HYPOTHESIS TESTING IN FINITE MIXTURE OF REGRESSIONS: SPARSITY AND MODEL SELECTION UNCERTAINTY

Abbas Khalili

Department of Mathematics and Statistics

McGill University, Canada

(Joint work with Anand N. Vidyashankar)
(Department of Statistics, George Mason University)

June 21, 2018
S4D 2018: Research Summer School on Statistics for Data Science, Caen, France

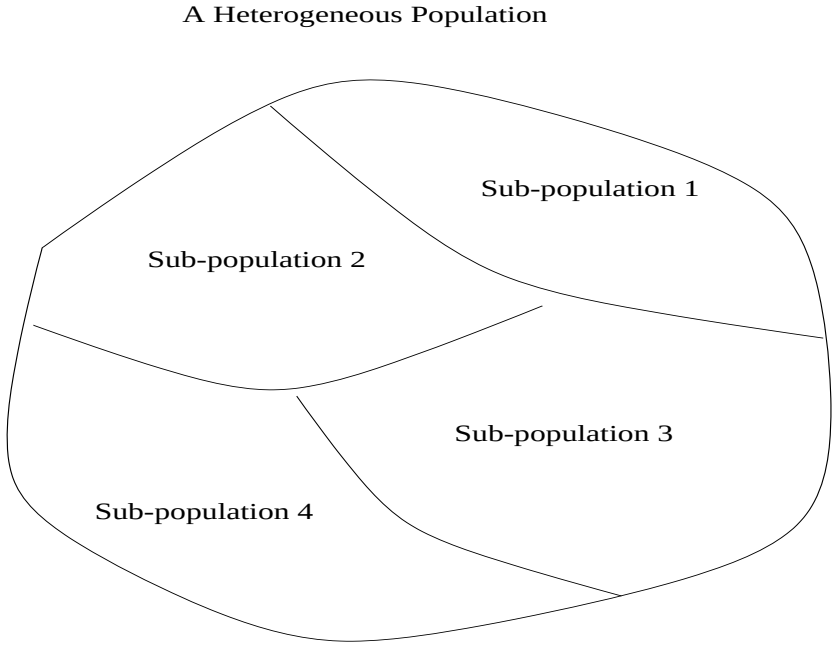
Outline

- Introduction
- Finite mixture of regression (FMR) models
- Sparsity and variable selection methods in FMR models
- Post-selection inference in FRM models
- A method for hypothesis testing after model selection
- Properties of the proposed method: theory and simulations

- Recent advancements in medical and other fields of scientific research has led to the collection of data of unprecedented size and complexity.
- One common statistical problem of interest in such applications is to model a response (or output) variable Y as a function of a small subset of large number of features $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$.
- In regression and classification, this is commonly referred to as variable (feature) selection problem which aims at building a *sparse* regression model or classifier.

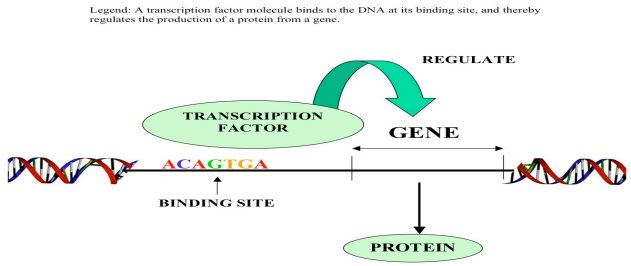
Heterogeneous populations

- The feature selection problem becomes even more complex when the population of interest is made-up of *hidden* sub-populations, i.e. a *heterogeneous* population:



Examples of heterogeneity in regression:

- In *motif discovery problem*, where motif-gene expression data are studied, the set of regulating motifs varies from one group of genes to another.



- In *market segmentation* research, consumers or suppliers rate the quality of products. Markets can be segmented by finding sub-groups with respect to the relationship between rating and the features of a product.

Regression modeling in heterogeneous populations

- When the population under study is heterogeneous:
  1. Unobserved heterogeneity.
  2. Relationship between Y and $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ varies across sub-populations.
  3. Each sub-population calls for its own regression model.
- Finite mixture of regression (FMR) models provide a natural tool to handle 1-3.

- In an FMR model with K components, the conditional density function of Y given $\boldsymbol{x}$ is

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

  with a known link function $\eta_j(\boldsymbol{x}) = \mathcal{L}(\beta_{0j} + \boldsymbol{\beta}_j^\top \boldsymbol{x})$, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd})^\top$, for $j = 1, \ldots, K$.

- The vector of all unknown parameters:

$$\boldsymbol{\Psi} = (\beta_{01}, \boldsymbol{\beta}_1, \ldots, \beta_{0K}, \boldsymbol{\beta}_K, \phi_1, \ldots, \phi_K, \pi_1, \ldots, \pi_K).$$

- In an FMR model with K components, the conditional density function of Y given $\boldsymbol{x}$ is

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

  with a known link function $\eta_j(\boldsymbol{x}) = \mathcal{L}(\beta_{0j} + \boldsymbol{\beta}_j^\top \boldsymbol{x})$, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd})^\top$, for $j = 1, \ldots, K$.

- The vector of all unknown parameters:

$$\boldsymbol{\Psi} = (\beta_{01}, \boldsymbol{\beta}_1, \ldots, \beta_{0K}, \boldsymbol{\beta}_K, \phi_1, \ldots, \phi_K, \pi_1, \ldots, \pi_K).$$

Sparse FMR models

- Often, at the beginning of a study a long list of potential explanatory variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ are available in the data. But not all the x_l's have effect on Y !
- In practice, fitting a large and complex model via MLE is undesirable (estimation problems, interpretation, ...).
- We assume that the FMR underlying the data is SPARSE, i.e. for some $l = 1, 2, \ldots, d$, and $j = 1, 2, \ldots, K$,

$$\beta_{jl} = 0$$

- Thus, when fitting an FMR model to a data set some FEATURE SELECTION decisions need to be made.

Sparse FMR models

- Often, at the beginning of a study a long list of potential explanatory variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ are available in the data. But not all the x_l's have effect on Y !
- In practice, fitting a large and complex model via MLE is undesirable (estimation problems, interpretation, ...).
- We assume that the FMR underlying the data is SPARSE, i.e. for some $l = 1, 2, \ldots, d$, and $j = 1, 2, \ldots, K$,

$$\beta_{jl} = 0$$

- Thus, when fitting an FMR model to a data set some FEATURE SELECTION decisions need to be made.

Sparse FMR models

- Often, at the beginning of a study a long list of potential explanatory variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ are available in the data. But not all the x_l's have effect on Y !
- In practice, fitting a large and complex model via MLE is undesirable (estimation problems, interpretation, ...).
- We assume that the FMR underlying the data is SPARSE, i.e. for some $l = 1, 2, \ldots, d$, and $j = 1, 2, \ldots, K$,

$$\beta_{jl} = 0$$

- Thus, when fitting an FMR model to a data set some FEATURE SELECTION decisions need to be made.

Parameter estimation in FRM models: K is known

- Maximum likelihood is the most popular method of estimation in FMR models. (An alternative would be the generalized method of moments).

- The log-likelihood based on a sample of observations $(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n$, from a K-components FMR model:

$$\ell_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{K} \pi_j \, f(y_i; \eta_j(\boldsymbol{x}_i), \phi_j) \right\}.$$

Maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}$:

$$\widetilde{\boldsymbol{\Psi}}_n = \text{argmax}_{\boldsymbol{\Psi}} \, \ell_n(\boldsymbol{\Psi})$$

- But MLE does not provide a sparse model as postulated.

Parameter estimation in FRM models: K is known

- Maximum likelihood is the most popular method of estimation in FMR models. (An alternative would be the generalized method of moments).

- The log-likelihood based on a sample of observations $(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n$, from a K-components FMR model:

$$\ell_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{K} \pi_j \, f(y_i; \eta_j(\boldsymbol{x}_i), \phi_j) \right\}.$$

Maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}$:

$$\widetilde{\boldsymbol{\Psi}}_n = \text{argmax}_{\boldsymbol{\Psi}} \, \ell_n(\boldsymbol{\Psi})$$

- But MLE does not provide a sparse model as postulated.

Parameter estimation in FRM models: K is known

- Maximum likelihood is the most popular method of estimation in FMR models. (An alternative would be the generalized method of moments).

- The log-likelihood based on a sample of observations $(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n$, from a K-components FMR model:

$$\ell_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{K} \pi_j \, f(y_i; \eta_j(\boldsymbol{x}_i), \phi_j) \right\}.$$

Maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}$:

$$\widetilde{\boldsymbol{\Psi}}_n = \text{argmax}_{\boldsymbol{\Psi}} \, \ell_n(\boldsymbol{\Psi})$$

- But MLE does not provide a sparse model as postulated.

Parameter estimation in FRM models: K is known

- Maximum likelihood is the most popular method of estimation in FMR models. (An alternative would be the generalized method of moments).

- The log-likelihood based on a sample of observations $(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n$, from a K-components FMR model:

$$\ell_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{K} \pi_j \, f(y_i; \eta_j(\boldsymbol{x}_i), \phi_j) \right\}.$$

Maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}$:

$$\widetilde{\boldsymbol{\Psi}}_n = \text{argmax}_{\boldsymbol{\Psi}} \, \ell_n(\boldsymbol{\Psi})$$

- But MLE does not provide a sparse model as postulated.

Estimation and Feature selection in FMR models

1) Estimation and feature selection when (K, d) are small:

- The Bayesian information criterion (BIC):

$$\text{BIC}(M) = \ell_n(\widetilde{\boldsymbol{\Psi}}_{n,M}) - 0.5 \dim(M) \times \log n$$

  for any FMR sub-model $M \in \mathcal{M}$.

- BIC examines $2^{K \times d}$ submodel for selecting the best one.

- Given the true K or a consistent estimator of K, and under STANDARD REGULARITY CONDITIONS, the BIC selects the true sparse FMR model with probability tending to one, as $n \to \infty$. (CONSISTENT MODEL SELECTOR).

- However, the BIC is computationally expensive for large (K, d), and thus alternative methods are required.

1) Estimation and feature selection when (K, d) are small:

- The Bayesian information criterion (BIC):

$$\text{BIC}(M) = \ell_n(\widetilde{\boldsymbol{\Psi}}_{n,M}) - 0.5 \dim(M) \times \log n$$

  for any FMR sub-model $M \in \mathcal{M}$.

- BIC examines $2^{K \times d}$ submodel for selecting the best one.

- Given the true K or a consistent estimator of K, and under STANDARD REGULARITY CONDITIONS, the BIC selects the true sparse FMR model with probability tending to one, as $n \to \infty$. (CONSISTENT MODEL SELECTOR).

- However, the BIC is computationally expensive for large (K, d), and thus alternative methods are required.

1) Estimation and feature selection when (K, d) are small:

- The Bayesian information criterion (BIC):

$$\text{BIC}(M) = \ell_n(\widetilde{\boldsymbol{\Psi}}_{n,M}) - 0.5 \dim(M) \times \log n$$

  for any FMR sub-model $M \in \mathcal{M}$.

- BIC examines $2^{K \times d}$ submodel for selecting the best one.

- Given the true K or a consistent estimator of K, and under STANDARD REGULARITY CONDITIONS, the BIC selects the true sparse FMR model with probability tending to one, as $n \to \infty$. (CONSISTENT MODEL SELECTOR).

- However, the BIC is computationally expensive for large (K, d), and thus alternative methods are required.

2) Estimation and feature selection via regularization when (K, d) are large:

- Motivated by the regularization techniques such as the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), ADAPTIVE LASSO (Zou, 2006), and MCP (Zhang, 2010), one may estimate $\boldsymbol{\Psi}$ using

  A PENALIZED (REGULARIZED) LIKELIHOOD APPROACH

- Simultaneous estimation and feature selection without an exhaustive search of the model space. Thus, computationally very efficient, compared to the BIC.

- Theoretical properties to be discussed soon.

2) Estimation and feature selection via regularization when (K, d) are large:

- Motivated by the regularization techniques such as the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), ADAPTIVE LASSO (Zou, 2006), and MCP (Zhang, 2010), one may estimate $\boldsymbol{\Psi}$ using

  A PENALIZED (REGULARIZED) LIKELIHOOD APPROACH

- Simultaneous estimation and feature selection without an exhaustive search of the model space. Thus, computationally very efficient, compared to the BIC.

- Theoretical properties to be discussed soon.

2) Estimation and feature selection via regularization when (K, d) are large:

- Motivated by the regularization techniques such as the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), ADAPTIVE LASSO (Zou, 2006), and MCP (Zhang, 2010), one may estimate $\boldsymbol{\Psi}$ using

  A PENALIZED (REGULARIZED) LIKELIHOOD APPROACH

- Simultaneous estimation and feature selection without an exhaustive search of the model space. Thus, computationally very efficient, compared to the BIC.

- Theoretical properties to be discussed soon.

The regularization approach:

- Given a tuning parameter λ, the maximum penalized likelihood estimate, $\widehat{\boldsymbol{\Psi}}_n(\lambda)$, of $\boldsymbol{\Psi}$ is obtained by maximizing:

$$\tilde{\ell}_n(\boldsymbol{\Psi}; \lambda) = \ell_n(\boldsymbol{\Psi}) - \sum_{k=1}^{K} \pi_k \sum_{j=1}^{p} r_n(\beta_{kj}; \lambda)$$

- Examples of $r_n(\beta; \lambda)$: LASSO, ADLASSO, SCAD, MCP.

- SELECTION CONSISTENCY & ORACLE PROPERTIES studied:

  Hui, Warton & Foster (2015); Städler, Bühlmann & van de Geer (2010); Khalili & Chen (2007), Khalili (2010), Khalili & Lin (2013).

- In practice, a data-driven choice of the tuning parameter λ is required. Khalili and Vidyashankar (2018) show that by choosing λ based on BIC, say $\hat{\lambda}_n$, the regularized estimator $\widehat{\boldsymbol{\Psi}}_n(\hat{\lambda}_n)$ has the selection consistency property, and

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

STATISTICAL INFERENCE AFTER VARIABLE SELECTION

Statistical inference after variable selection

- How about statistical inference such as testing hypotheses for regression coefficients of a selected model?
- Specifically, what statistical guarantees can be given to the regression coefficients of a final selected model?
- This is called

POST-SELECTION INFERENCE

Statistical inference after variable selection

- How about statistical inference such as testing hypotheses for regression coefficients of a selected model?
- Specifically, what statistical guarantees can be given to the regression coefficients of a final selected model?
- This is called

POST-SELECTION INFERENCE

Statistical inference after variable selection

- How about statistical inference such as testing hypotheses for regression coefficients of a selected model?
- Specifically, what statistical guarantees can be given to the regression coefficients of a final selected model?
- This is called

POST-SELECTION INFERENCE

Motivation towards problems investigated in this talk

- In functional genomics, 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.

- In market segmentation, a goal is to identify subgroups of consumers to target products and services for each segment separately. Of interest is to evaluate the statistical significance of the attributes between and within segments of the market which is important for the industry (Wedel and Kamakura, 2000).

- Beyond the works of Redner & Walker (1984) and Chen (2017), inferential aspects of FMRs are largely unknown.

Motivation towards problems investigated in this talk

- In functional genomics, 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.

- In market segmentation, a goal is to identify subgroups of consumers to target products and services for each segment separately. Of interest is to evaluate the statistical significance of the attributes between and within segments of the market which is important for the industry (Wedel and Kamakura, 2000).

- Beyond the works of Redner & Walker (1984) and Chen (2017), inferential aspects of FMRs are largely unknown.

Post-selection inference

- While SPARSIFICATION is useful in obtaining parsimonious models, current methods 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 a model whose active covariate set is chosen using a data-driven method.

Post-selection inference

- While SPARSIFICATION is useful in obtaining parsimonious models, current methods 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 a model whose active covariate set is chosen using a data-driven method.

Post-selection inference

- While SPARSIFICATION is useful in obtaining parsimonious models, current methods 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 a model whose active covariate set is chosen using a data-driven method.

Invalid post-selection inference

- Often the selection process is ignored and inference is performed as if there has been no selection involved;

  a standard text book practice which is not statistically valid.

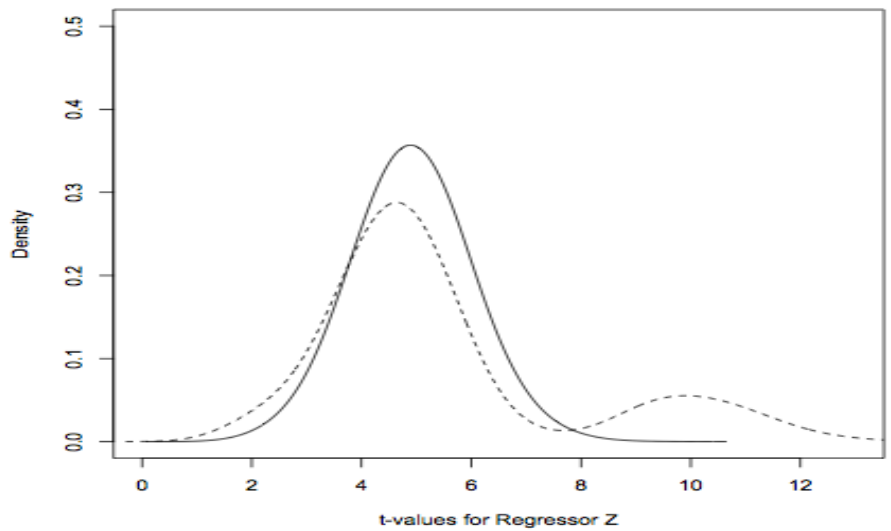
  "a quiet scandal in the statistical community" as phrased by Breiman (1992).

- The reason for invalidation of classical inference is that the randomness induced by a data-driven model selection procedure is not accounted for by classical theory.

  Berk et al. (2013)

Example: effect of the variable selection on inference

- This plot, taken from Berk et al. (2009), depicts the sampling distribution (broken line) of $\hat{\beta}_1/\text{SE}(\hat{\beta}_1)$ after a model selection process in a linear regression model.



- Clearly NOT a *t-student* distribution as it would be in a classical setting.

- This randomness needs to be taken into account for further inference and the issue is part of a general post-model selection inference problem:

  Dijkstra and Veldkamp, 1988; Kabaila, 1995; Leeb and Pöstcher, 2003–2008; Danilov and Magnus, 2004, among others.

- These authors remark that consistent model selectors usually produce super-efficient estimators, where non-uniformity (with respect to the true parameter Ψ_0) is observed in the convergence of finite-sample distributions to their asymptotic counterparts.

- This randomness needs to be taken into account for further inference and the issue is part of a general post-model selection inference problem:

  Dijkstra and Veldkamp, 1988; Kabaila, 1995; Leeb and Pöstcher, 2003–2008; Danilov and Magnus, 2004, among others.

- These authors remark that consistent model selectors usually produce super-efficient estimators, where non-uniformity (with respect to the true parameter $\boldsymbol{\Psi}_0$) is observed in the convergence of finite-sample distributions to their asymptotic counterparts.

Valid post-selection inference in multiple linear and GLM regression

- Recent developments on post-selection inference:
  - A significance test for the lasso: Lockhart et al. (2014)
  - Bootstrapping: Efron (2014)
  - De-sparsifying:

    Zhang & Zhang (2014); van de Geer et al. (2014)
  - Screening & cleaning based on sample splitting:

    Wasserman & Roeder (2009), Meinshausena et al. (2009)
  - Berk et al. (2013)
- An integrative review of post-selection inference by:

  Zhang et al (2018).

Note:

- Since the true sparse model is unknown, formulation of hypotheses concerning regression coefficients is unclear.
- As in Meinshausen et al. (2009), we may assign p-values to all the variables under study. However, rigorous statistical justification of such an approach raises fundamental questions about the meaning of the underlying true model.
- Thus, hypotheses of interest can only be formulated using an estimated sparse model.

Note:

- Since the true sparse model is unknown, formulation of hypotheses concerning regression coefficients is unclear.
- As in Meinshausen et al. (2009), we may assign p-values to all the variables under study. However, rigorous statistical justification of such an approach raises fundamental questions about the meaning of the underlying true model.
- Thus, hypotheses of interest can only be formulated using an estimated sparse model.

Note:

- Since the true sparse model is unknown, formulation of hypotheses concerning regression coefficients is unclear.
- As in Meinshausen et al. (2009), we may assign p-values to all the variables under study. However, rigorous statistical justification of such an approach raises fundamental questions about the meaning of the underlying true model.
- Thus, hypotheses of interest can only be formulated using an estimated sparse model.

Post-selection inference in sparse FMR models:

- Our proposed method involves:

  (i) estimating the active predictor set of the true sparse model using a consistent model selector,

  (ii) testing hypotheses for the regression coefficients associated with the estimated active predictor set (EAPS).

- The method asymptotically controls the family wise error rate (FWER, the probability of rejecting at least one hypothesis when it is true) at a pre-specified nominal level ($0 < \alpha < 1$), while accounting for selection uncertainty.

- We also provide examples of consistent model selectors and describe methods for finite sample improvements.

Post-selection inference in sparse FMR models:

- Our proposed method involves:

  (i) estimating the active predictor set of the true sparse model using a consistent model selector,

  (ii) testing hypotheses for the regression coefficients associated with the estimated active predictor set (EAPS).

- The method asymptotically controls the family wise error rate (FWER, the probability of rejecting at least one hypothesis when it is true) at a pre-specified nominal level ($0 < \alpha < 1$), while accounting for selection uncertainty.

- We also provide examples of consistent model selectors and describe methods for finite sample improvements.

Post-selection inference in sparse FMR models:

- Our proposed method involves:

  (i) estimating the active predictor set of the true sparse model using a consistent model selector,

  (ii) testing hypotheses for the regression coefficients associated with the estimated active predictor set (EAPS).

- The method asymptotically controls the family wise error rate (FWER, the probability of rejecting at least one hypothesis when it is true) at a pre-specified nominal level ($0 < \alpha < 1$), while accounting for selection uncertainty.

- We also provide examples of consistent model selectors and describe methods for finite sample improvements.

The proposed method

- Data: $(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n$.
- We split the data randomly into two parts $\mathcal{D}_{1n}$ and $\mathcal{D}_{2n}$ (approximately of the same size $n/2$), where we use $\mathcal{D}_{1n}$ to select a sparse model via a consistent selector T_n yielding an estimated active predictor set (EAPS), $\widehat{S}(T_n)$.
- Then, tests of hypotheses for regression coefficients of the selected model are performed using $\mathcal{D}_{2n}$, based on student-type statistics.

The proposed method

- Data: $(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n$.
- We split the data randomly into two parts $\mathcal{D}_{1n}$ and $\mathcal{D}_{2n}$ (approximately of the same size $n/2$), where we use $\mathcal{D}_{1n}$ to select a sparse model via a consistent selector T_n yielding an estimated active predictor set (EAPS), $\widehat{S}(T_n)$.
- Then, tests of hypotheses for regression coefficients of the selected model are performed using $\mathcal{D}_{2n}$, based on student-type statistics.

Note:

- 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.

Note:

- 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.

Some notations

- T_n is a consistent model selector: $\lim_{n\to\infty} P(T_n = \mathcal{M}_0) = 1$, where $\mathcal{M}_0$ is the true sparse FMR model.

- We apply T_n to $\mathcal{D}_{1n}$ and obtain an FMR sub-model with the EAPS $\widehat{\mathcal{S}}(T_n) = \bigcup_{j=1}^{K} \widehat{S}_j(T_n)$, where $\widehat{S}_j(T_n)$ is the active set selected by T_n in the j^{th} mixture component.

- The FMR sub-model associated with $\widehat{S}_j(T_n)$ is given by

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

where $\tilde{\eta}_j(\boldsymbol{x}) = \mathcal{L}(\beta_{j0} + \sum_{(j,l)\in\widehat{S}_j(T_n)} x_l \beta_{jl})$, and $\boldsymbol{\Psi}(\widehat{\mathcal{S}}(T_n))$ is a sub-vector of $\boldsymbol{\Psi}$. We will use $\widetilde{\boldsymbol{\Psi}}$ to denote $\boldsymbol{\Psi}(\widehat{\mathcal{S}}(T_n))$.

Some notations

- T_n is a consistent model selector: $\lim_{n\to\infty} P(T_n = \mathcal{M}_0) = 1$, where $\mathcal{M}_0$ is the true sparse FMR model.
- We apply T_n to $\mathcal{D}_{1n}$ and obtain an FMR sub-model with the EAPS $\widehat{\mathcal{S}}(T_n) = \bigcup_{j=1}^{K} \widehat{S}_j(T_n)$, where $\widehat{S}_j(T_n)$ is the active set selected by T_n in the j^{th} mixture component.
- The FMR sub-model associated with $\widehat{S}_j(T_n)$ is given by

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

where $\tilde{\eta}_j(\boldsymbol{x}) = \mathcal{L}(\beta_{j0} + \sum_{(j,l)\in\widehat{S}_j(T_n)} x_l \beta_{jl})$, and $\boldsymbol{\Psi}(\widehat{\mathcal{S}}(T_n))$ is a sub-vector of $\boldsymbol{\Psi}$. We will use $\widetilde{\boldsymbol{\Psi}}$ to denote $\boldsymbol{\Psi}(\widehat{\mathcal{S}}(T_n))$.

- T_n is a consistent model selector: $\lim_{n\to\infty} P(T_n = \mathcal{M}_0) = 1$, where $\mathcal{M}_0$ is the true sparse FMR model.

- We apply T_n to $\mathcal{D}_{1n}$ and obtain an FMR sub-model with the EAPS $\widehat{\mathcal{S}}(T_n) = \bigcup_{j=1}^{K} \widehat{S}_j(T_n)$, where $\widehat{S}_j(T_n)$ is the active set selected by T_n in the j^{th} mixture component.

- The FMR sub-model associated with $\widehat{S}_j(T_n)$ is given by

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

where $\tilde{\eta}_j(\boldsymbol{x}) = \mathcal{L}(\beta_{j0} + \sum_{(j,l)\in\widehat{S}_j(T_n)} x_l \beta_{jl})$, and $\boldsymbol{\Psi}(\widehat{\mathcal{S}}(T_n))$ is a sub-vector of $\boldsymbol{\Psi}$. We will use $\widetilde{\boldsymbol{\Psi}}$ to denote $\boldsymbol{\Psi}(\widehat{\mathcal{S}}(T_n))$.

Some hypotheses of interest

- For the selected model, we focus on the hypotheses:

(1) For all $1 \leq j \leq K$ and $l \in \widehat{S}_j(T_n)$:

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

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

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

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

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

To perform the tests:

- We use $\mathcal{D}_{2n}$ to obtain the MLE of $\widetilde{\boldsymbol{\Psi}}$, say $\overline{\widetilde{\boldsymbol{\Psi}}}_n$, 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 the hypothesis (1), compute the student-statistic

$$t_{jl,n} = \bar{\bar{\beta}}_{jl} / \mathrm{SE}(\bar{\bar{\beta}}_{jl}),$$

and $\mathrm{SE}(\bar{\bar{\beta}}_{jl})$ comes from the observed "information matrix".

- We show that asymptotically $\overline{\widetilde{\boldsymbol{\Psi}}}_n \sim$ Gaussian; from this, for small n, the distribution of $t_{jl,n}$ can then be approximated by a t-distribution with $\frac{n}{2} - \hat{q}_n - (3K - 1)$ degrees of freedom, where $\hat{q}_n = |\widehat{\mathcal{S}}(T_n)|$. We account for multiple comparisons using a Bonferroni-type adjustment.

To perform the tests:

- We use $\mathcal{D}_{2n}$ to obtain the MLE of $\widetilde{\boldsymbol{\Psi}}$, say $\overline{\widetilde{\boldsymbol{\Psi}}}_n$, 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 the hypothesis (1), compute the student-statistic

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

and $\mathrm{SE}(\bar{\tilde{\beta}}_{jl})$ comes from the observed "information matrix".

- We show that asymptotically $\overline{\widetilde{\boldsymbol{\Psi}}}_n \sim$ Gaussian; from this, for small n, the distribution of $t_{jl,n}$ can then be approximated by a t-distribution with $\frac{n}{2} - \hat{q}_n - (3K - 1)$ degrees of freedom, where $\hat{q}_n = |\widehat{\mathcal{S}}(T_n)|$. We account for multiple comparisons using a Bonferroni-type adjustment.

To perform the tests:

- We use $\mathcal{D}_{2n}$ to obtain the MLE of $\widetilde{\boldsymbol{\Psi}}$, say $\overline{\widetilde{\boldsymbol{\Psi}}}_n$, 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 the hypothesis (1), compute the student-statistic

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

and $\mathrm{SE}(\bar{\tilde{\beta}}_{jl})$ comes from the observed "information matrix".

- We show that asymptotically $\overline{\widetilde{\boldsymbol{\Psi}}}_n \sim$ Gaussian; from this, for small n, the distribution of $t_{jl,n}$ can then be approximated by a t-distribution with $\frac{n}{2} - \hat{q}_n - (3K - 1)$ degrees of freedom, where $\hat{q}_n = |\widehat{\mathcal{S}}(T_n)|$. We account for multiple comparisons using a Bonferroni-type adjustment.

In summary, here is the Algorithm:

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 $\mathcal{D}_{2n}$.

Theoretical justification: first we need some notation

- Given $\mathcal{D}_{1n}$, let p_{jl} be the p-value associated with the test in (1) which is of size $\alpha/\hat{q}_n$, for some $\alpha \in (0,1)$.
- 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.

- Furthermore, let

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

  denote the set of indices of regression coefficients of the selected covariates whose corresponding null hypotheses $H_{0,jl}$'s are rejected when they are true.

Theoretical justification: first we need some notation

- Given $\mathcal{D}_{1n}$, let p_{jl} be the p-value associated with the test in (1) which is of size $\alpha/\hat{q}_n$, for some $\alpha \in (0,1)$.

- 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.

- Furthermore, let

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

  denote the set of indices of regression coefficients of the selected covariates whose corresponding null hypotheses $H_{0,jl}$'s are rejected when they are true.

Theoretical justification: first we need some notation

- Given $\mathcal{D}_{1n}$, let p_{jl} be the p-value associated with the test in (1) which is of size $\alpha/\hat{q}_n$, for some $\alpha \in (0,1)$.

- 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.

- Furthermore, let

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

  denote the set of indices of regression coefficients of the selected covariates whose corresponding null hypotheses $H_{0,jl}$'s are rejected when they are true.

- 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 $n \to \infty$,

$$\mathcal{W}_n[\boldsymbol{I}_1(\boldsymbol{\Psi}_0(\hat{q}_n))]\mathcal{W}_n^\top \xrightarrow{p} \mathcal{W}[\boldsymbol{I}_1(\boldsymbol{\Psi}_0)]\mathcal{W}^\top, \tag{1}$$

  where $\boldsymbol{I}_1(\cdot)$ is the Fisher information matrix.

- The validity of (1) is guaranteed by the consistency property of the model selector T_n.

- Note: $\kappa_0 = q_0 + 3K - 1$ and $\dim(\widetilde{\boldsymbol{\Psi}}_n) \equiv \hat{\kappa}_n = \hat{q}_n + 3K - 1$, which could be different.

- 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 $n \to \infty$,

$$\mathcal{W}_n[\boldsymbol{I}_1(\boldsymbol{\Psi}_0(\hat{q}_n))]\mathcal{W}_n^\top \xrightarrow{p} \mathcal{W}[\boldsymbol{I}_1(\boldsymbol{\Psi}_0)]\mathcal{W}^\top, \tag{1}$$

  where $\boldsymbol{I}_1(\cdot)$ is the Fisher information matrix.

- The validity of (1) is guaranteed by the consistency property of the model selector T_n.

- Note: $\kappa_0 = q_0 + 3K - 1$ and $\dim(\widetilde{\boldsymbol{\Psi}}_n) \equiv \hat{\kappa}_n = \hat{q}_n + 3K - 1$, which could be different.

- 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 $n \to \infty$,

$$\mathcal{W}_n[\boldsymbol{I}_1(\boldsymbol{\Psi}_0(\hat{q}_n))]\mathcal{W}_n^\top \xrightarrow{p} \mathcal{W}[\boldsymbol{I}_1(\boldsymbol{\Psi}_0)]\mathcal{W}^\top, \tag{1}$$

  where $\boldsymbol{I}_1(\cdot)$ is the Fisher information matrix.

- The validity of (1) is guaranteed by the consistency property of the model selector T_n.

- Note: $\kappa_0 = q_0 + 3K - 1$ and $\dim(\widetilde{\boldsymbol{\Psi}}_n) \equiv \hat{\kappa}_n = \hat{q}_n + 3K - 1$, which could be different.

Dimension matching:

- 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.

  ($\mathcal{N}_0$: true inactive set; $\mathcal{S}_0$: true active set).

- 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 $\boldsymbol{\Psi}_0(\hat{q}_n) = (\pi_0, \phi_0, \beta_0^0, \boldsymbol{B}_{10}(\hat{q}_n))$.

Dimension matching:

- 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.

  ($\mathcal{N}_0$: true inactive set; $\mathcal{S}_0$: true active set).

- 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 $\boldsymbol{\Psi}_0(\hat{q}_n) = (\pi_0, \phi_0, \beta_0^0, \boldsymbol{B}_{10}(\hat{q}_n))$.

Dimension matching:

- 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.

  ($\mathcal{N}_0$: true inactive set; $\mathcal{S}_0$: true active set).

- 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 $\boldsymbol{\Psi}_0(\hat{q}_n) = (\pi_0, \phi_0, \beta_0^0, \boldsymbol{B}_{10}(\hat{q}_n))$.

Theorem 2

- Let T_n be a consistent model selector and $\alpha \in (0,1)$ be a nominal significance level. Under standard REGULARITY CONDITIONS, the following hold:

  (i) asymptotic normality: (as $n \to \infty$)

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

  (ii) FWER **control**:

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

Theorem 2

- Let T_n be a consistent model selector and $\alpha \in (0,1)$ be a nominal significance level. Under standard REGULARITY CONDITIONS, the following hold:

  (i) asymptotic normality: (as $n \to \infty$)

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

  (ii) FWER **control**:

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

Finite sample improvement: multiple splitting

- 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.
- A natural option is to split the data into two parts B times:

$$(\mathcal{D}_{1n}^1, \mathcal{D}_{2n}^1), (\mathcal{D}_{1n}^2, \mathcal{D}_{2n}^2), \cdots, (\mathcal{D}_{1n}^B, \mathcal{D}_{2n}^B).$$

- Accordingly, the EAPS for the b^{th} split is given by $\widehat{\mathcal{S}}_b$; then the EAPS based on all the splits is given by:

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

- By the choice of the consistent model selector, as $n \to \infty$, with probability tending to one, $\mathcal{S}_{B,n} = \mathcal{S}_0$, for any fixed B.

Finite sample improvement: multiple splitting

- 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.

- A natural option is to split the data into two parts B times:

$$(\mathcal{D}_{1n}^1, \mathcal{D}_{2n}^1), (\mathcal{D}_{1n}^2, \mathcal{D}_{2n}^2), \cdots, (\mathcal{D}_{1n}^B, \mathcal{D}_{2n}^B).$$

- Accordingly, the EAPS for the b^{th} split is given by $\widehat{\mathcal{S}}_b$; then the EAPS based on all the splits is given by:

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

- By the choice of the consistent model selector, as $n \to \infty$, with probability tending to one, $\mathcal{S}_{B,n} = \mathcal{S}_0$, for any fixed B.

Finite sample improvement: multiple splitting

- 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.

- A natural option is to split the data into two parts B times:

$$(\mathcal{D}_{1n}^1, \mathcal{D}_{2n}^1), (\mathcal{D}_{1n}^2, \mathcal{D}_{2n}^2), \cdots, (\mathcal{D}_{1n}^B, \mathcal{D}_{2n}^B).$$

- Accordingly, the EAPS for the b^{th} split is given by $\widehat{\mathcal{S}}_b$; then the EAPS based on all the splits is given by:

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

- By the choice of the consistent model selector, as $n \to \infty$, with probability tending to one, $\mathcal{S}_{B,n} = \mathcal{S}_0$, for any fixed B.

Hypotheses of interest

- Consider testing: $H_{0,jl} : \beta_{jl} = 0, \quad \text{for all} \quad (j, l) \in \mathcal{S}_{B,n}$.
- As before, we use the test statistic

$$t^b_{jl,n} = \bar{\tilde{\beta}}_{jl,b} / \mathrm{SE}(\bar{\tilde{\beta}}_{jl,b}),$$

  where 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 $\hat{q}_{n,b} = |\widehat{S}_b|$ p-values. For those indices in $\mathcal{S}_{B,n}$ but not in $\widehat{S}_b$ we assign p-value one.
- Multiple p-values: $\{p^b_{jl}, b = 1, 2 \cdots B\}$ which are correlated. We then need a method to aggregate dependent p-values.

Hypotheses of interest

- Consider testing: $H_{0,jl} : \beta_{jl} = 0, \quad \text{for all} \quad (j, l) \in \mathcal{S}_{B,n}$.
- As before, we use the test statistic

$$t^b_{jl,n} = \bar{\tilde{\beta}}_{jl,b} / \mathrm{SE}(\bar{\tilde{\beta}}_{jl,b}),$$

  where 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 $\hat{q}_{n,b} = |\widehat{\mathcal{S}}_b|$ p-values. For those indices in $\mathcal{S}_{B,n}$ but not in $\widehat{\mathcal{S}}_b$ we assign p-value one.

- Multiple p-values: $\{p^b_{jl}, b = 1, 2 \cdots B\}$ which are correlated. We then need a method to aggregate dependent p-values.

Hypotheses of interest

- Consider testing: $H_{0,jl} : \beta_{jl} = 0, \quad \text{for all} \quad (j,l) \in \mathcal{S}_{B,n}$.
- As before, we use the test statistic

$$t^b_{jl,n} = \bar{\tilde{\beta}}_{jl,b} / \mathrm{SE}(\bar{\tilde{\beta}}_{jl,b}),$$

  where 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 $\hat{q}_{n,b} = |\widehat{\mathcal{S}}_b|$ p-values. For those indices in $\mathcal{S}_{B,n}$ but not in $\widehat{\mathcal{S}}_b$ we assign p-value one.

- Multiple p-values: $\{p^b_{jl}, b = 1, 2 \cdots B\}$ which are correlated. We then need a method to aggregate dependent p-values.

- 1. Aggregation using quantiles:

$$Q_{jl}(\delta; \mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \equiv Q_{jl}(\delta) = \mathcal{Q}_\delta(\delta^{-1} p_{jl}^b : b = 1, ..., B),$$

where $\mathcal{Q}_\delta(\cdot)$ is the δ^{th} empirical quantile function.

- 2. Averaging: $\bar{Q}_{jl}(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) = B^{-1} \sum_{b=1}^{B} p_{jl}^b$ and set

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

- 1. Aggregation using quantiles:

$$Q_{jl}(\delta; \mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) \equiv Q_{jl}(\delta) = \mathcal{Q}_{\delta}(\delta^{-1} p_{jl}^{b} : b = 1, ..., B),$$

  where $\mathcal{Q}_{\delta}(\cdot)$ is the δ^{th} empirical quantile function.

- 2. Averaging: $\bar{Q}_{jl}(\mathcal{D}_{1n}^{1:B}, \mathcal{D}_{2n}^{1:B}) = B^{-1} \sum_{b=1}^{B} p_{jl}^{b}$ and set

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

In summary, the algorithm for multiple splitting (Msplit) is:

Step 1: Divide the data set randomly into two parts B times: $(\mathcal{D}_{1n}^1, \mathcal{D}_{2n}^1), (\mathcal{D}_{1n}^2, \mathcal{D}_{2n}^2), \cdots, (\mathcal{D}_{1n}^B, \mathcal{D}_{2n}^B)$.

Step 2: For each $1 \leq b \leq B$, obtain the EAPS $\widehat{\mathcal{S}}_b$, and set $\mathcal{S}_{B,n} = \bigcup_{1 \leq b \leq B} \widehat{\mathcal{S}}_b$.

Step 3: Using $\mathcal{D}_{2n}^{1:B}$, 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 and obtain the p-values.

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

In summary, the algorithm for multiple splitting (Msplit) is:

Step 1: Divide the data set randomly into two parts B times: $(\mathcal{D}_{1n}^1, \mathcal{D}_{2n}^1), (\mathcal{D}_{1n}^2, \mathcal{D}_{2n}^2), \cdots, (\mathcal{D}_{1n}^B, \mathcal{D}_{2n}^B)$.

Step 2: For each $1 \leq b \leq B$, obtain the EAPS $\widehat{\mathcal{S}}_b$, and set $\mathcal{S}_{B,n} = \bigcup_{1 \leq b \leq B} \widehat{\mathcal{S}}_b$.

Step 3: Using $\mathcal{D}_{2n}^{1:B}$, 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 and obtain the p-values.

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

In summary, the algorithm for multiple splitting (Msplit) is:

Step 1: Divide the data set randomly into two parts B times: $(\mathcal{D}_{1n}^1, \mathcal{D}_{2n}^1), (\mathcal{D}_{1n}^2, \mathcal{D}_{2n}^2), \cdots, (\mathcal{D}_{1n}^B, \mathcal{D}_{2n}^B)$.

Step 2: For each $1 \leq b \leq B$, obtain the EAPS $\widehat{\mathcal{S}}_b$, and set $\mathcal{S}_{B,n} = \bigcup_{1 \leq b \leq B} \widehat{\mathcal{S}}_b$.

Step 3: Using $\mathcal{D}_{2n}^{1:B}$, 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 and obtain the p-values.

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

In summary, the algorithm for multiple splitting (Msplit) is:

Step 1: Divide the data set randomly into two parts B times: $(\mathcal{D}_{1n}^1, \mathcal{D}_{2n}^1), (\mathcal{D}_{1n}^2, \mathcal{D}_{2n}^2), \cdots, (\mathcal{D}_{1n}^B, \mathcal{D}_{2n}^B)$.

Step 2: For each $1 \leq b \leq B$, obtain the EAPS $\widehat{\mathcal{S}}_b$, and set $\mathcal{S}_{B,n} = \bigcup_{1 \leq b \leq B} \widehat{\mathcal{S}}_b$.

Step 3: Using $\mathcal{D}_{2n}^{1:B}$, 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 and obtain the p-values.

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

In summary, the algorithm for multiple splitting (Msplit) is:

Step 1: Divide the data set randomly into two parts B times: $(\mathcal{D}_{1n}^1, \mathcal{D}_{2n}^1), (\mathcal{D}_{1n}^2, \mathcal{D}_{2n}^2), \cdots, (\mathcal{D}_{1n}^B, \mathcal{D}_{2n}^B)$.

Step 2: For each $1 \leq b \leq B$, obtain the EAPS $\widehat{\mathcal{S}}_b$, and set $\mathcal{S}_{B,n} = \bigcup_{1 \leq b \leq B} \widehat{\mathcal{S}}_b$.

Step 3: Using $\mathcal{D}_{2n}^{1:B}$, 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 and obtain the p-values.

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

Theoretical justification:

- We prove that the resulting quantities ($Q_{jl}(\delta)$ and $\bar{Q}^*_{jl}$) from any of the aggregation methods are indeed p-values.

- Furthermore, using the $Q_{jl}(\delta)$s, consider the sets

$$\mathcal{S}^*_{B,n}(\delta) = \bigcup_{(j,l)\in\mathcal{S}_{B,n}} \{Q_{jl}(\delta) \leq \alpha\} \ , \ \mathcal{E}(\delta) = \mathcal{N}_0 \bigcap \mathcal{S}^*_{B,n}(\delta).$$

- We show that:

$$\limsup_{n\to\infty} \mathrm{P}\left(\mathcal{E}(\delta) \neq \emptyset\right) \leq \alpha$$

i.e. the FWER control !

Simulations:

- The $\boldsymbol{x}_i$ are generated from multivariate normal with mean zero and an autoregressive-type covariance matrix Σ.
- 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 $\sigma^2 = 1, 4, 9, 25, 36$, yielding 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

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

$$\boldsymbol{\beta}_2^\top = (-1.7, 0.0, 2.5, -2.5, -2.0, 0.0, \ldots, 0.0)$$

containing $q_1 = 5$ and $q_2 = 4$ non-zero β_j's, respectively.

Simulations:

- The $\boldsymbol{x}_i$ are generated from multivariate normal with mean zero and an autoregressive-type covariance matrix Σ.
- 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 $\sigma^2 = 1, 4, 9, 25, 36$, yielding 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

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

$$\boldsymbol{\beta}_2^\top = (-1.7, 0.0, 2.5, -2.5, -2.0, 0.0, \ldots, 0.0)$$

containing $q_1 = 5$ and $q_2 = 4$ non-zero β_j's, respectively.

Performance measures:

- The proposed method (Msplit) is compared with the standard regularization techniques based on ADLASSO and SCAD penalties (Theorem 1), using the following 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.

Simulation results for the Gaussian model: $K = 2, d = 30$

		E(TP)			E(FP)			EFWER		
SNR	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
	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
6.45	Com_1	4.84	5.00	5.00	.000	.520	.335	.000	.345	.275
	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
2.87	Com_1	3.27	4.82	4.89	.000	.770	.955	.000	.485	.575
	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
1.03	Com_1	1.43	3.93	4.08	.000	4.67	3.85	.000	.970	.955
	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
0.72	Com_1	.905	3.67	3.77	.000	7.09	5.65	.000	1.00	1.00
	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

Simulation results for the Gaussian model: $K = 2, d = 50$

		E(TP)			E(FP)			EFWER		
SNR	Mixture	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO
25.8	Com_1	5.00	5.00	5.00	.000	.095	.000	.000	.075	.000
	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
6.45	Com_1	4.80	5.00	5.00	.000	1.04	.700	.000	.555	.495
	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
2.87	Com_1	3.27	4.76	4.75	.000	1.21	1.56	.000	.560	.740
	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
1.03	Com_1	1.55	3.45	3.91	.000	4.41	6.42	.000	.950	.995
	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
0.72	Com_1	1.12	3.14	3.61	.000	6.94	9.12	.000	1.00	1.00
	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

Simulation results for the Gaussian model: $K = 2, d = 70$

		E(TP)			E(FP)			EFWER		
SNR	Mixture	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO	Msplit	SCAD	ADLASSO
25.8	Com_1	5.00	5.00	5.00	.000	.115	.005	.000	.095	.005
	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	4.71	5.00	4.95	.000	1.58	1.27	.000	.685	.660
	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	3.44	4.74	4.56	.000	1.61	2.65	.000	.655	.875
	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	1.90	3.36	3.78	.010	5.69	8.39	.010	.985	1.00
	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	1.47	2.95	3.41	.010	9.23	12.7	.010	1.00	1.00
	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

thank you!