Hypothesis Testing in Finite Mixture of Regressions: Sparsity and Model Selection Uncertainty

Abbas Khalili[1]*, Anand N. Vidyashankar[2]

[1]Department of Mathematics and Statistics, McGill University

[2]Department of Statistics, Volgeneau School of Engineering,

George Mason University

May 8, 2018

Abstract

Sparse finite mixture of regression models arise in several scientific applications and testing hypotheses concerning regression coefficients in such models is fundamental to data analysis. In this paper, we describe an approach for hypothesis testing of regression coefficients that take into account model selection uncertainty. The proposed methods involve (i) estimating the active predictor set of the sparse model using a consistent model selector and (ii) testing hypotheses concerning the regression coefficients associated with the estimated active predictor set. The methods asymptotically control the family wise error rate at a pre-specified nominal level, while accounting for variable selection uncertainty. Additionally, we provide examples of consistent model selectors and describe methods for finite sample improvements. Performance of the methods are also illustrated using simulations. A real data analysis is included to illustrate the applicability of the methods.

KEY WORDS: Adjusted p-value, BIC-enhanced tuning parameter, Data splitting, Family-wise error rate, Model selection consistency.

*Corresponding author.

1 INTRODUCTION

Finite mixture of regression (FMR) models (McLachlan and Peel, 2000) arise in a variety of scientific disciplines and are used to account for hidden subgroups present within a heterogeneous population; for instance, when studying the relationship between a response variable Y and a vector of covariates $\boldsymbol{x} = (x_1, x_2, \dots, x_d)^\top$.

Despite their flexibility and usefulness, these models can be challenging to fit to a given dataset when the number of covariates d is large compared to the sample size. To address this issue, Khalili and Chen (2007) and Städler et al. (2010) introduced sparse modeling frameworks and developed regularization approaches for simultaneous parameter estimation and variable selection in FMR models. However, inferential problems such as hypothesis testing concerning regression coefficients have not received much attention in the literature. The primary purpose of this paper is to address this shortcoming and develop an approach for testing statistical hypotheses accounting for variable selection uncertainty.

For a motivation towards problems investigated in this paper, consider a data set containing a set of covariates potentially associated with a response variable, modeled using a mixture distribution. Assuming a sparse FMR model, one may be interested in (a) identifying the covariates associated with the response within each subgroup, and (b) testing whether their effects are significant within each subgroup or between subgroups.
As an example, in functional genomics several candidate motifs (x_j's) are examined to find a small subset that contributes substantially to variations in gene expression (y). It is known that the set of regulating motifs differ from one subgroup of genes to another (Conlon et al., 2003). Here, it is of interest to evaluate the statistical significance of the selected motifs within/between subgroups of genes. As a second example, in market segmentation research a goal is to identify different groups of consumers to target products and services for each segment separately. The attributes of products and services (x_j's) along with the preferences (y) of consumers can be modeled by a sparse FMR model and the statistical significance of the attributes between and within segments of the market is an important problem for the industry (Wedel and Kamakura, 2000). Beyond the works of Redner and Walker (1984) and

Chen (2016), inferential aspects of FMR models are largely unknown. This paper will provide rigorous statistical methodologies to address such questions.

Turning to sparse FMRs, while sparsification is useful in obtaining parsimonious models, current method for joint estimation and variable selection are fraught with multiple challenges. Specifically, due to the uncertainty inherited from variable selection, one encounters a "random model" when performing hypothesis tests; this must be distinguished from the case when a model is pre-specified, as is typical in classical statistical theory. By a "random model", we mean an FMR model whose active covariate set is chosen using a data-driven method. This randomness needs to be taken into account for further inference and the issue is part of a general post-model selection inference problem (Danilov and Magnus, 2004; Dijkstra and Veldkamp, 1988; Leeb and Pötscher, 2003; Kabaila, 1995).

For testing in sparse FMRs, the hypotheses of interest can only be formulated using an estimated sparse model. Hence, in our approach we split the data into two parts $\mathcal{D}_{1n}$ and $\mathcal{D}_{2n}$, where we use $\mathcal{D}_{1n}$ to select a sparse model via a consistent selector T_n yielding an estimated active predictor set (EAPS), $\widehat{\mathcal{S}}(T_n)$. The idea of data splitting has been used in the statistics literature and in the context of high-dimensional regression by Wasserman and Roeder (2009) who coined the term "Screen and Clean". It is pertinent to notice that our methods *do not involve any screening or cleaning.* Alternative approaches for inference are developed in Lockhart et al. (2014), Zhang and Zhang (2014), Berk et al. (2013), Efron (2014) and Van de Geer et al. (2014).

Tests of hypotheses are performed using $\mathcal{D}_{2n}$ based on student-type statistics. To provide p-values, we establish asymptotic normality of the estimated regression coefficients with indices in the EAPS. To address multiple testing problems encountered, we show that the family-wise error rate (FWER), the probability of rejecting at least one hypothesis when it is true, is asymptotically controlled at a given level α, say. We summarize our contributions:

1. We develop a new hypothesis testing framework for *selected* (random) sparse FMRs which, to the best of our knowledge, is the first work in the field. To address technical challenges due to random number of parameters, we introduce a new dimension

matching technique. The proposed framework also applies to linear and generalized linear regression models.

2. We establish theoretical guarantees concerning the asymptotic control of FWER at a level α. Our simulations show that the empirical FWER is controlled at the nominal 5% level as signal-to-noise ratio varies from strong to low.

3. Our third contribution is in the implementation of the proposed methodology (Algorithm 1) via the EM algorithm. Additionally, we also describe approaches that enhance finite sample performance of the proposed methods (Algorithm 2).

The rest of the paper is organized as follows: Section 2 describes FMR models and the variable selection problem. Section 3 deals with hypothesis testing problems, while Section 4 is devoted to numerical strategies for implementation, simulations and a real data analysis. Section 5 contains a summary and concluding remarks. Regularity conditions and proofs are provided in Appendices A and B, respectively. Additional numerical and theoretical results are given in the Supplementary Material.

2 DEFINITIONS, NOTATIONS AND TERMINOLOGY

Consider a response variable Y whose probability density or mass function is postulated to depend on a potential vector of covariates $\boldsymbol{X} = (X_1, \cdots, X_d)^\top$ with its observed value denoted by $\boldsymbol{x} \in \mathbb{R}^d$.

Definition 1 *A pair $(\boldsymbol{X}, Y)$ is said to follow an* FMR *model of order K if the conditional density (or mass) function of Y given $\boldsymbol{X} = \boldsymbol{x}$ is*

$$f(y; \boldsymbol{x}, \boldsymbol{\Psi}_F) = \sum_{j=1}^{K} \pi_j h(y; \theta_j(\boldsymbol{x}), \phi_j), \tag{1}$$

where $\pi_j > 0$ are mixing probabilities with $\sum_{j=1}^{K} \pi_j = 1$, and $h(\cdot; \theta_j(\boldsymbol{x}), \phi_j)$ belongs to a parametric family of density (or mass) functions, such that $\theta_j(\boldsymbol{x}) = g(\beta_{j0} + \boldsymbol{x}^\top \boldsymbol{\beta}_j)$ for a known link function $g(\cdot)$; β_{j0}, $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \cdots \beta_{jd})^\top$, and ϕ_j are respectively the intercepts, regression coefficients and dispersion parameters. In this paper we assume that K is known.

The vector of parameters in (1) is represented by $\boldsymbol{\Psi}_{\mathrm{F}} = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\beta}_0, \boldsymbol{B})$ with the sub-vectors $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)^\top$, $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_K)^\top$, $\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20}, \ldots, \beta_{K0})^\top$, and regression coefficients $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$. Notice that $\boldsymbol{\Psi}_{\mathrm{F}} \in \boldsymbol{\Theta} \subset \mathbb{R}^{K(d+3)-1}$, where $\boldsymbol{\Theta}$ is the parameter space. We assume throughout the paper that $0 < \pi_1 \leq \pi_2 \leq \ldots \leq \pi_K < 1$.

Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ be a random sample of observations from the FMR model (1). The log-likelihood function of $\boldsymbol{\Psi}_{\mathrm{F}}$ is given by

$$\ell_n(\boldsymbol{\Psi}_{\mathrm{F}}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{K} \pi_j h(y_i; \theta_j(\boldsymbol{x}_i), \phi_j) \right]. \tag{2}$$

When $h(\cdot; \theta_j(\boldsymbol{x}_i), \phi_j)$ is Gaussian and the component variances ϕ_j's are different, (2) becomes unbounded when some ϕ_j tends to zero. In practice, this can be avoided by introducing a positive lower bound on the smallest ratios of variances as in Hathaway (1985) or by adding a penalty on ϕ_j as in Chen et al. (2008).

When the number of regression coefficients is large compared to n, the maximum likelihood estimator (MLE) of $\boldsymbol{\Psi}_{\mathrm{F}}$ can be unstable and have large variance. It is then preferable to fit *sparse* FMR models, which involves the concepts of *active* and *inactive* predictor sets.

Definition 2 **(i)** *For each $1 \leq j \leq K$, the inactive and active predictor sets of the j^{th} component of an* FMR *are $N_j = \bigcup_{l=1}^{d}\{(j, l) : \beta_{jl} = 0\}$ and $S_j = \bigcup_{r=1}^{d}\{(j, r) : \beta_{jr} \neq 0\}$ respectively. Let $q_j = |S_j|$ be the active number of regression coefficients in the j^{th} component.* **(ii)** *The inactive and active predictor sets of an* FMR *are given by $\mathcal{N} = \bigcup_{j=1}^{K} N_j$ and $\mathcal{S} = \bigcup_{j=1}^{K} S_j$ respectively. The active number of regression coefficients is $q = |\mathcal{S}| = \sum_{j=1}^{K} q_j$.*

An FMR is said to be *sparse* if $q < Kd$, where Kd is the total number of regression coefficients of the full model. We refer to any such sparse model as a candidate FMR sub-model and denote it by $\mathcal{M}_C$. Also, we represent the full model by $\mathcal{M}_F$ with the corresponding inactive and active predictor sets $\mathcal{N}_F$ and $\mathcal{S}_F$ such that $\mathcal{N}_F = \emptyset$ and $\mathcal{S}_F = \bigcup_{j=1}^{K}\{(j, 1), (j, 2), \ldots, (j, d)\}$. The collection of all candidate sub-models is denoted by

$$\mathcal{A} = \left\{ \mathcal{M}_C : \mathcal{S} = \bigcup_{j=1}^{K} S_j \ , \ S_j \subseteq \{(j, 1), (j, 2), \ldots, (j, d)\}, 1 \leq j \leq K \right\}. \tag{3}$$

We note that $\mathcal{A}$ includes $\mathcal{M}_F$. It is clear that the active set of any sub-model is a subset of that of the full model, and with the abuse of notation we also write $\mathcal{M}_C \subseteq \mathcal{M}_F$ when referring to a sub-model. We denote the size of the active set of $\mathcal{M}_C$ by $q(\mathcal{M}_C)$.

Analogous to Definition 1, we denote by $\boldsymbol{\Psi}_C$ the parameter vector corresponding to sub-model $\mathcal{M}_C$. More specifically, we may express the regression vector of the j^{th} component of $\mathcal{M}_C$ as $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j1}, \boldsymbol{\beta}_{j2})$, for $1 \leq j \leq K$, where without loss of generality $\boldsymbol{\beta}_{j1} = (\beta_{j1}, \ldots, \beta_{j,q_j})^\top$ is a vector of non-zero regression coefficients whose indices belong to the active set S_j, and $\boldsymbol{\beta}_{j2} = (\beta_{j,q_j+1}, \ldots, \beta_{jd})^\top$ is a vector of zero coefficients, corresponding to the indices in the inactive set N_j. We notice that the active predictor set and consequently the above partitioning for each vector $\boldsymbol{\beta}_j$ may be different across the mixture components. Now, by setting $\boldsymbol{B}_1 = (\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{21}, \cdots, \boldsymbol{\beta}_{K1})$ and $\boldsymbol{B}_2 = (\boldsymbol{\beta}_{12}, \boldsymbol{\beta}_{22}, \cdots, \boldsymbol{\beta}_{K2})$, an FMR sub-model is characterized by the parameter vector $\boldsymbol{\Psi}_C = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\beta}_0, \boldsymbol{B}_1)$. Note that $\boldsymbol{B}_2 = \mathbf{0}$ and $\dim(\boldsymbol{B}_2) = Kd - q(\mathcal{M}_C)$.

In this paper, we denote the true sparse FMR sub-model by $\mathcal{M}_0 \in \mathcal{A}$, and the corresponding parameter vector is given by $\boldsymbol{\Psi}_0 = (\boldsymbol{\pi}_0, \boldsymbol{\phi}_0, \boldsymbol{\beta}_0^0, \boldsymbol{B}_{10})$, whose dimension equals $\kappa_0 = q_0 + 3K - 1$, where $q_0 \equiv q(\mathcal{M}_0) = \sum_{j=1}^K q_j^0$ and $\dim(\boldsymbol{\beta}_{j1}^0) = q_j^0$. We denote by N_j^0 and S_j^0, respectively, the inactive and active predictor sets of the j^{th} component of the model $\mathcal{M}_0$. Consequently, $\mathcal{N}_0 = \bigcup_{j=1}^K N_j^0$ and $\mathcal{S}_0 = \bigcup_{j=1}^K S_j^0$ are respectively the inactive and active predictor sets of $\mathcal{M}_0$. In the rest of the paper, we use $P(\cdot)$ to denote the probability distribution associated with $\mathcal{M}_0$, which has a probability density or mass function (from now on referred to as pdf)

$$f(y; \boldsymbol{x}, \boldsymbol{\Psi}_0) = \sum_{j=1}^K \pi_j^0 h(y; \theta_j^0(\boldsymbol{x}), \phi_j^0), \tag{4}$$

where $\theta_j^0(\boldsymbol{x}) = g(\beta_{j0}^0 + \sum_{(j,l)\in S_j^0} x_l \beta_{jl}^0)$. As explained in the Introduction, estimation of the active predictor set $\mathcal{S}_0$ (i.e., variable selection) is the first step towards formulating hypotheses concerning the regression coefficients of $\mathcal{M}_0$ which is described in the next section together with our testing procedure.

3 HYPOTHESIS TESTING ACCOUNTING FOR MODEL SELECTION UNCERTAINTY

In this section, we study hypothesis testing in sparse FMR models. However, since the true active predictor set is unknown, formulation of hypotheses concerning the regression coefficients is unclear. Specifically, consider the problem of testing if the regression coefficients corresponding to $\mathcal{S}_0$ are zero. Assigning p-values to such tests is a formidable problem and mostly ad hoc solutions have been provided. It is natural, as in Meinshausen et al. (2009), to assign p-values for all the variables under study. However, rigorous statistical justification of such an approach raises fundamental questions about the meaning of the underlying true model. In this section, we address this issue and develop an approach to test various hypotheses of interest. Our methodology involves a *dimension matching technique* and combines it with a data splitting idea which facilitates hypotheses formulation. Incidentally, it also accounts for variable selection uncertainty.

3.1 Data splitting method

As a first step, we divide the data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ randomly into two parts, $\mathcal{D}_{1n}$ and $\mathcal{D}_{2n}$, of approximately equal size $\frac{n}{2}$. We may refer to $\mathcal{D}_{1n}$ and $\mathcal{D}_{2n}$ respectively as the training and testing data, even though there are subtle differences in the terminology as used in machine learning literature. Using $\mathcal{D}_{1n}$ and a model selection method, we estimate the active predictor set and set-up hypotheses of interest. We then use $\mathcal{D}_{2n}$ to perform tests.

While the above approach seems natural and plausible, several subtle issues arise. First, we seek a consistent model selection mechanism; that is, as $n \to \infty$, the selected model estimates the true model $\mathcal{M}_0$ with probability approaching one. However, for an estimated model, the dimension of the parameter vector is random. Hence, direct comparison of the estimates of the parameter vector of a selected model to that of the "true model" is not feasible. To address this issue, we introduce a dimension matching technique.

Definition 3 *A mapping T_n from the sample space $\mathcal{X}$ into $\mathcal{A}$ is a consistent* selector *if*

$$\lim_{n\to\infty} P(T_n = \mathcal{M}_0) = 1.$$

Examples of such T_n are provided in Section 3.2. We apply T_n to $\mathcal{D}_{1n}$ and obtain an FMR sub-model with estimated active predictor set (EAPS) $\widehat{\mathcal{S}}(T_n) = \bigcup_{j=1}^{n} \widehat{S}_j(T_n)$, where $\widehat{S}_j(T_n)$ represents the set of indices selected by T_n in the j^{th} mixture component. The FMR sub-model associated with this EAPS is given by

$$f(y; \boldsymbol{x}, \boldsymbol{\Psi}(\widehat{\mathcal{S}}(T_n))) = \sum_{j=1}^{K} \pi_j h(y; \tilde{\theta}_j(\boldsymbol{x}), \phi_j), \tag{5}$$

where $\boldsymbol{\Psi}(\widehat{\mathcal{S}}(T_n))$ is a sub-vector of $\boldsymbol{\Psi}_{\text{F}}$, and $\tilde{\theta}_j(\boldsymbol{x}) = g(\beta_{j0} + \sum_{(j,l)\in\widehat{S}_j(T_n)} x_l \beta_{jl})$. In the following, we use $\widetilde{\boldsymbol{\Psi}}$ to denote $\boldsymbol{\Psi}(\widehat{\mathcal{S}}(T_n))$. We focus on the following sets of hypotheses:

(1) For all $1 \le j \le K$ and $l \in \widehat{S}_j(T_n)$, consider testing

$$H_{0,jl} : \beta_{jl} = 0. \tag{6}$$

(2) For any fixed $1 \le j \le K$, let $\mathcal{G}_j \subseteq \widehat{S}_j(T_n)$. For all $l \in \mathcal{G}_j$, consider testing

$$H_{0,jl} : \beta_{jl} = 0. \tag{7}$$

(3) Let $\mathcal{G} = \cup_{j=1}^{K} \mathcal{G}_j$, where $\mathcal{G}_j \subseteq \widehat{S}_j(T_n)$. For all $(j,l) \in \mathcal{G}$ we may also test

$$H_{0,jl} : \beta_{jl} = 0. \tag{8}$$

Next, we use $\mathcal{D}_{2n}$ to fit (5) using the maximum likelihood method. The MLE of $\widetilde{\boldsymbol{\Psi}}$, denoted by $\bar{\bar{\widetilde{\boldsymbol{\Psi}}}}_n$, is obtained by maximizing

$$\ell_n(\widetilde{\boldsymbol{\Psi}}) = \sum_{i\in\mathcal{D}_{2n}} \log \left[\sum_{j=1}^{K} \pi_j h(y_i; \tilde{\theta}_j(\boldsymbol{x}_i), \phi_j) \right].$$

Turning to (6), we consider the student-type statistic

$$t_{jl,n} = \bar{\bar{\tilde{\beta}}}_{jl} / \text{SE}(\bar{\bar{\tilde{\beta}}}_{jl}), \tag{9}$$

where $\text{SE}(\bar{\bar{\tilde{\beta}}}_{jl})$ is obtained from the observed “information matrix”. Details on the computation of $\text{SE}(\bar{\bar{\tilde{\beta}}}_{jl})$ are provided in Section 4.1. Note that $t_{jl,n}$ also depends on $\mathcal{D}_{1n}$, since $(j,l) \in \widehat{\mathcal{S}}_j(T_n)$.

We establish in Theorem 1 that the asymptotic distribution of $\overline{\widetilde{\mathbf{\Psi}}}_n$ can be approximated by a Normal distribution with appropriate mean and covariance matrix; from this it follows that, for small sample sizes, the distribution of (9) can be approximated by a t-distribution with $\frac{n}{2} - \hat{q}_n - (3K-1)$ degrees of freedom, where $\hat{q}_n = |\widehat{\mathcal{S}}(T_n)|$. Also, we account for multiple comparisons using a Bonferroni-type adjustment.

Given $\mathcal{D}_{1n}$, let p_{jl} be the p-value associated with the test in (6) which is of size $\alpha/\hat{q}_n$, for some $\alpha \in (0,1)$. The size of the above test is random, but conditioned on $\mathcal{D}_{1n}$ it is deterministic; additionally, conditioned on $\mathcal{D}_{1n}$, $t_{jl,n}$ is independent of $\hat{q}_n$. Define

$$\mathcal{S}_n^*(\mathcal{D}_{1n}, \mathcal{D}_{2n}) = \bigcup_{(j,l)\in\widehat{\mathcal{S}}(T_n)} \{(j,l) : p_{jl} \leq \alpha/\hat{q}_n\}$$

to be the set of all indices $(j,l) \in \widehat{\mathcal{S}}(T_n)$ for which the hypothesis $H_{0,jl}$ is rejected. Theorem 1 establishes that for testing (6) the FWER is asymptotically controlled at level α. Let

$$\mathcal{E}(\mathcal{D}_{1n}, \mathcal{D}_{2n}) = \mathcal{N}_0 \bigcap \mathcal{S}_n^*(\mathcal{D}_{1n}, \mathcal{D}_{2n}) \tag{10}$$

denote the set of indices of regression coefficients of the selected covariates whose corresponding null hypothesis $\mathrm{H}_{0,jl}$ is rejected when it is true.

Theoretical Justification:

To study the asymptotic properties of $\overline{\widetilde{\mathbf{\Psi}}}_n$ with respect to the true sparse FMR model in (4), we need to compare it to $\mathbf{\Psi}_0$ which is of dimension $\kappa_0 = q_0 + 3K - 1$ (potentially different from $\dim(\overline{\widetilde{\mathbf{\Psi}}}_n) \equiv \hat{\kappa}_n = \hat{q}_n + 3K - 1$). To address this issue, suppose $(j,l) \in \widehat{\mathcal{S}}(T_n) \bigcap \mathcal{N}_0$, then set $\beta_{jl}^0 = 0$; otherwise, if $(j,l) \in \widehat{\mathcal{S}}(T_n) \bigcap \mathcal{S}_0$, then the true value is β_{jl}^0. Thus for every $(j,l) \in \widehat{\mathcal{S}}(T_n)$, the true value of β_{jl} is defined. This yields a new regression coefficients vector $\boldsymbol{B}_{10}(\hat{q}_n)$ which we refer to as the dimension-adjusted true regression coefficients vector. Now, we denote the new dimension-adjusted true parameter vector by $\mathbf{\Psi}_0(\hat{q}_n) = (\boldsymbol{\pi}_0, \boldsymbol{\phi}_0, \boldsymbol{\beta}_0^0, \boldsymbol{B}_{10}(\hat{q}_n))$.

Denote the joint pdf of $\boldsymbol{Z} = (\boldsymbol{X}, Y)$ by $f^*(\cdot;\cdot)$ assumed to satisfy the regularity conditions (RC1)-(RC5) in Appendix A; additionally let

$$\boldsymbol{I}_1(\mathbf{\Psi}_C) = E\left\{ \left[\frac{\partial}{\partial \mathbf{\Psi}_C} \log f^*(\boldsymbol{Z}; \mathbf{\Psi}_C)\right] \left[\frac{\partial}{\partial \mathbf{\Psi}_C} \log f^*(\boldsymbol{Z}; \mathbf{\Psi}_C)\right]^T \right\}. \tag{11}$$

Assume $\mathcal{W}_n$ and $\mathcal{W}$ are constant matrices of dimensions $m \times \hat{\kappa}_n$ and $m \times \kappa_0$ (for some fixed $m \geq 1$) respectively and satisfying, as $n \to \infty$,

$$\mathcal{W}_n[\boldsymbol{I}_1(\boldsymbol{\Psi}_0(\hat{q}_n))]\mathcal{W}_n^\top \xrightarrow{p} \mathcal{W}[\boldsymbol{I}_1\left(\boldsymbol{\Psi}_0\right)]\mathcal{W}^\top, \tag{12}$$

where $\boldsymbol{I}_1(\cdot)$ is the Fisher information matrix defined in (11) with $\boldsymbol{\Psi}_C$ replaced by $\boldsymbol{\Psi}_0$ or $\boldsymbol{\Psi}_0(\hat{q}_n)$. The validity of (12) is guaranteed by the consistency property of the model selector T_n. It is worth noticing here that $\boldsymbol{\Psi}_0(\hat{q}_n)$ corresponds to the true value of $\widetilde{\boldsymbol{\Psi}}$. Furthermore, if $\hat{q}_n = q_0$ it follows that $\mathcal{W}_n = \mathcal{W}$. The matrices $\mathcal{W}_n$ and $\mathcal{W}$ facilitate convergence (as $n \to \infty$) of elements of the information matrix which is of random dimension for any fixed n.

We now state our results concerning the asymptotic distribution of $\overline{\widetilde{\boldsymbol{\Psi}}}_n$ and the asymptotic control of FWER.

Theorem 1 *Let T_n be a consistent model selector and $\alpha \in (0, 1)$ be a nominal significance level. Under the regularity conditions (RC1)-(RC5) in Appendix A, the following hold:*

(i) asymptotic normality*:*

$$\sqrt{\frac{n}{2}}\left\{\mathcal{W}_n\left(\overline{\widetilde{\boldsymbol{\Psi}}}_n - \boldsymbol{\Psi}_0(\hat{q}_n)\right)\right\} \xrightarrow{d} \mathcal{N}_m\left(0, [\mathcal{W}\boldsymbol{I}_1\left(\boldsymbol{\Psi}_0\right)\mathcal{W}^\top]^{-1}\right) \quad \textit{as } n \to \infty;$$

(ii) FWER **control***:*

$$\limsup_{n\to\infty} P\Bigg(\mathcal{E}(\mathcal{D}_{1n}, \mathcal{D}_{2n}) \neq \emptyset\Bigg) \leq \alpha.$$

We now provide two examples of the selectors T_n that allow the student-type inference described above.

3.2 Examples of Consistent Model Selectors

It is folklore amongst statisticians that parsimonious models yield stable inference. In the context of regression problems, parsimonious modelling amounts to identifying a set of active predictors that influence the response variable Y of interest. Here we outline two approaches that yield consistent model selectors as described in Definition 3.

1. BIC-based selector: One of the well-known methods for variable selection in regression consists in choosing that sub-model for which an information criterion is minimized. In the context of our problem, for any candidate model $\mathcal{M}_C \in \mathcal{A}$, the criterion based on BIC is

$$\text{BIC}(\mathcal{M}_C) = -2\ell_n(\overline{\boldsymbol{\Psi}}_{n,C}) + q(\mathcal{M}_C)\log n, \tag{13}$$

where $\ell_n(\cdot)$ is the log-likelihood associated with the sub-model $\mathcal{M}_C$, $\overline{\boldsymbol{\Psi}}_{n,C}$ is the MLE of $\boldsymbol{\Psi}_C$, and $q(\mathcal{M}_C)$ is the size of the active predictor set of $\mathcal{M}_C$. Let

$$T_n = \operatorname*{argmin}_{\mathcal{M}_C \in \mathcal{A}} \text{BIC}(\mathcal{M}_C). \tag{14}$$

In the proposed method, we apply T_n to the data $\mathcal{D}_{1n}$ (with $\log n$ replaced by $\log(n/2)$) yielding an FMR sub-model with EAPS $\widehat{\mathcal{S}}(T_n)$. It is well-known that, under the regularity conditions (RC1)-(RC5), T_n is a consistent model selector (Konishi and Kitagawa, 2008). In practice, BIC requires searching through 2^{Kd} FMR sub-models in the model space $\mathcal{A}$, which is computationally manageable for moderate values of (K, d). However, the BIC is not numerically feasible for large model spaces $\mathcal{A}$ and thus alternative methods are required.

2. Adaptive regularization-based selector: This approach involves maximization of the penalized log-likelihood function

$$p\ell_n(\boldsymbol{\Psi}_\text{F}; \boldsymbol{\gamma}) = \ell_n(\boldsymbol{\Psi}_\text{F}) - \boldsymbol{p}_n(\boldsymbol{\Psi}_\text{F}; \boldsymbol{\gamma}),$$

where $\ell_n(\boldsymbol{\Psi}_\text{F})$ is the log-likelihood in (2), and the penalty function is

$$\boldsymbol{p}_n(\boldsymbol{\Psi}_\text{F}; \boldsymbol{\gamma}) = \sum_{j=1}^{K} \pi_j \sum_{l=1}^{d} p_n(\beta_{jl}; \gamma_j), \tag{15}$$

where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)^\top$. Examples of $p_n(\beta_{jl}; \gamma_j)$ are adaptive LASSO (ADLASSO), SCAD, and MCP. Given $\boldsymbol{\gamma}$, the maximum penalized likelihood estimator (MPLE) of $\boldsymbol{\Psi}_\text{F}$ is defined to be

$$\widehat{\boldsymbol{\Psi}}_{\text{F},n}(\boldsymbol{\gamma}) \equiv \widehat{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma}) = \operatorname*{argmax}_{\boldsymbol{\Psi}_\text{F} \in \boldsymbol{\Theta}} p\ell_n(\boldsymbol{\Psi}_\text{F}; \boldsymbol{\gamma}). \tag{16}$$

By the properties of the penalty in (15) and tuning $\boldsymbol{\gamma}$, one can encourage estimates of some regression coefficients to be zero. Hence, using (16) we obtain an FMR sub-model with EAPS

$$\widehat{\mathcal{S}}(\boldsymbol{\gamma}) = \bigcup_{j=1}^{K} \widehat{S}_j(\boldsymbol{\gamma}),$$

where $\widehat{S}_j(\boldsymbol{\gamma}) = \bigcup_{r=1}^{d}\{(j,r) : \hat{\beta}_{jr}(\boldsymbol{\gamma}) \neq 0\}$; and the corresponding sub-model is $\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{A}$.

Since the candidate sub-models are indexed by $\boldsymbol{\gamma}$, the variable selection problem involves identifying an appropriate value for it. To this end, we adopt a data adaptive strategy which is analogous to the BIC approach described in Example 1 above; see also Zhang et al. (2010). Specifically, we choose the tuning parameter $\widehat{\boldsymbol{\gamma}}_n$ as

$$\widehat{\boldsymbol{\gamma}}_n = \underset{\boldsymbol{\gamma}\in[0,\gamma_n^*]^K}{\operatorname{argmin}} \ \text{BIC}(\boldsymbol{\gamma}), \tag{17}$$

$$\text{BIC}(\boldsymbol{\gamma}) = -2\ell_n(\widehat{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma})) + q(\boldsymbol{\gamma})\log n. \tag{18}$$

The role of γ_n^* is to ensure that the selected $\widehat{\boldsymbol{\gamma}}_n$ satisfies conditions (RCP1)-(RCP3) in Appendix A, in probability. Notice that $q(\boldsymbol{\gamma}) \equiv q(\mathcal{M}_{\boldsymbol{\gamma}}) = \sum_{j=1}^{K}\sum_{l=1}^{d} I(\hat{\beta}_{jl}(\boldsymbol{\gamma}) \neq 0)$. Let $\mathcal{M}_{\widehat{\boldsymbol{\gamma}}_n}$ be the corresponding sub-model with EAPS $\widehat{\mathcal{S}}(\widehat{\boldsymbol{\gamma}}_n)$. We refer to $\widehat{\boldsymbol{\gamma}}_n$ as the BIC-enhanced tuning parameter. Our next result establishes that the selector $T_n = \mathcal{M}_{\widehat{\boldsymbol{\gamma}}_n}$ is consistent.

Proposition 1 *Under the regularity conditions in Appendix A,* $\lim_{n\to\infty} P(\mathcal{M}_{\widehat{\boldsymbol{\gamma}}_n} = \mathcal{M}_0) = 1$.

Proposition 1 implies that if some rough properties of $\boldsymbol{\gamma}_n = (\gamma_{n1}, \gamma_{n2}, \ldots, \gamma_{nK})$ are known, such as those indicated by (RCP1)-(RCP4), then they can be input into (17) to find the "optimal" (BIC-enhanced) sequence, which in turn yields a consistent model selector.

Returning to our proposed method, we apply the above model selector to the data $\mathcal{D}_{1n}$ with $\log n$ replaced by $\log(n/2)$ in (18). We now describe Algorithm 1 that summarizes the steps for hypothesis testing using either of the selectors T_n described above.

Algorithm 1

Step 1: Divide the data randomly into $(\mathcal{D}_{1n}, \mathcal{D}_{2n})$ of approximately equal size $n/2$.

Step 2: Using $\mathcal{D}_{1n}$ and a consistent mode selector T_n, obtain the EAPS $\widehat{\mathcal{S}}(T_n)$.

Step 3: Using $\mathcal{D}_{2n}$, obtain the MLE $\overline{\widetilde{\boldsymbol{\Psi}}}_n$ of the parameter $\widetilde{\boldsymbol{\Psi}}$ of the selected FMR sub-model corresponding to $\widehat{\mathcal{S}}(T_n)$ in **Step 2**.

Step 4: Perform hypothesis testing using student-type statistics for the regression coefficients of the estimated sparse FMR model using $\mathcal{D}_{2n}$.

What happens if we do not account for model selection uncertainty?

It is a legitimate to wonder how the inference about the regression coefficients would be affected if we do not account for model selection uncertainty when using (16). To this end, we partition $\widehat{\boldsymbol{\Psi}}_n(\widehat{\boldsymbol{\gamma}}_n)$ in (16) into $(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n), \widehat{\boldsymbol{\Psi}}_{2,n}(\widehat{\boldsymbol{\gamma}}_n))$ such that $\dim(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n)) = \dim(\boldsymbol{\Psi}_0)$. This partitioning is based on the oracle's perspective. By Proposition 1, $\widehat{\boldsymbol{\Psi}}_{2,n}(\widehat{\boldsymbol{\gamma}}_n) = \mathbf{0}$ with probability tending to one as $n \to \infty$. As for the limit distribution of $\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n)$, we have

Theorem 2 *Assume that the conditions of Proposition 1 hold. Then, as $n \to \infty$, we have*

$$\sqrt{n}\left\{\left[\boldsymbol{I}_1\left(\boldsymbol{\Psi}_0\right) - \frac{\boldsymbol{p}_n''(\boldsymbol{\Psi}_0;\widehat{\boldsymbol{\gamma}}_n)}{n}\right]\left(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n) - \boldsymbol{\Psi}_0\right) + \frac{\boldsymbol{p}_n'(\boldsymbol{\Psi}_0;\widehat{\boldsymbol{\gamma}}_n)}{n}\right\} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_1\left(\boldsymbol{\Psi}_0\right)\right).$$

Theorem 2 may suggest a hypothesis testing procedure for the regression coefficients of the selected model, as is typically carried out in conventional data analysis. However, such a procedure will not account for model selection uncertainty since Theorem 2 is obtained from an oracle's perspective. A standard approach is to use an estimate of the model from the data (called selected model); this then leads to variability due to the model selection that needs to be taken into account in data analysis. We address this issue in Theorem 1 using data splitting and dimension matching techniques. Returning to Theorem 2, we observe that the estimator involves a bias term and a normalization factor which are explicit functions of the first and second derivatives of the penalty. While the effects of some penalties vanish asymptotically, they persist in finite samples which is confirmed in our simulations (Table S4 and Table S5 of the Supplementary Material). Also, due to the heterogeneity structure of the model, we observe accumulation of the false positives (see Section 4.3) across the mixture components which in turn results in high false positive in the overall model.

4 NUMERICAL STRATEGIES AND DATA ANALYSES

In this section, we provide computational strategies for implementing Algorithm 1. In Section 4.1, we focus on the EM algorithm for identifying the EAPS and the MLEs of the selected model. We then use these estimates to obtain the p-values described in Section 3.1. Additionally, in this section we describe methods for finite sample improvements.

4.1 Implementation via EM algorithm

Step 2 of Algorithm 1 involves identifying the EAPS. Depending on the choice of the selector T_n, the implementation of EM algorithm varies. We start with the BIC-based selector.

Consider the log-likelihood function for a candidate sub-model $\mathcal{M}_C \in \mathcal{A}$ with the corresponding parameter vector $\boldsymbol{\Psi}_C$, using $\mathcal{D}_{1n}$. To obtain the MLE of $\boldsymbol{\Psi}_C$, we apply the EM algorithm as follows. The complete data log-likelihood function is given by

$$\ell_n^c(\boldsymbol{\Psi}_{\mathrm{C}}) = \sum_{j=1}^{K} \sum_{i \in \mathcal{D}_{1n}} Z_{ij} \left\{ \log \pi_k + \log h(y_i; \theta_j(\boldsymbol{x}_i), \phi_j) \right\}, \tag{19}$$

where $\theta_j(\boldsymbol{x}) = g(\beta_{j0} + \sum_{(j,l) \in S_j} x_l \beta_{jl})$, and Z_{ij} are latent indicator variables representing mixture-component membership of observations in $\mathcal{D}_{1n}$. At the $(m+1)^{th}$ iteration of the algorithm, the expected value of the complete log-likelihood is given by

$$Q_1(\boldsymbol{\Psi}_C; \boldsymbol{\Psi}_C^{(m)}) = \sum_{j=1}^{K} \sum_{i \in \mathcal{D}_{1n}} \omega_{ij}^{(m)} \left\{ \log \pi_k + \log h(y_i; \theta_j(\boldsymbol{x}_i), \phi_j) \right\},$$

where $\omega_{ij}^{(m)} = E\{Z_{ij} | (Y_i, \boldsymbol{x}_i) \in \mathcal{D}_{1n}, \boldsymbol{\Psi}_C^{(m)}\}$. The parameter estimates are updated as follows:

E-step: In this step compute the function $Q_1(\cdot)$ or equivalently the weights

$$\omega_{ij}^{(m)} = \frac{\pi_j^{(m)} h(y_i; \theta_j^{(m)}(\boldsymbol{x}_i), \phi_j^{(m)})}{\sum_{k=1}^{K} \pi_k^{(m)} h(y_i; \theta_k^{(m)}(\boldsymbol{x}_i), \phi_k^{(m)})} \quad ; \quad j = 1, 2, \ldots, K, i \in \mathcal{D}_{1n}.$$

M-step: Here, obtain the updated parameter estimates by

$$\boldsymbol{\Psi}_C^{(m+1)} = \underset{\boldsymbol{\Psi}_C}{\operatorname{argmax}}\, Q_1(\boldsymbol{\Psi}_C; \boldsymbol{\Psi}_C^{(m)}). \tag{20}$$

The two steps are repeated until convergence, yielding a numerical approximation to the MLE $\overline{\boldsymbol{\Psi}}_{n,C}$. We repeat the EM steps for every candidate model $\mathcal{M}_C \in \mathcal{A}$ and compute the BIC in (13) using $\overline{\boldsymbol{\Psi}}_{n,C}$. Finally, the EAPS is $\widehat{\mathcal{S}}(T_n)$, where T_n is given in (14).

While the above algorithm is useful if the model space $\mathcal{A}$ is not large, one could alternatively use the computationally more efficient adaptive regularization technique which we now describe. Here, we consider the penalized log-likelihood

$$p\ell_n(\boldsymbol{\Psi}_{\mathrm{F}}; \boldsymbol{\gamma}) = \sum_{i \in \mathcal{D}_{1n}} \log \left[\sum_{j=1}^{K} \pi_j h(y_i; \theta_j(\boldsymbol{x}_i), \phi_j) \right] - \boldsymbol{p}_n(\boldsymbol{\Psi}_{\mathrm{F}}; \boldsymbol{\gamma}),$$

where $\theta_j(\boldsymbol{x}) = g(\beta_{j0} + \boldsymbol{x}^\top \boldsymbol{\beta}_j)$, and the penalty is given in (15). As before, by the EM principles, the complete penalized log-likelihood function is

$$p\ell_n^c(\boldsymbol{\Psi}_{\mathrm{F}}; \boldsymbol{\gamma}) = \sum_{j=1}^{K} \sum_{i \in \mathcal{D}_{1n}} Z_{ij} \left\{ \log \pi_k + \log h(y_i; \theta_j(\boldsymbol{x}_i), \phi_j) \right\} - \boldsymbol{p}_n(\boldsymbol{\Psi}_{\mathrm{F}}; \boldsymbol{\gamma}).$$

For a fixed $\boldsymbol{\gamma}$, at the $(m+1)^{th}$ iteration of the algorithm, the conditional expected value of the complete penalized log-likelihood is given by

$$pQ(\boldsymbol{\Psi}_{\mathrm{F}}; \boldsymbol{\Psi}_{\mathrm{F}}^{(m)}, \boldsymbol{\gamma}) \ = \ \sum_{j=1}^{K} \sum_{i \in \mathcal{D}_{1n}} \tau_{ij}^{(m)} \left\{ \log \pi_k + \log h(y_i; \theta_j(\boldsymbol{x}_i), \phi_j) \right\} - \boldsymbol{p}_n(\boldsymbol{\Psi}_{\mathrm{F}}; \boldsymbol{\gamma}),$$

where $\tau_{ij}^{(m)} = E\{Z_{ij} | (Y_i, \boldsymbol{x}_i) \in \mathcal{D}_{1n}, \boldsymbol{\Psi}_{\mathrm{F}}^{(m)}\}$. The parameter estimates are updated as follows:
E-step: In this step compute the function $pQ(\cdot)$ or equivalently the weights

$$\tau_{ij}^{(m)} = \frac{\pi_j^{(m)} h(y_i; \theta_j^{(m)}(\boldsymbol{x}_i), \phi_j^{(m)})}{\sum_{k=1}^{K} \pi_k^{(m)} h(y_i; \theta_k^{(m)}(\boldsymbol{x}_i), \phi_k^{(m)})} \quad ; \quad j = 1, 2, \ldots, K, i \in \mathcal{D}_{1n}.$$

M-step: Here, obtain the updated parameter estimates by

$$\boldsymbol{\Psi}_{\mathrm{F}}^{(m+1)} = \underset{\boldsymbol{\Psi}_{\mathrm{F}}}{\operatorname{argmax}}\, pQ(\boldsymbol{\Psi}_{\mathrm{F}}; \boldsymbol{\Psi}_{\mathrm{F}}^{(m)}, \boldsymbol{\gamma}). \tag{21}$$

The two steps are repeated until convergence, yielding a numerical approximation to the MPLE $\widehat{\boldsymbol{\Psi}}_{\mathrm{F},n}(\boldsymbol{\gamma}) \equiv \widehat{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma})$ in (16). To solve the optimization problem in (21) (and also in (20)), depending on the form of $h(y; \theta_k(\boldsymbol{x}_i), \phi_k)$, one may use Newton-like methods to approximate the leading terms by quadratic functions of $(\beta_{j0}, \phi_j, \beta_{jl})$ and then perform the optimization.

Now to identify the EAPS, we obtain (using (17)) the BIC-enhanced tuning parameter by applying the above algorithm to a sequence of tuning parameter values, say, $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_M$. This yields the EAPS $\widehat{\mathcal{S}}(\widehat{\boldsymbol{\gamma}}_n)$, which is indeed $\widehat{\mathcal{S}}(T_n)$.

In summary, having applied either of the above EM algorithms we arrive at a sub-model with EAPS $\widehat{\mathcal{S}}(T_n)$, and the parameter vector $\widetilde{\boldsymbol{\Psi}}$. This completes the **Step 2** of Algorithm 1.

For **Step 3**, to obtain the MLE of $\widetilde{\boldsymbol{\Psi}}$ we use the EM algorithm based on the complete log-likelihood function (19) with $(\mathcal{D}_{1n}, \boldsymbol{\Psi}_C)$ replaced by $(\mathcal{D}_{2n}, \widetilde{\boldsymbol{\Psi}})$. This yields the MLE $\overline{\widetilde{\boldsymbol{\Psi}}}_n$.

Finally, for **Step 4**, we describe the calculation of the standard errors for computing the t-ratios $t_{jl,n}$ in (9). We use the empirical observed information matrix to approximate the observed information matrix (McLachlan and Peel, 2000). For FMRs this matrix is given by

$$\boldsymbol{I}_e(\overline{\widetilde{\boldsymbol{\Psi}}}_n) = \sum_{i\in\mathcal{D}_{2n}} \boldsymbol{s}(y_i, \boldsymbol{x}_i; \overline{\widetilde{\boldsymbol{\Psi}}}_n)\boldsymbol{s}^\top(y_i, \boldsymbol{x}_i; \overline{\widetilde{\boldsymbol{\Psi}}}_n), \tag{22}$$

where $\boldsymbol{s}(\cdot)$ is the gradient of the complete log-likelihood and

$$\boldsymbol{s}(y_i, \boldsymbol{x}_i; \widetilde{\boldsymbol{\Psi}}) = \sum_{j=1}^{K} \omega_{ij} \frac{\partial}{\partial \widetilde{\boldsymbol{\Psi}}} \left\{ \log \pi_k + \log h(y_i; \tilde{\theta}_j(\boldsymbol{x}_i), \phi_j) \right\} \quad , \quad i \in \mathcal{D}_{2n}$$

and $\omega_{ij} = E\{Z_{ij} | (Y_i, \boldsymbol{x}_i) \in \mathcal{D}_{2n}, \overline{\widetilde{\boldsymbol{\Psi}}}_n\}$. Square root of the diagonal elements of the inverse matrix $\boldsymbol{I}_e^{-1}(\overline{\widetilde{\boldsymbol{\Psi}}}_n)$ are used as the approximate standard errors.

4.2 Finite Sample Improvements

The EAPS obtained from a single split of the data may not be a good representative of the true active predictor set due to the randomness in the split. Hence, a natural option is to split the data into two parts B times obtaining $(\mathcal{D}_{1n}^1, \mathcal{D}_{2n}^1), (\mathcal{D}_{1n}^2, \mathcal{D}_{2n}^2), \cdots, (\mathcal{D}_{1n}^B, \mathcal{D}_{2n}^B)$. Common choices for B are 25 or 50.

If one were to use the BIC-based selector T_n on each split, even for a moderately sized d the algorithm becomes computationally prohibitive, unless additional structures are assumed on the elements of the model space $\mathcal{A}$ in (3). Hence, in this case we recommend using the adaptive regularization-based selector and we focus on this approach in the rest of this section. Accordingly, the EAPS for the b^{th} split is given by $\widehat{\mathcal{S}}_b(\widehat{\gamma}_{nb})$, $b = 1, 2, \ldots, B$, where $\widehat{\gamma}_{nb}$ is the BIC-enhanced tuning parameter obtained from $\mathcal{D}_{1n}^b$; also set

$$\mathcal{S}_{B,n} = \bigcup_{b=1}^{B} \widehat{\mathcal{S}}_b(\widehat{\gamma}_{nb}).$$

Note that $\mathcal{S}_{B,n}$ is the set of all pairs (j, l) selected in various splits, where j is the index of the mixture component and l is the index of the selected covariate. By Proposition 1, as $n \to \infty$, with probability tending to one, $\mathcal{S}_{B,n} = \mathcal{S}_0$, for any fixed B. In finite samples, it is possible that $\mathcal{S}_{B,n}$ equals the set of indices of regression coefficients in the full FMR model. Even though we have used the unions of EAPS from every split to construct $\mathcal{S}_{B,n}$, it is possible to retain only those covariates that appear a certain percentage of times amongst the B splits.

We refer to the above procedure as the "Msplit" (multiple split) which we now use to test hypotheses analogous to (6) using the covariates associated with the indices in $\mathcal{S}_{B,n}$; for instance, consider

$$H_{0,jl} : \beta_{jl} = 0, \quad \text{for all} \quad (j,l) \in \mathcal{S}_{B,n}. \tag{23}$$

As described previously in (9), we use the test statistic

$$t^b_{jl,n} = \bar{\bar{\beta}}_{jl,b}/\text{SE}(\bar{\bar{\beta}}_{jl,b}), \tag{24}$$

where the index b represents the split, for testing $H_{0,jl}$ at level α. Let p^b_{jl} denote the corresponding p-value obtained by using the student-t approximation to the distribution of $t^b_{jl,n}$. Hence for every split b, we have $\hat{q}_{n,b} = |\widehat{\mathcal{S}}_b(\widehat{\boldsymbol{\gamma}}_{nb})|$ p-values. For those indices in $\mathcal{S}_{B,n}$ but not in $\widehat{\mathcal{S}}_b(\widehat{\boldsymbol{\gamma}}_{nb})$ we assign p-value to be 1. Let $\tilde{\mathcal{S}}_{B,n}$ denote this extended version of $\mathcal{S}_{B,n}$. Let $r_n = |\mathcal{S}_{B,n}|$, and $r^*_n = \max_{1\leq j\leq K, 1\leq b\leq B} \hat{q}_{n,b,j}$, where $\hat{q}_{n,b,j}$ is the size of the estimated active set of the j^{th} component in the b^{th} split. Let $\boldsymbol{p}^{(b)}$ denote the *matrix of the p-values* associated with the hypotheses (23), corresponding to the b^{th} split, as

$$\boldsymbol{p}^{(b)} = \begin{bmatrix} p^b_{11} & p^b_{12} & \cdots & p^b_{1,r^*_n} \\ \vdots & \vdots & \ddots & \vdots \\ p^b_{K1} & p^b_{K2} & \cdots & p^b_{K,r^*_n} \end{bmatrix}, \quad b = 1, 2, \ldots, B. \tag{25}$$

Since the p-values in (25) are not adjusted for multiple testing, as described in Section 4.1, we define the multiplicity adjusted p-value for the b^{th} split for testing (23) to be

$$\bar{p}^b_{jl} = \min(p^b_{jl} \times \hat{q}_{n,b}, 1), j = 1, \ldots, K; l = 1, \ldots, r^*_n; \tag{26}$$

we represent the corresponding matrix of the adjusted p-values by $\bar{\boldsymbol{p}}^{(b)}$, $b = 1, \ldots, B$.

Turning our attention to the the random variables $\{t^b_{jl,n}, b = 1, 2, \cdots B\}$ and their limit distributions, it is obvious that they are correlated. However, the correlation structure is unknown. Evidently, the p-values $\{p^b_{jl}, b = 1, 2 \cdots B\}$ are also correlated. Thus, one needs a method to aggregate dependent p-values.

Aggregating p-values from several independent and dependent tests have long been considered in the literature. Indeed, if the p-values were independent, Fisher's approach suggests combining them by considering -2 times the sum of the logarithm of the p-values. Since this quantity would have a χ^2_{2B} distribution, the aggregate p-value for the hypothesis can be obtained by calibrating this distribution. However, if the p-values are dependent Fisher's approach will not yield a χ^2_{2B} distribution and alternative methods are required. While this has also been studied in the literature, we describe two of the recently suggested methods (Meinshausen et al., 2009; Vovk, 2012).

For $b = 1, 2, \cdots B$, let $\tilde{p}^b_{jl}$ be any p-value for the b^{th} split. A typical choice for $\tilde{p}^b_{jl}$ is (26) *without* the multiplicity adjustment factor $\hat{q}_{n,b}$. Indeed, this would be a p-value for testing a single hypothesis concerning a regression coefficient whose index belongs to the EAPS.

1. **Aggregation using quantiles:** For any $\delta \in (0, 1)$, the aggregated p-value across B splits corresponding to the hypothesis $H_{0,jl}$ is given by

$$Q_{jl}(\delta; \mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n}) = \mathcal{Q}_\delta(\delta^{-1}\tilde{p}^b_{jl} : b = 1, ..., B), \tag{27}$$

   where $\mathcal{Q}_\delta(\cdot)$ is the δ^{th} empirical quantile function. Since the choice of δ is arbitrary, one needs to identify an appropriate value for it in practice, which frequently can be challenging. As an alternative one can find the minimum value of the function $\mathcal{Q}_\delta(\cdot)$ over an interval not including 0, *viz.* $(\delta_{min}, 1)$, for a pre-specified value of $\delta_{\min}$; however, this does not yield FWER control, and hence an adjustment factor is required. Thus, the modified quantile-based p-value is given by

$$Q^*_{jl}(\delta_{\min}; \mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n}) = \min\{1, (1 - \log \delta_{\min}) \inf_{\delta \in (\delta_{\min}, 1)} Q_{jl}(\delta; \mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n})\}, \tag{28}$$

   where $\delta_{min} \in (0, 1)$ acts as a calibration parameter.

2. **Averaging:** Even though the simplest approach to aggregation is averaging, it is known that the average of p-values is not always a p-value. To define a p-value using the averaging, let $\bar{Q}_{jl}(\mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n}) = B^{-1} \sum_{b=1}^{B} \tilde{p}^b_{jl}$ and set

$$\bar{Q}^*_{jl}(\mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n}) = \min(2\bar{Q}_{jl}(\mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n}), 1). \tag{29}$$

Below when there is no scope for confusion, we will denote the left hand side of (27), (28), and (29) respectively, by $Q_{jl}(\delta)$, $Q^*_{jl}(\delta_{\min})$, and $\bar{Q}^*_{jl}$. Note that the averaging method does not involve any calibration parameter such as $\delta_{\min}$ needed in the quantile method, and hence it is easy to use in practice. For a further comparison of the two methods see Section 4.3. Proposition 2, which is of independent interest, establishes that the above aggregation methods yield quantities which are p-values.

Proposition 2 *Let $p_1, p_2, \cdots p_B$ denote B p-values (possibly dependent) for testing a null hypothesis H_0 at level $\alpha \in (0,1)$. Let $\mathcal{Q}_\delta(\cdot)$ be the empirical quantile function as in (27), $Q_B(\delta) = \mathcal{Q}_\delta(\delta^{-1}p_b; b = 1, 2, \dots B)$ and $Q^*_B(\delta_{min}) = \min\{1, (1 - \log \delta_{min}) \inf_{\delta \in (\delta_{min}, 1)} Q_B(\delta)\}$, where $0 < \delta, \delta_{min} < 1$. Additionally, let $\bar{Q}^*_B = \min\{2\bar{Q}_B, 1\}$ where $\bar{Q}_B = B^{-1} \sum_{b=1}^{B} p_b$. Then, the quantities $Q(\delta)$, $Q(\delta_{min})$, and $\bar{Q}_B$ are p-values.*

As an immediate corollary of Proposition 2, it follows that expressions in (27), (28), and (29) are asymptotically p-values. We next investigate the behavior of the FWER for testing (23). To this end, let

$$\mathcal{S}^*_{B,n}(\delta) = \bigcup_{(j,l) \in \mathcal{S}_{B,n}} \{(j,l) : Q_{jl}(\delta) \leq \alpha\}, \tag{30}$$

where $Q_{jl}(\delta)$ is defined using $\bar{p}^b_{jl}$ as in (26). Finally, let $\mathcal{E}(\delta; \mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n}) = \mathcal{N}_0 \cap \mathcal{S}^*_{B,n}(\delta)$. Replacing $Q_{jl}(\delta)$ by $Q^*_{jl}(\delta_{\min})$ and $\bar{Q}^*_{jl}$ in (30), we obtain $\mathcal{E}^*(\delta_{min}; \mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n})$ and $\mathcal{E}^{**}(\mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n})$, respectively. We consider testing the hypotheses in (23), for any given $0 < \alpha, \delta, \delta_{\min} < 1$.

Theorem 3 *Assume that the conditions of Proposition 1 hold. We have that*

(i) $\limsup_{n\to\infty} P\Big(\mathcal{E}(\delta; \mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n}) \neq \emptyset\Big) \leq \alpha;$

(ii) $\limsup_{n\to\infty} P\Big(\mathcal{E}^*(\delta_{\min}; \mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n}) \neq \emptyset\Big) \leq \alpha;$

(iii) $\limsup_{n\to\infty} P\Big(\mathcal{E}^{**}(\mathcal{D}^{1:B}_{1n}, \mathcal{D}^{1:B}_{2n}) \neq \emptyset\Big) \leq \alpha.$

Remark 1. The Bonferroni adjustment used in the computation of the adjusted p-values in (26) is split-dependent. It is possible to replace it by r^*_n, described in the paragraph after equation (24), and the conclusions of Theorem 3 will continue to hold.

While Theorem 3 is concerned with the hypotheses in (23), the assertions of the Theorem continue to hold for hypotheses analogous to (7) and (8) (based on $\mathcal{S}_{B,n}$) if the adjustment factor $\hat{q}_{n,b}$ in (26) is replaced by split-dependent versions of $|\mathcal{G}_j|$ and $|\mathcal{G}|$, respectively. In particular, the proposed method facilitates testing further hypotheses concerning regression coefficients of the selected covariates within and between the mixture components. For example, a local test of the effect of a selected covariate is concerned with the question: is its effect present in at least one mixture component? i.e., testing $H_0 : \beta_{jl} = 0, \text{ for some } 1 \leq j \leq K$. Furthermore, a global test of a selected covariate is concerned with the question: is its effect present in all the mixture components? i.e., testing $H_0 : \beta_{jl} = 0, \text{ for all } 1 \leq j \leq K$. Algorithm 2 summarizes the Msplit method for testing (23).

Algorithm 2 (The Msplit Method for Hypothesis Testing)

Step 1: Divide the data set randomly into two parts of approximately equal size $n/2$, B times; call the resulting data $(\mathcal{D}^1_{1n}, \mathcal{D}^1_{2n}), (\mathcal{D}^2_{1n}, \mathcal{D}^2_{2n}), \cdots, (\mathcal{D}^B_{1n}, \mathcal{D}^B_{2n})$.

Step 2: For each $1 \leq b \leq B$, using $\mathcal{D}^b_{1n}$, the regularization method and the BIC tuning parameter selector, obtain the EAPS $\widehat{\mathcal{S}}_b(\widehat{\gamma}_{nb})$, and set $\mathcal{S}_{B,n} = \bigcup_{1 \leq b \leq B} \widehat{\mathcal{S}}_b(\widehat{\gamma}_{nb})$.

Step 3: Using $\mathcal{D}^{1:B}_{2n}$, obtain the MLE of all the β_{jl} of the selected covariates in **Step 2**.

Step 4: Using the MLEs in **Step 3**, calculate the student-type statistics using (24). Obtain the p-value matrix in (25) and the corresponding adjusted p-values in (26).

Step 5: Use one of the aggregation methods to find the overall p-values.

We now provide the details for implementing Algorithm 2. First, note that **Steps 2** through **4** of Algorithm 2 involve running the EM algorithms described in Section 4.1 for each split $(\mathcal{D}^b_{1n}, \mathcal{D}^b_{2n}), b = 1, 2, \ldots, B$. The procedure is exactly the same as in Algorithm 1 but repeated for each split. This results in the EAPS $\mathcal{S}_{B,n} = \bigcup_{1 \leq b \leq B} \widehat{\mathcal{S}}_b(\widehat{\gamma}_{nb})$. Now, to compute the t-ratios in (24), we use the (approximate) standard errors given in (22) for each split, and thus obtain the matrix of p-values. Finally, in **Step 5**, we use one of the p-value aggregation methods to obtain a single p-value corresponding to each covariate whose index belongs to the EAPS $\mathcal{S}_{B,n}$.

4.3 Numerical Experiments and Data Analysis

In this section we evaluate the finite sample performance of the proposed methods via simulations. The methods are compared with the standard regularization techniques based on ADLASSO and SCAD penalties, using the following three criteria:

1. Empirical family-wise error rate (EFWER): the empirical probability of including at least one covariate with a true zero regression coefficient;

2. Empirical expected number of true positives, E(TP): average number of correctly estimated non-zero regression coefficients;

3. Empirical expected number of false positives, E(FP): average number of incorrectly estimated non-zero regression coefficients.

In the implementation of all methods, we use the BIC-enhanced tuning parameter selector described in Section 3.2. Due to the dimensions (d) considered in our simulations, we choose the adaptive regularization-based selector T_n. We emphasize here that while our methods provide p-values by accounting for variable selection uncertainty, the other two methods only yield a sparse FMR model and no valid inference is feasible without additional work.

The vector of covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^\top$, with realized value $\boldsymbol{x}$, is generated from a multivariate normal with mean zero and a variance-covariance matrix Σ whose diagonal elements equal one and the off-diagonal elements have an autoregressive-type correlation structure; that is, $\mathrm{corr}(X_i, X_j) = .5^{|i-j|}$. Once $\boldsymbol{x}$ is generated it remains fixed throughout the simulations. We have considered FMR models with $K = 2$ mixture components and three covariate dimensions $d = 30, 50, 70$. The intercepts are set to $\beta_{10} = 1, \beta_{20} = 2$.

The results are based on 200 simulated random samples of size $n = 300$ from Gaussian and Binomial FMRs. We present results for the above mentioned criteria for each of the mixture components, Com_1 and Com_2, and the entire mixture, Both. These are provided in Table 1 (Gaussian FMR) and Table S1 of the Supplementary Material (Binomial FMR) using the quantile-based aggregation; Table S2 and Table S3 of the Supplementary Material contain the corresponding results using the average-based aggregation. The Msplit results are based on $B = 50$ and $\delta_{min} = .125$, and SCAD was used in the first-stage of Msplit; finally, all the hypothesis tests were carried out at a significance level of 0.05. The numerical results for single split are not included but, some discussion is provided at the end of this Section. The computational codes for all numerical work in this paper are written in C++ and R software and are placed in the Supplementary Material.

Gaussian FMR. Given $\boldsymbol{x}_i$, the response Y_i is generated from the mixture

$$\pi N(\beta_{10} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_1, \sigma^2) + (1 - \pi) N(\beta_{20} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_2, \sigma^2)$$

with $\pi = .45$, and variances $\sigma^2 = 1, 4, 9, 25, 36$ which yield the signal-to-noise ratio (SNR) values: $25.8, 6.45, 2.87, 1.03, 0.72$. The d-dimensional vector of regression coefficients are

$$\boldsymbol{\beta}_1^\top = (1.8, 1.6, 2.3, 0.0, 2.5, 1.7, 0.0, \ldots, 0.0) \text{ and } \boldsymbol{\beta}_2^\top = (-1.7, 0.0, 2.5, -2.5, -2.0, 0.0, \ldots, 0.0)$$

which contain $q_1 = 5$ and $q_2 = 4$ non-zero regression coefficients, respectively.

From Table 1 we observe that the EFWER for ADLASSO and SCAD increases from 0 to 1.00 as the signal becomes weaker while the Msplit remains stable at the significance level $\alpha = 0.05$ with a maximum EFWER .015 for the weakest signal (SNR $= 0.72$) and $Kd = 140$.

Next, when considering the (average) empirical number of false positives E(FP), we note that for the Msplit method it is almost zero (varies between .000 and .015) in all cases. The ADLASSO and SCAD also perform reasonably well by including, respectively, between .000 to 4.16 and .065 to 2.85 number of false positives as the signal changes from strong (SNR=25.8) to medium (SNR $= 2.87$) across different dimensions. However, as the signal becomes very weak, ADLASSO, on average, includes between 6.64 to 24, while SCAD, on average, includes 8.49 to 17.2 additional covariates with true zero coefficients.

Finally, we compare the three methods in terms of the (average) empirical number of true positives E(TP): we notice that for relatively strong signals (SNR $= 25.8, 6.45$), all three methods perform well, led by SCAD and closely followed by ADLASSO and the Msplit. For the medium signal (SNR $= 2.87$), the SCAD and ADLASSO are still able to pick most of the true non-zero coefficients while the Msplit misses, on average, one true active covariate in each of the mixture components. However, as the signal becomes weaker the task of identifying true non-zero regression coefficients becomes challenging for all three methods.

Table 1 and Table S1 were obtained using the quantile-based aggregation methods wherein the choice of δ_{min} is critical. In Table S2 and Table S3 we provide a comparison of the quantile- and averaging-based aggregation methods. We notice that for strong signals in both the Gaussian and Binomial FMRs, the averaging works as well as the modified quantile method with a carefully chosen δ_{min}. Indeed, in our simulations we noticed that if one were to use the recommended $\delta_{min} = .05$ in Meinshausen et al. (2009), the resulting EFWER exceeds the nominal 5% level. Thus, if the goal of the experiment is to be conservative in the identification of covariates or if the signal is known to be strong, then the averaging method may be desirable due to ease of implementation.

It is well-known that FWER control yields a conservative testing procedure, as is also seen from Table 1 and Table S1. Additionally, turning to E(TP), it is clear that both ADLASSO and SCAD (with the BIC-enhanced tuning parameter) yield higher true positive rates within each mixture component than the splitting methods. Thus, if the goal is to only identify possible covariates affecting a response in each mixture component, it may be worthwhile to use ADLASSO and/or SCAD with the BIC-enhanced tuning parameter selector. However, if the goal is to test the significance of an important effect in an FMR model then controlling FWER may be a better option.

Finally, regarding the single split method our analysis showed that (not presented here) EFWER exhibited high variability between splits, specially when the signal was weak.

Boston Housing data. The data for this example is posted on the UC Irvine Machine Learning Repository and concerns housing values in the suburbs of Boston (also available in the Supplementary Material). There are 506 observations on 14 variables: Per capita crime rate by town (x_1); proportion of residential land zoned for lots over 25,000 sq.ft. (x_2); proportion of nonretail business acres per town (x_3); Charles River dummy variable (x_4); nitric oxide concentration (parts per 10 million; x_5); average number of rooms per dwelling (x_6); proportion of owner occupied units built prior to 1940 (x_7); weighted distances to five Boston employment centres (x_8); index of accessibility to radial highways (x_9); full-value property-tax rate per 10,000 (x_{10}); pupil-teacher ratio by town (x11); 1000(Bk - 0.63)^2

where Bk is the proportion of blacks by town (x_{12}); a numeric vector of percentage values of lower status population (x_{13}); and the median value of owner occupied homes in thousands (MEDV).

Figure 1 in the Supplementary Material shows the histogram of a response variable $y = \text{MEDV}/\text{sd}(\text{MEDV})$. The histogram seems to suggest that a certain proportion of houses have relatively high y values. This motivated us to investigate a 2-component Gaussian FMR model for analyzing the data. Our goal is to identify the set of covariates that are associated with y within the two components, namely the houses with high median value and those that do not have high value. We also would like to test whether the selected covariates are significantly associated with the response variable y.

We used the Msplit (based on SCAD penalty) with $B = 50$ random splits of the data where each split contains two sub-samples each of size $n/2 = 253$. The average estimated value of π over $B = 50$ splits was approximately .36. Thus in the rest of our analysis we refer to the smaller component of the FMR model, which probably describes the higher priced houses, as Com1 and the second component as Com2. The calibration parameter for the aggregated p-values in (28) is $\delta_{\min} = .125$. The Msplit resulted in the EAPS: $\mathcal{S}_{B,n} = \{(1,2),(1,4),(1,6),(1,8),(1,9),(1,11),(1,12),(1,13)\} \bigcup \{(2,1),(2,6),(2,7),(2,10),(2,11),(2,12),(2,13)\}$. We retained only those covariates that occurred at least 10% of the time over $B = 50$ splits. The set of covariates with p-values less than .05 are Com1: $\{x_6, x_9, x_{13}\}$ and Com2: $\{x_6, x_{11}, x_{13}\}$. One possible interpretation of the results of the analysis is as follows: the average number of rooms in a house (x_6, positive effect) and the lower status population (x_{13}, negative effect) are significant factors affecting both high and low priced houses. In addition, for the high priced houses access to highway (x_9, positive effect) is another significant factor whereas for the lower priced houses pupil-teacher ratio by town (x_{11}, negative effect) is a significant factor.

5 SUMMARY AND CONCLUSION

There has been much recent work on post-selection inference problems in sparse linear and generalized linear regression models. In this paper, we have addressed this problem and proposed a new methodology for general hypothesis tests concerning regression coefficients of a selected sparse FMR model. The method naturally accounts for variable selection uncertainty. Our theoretical developments show that under model selection consistency, the family-wise error rate (FWER) can be controlled at a pre-specified nominal level and thus providing sound statistical underpinnings to a practical problem of data analysis using sparse FMR models. Numerical results support our theoretical findings.

Our work opens up opportunities for several research problems concerning post-selection inference in sparse FMRs. Specifically, our methods are not immediately applicable in the high-dimensional settings when the number of parameters increases with the sample size. This is a challenging problem in the context of mixture models and hence new ideas may be needed to provide stable inferential methods. Additionally, computational algorithms in this context needs to be developed since routine application of the EM algorithm described in the paper may not be feasible in high-dimensional settings. These issues are further complicated when the number of mixture components K is unknown.

ACKNOWLEDGEMENTS

The authors would like to thank the editor, associate editor, and two anonymous referees for their constructive comments and suggestions which improved the paper. The research of the first author was partly supported by a grant from The Natural Sciences and Engineering Research Council of Canada.

APPENDIX A: REGULARITY CONDITIONS

In this section we provide the regularity conditions that facilitate our theoretical study. For any candidate model $\mathcal{M}_C \in \mathcal{A}$, where $\mathcal{A}$ is given in (3), with corresponding parameter vector $\boldsymbol{\Psi}_C \in \boldsymbol{\Theta}_C \subseteq \mathbb{R}^{[q(\mathcal{M}_C)+3K-1]}$, consider the random vector $\boldsymbol{Z} = (\boldsymbol{X}, Y)$ with a joint pdf

$f^*(\boldsymbol{z};\boldsymbol{\Psi}_C)$ such that the conditional distribution of $(Y|\boldsymbol{X}=\boldsymbol{x})$ follows the FMR sub-model with parameter $\boldsymbol{\Psi}_C$. Here we assume that the marginal distribution of the covariates $\boldsymbol{X}$ does not depend on $\boldsymbol{\Psi}_C$. Also, we assume that $\boldsymbol{\Psi}_0$ is an interior point of the parameter space.

Regularity Conditions (RC):

(RC1) The support of $f^*(\boldsymbol{z};\boldsymbol{\Psi}_C)$ does not depend on $\boldsymbol{\Psi}_C$, for any candidate model $\mathcal{M}_C \in \mathcal{A}$. Further, $f^*(\boldsymbol{z};\boldsymbol{\Psi}_C)$ is identifiable up to the mixture components permutation.

(RC2) For all $\boldsymbol{\Psi}_C \in \boldsymbol{\Theta}_C$, $f^*(\boldsymbol{z};\boldsymbol{\Psi}_C)$ is three times continuously differentiable w.r.t $\boldsymbol{\Psi}_C$ for almost all $\boldsymbol{z}$.

(RC3) For any sub-model $\mathcal{M}_C \in \mathcal{A}$, the Kullback-Leibler distance

$$KL(\boldsymbol{\Psi}_C;\boldsymbol{\Psi}_0) = E_0\left\{\log\frac{f^*(\boldsymbol{Z};\boldsymbol{\Psi}_0)}{f^*(\boldsymbol{Z};\boldsymbol{\Psi}_C)}\right\} = \int\left[\log\frac{f^*(\boldsymbol{z};\boldsymbol{\Psi}_0)}{f^*(\boldsymbol{z};\boldsymbol{\Psi}_C)}\right]f^*(\boldsymbol{z};\boldsymbol{\Psi}_0)d\boldsymbol{z}$$

is well defined. Further, $\boldsymbol{\Psi}_C^0$ which is the minimizer of the KL with respect to $\boldsymbol{\Psi}_C$, satisfies

$$\int\left[\frac{\partial\log f^*(\boldsymbol{z};\boldsymbol{\Psi}_C)}{\partial\boldsymbol{\Psi}_C}\right]f^*(\boldsymbol{z};\boldsymbol{\Psi}_0)d\boldsymbol{z} = \boldsymbol{0}.$$

(RC4) For any $\boldsymbol{\Psi}_C^0 \in \boldsymbol{\Theta}_C$, and $\boldsymbol{\Psi}_C \in N(\boldsymbol{\Psi}_C^0)$ (where N($\boldsymbol{\Psi}_C^0$) is a neighborhood around $\boldsymbol{\Psi}_C^0$), there exist functions $M_1(\boldsymbol{z}), M_2(\boldsymbol{z})$, and $M_3(\boldsymbol{z})$ (possibly depending on $\boldsymbol{\Psi}_C^0$) such that

$$\left|\frac{\partial f^*(\boldsymbol{z};\boldsymbol{\Psi}_C)}{\partial\psi_j}\right| \le M_1(\boldsymbol{z}) \quad , \quad \left|\frac{\partial^2 f^*(\boldsymbol{z};\boldsymbol{\Psi}_C)}{\partial\psi_j\partial\psi_k}\right| \le M_2(\boldsymbol{z}) \quad , \quad \left|\frac{\partial^3 f^*(\boldsymbol{z};\boldsymbol{\Psi}_C)}{\partial\psi_j\partial\psi_k\partial\psi_l}\right| \le M_3(\boldsymbol{z})$$

where ψ_j represents the elements of the parameter vector $\boldsymbol{\Psi}_C$, such that $E_0\{M_3(\boldsymbol{Z})\} < \infty$, and $\int M_i(\boldsymbol{z})d\boldsymbol{z} < \infty$, for $i = 1, 2$.

(RC5) The matrices

$$\begin{aligned}
\boldsymbol{I}_1(\boldsymbol{\Psi}_C) &= E_0\left\{\left[\frac{\partial}{\partial\boldsymbol{\Psi}_C}\log f^*(\boldsymbol{Z};\boldsymbol{\Psi}_C)\right]\left[\frac{\partial}{\partial\boldsymbol{\Psi}_C}\log f^*(\boldsymbol{Z};\boldsymbol{\Psi}_C)\right]^\top\right\} \\
\boldsymbol{J}_1(\boldsymbol{\Psi}_C) &= -E_0\left\{\left[\frac{\partial^2}{\partial\boldsymbol{\Psi}_C\partial\boldsymbol{\Psi}_C}\log f^*(\boldsymbol{Z};\boldsymbol{\Psi}_C)\right]\right\}
\end{aligned}$$

are finite and positive definite for each $\boldsymbol{\Psi}_C \in \boldsymbol{\Theta}_C$.

These are standard conditions that one adopts when studying the asymptotic properties of the MLEs in parametric models. (RC1) is an identifiability condition on the true model and possible candidate sub-models $\mathcal{M}_C$. Additionally, the common support condition facilitates interchanging differentiation and integration operations on the log-likelihood. (RC2) is a smoothness condition on the density required for asymptotic analyses while (RC3) guarantees the asymptotic existence of the MLEs of the parameters of a sub-model. (RC4) allows interchanging of the expectation and the limits while (RC5) posits the finiteness of the Fisher information for the considered models.

Regularity Conditions on the Penalty (RCP):

We now state the regularity conditions on $p_n(x; \gamma_{nj})$. First, define

$$\begin{aligned} a_n &= \max_{j,l} \left\{ p_n(\beta^0_{jl}; \gamma_{nj})/\sqrt{n} : \beta^0_{jl} \neq 0 \right\} && (31) \\ b_n &= \max_{j,l} \left\{ p'_n(\beta^0_{jl}; \gamma_{nj})/\sqrt{n} : \beta^0_{jl} \neq 0 \right\} && (32) \\ c_n &= \max_{j,l} \left\{ p''_n(\beta^0_{jl}; \gamma_{nj})/n : \beta^0_{jl} \neq 0 \right\} && (33) \end{aligned}$$

where $p'(\cdot; \gamma)$ and $p''_n(\cdot; \gamma)$ are the first and second derivatives of $p_n(x; \gamma)$ with respect to x.

(RCP1) For all n and γ, the penalty function $p_n(x; \gamma)$ is symmetric, nonnegative, nondecreasing and it has first derivative for all $x \in (0, \infty)$. The function is also continuously twice differentiable for all $x \in (c\gamma, \infty)$, and some constant $c > 0$. In addition, $p_n(0; \gamma) = 0$.

(RCP2) $\lim\limits_{n\to\infty} \frac{a_n}{1+b_n} = 0$, and $\lim\limits_{n\to\infty} c_n = 0$.

(RCP3) Define $N_n = \{x | 0 < x < \log n/\sqrt{n}\}$. Then $\lim\limits_{n\to\infty} \inf\limits_{x \in N_n} \frac{p'_n(x;\gamma_{nj})}{\sqrt{n}} = +\infty$.

(RCP4) Assume that, for all $1 \leq j \leq K$, $\gamma_{nj} \in [0, \gamma^*_n]$, and that $\gamma^*_n \to 0$ as $n \to \infty$.

(RCP1) is a standard smoothness condition on the penalty which facilitates obtaining estimators by differentiating the objective function and for studying the asymptotic properties of the estimators of the true non-zero regression coefficients. (RCP2) is required to obtain the $\sqrt{n}$-consistency of the estimators of the true non-zero regression coefficients while (RCP3) is required for sparsistency of the estimators. (RCP4) ensures that the data-dependent choice (17) of the tuning parameter satisfies the conditions (RCP1)-(RCP3).

APPENDIX B: PROOFS.

Proof of Theorem 1. (i) asymptotic normality: Let $\mathcal{L}_n \equiv \sqrt{\frac{n}{2}}\left\{\mathcal{W}_n\left(\overline{\widetilde{\boldsymbol{\Psi}}}_n - \boldsymbol{\Psi}_0(\hat{q}_n)\right)\right\}$ and $\boldsymbol{x} \in \mathcal{R}^m$. Define the two events A_n and B_n as follows:

$$A_n = [\mathcal{L}_n \leq \boldsymbol{x}] \quad \text{and} \quad B_n = \left[\widehat{\mathcal{S}}(T_n) = \mathcal{S}_0\right],$$

where for two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we say $\boldsymbol{x} \leq \boldsymbol{y}$ if and only if every component of $\boldsymbol{x}$ is less than or equal to every component of $\boldsymbol{y}$. Now,

$$P(A_n) = P(A_n|B_n)P(B_n) + P(A_n \cap B_n^c). \tag{34}$$

Notice that

$$P(A_n|B_n) = P\Bigg(\mathcal{L}_n \leq \boldsymbol{x}\Bigg|\widehat{\mathcal{S}}(T_n) = \mathcal{S}_0\Bigg). \tag{35}$$

Under the conditioning $\widehat{\mathcal{S}}(T_n) = \mathcal{S}_0$, the dimension of $\mathcal{W}_n$ reduces to $q_0 + 3K - 1$ and hence $\mathcal{W}_n = \mathcal{W}$. Furthermore, the dimension of $\overline{\widetilde{\boldsymbol{\Psi}}}_n$ and $\boldsymbol{\Psi}_0(\hat{q}_n)$ also reduce to $q_0 + 3K - 1$, and $\boldsymbol{\Psi}_0(\hat{q}_n) = \boldsymbol{\Psi}_0$. Thus, writing $\overline{\widetilde{\boldsymbol{\Psi}}}_n$ as $\overline{\widetilde{\boldsymbol{\Psi}}}_n(\mathcal{D}_{2n})$, (35) becomes

$$P\left(\sqrt{\frac{n}{2}}\mathcal{W}\left(\overline{\widetilde{\boldsymbol{\Psi}}}_n(\mathcal{D}_{2,n}) - \boldsymbol{\Psi}_0\right) \leq \boldsymbol{x}\middle|\widehat{\mathcal{S}}(T_n) = \mathcal{S}_0\right). \tag{36}$$

Now, observing that $\widehat{\mathcal{S}}(T_n)$ depends on $\mathcal{D}_{1n}$, it follows from the independence of $\mathcal{D}_{1n}$ and $\mathcal{D}_{2n}$ that (36) reduces to

$$P\left(\sqrt{\frac{n}{2}}\mathcal{W}\left(\overline{\widetilde{\boldsymbol{\Psi}}}_n(\mathcal{D}_{2,n}) - \boldsymbol{\Psi}_0\right) \leq \boldsymbol{x}\right).$$

Now, under the regularity conditions (RC1)-(RC5), it follows upon noticing that $\overline{\widetilde{\boldsymbol{\Psi}}}_n(\mathcal{D}_{2n})$ is the MLE of $\boldsymbol{\Psi}_0$, that the expression in (36) converges to $P\left(N_m\left(\mathbf{0}, [\mathcal{W}I_1(\boldsymbol{\Psi}_0)\mathcal{W}^\tau]^{-1}\right) \leq \boldsymbol{x}\right)$, as $n \to \infty$. Also, $\lim_{n\to\infty} P(B_n) = 1$ by consistency of the model selector T_n. Thus, combining the results and taking limits in (34), the proof of **(i)** follows.

(ii) FWER **control**: By the definition of the event $\mathcal{E}(\mathcal{D}_{1n},\mathcal{D}_{2n})$ in (10),

$$\begin{aligned}
P\Big\{\mathcal{E}(\mathcal{D}_{1n},\mathcal{D}_{2n}) \neq \emptyset\Big\} &= E_0\Big\{\mathbf{1}_{[\mathcal{E}(\mathcal{D}_{1n},\mathcal{D}_{2n})\neq\emptyset]}\Big\} = E_0\Big[E_0\Big\{\mathbf{1}_{[\mathcal{E}(\mathcal{D}_{1n},\mathcal{D}_{2n})\neq\emptyset]}\Big|\mathcal{D}_{1n}\Big\}\Big] \\
&\leq E_0\left[\sum_{j=1}^{K}\sum_{l\in\widehat{S}_j(T_n)} P\Big\{(j,l): p_{jl} \leq \alpha/\hat{q}_n \text{ and } \mathrm{H}_{0,jl}: \beta_{jl}^0 = 0 \text{ is true}\Big|\mathcal{D}_{1n}\Big\}\right] \\
&= E_0\left[\sum_{j=1}^{K}\sum_{l\in\widehat{S}_j(T_n)} P\Big\{(j,l): p_{jl} \leq \alpha/\hat{q}_n\Big|\mathcal{D}_{1n},\mathrm{H}_{0,jl}\Big\}\right] \\
&\leq E_0\left[\sum_{j=1}^{K}\sum_{l\in\widehat{S}_j(T_n)} \frac{\alpha}{\hat{q}_n}\right] = \alpha\ E_0\left[\hat{q}_n^{-1}\sum_{j=1}^{K}\sum_{l\in\widehat{S}_j(T_n)} 1\right] \leq \alpha E_0\left[\hat{q}_n^{-1}\times\hat{q}_n\right] = \alpha,
\end{aligned}$$

where we used the fact that p_{jl} is independent of $\mathcal{D}_{1n}$ and that it is a p-value. ■

Proof of Proposition 1. Before we embark on the proof, we introduce the following definition which is concerned with the concepts of under-fitted, over-fitted, and exactly-fitted FMR models. We state two lemmas that play a critical role in the proof of the proposition.

Definition 4 *A candidate* FMR *sub-model* $\mathcal{M}_C \in \mathcal{A}$ *is said to be:* **(i)** *under-fitted if* $S_j \not\supseteq S_j^0$ *for at least one* $1 \leq j \leq K$*, and we write* $\mathcal{M}_C \not\supseteq \mathcal{M}_0$*.* **(ii)** *over-fitted if* $S_j \supseteq S_j^0$ *for all* $1 \leq j \leq K$*, and* $S_j \supset S_j^0$ *for at least one* $1 \leq j \leq K$*, and we write* $\mathcal{M}_C \supset \mathcal{M}_0$*.* **(iii)** *exactly-fitted if* $S_j = S_j^0$*, for all* $1 \leq j \leq K$*, and we write* $\mathcal{M}_C = \mathcal{M}_0$*.*

Lemma 1 below is critical to remove condition C4 in Theorem 1 of Zhang et al. (2010). We note that it is hard to verify condition C4 in practice.

Lemma 1 *Assume that the regularity conditions (RC1)-(RC5) are satisfied. As* $n \to \infty$*,*

(i) *for any candidate sub-model* $\mathcal{M}_C \in \mathcal{A}$*, there exists a maximizer* $\overline{\mathbf{\Psi}}_{n,C}$ *of the log-likelihood* $l_n(\mathbf{\Psi}_C)$ *given in* (2)*, such that* $\overline{\mathbf{\Psi}}_{n,C} \xrightarrow{a.s} \mathbf{\Psi}_C^0$*, and furthermore,*

$$\sqrt{n}(\overline{\mathbf{\Psi}}_{n,C} - \mathbf{\Psi}_C^0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{V}(\mathbf{\Psi}_C^0)),$$

where $\boldsymbol{V}(\mathbf{\Psi}_C^0) = \boldsymbol{J}_1^{-1}(\mathbf{\Psi}_C^0)\boldsymbol{I}_1(\mathbf{\Psi}_C^0)\boldsymbol{J}_1^{-1}(\mathbf{\Psi}_C^0)$*. Also, if* $\mathcal{M}_C = \mathcal{M}_0$ *then* $\boldsymbol{I}_1(\mathbf{\Psi}_0) = \boldsymbol{J}_1(\mathbf{\Psi}_0)$*.*

(ii) *let $\mathcal{M}_C \in \mathcal{A}$ be any under-fitted sub-model. Then,*

$$\frac{1}{n}\left[\ell_n(\overline{\boldsymbol{\Psi}}_{n,C}) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0})\right] \xrightarrow{a.s} E_0[\log f^*(\boldsymbol{Z}_i; \boldsymbol{\Psi}_C^0)] - E_0[\log f^*(\boldsymbol{Z}_i; \boldsymbol{\Psi}_0)] < 0,$$

where $\overline{\boldsymbol{\Psi}}_{n,0}$ is the MLE *of $\boldsymbol{\Psi}_0$.*

(iii) *let $\mathcal{M}_C \in \mathcal{A}$ be any over-fitted sub-model. Then,*

$$2\left[\ell_n(\overline{\boldsymbol{\Psi}}_{n,C}) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0})\right] \xrightarrow{d} \chi^2_{[q(\mathcal{M}_C)-q_0]}.$$

Proof of Lemma 1. **(i)** Under the regularity conditions (RC1)-(RC5), the existence and strong consistency of a maximizer $\overline{\boldsymbol{\Psi}}_{n,C}$ of the log-likelihood $l_n(\boldsymbol{\Psi}_C)$ satisfying

$$\ell_n'(\overline{\boldsymbol{\Psi}}_{n,C}) = \left.\frac{\partial \ell_n(\boldsymbol{\Psi}_C)}{\partial \boldsymbol{\Psi}_C}\right|_{\boldsymbol{\Psi}_C = \overline{\boldsymbol{\Psi}}_{n,C}} = \mathbf{0} \tag{37}$$

follows from Chanda (1954) and White (1982). Also, by the central limit theorem, the score function $\ell_n'(\boldsymbol{\Psi}_C^0)/\sqrt{n}$ has an asymptotic normal distribution with mean zero and the variance-covariance matrix $\boldsymbol{V}(\boldsymbol{\Psi}_C^0) = \boldsymbol{J}_1^{-1}(\boldsymbol{\Psi}_C^0)\boldsymbol{I}_1(\boldsymbol{\Psi}_C^0)\boldsymbol{J}_1^{-1}(\boldsymbol{\Psi}_C^0)$. By a third-order Taylor's expansion of the right hand side of (37) around $\boldsymbol{\Psi}_C^0$, the strong consistency of $\overline{\boldsymbol{\Psi}}_{n,C}$, and the boundedness condition of the third-order derivative in RC3, $\sqrt{n}(\overline{\boldsymbol{\Psi}}_{n,C} - \boldsymbol{\Psi}_C^0)$ has a mean-zero asymptotic normal distribution with variance-covariance matrix $\boldsymbol{V}(\boldsymbol{\Psi}_C^0)$.

(ii) As $n \to \infty$, by part **(i)**, for any candidate model $\mathcal{M}_C$ and the true model $\mathcal{M}_0$ we have that $\overline{\boldsymbol{\Psi}}_{n,C} \xrightarrow{a.s} \boldsymbol{\Psi}_C^0$ and $\overline{\boldsymbol{\Psi}}_{n,0} \xrightarrow{a.s} \boldsymbol{\Psi}_0$, respectively. By (RC2), (RC3), and the strong law of large numbers, it follows that as $n \to \infty$

$$\frac{1}{n}\ell_n(\overline{\boldsymbol{\Psi}}_{n,C}) - \frac{1}{n}\ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) \xrightarrow{a.s} E_0\{\log f^*(\boldsymbol{Z}_i; \boldsymbol{\Psi}_C^0)\} - E_0\{\log f^*(\boldsymbol{Z}_i; \boldsymbol{\Psi}_0)\} < 0.$$

(iii) Since the candidate model $\mathcal{M}_C$ is over-fitted, the true model $\mathcal{M}_0$ is nested in $\mathcal{M}_C$. Hence, by a verification of the conditions of Corollary 3.4 of Vuong (1989), it follows that the likelihood ratio statistic comparing these two models, *viz.* $2\left[\ell_n(\overline{\boldsymbol{\Psi}}_{n,C}) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0})\right]$, has a chi-squared distribution with degrees of freedom $q(\mathcal{M}_C) - q_0$. ■

Lemma 2 *Assume that the regularity conditions (RC1)-(RC5) and (RCP1)-(RCP3) hold. Also assume that γ_n^* satisfies (RCP4). Then, $\lim_{n\to\infty} P\{q(\boldsymbol{\gamma}_n) = q_0\} = 1$, where $q(\boldsymbol{\gamma}_n)$ is defined below.*

Proof of Lemma 2. The proof essentially follows from the estimation and selection consistency of the MPLE. Specifically, consider the MPLE $\widehat{\mathbf{\Psi}}_n(\boldsymbol{\gamma}_n)$ corresponding to the tuning parameter $\boldsymbol{\gamma}_n$ satisfying (RCP1)-(RCP4). Note that $q(\boldsymbol{\gamma}_n) = \sum_{j=1}^{K}\sum_{l=1}^{d} I(\hat{\beta}_{jl}(\boldsymbol{\gamma}_n) \neq 0)$ is the total number of estimated non-zero regression coefficients $\hat{\beta}_{jl}(\boldsymbol{\gamma}_n)$ contained in the MPLE $\widehat{\mathbf{\Psi}}_n(\boldsymbol{\gamma}_n)$. On the other hand, there are q_0 non-zero regression coefficients in the true sparse model. Under the regularity conditions (RC1)-(RC5) and (RCP1)-(RCP3), by Theorems 1 and 2 of Khalili and Chen (2007), for all (j, l) such that $\hat{\beta}_{jl}(\boldsymbol{\gamma}_n) \neq 0$, $\hat{\beta}_{jl}(\boldsymbol{\gamma}_n) \longrightarrow \beta_{jl}^0 \neq 0$, with probability tending to one, as $n \to \infty$. Hence, the Lemma follows. ■

We now turn to the proof of **Proposition 1**. The main idea of the proof is to show that the BIC$(\boldsymbol{\gamma})$ in (18) will not choose $\boldsymbol{\gamma}$ that corresponds to the under-fitted and over-fitted FMR models as described in Definition 4, and hence selects the true model. We divide the proof into three main steps.

Step 1: (**Lower bound for** BIC$(\boldsymbol{\gamma})$). For any $\boldsymbol{\gamma}$ the BIC is given by

$$\text{BIC}(\boldsymbol{\gamma}) = \frac{1}{n}\left\{-2\ell_n(\widehat{\mathbf{\Psi}}_n(\boldsymbol{\gamma})) + q(\boldsymbol{\gamma})\log n\right\}, \tag{38}$$

where $\widehat{\mathbf{\Psi}}_n(\boldsymbol{\gamma})$ is the MPLE. Note that $\mathcal{M}_{\boldsymbol{\gamma}}$ denotes the FMR sub-model dictated by the MPLE; let the corresponding parameter vector be $\mathbf{\Psi}_{\mathcal{M}_{\boldsymbol{\gamma}}} \equiv \mathbf{\Psi}(\boldsymbol{\gamma})$. Further, let $\overline{\mathbf{\Psi}}_{n,\mathcal{M}_{\boldsymbol{\gamma}}} \equiv \overline{\mathbf{\Psi}}_n(\boldsymbol{\gamma})$ be the MLE of the parameters of the sub-model $\mathcal{M}_{\boldsymbol{\gamma}}$, which maximizes the log-likelihood $\ell_n(\mathbf{\Psi}(\boldsymbol{\gamma}))$. Thus,

$$\ell_n(\overline{\mathbf{\Psi}}_n(\boldsymbol{\gamma})) \geq \ell_n(\widehat{\mathbf{\Psi}}_n(\boldsymbol{\gamma})). \tag{39}$$

Now by plugging (39) in (38), for any $\boldsymbol{\gamma}$

$$\text{BIC}(\boldsymbol{\gamma}) \geq \frac{1}{n}\left[-2\ell_n(\overline{\mathbf{\Psi}}_n(\boldsymbol{\gamma})\right]. \tag{40}$$

Step 2: (**Oracle-type property for** BIC$(\boldsymbol{\gamma}_n)$). Turning to the MPLE $\widehat{\mathbf{\Psi}}_n(\boldsymbol{\gamma}_n)$, where $\boldsymbol{\gamma}_n$ is chosen according to the regularity conditions (RCP1)-(RCP3), by Lemma 2, $q(\boldsymbol{\gamma}_n)$ equals to q_0 with probability tending to one, as $n \to \infty$. Denote by $\widehat{\mathbf{\Psi}}_A(\boldsymbol{\gamma}_n)$ the sub-vector of $\widehat{\mathbf{\Psi}}_n(\boldsymbol{\gamma}_n)$, with dimension equal to $\dim(\mathbf{\Psi}_0) = q_0 + 3K - 1$, which contains the estimated non-zero regression coefficients, mixing probabilities and dispersion parameters. Then, for large n, with a probability close to one, the following normal equations hold;

$$\nabla_{\boldsymbol{\Psi}_A}\left[\frac{1}{n}\ell_n(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n)) - \frac{1}{n}\boldsymbol{p}_n(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n);\boldsymbol{\gamma}_n)\right] = \mathbf{0}.$$

For the MLE $\overline{\boldsymbol{\Psi}}_{n,0}$, the following normal equations

$$\nabla_{\boldsymbol{\Psi}_A}\left[\frac{1}{n}\ell_n(\overline{\boldsymbol{\Psi}}_{n,0})\right] = \mathbf{0}$$

hold. For simplicity in notation we replace the gradient operator ∇ by $'$. By subtracting the two normal equations we have that, for large n,

$$\frac{1}{n}\left[\ell_n'(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n)) - \ell_n'(\overline{\boldsymbol{\Psi}}_{n,0})\right] = \frac{1}{n}\left[\boldsymbol{p}_n'(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n);\boldsymbol{\gamma}_n)\right].$$

Using a first-order Taylor's expansion of the left hand side of the above equation,

$$\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n) - \overline{\boldsymbol{\Psi}}_{n,0} = \left[\frac{1}{n}\ell''(\widetilde{\boldsymbol{\Psi}}_A)\right]^{-1} \times \left[\frac{1}{n}\boldsymbol{p}_n'(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n);\boldsymbol{\gamma}_n)\right],$$

where $\widetilde{\boldsymbol{\Psi}}_A$ is on a line segment joining $\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n)$ and $\overline{\boldsymbol{\Psi}}_{n,0}$. By the regularity condition (RC4) and the consistency of the estimator $\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n)$ it follows that, as $n \to \infty$,

$$\|\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n) - \overline{\boldsymbol{\Psi}}_{n,0}\|^2 = O_p(\|\boldsymbol{p}_n'(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n);\boldsymbol{\gamma}_n)/n\|^2) = O_p(\|\boldsymbol{p}_n'(\boldsymbol{\Psi}_0;\boldsymbol{\gamma}_n)/n\|^2).$$

On the other hand, by the structure of the penalty function in (15), and its dependence on the regression coefficients and the mixing proportions only,

$$\|\boldsymbol{p}_n'(\boldsymbol{\Psi}_0;\boldsymbol{\gamma}_n)/n\|^2 = \|\nabla_{\boldsymbol{B}_1}\boldsymbol{p}_n(\boldsymbol{\Psi}_0;\boldsymbol{\gamma}_n)/n\|^2 + \sum_{j=1}^{K-1}\left(\sum_{l=1}^{q_j^0} p_n(\beta_{jl}^0;\gamma_{nj})/n\right)^2.$$

Therefore, combining the last two equations, as $n \to \infty$, by (RCP2),

$$\|\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n) - \overline{\boldsymbol{\Psi}}_{n,0}\|^2 = O_p\left(\frac{b_n^2}{n} + \frac{a_n^2}{n}\right) = o_p(n^{-1}), \tag{41}$$

where the quantities a_n and b_n are defined in (31) and (32), respectively.

Also, using a second-order Taylor's expansion, and by (41), for large n, with a probability close to one,

$$\begin{aligned}0 > \frac{1}{n}\left[\ell_n(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n)) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0})\right] &= -\frac{1}{2}\left(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n) - \overline{\boldsymbol{\Psi}}_{n,0}\right)^\top\left[-\frac{1}{n}\ell''(\widetilde{\boldsymbol{\Psi}}_A)\right]\left(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n) - \overline{\boldsymbol{\Psi}}_{n,0}\right)\\ &\geq -\frac{\rho_2}{2}\|\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n) - \overline{\boldsymbol{\Psi}}_{n,0}\|^2,\end{aligned}$$

where ρ_2 is the largest eigen-value of the information matrix $\boldsymbol{I}_1(\boldsymbol{\Psi}_0)$, which by (RC5) is positive and finite. Hence, by (41), as $n \to \infty$, with probability tending to one,

$$\frac{1}{n}\left[\ell_n(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n)) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0})\right] = o(1). \tag{42}$$

Now consider the classical $\text{BIC}^*_{\kappa_0}$ evaluated based on the MLE of parameters of the true model $\mathcal{M}_0$, *viz.*,

$$\text{BIC}^*_{\kappa_0} = \frac{1}{n}\left\{-2\ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) + \kappa_0 \log n\right\},$$

where $\kappa_0 = q_0 + 3K - 1$. Thus, by (42),

$$\lim_{n\to\infty} P\left\{\frac{2}{n}\ell_n(\widehat{\boldsymbol{\Psi}}_A(\boldsymbol{\gamma}_n)) - \frac{2}{n}\ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) + \frac{\log n}{n}(q(\boldsymbol{\gamma}_n) - \kappa_0) = 0\right\} = 1$$

which implies that

$$\lim_{n\to\infty} P\left\{\text{BIC}(\boldsymbol{\gamma}_n) = \text{BIC}^*_{\kappa_0}\right\} = 1. \tag{43}$$

Hence, it follows that if $\boldsymbol{\gamma}_n$ is chosen according to the regularity conditions (RCP1)-(RCP4), then the difference between the BIC evaluated at the MPLE and the BIC evaluated at the MLE based on the true FMR model $\mathcal{M}_0$ equals zero, with probability tending to one, as $n \to \infty$.

Step 3: (**Selection consistency of $\mathcal{M}_{\widehat{\boldsymbol{\gamma}}_n}$**). We next show that the $\boldsymbol{\gamma}$ that fail to identify the true model cannot be selected by the BIC. That is, such a $\boldsymbol{\gamma}$ cannot be the minimizer of $\text{BIC}(\boldsymbol{\gamma})$ in (18). To this end, let

$$\begin{aligned}
\boldsymbol{\Gamma}_n^- &= \{\boldsymbol{\gamma} \in [0, \gamma_n^*]^K : \mathcal{M}_{\boldsymbol{\gamma}} \not\supset \mathcal{M}_0\},\\
\boldsymbol{\Gamma}_n^+ &= \{\boldsymbol{\gamma} \in [0, \gamma_n^*]^K : \mathcal{M}_{\boldsymbol{\gamma}} \supset \mathcal{M}_0 \;\;,\;\; \mathcal{M}_{\boldsymbol{\gamma}} \neq \mathcal{M}_0\}, \quad \text{and}\\
\boldsymbol{\Gamma}_n^0 &= \{\boldsymbol{\gamma} \in [0, \gamma_n^*]^K : \mathcal{M}_{\boldsymbol{\gamma}} = \mathcal{M}_0\}
\end{aligned}$$

denote the collection of those $\boldsymbol{\gamma}$ whose MPLE $\widehat{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma})$ yields respectively under-fitted, over-fitted, and exactly-fitted FMR sub-models as described in Definition 4, and γ_n^* is chosen according to condition (RCP4). We will show that

$$\lim_{n\to\infty} P\left(\inf_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_n^-} \text{BIC}(\boldsymbol{\gamma}) > \text{BIC}(\boldsymbol{\gamma}_n)\right) = 1, \quad \text{and} \tag{44}$$

$$\lim_{n\to\infty} P\left(\inf_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_n^+} \text{BIC}(\boldsymbol{\gamma}) > \text{BIC}(\boldsymbol{\gamma}_n)\right) = 1. \tag{45}$$

This will then imply that $\widehat{\boldsymbol{\gamma}}_n \notin \boldsymbol{\Gamma}_n^- \cup \boldsymbol{\Gamma}_n^+$ for sufficiently large n, where $\widehat{\boldsymbol{\gamma}}_n$ is defined in (17). Hence, $\widehat{\boldsymbol{\gamma}}_n \in \boldsymbol{\Gamma}_n^0$ for large n, and this completes the proof of Proposition 1.

We next turn to the proof of (44). For $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_n$ satisfying (40) and (43), with probability tending to one, as $n \to \infty$,

$$\text{BIC}(\boldsymbol{\gamma}) - \text{BIC}(\boldsymbol{\gamma}_n) \geq -\frac{2}{n}\ell_n(\overline{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma})) + \frac{2}{n}\ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) - \frac{\log n}{n} \times \kappa_0.$$

Now, by taking the infimum over $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_n^-$ it follows that

$$\inf_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_n^-} \text{BIC}(\boldsymbol{\gamma}) - \text{BIC}(\boldsymbol{\gamma}_n) \geq \frac{2}{n} \times \inf_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_n^-} \left[-\ell_n(\overline{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma}))\right] + \frac{2}{n}\ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) - \frac{\log n}{n} \times \kappa_0.$$

Since $d < \infty$, the number of under-fitted models corresponding to $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_n^-$ is finite, and thus the infimum in the above can be replaced by the minimum over the candidate under-fitted models. Hence, for large n,

$$\inf_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_n^-} \text{BIC}(\boldsymbol{\gamma}) - \text{BIC}(\boldsymbol{\gamma}_n) \geq \frac{2}{n} \min_{\mathcal{M}_0 \not\subset \mathcal{M}_C \in \mathcal{A}} \left[-\ell_n(\overline{\boldsymbol{\Psi}}_{n,C})\right] + \frac{2}{n}\ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) - \frac{\log n}{n} \times \kappa_0.$$

By Lemma 1-(ii), for any candidate under-fitted model $\mathcal{M}_C \in \mathcal{A}$, as $n \to \infty$, the right hand side of the above inequality converges to

$$2 \min_{\mathcal{M}_0 \not\subset \mathcal{M}_C \in \mathcal{A}} \left\{E_0[\log f^*(\boldsymbol{Z}_i; \boldsymbol{\Psi}_0) - E_0[\log f^*(\boldsymbol{Z}_i; \boldsymbol{\Psi}_C^0)]\right\} > 0$$

with probability tending to one. This completes the proof of (44).

We now prove (45). For any $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_n^+$, by Lemma 2 for large n, it follows that $q(\boldsymbol{\gamma}) - q(\boldsymbol{\gamma}_n) > \eta + o_p(1)$, for some $\eta > 0$. Thus by (39), for large n,

$$\begin{aligned} n\Big[\text{BIC}(\boldsymbol{\gamma}) - \text{BIC}(\boldsymbol{\gamma}_n)\Big] &= -2\ell_n(\widehat{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma})) + 2\ell_n(\widehat{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma}_n)) + \log n\Big[q(\boldsymbol{\gamma}) - q(\boldsymbol{\gamma}_n)\Big] \\ &\geq -2\ell_n(\overline{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma})) + 2\ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) + \log n\Big[q(\boldsymbol{\gamma}) - q(\boldsymbol{\gamma}_n)\Big] \\ &\geq -2[\ell_n(\overline{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma})) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0})] + (\eta + o_p(1))\log n. \end{aligned}$$

By taking the infimum over $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_n^+$ on both sides of the above inequality,

$$\inf_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_n^+} n\Big[\text{BIC}(\boldsymbol{\gamma}) - \text{BIC}(\boldsymbol{\gamma}_n)\Big] \geq \inf_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_n^+} \left\{-2\left[\ell_n(\overline{\boldsymbol{\Psi}}_n(\boldsymbol{\gamma})) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0})\right]\right\} + (\eta + o_p(1))\log n.$$

Once again, since $d < \infty$ the number of over-fitted models is also finite, we can write

$$\inf_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_n^+} n \Big[\text{BIC}(\boldsymbol{\gamma}) - \text{BIC}(\boldsymbol{\gamma}_n) \Big] \geq \min_{\mathcal{M}_0 \subsetneq \mathcal{M}_C \in \mathcal{A}} \left\{ -2 \left[\ell_n(\overline{\boldsymbol{\Psi}}_{n,C}) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) \right] \right\} + (\eta + o_p(1)) \log n. \quad (46)$$

By Lemma 1-(iii), for any over-fitted candidate FMR model $\mathcal{M}_0 \subsetneq \mathcal{M}_C$, as $n \to \infty$, $2 \left[\ell_n(\overline{\boldsymbol{\Psi}}_{n,C}) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) \right] \xrightarrow{d} \chi^2_{[q(\mathcal{M}_C) - q_0]}$. Hence,

$$\min_{\mathcal{M}_0 \subsetneq \mathcal{M}_C} \left\{ -2 \left[\ell_n(\overline{\boldsymbol{\Psi}}_{n,C}) - \ell_n(\overline{\boldsymbol{\Psi}}_{n,0}) \right] \right\} = O_p(1) \quad , \quad \text{as} \quad n \to \infty.$$

Thus, the right hand side of the inequality in (46) diverges to $+\infty$, as $n \to \infty$. Hence,

$$\lim_{n \to \infty} P\{ \inf_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_n^+} n[\text{BIC}(\boldsymbol{\gamma}) - \text{BIC}(\boldsymbol{\gamma}_n)] > 0 \} = 1$$

and this completes the proof of (45). ■

Proof of Proposition 2. To establish the claim, it is sufficient to show that under H_0 the following hold:

(i) $P(Q_B(\delta) \leq \alpha) \leq \alpha$;

(ii) $P(Q_B^*(\delta_{min}) \leq \alpha) \leq \alpha$;

(iii) $P(\bar{Q}_B^* \leq \alpha) \leq \alpha$.

We begin with the proof of Part **(i)**. Using the definition of $Q_B(\delta)$, note that

$$[Q_B(\delta) \leq \alpha] = [\mu_B(\alpha\delta) \geq \delta] \,, \quad \text{where} \quad \mu_B(x) = \frac{1}{B} \sum_{b=1}^{B} 1_{[p_b \leq x]} \quad (47)$$

is the empirical distribution of the given p-values. Now, by using the Markov's inequality

$$P\left(Q_B(\delta) \leq \alpha\right) = P\left(\mu_B(\alpha\delta) \geq \delta\right) \leq \frac{1}{\delta} E\left(\mu_B(\alpha\delta)\right) = \delta^{-1} P\left(p_b \leq \alpha\delta\right) \leq \alpha$$

where, in the last step, we have used that p_b for all $1 \leq b \leq B$ is a p-value.

Turning to Part **(ii)**, we notice that it is sufficient to establish that

$$P\left(\inf_{\delta \in (\delta_{min}, 1)} Q_B(\delta) \leq \alpha \right) \leq \alpha(1 - \log \delta_{min}).$$

To this end, again using (47) and Markov's inequality it follows that

$$\begin{aligned} P\left(\inf_{\delta\in(\delta_{min},1)} Q_B(\delta)\leq\alpha\right) &= P\left(\inf_{\delta\in(\delta_{min},1)} \delta^{-1}\mu_B(\alpha\delta)\geq 1\right) && (48)\\ &\leq E\left(\sup_{\delta\in(\delta_{min},1)} \frac{\mu_B(\alpha\delta)}{\delta}\right) && (49)\\ &\leq E\left(\sup_{\delta\in(\delta_{min},1)} \frac{1_{[p_1\leq\alpha\delta]}}{\delta})\right). && (50)\end{aligned}$$

Now, to deal with (50) (see also Meinshausen et al. (2009)), note that for any random variable taking values in (0,1) the following holds:

$$\sup_{\delta\in(\delta_{min},1)} \frac{1_{[U\leq\alpha\delta]}}{\delta} = \begin{cases} 0 & \text{if } U\geq\alpha\\ \frac{\alpha}{U} & \text{if } \alpha\delta_{min}\leq U<\alpha\\ \frac{1}{\delta_{min}} & \text{if } U\leq\alpha\delta_{min}. \end{cases} \quad (51)$$

Hence for a uniform (0, 1) valued random variable using (51) we obtain

$$E\left[\sup_{\delta\in(\delta_{min},1)} \frac{1_{[U\leq\alpha\delta]}}{\delta}\right]\leq\alpha(1-\log\delta_{min}).$$

Now, Part **(ii)** follows from the fact that p_1 in (50) is a p-value.

Finally, Part **(iii)** follows from Theorem 1 of Vovk (2012). ■

Proof of Theorem 3. For the ease of notation, we set $Q_{jl} = Q_{jl}(\delta;\mathcal{D}_{1n}^{1:B},\mathcal{D}_{2n}^{1:B})$, $Q_{jl}^* = Q_{jl}^*(\delta_{min};\mathcal{D}_{1n}^{1:B},\mathcal{D}_{2n}^{1:B})$, and $\mathcal{E}(\mathcal{D}_{1n}^{1:B},\mathcal{D}_{2n}^{1:B}) = \mathcal{E}(\delta;\mathcal{D}_{1n}^{1:B},\mathcal{D}_{2n}^{1:B})$, $\mathcal{E}^*(\mathcal{D}_{1n}^{1:B},\mathcal{D}_{2n}^{1:B}) = \mathcal{E}^*(\delta_{min};\mathcal{D}_{1n}^{1:B},\mathcal{D}_{2n}^{1:B})$.

We start with the proof of Part **(i)**. Note that

$$\begin{aligned} P\Big\{\mathcal{E}(\mathcal{D}_{1n}^{1:B},\mathcal{D}_{2n}^{1:B})\neq\emptyset\Big\} &= E_0\Big\{\mathbf{1}_{[\mathcal{E}(\mathcal{D}_{1n}^{1:B},\mathcal{D}_{2n}^{1:B})\neq\emptyset]}\Big\} = E_0\Big[E_0\Big\{\mathbf{1}_{[\mathcal{E}(\mathcal{D}_{1n}^{1:B},\mathcal{D}_{2n}^{1:B})\neq\emptyset]}\Big|\mathcal{D}_{1n}^{1:B}\Big\}\Big]\\ &\leq E_0\left[\sum_{(j,l)\in\mathcal{S}_{B,n}} P\Big\{Q_{jl}\leq\alpha \text{ and } \mathrm{H}_{0,jl}:\beta_{jl}^0=0 \text{ is true}\Big|\mathcal{D}_{1n}^{1:B}\Big\}\right].\end{aligned}$$

Now, as in the proof of Proposition 2, for any $(j,l)\in\mathcal{S}_{B,n}$ by letting

$$\mu_{jl,B}(x) = B^{-1}\sum_{b=1}^{B} 1_{[\bar{p}_{jl}^b\leq x]}$$

be the empirical distribution of the adjusted p-values $\bar{p}_{jl}^{b}$s in (26), it follows that

$$\begin{aligned} P\Big\{\mathcal{E}(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset\Big\} &\leq E_0\left[\sum_{(j,l)\in\mathcal{S}_{B,n}} P\Big\{\mu_{jl,B}(\alpha\delta) \geq \delta \text{ and } \mathrm{H}_{0,jl} : \beta_{jl}^0 = 0 \text{ is true}\Big|\mathcal{D}_{1n}^{1:B}\Big\}\right] \\ &\leq \delta^{-1}E_0\left[\sum_{(j,l)\in\mathcal{S}_{B,n}} E\left[\mu_{jl,B}(\alpha\delta)\Big|H_{0,jl} : \beta_{jl}^0 = 0 \text{ is true}, \mathcal{D}_{1n}^{1:B}\right]\right], \quad (52) \end{aligned}$$

where the last inequality is obtained by applying the conditional Markov's inequality. Now, using the definition of $\mu_{jl,B}(\cdot)$ and $\bar{p}_{jl}^{b}$ notice that

$$\begin{aligned} E\left[\mu_{jl,B}(\alpha\delta)\Big|H_{0,jl} : \beta_{jl}^0 = 0 \text{ is true}, \mathcal{D}_{1n}^{1:B}\right] &= \frac{1}{B}\sum_{b=1}^{B} P\left(p_{jl}^{b} \leq \hat{q}_{n,b}^{-1}\alpha\delta\Big|H_{0,jl} : \beta_{jl}^0 = 0 \text{ is true}, \mathcal{D}_{1n}^{1:B}\right) \\ &\leq \frac{1}{B}\sum_{b=1}^{B} \hat{q}_{n,b}^{-1}\alpha\delta. \qquad (53) \end{aligned}$$

Using (53) in (52) it follows that

$$P\Big\{\mathcal{E}(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset\Big\} \leq \frac{\alpha}{B}\sum_{b=1}^{B} E_0\left[\sum_{(j,l)\in\mathcal{S}_{B,n}} \hat{q}_{n,b}^{-1}\right] = E_0\left[\frac{\alpha}{B}\sum_{b=1}^{B} r_n\hat{q}_{n,b}^{-1}\right].$$

By taking the limit superior as $n \to \infty$ on both sides and using $r_n \to q_0$ and $\hat{q}_{n,b} \to q_0$ in probability as $n \to \infty$, it follows using the boundedness of $r_n\hat{q}_{n,b}^{-1}$ that

$$\limsup_{n\to\infty} \mathrm{P}\left(\mathcal{E}(\delta; \mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset\right) \leq \alpha.$$

We turn to the proof of Part **(ii)**. Note that,

$$\begin{aligned} P\Big\{\mathcal{E}^*(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset\Big\} &= E_0\Big\{\mathbf{1}_{[\mathcal{E}^*(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B})\neq\emptyset]}\Big\} = E_0\left[E_0\Big\{\mathbf{1}_{[\mathcal{E}^*(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B})\neq\emptyset]}\Big|\mathcal{D}_{1n}^{1:B}\Big\}\right] \\ &\leq E_0\left[\sum_{(j,l)\in\mathcal{S}_{B,n}} P\Big\{Q_{jl}^* \leq \alpha \text{ and } \mathrm{H}_{0,jl} : \beta_{jl}^0 = 0 \text{ is true}\Big|\mathcal{D}_{1n}^{1:B}\Big\}\right] \\ &\leq E_0\left[\sum_{(j,l)\in\mathcal{S}_{B,n}} P\Big\{\inf_{\delta\in(\delta_{min},1)} Q_{jl}(\delta) \leq \eta \text{ and } \mathrm{H}_{0,jl} : \beta_{jl}^0 = 0 \text{ is true}\Big|\mathcal{D}_{1n}^{1:B}\Big\}\right], \end{aligned}$$

where $\eta = \alpha(1 - \log\delta_{\min})^{-1}$. Using (48)–(50) we obtain

$$P\Big\{\mathcal{E}^*(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset\Big\} \leq E_0\left[\sum_{(j,l)\in\mathcal{S}_{B,n}} \frac{1}{B}\sum_{b=1}^{B} \eta(1 - \log\delta_{\min})\hat{q}_{n,b}^{-1}\right] \leq E_0\Big[\frac{\alpha r_n}{\min_{1\leq b\leq B}\hat{q}_{n,b}}\Big].$$

Taking the limit superior as $n \to \infty$ on both sides and using $r_n \to q_0$ and $\hat{q}_{n,b} \to q_0$ in probability as $n \to \infty$, it follows using the boundedness of $r_n \hat{q}_{n,b}^{-1}$ that

$$\limsup_{n\to\infty} \mathrm{P}\bigg(\mathcal{E}^*(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset \bigg) \leq \alpha.$$

The proof of Part **(iii)** uses Proposition 1 of Vovk (2012). Note that,

$$\begin{aligned} P\bigg\{ \mathcal{E}^{**}(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset \bigg\} &= E_0 \bigg\{ \mathbf{1}_{[\mathcal{E}^{**}(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset]} \bigg\} = E_0 \bigg[E_0 \bigg\{ \mathbf{1}_{[\mathcal{E}^{**}(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset]} \bigg| \mathcal{D}_{1n}^{1:B} \bigg\} \bigg] \\ &\leq E_0 \left[\sum_{(j,l) \in \mathcal{S}_{B,n}} P\bigg\{ \bar{Q}_{jl}^* \leq \alpha \text{ and } \mathrm{H}_{0,jl} : \beta_{jl}^0 = 0 \text{ is true} \bigg| \mathcal{D}_{1n}^{1:B} \bigg\} \right] \\ &\leq E_0 \left[\sum_{(j,l) \in \widehat{\mathcal{S}}_{B,n}} \frac{\alpha}{\min_{1 \leq b \leq B} \hat{q}_{n,b}} \right] \leq E_0 \left[\frac{\alpha r_n}{\min_{1 \leq b \leq B} \hat{q}_{n,b}} \right]. \end{aligned}$$

Finally, taking the limit superior as $n \to \infty$ on both sides and using $r_n \to q_0$ and $\hat{q}_{n,b} \to q_0$ in probability as $n \to \infty$, it follows using the boundedness of $r_n \hat{q}_{n,b}^{-1}$ that

$$\limsup_{n\to\infty} \mathrm{P}\bigg(\mathcal{E}^{**}(\delta; \mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \neq \emptyset \bigg) \leq \alpha. \qquad \blacksquare$$

References

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

Chanda, S. (1954). A notes on the consistency and maxima of the roots of the likelihood equations. *Biometrika 41*, 56–61.

Chen, J. (2016). Consistency of the MLE under mixture models. *Statistical Science 32*, 47–63.

Chen, J., X. Tan, and R. Zhang (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica 18*, 443–465.

Conlon, E., X. Liu, J. Lieb, and J. Liu (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences 100*, 3339–3344.

Danilov, D. and J. R. Magnus (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics 122*, 27–46.

Dijkstra, T. and J. Veldkamp (1988). Data-driven selection of regressors and the bootstrap. In *On Model Uncertainty and Its Statistical Implications*, pp. 17–38. Springer.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association 109*, 991–1007.

Hathaway, R. J. (1985). A constraint formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics 13*, 795–800.

Kabaila, P. (1995). The effect of model selection on confidence regions and prediction regions. *Econometric Theory 11*, 537–549.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical association 102*, 1025–1038.

Konishi, S. and G. Kitagawa (2008). *Information Criteria and statistical modeling.* Springer.

Leeb, H. and B. M. Pötscher (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory 19*, 100–142.

Lockhart, R., J. Taylor, R. J. Tibshirani, and R. Tibshirani (2014). A significance test for the lasso. *The Annals of Statistics 42*, 413.

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

Meinshausen, N., L. Meier, and P. Bühlmann (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association 104*, 1671–1681.

Redner, R. A. and H. F. Walker (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review 26*, 195–239.

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

Van de Geer, S., P. Bühlmann, Y. Ritov, R. Dezeure, et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*, 1166–1202.

Vovk, V. (2012). Combining p-values via averaging. *ArXiv Preprint, arXiv:1212.4966*.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*, 307–333.

Wasserman, L. and K. Roeder (2009). High dimensional variable selection. *The Annals of Statistics 37*, 2178–2201.

Wedel, M. and W. A. Kamakura (2000). *Market Segmentation: Conceptual and Methodological Foundations (2nd ed.)*. Boston: Kluwer Academic Publishers.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*, 217–242.

Zhang, Y., R. Li, and C.-L. Tsai (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association 105*, 312–323.

Table 1: Simulation results for the Gaussian FMR model.

			E(TP)			E(FP)			EFWER		
SNR	$K \times d$	Mixture	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO
25.8		Com_1	5.00	5.00	5.00	.000	.050	.000	.000	.045	.000
	60	Com_2	4.00	4.00	4.00	.000	.015	.000	.000	.015	.000
		Both	9.00	9.00	9.00	.000	.065	.000	.000	.060	.000
		Com_1	5.00	5.00	5.00	.000	.095	.000	.000	.075	.000
	100	Com_2	4.00	4.00	4.00	.000	.030	.000	.000	.025	.000
		Both	9.00	9.00	9.00	.000	.125	.000	.000	.100	.000
		Com_1	5.00	5.00	5.00	.000	.115	.005	.000	.095	.005
	140	Com_2	4.00	4.00	4.00	.000	.030	.000	.000	.025	.000
		Both	9.00	9.00	9.00	.000	.145	.005	.000	.120	.005
6.45		Com_1	4.84	5.00	5.00	.000	.520	.335	.000	.345	.275
	60	Com_2	4.00	4.00	4.00	.000	.355	.175	.000	.295	.165
		Both	8.84	9.00	9.00	.000	.875	.510	.000	.510	.395
		Com_1	4.80	5.00	5.00	.000	1.04	.700	.000	.555	.495
	100	Com_2	4.00	4.00	4.00	.000	.680	.320	.000	.440	.270
		Both	8.80	9.00	9.00	.000	1.72	1.02	.000	.715	.620
		Com_1	4.71	5.00	4.95	.000	1.58	1.27	.000	.685	.660
	140	Com_2	4.00	4.00	4.00	.000	.975	.615	.000	.505	.420
		Both	8.71	9.00	8.94	.000	2.55	1.89	.000	.780	.795
2.87		Com_1	3.27	4.82	4.89	.000	.770	.955	.000	.485	.575
	60	Com_2	3.63	4.00	4.00	.000	.510	.580	.000	.335	.405
		Both	6.90	8.82	8.89	.000	1.28	1.54	.000	.620	.740
		Com_1	3.27	4.76	4.75	.000	1.21	1.56	.000	.560	.740
	100	Com_2	3.63	3.99	3.97	.000	.915	.940	.000	.480	.540
		Both	6.89	8.75	8.72	.000	2.12	2.50	.000	.695	.850
		Com_1	3.44	4.74	4.56	.000	1.61	2.65	.000	.655	.875
	140	Com_2	3.60	3.99	3.91	.000	1.24	1.51	.000	.565	.695
		Both	7.03	8.73	8.47	.000	2.85	4.16	.000	.770	.960
1.03		Com_1	1.43	3.93	4.08	.000	4.67	3.85	.000	.970	.955
	60	Com_2	1.33	3.46	3.46	.000	3.83	2.79	.000	.980	.935
		Both	2.76	7.39	7.53	.000	8.49	6.64	.000	.995	.995
		Com_1	1.55	3.45	3.91	.000	4.41	6.42	.000	.950	.995
	100	Com_2	1.58	3.09	3.36	.000	3.29	5.09	.000	.975	.990
		Both	3.13	6.53	7.27	.000	7.70	11.5	.000	1.00	1.00
		Com_1	1.90	3.36	3.78	.010	5.69	8.39	.010	.985	1.00
	140	Com_2	1.97	3.04	3.28	.000	4.76	7.29	.000	.985	1.00
		Both	3.86	6.40	7.06	.010	10.45	15.7	.010	1.00	1.00
0.72		Com_1	.905	3.67	3.77	.000	7.09	5.65	.000	1.00	1.00
	60	Com_2	.810	3.19	3.21	.000	6.14	4.75	.000	.995	1.00
		Both	1.72	6.85	6.97	.000	13.22	10.4	.000	1.00	1.00
		Com_1	1.12	3.14	3.61	.000	6.94	9.12	.000	1.00	1.00
	100	Com_2	.970	2.82	3.16	.000	5.78	8.19	.000	.990	1.00
		Both	2.09	5.96	6.76	.000	12.7	17.3	.000	1.00	1.00
		Com_1	1.47	2.95	3.41	.010	9.23	12.7	.010	1.00	1.00
	140	Com_2	1.40	2.67	3.04	.005	7.98	11.3	.005	1.00	1.00
		Both	2.87	5.62	6.45	.015	17.2	24.0	.015	1.00	1.00

Supplementary Material:

Hypothesis Testing in Finite Mixture of Regressions: Sparsity and Model Selection Uncertainty

Abbas Khalili[1]*, Anand N. Vidyashankar[2]

[1]Department of Mathematics and Statistics, McGill University

[2]Department of Statistics, Volgeneau School of Engineering,

George Mason University

May 8, 2018

1 PROOF OF THEOREM 2

Consider the penalized log-likelihood function $\tilde{l}_n(\boldsymbol{\Psi}_A; \widehat{\boldsymbol{\gamma}}_n)$ as function of the active parameters $\boldsymbol{\Psi}_A = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\beta}_0, \boldsymbol{B}_1)$ where $\widehat{\boldsymbol{\gamma}}_n$ is the tuning parameter selected by the BIC. We divide the proof into two steps.

Step 1: *Existence of a root-n local maximizer.* Let $r_n = n^{-1/2}(1 + b_n)$, where b_n is given in the regularity condition (RCP2). It suffices to show that for any $\epsilon > 0$, there exists a large number C_ϵ such that

$$\lim_{n\to\infty} P\left\{ \sup_{\|\boldsymbol{u}\|=C_\epsilon} p\ell_n(\boldsymbol{\Psi}_0 + r_n\boldsymbol{u}; \widehat{\boldsymbol{\gamma}}_n) < p\ell_n(\boldsymbol{\Psi}_0; \widehat{\boldsymbol{\gamma}}_n) \right\} \geq 1 - \epsilon$$

where $\boldsymbol{\Psi}_0$ is the true vector of parameters.

*Corresponding author.

Recall the quantities

$$\begin{aligned}
\hat{a}_n &= \max_{j,l} \left\{ p_{nj}(\beta^0_{jl}; \hat{\gamma}_{nj})/\sqrt{n} : \beta^0_{jl} \neq 0 \right\} \\
\hat{b}_n &= \max_{j,l} \left\{ p'_{nj}(\beta^0_{jl}; \hat{\gamma}_{nj})/\sqrt{n} : \beta^0_{jl} \neq 0 \right\} \\
\hat{c}_n &= \max_{j,l} \left\{ p''_{nj}(\beta^0_{jl}; \hat{\gamma}_{nj})/n : \beta^0_{jl} \neq 0 \right\},
\end{aligned}$$

which are random sequences due to the randomness of $\widehat{\boldsymbol{\gamma}}_n$. The non-random version of these quantities are given in (31)-(33) of the main paper. Consider the difference

$$\begin{aligned}
D_n(\boldsymbol{u}) &= p\ell_n(\boldsymbol{\Psi}_0 + r_n \boldsymbol{u}; \widehat{\boldsymbol{\gamma}}_n) - p\ell_n(\boldsymbol{\Psi}_0; \widehat{\boldsymbol{\gamma}}_n) \\
&= [\ell_n(\boldsymbol{\Psi}_0 + r_n \boldsymbol{u}) - \ell_n(\boldsymbol{\Psi}_0)] - [\boldsymbol{p}_n(\boldsymbol{\Psi}_0 + r_n \boldsymbol{u}; \widehat{\boldsymbol{\gamma}}_n) - \boldsymbol{p}_n(\boldsymbol{\Psi}_0; \widehat{\boldsymbol{\gamma}}_n)]
\end{aligned}$$

By the regularity conditions (RC1)-(RC5), and a second-order Taylor's expansion, for large n, we have that

$$\begin{aligned}
\ell_n(\boldsymbol{\Psi}_0 + r_n \boldsymbol{u}) - \ell_n(\boldsymbol{\Psi}_0) &= \sqrt{n} r_n \left\{ l'_n(\boldsymbol{\Psi}_0)/\sqrt{n} \right\}^\top \boldsymbol{u} - \frac{n r_n^2}{2} \left\{ \boldsymbol{u}^\top \boldsymbol{I}_1 \left(\boldsymbol{\Psi}_0 \right) \boldsymbol{u} \right\} (1 + o_p(1)) \\
&= O_p(1)(1 + b_n)\|\boldsymbol{u}\| - \frac{(1 + b_n^2)}{2} \left\{ \boldsymbol{u}^\top \boldsymbol{I}_1 \left(\boldsymbol{\Psi}_0 \right) \boldsymbol{u} \right\} (1 + o_p(1)).
\end{aligned}$$

Now turning to the difference in the penalty terms, by a second-order Taylor's expansion we have that

$$\begin{aligned}
|\boldsymbol{p}_n(\boldsymbol{\Psi}_0 + r_n \boldsymbol{u}; \widehat{\boldsymbol{\gamma}}_n) - \boldsymbol{p}_n(\boldsymbol{\Psi}_0; \widehat{\boldsymbol{\gamma}}_n)| &\leq (\max_{1 \leq j \leq K} q^0_j)(\sqrt{n} r_n \hat{b}_n)\|\boldsymbol{u}\| + \frac{1}{2}(n r_n^2 \hat{c}_n)\|\boldsymbol{u}\|^2 + \sqrt{K}(\sqrt{n} r_n \hat{a}_n)\|\boldsymbol{u}\| \\
&= q^0_m \hat{b}_n (1 + b_n)\|\boldsymbol{u}\| + \frac{1}{2}(1 + b_n)^2 \hat{c}_n \|\boldsymbol{u}\|^2 + \sqrt{K}(1 + b_n)\hat{a}_n \|\boldsymbol{u}\|,
\end{aligned}$$

where $q^0_m = \max_{1 \leq j \leq K} q^0_j$.

Now by comparing the orders of the above two expressions and using (RCP2) it follows that, for large n,

$$D_n(\boldsymbol{u}) \leq -\frac{(1 + b_n^2)}{2} \left\{ \boldsymbol{u}^\top \boldsymbol{I}_1 \left(\boldsymbol{\Psi}_0 \right) \boldsymbol{u} \right\} (1 + o_p(1)) < 0.$$

This completes the proof of **Step 1**.

Step 2: *Asymptotic normality.* By (RCP2) $b_n = O(1)$, and hence using **Step 1**, there exists a local maximizer $\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n)$ of the penalized log-likelihood $p\ell_n(\boldsymbol{\Psi}_A; \widehat{\boldsymbol{\gamma}}_n)$ that is root-n consistent estimator of $\boldsymbol{\Psi}_0$ and it satisfies the normal equation

$$\left.\frac{\partial p\ell_n(\boldsymbol{\Psi}_A;\widehat{\boldsymbol{\gamma}}_n)}{\partial \boldsymbol{\Psi}_A}\right|_{\boldsymbol{\Psi}_A=\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n)} \equiv p\ell_n'(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n);\widehat{\boldsymbol{\gamma}}_n) = \ell_n'(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n)) - \boldsymbol{p}_n'(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n)) = \mathbf{0}.$$

Now using the first-order Taylor's expansion it follows that

$$\begin{aligned}\ell_n'(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n)) &= \ell_n'(\boldsymbol{\Psi}_0) + \{\ell_n''(\boldsymbol{\Psi}_0) + o_p(n)\}\left(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n) - \boldsymbol{\Psi}_0\right) \quad \text{and}\\ \boldsymbol{p}_n'(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n)) &= \boldsymbol{p}_n'(\boldsymbol{\Psi}_0;\widehat{\boldsymbol{\gamma}}_n) + \{\boldsymbol{p}_n''(\boldsymbol{\Psi}_0;\widehat{\boldsymbol{\gamma}}_n) + o_p(n)\}\left(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n) - \boldsymbol{\Psi}_0\right).\end{aligned}$$

Therefore,

$$\sqrt{n}\left\{\left[\frac{1}{n}\ell_n''(\boldsymbol{\Psi}_0) - \frac{1}{n}\boldsymbol{p}_n''(\boldsymbol{\Psi}_0;\widehat{\boldsymbol{\gamma}}_n) + o_p(1)\right]\left(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n) - \boldsymbol{\Psi}_0\right) - \frac{1}{n}\boldsymbol{p}_n'(\boldsymbol{\Psi}_0;\widehat{\boldsymbol{\gamma}}_n)\right\} = -\frac{\ell_n'(\boldsymbol{\Psi}_0)}{\sqrt{n}}.$$

By the regularity conditions (RC1)-(RC5) it follows that as $n \to \infty$,

$$\frac{1}{n}\ell_n''(\boldsymbol{\Psi}_0) \xrightarrow{p} \boldsymbol{I}_1(\boldsymbol{\Psi}_0) \quad , \quad -\frac{\ell_n'(\boldsymbol{\Psi}_0)}{\sqrt{n}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}_1(\boldsymbol{\Psi}_0)) \tag{A.1}$$

Finally, using Slutsky's theorem and (A.1) it follows that

$$\sqrt{n}\left\{\left[\boldsymbol{I}_1\left(\boldsymbol{\Psi}_0\right) - \frac{\boldsymbol{p}_n''(\boldsymbol{\Psi}_0;\widehat{\boldsymbol{\gamma}}_n)}{n}\right]\left(\widehat{\boldsymbol{\Psi}}_{1,n}(\widehat{\boldsymbol{\gamma}}_n) - \boldsymbol{\Psi}_0\right) + \frac{\boldsymbol{p}_n'(\boldsymbol{\Psi}_0;\widehat{\boldsymbol{\gamma}}_n)}{n}\right\} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_1\left(\boldsymbol{\Psi}_0\right)\right).$$

This completes the proof of Theorem 2. ■

2 ADDITIONAL NUMERICAL RESULTS

Binomial FMR. Given $\boldsymbol{x}_i$, the response Y_i is generated from the mixture

$$\pi \ \text{Bin}(15, p(\boldsymbol{x}_i;\beta_{10},\boldsymbol{\beta}_1)) + (1-\pi)\ \text{Bin}(15, p(\boldsymbol{x}_i;\beta_{20},\boldsymbol{\beta}_2))$$

with $\pi = .45$, and the probability of success within each component is modelled as $\text{logit}\{p(\boldsymbol{x}_i;\beta_{j0},\boldsymbol{\beta}_j)\} = \beta_{j0} + \boldsymbol{x}_i^\top\boldsymbol{\beta}_j,\ j = 1,2$. The sparse vector of regression coefficients are

$$\begin{aligned}\mathbf{M1}: \quad \boldsymbol{\beta}_1 &= (1.0, 0.0, -1.5, 1.8, 1.5, 0.0, 0.0, 1.2, 0.0, \ldots, 0.0)/\sqrt{2} \text{ and}\\ \boldsymbol{\beta}_2 &= (-1.0, 1.0, 0.0, 0.0, -1.5, 1.3, 0.0, -1.3, 1.4, 0.0, \ldots, 0.0)/\sqrt{2}.\end{aligned}$$

These vectors contain $q_1 = 5$ and $q_2 = 6$ non-zero regression coefficients respectively. We also consider a second Binomial model, that we call **M2**, by setting $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1/\sqrt{2}$ and $\boldsymbol{\beta}_2^* = \boldsymbol{\beta}_2/\sqrt{2}$. Model **M2** has a sparsity structure similar to **M1** but with a weaker signal.

From Table S1 we see that the EFWER of Msplit is controlled at 5% for both models **M1** and **M2** while the corresponding values for ADLASSO vary between 5.5% and 35.5% and that for SCAD, vary between 8% and 20%. It is worthwhile to notice that in **M1**, which has a stronger signal, the EFWER of ADLASSO and SCAD are less than 5% in one of the components (Com$_1$ or Com$_2$) but not for the entire mixture model (Both). In terms of E(FP), the performance of all three methods is similar and very good; the Msplit leads the performance. Finally, concerning E(TP) all three methods identify most of the true non-zero regression coefficients in **M1**, whereas in **M2** for dimension $Kd = 140$, on average, Msplit misses one additional true non-zero coefficient relative to ADLASSO and SCAD.

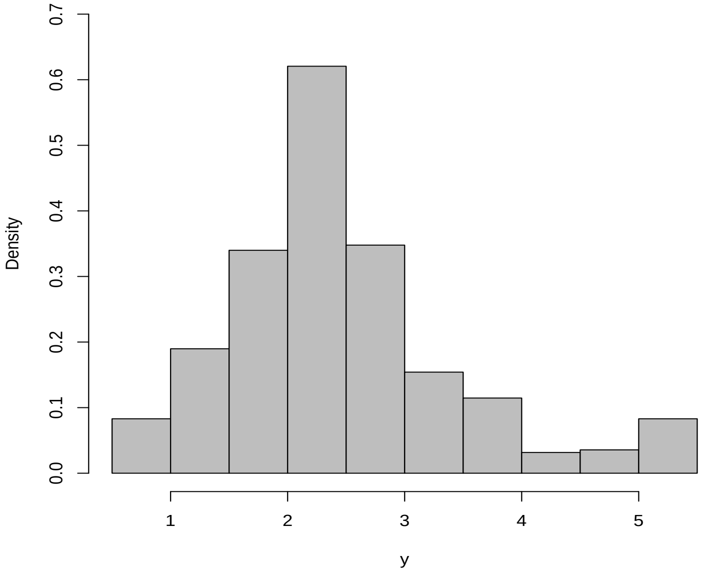

Figure 1: Histogram of $y =$ MEDV/sd(MEDV) in Example 3 of the main paper.

Table S1: Simulation results for the Binomial FMR model using Theorem 1.

			E(TP)			E(FP)			EFWER		
Model	$K \times d$	Mixture	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO
M1	60	Com$_1$	4.99	5.00	4.99	.000	.130	.050	.000	.040	.035
		Com$_2$	5.92	5.97	6.00	.000	.075	.030	.000	.040	.020
		Both	10.9	11.0	11.0	.000	.205	.080	.000	.080	.055
	100	Com$_1$	4.98	5.00	4.99	.000	.430	.270	.000	.070	.145
		Com$_2$	5.90	5.93	6.00	.000	.060	.060	.000	.025	.060
		Both	10.9	10.9	11.0	.000	.490	.330	.000	.095	.185
	140	Com$_1$	4.99	4.98	4.97	.010	.515	.685	.010	.080	.305
		Com$_2$	5.87	5.92	5.96	.000	.015	.185	.000	.015	.105
		Both	10.8	10.9	10.9	.010	.530	.870	.010	.095	.355
M2	60	Com$_1$	4.72	4.92	4.97	.005	.120	.045	.005	.055	.030
		Com$_2$	4.66	5.51	5.80	.005	.165	.045	.005	.085	.045
		Both	9.38	10.4	10.8	.010	.285	.090	.010	.135	.070
	100	Com$_1$	4.63	4.74	4.87	.010	.525	.190	.010	.100	.120
		Com$_2$	4.50	4.95	5.52	.005	.330	.140	.005	.105	.100
		Both	9.12	9.68	10.4	.015	.855	.330	.015	.200	.210
	140	Com$_1$	4.61	4.73	4.76	.040	.440	.390	.040	.085	.180
		Com$_2$	4.34	4.80	5.31	.005	.480	.275	.005	.100	.140
		Both	8.95	9.53	10.1	.045	.920	.655	.045	.170	.290

Table S2: Comparison of the aggregation methods for the Gaussian FMR model.

			E(TP)		E(FP)		EFWER	
SNR	Kd	Mixture	Quantile	Average	Quantile	Average	Quantile	Average
25.8	60	Com$_1$	5.00	4.94	.000	.000	.000	.000
		Com$_2$	4.00	4.00	.000	.000	.000	.000
		Both	9.00	8.94	.000	.000	.000	.000
	100	Com$_1$	5.00	5.00	.000	.000	.000	.000
		Com$_2$	4.00	4.00	.000	.000	.000	.000
		Both	9.00	9.00	.000	.000	.000	.000
	140	Com$_1$	5.00	4.99	.000	.000	.000	.000
		Com$_2$	4.00	4.00	.000	.000	.000	.000
		Both	9.00	8.99	.000	.000	.000	.000
6.45	60	Com$_1$	4.84	3.37	.000	.000	.000	.000
		Com$_2$	4.00	3.84	.000	.000	.000	.000
		Both	8.84	7.21	.000	.000	.000	.000
	100	Com$_1$	4.80	3.09	.000	.000	.000	.000
		Com$_2$	4.00	3.86	.000	.000	.000	.000
		Both	8.80	6.95	.000	.000	.000	.000
	140	Com$_1$	4.71	1.97	.000	.000	.000	.000
		Com$_2$	4.00	3.34	.000	.000	.000	.000
		Both	8.71	5.31	.000	.000	.000	.000

Table S3: Comparison of the aggregation methods for the Binomial FMR model.

			E(TP)		E(FP)		EFWER	
Model	$K \times d$	Mixture	Quantile	Average	Quantile	Average	Quantile	Average
M1	60	Com_1	4.99	4.49	.000	.000	.000	.000
		Com_2	5.92	4.72	.000	.000	.000	.000
		Both	10.9	9.21	.000	.000	.000	.000
	100	Com_1	4.98	4.24	.000	.000	.000	.000
		Com_2	5.90	4.07	.000	.000	.000	.000
		Both	10.9	8.31	.000	.000	.000	.000
	140	Com_1	4.99	3.77	.010	.000	.010	.000
		Com_2	5.87	3.56	.000	.000	.000	.000
		Both	10.8	7.33	.010	.000	.010	.000

Table S4: Hypothesis testing results for the Gaussian FMR model using Theorem 2.

			E(TP)			E(FP)			Pr(FP > 0)		
SNR	Kd	Mixture	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO
25.8	60	Com$_1$	5.00	5.00	5.00	.000	.000	.000	.000	.000	.000
		Com$_2$	4.00	4.00	4.00	.000	.005	.000	.000	.005	.000
		Both	9.00	9.00	9.00	.000	.005	.000	.000	.005	.000
	100	Com$_1$	5.00	5.00	5.00	.000	.000	.000	.000	.000	.000
		Com$_2$	4.00	4.00	4.00	.000	.020	.000	.000	.015	.000
		Both	9.00	9.00	9.00	.000	.020	.000	.000	.015	.000
	140	Com$_1$	5.00	5.00	5.00	.000	.010	.005	.000	.010	.005
		Com$_2$	4.00	4.00	4.00	.000	.015	.000	.000	.015	.000
		Both	9.00	9.00	9.00	.000	.025	.005	.000	.025	.005
6.45	60	Com$_1$	4.84	4.96	5.00	.000	.170	.335	.000	.150	.275
		Com$_2$	4.00	4.00	4.00	.000	.230	.175	.000	.200	.165
		Both	8.84	8.96	9.00	.000	.400	.510	.000	.325	.350
	100	Com$_1$	4.80	4.96	5.00	.000	.395	.700	.000	.330	.295
		Com$_2$	4.00	4.00	4.00	.000	.490	.320	.000	.370	.270
		Both	8.80	8.96	9.00	.000	.885	1.02	.000	.580	.620
	140	Com$_1$	4.71	4.95	4.95	.000	.680	1.26	.000	.485	.655
		Com$_2$	4.00	4.00	4.00	.000	.630	.615	.000	.435	.420
		Both	8.71	8.95	8.94	.000	1.31	1.87	.000	.670	.790
2.87	60	Com$_1$	3.27	4.15	4.89	.000	.280	.955	.000	.240	.575
		Com$_2$	3.63	3.96	4.00	.000	.295	.580	.000	.235	.405
		Both	6.90	8.11	8.89	.000	.575	1.54	.000	.395	.740
	100	Com$_1$	3.27	4.19	4.75	.000	.650	1.56	.000	.445	.740
		Com$_2$	3.63	3.97	3.97	.000	.635	.940	.000	.450	.540
		Both	6.89	8.16	8.72	.000	1.29	2.50	.000	.625	.850
	140	Com$_1$	3.44	4.20	4.56	.000	.950	2.65	.000	.545	.875
		Com$_2$	3.60	3.97	3.91	.000	1.03	1.51	.000	.545	.695
		Both	7.03	8.17	8.47	.000	1.98	4.16	.000	.725	.960
1.03	60	Com$_1$	1.43	2.23	4.08	.000	1.11	3.85	.000	.690	.955
		Com$_2$	1.33	2.56	3.46	.000	1.23	2.79	.000	.715	.935
		Both	2.76	4.79	7.53	.000	2.33	6.64	.000	.880	.995
	100	Com$_1$	1.55	2.32	3.91	.000	2.35	6.42	.000	.885	.995
		Com$_2$	1.58	2.55	3.36	.000	2.29	5.09	.000	.925	.990
		Both	3.13	4.86	7.27	.000	4.63	11.5	.000	1.00	1.00
	140	Com$_1$	1.90	2.41	3.78	.010	3.20	8.39	.010	.950	1.00
		Com$_2$	1.97	2.48	3.28	.000	3.21	7.29	.000	.945	1.00
		Both	3.86	4.89	7.06	.010	6.40	15.7	.010	1.00	1.00
0.72	60	Com$_1$	.905	1.79	3.77	.000	1.60	5.65	.000	.825	1.00
		Com$_2$	.810	1.82	3.21	.000	1.68	4.75	.000	.805	1.00
		Both	1.72	3.61	6.97	.000	3.28	10.4	.000	.955	1.00
	100	Com$_1$	1.12	1.97	3.61	.000	3.08	9.12	.000	.940	1.00
		Com$_2$	.970	1.75	3.16	.000	3.09	8.19	.000	.965	1.00
		Both	2.09	3.72	6.76	.000	6.17	17.3	.000	1.00	1.00
	140	Com$_1$	1.47	2.04	3.41	.010	4.52	12.7	.010	.970	1.00
		Com$_2$	1.40	1.90	3.04	.005	4.39	11.3	.005	.995	1.00
		Both	2.87	3.94	6.45	.015	8.91	24.0	.015	1.00	1.00

Table S5: Hypothesis testing results for the Binomial FMR model using Theorem 2.

			E(TP)			E(FP)			Pr(FP > 0)		
Model	$K \times d$	Mixture	Msplit	HSCAD	HADLASSO	Msplit	HSCAD	HADLASSO	Msplit	HSCAD	HADLASSO
M1	60	Com_1	4.99	4.97	4.97	.000	.010	.010	.000	.005	.005
		Com_2	5.92	5.94	5.98	.000	.010	.005	.000	.010	.005
		Both	10.9	10.9	10.9	.000	.020	.015	.000	.015	.010
	100	Com_1	4.98	4.97	4.98	.000	.115	.110	.000	.035	.065
		Com_2	5.90	5.91	5.98	.000	.030	.040	.000	.010	.040
		Both	10.9	10.9	11.0	.000	.145	.150	.000	.045	.095
	140	Com_1	4.99	4.96	4.95	.010	.205	.280	.010	.035	.175
		Com_2	5.87	5.90	5.93	.000	.010	.120	.000	.010	.070
		Both	10.8	10.9	10.9	.010	.215	.400	.010	.045	.230
M2	60	Com_1	4.72	4.67	4.80	.005	.010	.010	.005	.010	.010
		Com_2	4.66	5.06	5.31	.005	.115	.005	.005	.040	.005
		Both	9.38	9.73	10.1	.010	.125	.015	.010	.050	.015
	100	Com_1	4.63	4.60	4.78	.010	.215	085	.010	.075	.045
		Com_2	4.50	4.68	5.23	.005	.175	.060	.005	.075	.060
		Both	9.12	9.28	10.0	.015	.390	.145	.015	.150	.105
	140	Com_1	4.61	4.62	4.68	.040	.215	.215	.040	.070	.100
		Com_2	4.34	4.53	5.03	.005	.335	.180	.005	.075	.125
		Both	8.95	9.14	9.70	.045	.550	.395	.045	.135	.200