dimensions and sample sizes considered here. When the covariates are discrete, a larger n is required to achieve similar performance as those with continuous covariates. We observed that the method based on all four penalties performs similarly, and none of the penalties universally dominates the others. Thus, in practice, we suggest analyzing a dataset using all the penalties and choosing a final fitted sparse model that optimizes a selection criterion (e.g., BIC) and has more stable estimates based on different initial values for the EM, as done in our real data analysis.

Indeed, in our OCN data analysis, we explored several ways of analyzing the data. We fitted an FMR model which showed a lack of fit (Web Figure 2), motivating the use of an FM-VCR model. Additional details are provided in Web Appendix I. In our FM-VCR analysis, we used the penalties investigated in the simulation and with multiple initial values for the EM algorithm. Our results are based on the SCAD penalty, which provided the most stable outcome and was selected by the BIC. The fit of the model to the data is seen to improve (Web Figure 2) over the FMR, and a number of age-dependent covariates were also selected. Among them, SNP rs7456421, implicated in the literature for its link to OCN level, was selected in the high OCN subpopulation and shown to have a larger impact on younger people. To further substantiate the finding of the varying effect of the SNP over the age on OCN, we added the point-wise error bars to the estimated varying coefficient plot of the SNP

(Web Figure 3(a)). We also plotted the derivative curve (approximated using Matlab) of the estimated coefficient effect over age (Web Figure 3(b)). It is evident from these plots that the underlying true derivate function is unlikely to be 0 over the entire age range, substantiating the finding of a varying coefficient for the SNP effect.

To gauge the performance of our proposed method under a number of conditions that deviate from our theoretical developments, we conducted additional simulations. First, we studied the effect of model misspecifications in terms of mixture components or the order C in (3). When the data were generated from a t FM-VCR but fitted a misspecified Gaussian FM-VCR, the results show that the performance of the method does not degrade much in variable selection and estimation error for the setting considered (Web Table 7). The details and discussion of the results are given in Web Appendix H.1. On the other hand, for a model with true $C = 2$, if the order was incorrectly specified as $C = 1$, then the one-component underfitted model resembles the behavior of the larger component of the correct model, but with lower sensitivity and specificity and larger estimation errors for the corresponding measures in Component 2 when $C = 2$ was fitted (Web Table 8). In contrast, if C was incorrectly specified as larger than the true value, then the behavior of the overfitted models is similar to those with the correct model and with lower estimation errors. These simulation results are in line with the theoretical properties of over-fitted finite mixture models (Ho et al., 2022). More discussion is provided in Web Appendix H.1. We recognize that our observations are based on limited simulations; thus, further study on the effect of model misspecification on our method, including properties of over-fitted FM-VCR, is warranted.

We also studied the effect of fitting an FM-VCR model when data were generated from an FMR model (without allowing for varying coefficients). The parameter setting is given in Web Appendix H.2. We observe that for sensitivity, FM-VCR (the wrong model) is generally worse (on average based on 100 runs) than FMR (the true underlying model), with larger standard

deviations (SDs) (Web Table 9). On the other hand, for specificity, FM-VCR is consistently better and with much smaller SDs. This observation implies that FM-VCR selected fewer variables: fewer true positives and fewer false positives, hence larger specificity and smaller sensitivity. This observation is intuitively sensible: with a more complex model and a fixed sample size, variable selection is harder, especially for the parameter setting considered in this simulation where the effects of covariates (β_{jl}'s) are weak. Thus, in practice, one may fit both models to a data and assess the results as done in our data analysis.

In another simulation study, we also assessed the performance of the penalization method when the dimension d exceeds the sample size n, although this case is not covered in our theoretical results where d is fixed as n grows. We considered a setting with $d = 500$ and $n = 200, 400$ (Web Appendix H.4). Our results (Web Table 12) show that, while the specificity remains high and similar to the dimensions already considered in Section 6 (Table 2 and Web Table 2), there is some reduction in sensitivity, although the difference is quite small when compared to $d = 50$, indicating that reasonable results are likely to be obtained by the method even when the number of variables exceeds the sample size. Alternatively, one may screen the covariates to reduce the dimension, which appears to be reasonable as a preliminary step (Web Appendix H.4 and Web Tables 13–14). Nevertheless, rigorous study of high-dimensional settings requires new theoretical and numerical tools beyond the scope of the current paper and is a topic of future research. Finally, the mixing proportions π_j may also be considered as functions of some covariates in addition to the index variable; however, modeling and estimation become unwieldy and thus not pursued in the current study.

Acknowledgements

We would like to thank the editor, an associate editor, and two referees for their thoughtful comments. The authors also thank Professors Zengnan Mo and Ming Liao at Guangxi Medical University (Nanning, China) for their generosity of sharing the OCN data. A. Khalili and M. Asgharian are supported by the Natural Science and Engineering Research Council

of Canada (NSERC RGPIN-2020-05011) and (NSERC RGPIN-2018-05618). F. Shokoohi is supported by University of Nevada, Las Vegas, through the Startup Grant (PG18929). S. Lin is supported by NIH National Institute of General Medical Sciences (R01GM114142).

Data Availability

Data sharing is not applicable to this article as no new data were created or analyzed in this paper.

References

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* **41,** 802 – 837.

Bharadwaj, S., Naidu, A. G. T., Betageri, G. V., Prasadarao, N. V., and Naidu, A. S. (2009). Milk ribonuclease-enriched lactoferrin induces positive effects on bone turnover markers in postmenopausal women. *Osteoporosis International* **20,** 1603–1611.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association* **83,** 173–178.

Demontiero, O., Vidal, C., and Duque, G. (2012). Aging and bone loss: new insights for the clinician. *Therapeutic advances in musculoskeletal disease* **4,** 61–76.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* **39,** 1–38.

Fan, J. and Gigbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics* **27,** 1491–1518.

Feng, Y., Zhou, L., Sun, X., and Li, Q. (2017). Homeodomain-interacting protein kinase 2 (hipk2): a promising target for anti-cancer therapies. *Oncotarget* **8,** 20452–20461.

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P. (2018). *Handbook of Mixture Analysis*. Chapman and Hall/CRC, 1st edition.

Harada, J., Kokura, K., Kanei-Ishii, C., Nomura, T., Khan, M. M., Kim, Y., et al. (2003). Requirement of the co-repressor homeodomain-interacting protein kinase 2 for ski-mediated inhibition of bone morphogenetic protein-induced transcriptional activation. *Journal of Biological Chemistry* **278,** 38998–39005.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)* **55,** 757–796.

Ho, N., Yang, C.-Y., and Jordan, M. I. (2022). Convergence rates for gaussian mixtures of experts. *Journal of Machine Learning Research* .

Huang, M., Li, R., and Wang, S. (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association* **108,** 929–941.

Huang, M., Yao, W., Wang, S., and Chen, Y. (2018). Statistical inference and applications of mixture of varying coefficient models. *Scandinavian Journal of Statistics* **45,** 618–643.

Janson, S. (1987). Maximal spacing in several dimensions. *Annals of Probability* **15,** 274–80.

Khalili, A. and Chen, J. (2007). Variable Selection in Finite Mixture of Regression Models. *Journal of the American Statistical Association* **102,** 1025–1038.

Lee, N. K., Sowa, H., Hinoi, E., Ferron, M., Ahn, J. D., Confavreux, C., et al. (2007). Endocrine regulation of energy metabolism by the skeleton. *Cell* **130,** 456–469.

Liao, M., Shi, J., Huang, L., Gao, Y., Tan, A., Wu, C., et al. (2014). Genome-wide association study identifies variants in pms1 associated with serum ferritin in a chinese population. *PloS one* **9,** e105844.

Liu, D. M., Guo, X. Z., Tong, H. J., Tao, B., Sun, L.-H., Zhao, H.-Y., et al. (2015). Association between osteocalcin and glucose metabolism: a meta-analysis. *Osteoporosis International* **26,** 2823–2833.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. Wiley & Sons, New York.

Shokoohi, F., Khalili, A., Asgharian, M., and Lin, S. (2019). Capturing heterogeneity of covariate effects in hidden subpopulations in the presence of censoring and large number of covariates. *Annals of Applied Statistics* **13,** 444–465.

Silverman, B. W. (1984). Spline Smoothing: The Equivalent Variable Kernel Method. *The Annals of Statistics* **12,** 898 – 916.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). L_1-penalization for Mixture Regression Models. *Test* **19,** 209–256.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58,** 267–288.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* **104,** 747–757.

Wei, F., Huang, J., and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica* **21,** 1515–1540.

Xiang, S., Yao, W., and Yang, G. (2019). An overview of semiparametric extensions of finite mixture models. *Statistical Science* **34,** 391–404.

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

Zhang, C.-H. (2010). Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics* **38,** 894–942.

Zhang, H. (2017). *Detecting rare haplotype-environmental interaction and nonlinear effects of rare haplotypes using Bayesian lasso on quantitative traits*. PhD thesis, Ohio State U.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

Supporting Information

Web Appendices, Tables, and Figures referenced in Sections 2-8, and codes used for simulation and data analysis, are available with this paper at the Biometrics website on Wiley Online Library.

Received January 2022. *Revised August* 2022. *Accepted September* 2022.

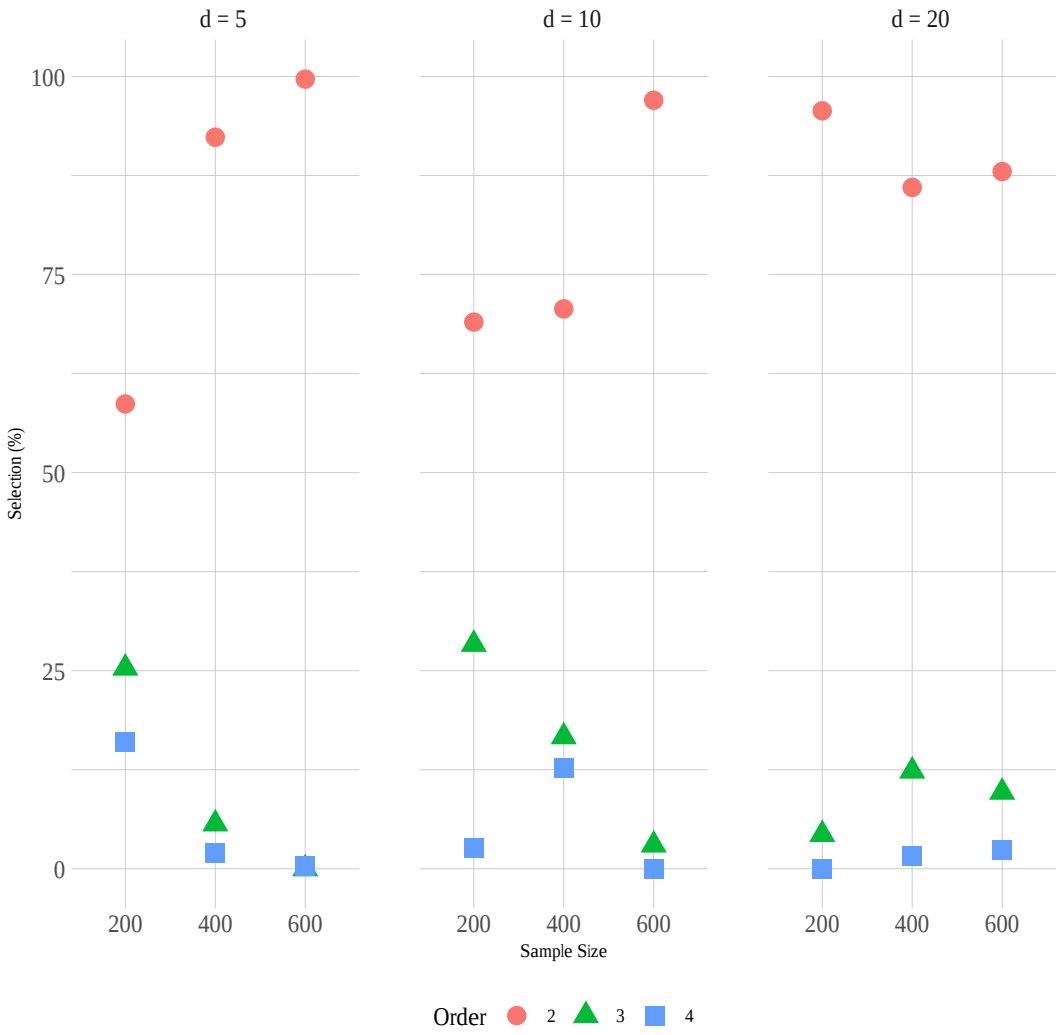


Figure 1: Example 1: Order selection results for the model with true order $C = 2$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

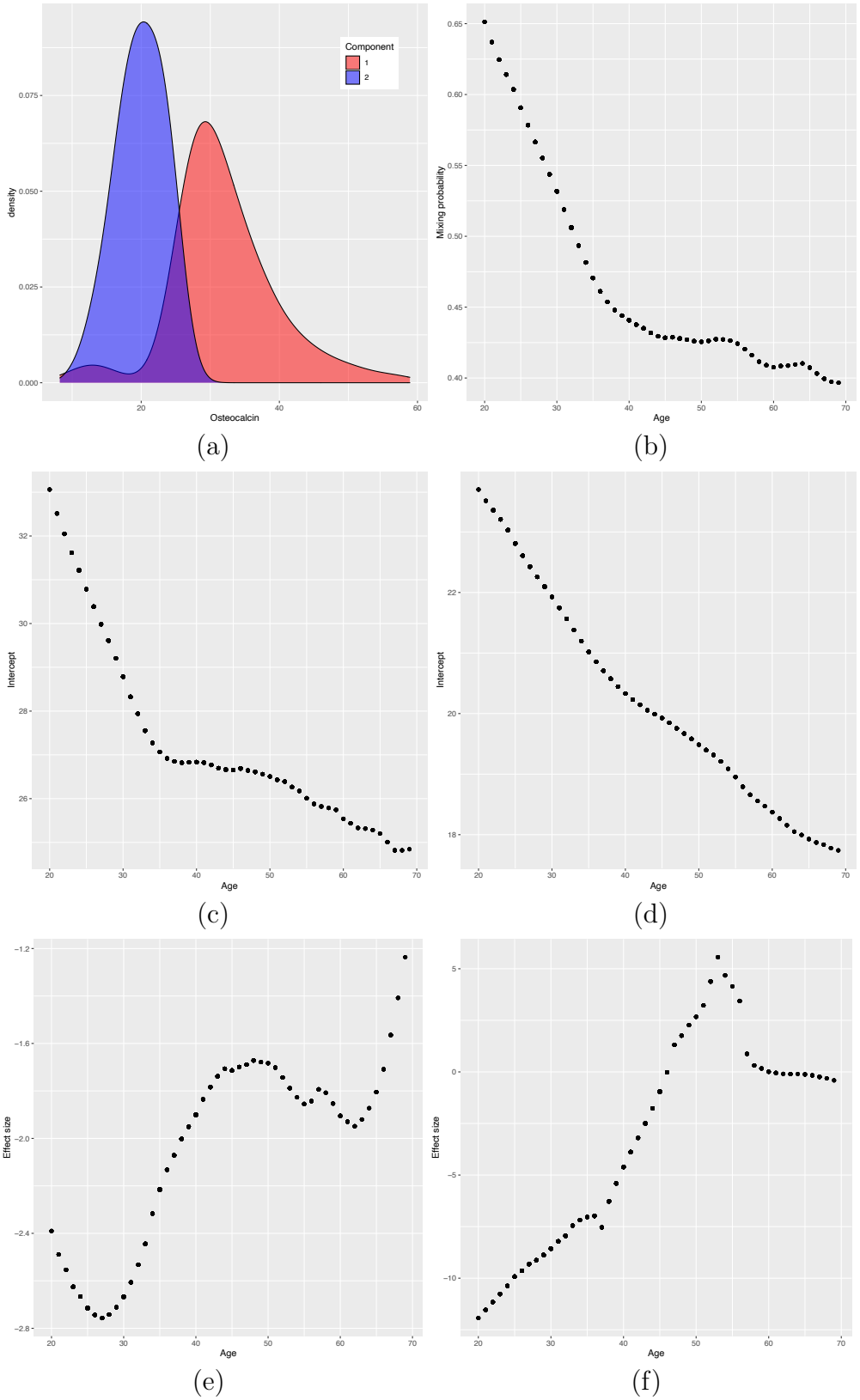


Figure 2: Osteocalcin data analysis; (a) Density plots: Component 1 (hOCN, red), Component 2 (lOCN, blue); (b) Estimated mixing probabilities over time in hOCN; (c) Estimated intercept over time in hOCN; (d) Estimated intercept over time in lOCN; (e) Estimated effect of log(BMI) over time in hOCN; (f) Estimated effect of rs7456421 over time in hOCN. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Table 1: Parameters settings for the Gaussian FM-VCR model with $C = 2$.

Component(j):	1	2
parameters	$d(= 5, 10, 20, 50)$[1]	
$\beta_{j0}(u)$	-2	-1
$\beta_{j1}(u)$	$1 + 0.5\cos(\pi u)$	$1.5\sin(\pi u)$
$\beta_{j2}(u)$	$1 + 0.5\cos(2\pi u)$	0
$\beta_{j3}(u)$	0	$1.5 - 0.5\sin(\pi u/2)$
$\beta_{j4}(u)$	$\sin(6\pi u)$	0
$\beta_{j5}(u)$	0	0
$\vdots$	$\vdots$	$\vdots$
$\beta_{j,50}(u)$	0	0
$\sigma_j(u)$	$0.3e^{(0.5u)}$	$0.5e^{(-0.2u)}$
$\pi_j(u)$	$e^{0.5u}/(1 + e^{0.5u})$	$(1 + e^{0.5u})^{-1}$

[1] We keep the non-zero coefficient functions $\beta_{jl}(\cdot)$ the same for different values of the dimension d.

Table 2: Results of Example 1: average (SD) sensitivity, specificity, and estimation errors.

$C=2$		Criteria	Sensitivity		Specificity		$L_2(\widehat{\boldsymbol{\beta}}_j)$		$L_2(\widehat{\boldsymbol{\sigma}}_j^2)$		$L_2(\widehat{\boldsymbol{\pi}}_j)$	
d	n	Component	1	2	1	2	1	2	1	2	1	2
5	200	Oracle	—	—	—	—	.650	.162	.383	.104	.321	.104
		MLLE	—	—	—	—	1.04	.243	.748	.127	.370	.127
		AdpLASSO	$.687_{(.203)}$	$.948_{(.191)}$	$.967_{(.125)}$	$.962_{(.122)}$	.934	.208	.612	.118	.500	.118
		LASSO	$.701_{(.223)}$	$.847_{(.347)}$	$.813_{(.265)}$	$.824_{(.214)}$	.983	.256	.812	.167	.524	.167
		MCP	$.682_{(.127)}$	$.952_{(.153)}$	$.978_{(.110)}$	$.972_{(.120)}$	.888	.163	.561	.094	.505	.094
		SCAD	$.684_{(.127)}$	$.957_{(.147)}$	$.980_{(.106)}$	$.974_{(.118)}$	.882	.165	.554	.093	.507	.093
	400	Oracle	—	—	—	—	.399	.083	.214	.080	.273	.080
		MLLE	—	—	—	—	.547	.097	.389	.075	.278	.075
		AdpLASSO	$.978_{(.083)}$	$.997_{(.058)}$	$.998_{(.030)}$	$.997_{(.043)}$	.499	.082	.225	.074	.311	.074
		LASSO	$.989_{(.060)}$	$.995_{(.064)}$	$.938_{(.179)}$	$.910_{(.180)}$	.582	.116	.376	.086	.338	.086
		MCP	$.983_{(.073)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	$.999_{(.019)}$	.458	.078	.208	.073	.313	.073
		SCAD	$.984_{(.070)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	$.999_{(.019)}$	.458	.077	.207	.073	.313	.073
10	200	Oracle	—	—	—	—	.698	.162	.397	.089	.314	.089
		MLLE	—	—	—	—	1.23	.287	1.02	.138	.335	.138
		AdpLASSO	$.589_{(.239)}$	$.893_{(.259)}$	$.970_{(.067)}$	$.966_{(.061)}$	1.09	.288	.848	.161	.501	.161
		LASSO	$.470_{(.321)}$	$.775_{(.401)}$	$.928_{(.100)}$	$.917_{(.087)}$	1.32	.383	1.15	.294	.505	.294
		MCP	$.581_{(.194)}$	$.895_{(.245)}$	$.984_{(.045)}$	$.981_{(.048)}$	1.04	.238	.763	.106	.503	.106
		SCAD	$.594_{(.188)}$	$.892_{(.243)}$	$.983_{(.049)}$	$.980_{(.050)}$	1.03	.237	.761	.107	.501	.107
	400	Oracle	—	—	—	—	.439	.089	.227	.082	.273	.082
		MLLE	—	—	—	—	.878	.149	.607	.087	.337	.087
		AdpLASSO	$.690_{(.127)}$	$.993_{(.057)}$	$.997_{(.023)}$	$.995_{(.029)}$	.808	.165	.390	.069	.411	.069
		LASSO	$.766_{(.230)}$	$.970_{(.166)}$	$.970_{(.066)}$	$.962_{(.064)}$	.862	.232	.518	.108	.425	.108
		MCP	$.692_{(.104)}$	$.992_{(.064)}$	$.997_{(.020)}$	$.998_{(.014)}$	.789	.145	.374	.06	.412	.062
		SCAD	$.688_{(.098)}$	$.992_{(.064)}$	$.996_{(.023)}$	$.998_{(.016)}$	.792	.146	.377	.062	.412	.062
20	200	Oracle	—	—	—	—	.615	.095	.297	.079	.313	.079
		MLLE	—	—	—	—	1.37	.225	1.30	.082	.383	.082
		AdpLASSO	$.646_{(.115)}$	$.943_{(.202)}$	$.981_{(.043)}$	$.984_{(.042)}$	.914	.269	.713	.056	.488	.056
		LASSO	$.657_{(.120)}$	$.953_{(.187)}$	$.963_{(.059)}$	$.952_{(.053)}$	.924	.336	.896	.056	.494	.056
		MCP	$.643_{(.121)}$	$.953_{(.199)}$	$.994_{(.028)}$	$.995_{(.022)}$	.915	.220	.619	.052	.467	.052
		SCAD	$.641_{(.127)}$	$.947_{(.206)}$	$.994_{(.028)}$	$.996_{(.021)}$	.918	.227	.633	.052	.472	.052
	400	Oracle	—	—	—	—	.353	.072	.211	.067	.251	.067
		MLLE	—	—	—	—	.912	.102	.616	.038	.423	.038
		AdpLASSO	$.668_{(.019)}$	$.998_{(.038)}$	$.999_{(.015)}$	$1.00_{(.005)}$	.798	.170	.417	.036	.484	.036
		LASSO	$.716_{(.118)}$	$1.00_{(.000)}$	$.978_{(.035)}$	$.969_{(.041)}$	.823	.237	.523	.036	.499	.036
		MCP	$.667_{(.047)}$	$.997_{(.058)}$	$.999_{(.010)}$	$1.00_{(.007)}$	.796	.159	.433	.036	.481	.036
		SCAD	$.668_{(.051)}$	$.997_{(.058)}$	$.999_{(.011)}$	$1.00_{(.007)}$	.796	.159	.433	.036	.481	.036

BIOMETRICS ??, 1–55 DOI: x.x/j.x-x.x.x.x
January 2022

Supporting Information for "Sparse Estimation in Semi-parametric Finite Mixture of Varying Coefficient Regression Models" by

Abbas Khalili[1,*], Farhad Shokoohi[2,], Masoud Asgharian[3,***], and Shili Lin[4,****]**

[1,3]Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada

[2]Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA

[4]Department of Statistics, Ohio State University, Columbus, OH 43210, USA

*_email:_ abbas.khalili@mcgill.ca

**_email:_ farhad.shokoohi@unlv.edu

***_email:_ masoud.asgharian2@mcgill.ca

****_email:_ shili@stat.osu.edu

This paper has been submitted for consideration for publication in *Biometrics*

Web Appendix A. Identifiability conditions of f_C^*

The following conditions are considered for an FM-VCR model with probability density (mass) function f_C^* in (3) of the paper to be identifiable (see Theorem 1 of Huang et al. (2018).).

(C1) For $j = 1, \ldots, C$, $\pi_j(u) > 0$ is a continuous function, and $\boldsymbol{\beta}_j(u)$ and $\phi_j(u)$ are differentiable functions of u.

(C2) Any two curves $\boldsymbol{a}_1(u) = (\boldsymbol{\beta}_i(u), \phi_i(u))$ and $\boldsymbol{a}_2(u) = (\boldsymbol{\beta}_j(u), \phi_j(u))$, $i \neq j$, are transversal; i.e., for any $u \in \mathcal{U}$, $\|\boldsymbol{a}_1(u) - \boldsymbol{a}_2(u)\|^2 + \|\boldsymbol{a}_1'(u) - \boldsymbol{a}_2'(u)\|^2 \neq 0$, and $\|\cdot\|$ is the Euclidean norm.

(C3) The support for U, denoted by $\mathcal{U}$, is a compact subset of the real numbers $\mathbb{R}$.

(C4) The density $g(\boldsymbol{x}|u)$ of $\boldsymbol{X}$ given $U = u$ is a full dimensional (d-dimensional) density function.

Web Appendix B. Regularity Conditions

In the following three subsections, we state Regularity Conditions (RC.1)-(RC.5) on the distributions f_C^*, g, and m which is the density of U; Conditions (KC.1)-(KC.2) on the kernel K, and Conditions (PC.1)-(PC.3) on the penalty p_n and the smoothing/tuning parameters h_n and λ_n.

Web Appendix B.1 *Conditions on f_C^*, g, and m*

(RC.1) The set $\{(U_i, \boldsymbol{X}_i, Y_i),\ i = 1, \ldots, n\}$ is a sample of independent and identically distributed variables drawn from its population $\boldsymbol{V} = (U, \boldsymbol{X}, Y)$ with the probability density function $f_{\boldsymbol{V}}(u, \boldsymbol{x}, y) = f_C^*\big(y|\boldsymbol{\psi}(u), \boldsymbol{x}\big) g(\boldsymbol{x}|u) m(u)$. The density has a common support in $\boldsymbol{v}$ for all the parameter values $\boldsymbol{\psi} \in \Omega$, and $f_{\boldsymbol{V}}$ or equivalently the mixture density f_C^* in (3) of the paper is identifiable up to a permutation of the mixture components. (See Section Web Appendix A for identifiability.)

(RC.2) The density $m(u)$ of U is positively bounded away from zero over its compact support $\mathcal{U}$; it is continuously twice differentiable, and these derivatives are uniformly bounded over $\mathcal{U}$. The density f_C^* admits third partial derivatives with respect to its parameters $\boldsymbol{\psi}$ for all $(u, \boldsymbol{x}, y)$; g admits its third derivatives with respect to u, for all $\boldsymbol{x}$. The function $\boldsymbol{\psi}(u)$ has continuous third derivatives over $\mathcal{U}$.

(RC.3) There exist functions $M_k(\boldsymbol{x}, y)$, $k = 1, 2$, with $E_{(\boldsymbol{X},Y)}\{M_1^{m_0}(\boldsymbol{X}, Y)\} < \infty$ for some $m_0 > 2$, and $E_{(\boldsymbol{X},Y)}\{M_2(\boldsymbol{X}, Y)\} < \infty$, and $\int_{\boldsymbol{x},y} M_1(\boldsymbol{x}, y)d\boldsymbol{x}dy < \infty$, such that for all $\boldsymbol{x}$ and y, and all $\boldsymbol{\psi}(u)$ in a neighbourhood of $\boldsymbol{\psi}^0(u)$,

$$\left|\frac{\partial l(\boldsymbol{\psi}, \boldsymbol{x}, y)}{\partial \psi_j}\right| < M_1(\boldsymbol{x}, y) \ , \quad \left|\frac{\partial^2 l(\boldsymbol{\psi}, \boldsymbol{x}, y)}{\partial \psi_j \partial \psi_k}\right| < M_1(\boldsymbol{x}, y),$$

and

$$\left|\frac{\partial^3 l(\boldsymbol{\psi}, \boldsymbol{x}, y)}{\partial \psi_j \partial \psi_k \partial \psi_l}\right| < M_2(\boldsymbol{x}, y).$$

(RC.4) The matrix $\boldsymbol{I}(u)$ in (12) of the paper is continuous in u and positive definite for all $u \in \mathcal{U}$; the vector $\Delta(u; u_t)$ in (13) of the paper is continuously third time differentiable with respect to u over $\mathcal{U}$.

(RC.5) The conditions

$$E_0\left(\left|\frac{\partial l(\boldsymbol{\psi}, \boldsymbol{X}, Y)}{\partial \psi_j}\right|^3\right) < \infty \quad \text{and} \quad E_0\left(\left|\frac{\partial^2 l(\boldsymbol{\psi}, \boldsymbol{X}, Y)}{\partial \psi_i \partial \psi_j}\right|^{2+\delta}\right) < \infty$$

hold for all i and j and some $\delta > 0$, where E_0 is the expectation with respect to the joint distribution of $(Y, \boldsymbol{X}, U)$.

Web Appendix B.2 *Conditions on the kernel K*

(KC.1) K is a Lipschitz continuous and symmetric pdf with a compact support $\mathcal{U}$.

(KC.2) $\mathcal{K}_i = \int_{\mathcal{U}} t^i K(t)dt$, $\mathcal{V}_i = \int_{\mathcal{U}} t^i K^2(t)dt$, $i = 0, \ldots, 4$, are finite; for i odd they are zero.

Web Appendix B.3 *Conditions on the penalty p_n, and the tuning parameters (λ_n, h_n)*

With regard to the penalty function p_n, we denote the quantities

$$r_{1n} = \max_{\substack{1\leqslant l\leqslant d_j^0 \\ 1\leqslant j\leqslant C}} \left\{ \frac{h_n^{1/2}|p_n'(\theta_{jl}^0;\lambda_n)|}{n^{3/2}} : \theta_{jl}^0 = \sqrt{E_0\{\beta_{jl}^0(U_t)\}^2} \neq 0 \right\}, \text{ and}$$

$$r_{2n} = \max_{\substack{1\leqslant l\leqslant d_j^0 \\ 1\leqslant j\leqslant C}} \left\{ \frac{|p_n''(\theta_{jl}^0;\lambda_n)|}{n^{2}} : \theta_{jl}^0 = \sqrt{E_0\{\beta_{jl}^0(U_t)\}^2} \neq 0 \right\},$$

where $p_n'(\theta_{jl}^0;\lambda_n)$ and $p_n''(\theta_{jl}^0;\lambda_n)$ are the first and second derivatives of $p_n(\theta_{jl};\lambda_n)$ evaluated at $\theta_{jl} = \theta_{jl}^0 \neq 0$.

(PC.1) $p_n(\theta;\lambda_n)$ is nonnegative and symmetric in $\theta \in \mathbb{R}$, and $p_n(0;\lambda_n) = 0$. It is also nondecreasing and twice continuously differentiable for all but finitely many values $\theta \in (0,\infty)$.

(PC.2) As $n \to \infty$, $\lambda_n = o(1)$ such that $\min_{j,l}\theta_{jl}^0/\lambda_n \to \infty$; and $h_n = o(1)$ such that $nh_n \to \infty$, and $nh_n^5 = O(1)$. Also, $r_{1n} = o(\sqrt{nh_n})$ and $r_{2n} = o(1)$.

(PC.3) Let $N_{h_n} = \left\{\theta;\ 0 < |\theta| \leqslant \frac{\log(nh_n)}{\sqrt{nh_n}}\right\}$. Then, $\lim_{n\to\infty}\inf_{\theta\in N_{h_n}} h_n^{1/2}\ \frac{p_n'(\theta;\lambda_n)}{n^{3/2}} = \infty$.

The smoothness Condition (RC.1) facilitates obtaining estimators of $\beta_{jl}(\cdot)$ by differentiating $\widetilde{L}_n(\boldsymbol{\Psi};\boldsymbol{\lambda},h)$ and for studying the asymptotic properties of the estimators of the true non-zero $\beta_{jl}(\cdot)$. (PC.2) is required to obtain $\sqrt{nh_n}$-consistent estimators of the true non-zero $\beta_{jl}(\cdot)$, while (PC.3) is required for sparsity.

Web Appendix C. Examples of the penalty function p_n and the kernel K

We used the following penalties and kernel function in our simulations and analysis of the real data.

- LASSO: $p_n(\theta;\lambda)/n^2 = \lambda|\theta|$;
- AdpLASSO: $p_n(\theta;\lambda)/n^2 = \lambda w|\theta|$, for some (possibly random) known weights w;
- MCP: $p_n'(\theta;\lambda)/n^2 = \text{sgn}(\theta)\ \frac{(a\lambda-|\theta|)_+}{a}$;

- SCAD: $p_n'(\theta;\lambda)/n^2 = \text{sgn}(\theta)\ \left\{\lambda I(|\theta| \leqslant \lambda) + \frac{(a\lambda - |\theta|)_+}{a-1} I(|\theta| > \lambda)\right\}$,

where $p_n'(\cdot;\lambda)$ is the first derivative of a penalty with respect to θ, and $(x)_+ = \max\{0, x\}$.

The Epanechnikov Kernel: $K(t) = \frac{3}{4}(1-u^2)_+$.

Web Appendix D. Proofs

The result of the following lemma is used in the sequel.

LEMMA 1: **(Barlett's first and second identities)** *Under Conditions (RC.1)-(RC.4) in Web Appendix B, and for any interior point u_t of $\mathcal{U}$, it holds that*

$$E_{(Y,\boldsymbol{X})|U}\left[l'(\boldsymbol{\psi}^0(U), \boldsymbol{X}, Y)|U = u_t\right] = 0,$$

and

$$\begin{aligned} &E_{(Y,\boldsymbol{X})|U}\left[l''(\boldsymbol{\psi}^0(U), \boldsymbol{X}, Y)|U = u_t\right] \\ &\quad = -E_{(Y,\boldsymbol{X})|U}\left[l'(\boldsymbol{\psi}^0(U), \boldsymbol{X}, Y) l'^{\top}(\boldsymbol{\psi}^0(U), \boldsymbol{X}, Y)|U = u_t\right] = \boldsymbol{I}(u_t). \end{aligned}$$

Proof. Conditioning on $U = u_t$, we prove Barlett's first identity as follows.

$$\begin{aligned} &E_{(Y,\boldsymbol{X})|U}\left[l'(\boldsymbol{\psi}^0(U), \boldsymbol{X}, Y)|U = u_t\right] \\ &\quad = \int_{\mathcal{Y}}\int_{\mathcal{X}} l'(\boldsymbol{\psi}^0(u_t), \boldsymbol{x}, y) f_C^*(y|\boldsymbol{\psi}^0(u_t), \boldsymbol{x}) g(\boldsymbol{x}|u_t) d\boldsymbol{x} dy \\ &\quad = \int_{\mathcal{Y}}\int_{\mathcal{X}} \frac{\partial l(\boldsymbol{\psi}, \boldsymbol{x}, y)}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0(u_t)} f_C^*(y|\boldsymbol{\psi}^0(u_t), \boldsymbol{x}) g(\boldsymbol{x}|u_t) d\boldsymbol{x} dy \\ &\quad = \int_{\mathcal{Y}}\int_{\mathcal{X}} \frac{\partial \log f_C^*(y|\boldsymbol{\psi}, \boldsymbol{x})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0(u_t)} f_C^*(y|\boldsymbol{\psi}^0(u_t), \boldsymbol{x}) g(\boldsymbol{x}|u_t) d\boldsymbol{x} dy \\ &\quad = \int_{\mathcal{Y}}\int_{\mathcal{X}} \frac{\partial f_C^*(y|\boldsymbol{\psi}, \boldsymbol{x})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0(u_t)} g(\boldsymbol{x}|u_t) d\boldsymbol{x} dy \\ &\quad = \frac{\partial}{\partial \boldsymbol{\psi}}\int_{\mathcal{Y}}\int_{\mathcal{X}} \left[\sum_{c=1}^{C} \pi_c f(y; \theta_c(\boldsymbol{x}), \phi_c)\right] g(\boldsymbol{x}|u_t) d\boldsymbol{x} dy\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0(u_t)} \\ &\quad = \frac{\partial}{\partial \boldsymbol{\psi}}(1) = 0. \end{aligned}$$

For a proof of Barlett's second identity note that

$$
\begin{aligned}
&- E_{(Y,\boldsymbol{X})|U}\left[l''(\boldsymbol{\psi}^0(U),\boldsymbol{X},Y)|U=u_t\right] \\
&= -\int_{\mathcal{Y}}\int_{\mathcal{X}} \frac{\partial^2 \log f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0(u_t)} f_C^*(y|\boldsymbol{\psi}^0(u_t),\boldsymbol{x})g(\boldsymbol{x}|u_t)d\boldsymbol{x}dy \\
&= -\int_{\mathcal{Y}}\int_{\mathcal{X}} \frac{\frac{\partial^2 f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top} f_C^*(y|\boldsymbol{\psi},\boldsymbol{x}) - \left\{\frac{\partial f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})}{\partial\boldsymbol{\psi}}\right\}\left\{\frac{\partial f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})}{\partial\boldsymbol{\psi}}\right\}^\top}{\left\{f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})\right\}^2}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0(u_t)} \\
&\qquad \times f_C^*(y|\boldsymbol{\psi}^0(u_t),\boldsymbol{x})g(\boldsymbol{x}|u_t)d\boldsymbol{x}dy \\
&= -\int_{\mathcal{Y}}\int_{\mathcal{X}} \frac{\partial^2}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}\left\{f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})\right\}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0(u_t)} g(\boldsymbol{x}|u_t)d\boldsymbol{x}dy \\
&\quad + \int_{\mathcal{Y}}\int_{\mathcal{X}} \frac{\left\{\frac{\partial}{\partial\boldsymbol{\psi}} f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\psi}} f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})\right\}^\top}{\left\{f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})\right\}^2}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0(u_t)} f_C^*(y|\boldsymbol{\psi}^0(u_t),\boldsymbol{x})g(\boldsymbol{x}|u_t)d\boldsymbol{x}dy \\
&= -\frac{\partial^2}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}\int_{\mathcal{Y}}\int_{\mathcal{X}} \left\{f_C^*(y|\boldsymbol{\psi},\boldsymbol{x})\right\}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0(u_t)} g(\boldsymbol{x}|u_t)d\boldsymbol{x}dy \\
&\quad + \int_{\mathcal{Y}}\int_{\mathcal{X}} l'(\boldsymbol{\psi}^0(u_t),\boldsymbol{x},y)l'^\top(\boldsymbol{\psi}^0(u_t),\boldsymbol{x},y)f_C^*(y|\boldsymbol{\psi}^0(u_t),\boldsymbol{x})g(\boldsymbol{x}|u_t)d\boldsymbol{x}dy \\
&= E_{(Y,\boldsymbol{X})|U}\left\{l'(\boldsymbol{\psi}^0(u_t),\boldsymbol{X},Y)l'^\top(\boldsymbol{\psi}^0(u_t),\boldsymbol{X},Y)|U=u_t\right\}.
\end{aligned}
$$

We now provide the proofs of our main results, starting with a proposition on point-wise consistency of the maximum local-kernel log-likelihood estimator (MLLE) of the parameter vector of $\boldsymbol{\psi}(u)$ in the FM-VCR model in (3) of the paper .

PROPOSITION 1: Suppose that Regularity Conditions (RC.1)-(RC.4) on the parametric family and Conditions (KC.1)-(KC.2) on the kernel K stated in Web Appendix B of this document hold, and $h_n = o(1)$ such that $nh_n \to \infty$, and $nh_n^5 = O(1)$. Then for any $u \in \mathcal{U}$, there exists a local maximizer $\breve{\boldsymbol{\psi}}_n(u)$ of $\ell_n\big(\boldsymbol{\psi}(u); h_n\big)$ in (7) of the paper such that: $\|\breve{\boldsymbol{\psi}}_n(u) - \boldsymbol{\psi}^0(u)\|_2 = O_p\{(nh_n)^{-1/2}\}$.

Proof of Proposition 1

Proof. The proof is a special case of the proof of Theorem 1 (without the penalty) and thus omitted.

Proof of Theorem 1

Proof. First note that, for any $u \in \mathcal{U}$,

$$\boldsymbol{\psi}(u) = \left(\boldsymbol{\pi}^\top(u), \boldsymbol{\phi}^\top(u), \boldsymbol{\beta}_1^\top(u), \ldots, \boldsymbol{\beta}_C^\top(u)\right)^\top \in \mathbb{R}^{(C(d+2)-1)\times 1},$$

and whenever necessary, we rewrite

$$\boldsymbol{\psi}(u) = \left(\psi_1(u), \ldots, \psi_{C(d+2)-1}(u)\right)^\top$$

without changing the order of $\boldsymbol{\phi}^\top(u), \boldsymbol{\pi}^\top(u), \boldsymbol{\beta}_1^\top(u), \ldots, \boldsymbol{\beta}_C^\top(u)$. Otherwise, we will use the same notation as defined in Section 2 of the paper . Let

$$\boldsymbol{w}_{j,I} = \begin{pmatrix} w_{j1,I,1} & \ldots & w_{jd,I,1} \\ \vdots & \ddots & \vdots \\ w_{j1,I,n} & \ldots & w_{jd,I,n} \end{pmatrix} \in \mathbb{R}^{n\times d},$$

and the vectors

$$\boldsymbol{w}_{j,II} = (w_{j,II,1}, \ldots, w_{j,II,n})^\top \in \mathbb{R}^{n\times 1}, \quad j = 1, \ldots, C,$$

$$\boldsymbol{w}_{j,III} = (w_{j,III,1}, \ldots, w_{j,III,n})^\top \in \mathbb{R}^{n\times 1}, \quad j = 1, \ldots, C-1.$$

We set

$$\boldsymbol{W}_n = (\boldsymbol{w}_{1,I}, \ldots, \boldsymbol{w}_{C,I}, \boldsymbol{w}_{1,II}, \ldots, \boldsymbol{w}_{C,II}, \boldsymbol{w}_{1,III}, \ldots, \boldsymbol{w}_{C-1,III}) \in \mathbb{R}^{n\times(C(d+2)-1)}$$

which is an arbitrary matrix with rows $\boldsymbol{w}_t^\top$, and $\boldsymbol{w}_{j,I,t}$ is denoted as the t-th row of $\boldsymbol{w}_{j,I}$, for $t = 1, \ldots, n$. The first Cd columns of $\boldsymbol{w}$ are denoted by $\boldsymbol{v}_{11}, \ldots, \boldsymbol{v}_{1d}, \ldots, \boldsymbol{v}_{C1}, \ldots, \boldsymbol{v}_{Cd}$. For an arbitrary matrix $\boldsymbol{A} = (a_{ij})$, we define its L_2 norm (i.e. the Frobenius norm) as

$$\|\boldsymbol{A}\|_{\mathrm{F}} = \left(\sum a_{ij}^2\right)^{1/2}.$$

Note that $\|\boldsymbol{W}_n\|_{\mathrm{F}}^2 = \sum_{t=1}^n \|\boldsymbol{w}_t\|^2$ and $\sum_{j=1}^C \sum_{l=1}^d \|\boldsymbol{v}_{jl}\|^2 \leqslant \|\boldsymbol{W}_n\|_{\mathrm{F}}^2$. Whenever necessary, we rewrite $\boldsymbol{w}_t = (w_{1,t}, \ldots, w_{C(d+2)-1,t})^\top$; otherwise, we will use the same notation as defined by

$$\boldsymbol{w}_t = (w_{11,I,t}, \ldots, w_{1d,I,t}, \ldots, w_{C1,I,t}, \ldots, w_{Cd,I,t}, w_{1,II,t}, \ldots, w_{C,II,t}, w_{1,III,t}, \ldots, w_{C-1,III,t})^\top.$$

To prove the claim of the theorem, it suffices to show that for any small $\epsilon > 0$, there exists a constant M_ϵ such that

$$\lim_{n\to\infty} P\left\{ \sup_{n^{-1}\|\boldsymbol{W}_n\|_2^2 = M_\epsilon^2} \widetilde{L}_n\left(\boldsymbol{\Psi}_n^0 + \gamma_n \boldsymbol{W}_n; \boldsymbol{\lambda}_n, h_n\right) < \widetilde{L}_n\left(\boldsymbol{\Psi}_n^0; \boldsymbol{\lambda}_n, h_n\right) \right\} \geqslant 1 - \epsilon, \tag{A.1}$$

where $\gamma_n = (1 + r_{1n})(nh_n)^{-1/2}$.

To show (A.1), we proceed as follows. Let

$$D_n(\boldsymbol{W}_n) = h_n n^{-1} \left\{ \widetilde{L}_n\left(\boldsymbol{\Psi}_n^0 + \gamma_n \boldsymbol{W}_n; \boldsymbol{\lambda}_n, h_n\right) - \widetilde{L}_n\left(\boldsymbol{\Psi}_n^0; \boldsymbol{\lambda}_n, h_n\right) \right\}.$$

By the definition of penalized local log-likelihood $\widetilde{L}_n$ in (9) of the paper, we have

$$\widetilde{L}_n\left(\boldsymbol{\Psi}_n^0 + \gamma_n \boldsymbol{W}_n; \boldsymbol{\lambda}_n, h_n\right) = \sum_{t=1}^n \sum_{i=1}^n l\left(\boldsymbol{\psi}^0(U_t) + \gamma_n \boldsymbol{w}_t, \boldsymbol{X}_i, Y_i\right) K_{h_n}(U_i - U_t) - \mathbb{P}_n(\boldsymbol{\Psi}_n^0 + \gamma_n \boldsymbol{W}_n; \boldsymbol{\lambda}_n)$$

and

$$\widetilde{L}_n\left(\boldsymbol{\Psi}_n^0; \boldsymbol{\lambda}_n, h_n\right) = \sum_{t=1}^n \sum_{i=1}^n l\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i\right) K_{h_n}(U_i - U_t) - \mathbb{P}_n(\boldsymbol{\Psi}_n^0; \boldsymbol{\lambda}_n),$$

where $l(\boldsymbol{\psi}^0(u), \boldsymbol{x}, y) = \log\left\{f_C^*\left(y | \boldsymbol{\psi}^0(u), \boldsymbol{x}\right)\right\}$, f_C^* is the mixture density given in (3) of the paper, and the penalty

$$\mathbb{P}_n(\boldsymbol{\Psi}_n; \boldsymbol{\lambda}_n) = \sum_{j=1}^C \sum_{l=1}^d p_n(\|\boldsymbol{b}_{jl}\|_2 / \sqrt{n}; \lambda_{nj}).$$

By Condition (PC.1) on the penalty, we have $p_n(0; \lambda_{nj}) = 0$. Thus,

$$\begin{aligned}
D_n(\boldsymbol{W}_n) &= h_n n^{-1} \sum_{t=1}^n \sum_{i=1}^n \left\{ l\left(\boldsymbol{\psi}^0(U_t) + \gamma_n \boldsymbol{w}_t, \boldsymbol{X}_i, Y_i\right) - l\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i\right) \right\} K_{h_n}(U_i - U_t) \\
&\quad - h_n n^{-1} \sum_{j=1}^C \sum_{l=1}^d \left\{ p_n\left(\left\|\boldsymbol{b}_{jl}^0 + \gamma_n \boldsymbol{v}_{jl}\right\|_2 / \sqrt{n}; \lambda_{nj}\right) - p_n\left(\left\|\boldsymbol{b}_{jl}^0\right\|_2 / \sqrt{n}; \lambda_{nj}\right) \right\} \\
&\leqslant h_n n^{-1} \sum_{t=1}^n \sum_{i=1}^n \left\{ l\left(\boldsymbol{\psi}^0(U_t) + \gamma_n \boldsymbol{w}_t, \boldsymbol{X}_i, Y_i\right) - l\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i\right) \right\} K_{h_n}(U_i - U_t) \\
&\quad - h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \left\{ p_n\left(\left\|\boldsymbol{b}_{jl}^0 + \gamma_n \boldsymbol{v}_{jl}\right\|_2 / \sqrt{n}; \lambda_{nj}\right) - p_n\left(\left\|\boldsymbol{b}_{jl}^0\right\|_2 / \sqrt{n}; \lambda_{nj}\right) \right\} \\
&= D_{n,I}(\boldsymbol{W}_n) - D_{n,II}(\boldsymbol{W}_n),
\end{aligned} \tag{A.2}$$

where $D_{n,I}(\boldsymbol{W}_n)$ and $D_{n,II}(\boldsymbol{W}_n)$ are respectively the differences in the local-kernel log-likelihood and the penalty functions. In what follows, we first perform an order assessment of the two differences for large n, in two steps:

Step 1: (Order assessment of the local log-likelihood difference)

Following the first part of (A.2), we rewrite the difference as

$$D_{n,I}(\boldsymbol{W}_n) = n^{-1}\sum_{t=1}^{n} d_{n,I}(U_t, \boldsymbol{w}_t), \tag{A.3}$$

where for each $t = 1, \ldots, n$,

$$d_{n,I}(U_t, \boldsymbol{w}_t) = h_n \sum_{i=1}^{n} \left\{ l\left(\boldsymbol{\psi}^0(U_t) + \gamma_n \boldsymbol{w}_t, \boldsymbol{X}_i, Y_i\right) - l\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i\right) \right\} K_{h_n}(U_i - U_t).$$

By the differentiability Condition (RC.2) on f_C^* and using a third-order Taylor expansion,

$$\begin{aligned} d_{n,I}(U_t, \boldsymbol{w}_t) \;=\; & h_n \sum_{i=1}^{n} \Bigg\{ \gamma_n \boldsymbol{w}_t^\top l'\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i\right) + \frac{1}{2}\gamma_n^2 \left[\boldsymbol{w}_t^\top\, l''\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i\right)\, \boldsymbol{w}_t\right] \\ & + \frac{1}{6}\gamma_n^3 \sum_{j,k,l=1}^{\mathcal{C}(d+2)-1} \frac{\partial^3}{\partial\psi_j \partial\psi_k \partial\psi_l} l(\widetilde{\boldsymbol{\psi}}^0(U_t), \boldsymbol{X}_i, Y_i)\; w_{j,t} w_{k,t} w_{l,t} \Bigg\} K_{h_n}(U_i - U_t), \end{aligned}$$

where $\widetilde{\boldsymbol{\psi}}^0(U_t)$ is between $\boldsymbol{\psi}^0(U_t)$ and $\boldsymbol{\psi}^0(U_t) + \gamma_n \boldsymbol{w}_t$, for $t = 1, \ldots, n$. We then have

$$d_{n,I}(U_t, \boldsymbol{w}_t) = \boldsymbol{w}_t^\top l'_{n,h_n}(U_t) + \frac{1}{2}\boldsymbol{w}_t^\top \{l''_{n,h_n}(U_t)\}\boldsymbol{w}_t + \frac{h_n\gamma_n^3}{6}\sum_{i=1}^{n} R\left(\widetilde{\boldsymbol{\psi}}^0(U_t), \boldsymbol{X}_i, Y_i\right), \tag{A.4}$$

where

$$l'_{n,h_n}(U_t) = \sqrt{\frac{h_n}{n}}(1 + r_{1n}) \sum_{i=1}^{n} l'\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i\right) K_{h_n}(U_i - U_t), \tag{A.5}$$

$$l''_{n,h_n}(U_t) = \frac{1}{n}(1 + r_{1n})^2 \sum_{i=1}^{n} l''(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - U_t)$$

and

$$R\left(\widetilde{\boldsymbol{\psi}}^0(U_t), \boldsymbol{X}_i, Y_i\right) = \sum_{j,k,l=1}^{\mathcal{C}(d+2)-1} \frac{\partial^3}{\partial\psi_j \partial\psi_k \partial\psi_l} l\left(\widetilde{\boldsymbol{\psi}}^0(U_t), \boldsymbol{X}_i, Y_i\right) K_{h_n}(U_i - U_t) w_{j,t} w_{k,t} w_{l,t}.$$

We now perform an order assessment of the three terms in (A.4), by first focusing on $l'_{n,h_n}(U_t)$. Under our regularity conditions, we will show that $l'_{n,h_n}(U_t) = O_p(1)$. Using Prohorov's Theorem and Example 2.6 on pages 8-10 of van der Vaart (1998), it is sufficient to show that $E_0\|l'_{n,h_n}(U_t)\|^2 = O(1)$, where E_0 is the expectation with respect to the true joint

distribution of $(U_t, U_i, \boldsymbol{X}_i, Y_i)$. Note that,

$$
\begin{aligned}
& E_0\|l'_{n,h_n}(U_t)\|^2 \\
&= E_0\left\{\left\|(1+r_{1n})\sqrt{\frac{h_n}{n}}\sum_{i=1}^{n} l'(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i-U_t)\right\|^2\right\} \\
&= (1+r_{1n})^2\frac{h_n}{n}E_0\left\{\left\|\sum_{i=1}^{n} l'(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i-U_t)\right\|^2\right\} \\
&= (1+r_{1n})^2\frac{h_n}{n}\sum_{i=1}^{n} E_0\left\{\left\|l'\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i\right)\right\|^2 K_{h_n}^2(U_i-U_t)\right\} + 2(1+r_{1n})^2\frac{h_n}{n} \\
&\quad \times \sum_{i<j}^{n} E_0\left\{\left[l'(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i)\right]^\top\left[l'\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_j, Y_j\right)\right]K_{h_n}(U_i-U_t)K_{h_n}(U_j-U_t)\right\} \\
&= (1+r_{1n})^2\frac{h_n}{n}\sum_{i=1}^{n} E_0[\mathcal{Q}_{i,t}] + 2(1+r_{1n})^2\frac{h_n}{n}\sum_{i<j}^{n} E_0[\mathcal{P}_{i,j,t}], \qquad \text{(A.6)}
\end{aligned}
$$

where $\mathcal{Q}_{i,t}$ and $\mathcal{P}_{i,j,t}$ are the quadratic and cross-product terms, respectively. The s^{th} element of the vector $l'(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i)$ is denoted by $l'^{(s)}(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i)$, where $s = 1, 2, \ldots, d^* = \mathcal{C}(d+2) - 1$.

We now focus on $E_0[\mathcal{Q}_{i,t}]$ for the cases $i = t$ and $i \neq t$. When $i = t$, by the definition of the conditional information matrix $\boldsymbol{I}(u_t)$ in (12) of the paper and denoting $\mathrm{I}^{(s,s)}(u_t)$ as its $(s,s)^{\text{th}}$ element, we obtain

$$
\begin{aligned}
E_0[\mathcal{Q}_{t,t}] &= E_0\left\{\left\|l'\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t\right)\right\|^2 K_{h_n}^2(0)\right\} \\
&= K_{h_n}^2(0)\sum_{s=1}^{d^*} E_0\left\{l'^{(s)}(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t)\right\}^2 \\
&= K_{h_n}^2(0)\sum_{s=1}^{d^*} E_U\left\{E_{(\boldsymbol{X},Y)|U_t}\left\{[l'^{(s)}(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t)]^2 \mid U_t\right\}\right\} \\
&= K_{h_n}^2(0)\sum_{s=1}^{d^*} E_U\left\{I^{(s,s)}(U_t)\right\}.
\end{aligned}
$$

By (RC.4), the expected value in the last sum is finite. Thus, for some constant $M > 0$,

$$
E_0[\mathcal{Q}_{t,t}] \leqslant K_{h_n}^2(0)Md^* = \frac{1}{h_n^2}K^2(0)Md^* = \frac{C_0}{h_n^2}. \qquad \text{(A.7)}
$$

For $i \neq t$, we have

$$\begin{aligned} E_0[\mathcal{Q}_{i,t}] &= E_0\left\{\left\|l'(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i)\right\|^2 K_{h_n}^2(U_i - U_t)\right\} \\ &= \sum_{s=1}^{d^*} E_0\left\{l'^{(s)}(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - U_t)\right\}^2. \end{aligned}$$

Next, we evaluate the expected value inside the sum in the above equation. Note that $(\boldsymbol{X}_i, Y_i, U_i)$ and U_t are independent when $i \neq t$. Thus,

$$\begin{aligned} & E_0\left\{l'^{(s)}(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - U_t)\right\}^2 \\ & \leqslant 2E_0\left\{l'^{(s)}(\boldsymbol{\psi}^0(U_i), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - U_t)\right\}^2 \\ & \quad + 2E_0\left\{\left[l'^{(s)}(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i) - l'^{(s)}(\boldsymbol{\psi}^0(U_i), \boldsymbol{X}_i, Y_i)\right] K_{h_n}(U_i - U_t)\right\}^2 \\ & = \mathcal{T}_{n,1} + \mathcal{T}_{n,2}, \end{aligned}$$

where $\mathcal{T}_{n,1}$ and $\mathcal{T}_{n,2}$ respectively represent the two expectations on the right hand side of the above inequality. Order assessment of each term is given below.

By Condition (RC.4) on $\boldsymbol{I}(u)$, it follows that for some constant $M > 0$,

$$\begin{aligned} \mathcal{T}_{n,1} &= 2\int_{\mathcal{U}}\int_{\mathcal{U}}\int_{\mathcal{X}}\int_{\mathcal{Y}}\{l'^{(s)}(\boldsymbol{\psi}^0(u_i), \boldsymbol{x}_i, y_i)\}^2 f_C^*(y_i|\boldsymbol{\psi}^0(u_i), \boldsymbol{x}_i) g(\boldsymbol{x}_i|u_i) \\ &\qquad \times K_{h_n}^2(u_i - u_t) m(u_i) m(u_t) dy_i d\boldsymbol{x}_i du_i du_t \\ &= 2\int_{\mathcal{U}}\int_{\mathcal{U}} I^{(s,s)}(u_i) K_{h_n}^2(u_i - u_t) m(u_i) m(u_t) du_i du_t \\ &\leqslant 2M\int_{\mathcal{U}}\int_{\mathcal{U}} K_{h_n}^2(u_i - u_t) m(u_i) m(u_t) du_i du_t. \end{aligned}$$

Applying the change of variable $u_i^* = (u_i - u_t)/h_n$ to the above integration, we have

$$\mathcal{T}_{n,1} \leqslant 2Mh_n^{-1}\int_{\mathcal{U}}\int_{\mathcal{U}^*} K^2(u_i^*) m(u_t + h_n u_i^*) m(u_t) du_i^* du_t.$$

Using a second-order Taylor expansion,

$$m(u_t + h_n u_i^*) = m(u_t) + h_n u_i^*\, m'(u_t) + (h_n u_i^*)^2\, m''(\tilde{u}_{i,t}), \tag{A.8}$$

where $\tilde{u}_{i,t}$ is between u_t and $u_t + h_n u_i^*$. Using the boundedness Condition (RC.2) of $m(u)$ and its first and second derivatives over $\mathcal{U}$, and the finite-moment Condition (KC.1) of $K^2(u)$,

for some positive constants C_1 and C_2, we arrive at

$$\mathcal{T}_{n,1} \leqslant C_1 h_n^{-1} + C_2 h_n. \tag{A.9}$$

Using a first-order Taylor expansion and by Conditions (RC.2) and (RC.3) for some constants C_3 and C_4,

$$\begin{aligned}
\mathcal{T}_{n,2} &= 2E_0\left\{\left[l'^{(s)}(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i) - l'^{(s)}(\boldsymbol{\psi}^0(U_i), \boldsymbol{X}_i, Y_i)\right] K_{h_n}(U_i - U_t)\right\}^2 \\
&\leqslant C_3 E_0\left\{M_1(\boldsymbol{X}_i, Y_i)(U_t - U_i) K_{h_n}(U_i - U_t)\right\}^2 \\
&= C_3 \int_{\mathcal{U}}\int_{\mathcal{U}}\left[\int_{\mathcal{X}}\int_{\mathcal{Y}} M_1^2(\boldsymbol{x}_i, y_i) f_C^*(y_i|\boldsymbol{\psi}^0(u_i), \boldsymbol{x}_i) g(\boldsymbol{x}_i|u_i) dy_i d\boldsymbol{x}_i\right](u_i - u_t)^2 \\
&\quad \times K_{h_n}^2(u_i - u_t) m(u_i) m(u_t) du_i du_t \\
&\leqslant C_3 C_4 \int_{\mathcal{U}}\int_{\mathcal{U}} (u_i - u_t)^2 K_{h_n}^2(u_i - u_t) m(u_i) m(u_t) du_i du_t.
\end{aligned}$$

Applying the change of variable $u_i^* = (u_i - u_t)/h_n$ and the expansion (A.8), we have

$$\begin{aligned}
\mathcal{T}_{n,2} \leqslant C_3 C_4 \Bigg\{ & h_n \int_{\mathcal{U}}\int_{\mathcal{U}^*} (u_i^*)^2 K^2(u_i^*) m^2(u_t) du_i^* du_t \\
& + h_n^2 \int_{\mathcal{U}}\int_{\mathcal{U}^*} (u_i^*)^3 K^2(u_i^*) m'(u_t) m(u_t) du_i^* du_t \\
& + h_n^3 \int_{\mathcal{U}}\int_{\mathcal{U}^*} (u_i^*)^4 K^2(u_i^*) m''(\tilde{u}_{i,t}) m(u_t) du_i^* du_t \Bigg\}.
\end{aligned}$$

Thus, by Condition (RC.2) on $m(u)$ and its first and second derivatives, and the finite-moment Condition (KC.1) of $K^2(u)$, for some positive constants C_5 and C_6, we arrive at

$$\mathcal{T}_{n,2} \leqslant C_5 h_n + C_6 h_n^3. \tag{A.10}$$

Hence, using (A.7), (A.9), and (A.10), the first sum on the right hand side of (A.6) becomes

$$\begin{aligned}
(1 + r_{1n})^2 \frac{h_n}{n} \sum_{i=1}^{n} E_0[\mathcal{Q}_{i,t}] &= (1 + r_{1n})^2 \frac{h_n}{n}\left\{\frac{C_0}{h_n^2} + (n-1)\left(C_1 h_n^{-1} + C_2 h_n + C_5 h_n + C_6 h_n^3\right)\right\} \\
&= (1 + r_{1n})^2 \left\{\frac{C_0}{n h_n} + \frac{(n-1)}{n}\left[C_1 + (C_2 + C_5) h_n^2 + C_6 h_n^4\right]\right\}.
\end{aligned}$$

Under Condition (PC.2), we have $h_n \to 0$ and $n h_n \to \infty$, as $n \to \infty$. Thus,

$$(1 + r_{1n})^2 \frac{h_n}{n} \sum_{i=1}^{n} E_0[\mathcal{Q}_{i,t}] = O\{(1 + r_{1n})^2\}. \tag{A.11}$$

We now assess the second sum in (A.6). The expectation of $\mathcal{P}_{i,j,t}$, for $i > j$, $i \neq t$ and $j \neq t$

is given by

$$\begin{aligned} E_0[\mathcal{P}_{i,j,t}] &= E_0\left\{[l'(\boldsymbol{\psi}^0(U_t),\boldsymbol{X}_i,Y_i)]^\top l'(\boldsymbol{\psi}^0(U_t),\boldsymbol{X}_j,Y_j)K_{h_n}(U_i-U_t)K_{h_n}(U_j-U_t)\right\} \\ &= \sum_{s=1}^{d^*} E_0\{l'^{(s)}(\boldsymbol{\psi}^0(U_t),\boldsymbol{X}_i,Y_i)l'^{(s)}(\boldsymbol{\psi}^0(U_t),\boldsymbol{X}_j,Y_j)K_{h_n}(U_i-U_t)K_{h_n}(U_j-U_t)\} \\ &= \sum_{s=1}^{d^*} E_{U_t}\left\{E_{(Y,\boldsymbol{X},U_j,U_i)|U_t}\left\{l'^{(s)}(\boldsymbol{\psi}^0(U_t),\boldsymbol{X}_i,Y_i)l'^{(s)}(\boldsymbol{\psi}^0(U_t),\boldsymbol{X}_j,Y_j)\right.\right. \\ &\quad \left.\left.\times\ K_{h_n}(U_i-U_t)K_{h_n}(U_j-U_t)\right\}\mid U_t\right\}. \end{aligned} \tag{A.12}$$

Conditioning on U_t, since $i \neq j$, $(U_i, \boldsymbol{X}_i, Y_i)$ and $(U_j, \boldsymbol{X}_j, Y_j)$ are independent and identically distributed. In what follows, we first evaluate the conditional expectations.

$$\begin{aligned} &E_{(Y,\boldsymbol{X},U_i,U_j)|U_t}\left\{l'^{(s)}(\boldsymbol{\psi}^0(u_t),\boldsymbol{X}_i,Y_i)l'^{(s)}(\boldsymbol{\psi}^0(u_t),\boldsymbol{X}_j,Y_j)K_{h_n}(U_i-u_t)K_{h_n}(U_j-u_t)\mid U_t=u_t\right\} \\ &= E^2_{(Y,\boldsymbol{X},U_i)|U_t}\left\{l'^{(s)}(\boldsymbol{\psi}^0(u_t),\boldsymbol{X}_i,Y_i)K_{h_n}(U_i-u_t)\mid U_t=u_t\right\}, \quad i\neq t. \end{aligned}$$

Thus,

$$\begin{aligned} &E_{(Y,\boldsymbol{X},U_i)|U_t}\left\{l'^{(s)}(\boldsymbol{\psi}^0(u_t),\boldsymbol{X}_i,Y_i)K_{h_n}(U_i-u_t)|U_t=u_t\right\} \\ &= \int_{\mathcal{Y}}\int_{\mathcal{X}}\int_{\mathcal{U}} l'(\boldsymbol{\psi}^0(u_t),\boldsymbol{x}_i,y_i)K_{h_n}(u_i-u_t)f^*_C(y_i|\boldsymbol{\psi}^0(u_i),\boldsymbol{x}_i)g(\boldsymbol{x}_i|u_i)m(u_i)du_i d\boldsymbol{x}_i dy_i \\ &= \int_{\mathcal{U}}\left[\int_{\mathcal{Y}}\int_{\mathcal{X}} l'(\boldsymbol{\psi}^0(u_t),\boldsymbol{x}_i,y_i)f^*_C(y_i|\boldsymbol{\psi}^0(u_i),\boldsymbol{x}_i)g(\boldsymbol{x}_i|u_i)d\boldsymbol{x}_i dy_i\right]K_{h_n}(u_i-u_t)m(u_i)du_i \\ &= \int_{\mathcal{U}}\Delta(u_i;u_t)K_{h_n}(u_i-u_t)m(u_i)du_i \\ &= \int_{\mathcal{U}}K_{h_n}(u_i-u_t)\Delta(u_i;u_t)m(u_i)du_i, \end{aligned}$$

where $\Delta(u_i;u_t)$ is given in (13) of the paper. Note that by the first Bartlett identity verified in Lemma 1, $\Delta(u_t;u_t)=0$. Denote $H(u_i)=\Delta(u_i;u_t)m(u_i)$, then

$$E_{(Y,\boldsymbol{X},U_i)|U_t}\left\{l'^{(s)}\left(\boldsymbol{\psi}^0(u_t),\boldsymbol{X}_i,Y_i\right)K_{h_n}(U_i-u_t)|U_t=u_t\right\} = \int_{\mathcal{U}}K_{h_n}(u_i-u_t)H(u_i)du_i. \tag{A.13}$$

By the differentiability conditions of (f^*_C, g, m) in (RC.2), and Condition (RC.4) on Δ, the first and second derivatives of $H(u_i)$ with respect to u_i, evaluated at $u_i=u_t$, are

$$H'(u_t) = \Delta'(u_t;u_t)m(u_t)$$

and

$$H''(u_t) = \Delta''(u_t; u_t)m(u_t) + 2\Delta'(u_t; u_t)m'(u_t),$$

where $\Delta'(\cdot; u_t)$ and $\Delta''(\cdot; u_t)$ are the first and second derivatives of $\Delta(u_i; u_t)$ with respect to u_i. Similarly, the third derivative of $H(u_i)$ with respect to u_i is calculated. Using a third-order Taylor expansion,

$$H(u_i) = \Delta'(u_t; u_t)m(u_t)(u_i - u_t) + \frac{1}{2}\left\{\Delta''(u_t; u_t)m(u_t) + 2\Delta'(u_t; u_t)m'(u_t)\right\}(u_i - u_t)^2$$
$$+ \frac{1}{6}\left\{\Delta'''(\tilde{u}_t; u_t)m(\tilde{u}_t) + \Delta''(\tilde{u}_t; u_t)m'(\tilde{u}_t) + 2\Delta''(\tilde{u}_t; u_t)m'(\tilde{u}_t) + 2\Delta'(\tilde{u}_t; u_t)m''(\tilde{u}_t)\right\}(u_i - u_t)^3,$$

where $\tilde{u}_t$ is between u_t and u_i, and $\Delta'''(\cdot; u_t)$ is the third derivative of $\Delta(u_i; u_t)$ with respect to u_i. By replacing the above expansion in (A.13), and Condition (KC.2) on the kernel $K(\cdot)$,

$$E_{(Y,\boldsymbol{X},U_i)|U_t}\left\{l'^{(s)}(\boldsymbol{\psi}^0(u_t), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i - u_t)|U_t = u_t\right\}$$
$$= \frac{1}{2}\mathcal{K}_2\left\{\Delta''(u_t; u_t)m(u_t) + 2\Delta'(u_t; u_t)m'(u_t)\right\}h_n^2\{1 + o(1)\}. \tag{A.14}$$

Thus,

$$\begin{aligned} E_0[\mathcal{P}_{i,j,t}] &= \{1 + o(1)\}^2\frac{h_n^4\mathcal{K}_2^2}{4}\sum_{s=1}^{d^*}\int_{\mathcal{U}}\left\{\Delta''(u_t; u_t)m(u_t) + 2\Delta'(u_t; u_t)m'(u_t)\right\}^2 m(u_t)du_t \\ &= \{1 + o(1)\}^2\frac{d^*h_n^4\mathcal{K}_2^2}{4}\int_{\mathcal{U}}\left\{\Delta''(u_t; u_t)m(u_t) + 2\Delta'(u_t; u_t)m'(u_t)\right\}^2 m(u_t)du_t. \end{aligned}$$

Hence, for $i > j$, $i \neq t$ and $j \neq t$, by the boundedness Condition (RC.2) of $m(u)$ and its first derivative over $\mathcal{U}$, and Condition (RC.4) on Δ, it follows that, as $n \to \infty$,

$$E_0[\mathcal{P}_{i,j,t}] = O(h_n^4). \tag{A.15}$$

For $i > j$, and $i = t$, we have

$$\begin{aligned} & E_0[\mathcal{P}_{i,j,i}] \\ &= E_0\left\{[l'(\boldsymbol{\psi}^0(U_i), \boldsymbol{X}_i, Y_i)]^\top l'(\boldsymbol{\psi}^0(U_i), \boldsymbol{X}_j, Y_j)K_{h_n}(0)K_{h_n}(U_j - U_i)\right\} \\ &= K_{h_n}(0)\sum_{s=1}^{d^*}E_{U_i}\left\{E_{(Y,\boldsymbol{X},U_j)|U_i}\left\{l'^{(s)}(\boldsymbol{\psi}^0(U_i), \boldsymbol{X}_i, Y_i)l'^{(s)}(\boldsymbol{\psi}^0(U_i), \boldsymbol{X}_j, Y_j)K_{h_n}(U_j - U_i)|U_i\right\}\right\}. \end{aligned}$$

Now, we calculate the conditional expectations in the above sum. By the independence of

$(\boldsymbol{X}_i, Y_i)$ and $(U_j, \boldsymbol{X}_j, Y_i)$ for $i \neq j$, and the first Bartlett identity verified in Lemma 1,

$$\begin{aligned}
&E_{(Y,\boldsymbol{X},U_j)|U_i}\left\{l'^{(s)}(\boldsymbol{\psi}^0(u_i), \boldsymbol{X}_i, Y_i) l'^{(s)}(\boldsymbol{\psi}^0(u_i), \boldsymbol{X}_j, Y_j) K_{h_n}(U_j - u_i) | U_i = u_i\right\} \\
&\quad = E_{(Y,\boldsymbol{X})|U}\left\{l'^{(s)}(\boldsymbol{\psi}^0(u_i), \boldsymbol{X}_i, Y_i) | U_i = u_i\right\} \\
&\qquad \times E_{(Y,\boldsymbol{X},U_j)|U_i}\left\{l'^{(s)}(\boldsymbol{\psi}^0(u_i), \boldsymbol{X}_j, Y_j) K_{h_n}(U_j - u_i) | U_i = u_i\right\} \\
&\quad = 0 \times E_{(Y,\boldsymbol{X},U_j)|U_i}\left\{l'^{(s)}(\boldsymbol{\psi}^0(u_i), \boldsymbol{X}_j, Y_j) K_{h_n}(U_j - u_i) | U_i = u_i\right\} = 0.
\end{aligned}$$

Thus,

$$E_0[\mathcal{P}_{i,j,i}] = E_0[\mathcal{P}_{i,j,j}] = 0. \tag{A.16}$$

Using (A.15) and (A.16), the second sum in (A.6) can be written as

$$\begin{aligned}
2(1+r_{1n})^2 \frac{h_n}{n} \sum_{i<j}^{n} E_0[\mathcal{P}_{i,j,t}] &= 2(1+r_{1n})^2 \frac{h_n}{n} \left\{ \sum_{\substack{i<j \\ i,j\neq t}}^{n} E_0[\mathcal{P}_{i,j,t}] + E_0[\mathcal{P}_{i,j,i}] + E_0[\mathcal{P}_{i,j,j}] \right\} \\
&= 2(1+r_{1n})^2 \frac{h_n}{n} \left\{ \sum_{\substack{i<j \\ i,j\neq t}}^{n} O(h_n^4) + 0 + 0 \right\} \\
&= (1+r_{1n})^2 \frac{(n-1)(n-2)}{n} O(h_n^5) = (1+r_{1n})^2 O(nh_n^5).
\end{aligned}$$

Under conditions (PC.2), as $n \to \infty$, we have $nh_n^5 = O(1)$. Thus, for large n

$$2(1+r_{1n})^2 \frac{h_n}{n} \sum_{i<j}^{n} E_0[\mathcal{P}_{i,j,t}] = O\{(1+r_{1n})^2\}. \tag{A.17}$$

Hence, by putting together the order assessments in (A.11) and (A.17), and using (A.6), for large n we arrive at $E_0\|l'_{n,h_n}(U_t)\|^2 = O\{(1+r_{1n})^2\}$, which implies that

$$l'_{n,h_n}(U_t) = O_p\{(1+r_{1n})\}. \tag{A.18}$$

Next, we perform an order assessment of the third term in (A.4). Using Condition (RC.3),

$$
\begin{aligned}
& E_0\left[\left|\frac{h_n\gamma_n^3}{6}\sum_{i=1}^{n} R\left(\widetilde{\boldsymbol{\psi}}^0(U_t),\boldsymbol{X}_i,Y_i\right)\right|\right] \\
& \leqslant \frac{h_n\gamma_n^3}{6}\sum_{i=1}^{n} E_0\left|R\left(\widetilde{\boldsymbol{\psi}}^0(U_t),\boldsymbol{X}_i,Y_i\right)\right| \\
& = \frac{h_n\gamma_n^3}{6}\left[E_0\left\{\left|R\left(\widetilde{\boldsymbol{\psi}}^0(U_t),\boldsymbol{X}_t,Y_t\right)\right|\right\} + \sum_{(t\neq i)=1}^{n} E_0\left\{\left|R\left(\widetilde{\boldsymbol{\psi}}^0(U_t),\boldsymbol{X}_i,Y_i\right)\right|\right\}\right] \\
& = \frac{h_n\gamma_n^3}{6}\, K_{h_n}(0)\sum_{j,k,l=1}^{d^*} E_0\left[\left|\frac{\partial^3}{\partial\psi_j\partial\psi_k\partial\psi_l} l\left(\widetilde{\boldsymbol{\psi}}^0(U_t),\boldsymbol{X}_t,Y_t\right) w_{j,t}w_{k,t}w_{l,t}\right|\right] \\
& + \frac{h_n\gamma_n^3}{6} E_0\left[\left|\sum_{(t\neq i)=1}^{n}\sum_{j,k,l=1}^{d^*} \frac{\partial^3}{\partial\psi_j\partial\psi_k\partial\psi_l} l\left(\widetilde{\boldsymbol{\psi}}^0(U_t),\boldsymbol{X}_i,Y_i\right) K_{h_n}(U_i-U_t) w_{j,t}w_{k,t}w_{l,t}\right|\right] \\
& \leqslant \frac{(d^*)^3 h_n\gamma_n^3}{6}\, K_{h_n}(0)\; E_0\left[M_2(\boldsymbol{X}_t,Y_t)\right] \left\{\sum_{l=1}^{d^*}|w_{l,t}|^2\right\}^{3/2} \\
& + \frac{(d^*)^3 h_n\gamma_n^3}{6}\left[\sum_{(t\neq i)=1}^{n} E_0\left[M_2(\boldsymbol{X}_i,Y_i)K_{h_n}(U_i-U_t)\right]\right] \left\{\sum_{l=1}^{d^*}|w_{l,t}|^2\right\}^{3/2} \\
& \leqslant \frac{(d^*)^3 h_n\gamma_n^3}{6}\, K_{h_n}(0)\; E_0\left[M_2(\boldsymbol{X}_t,Y_t)\right] \left\{\sum_{l=1}^{d^*}|w_{l,t}|^2\right\}^{3/2} \\
& + \frac{(d^*)^3 h_n\gamma_n^3}{6}\, nE_0\left[M_2(\boldsymbol{X}_i,Y_i)K_{h_n}(U_i-U_t)\right] \left\{\sum_{l=1}^{d^*}|w_{l,t}|^2\right\}^{3/2} \\
& = \frac{(d^*)^3\gamma_n^3}{6}\, K(0)\; E_0\left[M_2(\boldsymbol{X}_t,Y_t)\right] \left\{\sum_{j=1}^{d^*}|w_{j,t}|^2\right\}^{3/2} \\
& + \frac{(d^*)^3 h_n\gamma_n^3}{6}\, nE_0\left[M_2(\boldsymbol{X}_i,Y_i)K_{h_n}(U_i-U_t)\right] \left\{\sum_{l=1}^{d^*}|w_{l,t}|^2\right\}^{3/2} \\
& = O(\gamma_n^3) + O(\gamma_n) = O(\gamma_n).
\end{aligned}
$$

Thus,

$$
\frac{h_n\gamma_n^3}{6}\sum_{i=1}^{n} R\left(\widetilde{\boldsymbol{\psi}}^0(U_t),\boldsymbol{X}_i,Y_i\right) = O_p(\gamma_n) \tag{A.19}
$$

uniformly for all $t = 1, 2, \ldots, n$. Thus, since $\gamma_n \to 0$, (A.4) reduces to

$$
d_{n,I}(U_t,\boldsymbol{w}_t) = \boldsymbol{w}_t^\top l'_{n,h_n}(U_t) + \frac{1}{2}\boldsymbol{w}_t^\top l''_{n,h_n}(U_t)\boldsymbol{w}_t + o_p(1). \tag{A.20}
$$

On the other hand,

$$\begin{aligned} l''_{n,h_n}(U_t) &= \frac{(1+r_{1n})^2}{n}\sum_{i=1}^{n} l''\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i\right) K_{h_n}(U_i - U_t) \\ &= \frac{K(0)}{nh}(1+r_{1n})^2 l''(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t) \\ &\quad + \frac{(1+r_{1n})^2}{n}\left\{\sum_{(t\neq)i=1}^{n} l''(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - U_t)\right\} \\ &= \mathcal{E}_{1n} + \mathcal{E}_{2n}(U_t). \end{aligned} \tag{A.21}$$

Note that

$$-E_0\left\{l''(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t)\right\} = E_U\left\{\boldsymbol{I}(U_t)\right\}$$

which by Condition (RC.4) is finite and positive definite. Thus, by Condition (RC.5),

$$\begin{aligned} & l''\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t\right) \\ &= \left\{l''\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t\right) - E_0\left\{l''\left(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t\right)\right\}\right\} + E_0\{l''(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t)\} \\ &= O_p(1) - E_0\{\boldsymbol{I}(U_t)\} = O_p(1). \end{aligned}$$

Hence, by conditions $r_{1n} = O(1)$ and $nh_n \to \infty$ (see (PC.2)), as $n \to \infty$,

$$\mathcal{E}_{1n} = \frac{K(0)}{nh}(1+r_{1n})^2 l''(\boldsymbol{\psi}^0(U_t), \boldsymbol{X}_t, Y_t) = \frac{(1+r_{1n})^2}{nh_n}\, O_p(1) = o_p(1). \tag{A.22}$$

Next, we focus on $\mathcal{E}_{2n}(U_t)$. Define the centralized (finite dimensional) matrix

$$Z_{i,n}(u) = l''(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - u) - E_0[l''(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - u)], \tag{A.23}$$

where $E_0[Z_{i,n}(u)] = 0$, for any $u \in \mathcal{U}$. Note that, by using a change of variable $\nu_i = (u_i - u)/h_n$,

$$\begin{aligned} & E_0[l''(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - u)] \\ &= \int_{\mathcal{Y}}\int_{\mathcal{X}}\int_{\mathcal{U}} l''(\boldsymbol{\psi}^0(u), \boldsymbol{x}_i, y_i) K_{h_n}(u_i - u) f(u_i, \boldsymbol{x}_i, y_i) du_i d\boldsymbol{x}_i dy_i \\ &= \int_{\mathcal{U}}\int_{\mathcal{Y}}\int_{\mathcal{X}} l''(\boldsymbol{\psi}^0(u), \boldsymbol{x}_i, y_i) K(\nu_i) f(u + h_n\nu_i, \boldsymbol{x}_i, y_i) d\boldsymbol{x}_i dy_i d\nu_i, \end{aligned}$$

where f is the joint pdf of $(U_i, \boldsymbol{X}_i, Y_i)$. By a second-order Taylor expansion,

$$f(u + h_n\nu_i, \boldsymbol{x}_i, y_i) = f(u, \boldsymbol{x}_i, y_i) + h_n\nu_i f'(u, \boldsymbol{x}_i, y_i) + \frac{h_n^2\nu_i^2}{2} f''(u_{ni}, \boldsymbol{x}_i, y_i) \tag{A.24}$$

such that u_{ni} lies between u and $u + h_n\nu_i$, and f' and f'' are the partial derivatives of $f(u, \boldsymbol{x}_i, y_i)$ with respect to u, which by Conditions (RC.1) and (RC.2) exist. Replacing (A.24) in the above integration and using the boundedness Condition (KC.2) on the kernel $K(\cdot)$ and by the second Bartlett identity verified in Lemma 1, we arrive at

$$E_0[l''(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i - u)] = -m(u)\boldsymbol{I}(u) + o_p(h_n)\ ,\ \forall u \in \mathcal{U}. \tag{A.25}$$

Turning to (A.23), we show that, as $n \to \infty$,

$$\sup_{u\in\mathcal{U}} \left\| \frac{1}{n}\sum_{i=1}^{n} Z_{i,n}(u) \right\|_2 = O_p\left\{ \left(\frac{\log(\frac{1}{h_n})}{nh_n} \right)^{\frac{1}{2}} \right\} \equiv O_p(\xi_n) = o_p(1). \tag{A.26}$$

Let $Z_{i,n}^{jl}(u)$ be the (j,l)-th element of the matrix $Z_{i,n}(u)$. Since $\mathcal{U}$ is compact, let $\bigcup_{k=1}^{\mathcal{N}_n} B(u_k; \eta_n)$ be a finite open cover of $\mathcal{U}$, where $B(u_k; \eta_n)$ are balls centred at u_k and with radius η_n, such that $\eta_n = o(1)$, as $n \to \infty$. Clearly, we have $\mathcal{N}_n = O(\eta_n^{-1})$. Then, we can write,

$$\begin{aligned}
\sup_{u\in\mathcal{U}} \left| \frac{1}{n}\sum_{i=1}^{n} Z_{i,n}^{jl}(u) \right| &\leqslant \max_{1\leqslant k\leqslant \mathcal{N}_n} \left| \frac{1}{n}\sum_{i=1}^{n} Z_{i,n}^{jl}(u_k) \right| + \max_{1\leqslant k\leqslant \mathcal{N}_n} \sup_{u\in\mathcal{U}} \left| \frac{1}{n}\sum_{i=1}^{n} \left[Z_{i,n}^{jl}(u) - Z_{i,n}^{jl}(u_k) \right] \right| \\
&\leqslant \max_{1\leqslant k\leqslant \mathcal{N}_n} \left| \frac{1}{n}\sum_{i=1}^{n} Z_{i,n}^{jl}(u_k) \right| + \max_{1\leqslant k\leqslant \mathcal{N}_n} \sup_{u\in B(u_k;\eta_n)} \left| \frac{1}{n}\sum_{i=1}^{n} \left[Z_{i,n}^{jl}(u) - Z_{i,n}^{jl}(u_k) \right] \right| \\
&= \mathcal{Z}_{n1} + \mathcal{Z}_{n2}.
\end{aligned} \tag{A.27}$$

We show that $\mathcal{Z}_{n1}$ and $\mathcal{Z}_{n2}$ follow the rate as in (A.26). We first focus on $\mathcal{Z}_{n2}$. Note that,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \left[Z_{i,n}^{jl}(u) - Z_{i,n}^{jl}(u_k) \right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\{ \left[l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i - u) - l''_{jl}(\boldsymbol{\psi}^0(u_k), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i - u_k) \right] \right. \\
&\qquad \left. - E_0\left[l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i - u) - l''_{jl}(\boldsymbol{\psi}^0(u_k), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i - u_k) \right] \right\},
\end{aligned}$$

where l''_{jl} is the (j,l)-th element of the matrix l''. The expression in the first square bracket can be rewritten as

$$\begin{aligned}
&\left[l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i - u) - l''_{jl}(\boldsymbol{\psi}^0(u_k), \boldsymbol{X}_i, Y_i)K_{h_n}(U_i - u_k) \right] \\
&= l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i)\left[K_{h_n}(U_i - u) - K_{h_n}(U_i - u_k) \right] \\
&\quad + \left[l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i) - l''_{jl}(\boldsymbol{\psi}^0(u_k), \boldsymbol{X}_i, Y_i) \right] K_{h_n}(U_i - u_k).
\end{aligned}$$

By the Lipschitz condition in (KC.1) on the kernel K, and Conditions (RC.3) and (RC.5) on $l(\boldsymbol{\psi}(u), \boldsymbol{x}, y)$, we have that, for large n,

$$\left|\frac{1}{n}\sum_{i=1}^{n} l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i)\left[K_{h_n}(U_i - u) - K_{h_n}(U_i - u_k)\right]\right| \leqslant \left|\frac{1}{n}\sum_{i=1}^{n} l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i)\right| \frac{C|u-u_k|}{h_n^2} = O_p\left(\frac{\eta_n}{h_n^2}\right).$$

Similarly, by the above conditions and also Condition (RC.2), the Condition (KC.1) on the kernel $K(\cdot)$, and using the mean value theorem, for large n,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\left|l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i) - l''_{jl}(\boldsymbol{\psi}^0(u_k), \boldsymbol{X}_i, Y_i)\right| K_{h_n}(U_i - u_k)\right\} \leqslant O_p(1)\frac{|u-u_k|}{h_n} = O_p\left(\frac{\eta_n}{h_n}\right).$$

Hence, by the putting together the above order assessments, we have that, for large n,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[Z^{jl}_{i,n}(u) - Z^{jl}_{i,n}(u_k)\right]\right| = O_p\left(\frac{r_n}{h_n^2}\right) + O_p\left(\frac{\eta_n}{h_n}\right) = O_p\left(\frac{\eta_n}{h_n^2}\right)$$

which implies that if we choose $\eta_n = h_n^2\xi_n$, we have, for large n,

$$\mathcal{Z}_{n2} = O_p\left(\frac{\eta_n}{h_n^2}\right) = O_p(\xi_n). \tag{A.28}$$

Next, we focus on the order assessment of $\mathcal{Z}_{n1}$ in (A.27). Denote the random variables,

$$\overline{Z}^{jl}_{i,n}(u) \;=\; Z^{jl}_{i,n}(u)\;\mathbf{1}\{|l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i)| \leqslant C_2 n^{1/m_0}\} \tag{A.29}$$

and

$$\widetilde{Z}^{jl}_{i,n}(u) \;=\; Z^{jl}_{i,n}(u)\;\mathbf{1}\{|l''_{jl}(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i)| > C_2 n^{1/m_0}\} \tag{A.30}$$

such that $Z^{jl}_{i,n}(u) = \overline{Z}^{jl}_{i,n}(u) + \widetilde{Z}^{jl}_{i,n}(u)$, for some constant $C_2 > 0$, and $\mathbf{1}\{\}$ is indicator function. Note that $E_0\{Z^{jl}_{i,n}(u)\} = 0$. Then, we have that

$$\begin{aligned}
\mathcal{Z}_{n1} &= \max_{1\leqslant k\leqslant \mathcal{N}_n}\left|\frac{1}{n}\sum_{i=1}^{n} Z^{jl}_{i,n}(u_k)\right| \\
&\leqslant \max_{1\leqslant k\leqslant \mathcal{N}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\overline{Z}^{jl}_{i,n}(u_k) - E\{\overline{Z}^{jl}_{i,n}(u_k)\}\right]\right| + \max_{1\leqslant k\leqslant \mathcal{N}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{Z}^{jl}_{i,n}(u_k) - E\{\widetilde{Z}^{jl}_{i,n}(u_k)\}\right]\right| \\
&= \mathcal{J}_{n1} + \mathcal{J}_{n2}.
\end{aligned} \tag{A.31}$$

For some constant $C_3 > 0$, any $\epsilon > 0$, by using the Markov inequality and Condition (RC.3),

$$\begin{aligned} P(\mathcal{J}_{n2} > C_3\xi_n) &\leqslant P\left(\max_{1\leqslant k\leqslant \mathcal{N}_n}\bigcup_{i=1}^{n}\left\{|l''_{jl}(\boldsymbol{\psi}^0(u_k), \boldsymbol{X}_i, Y_i)| > C_2 n^{1/m_0}\right\}\right) \\ &\leqslant P\left(\bigcup_{i=1}^{n}\left\{|M_1(\boldsymbol{X}_i, Y_i)| > C_2 n^{1/m_0}\right\}\right) \leqslant \sum_{i=1}^{n} P(M_1(\boldsymbol{X}_i, Y_i) > C_2 n^{1/m_0}) \\ &\leqslant \sum_{i=1}^{n} E_0\{M_1^{m_0}(\boldsymbol{X}_i, Y_i)\}\ (C_2 n^{1/m_0})^{-m_0} = E_0\{M_1^{m_0}(\boldsymbol{X}_i, Y_i)\}C_2^{-m_0} < \epsilon \end{aligned}$$

if we choose $C_2 > \epsilon^{-1/m_0}\ [E_0\{M_1^{m_0}(\boldsymbol{X}_i, Y_i)\}]^{1/m_0}$. Since $\epsilon > 0$ can be arbitrarily small, we then have that, for large n,

$$\mathcal{J}_{n2} = O_p(\xi_n). \tag{A.32}$$

We use the Bernstein inequality, Lemma 2.2.9 of van der Vaart and Wellner (1996), to assess the large sample behavior of $\mathcal{J}_{n1}$ in (A.31). First, we verify the conditions of the lemma. Note that, by the definitions in (A.23) and (A.29), and the boundedness of $K(\cdot)$,

$$\left|\overline{Z}_{i,n}^{jl}(u_k) - E\{\overline{Z}_{i,n}^{jl}(u_k)\}\right| \leqslant \frac{C_4 C_2 n^{1/m_0}}{h_n}, \quad Var\{\overline{Z}_{i,n}^{jl}(u_k)\} \leqslant \frac{C_4}{h_n} \tag{A.33}$$

for some constant $C4 > 0$. By the Bernstein inequality,

$$P(\mathcal{J}_{n1} > C_3\xi_n) \leqslant 2\mathcal{N}_n \exp\left\{-\frac{C_3^2 n^2 \xi_n^2}{2nC_4/h_n + 2nC_4\xi_n C_3 C_2 n^{1/m_0}/3h_n}\right\}.$$

If we choose C_3 such that $C_3 > 3C_4$, then by the definition of ξ_n,

$$\frac{C_3^2 n^2 \xi_n^2}{2nC_4/h_n + 2nC_4\xi_n C_3 C_2 n^{1/m_0}/3h_n} \geqslant \frac{(3C_3/2)\log h_n^{-1}}{1 + C_2 C_4 \xi_n n^{1/m_0}}.$$

Also if, for large n,

$$n^{1/m_0}\xi_n = \left(\frac{n^{2/m_0}\log(\frac{1}{h_n})}{nh_n}\right)^{\frac{1}{2}} = o(1) \quad , \quad \mathcal{N}_n \exp\{-C_3 \log h_n^{-1}\} = o(1) \tag{A.34}$$

Thus, for large n, we have that $P(\mathcal{J}_{n1} > C_3\xi_n) \leqslant o(1)$, which implies that

$$\mathcal{J}_{n1} = O_p(\xi_n). \tag{A.35}$$

Hence, in summary, by (A.28), and also applying (A.32) and (A.35) to (A.31), we have proved (A.26).

Therefore, for any $u \in \mathcal{U}$, and using (A.25),

$$\begin{aligned}
&\frac{1}{n}\left\{\sum_{i=1}^{n} l''(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - u)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} Z_{i,n}(u) + E_0\left[l''(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - u)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} Z_{i,n}(u) - m(u)\boldsymbol{I}(u) + o_p(h_n) \qquad \text{(A.36)}
\end{aligned}$$

and then, using (A.26), we have that

$$\begin{aligned}
&\sup_{u\in\mathcal{U}}\left\|\frac{1}{n}\sum_{i=1}^{n} l''(\boldsymbol{\psi}^0(u), \boldsymbol{X}_i, Y_i) K_{h_n}(U_i - u) + m(u)\boldsymbol{I}(u)\right\|_2 \\
&= \sup_{u\in\mathcal{U}}\left\|\frac{1}{n}\sum_{i=1}^{n} Z_{i,n}(u)\right\|_2 + o_p(h_n) = O_p(\xi_n + h_n). \qquad \text{(A.37)}
\end{aligned}$$

This implies that, for large n,

$$\begin{aligned}
&\sup_{u\in\mathcal{U}} \|l''_{n,h_n}(u) + (1 + r_{1n})^2 m(u)\boldsymbol{I}(u)\|_2 \\
&= \sup_{u\in\mathcal{U}} \|o_p\{(1 + r_{1n})^2\} + \mathcal{E}_{2n}(u) + (1 + r_{1n})^2 m(u)\boldsymbol{I}(u)\|_2 \\
&= (1 + r_{1n})^2 O_p(\xi_n + h_n) = o_p\{(1 + r_{1n})^2\} \qquad \text{(A.38)}
\end{aligned}$$

since, by Condition (PC.2), $nh_n \to \infty$.

Using (A.20) and taking into account (A.38), then (A.3) reduces to

$$\begin{aligned}
&D_{n,I}(\boldsymbol{W}_n) = \frac{1}{n}\sum_{t=1}^{n}\left[\boldsymbol{w}_t^\top l'_{n,h_n}(U_t) + \frac{1}{2}\boldsymbol{w}_t^\top l''_{n,h_n}(U_t)\boldsymbol{w}_t\right] + o_p(1) \\
&= \frac{1}{n}\sum_{t=1}^{n}\left[\boldsymbol{w}_t^\top l'_{n,h_n}(U_t) + \frac{1}{2}\boldsymbol{w}_t^\top\left\{l''_{n,h_n}(U_t) + (1 + r_{1n})^2 m(U_t)\boldsymbol{I}(U_t) - (1 + r_{1n})^2 m(U_t)\boldsymbol{I}(U_t)\right\}\boldsymbol{w}_t\right] \\
&\quad + o_p(1) \\
&\leqslant \frac{1}{n}\sum_{t=1}^{n}\left[\boldsymbol{w}_t^\top l'_{n,h_n}(U_t) - \frac{(1 + r_{1n})^2}{2}\boldsymbol{w}_t^\top\left\{m(U_t)\boldsymbol{I}(U_t) + o_p(1)\right\}\boldsymbol{w}_t\right] + o_p(1) \\
&= \frac{1}{n}\sum_{t=1}^{n}\left[\boldsymbol{w}_t^\top l'_{n,h_n}(U_t) - \frac{(1 + r_{1n})^2}{2}\boldsymbol{w}_t^\top\left\{m(U_t)\boldsymbol{I}(U_t)\right\}\boldsymbol{w}_t\right] + o_p(1). \qquad \text{(A.39)}
\end{aligned}$$

By conditions (RC.2) and (RC.4), the matrix $m(U_t)\boldsymbol{I}(U_t)$ is positive definite, and thus all of its eigenvalues are positive. Let $\lambda_t^{\min}$ be the smallest eigenvalue of $m(U_t)\boldsymbol{I}(U_t)$. Also, let $\lambda_{\min} = \min\{\lambda_t^{\min},\ t = 1, \ldots, n\}$ and $\lambda_0^{\min} = \inf_{u\in\mathcal{U}} \lambda_{\min}\{m(u)\boldsymbol{I}(u)\}$. Inequality in (A.39)

reduces to

$$
\begin{aligned}
& D_{n,I}(\boldsymbol{W}_n) \\
\leqslant\ & \frac{1}{n}\sum_{t=1}^{n}\left[\boldsymbol{w}_t^\top l'_{n,h_n}(U_t) - \frac{(1+r_{1n})^2}{2}\boldsymbol{w}_t^\top \lambda_t^{\min}\boldsymbol{w}_t\right] + o_p(1) \\
\leqslant\ & \frac{1}{n}\left[\sum_{t=1}^{n}\boldsymbol{w}_t^\top l'_{n,h_n}(U_t) - \frac{(1+r_{1n})^2}{2}\times\lambda_{\min}\times\sum_{t=1}^{n}\|\boldsymbol{w}_t\|^2\right] + o_p(1) \\
\leqslant\ & \left(n^{-1}\|\boldsymbol{W}_n\|^2\right)^{\frac{1}{2}}\left(n^{-1}\sum_{t=1}^{n}\|l'_{n,h_n}(U_t)\|^2\right)^{\frac{1}{2}} - \frac{(1+r_{1n})^2}{2}\times\lambda_{\min}\times\left(n^{-1}\|\boldsymbol{W}_n\|^2\right) + o_p(1) \\
=\ & M_\epsilon\left(n^{-1}\sum_{t=1}^{n}\|l'_{n,h_n}(U_t)\|^2\right)^{\frac{1}{2}} - \frac{(1+r_{1n})^2}{2}\times\lambda_{\min}\times M_\epsilon^2 + o_p(1) \\
\leqslant\ & M_\epsilon\left(n^{-1}\sum_{t=1}^{n}\|l'_{n,h_n}(U_t)\|^2\right)^{\frac{1}{2}} - \frac{(1+r_{1n})^2}{2}\times\lambda_0^{\min}\times M_\epsilon^2\ (1+o_p(1)). \qquad \text{(A.40)}
\end{aligned}
$$

As we showed in (A.18), we have

$$
n^{-1}\sum_{t=1}^{n}E\|l'_{n,h_n}(U_t)\|^2 = n^{-1}\sum_{t=1}^{n}O_p\{(1+r_{1n})^2\} = O_p\{(1+r_{1n})^2\}
$$

Note that $\lambda_0^{\min} > 0$. Therefore, we can choose M_ϵ large enough such that the second term in (A.40) and thus $D_{n,I}(\boldsymbol{W}_n)$ become negative, in probability, for large n.

Step 2: (Order assessment of the penalty difference) Following the penalty difference

in (A.2), $D_{n,II}(\boldsymbol{W}_n)$, which involves the grouping penalty and using Taylor expansion,

$$
\begin{aligned}
& D_{n,II}(\boldsymbol{W}_n) \qquad\qquad (A.41)\\
= \;& h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \left\{ p_n \left(\left\| \boldsymbol{b}_{jl}^0 + \gamma_n \boldsymbol{v}_{jl} \right\| / \sqrt{n}; \lambda_{nj} \right) - p_n \left(\left\| \boldsymbol{b}_{jl}^0 \right\| / \sqrt{n}; \lambda_{nj} \right) \right\} \\
= \;& h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \sum_{t=1}^{n} \frac{\gamma_n \beta_{jl}^0(U_t)}{\sqrt{n} \left\| \boldsymbol{b}_{jl}^0 \right\|} \; p_n' \left(\left\| \boldsymbol{b}_{jl}^0 \right\| / \sqrt{n}; \lambda_{nj} \right) \; w_{jl,I,t} \\
& + h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \sum_{t=1}^{n} \frac{\gamma_n^2}{2\sqrt{n}} \left(\frac{1}{\| \boldsymbol{b}_{jl}^0 \|} - \frac{\left[\beta_{jl}^0(U_t) \right]^2}{\| \boldsymbol{b}_{jl}^0 \|^3} \right) \; p_n' \left(\left\| \boldsymbol{b}_{jl}^0 \right\| / \sqrt{n}; \lambda_{nj} \right) \; w_{cj,I,t}^2 \\
& + h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \sum_{t=1}^{n} \frac{\gamma_n^2 \left[\beta_{cj}^0(U_t) \right]^2}{2n \left\| \boldsymbol{b}_{jl}^0 \right\|^2} \; p_n''(\| \boldsymbol{b}_{jl}^0 \| / \sqrt{n}; \lambda_{n\kappa}) \; w_{jl,I,t}^2 \\
& + h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \sum_{t \neq l=1}^{n} \frac{-\gamma_n^2 \beta_{jl}^0(U_t) \beta_{jl}^0(U_l)}{\sqrt{n} \left\| \boldsymbol{b}_{jl}^0 \right\|^3} p_n' \left(\left\| \boldsymbol{b}_{jl}^0 \right\| / \sqrt{n}; \lambda_{nj} \right) \; w_{jl,I,t} w_{jl,I,l} / 2 \\
& + h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \sum_{t \neq l=1}^{n} \frac{\gamma_n^2 \beta_{jl}^0(U_t) \beta_{jl}^0(U_l)}{n \left\| \boldsymbol{b}_{jl}^0 \right\|^2} \; p_n''(\left\| \boldsymbol{b}_{jl}^0 \right\| / \sqrt{n}; \lambda_{nj}) \; w_{jl,I,t} w_{jl,I,l} / 2 \\
= \;& d_{1,n,II}(\boldsymbol{W}_n) + d_{2,n,II}(\boldsymbol{W}_n) + d_{3,n,II}(\boldsymbol{W}_n) + d_{4,n,II}(\boldsymbol{W}_n) + d_{5,n,II}(\boldsymbol{W}_n).
\end{aligned}
$$

We now perform order assessment of the terms $d_{1,n,II}(\boldsymbol{w}), \ldots, d_{5,n,II}(\boldsymbol{w})$. For large n,

$$
\begin{aligned}
& d_{1,n,II}(\boldsymbol{W}_n) \\
\leqslant \;& (1 + r_{1n}) h_n^{\frac{1}{2}} n^{-\frac{4}{2}} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_\kappa^0} \sum_{t=1}^{n} \frac{|\beta_{jl}^0(U_t)|}{\left\| \boldsymbol{b}_{jl}^0 \right\|} \max_{\substack{1 \leqslant l \leqslant d_j^0 \\ 1 \leqslant j \leqslant \mathcal{C}}} \left| p_n' \left(\left\| \boldsymbol{b}_{jl}^0 \right\| / \sqrt{n}; \lambda_{nj} \right) \right| |w_{jl,I,t}| \\
\leqslant \;& (1 + r_{1n}) h_n^{\frac{1}{2}} \max_{\substack{1 \leqslant l \leqslant d_j^0 \\ 1 \leqslant j \leqslant \mathcal{C}}} \left| \frac{p_n' \left(\left\| \boldsymbol{b}_{jl}^0 \right\| / \sqrt{n}; \lambda_{nj} \right)}{n^{3/2}} \right| \; \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \left\{ \sum_{t=1}^{n} w_{jl,I,t}^2 / n \right\}^{1/2} \\
\leqslant \;& M_\epsilon \, (1 + r_{1n}) r_{1n}. \qquad\qquad (A.42)
\end{aligned}
$$

Following $d_{2,n,II}(\boldsymbol{W}_n)$, for large n,

$$
\begin{aligned}
&|d_{2,n,II}(\boldsymbol{W}_n)| \\
&\leqslant 2^{-1}h_n n^{-3/2}\gamma_n^2 \sum_{j=1}^{\mathcal{C}}\sum_{l=1}^{d_\kappa^0}\sum_{t=1}^{n}\frac{1}{\|\boldsymbol{b}_{jl}^0\|}\left(1+\frac{\left[\beta_{jl}^0(U_t)\right]^2}{\|\boldsymbol{b}_{jl}^0\|^2}\right)\ \left|p_n'\left(\left\|\boldsymbol{b}_{jl}^0\right\|/\sqrt{n};\lambda_{nj}\right)\right|\ w_{jl,I,t}^2 \\
&\leqslant (1+r_{1n})^2 n^{-1/2}\{\min_{j,l}\|\boldsymbol{b}_{jl}^0\|/\sqrt{n}\}^{-1}\max_{\substack{1\leqslant l\leqslant d_j^0\\ 1\leqslant j\leqslant\mathcal{C}}}\left|\frac{p_n'\left(\left\|\boldsymbol{b}_{jl}^0\right\|/\sqrt{n};\lambda_{nj}\right)}{n^{3/2}}\right|\left\{\sum_{t=1}^{n}w_{jl,I,t}^2/n\right\} \\
&= M_\epsilon^2\ r_{1n}(1+r_{1n})^2\{nh_n\}^{-1/2}\left\{\min_{j,l}E_0\left[\beta_{jl}^2(U)\right]\right\}^{-1/2}. \qquad \text{(A.43)}
\end{aligned}
$$

Following $d_{3,n,II}(\boldsymbol{W}_n)$, for large n,

$$
\begin{aligned}
|d_{3,n,II}(\boldsymbol{W}_n)| &\leqslant h_n n^{-1}\sum_{j=1}^{\mathcal{C}}\sum_{l=1}^{d_j^0}\sum_{t=1}^{n}\frac{\gamma_n^2\left[\beta_{jl}^0(U_t)\right]^2}{2n\left\|\boldsymbol{b}_{jl}^0\right\|^2}\ \left|p_n''\left(\left\|\boldsymbol{b}_{jl}^0\right\|/\sqrt{n};\lambda_{nj}\right)\right|\ w_{jl,I,t}^2 \\
&\leqslant (1+r_{1n})^2\max_{\substack{1\leqslant l\leqslant d_j^0\\ 1\leqslant j\leqslant\mathcal{C}}}\left|\frac{p_n''(\left\|\boldsymbol{b}_{jl}^0\right\|/\sqrt{n};\lambda_{nj})}{n^2}\right|\left\{\sum_{t=1}^{n}w_{jl,I,t}^2/n\right\} \\
&\leqslant M_\epsilon^2\ r_{2n}(1+r_{1n})^2. \qquad \text{(A.44)}
\end{aligned}
$$

Following $d_{4,n,II}(\boldsymbol{W}_n)$,

$$
\begin{aligned}
&|d_{4,n,II}(\boldsymbol{W}_n)| \\
&\leqslant h_n n^{-1}\sum_{j=1}^{\mathcal{C}}\sum_{l=1}^{d_\kappa^0}\sum_{t\neq l=1}^{n}\frac{\gamma_n^2\left|\beta_{jl}^0(U_t)\right|\left|\beta_{jl}^0(U_l)\right|}{\sqrt{n}\left\|\boldsymbol{b}_{jl}^0\right\|^3}\left|p_n'\left(\left\|\boldsymbol{b}_{jl}^0\right\|/\sqrt{n};\lambda_{n\kappa}\right)\right|\ \left|w_{jl,I,t}w_{jl,I,l}\right|/2 \\
&\leqslant h_n n^{-3/2}(1+r_{1n})^2(h_n n)^{-1}\left\{\min_{j,l}\left\|\boldsymbol{b}_{jl}^0\right\|\right\}^{-1}\max_{\substack{1\leqslant j\leqslant d_\kappa^0\\ 1\leqslant\kappa\leqslant\mathcal{C}}}\left|p_n'\left(\left\|\boldsymbol{b}_{jl}^0\right\|/\sqrt{n};\lambda_{nj}\right)\right|\left\{\sum_{t=1}^{n}w_{jl,I,t}^2\right\} \\
&= n^{-1/2}(1+r_{1n})^2\left\{\min_{j,l}\left\|\boldsymbol{b}_{jl}^0\right\|/\sqrt{n}\right\}^{-1}\ \max_{\substack{1\leqslant l\leqslant d_j^0\\ 1\leqslant j\leqslant\mathcal{C}}}\left|\frac{p_n'\left(\left\|\boldsymbol{b}_{jl}^0\right\|/\sqrt{n};\lambda_{nj}\right)}{n^{3/2}}\right|\left\{\sum_{t=1}^{n}w_{jl,I,t}^2/n\right\} \\
&= M_\epsilon^2\ \{nh_n\}^{-1/2}r_{1n}(1+r_{1n})^2\left\{\min_{j,l}E_0\left[\beta_{jl}^2(U)\right]\right\}^{-1/2}. \qquad \text{(A.45)}
\end{aligned}
$$

Following $d_{5,n,II}(\boldsymbol{W}_n)$,

$$\begin{aligned}
&|d_{5,n,II}(\boldsymbol{W}_n)| \\
&\leqslant h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \sum_{t\neq l=1}^{n} \frac{\gamma_n^2 \left|\beta_{jl}^0(U_t)\right| \left|\beta_{jl}^0(U_l)\right|}{n \left\|\boldsymbol{b}_{jl}^0\right\|^2} \, p_n''\left(\left\|\boldsymbol{b}_{jl}^0\right\| / \sqrt{n}; \lambda_{nj}\right) \, |w_{jl,I,t} w_{jl,I,l}|/2 \\
&\leqslant (1+r_{1n})^2 \max_{\substack{1\leqslant l\leqslant d_j^0 \\ 1\leqslant j\leqslant \mathcal{C}}} \left| \frac{p_n''\left(\left\|\boldsymbol{b}_{jl}^0\right\| / \sqrt{n}; \lambda_{nj}\right)}{n^2} \right| \left\{ \sum_{t=1}^{n} w_{jl,I,t}^2 / n \right\} \\
&= M_\epsilon^2 \, r_{2n} (1+r_{1n})^2. \qquad \text{(A.46)}
\end{aligned}$$

Thus, by Condition (PC.2) on the penalty p_n and the smoothing parameter h_n, the order assessment in (A.40) and those in (A.42)-(A.46) corresponding to the terms $D_{n,I}(\boldsymbol{W}_n)$ and $D_{n,II}(\boldsymbol{W}_n)$ in (A.2) imply that for a sufficiently large M_ϵ the expression

$$-\frac{(1+r_{1n})^2}{2} \lambda_0^{\min} M_\epsilon^2$$

is the sole leading term in the right side of $D_n(\boldsymbol{w})$ in (A.2). Therefore, for any given $\epsilon > 0$, there exists a sufficiently large M_ϵ, such that

$$\lim_{n\to\infty} P\left\{ \sup_{n^{-1}\|\boldsymbol{W}_n\|_2^2 = M_\epsilon^2} \widetilde{L}_n\left(\boldsymbol{\Psi}_n^0 + \gamma_n \boldsymbol{W}_n; \boldsymbol{\lambda}_n, h_n\right) < \widetilde{L}_n\left(\boldsymbol{\Psi}_n^0; \boldsymbol{\lambda}_n, h_n\right) \right\} \geqslant 1-\epsilon$$

as needed in (A.1). This completes the proof of Theorem 1.

Proof of Lemma 1

Proof. Recall the partitioning of the true parameter matrix $\boldsymbol{\Psi}_n^0 = (\boldsymbol{\Psi}_{n1}^0, \boldsymbol{\Psi}_{n2}^0)$ such that $\boldsymbol{\Psi}_{n2}^0$ corresponds to all the zero regression functions, while $\boldsymbol{\Psi}_{n1}^0$ corresponds to the nonzero regression functions, among other parameters, given in Section 4 of the paper. Also, consider any matrix $\boldsymbol{\Psi}_n = (\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2})$ in a neighbourhood of $\boldsymbol{\Psi}_n^0$ such that $n^{-1}\|\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^0\|_F^2 = O_p\big((nh_n)^{-1}\big)$, and $\dim(\boldsymbol{\Psi}_{n1}) = \dim(\boldsymbol{\Psi}_{n1}^0)$ and $\dim(\boldsymbol{\Psi}_{n2}) = \dim(\boldsymbol{\Psi}_{n2}^0)$. The existence of such neighbourhood is guaranteed by the result of Theorem 1. We provide the proof in two steps:

Step 1. We first prove that, with probability tending to one, as $n \to \infty$,

$$\widetilde{L}_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2}); \boldsymbol{\lambda}_n, h_n\big) - \widetilde{L}_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{0}); \boldsymbol{\lambda}_n, h_n\big) < 0. \qquad \text{(A.47)}$$

By the definition of the penalized likelihood function in (9) of the paper, we have that

$$\begin{aligned}\widetilde{L}_n\big((\boldsymbol{\Psi}_{n1},\boldsymbol{\Psi}_{n2});\boldsymbol{\lambda}_n,h_n\big)-\widetilde{L}_n\big((\boldsymbol{\Psi}_{n1},\mathbf{0});\boldsymbol{\lambda}_n,h_n\big)&\\ =\big[L_n\big((\boldsymbol{\Psi}_{n1},\boldsymbol{\Psi}_{n2});h_n\big)-L_n\big((\boldsymbol{\Psi}_{n1},\mathbf{0});h_n\big)\big]&\\ -\big[\mathbb{P}_n\big((\boldsymbol{\Psi}_{n1},\boldsymbol{\Psi}_{n2});\boldsymbol{\lambda}_n\big)-\mathbb{P}_n\big((\boldsymbol{\Psi}_{n1},\mathbf{0});\boldsymbol{\lambda}_n\big)\big]\,.&\end{aligned}\tag{A.48}$$

By the definition of the penalty in (10) of the paper, we have

$$\mathbb{P}_n\big((\boldsymbol{\Psi}_{n1},\boldsymbol{\Psi}_{n2});\boldsymbol{\lambda}_n\big)-\mathbb{P}_n\big((\boldsymbol{\Psi}_{n1},\mathbf{0});\boldsymbol{\lambda}_n\big)=\sum_{j=1}^{C}\sum_{l=d_j^0+1}^{d}p_n\left(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n};\lambda_{nj}\right).\tag{A.49}$$

On the other hand, using (7) and (8) of the paper, we can write the likelihood difference as

$$\begin{aligned}L_n\big((\boldsymbol{\Psi}_{n1},\boldsymbol{\Psi}_{n2});h_n\big)-L_n\big((\boldsymbol{\Psi}_{n1},\mathbf{0});h_n\big)&=\sum_{i=1}^{n}\big[\ell_n\big((\boldsymbol{\psi}_1(u_t),\boldsymbol{\psi}_2(u_t));h_n\big)-\ell_n\big((\boldsymbol{\psi}_1(u_t),\mathbf{0});h_n\big)\big]\\ &=\sum_{t=1}^{n}\sum_{i=1}^{n}\big[l\big((\boldsymbol{\psi}_1(U_t),\boldsymbol{\psi}_2(U_t)),\boldsymbol{X}_i,Y_i\big)-l\big((\boldsymbol{\psi}_1(U_t),\mathbf{0}),\boldsymbol{X}_i,Y_i\big)\big]K_{h_n}(U_i-U_t),\end{aligned}\tag{A.50}$$

where $l(\cdot)$ is the log of the mixture density as given in (6) of the paper. Note that $\boldsymbol{\psi}(u_t)=(\boldsymbol{\psi}_1(u_t),\boldsymbol{\psi}_2(u_t))$ is the t-th row of the matrix $\boldsymbol{\Psi}_n=(\boldsymbol{\Psi}_{n1},\boldsymbol{\Psi}_{n2})$. For any fixed $U_t=u\in\mathcal{U}$, we first assess the inner sum in (A.50). We thus consider

$$\Delta_n(u)=\sum_{i=1}^{n}\big[l\big((\boldsymbol{\psi}_1(u),\boldsymbol{\psi}_2(u)),\boldsymbol{X}_i,Y_i\big)-l\big((\boldsymbol{\psi}_1(u),\mathbf{0}),\boldsymbol{X}_i,Y_i\big)\big]K_{h_n}(U_i-u)\tag{A.51}$$

By the mean value theorem, there exists a vector $\boldsymbol{\xi}(u)$ on the segment between $\mathbf{0}$ and $\boldsymbol{\psi}_2(u)$ such that

$$\begin{aligned}&\Delta_n(u)\\ &=\sum_{i=1}^{n}K_{h_n}(U_i-u)\left[\frac{\partial l((\boldsymbol{\psi}_1(u),\boldsymbol{\xi}(u)),\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\psi}_2(u)}\right]^{\top}\boldsymbol{\psi}_2(u)\\ &=\sum_{i=1}^{n}K_{h_n}(U_i-u)\left[\frac{\partial l((\boldsymbol{\psi}_1(u),\boldsymbol{\xi}(u)),\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\psi}_2(u)}-\frac{\partial l((\boldsymbol{\psi}_1^0(u),\mathbf{0}),\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\psi}_2(u)}\right]^{\top}\boldsymbol{\psi}_2(u)\\ &+\sum_{i=1}^{n}K_{h_n}(U_i-u)\left[\frac{\partial l((\boldsymbol{\psi}_1^0(u),\mathbf{0}),\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\psi}_2(u)}\right]^{\top}\boldsymbol{\psi}_2(u)=\Delta_1(u)+\Delta_2(u),\end{aligned}\tag{A.52}$$

where $\Delta_1(u)$ and $\Delta_2(u)$ are respectively the two sums on the right hand side of the above equation. Similar to (A.18), as shown in the proof of Theorem 1 for the order assessment of

(A.5), we have that, for any $u \in \mathcal{U}$,

$$\sum_{i=1}^{n} \frac{\partial l((\boldsymbol{\psi}_1^0(u), \mathbf{0}), \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\psi}_2(u)} K_{h_n}(U_i - u) = O_p\{(nh_n^{-1})^{\frac{1}{2}}\}. \tag{A.53}$$

Thus, since $\dim(\boldsymbol{\psi}_2(u)) < \infty$, using the Cauchy-Schwarz inequality we have

$$\begin{aligned} \Delta_2(u) &= O_p\{(nh_n^{-1})^{\frac{1}{2}}\}\|\boldsymbol{\psi}_2(u)\|_1 \\ &= O_p\{(nh_n^{-1})^{\frac{1}{2}}\}\|\boldsymbol{\psi}_2(u)\|_2 \\ &= O_p\{(nh_n^{-1})^{\frac{1}{2}}\}\left(\sum_{j=1}^{C}\sum_{l=d_{j+1}^0}^{d} \beta_{jl}^2(u)\right)^{1/2} \end{aligned} \tag{A.54}$$

which then implies that

$$\sum_{t=1}^{n} \Delta_2(U_t) = O_p\{(nh_n^{-1})^{\frac{1}{2}}\} \sum_{t=1}^{n} \|\boldsymbol{\psi}_2(U_t)\|_2$$

or using the Cauchy-Schwarz inequality,

$$\begin{aligned} \sum_{t=1}^{n} \Delta_2(U_t) &= O_p\left(\frac{n}{\sqrt{h_n}}\right)\left\{\sum_{t=1}^{n} \|\boldsymbol{\psi}_2(U_t)\|_2^2\right\}^{1/2} \\ &= O_p\left(\frac{n^{3/2}}{\sqrt{h_n}}\right)\left\{\sum_{j=1}^{c}\sum_{l=d_{j+1}^0}^{d} \|\boldsymbol{b}_{jl}\|_2 / \sqrt{n}\right\}. \end{aligned} \tag{A.55}$$

For the first term in (A.52), we have that

$$\begin{aligned} &\Delta_1(u) \\ &= \sum_{i=1}^{n} K_{h_n}(U_i - u)\left[\frac{\partial l((\boldsymbol{\psi}_1(u), \boldsymbol{\xi}(u)), \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\psi}_2(u)} - \frac{\partial l((\boldsymbol{\psi}_1(u), \mathbf{0}), \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\psi}_2(u)}\right]^{\top} \boldsymbol{\psi}_2(u) \\ &\quad + \sum_{i=1}^{n} K_{h_n}(U_i - u)\left[\frac{\partial l((\boldsymbol{\psi}_1(u), \mathbf{0}), \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\psi}_2(u)} - \frac{\partial l((\boldsymbol{\psi}_1^0(u), \mathbf{0}), \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\psi}_2(u)}\right]^{\top} \boldsymbol{\psi}_2(u). \end{aligned}$$

By the mean value theorem, there exist vectors $\boldsymbol{\xi}^*(u)$ and $\boldsymbol{\psi}_1^*(u)$ respectively on segments between $\mathbf{0}$ and $\boldsymbol{\xi}(u)$, and $\boldsymbol{\psi}_1(u)$ and $\boldsymbol{\psi}_1^0(u)$ such that, by Condition (RC.3),

$$\begin{aligned} &\sum_{i=1}^{n} K_{h_n}(U_i - u)\left[\frac{\partial l((\boldsymbol{\psi}_1(u), \boldsymbol{\xi}(u)), \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\psi}_2(u)} - \frac{\partial l((\boldsymbol{\psi}_1(u), \mathbf{0}), \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\psi}_2(u)}\right] \\ &= \sum_{i=1}^{n} K_{h_n}(U_i - u)\left[l''\left((\boldsymbol{\psi}_1(u), \boldsymbol{\xi}^*(u)), \boldsymbol{X}_i, Y_i\right)\right]\boldsymbol{\xi}(u) = O_p(n)\|\boldsymbol{\xi}(u)\|_2 \end{aligned}$$

and

$$\sum_{i=1}^{n} K_{h_n}(U_i - u) \left[\frac{\partial l((\boldsymbol{\psi}_1(u), \mathbf{0}), \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\psi}_2(u)} - \frac{\partial l((\boldsymbol{\psi}_1^0(u), \mathbf{0}), \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\psi}_2(u)} \right]$$
$$= \sum_{i=1}^{n} K_{h_n}(U_i - u)\ l''((\boldsymbol{\psi}_1^*(u), \mathbf{0}), \boldsymbol{X}_i, Y_i) \left[\boldsymbol{\psi}_1(u) - \boldsymbol{\psi}_1^0(u) \right] = O_p(n) \left\| \boldsymbol{\psi}_1(u) - \boldsymbol{\psi}_1^0(u) \right\|_2 .$$

Thus, together with the above order assessment, and again the Cauchy-Schwarz inequality, we have that

$$\Delta_1(u) = O_p(n) \left\{ \|\boldsymbol{\xi}(u)\|_2 + \left\| \boldsymbol{\psi}_1(u) - \boldsymbol{\psi}_1^0(u) \right\|_2 \right\} \|\boldsymbol{\psi}_2(u)\|_2 . \tag{A.56}$$

This implies that

$$\sum_{t=1}^{n} \Delta_1(U_t) = O_p(n) \sum_{t=1}^{n} \left\{ \|\boldsymbol{\xi}(U_t)\|_2 + \left\| \boldsymbol{\psi}_1(U_t) - \boldsymbol{\psi}_1^0(U_t) \right\|_2 \right\} \|\boldsymbol{\psi}_2(U_t)\|_2 . \tag{A.57}$$

By the Cauchy–Schwarz inequality,

$$\sum_{t=1}^{n} \left\{ \|\boldsymbol{\xi}(U_t)\|_2 + \left\| \boldsymbol{\psi}_1(U_t) - \boldsymbol{\psi}_1^0(U_t) \right\|_2 \right\} \quad \|\boldsymbol{\psi}_2(U_t)\|_2$$
$$\leqslant \left\{ \sum_{t=1}^{n} \left[\|\boldsymbol{\xi}(U_t)\|_2^2 + \|\boldsymbol{\psi}_1(U_t) - \boldsymbol{\psi}_1^0(U_t)\|_2^2 \right] \right\}^{1/2} \left\{ \sum_{t=1}^{n} \|\boldsymbol{\psi}_2(U_t)\|_2^2 \right\}^{1/2}$$
$$= \{n(nh_n)^{-1}\}^{1/2} \left\{ \sum_{j=1}^{C} \sum_{l=d_j^0+1}^{d} \|\boldsymbol{b}_{jl}\|_2^2 \right\}^{1/2} \leqslant h_n^{-1/2} \left\{ \sum_{j=1}^{C} \sum_{l=d_j^0+1}^{d} \|\boldsymbol{b}_{jl}\|_2 \right\} .$$

Combing the above results, for some constant C_1, we arrive at,

$$\sum_{t=1}^{n} \Delta_1(U_t) = \frac{C_1 n^{3/2}}{\sqrt{h_n}} \left\{ \sum_{j=1}^{C} \sum_{l=d_j^0+1}^{d} \|\boldsymbol{b}_{jl}\|_2 / \sqrt{n} \right\} . \tag{A.58}$$

By (A.55), (A.58), and (A.50), we have that, for large n,

$$L_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2}); \lambda_n, h_n\big) - L_n\big((\boldsymbol{\Psi}_{n1}, \mathbf{0}); \lambda_n, h_n\big) = \frac{C_2 n^{3/2}}{\sqrt{h_n}} \left\{ \sum_{j=1}^{C} \sum_{j=d_j^0+1}^{d} \|\boldsymbol{b}_{jl}\|_2 / \sqrt{n} \right\} \tag{A.59}$$

for some constant C_2. Finally, by using (A.49) and (A.59) in (A.48), for large n,

$$\begin{aligned}
&\widetilde{L}_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2}); \lambda_n, h_n\big) - \widetilde{L}_n\big((\boldsymbol{\Psi}_{n1}, \mathbf{0}); \lambda_n, h_n\big) \\
&\quad = \frac{C_2 n^{3/2}}{\sqrt{h_n}} \left\{ \sum_{j=1}^{C} \sum_{l=d_j^0+1}^{d} \|\boldsymbol{b}_{jl}\|_2/\sqrt{n} \right\} - \sum_{j=1}^{C} \sum_{l=d_j^0+1}^{d} p_n\left(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}; \lambda_n\right) \\
&\quad = \sum_{j=1}^{C} \sum_{j=d_j^0+1}^{d} \left\{ \frac{C_2 n^{3/2}}{\sqrt{h_n}} \|\boldsymbol{b}_{jl}\|_2/\sqrt{n} - p_n\left(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}; \lambda_n\right) \right\} < 0.
\end{aligned}$$

The last inequality is due to Condition (PC.3) on the penalty function p_n. This completes the proof of (A.47).

Step 2. Consider the penalized local log-likelihood function $\widetilde{L}_n\big((\boldsymbol{\Psi}_{n1}, \mathbf{0}); \boldsymbol{\lambda}_n, h_n\big)$, which is only a function of $\boldsymbol{\Psi}_{n1}$. Let $(\widehat{\boldsymbol{\Psi}}_{n1}, \mathbf{0})$ be its maximizer. Then for any $\boldsymbol{\Psi}_n = (\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2})$ with the property (A.47) shown in **Step 1**, we have that,

$$\begin{aligned}
&L_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2}); \boldsymbol{\lambda}_n, h_n\big) - L_n\big((\widehat{\boldsymbol{\Psi}}_{n1}, \mathbf{0}); \boldsymbol{\lambda}_n, h_n\big) \\
&\qquad = \big[L_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2}); \lambda_n, h_n\big) - L_n\big((\boldsymbol{\Psi}_{n1}, \mathbf{0}); \boldsymbol{\lambda}_n, h_n\big)\big] \\
&\qquad\quad - \Big[L_n\big((\widehat{\boldsymbol{\Psi}}_{n1}, \mathbf{0}); \boldsymbol{\lambda}_n, h_n\big) - L_n\big((\boldsymbol{\Psi}_{n1}, \mathbf{0}); \lambda_n, h_n\big)\Big] \\
&\qquad \leqslant L_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2}); \boldsymbol{\lambda}_n, h_n\big) - L_n\big((\widehat{\boldsymbol{\Psi}}_{n1}, \mathbf{0}); \boldsymbol{\lambda}_n, h_n\big) < 0,
\end{aligned}$$

where the last two inequalities are due to the definition of $(\widehat{\boldsymbol{\Psi}}_{n1}, \mathbf{0})$ and the result of **Step 1**. Hence, in the $\sqrt{nh_n}$-neighbourhood of $\boldsymbol{\Psi}_n^0$ guaranteed by Theorem 1, the maximizer of the penalized local log-likelihood $L_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2}); \boldsymbol{\lambda}_n, h_n\big)$ satisfies $\widehat{\boldsymbol{\Psi}}_{n2} = \mathbf{0}$, with probability tending to one, as $n \to \infty$. This completes the proof of Lemma 1.

Proof of Theorem 2

Proof. **(i)** Recall the partitioning of the true parameter matrix $\boldsymbol{\Psi}_n^0 = (\boldsymbol{\Psi}_{n1}^0, \boldsymbol{\Psi}_{n2}^0)$ such that $\boldsymbol{\Psi}_{n2}^0$ corresponds to all the zero regression functions, while $\boldsymbol{\Psi}_{n1}^0$ corresponds to the nonzero regression functions, among other parameters, given in Section 4 of the paper. Consider any matrix $\boldsymbol{\Psi}_n = (\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2})$ in a neighbourhood of $\boldsymbol{\Psi}_n^0$ such that $n^{-1}\|\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^0\|_F^2 = O_p\big((nh_n)^{-1}\big)$, and $\dim(\boldsymbol{\Psi}_{n1}) = \dim(\boldsymbol{\Psi}_{n1}^0)$ and $\dim(\boldsymbol{\Psi}_{n2}) = \dim(\boldsymbol{\Psi}_{n2}^0)$. Note that such choice is guaranteed

by Theorem 1. Also, for the t-th row of this matrix we have that $\|\boldsymbol{\psi}(U_t) - \boldsymbol{\psi}^0(U_t)\|_2 = O_p\big((nh_n)^{-1/2}\big), t = 1, \ldots, n$. On the other hand, for any $u \in \mathcal{U}$, let u^* be its nearest neighbourhood among the observed index values $U_1, \ldots, U_n$, that is $u^* = \text{argmin}_{\tilde{u} \in \{U_1, U_2, \ldots, U_n\}} |u - \tilde{u}|$. Under Condition (RC.2), we have that $\|\boldsymbol{\psi}^0(u) - \boldsymbol{\psi}^0(u^*)\|_2 = O_p(\log n/n)$, (Janson, 1987). Using the triangle inequality, we then have that $\|\boldsymbol{\psi}(u) - \boldsymbol{\psi}^0(u)\|_2 \leqslant \|\boldsymbol{\psi}(u) - \boldsymbol{\psi}(u^*)\|_2 + \|\boldsymbol{\psi}(u^*) - \boldsymbol{\psi}^0(u^*)\|_2 + \|\boldsymbol{\psi}^0(u^*) - \boldsymbol{\psi}^0(u)\|_2$. Thus, for any $u \in \mathcal{U}$, we have that $\|\boldsymbol{\psi}(u) - \boldsymbol{\psi}^0(u)\|_2 = O_p\big((nh_n)^{-1/2}\big)$. We provide the proof in two steps as follows.

Step 1: For any $u \in \mathcal{U}$, using the partitioning $\boldsymbol{\psi}(u) = (\boldsymbol{\psi}_1(u), \boldsymbol{\psi}_2(u))$ for any choice $\boldsymbol{\psi}(u)$ in the above neighbourhood, consider the local log-likelihood given in (7) of the paper,

$$\ell_n\big((\boldsymbol{\psi}_1(u), \boldsymbol{\psi}_2(u)); h_n\big) = \sum_{i=1}^{n} l\big((\boldsymbol{\psi}_1(u), \boldsymbol{\psi}_2(u)), \boldsymbol{X}_i, Y_i\big) K_{h_n}(U_i - u).$$

We now denote the following two penalized local log-likelihoods,

$$\begin{aligned}
\tilde{\ell}_n\big(\boldsymbol{\psi}(u); h_n, \boldsymbol{\lambda}_n\big) &= \ell_n\big((\boldsymbol{\psi}_1(u), \boldsymbol{\psi}_2(u)); h_n\big) - \frac{1}{n} \sum_{j=1}^{C} \sum_{l=1}^{d} p_n\big(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}; \lambda_{nj}\big), \\
\tilde{\ell}_n\big(\boldsymbol{\psi}_1(u); h_n, \boldsymbol{\lambda}_n\big) &= \ell_n\big((\boldsymbol{\psi}_1(u), \mathbf{0}); h_n\big) - \frac{1}{n} \sum_{j=1}^{C} \sum_{l=1}^{d_j^0} p_n\big(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}; \lambda_{nj}\big).
\end{aligned}$$

In what follows, for any $u \in \mathcal{U}$, we assess the order of the following difference for large n,

$$\begin{aligned}
D_n(u) &= \tilde{\ell}_n\big(\boldsymbol{\psi}(u); h_n, \boldsymbol{\lambda}_n\big) - \tilde{\ell}_n\big(\boldsymbol{\psi}_1(u); h_n, \boldsymbol{\lambda}_n\big) \\
&= \big[\ell_n\big((\boldsymbol{\psi}_1(u), \boldsymbol{\psi}_2(u)); h_n\big) - \ell_n\big((\boldsymbol{\psi}_1(u), \mathbf{0}); h_n\big)\big] \\
&\quad - \frac{1}{n} \sum_{j=1}^{C} \sum_{l=d_j^0+1}^{d} p_n\big(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}; \lambda_{nj}\big) \\
&= \Delta_n(u) - \frac{1}{n} \sum_{j=1}^{C} \sum_{l=d_j^0+1}^{d} p_n\big(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}; \lambda_{nj}\big),
\end{aligned}$$

where $\Delta_n(u)$ is the difference in the local log-likelihood which is also used in the proof of Lemma 1, given in (A.51). Recall the following representation of $\Delta_n(u)$ given in (A.52),

$$\Delta_n(u) = \Delta_1(u) + \Delta_2(u).$$

From (A.54), we have that (up to a constant), for large enough n

$$|\Delta_2(u)| \leqslant \sqrt{\frac{n}{h_n}} \sum_{j=1}^{C} \sum_{l=d^0_{j+1}}^{d} |\beta_{jl}(u)|.$$

Under Condition (RC.2), we have that $\|\boldsymbol{\psi}^0(u) - \boldsymbol{\psi}^0(u^*)\|_2 = O_p(\frac{\log n}{n})$. Using (A.56), for large n, we also have that (up to a constant), for any $u \in \mathcal{U}$,

$$|\Delta_1(u)| \leqslant \sqrt{\frac{n}{h_n}} \sum_{j=1}^{C} \sum_{l=d^0_{j+1}}^{d} |\beta_{jl}(u)|.$$

Therefore, for any $u \in \mathcal{U}$, for large n we have

$$D_n(u) \leqslant \sqrt{\frac{n}{h_n}} \sum_{j=1}^{C} \sum_{l=d^0_{j+1}}^{d} \left\{ |\beta_{jl}(u)| - \frac{\sqrt{h_n}\; p_n\big(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}; \lambda_{nj}\big)}{n^{3/2}} \right\}.$$

Also, for any $u \in \mathcal{U}$ and its corresponding u^* as defined above, for large n we have that

$$\begin{aligned} |\beta_{jl}(u)| &\leqslant |\beta_{jl}(u) - \beta_{jl}(u^*)| + |\beta_{jl}(u^*)| \\ &= O_p\left(\log n/n\right) + \|\boldsymbol{b}_{jl}\|_2/\sqrt{n} + O_p(\{nh_n\}^{-1/2}) \\ &= \|\boldsymbol{b}_{jl}\|_2/\sqrt{n} + o_p(1). \end{aligned}$$

Therefore, for large n we have that by the Condition (PC.3) on the penalty, $D_n(u) < 0$, uniformly in $u \in \mathcal{U}$.

Step 2: For any $u \in \mathcal{U}$, let $\widehat{\boldsymbol{\psi}}_{n1}(u)$ be the maximizer of $\tilde{\ell}_n\big(\boldsymbol{\psi}_1(u); h_n, \boldsymbol{\lambda}_n\big)$ which is considered as only a function of $\boldsymbol{\psi}_1(u)$. Then, for any $u \in \mathcal{U}$,

$$\begin{aligned} &\tilde{\ell}_n\big(\boldsymbol{\psi}(u); h_n, \boldsymbol{\lambda}_n\big) - \tilde{\ell}_n\big(\widehat{\boldsymbol{\psi}}_1(u); h_n, \boldsymbol{\lambda}_n\big) \\ &= \left\{ \tilde{\ell}_n\big(\boldsymbol{\psi}(u); h_n, \boldsymbol{\lambda}_n\big) - \tilde{\ell}_n\big(\boldsymbol{\psi}_1(u); h_n, \boldsymbol{\lambda}_n\big) \right\} \\ &\quad - \left\{ \tilde{\ell}_n\big(\widehat{\boldsymbol{\psi}}_{n1}(u); h_n, \boldsymbol{\lambda}_n\big) - \tilde{\ell}_n\big(\boldsymbol{\psi}_1(u); h_n, \boldsymbol{\lambda}_n\big) \right\} < 0. \end{aligned}$$

This difference is indeed negative uniformly over $u \in \mathcal{U}$. This implies that with probability tending to one, as $n \to \infty$, $\widehat{\boldsymbol{\psi}}_{n2}(u) = \mathbf{0}$, uniformly in u. Similar to the result of Theorem 1, $\widehat{\boldsymbol{\psi}}_{n1}(u)$ is a consistent estimator of $\boldsymbol{\psi}^0_1(u)$, for any $u \in \mathcal{U}$.

(ii) For any $u \in \mathcal{U}$, recall the oracle estimator $\widehat{\boldsymbol{\psi}}_{n1,orc}(u)$ which is the MLLE of $\boldsymbol{\psi}^0_1(u)$, having known $\boldsymbol{\psi}^0_2(u) = \mathbf{0}$ a priori. This estimator is the maximizer of the local log-likelihood

$\ell\big(\boldsymbol{\psi}(u); h_n\big)$ in (7) of the paper, and it satisfies the estimating equation

$$\ell_n'(\widehat{\boldsymbol{\psi}}_{n1,orc}(u); h_n) = \mathbf{0}.$$

Using a Taylor's expansion around $\boldsymbol{\psi}_1^0(u)$, we arrive at

$$\ell_n'(\boldsymbol{\psi}_1^0(u); h_n) + \ell_n''(\tilde{\boldsymbol{\psi}}_1^0(u); h_n)\big(\widehat{\boldsymbol{\psi}}_{n1,orc}(u) - \boldsymbol{\psi}_1^0(u)\big) = \mathbf{0},$$

where $\tilde{\boldsymbol{\psi}}_1^0(u)$ lies on a segment between $\boldsymbol{\psi}_1^0(u)$ and $\widehat{\boldsymbol{\psi}}_{n1,orc}(u)$, for any $u \in \mathcal{U}$. By Proposition 1 of the paper, since $\widehat{\boldsymbol{\psi}}_{n1,orc}(u)$ is a consistent estimator of $\boldsymbol{\psi}_1^0(u)$, so is $\tilde{\boldsymbol{\psi}}_1^0(u)$. Thus, for large n,

$$\ell_n'(\boldsymbol{\psi}_1^0(u); h_n) + \big\{\ell_n''(\boldsymbol{\psi}_1^0(u); h_n) + o_p(n)\big\}\big(\widehat{\boldsymbol{\psi}}_{n1,orc}(u) - \boldsymbol{\psi}_1^0(u)\big) = \mathbf{0}. \tag{A.60}$$

Similarly, the MPLLE $\widehat{\boldsymbol{\psi}}_{n1}(u)$ also satisfies the estimation equation

$$\tilde{\ell}_n'\big(\widehat{\boldsymbol{\psi}}_{n1}(u); h_n, \boldsymbol{\lambda}_n\big) = \mathbf{0}$$

or equivalently,

$$\ell_n'\big(\widehat{\boldsymbol{\psi}}_{n1}(u); h_n\big) - \boldsymbol{P}_n'(\widehat{\boldsymbol{\Psi}}_{n1}; \boldsymbol{\lambda}_n) = \mathbf{0},$$

where

$$\boldsymbol{P}_n(\boldsymbol{\Psi}_{n1}; \boldsymbol{\lambda}_n) = \frac{1}{n}\sum_{j=1}^{C}\sum_{l=1}^{d_j^0} p_n\big(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}; \lambda_{nj}\big) = \mathbb{P}_n(\boldsymbol{\Psi}_{n1}; \boldsymbol{\lambda}_n)/n. \tag{A.61}$$

Using a Taylor's expansion around $\boldsymbol{\psi}_1^0(u)$, and consistency of $\widehat{\boldsymbol{\psi}}_{n1}(u)$, for large n,

$$\begin{aligned}\big\{\ell_n'(\boldsymbol{\psi}_1^0(u); h_n) \;&-\; [\boldsymbol{P}_n'(\boldsymbol{\theta}_1^0; \boldsymbol{\lambda}_n) + o_p(n)]\big\} \\ &+\; \big\{\ell_n''(\boldsymbol{\psi}_1^0(u); h_n) - \boldsymbol{P}_n''(\boldsymbol{\theta}_1^0; \boldsymbol{\lambda}_n) + o_p(n)\big\}\big(\widehat{\boldsymbol{\psi}}_{n1}(u) - \boldsymbol{\psi}_1^0(u)\big) \\ &=\; \mathbf{0},\end{aligned} \tag{A.62}$$

where $\boldsymbol{P}_n'$ and $\boldsymbol{P}_n''$ are the gradient and Hessian of the penalty $\boldsymbol{P}_n$ in (A.61) with respect to the parameter vector $\boldsymbol{\psi}_1(u)$, and $\boldsymbol{\theta}_1^0 = \big\{\theta_{jl}^0 = \sqrt{E_0\{\beta_{jl}^0(U)\}^2} : 1 \leqslant j \leqslant C, 1 \leqslant l \leqslant d_j^0\big\}$.

By (A.60) and (A.62), we have that

$$\begin{aligned}
&\widehat{\boldsymbol{\psi}}_{n1}(u) - \widehat{\boldsymbol{\psi}}_{n1,orc}(u) \\
&= \left\{\ell_n''(\boldsymbol{\psi}_1^0(u); h_n) + o_p(n)\right\}^{-1} \ell_n'(\boldsymbol{\psi}_1^0(u); h_n) \\
&\quad - \left\{\ell_n''(\boldsymbol{\psi}_1^0(u); h_n) - \boldsymbol{P}_n''(\boldsymbol{\theta}_1^0; \boldsymbol{\lambda}_n) + o_p(n)\right\}^{-1} \left\{\ell_n'(\boldsymbol{\psi}_1^0(u); h_n) - \boldsymbol{P}_n'(\boldsymbol{\theta}_1^0; \boldsymbol{\lambda}_n)\right\} \\
&= \left\{\left\{\ell_n''(\boldsymbol{\psi}_1^0(u); h_n)/n + o_p(1)\right\}^{-1} - \left\{\ell_n''(\boldsymbol{\psi}_1^0(u); h_n)/n - \boldsymbol{P}_n''(\boldsymbol{\theta}_1^0; \boldsymbol{\lambda}_n)/n + o_p(1)\right\}^{-1}\right\} \\
&\quad \times \left\{\frac{\ell_n'(\boldsymbol{\psi}_1^0(u); h_n)}{n}\right\} \\
&\quad + \left\{\ell_n''(\boldsymbol{\psi}_1^0(u); h_n)/n - \boldsymbol{P}_n''(\boldsymbol{\theta}_1^0; \boldsymbol{\lambda}_n)/n + o_p(1)\right\}^{-1} \left\{\frac{\boldsymbol{P}_n'(\boldsymbol{\theta}_1^0; \boldsymbol{\lambda}_n)}{n} + o_p(1)\right\} \\
&= \boldsymbol{V}_{1n}(u) + \boldsymbol{V}_{2n}(u),
\end{aligned}$$

where $\boldsymbol{V}_{1n}(u)$ and $\boldsymbol{V}_{2n}(u)$ are respectively the two vectors on the right hand side of the above equation. By (A.25), for any $u \in \mathcal{U}$ and for large n,

$$\ell_n''(\boldsymbol{\psi}_1^0(u); h_n)/n = -m(u)\boldsymbol{I}(u)(1 + o_p(1)). \tag{A.63}$$

Thus, due to the finiteness condition of $m(u)\boldsymbol{I}(u)$ and since $r_{2n} = o(1)$ by Condition (PC.2) on the penalty, we have (in a matrix form)

$$\left\{\ell_n''(\boldsymbol{\psi}_1^0(u); h_n)/n + o_p(1)\right\}^{-1} - \left\{\ell_n''(\boldsymbol{\psi}_1^0(u); h_n)/n - \boldsymbol{P}_n''(\boldsymbol{\theta}_1^0; \boldsymbol{\lambda}_n)/n + o_p(1)\right\}^{-1} = o_p(1). \tag{A.64}$$

Also, by (A.14) we have

$$E\left\{\frac{\ell_n'(\boldsymbol{\psi}_1^0(u); h_n)}{n}\right\} = \frac{1}{2}\mathcal{K}_2 m(u)\left\{\Delta''(u;u) + 2\Delta'(u;u)m'(u)/m(u)\right\}h_n^2\{1 + o(1)\} < \infty \tag{A.65}$$

for any $u \in \mathcal{U}$, and $\mathcal{K}_2 = \int_{\mathcal{U}} t^2 K(t)dt < \infty$. Using a similar approach to show (A.26) and the centralizing technique as in (A.36), and since $nh_n^5 = O(1)$, we have that

$$\sup_{u\in\mathcal{U}} \left\|\ell_n'(\boldsymbol{\psi}_1^0(u); h_n)/n\right\|_2 = O_p\left\{\left(\frac{\log(\frac{1}{h_n})}{nh_n}\right)^{\frac{1}{2}}\right\} + O_p(h_n^2) = O_p\left\{\left(\frac{\log(\frac{1}{h_n})}{nh_n}\right)^{\frac{1}{2}}\right\}. \tag{A.66}$$

Thus, by (A.64) and (A.66),

$$\sup_{u\in\mathcal{U}} \|\boldsymbol{V}_{1n}(u)\|_2 = o_p\left\{\left(\frac{\log(\frac{1}{h_n})}{nh_n}\right)^{\frac{1}{2}}\right\} = o_p(\xi_n). \tag{A.67}$$

Also, by the positive definiteness of $m(u)\boldsymbol{I}(u)$ and condition $r_{2n} = o(1)$, we have that

$$\sup_{u\in\mathcal{U}} \|\boldsymbol{V}_{2n}(u)\|_2 = O_p\left\{r_{1n}/\sqrt{nh_n}\right\}. \tag{A.68}$$

Thus, using (A.67) and (A.68) we have

$$\begin{aligned}
&\{(1+r_{1n})\xi_n\}^{-1} \sup_{u\in\mathcal{U}} \|\widehat{\boldsymbol{\psi}}_{n1}(u) - \widehat{\boldsymbol{\psi}}_{n1,orc}(u)\|_2 \\
&\quad\leqslant \{(1+r_{1n})\xi_n\}^{-1}\left\{\sup_{u\in\mathcal{U}} \|\boldsymbol{V}_{1n}(u)\|_2 + \sup_{u\in\mathcal{U}} \|\boldsymbol{V}_{2n}(u)\|_2\right\} \\
&\quad= \{(1+r_{1n})\}^{-1}o_p(1) + \frac{r_{1n}}{1+r_{1n}}\sqrt{\frac{1}{\log h_n^{-1}}} = o_p(1).
\end{aligned}$$

This completes the proof.

(iii) From (A.62) and using (A.63), and the relation (A.61) we have

$$\left\{m(u)\boldsymbol{I}(u) + \frac{\mathbb{P}_n''(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n)}{n^2} + o_p(1)\right\}\left(\widehat{\boldsymbol{\psi}}_{n1}(u) - \boldsymbol{\psi}_1^0(u)\right) + \frac{\mathbb{P}_n(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n)}{n^2} = \frac{\ell_n'(\boldsymbol{\psi}_1^0(u);h_n)}{n}. \tag{A.69}$$

Using the second-order moment calculations in the proof of Theorem 2, we have

$$Var\left\{\frac{\ell_n'(\boldsymbol{\psi}_1^0(u);h_n)}{n}\right\} = \frac{\mathcal{V}_0}{nh_n}\boldsymbol{I}(u)m(u)(1+o_p(1)) < \infty, \tag{A.70}$$

where $\mathcal{V}_0 = \int_{\mathcal{U}} K^2(t)dt < \infty$.

Using (A.65) and (A.70), we standardize both sides of the equation (A.69). Hence, by the multivariate central limit theorem we have that

$$\sqrt{nh_n}\left\{\left[\boldsymbol{I}(u) + m^{-1}(u)\frac{\mathbb{P}_n''(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n)}{n^2}\right]\left(\widehat{\boldsymbol{\psi}}_{n1}(u) - \boldsymbol{\psi}_1^0(u)\right) - \boldsymbol{B}_n(u)\right\} \sim N\left(\mathbf{0}, \mathcal{V}_0 m^{-1}(u)\boldsymbol{I}(u)\right),$$

where

$$\boldsymbol{B}_n(u) = \frac{1}{2}\mathcal{K}_2\{\Delta''(u;u) + 2\Delta'(u;u)m'(u)/m(u)\}h_n^2 - m^{-1}(u)\frac{\mathbb{P}_n'(\boldsymbol{\theta}_1^0;\boldsymbol{\lambda}_n)}{n^2} + o_p(h_n^2).$$

Web Appendix E. Remarks

REMARK 1: The sparsity assumption in Equation (5) of the paper is indeed the most common assumption in the variable selection literature. The implication is that the underlying data-generating process is a simple sparse model and is beneficial for interpretation

purposes, specifically in mixture regressions when the number of covariates is relatively large. Together with (5) and under Conditions (PC.2) and (PC.3) on the penalty function p_n in Web Appendix B , we achieve the selection consistency property as stated in Theorem 2(i). Essentially, Conditions (PC.2) and (PC.3) imply that only those coefficients that satisfy $E\{\beta_{jl}(U)\} > \lambda_n$, where $\lambda_n \to 0$ as $n \to \infty$, are detectable by the regularization methods and will be estimated as non-zero, while the ones below the threshold λ_n, so-called weak signals, will most likely be estimated as zero. In a sense, the aforementioned conditions are similar to the ones considered by Wei et al. (2011) with η_1 replaced by λ_n. Wei et al. (2011) refer to (5) as the narrow sparsity condition ($\eta_1 = 0$) under which they also show that, for example, AdpLASSO has the selection consistency property. By changing (5) to the one used in Wei et al. (2011) with $\eta_1 > 0$, it turns out that regularization methods only achieve certain estimation error bounds but not really selection consistency as those weak-signal regression parameters most likely will be estimated as zero. Hence, since our main result focuses on the variable selection consistency property, we have decided to keep the current definition of sparsity (5) in the FM-VCR models. Nevertheless, it is worth noting that Fang et al. (2021) proposed a two-step procedure based on both variable selection and ridge regression estimators that were shown to be capable of detecting weak signals and providing an estimation of both strong and weak signals.

REMARK 2: The asymptotic normality result in Theorem 2(iii) is obtained as if the true sparse structure of the model is known in advance, i.e., an oracle's perspective. In general, its use in practice to perform likelihood-ratio type inference is reserved (referred to as naive inference) as the true sparse structure of the model is not known in advance and it is estimated by the penalization method. As a result, in the partitioning $\widehat{\boldsymbol{\psi}}_n(u) = \big(\widehat{\boldsymbol{\psi}}_{n1}(u), \widehat{\boldsymbol{\psi}}_{n2}(u)\big)$ introduced before Theorem 2 of the paper (oracle perspective), in practice due to the variable selection stage the dimension of the sub-vector $\widehat{\boldsymbol{\psi}}_{n1}(u)$ is indeed random and may not be equal

to the dimension of true non-zero vector $\boldsymbol{\psi}_1^0(u)$, and hence asymptotically normal distribution may be distorted. The extra variability due to the variable selection needs to be taken into account for further inference and is a part of the general topic of post-selection inference (PoSI, Berk et al. (2013)), which sees a surge of research in recent years for (generalized) linear regression models (see Zhang et al. (2022) for a recent survey), although little for the FMR and FM-VCR models.

REMARK 3: Recall the penalty p_n in (10) of the paper. If the AdpLASSO is used, we have

$$p_n(\|\boldsymbol{b}_{jl}\|_2/\sqrt{n};\lambda_n)/n^2 = \lambda_n w_{jl}\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}, \quad j=1,\ldots,C, \quad l=1,\ldots,d.$$

In practice, one needs to specify the weights w_{jl}. As in Zou (2006), we use $w_{jl} = \{\|\breve{\boldsymbol{b}}_{jl}\|_2/\sqrt{n}\}^{-\gamma}$, for some $\gamma > 0$, where $\|\breve{\boldsymbol{b}}_{jl}\|_2^2 = \sum_{t=1}^n \breve{\beta}_{jl}^2(u_t)$ and $\breve{\beta}_{jl}(u_t)$'s are the consistent MLLE discussed in Section 3 of the paper. Thus, as $n \to \infty$, the condition $\lambda_n \max_{1\leqslant j\leqslant C, 1\leqslant l\leqslant d_j^0} w_{jl} = o_p(\{nh_n\}^{-1/2})$ in Theorem 1 becomes $\sqrt{nh_n}\lambda_n \to 0$, and the sparsity condition $\sqrt{nh_n}\lambda_n \min_{1\leqslant j\leqslant C, d_j^0+1\leqslant l\leqslant d} w_{jl} \to \infty$ in Theorem 2 becomes $(nh_n)^{(\gamma+1)/2}\lambda_n \to \infty$.

REMARK 4: For the AdpLASSO, the difference $D_{n,II}(\boldsymbol{W}_n)$ in (A.41) has a simpler upper bound as follows. By the definition of the penalty, we have that,

$$\begin{aligned} D_{n,II}(\boldsymbol{W}_n) &\leqslant h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{l=1}^{d_j^0} \left\{ p_n\left(\left\|\boldsymbol{b}_{jl}^0 + \gamma_n \boldsymbol{v}_{jl}\right\|/\sqrt{n};\lambda_{nj}\right) - p_n\left(\left\|\boldsymbol{b}_{jl}^0\right\|/\sqrt{n};\lambda_{nj}\right)\right\} \\ &\leqslant h_n n^{-1} \sum_{j=1}^{\mathcal{C}} \sum_{j=1}^{d_j^0} n^2 \gamma_n \lambda_{nj} w_{jl} \|\boldsymbol{v}_{jl}\|/\sqrt{n} \\ &\leqslant M_\epsilon \{nh_n\}^{1/2} \max_{1\leqslant j\leqslant \mathcal{C}, 1\leqslant l\leqslant d_j^0} \{\lambda_{nj} w_{jl}\} \end{aligned}$$

which is dominated by the likelihood difference $D_{n,I}(\boldsymbol{W}_n)$ in (A.2) under the condition

$$\{nh_n\}^{1/2} \max_{1\leqslant j\leqslant \mathcal{C}, 1\leqslant l\leqslant d_j^0} \{\lambda_{nj} w_{jl}\} = o_p(1).$$

REMARK 5: For the AdpLASSO, the penalty difference in (A.49) becomes,

$$\mathbb{P}_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{\Psi}_{n2}); \boldsymbol{\lambda}_n\big) - \mathbb{P}_n\big((\boldsymbol{\Psi}_{n1}, \boldsymbol{0}); \boldsymbol{\lambda}_n\big) = \sum_{j=1}^{C} \sum_{j=d_j^0+1}^{d} n^2 \lambda_{nj} w_{jl} (\|\boldsymbol{b}_{jl}\|/\sqrt{n}).$$

Thus, the difference in (A.48) is bounded by

$$\begin{aligned}&\widetilde{L}_n\big((\boldsymbol{\Psi}_{n1},\boldsymbol{\Psi}_{n2});\lambda_n,h_n\big)-\widetilde{L}_n\big((\boldsymbol{\Psi}_{n1},\mathbf{0});\lambda_n,h_n\big)\\&\leqslant\sum_{j=1}^{C}\sum_{l=d_j^0+1}^{d}\left\{\frac{C_2n^{3/2}}{\sqrt{h_n}}\|\boldsymbol{b}_{jl}\|_2/\sqrt{n}-n^2\min_{1\leqslant j\leqslant\mathcal{C},d_j^0+1\leqslant l\leqslant d}\{\lambda_{nj}w_{jl}\}(\|\boldsymbol{b}_{jl}\|/\sqrt{n})\right\}<0\end{aligned}$$

under the condition $\sqrt{nh_n}\min_{1\leqslant j\leqslant\mathcal{C},d_j^0+1\leqslant l\leqslant d}\{\lambda_{nj}w_{jl}\}\to\infty$, as $n\to\infty$.

Web Appendix F. Details of the numerical algorithm and implementation

Web Appendix F.1 *Local quadratic approximation of the penalty functions*

The Local Quadratic Approximation (LQA) of a penalty function is given as follows.

$$\begin{aligned}&\overline{\mathbb{P}}_n(\boldsymbol{\Psi};\boldsymbol{\lambda})\\&\quad=\sum_{j=1}^{C}\sum_{l=1}^{d}\left\{p_n\left(\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2/\sqrt{n};\lambda_j\right)\right.\\&\qquad\left.+\frac{p_n'\left(\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2/\sqrt{n};\lambda_j\right)}{2\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2/\sqrt{n}}\left(\boldsymbol{b}_{jl}^{\top}\boldsymbol{b}_{jl}/n-\boldsymbol{b}_{jl}^{(m)\top}\boldsymbol{b}_{jl}^{(m)}/n\right)\right\}.\end{aligned}\tag{A.71}$$

For each (j,l) the derivative is given as

$$\frac{\partial\overline{\mathbb{P}}_n(\boldsymbol{\Psi};\boldsymbol{\lambda})}{\partial\boldsymbol{b}_{jl}}=\frac{p_n'\left(\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2/\sqrt{n};\lambda_j\right)}{\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2/\sqrt{n}}\times\frac{\boldsymbol{b}_{jl}}{n}.$$

Specifically, for the well-known penalty functions that are used in the main paper, we have the following expressions:

- LASSO:

$$\frac{\partial\overline{\mathbb{P}}_n(\boldsymbol{\Psi};\boldsymbol{\lambda})}{\partial\boldsymbol{b}_{jl}}=\frac{n^{3/2}\lambda_j}{\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2}\boldsymbol{b}_{jl}.$$

- AdpLASSO: Given the weights w_{jl} discussed in Remark 2 of the paper, which are based in the MLLE,

$$\frac{\partial\overline{\mathbb{P}}_n(\boldsymbol{\Psi};\boldsymbol{\lambda})}{\partial\boldsymbol{b}_{jl}}=\frac{n^{3/2}\lambda_jw_{jl}}{\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2}\boldsymbol{b}_{jl}.$$

- MCP:

$$\frac{\partial \overline{\mathbb{P}}_n(\boldsymbol{\Psi};\boldsymbol{\lambda})}{\partial \boldsymbol{b}_{jl}} = \begin{cases} \left[\dfrac{n^{3/2}\lambda_j}{\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2} - \frac{n}{\gamma}\right]\boldsymbol{b}_{jl}, & \left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2 \leqslant \sqrt{n}\gamma\lambda_j \\ \mathbf{0}, & \left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2 > \sqrt{n}\gamma\lambda_j \end{cases}$$

- SCAD:

$$= \begin{cases} \dfrac{n^{3/2}\lambda_j}{\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2}\boldsymbol{b}_{jl} & \left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2 \leqslant \sqrt{n}\lambda_j \\ \dfrac{n}{\gamma-1}\left(\dfrac{\sqrt{n}\gamma\lambda_j}{\left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2} - 1\right)\boldsymbol{b}_{jl}, & \sqrt{n}\lambda_j < \left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2 \leqslant \sqrt{n}\gamma\lambda_j \\ \mathbf{0}, & \left\|\boldsymbol{b}_{jl}^{(m)}\right\|_2 > \sqrt{n}\gamma\lambda_j \end{cases} \tag{A.72}$$

Web Appendix F.2 *The* EM *algorithm for the* MLLE *and* MPLLE *in Gaussian* FM-VCR

For a Gaussian FM-VCR, at any $u \in \mathcal{U}$, the local-kernel log-likelihood in (7) of the paper is given by

$$\ell_n\big(\boldsymbol{\psi}(u); h\big) = \sum\nolimits_{i=1}^{n} \log\left\{\sum\nolimits_{j=1}^{C} \pi_j N(y_i; \boldsymbol{x}^\top\boldsymbol{\beta}_j, \sigma_j^2)\right\} K_h(u_i - u), \tag{A.73}$$

where $N(y;\mu,\sigma^2)$ is the pdf of a Gaussian distribution with mean μ and σ^2.

The locally constant vector of parameters is $\boldsymbol{\psi}(u) = (\boldsymbol{\pi}^\top, \boldsymbol{\phi}^\top, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_C^\top)$, where $\boldsymbol{\pi}^\top = (\pi_1, \ldots, \pi_{C-1})$, $\boldsymbol{\phi}^\top = (\sigma_1^2, \ldots, \sigma_C^2)$ and $\boldsymbol{\beta}_j^\top = (\beta_{j1}, \ldots, \beta_{jd})$ for $j = 1, \ldots, C$. Note that the entries of the vector $\boldsymbol{\psi}(u)$ are local constant approximations of the functions $\beta_{jl}(u), l = 1, \ldots, d$, $\pi_j(u)$, and $\sigma_j^2(u)$, for all $j = 1, \ldots, C$. These entries clearly depend on u, and for simplicity, we suppress u in the notation but keep it for $\boldsymbol{\psi}(u)$.

Using (A.73), the corresponding (total) local-kernel log-likelihood in (8) of the paper is

given by

$$\begin{aligned} L_n(\boldsymbol{\Psi}; h) &= \sum\nolimits_{t=1}^{n} \ell_n\big(\boldsymbol{\psi}(u_t); h\big) \\ &= \sum\nolimits_{t,i=1}^{n} \log\Big\{\sum\nolimits_{j=1}^{C} \pi_{j,t}\ N\big(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}_{j,t}, \sigma_{j,t}^2\big)\Big\}\, K_h(u_i - u_t), \end{aligned} \tag{A.74}$$

where $\boldsymbol{\Psi}$ is the $(n \times \{C(d+2)-1\})$-dimensional matrix of the locally constant vectors $\boldsymbol{\psi}(u_i), i = 1, \ldots, n$.

Web Appendix F.2.1 *The EM algorithm for the* MLLE. We view the observed data $\{(u_i, \boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ as incomplete, and introduce the unobserved Bernoulli random variables Z_{ij} to represent the membership of the i-th observation to the j-th component of the mixture model, $\forall j = 1, \ldots, C$, such that $\Pr\left(Z_{ij} = 1 | u_i, \boldsymbol{x}_i, y_i\right) = \pi_j(u_i)$. The complete data consist of $\{(u_i, \boldsymbol{x}_i, y_i, \boldsymbol{z}_i) : i = 1, \ldots, n\}$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iC})^\top$.

At any arbitrary point $u \in \mathcal{U}$, the complete local log-likelihood function is given by

$$\begin{aligned} \ell_n^c\big(\boldsymbol{\psi}(u); h\big) &= \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{C} Z_{ij}\big\{\log \pi_j + \log N(y_i; \boldsymbol{x}^\top \boldsymbol{\beta}_j, \sigma_j^2)\big\} K_h(u_i - u) \\ &= \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{C} Z_{ij}\big\{\log \pi_j - \frac{1}{2\sigma_j^2}\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j\right)^2 - \frac{1}{2}\log(\sigma_j^2)\big\} K_h(u_i - u) \end{aligned}$$

Given the current value of the parameter $\boldsymbol{\psi}^{(m)}(u)$, at the $(m+1)$-th iteration, the EM algorithm proceeds in two steps as follows.

E-step: Since the Z_{ij}'s are unobservable, we compute the expectation of the complete local log-likelihood $\widetilde{l}_n^c$ with respect to Z_{ij} conditional on the observations $\{(u_i, \boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ and the current parameter values $\boldsymbol{\psi}^{(m)}(u)$. This boils down to the computation of the conditional expectations

$$w_{ij}^{(m)} = \frac{\pi_j^{(m)} N(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(m)}, \sigma_j^{(m)2})}{\sum_{j=1}^{C} \pi_j^{(m)} N(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(m)}, \sigma_j^{(m)2})}$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, C$.

M-step: We maximize the objective function

$$\sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{C} w_{ij}^{(m)}\big\{\log \pi_j - \frac{1}{2\sigma_j^2}\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j\right)^2 - \frac{1}{2}\log(\sigma_j^2)\big\} K_h(u_i - u)$$

with respect to $(\pi_j, \boldsymbol{\beta}_j, \sigma_j^2)$. The locally-constant parameter updates are,

$$\begin{aligned}
\pi_j^{(m+1)}(u) \equiv \pi_j^{(m+1)} &= \sum_{i=1}^{n} w_{ij}^{(m)} K_h(u_i - u) \Big/ \sum\nolimits_{i=1}^{n} K_h(u_i - u) \\
\boldsymbol{\beta}_j^{(m+1)}(u) \equiv \boldsymbol{\beta}_j^{(m+1)} &= \left(\boldsymbol{X}^\top \boldsymbol{W}_j^{(m)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W}_j^{(m)} \boldsymbol{y} \\
\sigma_j^{2(m+1)}(u) \equiv \sigma_j^{2(m+1)} &= \frac{\sum_{i=1}^{n} w_{ij}^{(m)} K_h(u_i - u) \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j^{(m+1)}\right)^2}{\sum_{i=1}^{n} w_{ij}^{(m)} K_h(u_i - u)},
\end{aligned}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, and

$$\boldsymbol{W}_j^{(m)}(u) \equiv \boldsymbol{W}_j^{(m)} = \operatorname{diag}\left\{ w_{ij}^{(m)} K_h(u_i - u); i = 1, \ldots, n \right\}.$$

Web Appendix F.2.2 *The EM algorithm for the* MPLLE. The complete (total) local log-likelihood is given by

$$L_n^c(\boldsymbol{\Psi}; h) = \sum\nolimits_{t=1}^{n} \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{C} Z_{ij} \big\{ \log \pi_{j,t} + \log f(y_i; \theta_{j,t}(\boldsymbol{x}_i), \phi_{j,t}) \big\} K_h(u_i - u_t).$$

Given the current value of the parameter $\boldsymbol{\Psi}^{(m)}$, at the $(m+1)^{\text{th}}$ iteration, the EM algorithm proceeds in two steps as follows.

E-step: Since the Z_{ij}'s are unobservable, we compute the expectation of the penalized complete local log-likelihood $\widetilde{L}_n^c$ with respect to Z_{ij} conditional on the observations $\{(u_i, \boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ and the current parameter values $\boldsymbol{\Psi}^{(m)}$. This indeed boils down to the computation of the conditional expectations

$$w_{ij}^{(m)} = \frac{\pi_j^{(m)}(u_i) f\big(y_i; \theta_j^{(m)}(\boldsymbol{x}_i, u_i), \phi_j^{(m)}(u_i)\big)}{\sum_{j=1}^{C} \pi_j^{(m)}(u_i) f\big(y_i; \theta_j^{(m)}(\boldsymbol{x}_i, u_i)), \phi_j^{(m)}(u_i)\big)}$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, C$.

M-step: The objective function $\widetilde{Q}$ in M-step of the modified EM algorithm in Section 5 of the paper becomes

$$\begin{aligned}
&\widetilde{Q}(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) \\
&\quad = \sum_{j=1}^{C} \sum_{t=1}^{n} \Bigg\{ \sum_{i=1}^{n} w_{ij}^{(m)} \big\{ \log \pi_{j,t} - \frac{1}{2\sigma_{j,t}^2} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_{j,t}\right)^2 - \frac{1}{2} \log(\sigma_{j,t}^2) \big\} \\
&\qquad \times K_h(u_i - u_t) - \frac{1}{2} \boldsymbol{\beta}_{j,t}^\top \boldsymbol{\Sigma}_j^{(m)} \boldsymbol{\beta}_{j,t} \Bigg\},
\end{aligned}$$

where $\boldsymbol{\Sigma}_j^{(m)} = \text{diag}\{\tau_{jl}^{(m)} : l = 1, \ldots, d\}$ with

$$\tau_{jl}^{(m)} = \frac{1}{n} \frac{p_n'\big(\|\boldsymbol{b}_{jl}^{(m)}\|/\sqrt{n}; \lambda_j\big)}{\|\boldsymbol{b}_{jl}^{(m)}\|/\sqrt{n}}.$$

The locally-constant parameter updates are, $t = 1, \ldots, n$,

$$\begin{aligned}
\pi_{j,t}^{(m+1)} = \pi_j^{(m+1)}(u_t) &= \sum_{i=1}^{n} w_{ij}^{(m)} K_h(u_i - u_t) \Big/ \sum\nolimits_{i=1}^{n} K_h(u_i - u_t) \\
\boldsymbol{\beta}_{j,t}^{(m+1)} = \boldsymbol{\beta}_j^{(m+1)}(u_t) &= \arg\min_{\boldsymbol{\beta}_{j,t} \in \mathbb{R}^d} \Bigg\{ \sum_{i=1}^{n} w_{ij}^{(m)} \{ \frac{1}{\sigma_{j,t}^{2(m)}} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_{j,t}\right)^2 \} K_h(u_i - u_t) \\
&\qquad + \boldsymbol{\beta}_{j,t}^\top \boldsymbol{\Sigma}_j^{(m)} \boldsymbol{\beta}_{j,t} \Bigg\} \\
\sigma_{j,t}^{2(m+1)} \equiv \sigma_j^{2(m+1)}(u_t) &= \frac{\sum_{i=1}^{n} w_{ij}^{(m)} K_h(u_i - u_t) \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_{j,t}^{(m+1)}\right)^2}{\sum_{i=1}^{n} w_{ij}^{(m)} K_h(u_i - u)}.
\end{aligned}$$

Web Appendix F.3 *The* EM *algorithm for the* MLLE *and* MPLLE *in* t FM-VCR

We adapted the EM algorithm outlined in Yao et al. (2014) for t-distribution mixtures to our penalization method for the FM-VCR models.

In a t FM-VCR, the conditional density (mass) of $Y|(\boldsymbol{X} = \boldsymbol{x}, U = u)$ is given by

$$f_C^*(y|\boldsymbol{\psi}(u), \boldsymbol{x}) = \sum_{j=1}^{C} \pi_j(u) f_T(y - \mathbf{x}^\top \boldsymbol{\beta}_j(u); \sigma_j(u), v_j),$$

where f_T is the density of a t-distribution with v_k degrees of freedom and

$$f_T(\epsilon_j(u); \sigma_j(u), v_j) = \frac{\Gamma\left(\frac{v_j+1}{2}\right)}{(\pi v_j)^{1/2} \Gamma\left(\frac{v_j}{2}\right) \left\{1 + \frac{\epsilon_j^2(u)}{\sigma_j^2(u) v_j}\right\}^{(v_j+1)/2} \sigma_j(u)}$$

with $\epsilon_j(u) = y - \mathbf{x}^\top \boldsymbol{\beta}_j(u)$.

Let $(u_i, \boldsymbol{x}_i, y_i), i = 1, \ldots, n$ be the observations based on a random sample from above model. The (conditional) log-likelihood function is given by

$$\mathcal{L} = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{C} \pi_j(u_i) f_T(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j(u_i); \sigma_j(u_i), v_j) \right\}.$$

At any $u \in \mathcal{U}$, the local-kernel log-likelihood in (7) of the paper is given by

$$\ell_n\big(\boldsymbol{\psi}(u); h\big) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{C} \pi_j f_T(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j; \sigma_j, v_j) \right\} K_h(u_i - u), \tag{A.75}$$

The locally constant vector of parameters is $\boldsymbol{\psi}(u) = (\boldsymbol{\pi}^\top, \boldsymbol{\phi}^\top, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_C^\top)$, where $\boldsymbol{\pi}^\top =$

$(\pi_1, \ldots, \pi_{C-1})$, $\boldsymbol{\phi}^\top = (\sigma_1, \ldots, \sigma_C)$ and $\boldsymbol{\beta}_j^\top = (\beta_{j1}, \ldots, \beta_{jd})$ for $j = 1, \ldots, C$. Note that the entries of the vector $\boldsymbol{\psi}(u)$ are local constant approximations of the functions $\beta_{jl}(u), l = 1, \ldots, d$, $\pi_j(u)$, and $\sigma_j(u)$, for all $j = 1, \ldots, C$. These entries clearly depend on u, and for simplicity, we suppress u in the notation but keep it for $\boldsymbol{\psi}(u)$.

Using (A.75), the corresponding (total) local-kernel log-likelihood in (8) of the paper is given by

$$\begin{aligned} L_n(\boldsymbol{\Psi}; h) &= \sum\nolimits_{t=1}^{n} \ell_n\big(\boldsymbol{\psi}(u_t); h\big) \\ &= \sum\nolimits_{t,i=1}^{n} \log\left\{\sum_{j=1}^{C} \pi_{j,t} f_T(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{j,t}; \sigma_{j,t}, v_j)\right\} K_h(u_i - u_t), \end{aligned} \tag{A.76}$$

where $\boldsymbol{\Psi}$ is the $(n \times \{C(d+2) - 1\})$-dimensional matrix of the locally constant vectors $\boldsymbol{\psi}(u_i), i = 1, \ldots, n$.

As discussed in Yao et al. (2014), a t-distribution can be represented as a scale mixture of normal distributions as follows. Let g be the latent variable such that

$$\epsilon | g \sim N(0, \sigma^2/g), \qquad g \sim \text{Gamma}(v/2, v/2),$$

where $\text{Gamma}(\alpha, \gamma)$ has the density

$$f_G(g; \alpha, \gamma) = \frac{1}{\Gamma(\alpha)} \gamma^\alpha g^{\alpha-1} e^{-\gamma g}, g > 0.$$

Then ϵ has a marginal t-distribution with degrees of freedom v and scale parameter σ. The scale mixture is used in the EM algorithm outlined below.

Web Appendix F.3.1 *The* EM *algorithm.* We view the observed data $\{(u_i, \boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ as incomplete, and introduce the unobserved Bernoulli random variables Z_{ij} to represent the membership of the i-th observation to the j-th component of the mixture model, $\forall j = 1, \ldots, C$, such that $\Pr(Z_{ij} = 1 | u_i, \boldsymbol{x}_i, y_i) = \pi_j(u_i)$. The complete data consist of $\{(u_i, \boldsymbol{x}_i, y_i, \boldsymbol{z}_i, g_i) : i = 1, \ldots, n\}$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iC})^\top$, and g_i's are the latent variables in the scale mixture.

The complete local-kernel log-Likelihood is given by

$$\begin{aligned}
&L_n^c(\boldsymbol{\Psi}; h) \\
&= \sum_{t=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{C} z_{ij} \log\left\{\pi_{j,t} N\left(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}_{j,t}, \frac{\sigma_{j,t}^2}{g_i}\right) f_G\left(g_i, \frac{v_j}{2}, \frac{v_j}{2}\right)\right\} K_h(u_i - u_t) \\
&= \sum_{t=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{C} z_{ij} K_h(u_i - u_t) \log(\pi_{j,t}) \\
&\quad + \sum_{t=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{C} z_{ij} K_h(u_i - u_t) \log\left\{N\left(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}_{j,t}; \frac{\sigma_{j,t}^2}{g_i}\right)\right\} \\
&\quad + \sum_{t=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{C} z_{ij} K_h(u_i - u_t) \log\left\{f_G\left(g_i, \frac{v_j}{2}, \frac{v_j}{2}\right)\right\}
\end{aligned}$$

The EM algorithm first computes the expected value of $L_n^c(\boldsymbol{\Psi}; h)$, with respect to the latent variables (z_{ij}, g_i)'s conditional on the data, and the current parameter values $\boldsymbol{\Psi}^{(m)}$.

E-step: At the $(m+1)$-th iteration, we compute

$$w_{ij}^{(m)} = E(Z_{ij}|\boldsymbol{\Psi}^{(m)}, \text{data}) = \frac{\pi_j^{(m)}(u_i) f_T\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j^{(m)}(u_i); \sigma_j^{(m)}(u_i), v_j\right)}{\sum_{k=1}^{C} \pi_k^{(m)}(u_i) f_T\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^{(m)}(u_i); \sigma_k^{(m)}(u_i), v_k\right)}$$

and

$$g_{ij}^{(m)} = E(g_i|\boldsymbol{\Psi}^{(m)}, \text{data}, Z_{ij} = 1) = \frac{v_j + 1}{v_j + \left\{\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(m)}(u_i)}{\sigma_j^{(m)}(u_i)}\right\}^2}.$$

M-step: The MLLE parameter updates are given as follows.

$$\pi_{j,t}^{(m+1)} = \pi_j^{(m+1)}(u_t) = \frac{\sum_{i=1}^{n} w_{ij}^{(m)} K_h(u_i - u_t)}{\sum_{i=1}^{n} K_h(u_i - u_t)},$$

and

$$\boldsymbol{\beta}_{j,t}^{(m+1)} = \boldsymbol{\beta}_j^{(m+1)}(u_t) = \left(\boldsymbol{X}^\top W_j^{(m+1)} G_j^{(m+1)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top W_j^{(m+1)} G_j^{(m+1)} \boldsymbol{y},$$

where $W_j^{(m+1)} = \text{diag}\left\{w_{ij}^{(m+1)}, i = 1, \ldots, n\right\}$, $G_j^{(m+1)} = \text{diag}\left\{g_{ij}^{(m+1)}, i = 1, \ldots, n\right\}$, and

$$\sigma_{j,t}^{2(m+1)} = \sigma_j^{2(m+1)}(u_t) = \frac{\sum_{i=1}^{n} w_{ij}^{(m+1)} g_{ij}^{(m+1)} K_h(u_i - u_t)\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_{j,t}^{(m+1)}\right)^2}{\sum_{i=1}^{n} w_{ij}^{(m+1)} K_h(u_i - u_t)}.$$

The updates of the regression parameters based on ridge penalty are given as

$$\boldsymbol{\beta}_{j,t}^{(m+1)} = \boldsymbol{\beta}_j^{(m+1)}(u_t) = \left(\boldsymbol{X}^\top W_j^{(m+1)} G_j^{(m+1)} \boldsymbol{X} + \gamma \mathbf{I}_{d\times d}\right)^{-1} \boldsymbol{X}^\top W_j^{(m+1)} G_j^{(m+1)} \boldsymbol{y},$$

where $\gamma > 0$ is the ridge tuning parameter.

Finally, the MPLLE updates of the regression parameters are given as

$$\boldsymbol{\beta}_{j,t}^{(m+1)} = \boldsymbol{\beta}_j^{(m+1)}(u_t) = \left(\boldsymbol{X}^\top W_j^{(m+1)} G_j^{(m+1)} \boldsymbol{X} + \gamma \mathbf{I}_{d\times d} + \boldsymbol{\Sigma}_j^{(m+1)}\right)^{-1} \boldsymbol{X}^\top W_j^{(m+1)} G_j^{(m+1)} \boldsymbol{y},$$

where $\boldsymbol{\Sigma}_j^{(m)} = \mathrm{diag}\{\tau_{jl}^{(m)} : l = 1, \ldots, d\}$ and

$$\tau_{jl}^{(m)} = \frac{1}{n} \frac{p_n'\big(\|\boldsymbol{b}_{jl}^{(m)}\|/\sqrt{n}; \lambda_j\big)}{\|\boldsymbol{b}_{jl}^{(m)}\|/\sqrt{n}}.$$

Web Appendix F.4 *Initial value for the* EM *algorithm*

In our simulation, we obtained $\boldsymbol{\Psi}^{(0)}$ by adding perturbation Gaussian noises to the true value $\boldsymbol{\Psi}^0$. In another approach, we first fit a finite mixture of polynomial regression coefficients of a prespecified degree (five) with fixed (non-varying) mixing probabilities and dispersion parameters (π_j, ϕ_j). The resulting estimates are then used as the initial value $\boldsymbol{\Psi}^{(0)}$ in the modified EM algorithm. In our simulations, both approaches led to similar results. In our analysis of the real data, we adopted the second approach by the first fitting finite a mixture of polynomial regressions based on several randomly generated initial values.

When the dimension d is comparable to the sample size n, or there is a near-singularity among the covariates, a ridge penalty may be applied in fitting the finite mixture of polynomial regression models.

Web Appendix G. Tuning parameter and mixture order selection

In practice, one needs to choose the band-width h, the tuning parameters λ_j's, and the mixture order C. It is computationally infeasible to simultaneously choose appropriate values for all these parameters. We now describe data-adaptive strategies for their selection.

Web Appendix G.1 *Band-width selection for smoothing*

We use the idea of a multi-fold cross-validation (Geisser, 1975) for band-width selection.

Let $\mathcal{D} = \{(u_i, \boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ represent the full data. We partition $\mathcal{D}$ randomly

into subsets $\mathcal{P}_l, l = 1, \ldots, J$, each of size approximately n/J. For each l, we refer to $\mathcal{P}_l$ and $\mathcal{T}_{-l} = \mathcal{D} \setminus \mathcal{P}_l$ as test and training data, respectively. For each l, and given h, let $\widetilde{\boldsymbol{\Psi}}_{-l}(h) \equiv \widetilde{\boldsymbol{\Psi}}_{-l}$ be the MLLE of $\boldsymbol{\Psi}$, by maximizing the local-kernel log-likelihood in (7) of the paper and using the training data $\mathcal{T}_{-l}$. Inspired by Huang et al. (2018), we compute the predictive log-likelihood

$$Pl(h) = \sum_{l=1}^{J} \sum_{i \in \mathcal{P}_l} \log \left\{ \sum_{j=1}^{C} \tilde{\pi}_{j,-l}(u_i) f\left(y_i; \tilde{\theta}_{j,-l}(\boldsymbol{x}_i, u_i), \tilde{\phi}_{j,-l}(u_i)\right) \right\}, \tag{A.77}$$

where $\tilde{\theta}_{j,-l}(\boldsymbol{x}_i, u_i) = g(\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_{j,-l}(u_i))$. The predictive log-likelihood is computed over a grid of h values, and we choose a value of h that maximizes $Pl(h)$. To reduce the effect of random partitioning, we repeat the partitioning, say B times, and maximize an average predictive likelihood with respect to h. In our simulation, we used $J = 5$ and $B = 10$.

In what follows, we fix h at the value obtained based on the criterion in (A.77).

Web Appendix G.2 *Tuning parameter for variable selection*

We use a BIC-type criterion (Wang et al., 2007) for choosing a presumed common tuning parameter $\lambda_1 = \cdots = \lambda_C = \lambda$ for the variable selection via penalty p_n in (10) of the paper. Using this method, one may choose different λ_j's by searching over a C-dimensional grid which is computationally more expensive.

For a value λ, let $\widehat{\boldsymbol{\Psi}}_n(\lambda)$ be the MPLLE as in (11) of the paper. Note that to perform variable selection, each non-parametric function $\boldsymbol{\beta}_{jl}(u)$ is estimated at the points $u_1, \ldots, u_n$ by the vector $\widehat{\boldsymbol{b}}_{jl}(\lambda) = \left(\widehat{\beta}_{jl}(u_1), \ldots, \widehat{\beta}_{jl}(u_n)\right)^\top$. Thus, the total number of estimated non-zero regression functions is given by $\sum_{j=1}^{C} \sum_{l=1}^{d} 1\{\|\widehat{\boldsymbol{b}}_{jl}(\lambda)\|_2 > 0\}$. We compute

$$\text{BIC}_1(\lambda) = -2\mathcal{L}(\widehat{\boldsymbol{\Psi}}_n(\lambda)) + \log n \times \left\{ \sum\nolimits_{j=1}^{C} \sum\nolimits_{l=1}^{d} 1\{\|\widehat{\boldsymbol{b}}_{jl}(\lambda)\|_2 > 0\} \right\} \times \text{DF}_h, \tag{A.78}$$

where $\mathcal{L}$ is the log-likelihood in (6) of the paper evaluated at $\widehat{\boldsymbol{\Psi}}_n(\lambda)$, and as in Huang et al. (2013),

$$\text{DF}_h = \tau_K h^{-1} \, |\mathcal{U}| \left\{ K(0) - \frac{1}{2} \int_{\mathcal{U}} K^2(u) du \right\},$$

where $|\mathcal{U}|$ is equal to the length of the support of the index variable U, and

$$\tau_K = \left\{ K(0) - \frac{1}{2}\int_{\mathcal{U}} K^2(u)du \right\} \left\{ \int_{\mathcal{U}} \left\{ K(u) - \frac{1}{2} K * K(u) \right\}^2 du \right\}^{-1}$$

and $K * K(t) = \int_0^t K(u)K(t-u)du$. In Web Appendix G.4 that follows, we compute the degrees of freedom DF_h for the Epanechnikov Kernel. We compute the BIC over a grid of λ-values, say $\lambda_1, \ldots, \lambda_M$, and then choose a value of λ that minimizes (A.78).

In our simulation and data analysis, guided by the theory in Section 4 of the paper, for a given n, the smoothing parameter h is chosen by maximizing the predictive log-likelihood (A.77) over the range $[0.1, 2n^{-0.2}]$; the tuning parameter λ is chosen by minimizing the BIC in (A.78) over the range $[10^{-6}, 3n^{-0.45}]$ for the AdpLASSO and over the range $[10^{-6}, 3n^{-0.35}]$ for the LASSO, SCAD, MCP. The constants involved in the ranges are chosen by trial and error.

Web Appendix G.3 *Mixture order selection*

We also use a BIC for selection of the mixture order C, when it is unknown. Here, we need to take into account the total number of non-parametric mixing probabilities and dispersion parameters that are estimated. We compute the BIC

$$\text{BIC}_2(C) = -2\mathcal{L}(\widehat{\boldsymbol{\Psi}}_n(C)) + \log n \times \left\{ 2C - 1 + \sum\nolimits_{j=1}^{C} \sum\nolimits_{l=1}^{d} 1\{\|\widehat{\boldsymbol{b}}_{jl}(C)\|_2 > 0\} \right\} \times \text{DF}_h, \quad \text{(A.79)}$$

where $\widehat{\boldsymbol{\Psi}}_n(C)$ is the MPLLE obtained for the fitted mixture models of different orders $C = 1, \ldots, \mathcal{K}$, for some pre-specified upper bound $\mathcal{K}$. In Web Appendix G.4 that follows, we compute the degrees of freedom DF_h for the Epanechnikov Kernel. The mixture order is then estimated by $\widehat{C}_n$ that minimizes $\text{BIC}_2(C)$ over $1 \leqslant C \leqslant \mathcal{K}$. Theoretically, under the conditions of Theorem 2 in the paper, and similar to the results of Leroux (1992), the estimator $\widehat{C}_n$ does not underestimate the true mixture order with probability tending to one as $n \to \infty$. Our simulation results show that $\widehat{C}_n$ does not underestimate the mixture order, and also the percentage of overestimation decreases as n increases.

Web Appendix G.4 *Computing the degree of freedom for the BIC*

For the Epanechnikov Kernel in Web Appendix C, we have

$$\text{DF}_h = \tau_K h^{-1} \ |\mathcal{U}| \left\{ K(0) - \frac{1}{2} \int_{\mathcal{U}} K^2(u) du \right\},$$

and for the BIC given in Web Appendix G of the paper, where $\mathcal{U}$ is the support of the index variable U, $|\mathcal{U}|$ is equal to the length of the support, and

$$\tau_K = \frac{K(0) - \frac{1}{2}\int K^2(u)du}{\int_{\mathcal{U}} \left\{ K(u) - \frac{1}{2} K * K(u) \right\}^2 du},$$

with the convolution

$$K * K(t) = \int_0^t K(u)K(t-u)du.$$

For the Epanechnikov Kernel, we have

$$K * K(t) = \frac{3^2}{4^2} \int_0^t \frac{3}{4}(1-u^2)\frac{3}{4}(1-(t-u)^2)du = \frac{3^2}{4^2}\left(t - \frac{2}{3}t^3 + \frac{1}{30}t^5 \right),$$

and

$$\begin{aligned} \int_0^1 \left\{ K(u) - \frac{1}{2} K * K(u) \right\}^2 dt &= \int_0^1 \left\{ \frac{3}{4}(1-u^2) - \frac{1}{2}\frac{3^2}{4^2}\left(u - \frac{2}{3}u^3 + \frac{1}{30}u^5 \right) \right\}^2 du \\ &= \frac{3^2}{4^2} \int_0^1 \left(1 - \frac{3}{8}u - u^2 + \frac{1}{4}u^3 - \frac{1}{80}u^5 \right)^2 du \\ &= 0.2285. \end{aligned}$$

Thus,

$$K(0) = \frac{3}{4},$$

$$\int_0^1 K^2(u)du = \int_0^1 \frac{3^2}{4^2}\left(1-u^2\right)^2 du = \frac{3}{10},$$

and

$$\text{DF}_h = \frac{3/4 - 3/20}{0.2285}\frac{|\mathcal{U}|}{h}(3/4 - 3/20) = \frac{1.5755}{h}.$$

Web Appendix H. Additional simulation studies

Web Appendix H.1 *Model Misspecification*

Since our proposed estimation and variable selection method is likelihood-based, one expects that a misspecification of the parametric form of the mixture components would, in general, affect the performance of the method, although the degree of performance degradation depends on how different the misspecified and the true models are. In terms of the number of latent classes, if C is smaller than its true value, then the fitted model converges to an under-specified model that minimizes the Kullback-Leibler distance to the true FM-VCR model (Leroux, 1992). Thus, we cannot make statements about the consistent estimation of the parameters of the true model based on an under-specified model. On the other hand, if C is larger than its true value, then the overall behavior of the true FM-VCR model will be captured by the over-fitted mixture model. In particular, Ho and Nguyen (2016) and Ho et al. (2022) showed that the density function of an over-fitted standard finite mixture model and the finite mixture of regressions (both special cases of FM-VCRs) consistently estimates the true finite mixture density. We believe such property also holds for over-fitted FM-VCR, although further investigation is needed to study this theoretically.

Below we investigate, through simulation, two scenarios related to model misspecification: misspecifying the parametric form and misspecifying the number of components, with the results presented below as (i) and (ii), respectively. Our simulations are based on the dimension $d = 10$ and sample sizes $n = 200$ and 400, representative settings of those in the paper. The results are based on $R = 100$ replicates.

(i) Misspecification of the parametric form of the mixture component density

We generated random samples $(u_i, \boldsymbol{x}_i, y_i), i = 1, \ldots, n$, from a two-components FM-VCR model with each component density set to be a t-distribution with 10 degrees of freedom,

which has moderately heavier tails than a Gaussian distribution. The rest of the parameter settings are the same as those in Table 1 of the Simulation Section in the paper.

Using our proposed regularization method outlined in Section 3 of the paper, we fitted two models to the generated data: the correct t FM-VCR from which the data were generated and the misspecified Gaussian FM-VCR model. As in the paper, the simulation results are summarized in terms of sensitivity, specificity, and estimation error (L_2). The results are given in Web Table 7 below. Similar to the Gaussian FM-VCR model, the details of the numerical implementation of the penalization method for a t FM-VCR are now given in Web Appendix F.3 above.

From Web Table 7, we observe that, overall, the results based on the correct t FM-VCR (upper portion of the table) are reasonable in terms of all the performance measures under consideration. The results are roughly similar to those of Table 2 of the paper based on the correct Gaussian FM-VCR model. As expected, the performance of the method in Component 2 (the larger component) of the mixture is better. On the other hand, the results (lower portion of the table) based on the misspecified Gaussian model show that the performance of the method in terms of sensitivity and specificity generally degrades but not dramatically in terms of sensitivity, while specificity has not been affected much. The quality of the parameter estimates in terms of estimation error has been affected particularly in Component 1, the smaller component. As the sample size n increases, the loss in performance is less which is expected. Overall, the misspecification for the above parameter setting has not affected the performance much.

(ii) **Misspecification of the number of mixture components (C)**

We consider the Gaussian FM-VCR with the true number of components (order) $C = 2$ and the same parameter setting as in Table 1 of the paper. We generated random samples $(u_i, \boldsymbol{x}_i, y_i), i = 1, \ldots, n$, from this model and fitted Gaussian FM-VCR models with orders

$C = 1, 2, 3$ and 4, to the data. For the overfitted FM-VCR models with orders $C = 3, 4$, we first perform a component-matching in which we find the closest components of the over-fitted model to those of the true model with $C = 2$ and then compute the performance measures.

The simulation results are summarized in Web Table 8. The results for the fitted misspecified models with orders $C = 1, 3$ and 4 are compared with those from the correct order $C = 2$. The one-component underfitted model with $C = 1$ resembles the behavior of the larger component of the correct model, i.e. Component 2, but with lower sensitivity and specificity and larger estimation errors for the corresponding measures for Component 2 when $C = 2$ was fitted. On the other hand, the behavior of the models with $C = 3, 4$ are similar to those with the correct model but with lower estimation errors. These simulation results are in line with the theoretical properties of over-fitted finite mixture models (Ho and Nguyen, 2016; Ho et al., 2022), although further work is necessary to theoretically investigate such properties for the over-fitted FM-VCR models.

Web Appendix H.2 *Comparison of* FMR *and* FM-VCR *models*

We generated data from a Gaussian finite mixture of regression (FMR) model (without allowing for varying coefficients) with the number of components $C = 2$, dimension $d = 10$, and samples sizes $n = 200$ or 400. The parameter setting for the FMR model is as follows:

$$\boldsymbol{\beta}_1^\top = [-0.5, 0.25, 0.25, 0, -0.25, 0, 0, 0, 0, 0, 0],$$

$$\boldsymbol{\beta}_2^\top = [-0.25, 0.25, 0, 0.25, 0, 0, 0, 0, 0, 0, 0],$$

$$\boldsymbol{\sigma}^\top = [0.39, 0.45], \quad \boldsymbol{\pi}^\top = [0.55, 0.45].$$

We then fitted both Gaussian FMR and FM-VCR models to each simulated sample. The results are given in Web Table 9. We observe that, for sensitivity, FM-VCR (the wrong model) is generally worse (on average based on 100 replicates) than FMR (the true underlying

model), with larger standard deviations (SD). On the other hand, for specificity, FM-VCR is consistently better and with much smaller SD. This implies that FM-VCR selected fewer variables, hence fewer true positives and fewer false positives, thus larger specificity and smaller sensitivity. This result makes sense because with a more complex model and more parameters but a fixed sample size, variables are harder to be selected as significant, especially for the parameter setting considered above where the effects of covariates (β_{jl}'s) are weak. Thus, in practice, one may fit both models to a dataset and assess the results as we have done for our real data.

Web Appendix H.3 *Alternative tuning parameter selection*

In general, it is challenging to provide a universally dominant method for a data-dependent choice of the three tuning parameters. In addition to the implementation of the sequential selection of h and then λ for a prespecified mixture order C (Web Appendix G), here we investigate an alternative tuning parameter selection approach, where h and λ are selected simultaneously, to see whether performance can be improved although at the expense of greater computational intensity. Specifically, we investigate the scenario in which $C = 2$, dimension $d = 10$, and sample sizes $n = 200, 400$, which was considered in the paper. The variable selection and estimation error results are given in Web Table 10, and the time (in seconds) are given in Web Table 11.

As we can see from the computational time results, on average, the MPLLE based on selecting h and then λ is at least 18 times faster than the MPLLE when (h, λ) are selected simultaneously, which is expected. On the other hand, comparing the results of Web Table 10 with those in Table 2 of the paper, despite the increase in computational time, the gain in performance of the MPLLE based on the simultaneous selection of (h, λ) both in terms of variable selection and estimation error is negligible.

Web Appendix H.4 *High-dimensions*

In this section, we assess the performance of the penalization method when the dimension d exceeds the sample size n, although this case is not covered in our theoretical results where d is fixed as n grows.

We have considered the two-component ($C = 2$) Gaussian FM-VCR model with the parameter setting given in Table 1 of the paper, dimension $d = 500$, and sample sizes $n = 200, 400$. In this model, covariates (x_1, x_2, x_4) in Component 1 and covariates (x_1, x_3) in Component 2 have non-zero $\beta_{jk}(u)$, and the rest have zero coefficients. There are a total of $p = Cd + 2C - 1 = 1003$ non-parametric functions to be estimated by the proposed penalization method. The results are summarized in Web Table 12.

The results show that while the specificity remains high and similar to dimensions $d = 5, 10, 20$, and 50 already considered in Example 1 of Section 6 in the main paper (Table 2, and see also Web Table 2), there is some reduction in sensitivity, although the difference is quite small when compared to $d = 50$, indicating that reasonable results are likely to be obtained by the method even when the number of variables exceed the sample size.

Nevertheless, rigorous study of high-dimensional setting requires new theoretical and numerical tools, which are not covered in the current paper.

In the second part of our simulation to investigate high-dimensional settings, we examined two popular screening techniques that one could use when fitting an FM-VCR to a dataset.

(1) Screening using the Pearson's Correlation.

We computed the sample correlation between the response and $d = 500$ covariates for each simulated sample, and we kept $d^* = 50$ top covariates with the highest correlation values. This procedure was repeated for 100 replicated samples. Web Table 13 shows the top 20 covariates that have the highest frequencies of being selected out of the 100 replications. Note that in our simulation, the covariates (x_1, x_2, x_3, x_4) have non-zero $\beta_{jl}(u)$'s in the true

data generating FM-VCR, and they all survived the screening process in all 100 replicates. Some other covariates (those with zero $\beta_{jl}(u)$'s, essentially noise) also survived with high frequencies, necessitating the use of our FM-VCR method for further in-depth variable selection and estimation.

(2) Screening using marginal likelihood of single-covariate FM-VCR model with $C = 2$.

In this approach, via MLE, we fitted $d = 500$ single-covariate FM-VCR models each of which includes intercepts and only one covariate in both components of the mixture model. We then rank the top $d^* = 50$ covariates based on the likelihood value of their corresponding fitted model. This procedure was repeated for 100 replicated samples. Web Table 14 below shows the top 20 covariates with the highest frequencies of being among the top 50's out of the 100 replications. Again, we can see that the covariates (x_1, x_2, x_3, x_4) that have non-zero $\beta_{jl}(u)$'s in the true data generating FM-VCR survived the screening process in all 100 replicates, among other (noise) covariates that were in fact selected less.

Web Appendix I. More on the data analysis from the OCN study

To substantiate the need for an FM-VCR model, we carried out an analysis by fitting an FMR model to the OCN data to show the lack of fit of this model compared to the FM-VCR model. Specifically, using all four penalties studied in the paper, we analyzed the OCN data by fitting FMR models with mixture orders $C = 1, 2, 3, 4$, and 5. An FMR model with two components ($C = 2$), and based on the SCAD penalty, was selected by the BIC. Similar to the fitted sparse FM-VCR model, SNP rs7456421 was also selected in the fitted sparse FMR model, among several other covariates. The mixing proportion of Component 1 in the FMR model is 53% which is approximately equal to the average of the varying mixing proportion of Component 1 in the FM-VCR model. The variance of Component 1 in the FMR model is about 62, which is close to the average of varying variance of Component 1 in the FM-VCR

model. However, the second component of the FMR model has a much larger variance, about 149, compared to the variance of Component 2 in the FM-VCR model (which ranges from 13 to 19 as a function of age). This much larger variance of Component 2 compared to the variance of Component 1 in the FMR can be an indication of a misspecified model. To demonstrate the lack of fit using the FMR model and the subsequent improvement by the FM-VCR model, we analyzed their residuals by plotting the quantiles against their normal counterparts. As can be seen from the QQ-plots in Web Figure 2 below, the use of FM-VCR indeed improves the fit to the data compared to FMR, especially for Component 2.

We have also supplied two plots to provide further evidence of the varying coefficient effect of SNP rs7456421 over the age on osteocalcin. In Web Figure 3(a), we have added the point-wise error bars to the estimated varying coefficient plot of the SNP, where the error bars were calculated using the EM algorithm approximation method of **?**. This substantiates our conclusion that SNP rs7456421 has an age-dependent effect on osteocalcin, and that this SNP has a significantly negative effect at a younger age, and this effect diminishes as one ages. Web Figure 3(b) is the derivative curve (approximated using Matlab) of the estimated coefficient curve over age. It is evident from the plot that the underlying true derivatives are unlikely to be constant at 0 over the age range.

Web Appendix J. Additional tables and figures

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

[Table 10 about here.]

[Table 11 about here.]

[Table 12 about here.]

[Table 13 about here.]

[Table 14 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

References

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* **41,** 802 – 837.

Fang, F., Jiwei, Z., S. Ejaz, A., and Qu, A. (2021). A weak-signal-assisted procedure for variable selection and statistical inference with an informative subsample. *Biometrics* **77,** 996–1010.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70,** 320–328.

Ho, N. and Nguyen, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics* **10,** 271–307.

Ho, N., Yang, C.-Y., and Jordan, M. I. (2022). Convergence rates for gaussian mixtures of experts. *Journal of Machine Learning Research* .

Huang, M., Li, R., and Wang, S. (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association* **108,** 929–941.

Huang, M., Yao, W., Wang, S., and Chen, Y. (2018). Statistical inference and applications of mixture of varying coefficient models. *Scandinavian Journal of Statistics* **45,** 618–643.

Janson, S. (1987). Maximal spacing in several dimensions. *Annals of Probability* **15,** 274–80.

Leroux, B. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics* **20,** pp. 1350–1360.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94,** 553–568.

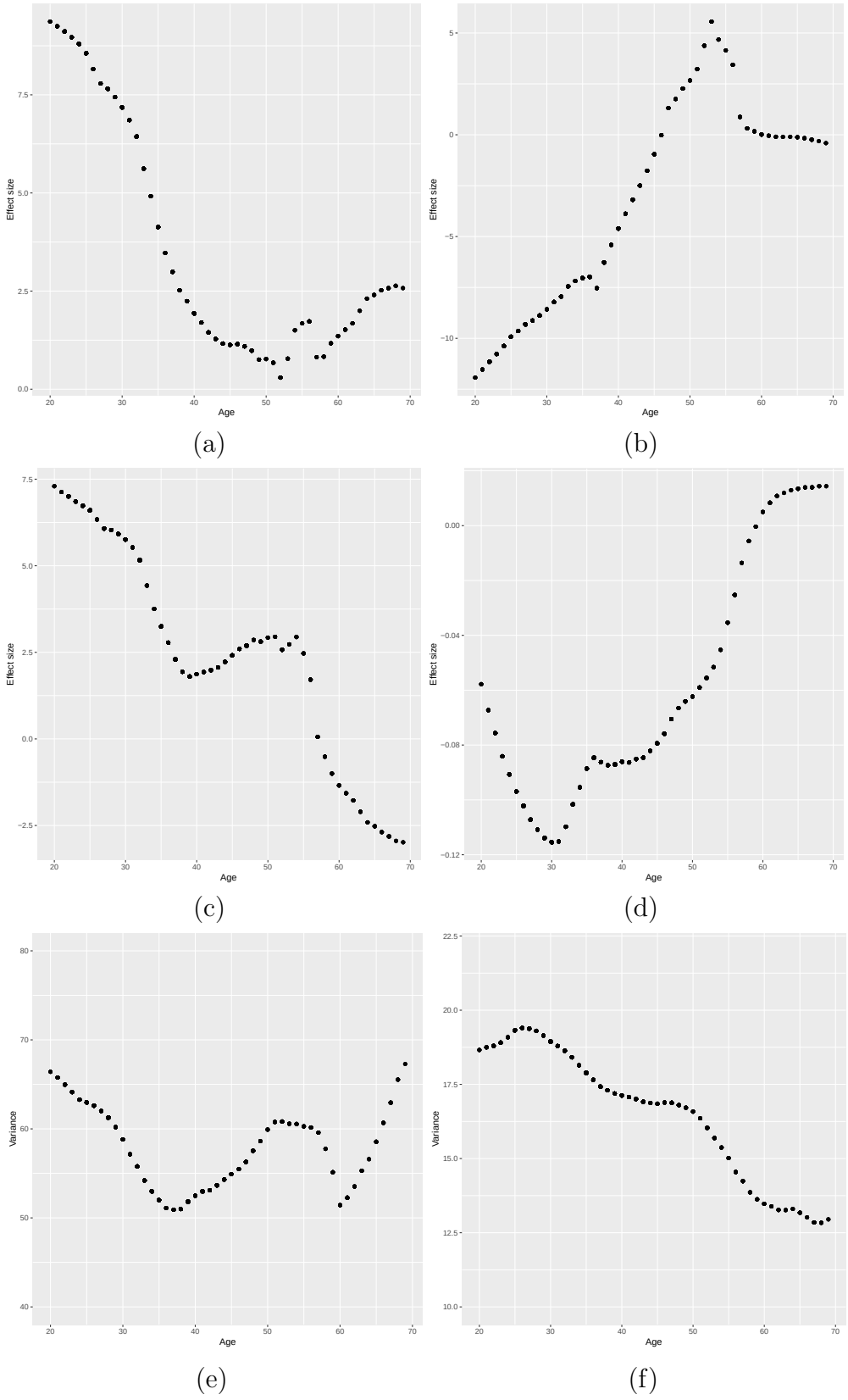
Wei, F., Huang, J., and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica* **21,** 1515–1540.

Yao, W., Wei, Y., and Yu, C. (2014). Robust mixture regression using the t-distribution. *Computational Statistics & Data Analysis* **71,** 116–127.

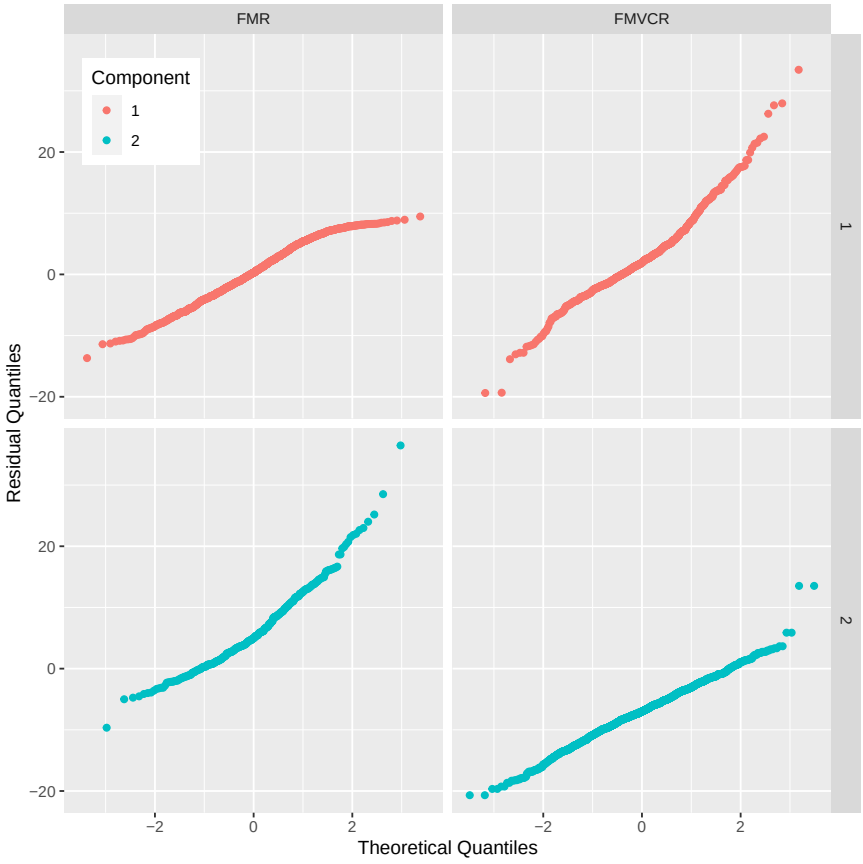
Zhang, D., Khalili, A., and Asgharian, M. (2022). Post-model-selection inference in linear regression models: An integrated review. *Statistics Surveys* **16,** 86–136.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

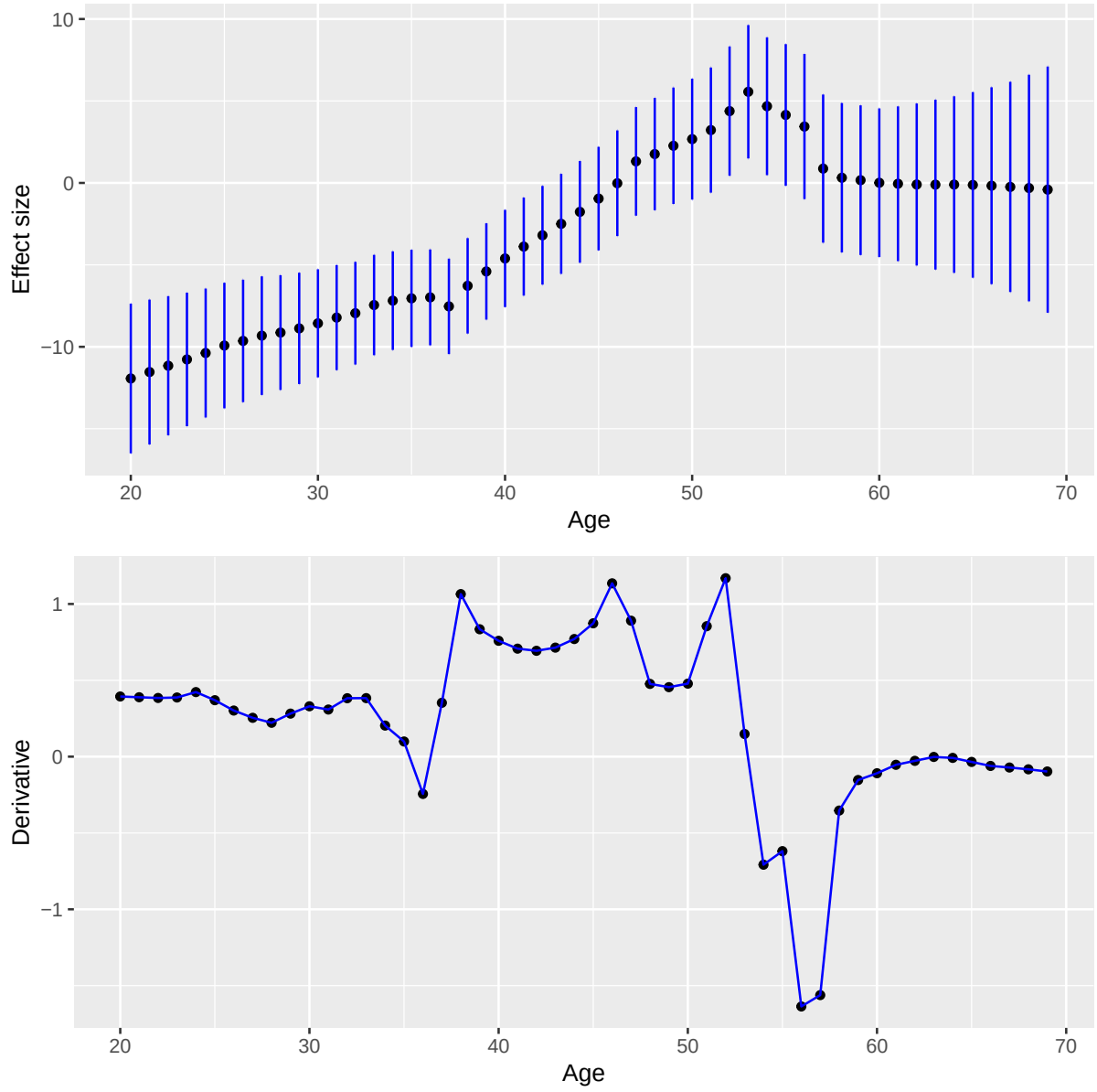
Received January 2022. *Revised ?* 2022. *Accepted ?* 2022.



Web Figure 1: Osteocalcin data analysis; (a) The estimated effect of rs109522346 over time in hOCN. (b) The estimated effect of rs7456421 over time in hOCN. (c) The estimated effect of rs4074826 over time in hOCN. (d) The estimated effect of folic acid over time in hOCN. (e) Estimated variance over time in hOCN. (f) Estimated variance over time in lOCN.



Web Figure 2: QQ-plots of the OCN data analyses; results of FM-VCR versus FMR model.



Web Figure 3: The non-constant effect of SNP rs7456421 on osterocalcin for Component 1 of the two-component FM-VCR model. (a) Error bar of the estimated effect size at each age; (b) Derivatives of the non-constant effect curve.

Web Table 1: Parameters settings for the Gaussian FM-VCR models with $C = 2, 3$.

Component(j)	1	2	3
Parameters	$d(= 5, 10, 20, 50)$[1]		
$\beta_{j0}(u)$	-2	-1	1
$\beta_{j1}(u)$	$1 + 0.5\cos(\pi u)$	$1.5\sin(\pi u)$	0
$\beta_{j2}(u)$	$1 + 0.5\cos(2\pi u)$	0	$-1.5(\cos(2\pi u))^2$
$\beta_{j3}(u)$	0	$1.5 - 0.5\sin(\pi u/2)$	0
$\beta_{j4}(u)$	$\sin(6\pi u)$	0	$1 - 0.5\cos(3u\pi/2)$
$\beta_{j5}(u)$	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\beta_{j,50}(u)$	0	0	0
$\sigma_j(u)$	$0.3e^{(0.5u)}$	$0.5e^{(-0.2u)}$	$0.8e^{(0.5u)}$
$\pi_j(u)$	$e^{0.5u}/(1 + e^{0.5u})$	$(1 + e^{0.5u})^{-1}$	—
$\pi_j(u)$	$e^{0.5u}/(1 + e^{0.5u} + e^{0.2u})$	$e^{0.2u}/(1 + e^{0.5u} + e^{0.2u})$	$(1 + e^{0.5u} + e^{0.2u})^{-1}$

[1] We keep the non-zero coefficient functions $\beta_{jl}(\cdot)$ the same for different values of the dimension d.

Web Table 2: Results of Example 1: average (SD) sensitivity, specificity, and estimation errors: $C = 2$.

Criteria	Sensitivity		Specificity		$L_2(\widehat{\boldsymbol{\beta}}_j)$		$L_2(\widehat{\boldsymbol{\sigma}}_j^2)$		$L_2(\widehat{\boldsymbol{\pi}}_j)$	
Component	1	2	1	2	1	2	1	2	1	2
d = 5	**n = 200**									
Oracle	—	—	—	—	0.6504	0.1619	0.3828	0.1037	0.3209	0.1037
MLLE	—	—	—	—	1.0348	0.2431	0.7476	0.1273	0.3702	0.1273
AdpLASSO	$0.6867_{(0.2030)}$	$0.9483_{(0.1914)}$	$0.9667_{(0.1249)}$	$0.9622_{(0.1221)}$	0.9337	0.2077	0.6119	0.1182	0.5001	0.1182
LASSO	$0.7011_{(0.2230)}$	$0.8467_{(0.3467)}$	$0.8133_{(0.2653)}$	$0.8244_{(0.2136)}$	0.9829	0.2560	0.8119	0.1670	0.5237	0.1670
MCP	$0.6822_{(0.1269)}$	$0.9517_{(0.1535)}$	$0.9783_{(0.1099)}$	$0.9722_{(0.1203)}$	0.8885	0.1632	0.5608	0.0938	0.5046	0.0938
SCAD	$0.6844_{(0.1266)}$	$0.9567_{(0.1467)}$	$0.9800_{(0.1063)}$	$0.9744_{(0.1176)}$	0.8819	0.1655	0.5536	0.0926	0.5074	0.0926
	n = 400									
Oracle	—	—	—	—	0.3987	0.0831	0.2144	0.0800	0.2730	0.0800
MLLE	—	—	—	—	0.5467	0.0969	0.3895	0.0749	0.2782	0.0749
AdpLASSO	$0.9778_{(0.0833)}$	$0.9967_{(0.0577)}$	$0.9983_{(0.0289)}$	$0.9967_{(0.0430)}$	0.4991	0.0819	0.2253	0.0743	0.3114	0.0743
LASSO	$0.9889_{(0.0599)}$	$0.9950_{(0.0645)}$	$0.9383_{(0.1793)}$	$0.9100_{(0.1799)}$	0.5819	0.1165	0.3759	0.0859	0.3377	0.0859
MCP	$0.9833_{(0.0728)}$	$1.0000_{(0.0000)}$	$1.0000_{(0.0000)}$	$0.9989_{(0.0192)}$	0.4582	0.0777	0.2079	0.0735	0.3128	0.0735
SCAD	$0.9844_{(0.0704)}$	$1.0000_{(0.0000)}$	$1.0000_{(0.0000)}$	$0.9989_{(0.0192)}$	0.4581	0.0774	0.2074	0.0734	0.3129	0.0734
d = 10	**n = 200**									
Oracle	—	—	—	—	0.6983	0.1616	0.3968	0.0887	0.3143	0.0887
MLLE	—	—	—	—	1.2296	0.2886	1.0160	0.1378	0.3355	0.1378
AdpLASSO	$0.5889_{(0.2390)}$	$0.8933_{(0.2592)}$	$0.9705_{(0.0667)}$	$0.9663_{(0.0610)}$	1.0894	0.2876	0.8483	0.1608	0.5007	0.1608
LASSO	$0.4700_{(0.3206)}$	$0.7750_{(0.4009)}$	$0.9276_{(0.1002)}$	$0.9167_{(0.0869)}$	1.3206	0.3829	1.1522	0.2940	0.5054	0.2940
MCP	$0.5811_{(0.1940)}$	$0.8950_{(0.2450)}$	$0.9838_{(0.0454)}$	$0.9812_{(0.0481)}$	1.0420	0.2378	0.7633	0.1056	0.5034	0.1056
SCAD	$0.5944_{(0.1878)}$	$0.8917_{(0.2435)}$	$0.9833_{(0.0488)}$	$0.9800_{(0.0503)}$	1.0294	0.2372	0.7612	0.1066	0.5010	0.1066
	n = 400									
Oracle	—	—	—	—	0.4393	0.0888	0.2270	0.0816	0.2728	0.0816
MLLE	—	—	—	—	0.8776	0.1489	0.6068	0.0867	0.3374	0.0867
AdpLASSO	$0.6900_{(0.1272)}$	$0.9933_{(0.0574)}$	$0.9971_{(0.0232)}$	$0.9954_{(0.0294)}$	0.8079	0.1654	0.3905	0.0692	0.4110	0.0692
LASSO	$0.7656_{(0.2302)}$	$0.9700_{(0.1659)}$	$0.9705_{(0.0657)}$	$0.9617_{(0.0638)}$	0.8617	0.2322	0.5176	0.1076	0.4249	0.1076
MCP	$0.6922_{(0.1042)}$	$0.9917_{(0.0641)}$	$0.9971_{(0.0200)}$	$0.9983_{(0.0144)}$	0.7892	0.1455	0.3745	0.0616	0.4119	0.0616
SCAD	$0.6878_{(0.0979)}$	$0.9917_{(0.0641)}$	$0.9962_{(0.0231)}$	$0.9979_{(0.0160)}$	0.7919	0.1463	0.3768	0.0620	0.4122	0.0620
d = 20	**n = 200**									
Oracle	—	—	—	—	0.6147	0.0952	0.2975	0.0790	0.3133	0.0790
MLLE	—	—	—	—	1.3671	0.2255	1.3035	0.0821	0.3835	0.0821
AdpLASSO	$0.6456_{(0.1153)}$	$0.9433_{(0.2025)}$	$0.9814_{(0.0434)}$	$0.9839_{(0.0416)}$	0.9140	0.2692	0.7131	0.0563	0.4878	0.0563
LASSO	$0.6567_{(0.1200)}$	$0.9533_{(0.1870)}$	$0.9629_{(0.0587)}$	$0.9524_{(0.0532)}$	0.9240	0.3363	0.8963	0.0563	0.4940	0.0563
MCP	$0.6433_{(0.1212)}$	$0.9533_{(0.1986)}$	$0.9941_{(0.0278)}$	$0.9955_{(0.0218)}$	0.9149	0.2200	0.6193	0.0520	0.4673	0.0520
SCAD	$0.6411_{(0.1268)}$	$0.9467_{(0.2061)}$	$0.9943_{(0.0277)}$	$0.9957_{(0.0211)}$	0.9183	0.2270	0.6333	0.0524	0.4721	0.0524
	n = 400									
Oracle	—	—	—	—	0.3533	0.0721	0.2108	0.0670	0.2515	0.0670
MLLE	—	—	—	—	0.9120	0.1019	0.6161	0.0377	0.4234	0.0377
AdpLASSO	$0.6678_{(0.0192)}$	$0.9978_{(0.0385)}$	$0.9986_{(0.0148)}$	$0.9996_{(0.0048)}$	0.7984	0.1697	0.4171	0.0356	0.4844	0.0356
LASSO	$0.7156_{(0.1181)}$	$1.0000_{(0.0000)}$	$0.9780_{(0.0351)}$	$0.9694_{(0.0412)}$	0.8226	0.2373	0.5234	0.0365	0.4995	0.0365
MCP	$0.6667_{(0.0472)}$	$0.9967_{(0.0577)}$	$0.9994_{(0.0102)}$	$0.9996_{(0.0068)}$	0.7962	0.1589	0.4331	0.0360	0.4808	0.0360
SCAD	$0.6678_{(0.0510)}$	$0.9967_{(0.0577)}$	$0.9992_{(0.0107)}$	$0.9996_{(0.0068)}$	0.7961	0.1588	0.4331	0.0360	0.4806	0.0360
d = 50	**n = 200**									
Oracle	—	—	—	—	0.7270	0.1672	0.3934	0.0968	0.2946	0.0968
MLLE	—	—	—	—	2.0960	0.2002	2.0600	0.0808	0.1598	0.0808
AdpLASSO	$0.5478_{(0.2188)}$	$0.7633_{(0.3307)}$	$0.9882_{(0.0150)}$	$0.9895_{(0.0157)}$	1.2587	0.2945	1.0778	0.1649	0.4606	0.1649
LASSO	$0.5067_{(0.2852)}$	$0.6983_{(0.4148)}$	$0.9873_{(0.0147)}$	$0.9872_{(0.0154)}$	1.3616	0.3737	1.2348	0.2622	0.4830	0.2622
MCP	$0.5567_{(0.1617)}$	$0.8217_{(0.2468)}$	$0.9923_{(0.0105)}$	$0.9925_{(0.0107)}$	1.2601	0.2437	1.0173	0.1137	0.3995	0.1137
SCAD	$0.5578_{(0.1589)}$	$0.8283_{(0.2413)}$	$0.9930_{(0.0102)}$	$0.9926_{(0.0109)}$	1.2481	0.2397	1.0028	0.1117	0.4045	0.1117
	n = 400									
Oracle	—	—	—	—	0.4701	0.1410	0.3067	0.0957	0.2988	0.0957
MLLE	—	—	—	—	1.8860	0.2506	1.6380	0.1881	0.2752	0.1881
AdpLASSO	$0.6122_{(0.1669)}$	$0.8783_{(0.2299)}$	$0.9906_{(0.0152)}$	$0.9926_{(0.0125)}$	1.1289	0.2968	0.8221	0.0953	0.5259	0.0953
LASSO	$0.6067_{(0.2234)}$	$0.8750_{(0.2590)}$	$0.9887_{(0.0156)}$	$0.9891_{(0.0164)}$	1.1621	0.3373	0.8745	0.1454	0.5231	0.1454
MCP	$0.6067_{(0.1593)}$	$0.8900_{(0.2075)}$	$0.9950_{(0.0092)}$	$0.9953_{(0.0089)}$	1.0893	0.2396	0.7811	0.0676	0.5090	0.0676
SCAD	$0.5978_{(0.1556)}$	$0.8833_{(0.2118)}$	$0.9948_{(0.0094)}$	$0.9951_{(0.0090)}$	1.1050	0.2428	0.7983	0.0688	0.5053	0.0688

Web Table 3: Results of Example 1: average (SD) sensitivity, specificity, and estimation errors: $C = 3$.

Criteria	Sensitivity			Specificity			$L_2(\widehat{\boldsymbol{\beta}}_j)$			$L_2(\widehat{\boldsymbol{\sigma}}_j^2)$			$L_2(\widehat{\boldsymbol{\pi}}_j)$		
Component	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
d = 5	**n = 200**														
Oracle	—	—	—	—	—	—	1.1339	0.4871	0.0875	0.7438	0.2902	0.0865	1.0024	0.2325	0.0517
MLLE	—	—	—	—	—	—	1.6590	0.4347	0.1254	1.8838	0.3828	0.1070	1.6658	0.2806	0.0709
AdpLASSO	$0.4700_{(0.2689)}$	$0.4000_{(0.4190)}$	$0.5567_{(0.3859)}$	$0.8350_{(0.2425)}$	$0.7844_{(0.2427)}$	$0.8722_{(0.2206)}$	1.6441	0.5469	0.2052	1.8992	0.4338	0.1766	1.5472	0.3074	0.1055
LASSO	$0.4644_{(0.3021)}$	$0.3483_{(0.4322)}$	$0.4833_{(0.3677)}$	$0.7417_{(0.2569)}$	$0.7978_{(0.2261)}$	$0.8544_{(0.2230)}$	1.7766	0.6021	0.2767	2.1098	0.4487	0.2314	1.6641	0.3258	0.1348
MCP	$0.4589_{(0.2364)}$	$0.4267_{(0.4105)}$	$0.5567_{(0.3816)}$	$0.8417_{(0.2365)}$	$0.7833_{(0.2363)}$	$0.8856_{(0.2091)}$	1.6661	0.5162	0.1561	1.8239	0.4479	0.1346	1.6039	0.3143	0.0860
SCAD	$0.4722_{(0.2337)}$	$0.4267_{(0.4085)}$	$0.5517_{(0.3812)}$	$0.8383_{(0.2413)}$	$0.7800_{(0.2402)}$	$0.8778_{(0.2195)}$	1.6562	0.5188	0.1483	1.8079	0.4530	0.1272	1.6232	0.3124	0.0861
	n = 400														
Oracle	—	—	—	—	—	—	0.7176	0.3614	0.0662	0.3425	0.1530	0.0659	0.6768	0.2011	0.0549
MLLE	—	—	—	—	—	—	1.2502	0.4340	0.0896	1.3935	0.3189	0.0766	1.2189	0.2185	0.0532
AdpLASSO	$0.6011_{(0.1535)}$	$0.6833_{(0.4086)}$	$0.8517_{(0.3122)}$	$0.8683_{(0.2206)}$	$0.8322_{(0.2384)}$	$0.9022_{(0.2024)}$	1.2151	0.4796	0.0988	1.2501	0.3418	0.0859	1.0994	0.2273	0.0541
LASSO	$0.6544_{(0.2152)}$	$0.5483_{(0.4592)}$	$0.8233_{(0.3329)}$	$0.7467_{(0.2602)}$	$0.7667_{(0.2593)}$	$0.7867_{(0.2536)}$	1.2838	0.5554	0.1691	1.5429	0.3721	0.1500	1.1995	0.2440	0.0691
MCP	$0.5900_{(0.1532)}$	$0.7117_{(0.3828)}$	$0.8600_{(0.3094)}$	$0.8750_{(0.2169)}$	$0.8333_{(0.2369)}$	$0.9167_{(0.1987)}$	1.2334	0.4681	0.0849	1.2412	0.3263	0.0755	1.0959	0.2326	0.0509
SCAD	$0.5956_{(0.1447)}$	$0.7133_{(0.3829)}$	$0.8617_{(0.3088)}$	$0.8783_{(0.2149)}$	$0.8333_{(0.2353)}$	$0.9111_{(0.2028)}$	1.2209	0.4633	0.0825	1.2322	0.3268	0.0730	1.0902	0.2345	0.0501
	n = 600														
Oracle	—	—	—	—	—	—	0.4931	0.3087	0.0734	0.3156	0.1236	0.0690	0.4868	0.1796	0.0543
MLLE	—	—	—	—	—	—	0.8025	0.3754	0.1115	0.9222	0.1933	0.0932	0.8227	0.1839	0.0588
AdpLASSO	$0.7344_{(0.1597)}$	$0.7983_{(0.3924)}$	$0.9217_{(0.1996)}$	$0.9150_{(0.1881)}$	$0.9289_{(0.1892)}$	$0.9789_{(0.0900)}$	0.8513	0.4528	0.1154	0.8243	0.2104	0.1012	0.7135	0.1739	0.0568
LASSO	$0.8589_{(0.1780)}$	$0.6617_{(0.4623)}$	$0.9167_{(0.2271)}$	$0.7717_{(0.2625)}$	$0.9056_{(0.2120)}$	$0.7444_{(0.2738)}$	0.8996	0.5195	0.1626	1.1124	0.2676	0.1385	0.9293	0.1838	0.0631
MCP	$0.8367_{(0.1839)}$	$0.8517_{(0.3330)}$	$0.9767_{(0.1133)}$	$0.9133_{(0.1896)}$	$0.8978_{(0.2279)}$	$0.9467_{(0.1394)}$	0.7811	0.4085	0.1024	0.7399	0.1825	0.0864	0.5831	0.1728	0.0554
SCAD	$0.8333_{(0.1839)}$	$0.8583_{(0.3258)}$	$0.9767_{(0.1133)}$	$0.9167_{(0.1867)}$	$0.8967_{(0.2282)}$	$0.9444_{(0.1464)}$	0.7975	0.4116	0.1033	0.7482	0.1825	0.0877	0.5821	0.1732	0.0553
	n = 800														
Oracle	—	—	—	—	—	—	0.4614	0.3326	0.0619	0.2413	0.0986	0.0584	0.4829	0.1581	0.0486
MLLE	—	—	—	—	—	—	0.9958	0.3702	0.0905	1.1395	0.2504	0.0812	0.8393	0.1799	0.0512
AdpLASSO	$0.7833_{(0.2347)}$	$0.7717_{(0.3523)}$	$0.9383_{(0.2133)}$	$0.8633_{(0.2232)}$	$0.8189_{(0.2660)}$	$0.9633_{(0.1298)}$	0.9786	0.4266	0.0942	0.9848	0.2471	0.0841	0.7121	0.1696	0.0537
LASSO	$0.8733_{(0.2220)}$	$0.7583_{(0.3599)}$	$0.9450_{(0.1990)}$	$0.7833_{(0.2581)}$	$0.7000_{(0.3231)}$	$0.7411_{(0.2671)}$	1.0108	0.4708	0.1242	1.1539	0.3030	0.1123	0.8866	0.1700	0.0579
MCP	$0.8411_{(0.2368)}$	$0.8500_{(0.2848)}$	$0.9633_{(0.1744)}$	$0.8383_{(0.2343)}$	$0.7856_{(0.2903)}$	$0.9167_{(0.1831)}$	0.9485	0.3770	0.0784	0.9249	0.2235	0.0716	0.6262	0.1631	0.0476
SCAD	$0.8378_{(0.2384)}$	$0.8517_{(0.2842)}$	$0.9600_{(0.1784)}$	$0.8450_{(0.2316)}$	$0.7867_{(0.2867)}$	$0.9222_{(0.1784)}$	0.9638	0.3764	0.0773	0.9287	0.2276	0.0713	0.6278	0.1629	0.0473
d = 10	**n = 200**														
Oracle	—	—	—	—	—	—	0.9505	0.4974	0.1277	0.9805	0.3235	0.1182	0.8308	0.2478	0.0637
MLLE	—	—	—	—	—	—	1.4183	0.4360	0.1719	2.4018	0.3248	0.1455	1.7147	0.4082	0.0742
AdpLASSO	$0.4978_{(0.2565)}$	$0.3367_{(0.4069)}$	$0.5333_{(0.4076)}$	$0.9595_{(0.0715)}$	$0.9362_{(0.0972)}$	$0.9625_{(0.0741)}$	1.3498	0.6338	0.2486	1.9354	0.4319	0.2061	1.3817	0.3126	0.0996
LASSO	$0.5011_{(0.2943)}$	$0.2700_{(0.4175)}$	$0.4517_{(0.3988)}$	$0.9124_{(0.0806)}$	$0.9333_{(0.0928)}$	$0.9483_{(0.0794)}$	1.4418	0.6559	0.3217	2.1788	0.4406	0.2604	1.5181	0.3373	0.1123
MCP	$0.4856_{(0.2007)}$	$0.4033_{(0.3909)}$	$0.5733_{(0.3875)}$	$0.9643_{(0.0652)}$	$0.9121_{(0.1185)}$	$0.9658_{(0.0810)}$	1.3085	0.5850	0.1633	1.9423	0.4015	0.1397	1.2839	0.3119	0.0833
SCAD	$0.4900_{(0.1992)}$	$0.4083_{(0.3911)}$	$0.5767_{(0.3932)}$	$0.9657_{(0.0644)}$	$0.9012_{(0.1317)}$	$0.9596_{(0.0902)}$	1.3065	0.5855	0.1556	1.9433	0.4122	0.1330	1.3086	0.3125	0.0836
	n = 400														
Oracle	—	—	—	—	—	—	0.7064	0.3682	0.1009	0.5835	0.1737	0.0958	0.5880	0.1974	0.0578
MLLE	—	—	—	—	—	—	1.1659	0.4723	0.1869	1.8652	0.2792	0.1559	1.0682	0.222	0.0531
AdpLASSO	$0.6333_{(0.1430)}$	$0.5383_{(0.4691)}$	$0.9083_{(0.2435)}$	$0.9376_{(0.0719)}$	$0.9438_{(0.1023)}$	$0.9792_{(0.0520)}$	1.0881	0.5651	0.1851	1.4778	0.2902	0.1525	0.8380	0.2063	0.0569
LASSO	$0.6389_{(0.2155)}$	$0.3733_{(0.4740)}$	$0.8283_{(0.3163)}$	$0.8990_{(0.0731)}$	$0.9454_{(0.0914)}$	$0.9312_{(0.1087)}$	1.1901	0.6027	0.2640	1.7628	0.3562	0.2239	0.9875	0.2344	0.0670
MCP	$0.6033_{(0.1392)}$	$0.5967_{(0.4431)}$	$0.9333_{(0.1975)}$	$0.9443_{(0.0727)}$	$0.9054_{(0.1385)}$	$0.9821_{(0.0535)}$	1.1059	0.5332	0.1566	1.5774	0.2718	0.1274	0.8055	0.1920	0.0517
SCAD	$0.6111_{(0.1303)}$	$0.6083_{(0.4356)}$	$0.9283_{(0.1979)}$	$0.9490_{(0.0743)}$	$0.9017_{(0.1475)}$	$0.9788_{(0.0606)}$	1.0858	0.5366	0.1554	1.5586	0.2647	0.1280	0.8064	0.1931	0.0491
	n = 600														
Oracle	—	—	—	—	—	—	0.6282	0.3691	0.0699	0.3349	0.1507	0.0656	0.6261	0.1922	0.0539
MLLE	—	—	—	—	—	—	1.5853	0.4621	0.1178	1.8785	0.4041	0.0933	1.6192	0.2246	0.0626
AdpLASSO	$0.5544_{(0.1879)}$	$0.5050_{(0.4337)}$	$0.7183_{(0.4005)}$	$0.9300_{(0.0734)}$	$0.8717_{(0.1205)}$	$0.9308_{(0.1012)}$	1.5225	0.5138	0.1287	1.7028	0.4685	0.0957	1.4546	0.2560	0.0718
LASSO	$0.5567_{(0.2395)}$	$0.4733_{(0.4564)}$	$0.6817_{(0.4204)}$	$0.8943_{(0.0817)}$	$0.8429_{(0.1596)}$	$0.8717_{(0.1317)}$	1.5624	0.5539	0.1808	1.8107	0.4890	0.1378	1.5342	0.2823	0.0951
MCP	$0.5467_{(0.1758)}$	$0.5250_{(0.4272)}$	$0.7400_{(0.3910)}$	$0.9319_{(0.0838)}$	$0.8692_{(0.1263)}$	$0.9308_{(0.1023)}$	1.5513	0.4919	0.1099	1.7320	0.4426	0.0853	1.4713	0.2486	0.0604
SCAD	$0.5444_{(0.1763)}$	$0.5233_{(0.4224)}$	$0.7367_{(0.3888)}$	$0.9262_{(0.0886)}$	$0.8704_{(0.1232)}$	$0.9308_{(0.1038)}$	1.5480	0.4922	0.1090	1.7370	0.4425	0.0843	1.4715	0.2517	0.0598
	n = 800														
Oracle	—	—	—	—	—	—	0.4334	0.3366	0.0730	0.3427	0.1056	0.0690	0.4247	0.1583	0.0494
MLLE	—	—	—	—	—	—	0.9388	0.4102	0.1100	1.0839	0.1892	0.0962	0.9063	0.1567	0.0424
AdpLASSO	$0.6878_{(0.1121)}$	$0.8150_{(0.3747)}$	$0.9700_{(0.1500)}$	$0.9714_{(0.0572)}$	$0.9658_{(0.0829)}$	$0.9942_{(0.0264)}$	0.8708	0.4872	0.0964	0.7564	0.2115	0.0852	0.6450	0.1744	0.0402
LASSO	$0.7378_{(0.1447)}$	$0.7983_{(0.3924)}$	$0.9567_{(0.1913)}$	$0.9619_{(0.0633)}$	$0.9579_{(0.0876)}$	$0.9012_{(0.1174)}$	0.9325	0.5133	0.1156	0.8413	0.2873	0.1076	0.7948	0.1757	0.0426
MCP	$0.7378_{(0.1546)}$	$0.8467_{(0.3369)}$	$0.9867_{(0.0905)}$	$0.9710_{(0.0643)}$	$0.9304_{(0.1610)}$	$0.9729_{(0.0666)}$	0.8441	0.4815	0.0949	0.8760	0.1912	0.0812	0.5877	0.1680	0.0416
SCAD	$0.7322_{(0.1510)}$	$0.8617_{(0.3168)}$	$0.9850_{(0.0947)}$	$0.9700_{(0.0649)}$	$0.9204_{(0.1853)}$	$0.9708_{(0.0836)}$	0.8497	0.4811	0.0947	0.8860	0.1905	0.0816	0.5908	0.1714	0.0411

Web Table 4: Results of Example 1: average (SD) sensitivity, specificity, and estimation errors: $C = 3$. (**Continued from Table 3**).

Criteria	Sensitivity			Specificity			$L_2(\widehat{\boldsymbol{\beta}}_j)$			$L_2(\widehat{\boldsymbol{\sigma}}_j^2)$			$L_2(\widehat{\boldsymbol{\pi}}_j)$		
Component	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
d = 20	**n = 200**														
Oracle	—	—	—	—	—	—	0.8847	0.4505	0.0692	0.6231	0.2354	0.0825	0.9026	0.3294	0.0928
MLLE	—	—	—	—	—	—	2.4150	0.2449	0.0865	3.0950	0.2591	0.0847	2.6280	0.5813	0.0955
AdpLASSO	$0.3922_{(0.2791)}$	$0.1333_{(0.2615)}$	$0.1222_{(0.2245)}$	$0.9855_{(0.0372)}$	$0.9486_{(0.0652)}$	$0.9496_{(0.0628)}$	1.5544	0.6677	0.1387	2.3229	0.5798	0.1672	2.1041	0.4115	0.1683
LASSO	$0.4956_{(0.2719)}$	$0.1422_{(0.2681)}$	$0.1789_{(0.2744)}$	$0.9847_{(0.0486)}$	$0.9431_{(0.0721)}$	$0.9294_{(0.0795)}$	1.4868	0.6605	0.1442	2.2727	0.5143	0.1858	2.0811	0.4245	0.1772
MCP	$0.4611_{(0.2476)}$	$0.2033_{(0.3128)}$	$0.1989_{(0.2795)}$	$0.9825_{(0.0457)}$	$0.9290_{(0.0654)}$	$0.9190_{(0.0640)}$	1.4237	0.5401	0.0959	2.3932	0.4939	0.1378	2.1098	0.3868	0.1477
SCAD	$0.4456_{(0.2506)}$	$0.1978_{(0.3026)}$	$0.1844_{(0.2630)}$	$0.9814_{(0.0480)}$	$0.9255_{(0.0678)}$	$0.9125_{(0.0679)}$	1.4476	0.5579	0.0970	2.3869	0.5199	0.1320	2.1923	0.3921	0.1427
	n = 400														
Oracle	—	—	—	—	—	—	0.6117	0.3402	0.0597	0.3602	0.1264	0.0694	0.6735	0.2347	0.0743
MLLE	—	—	—	—	—	—	1.5067	0.4131	0.0789	2.7800	0.3614	0.1365	2.2260	0.4047	0.1324
AdpLASSO	$0.5989_{(0.1933)}$	$0.3044_{(0.3864)}$	$0.3011_{(0.3798)}$	$0.9920_{(0.0303)}$	$0.8912_{(0.1041)}$	$0.8876_{(0.0693)}$	1.0889	0.5669	0.0635	2.0191	0.5990	0.1172	1.8476	0.3394	0.1215
LASSO	$0.6211_{(0.2210)}$	$0.2844_{(0.3775)}$	$0.3344_{(0.3822)}$	$0.9802_{(0.0492)}$	$0.8908_{(0.1065)}$	$0.8769_{(0.0840)}$	1.1582	0.5837	0.0752	2.0132	0.6125	0.1427	1.8449	0.3669	0.1568
MCP	$0.6256_{(0.1422)}$	$0.2944_{(0.3992)}$	$0.2900_{(0.3884)}$	$0.9931_{(0.0290)}$	$0.8990_{(0.0793)}$	$0.8878_{(0.0737)}$	0.9780	0.5352	0.0528	2.1209	0.5278	0.0948	2.0148	0.3719	0.0934
SCAD	$0.6278_{(0.1376)}$	$0.2889_{(0.3907)}$	$0.2911_{(0.3813)}$	$0.9929_{(0.0336)}$	$0.8951_{(0.0828)}$	$0.8876_{(0.0740)}$	0.9744	0.5253	0.0530	2.1295	0.5399	0.0920	2.0125	0.3692	0.0937
	n = 600														
Oracle	—	—	—	—	—	—	0.5206	0.3012	0.0575	0.2914	0.1123	0.0620	0.5303	0.2013	0.0652
MLLE	—	—	—	—	—	—	1.0740	0.3920	0.0515	2.2576	0.3473	0.1071	1.9200	0.3376	0.1099
AdpLASSO	$0.6456_{(0.1216)}$	$0.4567_{(0.4658)}$	$0.4567_{(0.4807)}$	$0.9933_{(0.0357)}$	$0.8822_{(0.1234)}$	$0.8873_{(0.0755)}$	0.9047	0.4913	0.0428	1.6734	0.5258	0.0617	1.6541	0.3080	0.0686
LASSO	$0.6567_{(0.1477)}$	$0.4600_{(0.4621)}$	$0.4756_{(0.4759)}$	$0.9810_{(0.0405)}$	$0.8604_{(0.1295)}$	$0.8539_{(0.0824)}$	0.9382	0.4920	0.0467	1.7238	0.6113	0.0763	1.6019	0.2769	0.0856
MCP	$0.6544_{(0.0916)}$	$0.4333_{(0.4742)}$	$0.4689_{(0.4763)}$	$0.9945_{(0.0311)}$	$0.8947_{(0.0972)}$	$0.8992_{(0.0851)}$	0.8688	0.4742	0.0426	1.8228	0.4525	0.0625	1.7317	0.3384	0.0647
SCAD	$0.6533_{(0.0894)}$	$0.4344_{(0.4751)}$	$0.4578_{(0.4777)}$	$0.9943_{(0.0312)}$	$0.8957_{(0.0991)}$	$0.9012_{(0.0851)}$	0.8666	0.4774	0.0415	1.8015	0.4540	0.0585	1.7368	0.3476	0.0598
	n = 800														
Oracle	—	—	—	—	—	—	0.4053	0.3002	0.0544	0.2182	0.0958	0.0538	0.4650	0.1791	0.0596
MLLE	—	—	—	—	—	—	0.9461	0.4178	0.0391	1.7950	0.3703	0.0663	1.8255	0.3372	0.0729
AdpLASSO	$0.6611_{(0.0745)}$	$0.5011_{(0.4835)}$	$0.4956_{(0.4842)}$	$0.9988_{(0.0144)}$	$0.8976_{(0.1083)}$	$0.8990_{(0.0827)}$	0.8429	0.4953	0.0361	1.4930	0.5105	0.0498	1.5919	0.2987	0.0569
LASSO	$0.6878_{(0.1216)}$	$0.5400_{(0.4669)}$	$0.4956_{(0.4850)}$	$0.9857_{(0.0362)}$	$0.8582_{(0.1325)}$	$0.8486_{(0.0800)}$	0.8838	0.5161	0.0363	1.5547	0.5969	0.0586	1.5479	0.2557	0.0662
MCP	$0.6656_{(0.0510)}$	$0.5000_{(0.4903)}$	$0.4956_{(0.4873)}$	$0.9988_{(0.0127)}$	$0.9065_{(0.0934)}$	$0.9051_{(0.0865)}$	0.8150	0.4965	0.0345	1.5933	0.4380	0.0495	1.6502	0.3203	0.0516
SCAD	$0.6656_{(0.0510)}$	$0.4989_{(0.4896)}$	$0.4933_{(0.4857)}$	$0.9984_{(0.0166)}$	$0.9045_{(0.0966)}$	$0.9051_{(0.0865)}$	0.8149	0.4966	0.0343	1.5900	0.4413	0.0505	1.6569	0.3208	0.0529
d = 50	**n = 200**														
Oracle	—	—	—	—	—	—	1.1880	0.4564	0.0926	0.7124	0.2652	0.0951	0.9451	0.2230	0.0597
MLLE	—	—	—	—	—	—	40.344	0.2321	0.0663	125.66	0.2723	0.0514	32.176	0.8198	0.0608
AdpLASSO	$0.4689_{(0.3352)}$	$0.4217_{(0.4065)}$	$0.4383_{(0.3797)}$	$0.8240_{(0.2484)}$	$0.8062_{(0.2585)}$	$0.7938_{(0.2526)}$	2.1951	0.4730	0.2227	2.3250	0.4201	0.1927	2.3168	0.3967	0.1750
LASSO	$0.4133_{(0.3101)}$	$0.3017_{(0.4108)}$	$0.3450_{(0.3117)}$	$0.9848_{(0.0211)}$	$0.9776_{(0.0265)}$	$0.9744_{(0.0256)}$	2.1437	0.5992	0.2835	2.3110	0.4974	0.2434	2.0452	0.3191	0.1897
MCP	$0.3933_{(0.2332)}$	$0.3350_{(0.3545)}$	$0.4033_{(0.3712)}$	$0.9841_{(0.0819)}$	$0.9703_{(0.0832)}$	$0.9708_{(0.0824)}$	2.5522	0.5669	0.1513	2.2695	0.4985	0.1327	2.4620	0.2859	0.1112
SCAD	$0.4000_{(0.2329)}$	$0.3317_{(0.3434)}$	$0.4050_{(0.3772)}$	$0.9811_{(0.0828)}$	$0.9702_{(0.0839)}$	$0.9703_{(0.0821)}$	2.5530	0.5914	0.1524	3.2542	0.5091	0.1344	2.5090	0.2837	0.1159
	n = 400														
Oracle	—	—	—	—	—	—	0.8292	0.4252	0.0808	0.5226	0.2399	0.0835	0.7582	0.2307	0.0641
MLLE	—	—	—	—	—	—	2.6560	0.1830	0.0777	3.2830	0.2087	0.0595	3.7560	0.7442	0.0463
AdpLASSO	$0.3767_{(0.2360)}$	$0.3017_{(0.3856)}$	$0.4450_{(0.3979)}$	$0.9895_{(0.0113)}$	$0.9801_{(0.0273)}$	$0.9820_{(0.0238)}$	1.7019	0.6798	0.2610	2.1450	0.6171	0.2076	1.5710	0.2986	0.1125
LASSO	$0.5156_{(0.2687)}$	$0.3183_{(0.4436)}$	$0.5983_{(0.3894)}$	$0.9841_{(0.0140)}$	$0.9753_{(0.0286)}$	$0.9850_{(0.0205)}$	1.5590	0.6283	0.3020	2.0440	0.5326	0.2386	1.4284	0.3122	0.1139
MCP	$0.4444_{(0.1934)}$	$0.3533_{(0.4050)}$	$0.6617_{(0.4086)}$	$0.9861_{(0.0207)}$	$0.9759_{(0.0261)}$	$0.9836_{(0.0269)}$	1.5994	0.5609	0.1893	2.0424	0.5460	0.1432	1.4424	0.2625	0.0808
SCAD	$0.4389_{(0.1898)}$	$0.3317_{(0.3954)}$	$0.6550_{(0.4132)}$	$0.9864_{(0.0169)}$	$0.9744_{(0.0263)}$	$0.9838_{(0.0254)}$	1.6486	0.5602	0.1897	2.0740	0.5373	0.1452	1.4756	0.2723	0.0823
	n = 600														
Oracle	—	—	—	—	—	—	0.7489	0.3810	0.0840	0.4102	0.1602	0.0902	0.6591	0.2147	0.0603
MLLE	—	—	—	—	—	—	1.9910	0.4178	0.1770	3.3130	0.3138	0.1285	3.3190	0.6631	0.0730
AdpLASSO	$0.4922_{(0.2029)}$	$0.3183_{(0.4021)}$	$0.5867_{(0.3646)}$	$0.9847_{(0.0114)}$	$0.9798_{(0.0239)}$	$0.9807_{(0.0284)}$	1.4228	0.6366	0.2343	2.1166	0.5214	0.1771	1.2459	0.2997	0.0858
LASSO	$0.5400_{(0.2489)}$	$0.2533_{(0.4099)}$	$0.6017_{(0.3971)}$	$0.9832_{(0.0092)}$	$0.9768_{(0.0328)}$	$0.9853_{(0.0234)}$	1.4331	0.6275	0.3046	2.1956	0.5131	0.2290	1.2370	0.3321	0.1055
MCP	$0.5078_{(0.1776)}$	$0.3450_{(0.3967)}$	$0.8283_{(0.2972)}$	$0.9855_{(0.0116)}$	$0.9649_{(0.0362)}$	$0.9872_{(0.0305)}$	1.4685	0.5357	0.1772	2.1265	0.4445	0.1354	1.1316	0.2397	0.0636
SCAD	$0.4867_{(0.1793)}$	$0.3633_{(0.3875)}$	$0.8217_{(0.3099)}$	$0.9855_{(0.0120)}$	$0.9632_{(0.0393)}$	$0.9859_{(0.0311)}$	1.4795	0.5362	0.1750	2.0962	0.4475	0.1354	1.1242	0.2492	0.0634
	n = 800														
Oracle	—	—	—	—	—	—	0.5805	0.3965	0.0993	0.4517	0.1583	0.0967	0.5172	0.1903	0.0560
MLLE	—	—	—	—	—	—	1.6290	0.5095	0.2281	3.2746	0.3564	0.1757	2.5160	0.5238	0.0697
AdpLASSO	$0.5678_{(0.1731)}$	$0.3733_{(0.4277)}$	$0.7950_{(0.3477)}$	$0.9853_{(0.0102)}$	$0.9795_{(0.0248)}$	$0.9884_{(0.0207)}$	1.2957	0.6338	0.2102	1.9223	0.4454	0.1576	0.9566	0.2544	0.0685
LASSO	$0.6178_{(0.1999)}$	$0.3567_{(0.4379)}$	$0.8117_{(0.3663)}$	$0.9836_{(0.0102)}$	$0.9823_{(0.0224)}$	$0.9803_{(0.0231)}$	1.2639	0.6388	0.2336	1.9158	0.4390	0.1775	1.0092	0.2894	0.0780
MCP	$0.5644_{(0.1656)}$	$0.4300_{(0.3945)}$	$0.9400_{(0.1997)}$	$0.9845_{(0.0149)}$	$0.9619_{(0.0469)}$	$0.9935_{(0.0188)}$	1.3264	0.5596	0.1822	2.0416	0.3556	0.1493	0.7944	0.1909	0.0509
SCAD	$0.5544_{(0.1669)}$	$0.4483_{(0.3790)}$	$0.9283_{(0.2180)}$	$0.9855_{(0.0133)}$	$0.9556_{(0.0543)}$	$0.9922_{(0.0227)}$	1.3555	0.5579	0.1771	2.0526	0.3569	0.1443	0.8089	0.1938	0.0515

Web Table 5: Results of Example 2: Average (SD) sensitivity, specificity, and estimation errors:
$(C = 2, d = 5, n = 200)$.

Criteria	Sensitivity		Specificity		$L_2(\widehat{\boldsymbol{\beta}}_j)$		$L_2(\widehat{\boldsymbol{\sigma}}_j^2)$		$L_2(\widehat{\boldsymbol{\pi}}_j)$	
Component	1	2	1	2	1	2	1	2	1	2
p = 0.05										
Oracle	—	—	—	—	1.1871	0.0771	0.7560	0.0919	0.0765	0.0919
MLLE	—	—	—	—	1.5363	0.0678	1.3980	0.0926	0.0740	0.0926
AdpLASSO	$0.0878_{(0.1831)}$	$0.5883_{(0.3630)}$	$0.8933_{(0.2092)}$	$0.9889_{(0.0763)}$	1.7500	0.1263	1.2046	0.1338	0.1192	0.1338
LASSO	$0.0822_{(0.1963)}$	$0.6117_{(0.4090)}$	$0.8983_{(0.2057)}$	$0.9578_{(0.1324)}$	1.7360	0.1348	1.2040	0.1390	0.1189	0.1390
MCP	$0.4211_{(0.3000)}$	$0.6517_{(0.2969)}$	$0.8917_{(0.2103)}$	$0.9467_{(0.1340)}$	1.5986	0.0888	1.2499	0.1029	0.0956	0.1029
SCAD	$0.4311_{(0.2975)}$	$0.6600_{(0.2995)}$	$0.8933_{(0.2092)}$	$0.9456_{(0.1264)}$	1.5994	0.0903	1.2441	0.1038	0.0950	0.1038
p = 0.1										
Oracle	—	—	—	—	0.9970	0.0825	0.6118	0.0724	0.0991	0.0724
MLLE	—	—	—	—	1.3279	0.0844	1.0951	0.0863	0.0976	0.0863
AdpLASSO	$0.2611_{(0.2917)}$	$0.7367_{(0.3307)}$	$0.8950_{(0.2198)}$	$0.9667_{(0.1377)}$	1.6326	0.1669	1.0047	0.1540	0.1623	0.1540
LASSO	$0.3211_{(0.3375)}$	$0.6983_{(0.3877)}$	$0.8667_{(0.2532)}$	$0.8889_{(0.2152)}$	1.6067	0.1777	1.1127	0.1582	0.1578	0.1582
MCP	$0.5511_{(0.2833)}$	$0.8167_{(0.2613)}$	$0.9067_{(0.1951)}$	$0.9667_{(0.1107)}$	1.3219	0.0977	0.9244	0.0872	0.1384	0.0872
SCAD	$0.5544_{(0.2853)}$	$0.8200_{(0.2572)}$	$0.9067_{(0.1951)}$	$0.9656_{(0.1121)}$	1.3189	0.0984	0.9256	0.0872	0.1370	0.0872
p = 0.2										
Oracle	—	—	—	—	0.9110	0.1351	0.4973	0.0793	0.2097	0.0793
MLLE	—	—	—	—	1.1351	0.1387	0.8888	0.0935	0.1902	0.0935
AdpLASSO	$0.6233_{(0.2898)}$	$0.8933_{(0.2459)}$	$0.9483_{(0.1681)}$	$0.9411_{(0.1562)}$	1.2894	0.1804	0.7608	0.1096	0.2777	0.1096
LASSO	$0.6611_{(0.3559)}$	$0.8550_{(0.3030)}$	$0.8750_{(0.2492)}$	$0.7433_{(0.2993)}$	1.3308	0.2158	0.9016	0.1253	0.2633	0.1253
MCP	$0.6400_{(0.2215)}$	$0.9100_{(0.1967)}$	$0.9550_{(0.1433)}$	$0.9589_{(0.1195)}$	1.0720	0.1371	0.7125	0.0820	0.2793	0.0820
SCAD	$0.6478_{(0.2232)}$	$0.9167_{(0.1911)}$	$0.9567_{(0.1409)}$	$0.9567_{(0.1218)}$	1.0636	0.1386	0.7111	0.0819	0.2787	0.0819
p = 0.3										
Oracle	—	—	—	—	0.8549	0.1253	0.5063	0.0858	0.2964	0.0858
MLLE	—	—	—	—	1.0499	0.1552	0.8202	0.0929	0.3081	0.0929
AdpLASSO	$0.5878_{(0.2283)}$	$0.8850_{(0.2605)}$	$0.9700_{(0.1189)}$	$0.9322_{(0.1371)}$	1.1651	0.2074	0.7959	0.1184	0.3692	0.1184
LASSO	$0.5822_{(0.3184)}$	$0.8567_{(0.2911)}$	$0.8933_{(0.2246)}$	$0.8422_{(0.2101)}$	1.2739	0.2632	0.9632	0.1585	0.3676	0.1585
MCP	$0.6022_{(0.1645)}$	$0.8900_{(0.2410)}$	$0.9683_{(0.1220)}$	$0.9733_{(0.0906)}$	0.9990	0.1418	0.6924	0.0799	0.3990	0.0799
SCAD	$0.6089_{(0.1624)}$	$0.9017_{(0.2191)}$	$0.9667_{(0.1249)}$	$0.9733_{(0.0946)}$	0.9956	0.1412	0.6828	0.0823	0.3946	0.0823
p = 0.4										
Oracle	—	—	—	—	0.9225	0.1649	0.5525	0.0980	0.3561	0.0980
MLLE	—	—	—	—	1.1974	0.2172	0.9813	0.1080	0.3574	0.1080
AdpLASSO	$0.4978_{(0.2993)}$	$0.8600_{(0.2720)}$	$0.9233_{(0.1805)}$	$0.9189_{(0.1581)}$	1.4347	0.3391	0.9183	0.1653	0.4358	0.1653
LASSO	$0.4611_{(0.3782)}$	$0.8283_{(0.3082)}$	$0.8800_{(0.2326)}$	$0.8278_{(0.2305)}$	1.6320	0.4374	1.1354	0.2329	0.4210	0.2329
MCP	$0.5667_{(0.1937)}$	$0.8767_{(0.2380)}$	$0.9250_{(0.1788)}$	$0.9467_{(0.1283)}$	1.1739	0.2178	0.7907	0.0997	0.4791	0.0997
SCAD	$0.5678_{(0.1972)}$	$0.8833_{(0.2307)}$	$0.9233_{(0.1805)}$	$0.9467_{(0.1283)}$	1.1718	0.2147	0.7915	0.0983	0.4769	0.0983
p = 0.5										
Oracle	—	—	—	—	0.9803	0.1717	0.5442	0.1024	0.4244	0.1024
MLLE	—	—	—	—	1.3809	0.2466	1.2807	0.1315	0.4193	0.1315
AdpLASSO	$0.4678_{(0.2659)}$	$0.6867_{(0.3660)}$	$0.8450_{(0.2316)}$	$0.9033_{(0.1722)}$	1.6332	0.3705	1.2317	0.1791	0.5217	0.1791
LASSO	$0.4300_{(0.3083)}$	$0.6383_{(0.4007)}$	$0.8383_{(0.2548)}$	$0.8744_{(0.2062)}$	1.9366	0.4615	1.5444	0.2358	0.5642	0.2358
MCP	$0.5144_{(0.2080)}$	$0.7417_{(0.3433)}$	$0.8400_{(0.2336)}$	$0.8956_{(0.1596)}$	1.4062	0.2975	1.0925	0.1239	0.5012	0.1239
SCAD	$0.5156_{(0.2061)}$	$0.7417_{(0.3383)}$	$0.8383_{(0.2343)}$	$0.8900_{(0.1728)}$	1.4023	0.2993	1.1004	0.1246	0.5083	0.1246

Web Table 6: Results of Example 2: Average (SD) sensitivity, specificity, and estimation errors:
$(C = 2, d = 5, n = 400)$.

Criteria	Sensitivity		Specificity		$L_2(\widehat{\boldsymbol{\beta}}_j)$		$L_2(\widehat{\boldsymbol{\sigma}}_j^2)$		$L_2(\widehat{\boldsymbol{\pi}}_j)$	
Component	1	2	1	2	1	2	1	2	1	2
p = 0.05										
Oracle	—	—	—	—	0.9785	0.0572	0.5666	0.0657	0.0609	0.0657
MLLE	—	—	—	—	1.3969	0.0650	1.1504	0.0748	0.0758	0.0748
AdpLASSO	$0.1967_{(0.2458)}$	$0.7067_{(0.3527)}$	$0.8867_{(0.2213)}$	$0.9789_{(0.1153)}$	1.6640	0.1458	1.0619	0.1949	0.1355	0.1949
LASSO	$0.2322_{(0.3069)}$	$0.6600_{(0.3740)}$	$0.8700_{(0.2582)}$	$0.8533_{(0.2360)}$	1.6322	0.1719	1.1591	0.2366	0.1429	0.2366
MCP	$0.6044_{(0.2404)}$	$0.8433_{(0.2497)}$	$0.8783_{(0.2149)}$	$0.9544_{(0.1240)}$	1.3575	0.0695	1.0012	0.0881	0.0883	0.0881
SCAD	$0.6244_{(0.2339)}$	$0.8583_{(0.2365)}$	$0.8767_{(0.2159)}$	$0.9433_{(0.1394)}$	1.3490	0.0681	1.0176	0.0861	0.0875	0.0861
p = 0.1										
Oracle	—	—	—	—	0.7820	0.0814	0.4509	0.0739	0.1127	0.0739
MLLE	—	—	—	—	1.2497	0.0990	0.9775	0.0916	0.1111	0.0916
AdpLASSO	$0.5433_{(0.2936)}$	$0.8217_{(0.2725)}$	$0.9033_{(0.2101)}$	$0.9367_{(0.1638)}$	1.3122	0.1192	0.8449	0.1247	0.1805	0.1247
LASSO	$0.5411_{(0.3670)}$	$0.7817_{(0.3320)}$	$0.8933_{(0.2283)}$	$0.8200_{(0.2442)}$	1.3590	0.1473	0.9677	0.1489	0.1835	0.1489
MCP	$0.6489_{(0.2447)}$	$0.8583_{(0.2330)}$	$0.9083_{(0.1938)}$	$0.9400_{(0.1394)}$	1.1660	0.0959	0.8370	0.0806	0.1529	0.0806
SCAD	$0.6656_{(0.2384)}$	$0.8667_{(0.2252)}$	$0.9100_{(0.1924)}$	$0.9278_{(0.1529)}$	1.1679	0.0956	0.8470	0.0806	0.1506	0.0806
p = 0.2										
Oracle	—	—	—	—	0.5718	0.0882	0.3571	0.0822	0.1707	0.0822
MLLE	—	—	—	—	0.9225	0.1090	0.6945	0.0853	0.1805	0.0853
AdpLASSO	$0.6789_{(0.2703)}$	$0.9233_{(0.1939)}$	$0.9650_{(0.1278)}$	$0.9644_{(0.1197)}$	1.0680	0.1257	0.6026	0.1044	0.2490	0.1044
LASSO	$0.4867_{(0.4440)}$	$0.7850_{(0.2856)}$	$0.9367_{(0.1810)}$	$0.9000_{(0.2103)}$	1.3737	0.2918	0.9929	0.2213	0.2328	0.2213
MCP	$0.6644_{(0.1442)}$	$0.9500_{(0.1503)}$	$0.9633_{(0.1306)}$	$0.9733_{(0.0946)}$	0.9195	0.0933	0.5101	0.0733	0.2413	0.0733
SCAD	$0.6667_{(0.1363)}$	$0.9517_{(0.1480)}$	$0.9633_{(0.1306)}$	$0.9733_{(0.0946)}$	0.9138	0.0927	0.5069	0.0731	0.2403	0.0731
p = 0.3										
Oracle	—	—	—	—	0.5387	0.0911	0.3518	0.0846	0.2383	0.0846
MLLE	—	—	—	—	0.7801	0.1099	0.5970	0.0855	0.2577	0.0855
AdpLASSO	$0.8533_{(0.2513)}$	$0.9833_{(0.1069)}$	$0.9917_{(0.0641)}$	$0.9922_{(0.0635)}$	0.8353	0.1166	0.4036	0.0924	0.3013	0.0924
LASSO	$0.7033_{(0.4308)}$	$0.9683_{(0.1411)}$	$0.9750_{(0.1092)}$	$0.9711_{(0.1184)}$	1.1928	0.2378	0.7511	0.1774	0.3229	0.1774
MCP	$0.7322_{(0.1583)}$	$0.9900_{(0.0701)}$	$0.9900_{(0.0701)}$	$0.9900_{(0.0631)}$	0.8046	0.0941	0.3946	0.0786	0.3138	0.0786
SCAD	$0.7344_{(0.1573)}$	$0.9900_{(0.0701)}$	$0.9883_{(0.0756)}$	$0.9900_{(0.0631)}$	0.8013	0.0949	0.3934	0.0787	0.3132	0.0787
p = 0.4										
Oracle	—	—	—	—	0.5475	0.0976	0.3481	0.0840	0.2983	0.0840
MLLE	—	—	—	—	1.0006	0.1478	0.8966	0.0999	0.2903	0.0999
AdpLASSO	$0.7189_{(0.3605)}$	$0.9233_{(0.2143)}$	$0.9433_{(0.1588)}$	$0.9478_{(0.1654)}$	1.1094	0.2103	0.7081	0.1482	0.3535	0.1482
LASSO	$0.6500_{(0.4008)}$	$0.9167_{(0.2482)}$	$0.9033_{(0.1978)}$	$0.8700_{(0.2209)}$	1.4012	0.3134	1.0900	0.2231	0.3691	0.2231
MCP	$0.7533_{(0.2262)}$	$0.9433_{(0.1588)}$	$0.9367_{(0.1666)}$	$0.9478_{(0.1513)}$	0.9008	0.1353	0.5396	0.0890	0.3476	0.0890
SCAD	$0.7489_{0.2229}$	$0.9417_{(0.1608)}$	$0.9367_{(0.1666)}$	$0.9478_{(0.1513)}$	0.9080	0.1382	0.5417	0.0889	0.3508	0.0889
p = 0.5										
Oracle	—	—	—	—	0.7021	0.1657	0.4427	0.1008	0.3494	0.1008
MLLE	—	—	—	—	1.0445	0.1685	0.9230	0.1127	0.3283	0.1127
AdpLASSO	$0.5489_{(0.4164)}$	$0.9017_{(0.2794)}$	$0.9433_{(0.1588)}$	$0.9756_{(0.1063)}$	1.3950	0.3236	0.9470	0.2220	0.4022	0.2220
LASSO	$0.5311_{(0.4216)}$	$0.8850_{(0.3156)}$	$0.9100_{(0.1967)}$	$0.8833_{(0.1968)}$	1.6108	0.3979	1.2915	0.2812	0.4112	0.2812
MCP	$0.7178_{(0.1994)}$	$0.9417_{(0.1608)}$	$0.9433_{(0.1588)}$	$0.9322_{(0.1620)}$	0.9891	0.1540	0.5544	0.0961	0.4057	0.0961
SCAD	$0.7078_{(0.1951)}$	$0.9400_{(0.1678)}$	$0.9433_{(0.1588)}$	$0.9433_{(0.1497)}$	0.9994	0.1551	0.5451	0.0947	0.4111	0.0947

Web Table 7: Average (SD) sensitivity, specificity, and estimation errors over 100 replicates.

$C=2$		Criteria	Sensitivity		Specificity		$L_2(\widehat{\boldsymbol{\beta}}_j)$		$L_2(\widehat{\boldsymbol{\sigma}}_j^2)$		$L_2(\widehat{\boldsymbol{\pi}}_j)$	
d	n	Component	1	2	1	2	1	2	1	2	1	2
						Correct model: t FM-VCR						
10	200	Oracle	—	—	—	—	.303	.214	.171	.249	.098	.098
		MLLE	—	—	—	—	.474	.420	.333	.326	.153	.153
		AdpLASSO	$.482_{(.235)}$	$.655_{(.388)}$	$.976_{(.059)}$	$.968_{(.062)}$	.402	.394	.234	.295	.265	.265
		LASSO	$.500_{(.265)}$	$.638_{(.399)}$	$.955_{(.072)}$	$.939_{(.075)}$	.407	.412	.273	.319	.328	.328
		MCP	$.443_{(.199)}$	$.700_{(.325)}$	$.993_{(.038)}$	$.985_{(.047)}$	.405	.350	.185	.241	.174	.174
		SCAD	$.446_{(.196)}$	$.702_{(.320)}$	$.992_{(.040)}$	$.985_{(.048)}$	.404	.346	.186	.235	.167	.167
	400	Oracle	—	—	—	—	.230	.153	.181	.320	.092	.092
		MLLE	—	—	—	—	.372	.329	.225	.274	.132	.132
		AdpLASSO	$.587_{(.190)}$	$.940_{(.195)}$	$.993_{(.031)}$	$.987_{(.038)}$	.327	.228	.180	.190	.172	.172
		LASSO	$.556_{(.250)}$	$.857_{(.319)}$	$.968_{(.061)}$	$.953_{(.063)}$	.371	.286	.233	.212	.270	.270
		MCP	$.556_{(.175)}$	$.945_{(.167)}$	$.997_{(.020)}$	$.995_{(.024)}$	.326	.221	.160	.192	.154	.154
		SCAD	$.571_{(.167)}$	$.942_{(.171)}$	$.997_{(.020)}$	$.995_{(.025)}$	.319	.220	.158	.195	.149	.149
d	n					Misspecified model: Gaussian FM-VCR						
10	200	Oracle	—	—	—	—	1.083	.342	.786	.181	.330	.181
		MLLE	—	—	—	—	1.748	.426	1.667	.234	.443	.234
		AdpLASSO	$.394_{(.272)}$	$.538_{(.414)}$	$.942_{(.085)}$	$.939_{(.076)}$	1.568	.394	1.489	.309	.387	.309
		LASSO	$.391_{(.335)}$	$.515_{(.476)}$	$.920_{(.090)}$	$.925_{(.086)}$	1.721	.588	1.721	.425	.558	.425
		MCP	$.441_{(.236)}$	$.588_{(.368)}$	$.943_{(.100)}$	$.942_{(.101)}$	1.508	.291	1.476	.218	.281	.218
		SCAD	$.444_{(.235)}$	$.583_{(.370)}$	$.937_{(.110)}$	$.937_{(.111)}$	1.509	.280	1.475	.204	.271	.204
	400	Oracle	—	—	—	—	.809	.282	.553	.111	.199	.111
		MLLE	—	—	—	—	1.450	.329	1.210	.209	.325	.209
		AdpLASSO	$.544_{(.247)}$	$.827_{(.342)}$	$.967_{(.073)}$	$.970_{(.061)}$	1.216	.275	.986	.226	.272	.226
		LASSO	$.453_{(.328)}$	$.650_{(.467)}$	$.933_{(.077)}$	$.932_{(.075)}$	1.525	.470	1.360	.373	.469	.373
		MCP	$.567_{(.176)}$	$.892_{(.250)}$	$.957_{(.113)}$	$.970_{(.070)}$	1.249	.201	.897	.177	.219	.177
		SCAD	$.573_{(.173)}$	$.892_{(.247)}$	$.953_{(.127)}$	$.970_{(.076)}$	1.234	.202	.895	.174	.212	.174

Web Table 8: Average (SD) sensitivity, specificity, and estimation errors for models with orders $C = 1, 2, 3, 4$.

$C=1$		Criteria	Sensitivity		Specificity		$L_2(\widehat{\boldsymbol{\beta}}_j)$		$L_2(\widehat{\boldsymbol{\sigma}}_j^2)$		$L_2(\widehat{\boldsymbol{\pi}}_j)$	
d	n	Component	1	2	1	2	1	2	1	2	1	2
10	200	Oracle	—	—	—	—	.357	—	.634	—	—	—
		MLLE	—	—	—	—	.396	—	.616	—	—	—
		AdpLASSO	$.772_{(.073)}$	—	$.831_{(.167)}$	—	.383	—	.672	—	—	—
		LASSO	$.802_{(.101)}$	—	$.708_{(.221)}$	—	.390	—	.673	—	—	—
		MCP	$.771_{(.069)}$	—	$.859_{(.151)}$	—	.375	—	.667	—	—	—
		SCAD	$.767_{(.064)}$	—	$.874_{(.137)}$	—	.371	—	.668	—	—	—
	400	Oracle	—	—	—	—	.339	—	.558	—	—	—
		MLLE	—	—	—	—	.365	—	.554	—	—	—
		AdpLASSO	$.984_{(.102)}$	—	$.712_{(.243)}$	—	.365	—	.587	—	—	—
		LASSO	$.972_{(.079)}$	—	$.738_{(.255)}$	—	.375	—	.593	—	—	—
		MCP	$.943_{(.105)}$	—	$.746_{(.235)}$	—	.358	—	.583	—	—	—
		SCAD	$.953_{(.098)}$	—	$.636_{(.287)}$	—	.354	—	.579	—	—	—
$C=2$					the true model							
10	200	Oracle	—	—	—	—	.698	.162	.397	.089	.314	.089
		MLLE	—	—	—	—	1.23	.287	1.02	.138	.335	.138
		AdpLASSO	$.589_{(.239)}$	$.893_{(.259)}$	$.970_{(.067)}$	$.966_{(.061)}$	1.09	.288	.848	.161	.501	.161
		LASSO	$.470_{(.321)}$	$.775_{(.401)}$	$.928_{(.100)}$	$.917_{(.087)}$	1.32	.383	1.15	.294	.505	.294
		MCP	$.581_{(.194)}$	$.895_{(.245)}$	$.984_{(.045)}$	$.981_{(.048)}$	1.04	.238	.763	.106	.503	.106
		SCAD	$.594_{(.188)}$	$.892_{(.243)}$	$.983_{(.049)}$	$.980_{(.050)}$	1.03	.237	.761	.107	.501	.107
	400	Oracle	—	—	—	—	.439	.089	.227	.082	.273	.082
		MLLE	—	—	—	—	.878	.149	.607	.087	.337	.087
		AdpLASSO	$.690_{(.127)}$	$.993_{(.057)}$	$.997_{(.023)}$	$.995_{(.029)}$	.808	.165	.390	.069	.411	.069
		LASSO	$.766_{(.230)}$	$.970_{(.166)}$	$.970_{(.066)}$	$.962_{(.064)}$	.862	.232	.518	.108	.425	.108
		MCP	$.692_{(.104)}$	$.992_{(.064)}$	$.997_{(.020)}$	$.998_{(.014)}$	.789	.145	.374	.06	.412	.062
		SCAD	$.688_{(.098)}$	$.992_{(.064)}$	$.996_{(.023)}$	$.998_{(.016)}$	.792	.146	.377	.062	.412	.062
$C=3$												
	200	Oracle	—	—	—	—	.201	.111	.314	.125	.083	.083
		MLLE	—	—	—	—	.371	.297	.173	.115	.216	.092
		AdpLASSO	$.654_{(.140)}$	$.962_{(.139)}$	$.987_{(.044)}$	$.992_{(.033)}$	.295	.199	.473	.182	.121	.096
		LASSO	$.641_{(.167)}$	$.938_{(.209)}$	$.967_{(.065)}$	$.963_{(.066)}$	.302	.226	.510	.239	.121	.096
		MCP	$.612_{(.180)}$	$.958_{(.138)}$	$.997_{(.025)}$	$.999_{(.010)}$	.319	.192	.407	.145	.157	.085
		SCAD	$.623_{(.177)}$	$.947_{(.160)}$	$1.00_{(.000)}$	$.999_{(.010)}$	.310	.192	.412	.153	.150	.084
10	400	Oracle	—	—	—	—	.125	.067	.260	.081	.081	.082
		MLLE	—	—	—	—	.319	.176	.206	.096	.247	.072
		AdpLASSO	$.722_{(.124)}$	$1.00_{(.000)}$	$.999_{(.014)}$	$1.00_{(.007)}$	.260	.136	.401	.169	.118	.062
		LASSO	$.717_{(.119)}$	$.997_{(.058)}$	$.995_{(.027)}$	$.994_{(.027)}$	.259	.141	.446	.213	.097	.068
		MCP	$.706_{(.117)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	.261	.138	.380	.146	.127	.059
		SCAD	$.708_{(.116)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	.258	.136	.374	.145	.135	.059
$C=4$												
10	200	Oracle	—	—	—	—	.199	.106	.313	.113	.083	.083
		MLLE	—	—	—	—	.389	.322	.119	.098	.294	.165
		AdpLASSO	$.616_{(.148)}$	$.963_{(.149)}$	$.994_{(.028)}$	$.993_{(.03)}$	.310	.204	.434	.191	.161	.102
		LASSO	$.642_{(.140)}$	$.922_{(.234)}$	$.980_{(.051)}$	$.977_{(.05)}$	.299	.225	.529	.234	.156	.135
		MCP	$.536_{(.188)}$	$.957_{(.141)}$	$1.00_{(.008)}$	$.999_{(.01)}$	.372	.205	.342	.133	.237	.090
		SCAD	$.561_{(.180)}$	$.948_{(.152)}$	$.999_{(.012)}$	$.999_{(.01)}$	.349	.203	.356	.123	.219	.092
	400	Oracle	—	—	—	—	.125	.067	.260	.081	.081	.082
		MLLE	—	—	—	—	.338	.204	.149	.076	.306	.132
		AdpLASSO	$.701_{(.105)}$	$1.00_{(.000)}$	$.999_{(.014)}$	$.999_{(.010)}$	.266	.147	.418	.175	.120	.068
		LASSO	$.702_{(.103)}$	$.998_{(.029)}$	$.994_{(.028)}$	$.994_{(.026)}$	.265	.150	.452	.226	.104	.076
		MCP	$.693_{(.116)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	.273	.153	.369	.120	.160	.071
		SCAD	$.696_{(.118)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	.268	.154	.356	.115	.176	.075

Web Table 9: Result of the FMR and FM-VCR models; average (SD) Sensitivity, Specificity, and estimation errors.

$C=2$		Criteria	Sensitivity		Specificity		$L_2(\widehat{\boldsymbol{\beta}}_j)$		$L_2(\widehat{\boldsymbol{\sigma}}_j)$		$L_2(\widehat{\boldsymbol{\pi}}_j)$	
d	n	Component	1	2	1	2	1	2	1	2	1	2
True Gaussian FMR model												
10	200	Oracle	—	—	—	—	.050	.055	.063	.074	.147	.147
		MLLE	—	—	—	—	.115	.140	.138	.158	.232	.232
		AdpLASSO	$.529_{(.364)}$	$.485_{(.389)}$	$.848_{(.348)}$	$.837_{(.347)}$	.125	.133	.138	.165	.352	.352
		LASSO	$.732_{(.361)}$	$.568_{(.436)}$	$.843_{(.279)}$	$.815_{(.280)}$	.101	.126	.112	.185	.321	.321
		MCP	$.766_{(.277)}$	$.702_{(.314)}$	$.864_{(.249)}$	$.806_{(.244)}$	.102	.125	.115	.153	.267	.267
		SCAD	$.762_{(.283)}$	$.725_{(.287)}$	$.839_{(.259)}$	$.775_{(.262)}$	.106	.125	.120	.145	.259	.259
	400	Oracle	—	—	—	—	.027	.031	.031	.039	.078	.078
		MLLE	—	—	—	—	.075	.107	.081	.115	.185	.185
		AdpLASSO	$.708_{(.312)}$	$.592_{(.412)}$	$.878_{(.318)}$	$.870_{(.320)}$	.099	.127	.088	.148	.306	.306
		LASSO	$.941_{(.185)}$	$.790_{(.374)}$	$.881_{(.191)}$	$.850_{(.194)}$	.060	.087	.053	.118	.209	.209
		MCP	$.943_{(.152)}$	$.888_{(.242)}$	$.944_{(.130)}$	$.885_{(.189)}$	.050	.075	.062	.094	.169	.169
		SCAD	$.951_{(.144)}$	$.897_{(.230)}$	$.922_{(.163)}$	$.865_{(.222)}$	.050	.076	.064	.097	.166	.166
Misspecified Gaussian FM-VCR model												
10	200	Oracle	—	—	—	—	.050	.055	.062	.072	.139	.139
		MLLE	—	—	—	—	.099	.132	.060	.098	.133	.133
		AdpLASSO	$.562_{(.323)}$	$.563_{(.356)}$	$.990_{(.042)}$	$.970_{(.060)}$	.100	.100	.102	.112	.255	.255
		LASSO	$.643_{(.323)}$	$.600_{(.359)}$	$.991_{(.034)}$	$.965_{(.065)}$	.093	.096	.086	.120	.243	.243
		MCP	$.588_{(.320)}$	$.545_{(.310)}$	$.991_{(.040)}$	$.976_{(.054)}$	.097	.105	.077	.092	.215	.215
		SCAD	$.583_{(.311)}$	$.548_{(.304)}$	$.991_{(.040)}$	$.975_{(.058)}$	.097	.100	.066	.080	.202	.202
	400	Oracle	—	—	—	—	.028	.031	.030	.040	.074	.074
		MLLE	—	—	—	—	.073	.096	.047	.070	.125	.125
		AdpLASSO	$.841_{(.252)}$	$.800_{(.309)}$	$.993_{(.032)}$	$.972_{(.058)}$	.066	.073	.046	.064	.172	.172
		LASSO	$.899_{(.211)}$	$.810_{(.325)}$	$.987_{(.043)}$	$.969_{(.061)}$	.059	.072	.048	.073	.177	.177
		MCP	$.876_{(.227)}$	$.810_{(.299)}$	$.993_{(.032)}$	$.979_{(.052)}$	.054	.062	.039	.055	.144	.144
		SCAD	$.872_{(.234)}$	$.812_{(.298)}$	$.994_{(.029)}$	$.981_{(.051)}$	.051	.059	.037	.050	.126	.126

Web Table 10: Results of simultaneous selection of (h, λ) with $C = 2$: average (SD) sensitivity, specificity, and estimation errors.

$C = 2$		Criteria	Sensitivity		Specificity		$L_2(\widehat{\boldsymbol{\beta}}_j)$		$L_2(\widehat{\boldsymbol{\sigma}}_j)$		$L_2(\widehat{\boldsymbol{\pi}}_j)$	
d	n	Component	1	2	1	2	1	2	1	2	1	2
10	200	Oracle	—	—	—	—	.699	.162	.400	.089	.314	.088
		MLLE	—	—	—	—	1.229	.287	1.015	.136	.335	.136
		AdpLASSO	$.604_{(.234)}$	$.905_{(.249)}$	$.965_{(.071)}$	$.962_{(.066)}$	1.080	.284	.834	.154	.495	.154
		LASSO	$.462_{(.319)}$	$.783_{(.396)}$	$.929_{(.097)}$	$.920_{(.081)}$	1.332	.392	1.139	.297	.504	.297
		MCP	$.593_{(.202)}$	$.912_{(.219)}$	$.981_{(.052)}$	$.979_{(.053)}$	1.035	.236	.743	.106	.497	.106
		SCAD	$.608_{(.194)}$	$.902_{(.234)}$	$.980_{(.062)}$	$.975_{(.054)}$	1.026	.237	.747	.106	.494	.106
	400	Oracle	—	—	—	—	.439	.089	.227	.081	.273	.081
		MLLE	—	—	—	—	.878	.149	.606	.086	.337	.086
		AdpLASSO	$.768_{(.167)}$	$.995_{(.050)}$	$.985_{(.050)}$	$.993_{(.037)}$	.779	.156	.378	.065	.406	.065
		LASSO	$.796_{(.252)}$	$.967_{(.175)}$	$.946_{(.088)}$	$.952_{(.071)}$	.865	.230	.518	.112	.419	.112
		MCP	$.766_{(.162)}$	$.992_{(.064)}$	$.973_{(.073)}$	$.997_{(.019)}$	.764	.141	.372	.060	.406	.060
		SCAD	$.753_{(.156)}$	$.992_{(.064)}$	$.963_{(.089)}$	$.995_{(.025)}$	.764	.141	.373	.061	.408	.061

Web Table 11: Comparison of the methods in terms of tuning parameter and band-width selection time for the adaptive Lasso (AdpLASSO) penalty; $C = 2$;

d	n	Elapsed time	$Pl(h)$ then $BIC(\lambda\|h)$	Simultaneous $BIC_1(h, \lambda)$
10	200	Time in seconds	3.64	65.45
		Ratio	1	18
	400		9.57	171.17
			1	18

Web Table 12: Average (SD) sensitivity, specificity, and estimation errors for a high-dimensional setting.

$C=$	2	Criteria	Sensitivity		Specificity		$L_2(\widehat{\boldsymbol{\beta}}_j)$		$L_2(\widehat{\boldsymbol{\sigma}}_j)$		$L_2(\widehat{\boldsymbol{\pi}}_j)$	
d	n	Component	1	2	1	2	1	2	1	2	1	2
500	200	Oracle	—	—	—	—	0.62	0.09	0.30	0.08	0.08	0.08
		MLLE	—	—	—	—	1.71	1.69	0.29	0.35	0.25	0.26
		LASSO	$0.50_{(0.31)}$	$0.67_{(0439)}$	$0.99_{(0.00)}$	$0.99_{(0.00)}$	1.69	1.59	0.58	0.44	0.54	0.41
		MCP	$0.58_{(0.23)}$	$0.80_{(0.28)}$	$0.99_{(0.02)}$	$0.99_{(0.02)}$	1.55	1.44	0.32	0.31	0.14	0.15
		SCAD	$0.55_{(0.23)}$	$0.76_{(0.31)}$	$0.99_{(0.02)}$	$0.99_{(0.02)}$	1.15	1.41	0.33	0.32	0.16	0.16
	400	Oracle	—	—	—	—	0.35	0.07	0.21	0.07	0.25	0.25
		MLLE	—	—	—	—	2.31	1.94	0.36	0.42	0.36	0.35
		LASSO	$0.54_{(0.26)}$	$0.72_{(0.35)}$	$0.99_{(0.00)}$	$0.99_{(0.00)}$	2.87	2.10	0.34	0.34	0.41	0.41
		MCP	$0.61_{(0.25)}$	$0.81_{(0.25)}$	$0.99_{(0.02)}$	$0.99_{(0.02)}$	2.99	2.20	0.29	0.26	0.40	0.40
		SCAD	$0.59_{(0.25)}$	$0.80_{(0.25)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	3.09	2.18	0.29	0.26	0.40	0.39

The runs using adpLASSO did not converge; thus, no results are provided.

Web Table 13: The top 20 covariates with the highest frequencies of being selected based on Pearson's correlation out of the 100 replications.

	True Covariates				Other Covariates															
x_j	1	2	3	4	74	153	155	5	269	218	268	66	79	329	330	154	118	156	193	103
%	100	100	100	100	99	99	98	96	94	92	92	86	86	80	78	73	70	70	70	62

Web Table 14: The top 20 covariates with the highest frequencies of being selected based on log-likelihood out of the 100 replications.

	True Covariates				Other Covariates															
x_j	1	2	3	4	101	469	352	68	415	118	67	377	74	270	406	414	247	429	5	169
%	100	100	100	100	79	79	78	70	70	65	62	60	56	54	52	52	49	49	48	48