Selective Inference for Sparse Multitask Regression with Applications in Neuroimaging

Snigdha Panigrahi¹, Natasha Stewart¹, Chandra Sripada²,³, and Elizaveta Levina¹

¹Department of Statistics, University of Michigan
²Department of Psychiatry, University of Michigan
³Department of Philosophy, University of Michigan

Multi-task learning is frequently used to model a set of related response variables from the same set of features, improving predictive performance and modeling accuracy relative to methods that handle each response variable separately. Despite the potential of multi-task learning to yield more powerful inference than single-task alternatives, prior work in this area has largely omitted uncertainty quantification. Our focus in this paper is a common multi-task problem in neuroimaging, where the goal is to understand the relationship between multiple cognitive task scores (or other subject-level assessments) and brain connectome data collected from imaging. We propose a framework for selective inference to address this problem, with the flexibility to: (i) jointly identify the relevant covariates for each task through a sparsity-inducing penalty, and (ii) conduct valid inference in a model based on the estimated sparsity structure. Our framework offers a new conditional procedure for inference, based on a refinement of the selection event that yields a tractable selection-adjusted likelihood. This gives an approximate system of estimating equations for maximum likelihood inference, solvable via a single convex optimization problem, and enables us to efficiently form confidence intervals with approximately the correct coverage. Applied to both simulated data and data from the Adolescent Brain Cognitive Development (ABCD) study, our selective inference methods yield tighter confidence intervals than commonly used alternatives, such as data splitting. We also demonstrate through simulations that multi-task learning with selective inference can more accurately recover true signals than single-task methods.

Keywords: Multi-task learning, multi-level Lasso, joint sparsity, post-selection inference, selective inference, neuroimaging, fMRI data
1 Introduction

Humans exhibit a diversity of cognitive abilities, which can be categorized as either fluid or crystallized. Fluid abilities are rooted in problem solving and manipulation of information, independent of prior learning (Gray et al., 2003; Blair, 2006). Examples include the ability to store and manipulate items in short-term memory (working memory) and detect subtle patterns in sequences (matrix reasoning). Crystallized abilities are rooted in determinate facts that are trained through prior learning (Cattell, 1943; Horn and Noll, 1997), including reading comprehension and vocabulary. According to proponents of a general ability model, although fluid tasks are superficially quite different from crystallized tasks, there is an underlying shared ability that drives performance on both kinds of tasks (Spearman, 1961; Humphreys, 1979; Carroll et al., 1993). The question of whether cognitive abilities are better understood in terms of a single, general ability or separate fluid and crystallized abilities has been debated for over 100 years without a firm consensus emerging.

Functional neuroimaging offers new opportunities to non-invasively investigate the neurological organization of cognitive abilities, avoiding the need to rely solely on behavioral data. Recent efforts have focused on using neuroimaging data to predict performance across cognitive tasks and enhance cartographic understanding by linking cognitive functions to particular brain networks. Cartographic maps for various cognitive abilities could add evidence for the general ability hypothesis if they show spatial overlap. Alternatively, these maps might show that individual cognitive abilities exhibit spatially distinct patterns in the brain, strengthening the case for neurally separate abilities.

In this paper, we leverage multi-task learning (MTL) to advance both predictive and cartographic goals in the study of cognitive abilities. MTL is an important tool for modeling related response variables, such as performance on cognitive tasks, that have common predictors and potentially common patterns of dependence between the predictors and responses. Algorithms that use MTL are known to improve predictive accuracy by accounting for shared information between the tasks. Many different MTL algorithms have been developed, including regression methods that impose inter-task dependence through shared sparsity or low-rank constraints on the multi-task regression coefficients; see Zhang and Yang (2021) and references therein for a thorough survey on the topic. We utilize a regression-based MTL method to identify a potential common neurological basis for fluid and crystallized abilities.

In addition to predicting performance across different cognitive tasks and visualizing the involvement of different brain regions, we also address an important question that has been neglected: statistical inference for the selected neurological models. We introduce a novel, two-step procedure that can recover shared signals in neuroimaging data and subsequently quantify uncertainty via selective inference. The first step in our procedure entails using a randomized MTL algorithm with a sparsity-inducing penalty to jointly identify a model for each of the cognitive tasks. In stage two, we conduct
selective inference in the model chosen from the estimated sparsity structure, enabling us to test the significance of the shared signals recovered in the first step. This hypothesis testing framework, which offers the flexibility of choosing the model under which to conduct inference after examining the data, has the potential to significantly expand our understanding of the relationship between cognition and neurological features extracted from imaging data.

1.1 Contributions to the Adolescent Brain Cognitive Development (ABCD) Study

The behavioral and brain data we study come from the Adolescent Brain Cognitive Development (ABCD) study, one of the most extensive efforts to track the brain development of a large cohort of children in the United States, with close to 12,000 children enrolled across 23 research sites. The ABCD study includes an 11-task neurocognitive battery (Luciana et al., 2018), primarily based on the NIH Toolbox for the Assessment of Neurological and Behavioral Function, with several additional tasks to ensure comprehensive coverage across cognitive domains. Importantly, the ABCD battery includes two classic crystallized ability tasks (picture vocabulary and reading comprehension) as well as two classic fluid ability tasks (list working memory and matrix reasoning). It also uses multiple imaging modalities to track the participants’ brain development (Hagler Jr et al., 2019). Our focus is resting-state fMRI, which captures patterns of spontaneous activation throughout the brain while subjects are at rest in the scanner, yielding maps of functional connectivity. Previous studies have investigated the role of functional connectivity in general cognitive ability (Finn et al., 2015; Hearne et al., 2016; Sripada et al., 2021; Anderson and Barbey, 2022; Tong et al., 2022) as well as specific abilities like working memory (Markett et al., 2018) and matrix reasoning (Fraenz et al., 2021). Few previous studies have jointly modeled performance across cognitive domains to improve predictive performance or to better understand the spatial relationships between tasks. One notable exception is Adeli et al. (2019), who used MTL to jointly predict the development of general and specific cognitive abilities in infants and young children over time. Other studies have implicitly leveraged joint model selection to predict performance on individual cognitive domains using connections associated with general ability (Anderson and Barbey, 2022; Tong et al., 2022). In contrast to these previous studies, we do not use any measure of general ability computed from behavioral data. Our method relies on MTL to uncover the relationships between tasks from brain-behavior patterns, providing a new framework to assess the shared and distinct neural contributions to fluid and crystallized intelligence.

A further limitation of the studies mentioned above, as well as most previous methodological developments for MTL, is an emphasis on prediction over inference. Prediction is typically not the final goal in cognitive neuroscience applications, however, since directly measuring cognitive ability from behavioral data is easier and less expensive than acquiring brain imaging data. Furthermore, predictive accuracy has limited explanatory value by itself since very different models can sometimes achieve
similar predictive performance. A combination of prediction and statistical inference could shed light on the neurobiological basis of cognition, identifying brain-behavior relationships and measuring both the strength and the degree of confidence for each relationship.

1.2 Novel Contributions to Selective Inference

Selective inference has been studied extensively for single-task prediction. A common approach to selective inference, described by Fithian et al. (2017), accounts for bias from model selection by conditioning on the chosen model. The key to this approach is characterizing the selection event in a sufficiently simple form. For models chosen based on the LASSO, Lee et al. (2016) developed the influential polyhedral method, reducing the selection event to a series of linear inequalities in the response variable. Many different model selection events for a Gaussian response variable have subsequently been identified with a similar set of inequalities, yielding conditional distributions that can be used for inference (Suzumura et al., 2017; Liu et al., 2018; Taylor and Tibshirani, 2018; Zhao and Panigrahi, 2019; Tanizaki et al., 2020, among others).

Despite the convenience of the polyhedral method, the resulting confidence intervals can have infinite expected length for Gaussian regression (Kivaranovic and Leeb, 2018). The loss in inferential power can be remedied by applying conditional inference to a randomized problem, for example, by adding random noise to the response variable (Tian and Taylor, 2018) or by holding out some samples during selection, known as data carving (Fithian et al., 2017; Panigrahi, 2018). Under randomized versions of single-task algorithms such as the LASSO, Tian and Taylor (2018) obtained a pivot for each selected parameter after eliminating other (nuisance) parameters. More recently, Panigrahi and Taylor (2022) and Panigrahi et al. (2018, 2021, 2022a) build on the polyhedral method to introduce a tractable likelihood that allows for both frequentist and Bayesian selective inference, using a prior in conjunction with the likelihood for the latter.

Unfortunately, the polyhedral methods for single-task algorithms do not generalize well to the multi-task setting because the usual conditioning event for single-task methods does not have a simple characterization under joint model selection. Remarkably, there is a proper subset of the usual conditioning event that makes selective inference feasible for our particular MTL procedure without sacrificing much power, leading to an easy-to-solve system of estimating equations. To the best of our knowledge, this is the first selective inference method for the multi-task setting.

The remaining paper is organized as follows. Section 2 presents our algorithm for estimating the shared sparsity structure. Section 3 develops our method for MLE-based selective inference. In Section 4, we apply our methods to the ABCD study data for identifying the neurobiological underpinnings of fluid and crystallized intelligence. A brief discussion in Section 5 concludes our paper.
2 Multi-Task Learning for Joint Model Selection

2.1 The Two-Stage Model Selection and Inference Protocol

For ease of presentation, we first review the two-stage protocol for model selection and subsequent selective inference for the (single-task) randomized LASSO described in Panigrahi and Taylor (2022). Suppose we observe a predictor matrix \( X \in \mathbb{R}^{n \times p} \) with fixed entries, a random response vector \( y \sim N(\mu, \Sigma) \in \mathbb{R}^n \) with \( \mu \) unknown and \( \Sigma \) known, and an independent randomization variable \( \omega \sim N(0, \Omega) \in \mathbb{R}^p \) with known \( \Omega \). In the first stage, we identify a sparse linear model through a randomized LASSO regression. We estimate the regression coefficients, \( \Theta \in \mathbb{R}^p \), by solving

\[
\hat{\Theta} = \arg\min_{\Theta} L(\Theta; y, X, \omega) + \|\Lambda \Theta\|_1,
\]

(1)

where \( \Lambda \in \mathbb{R}^{p \times p} \) represents a diagonal matrix of feature-specific tuning parameters. The randomized loss \( L(\cdot; y, X, \omega) \) is given by

\[
L(\Theta; y, X, \omega) = \frac{1}{2} ||y - X\Theta||_2^2 - \omega'\Theta + \frac{\epsilon}{2} \|\Theta\|_2^2.
\]

(2)

We use \( M' \) to denote the transpose of a matrix \( M \). The ridge term in the loss function, with a small \( \epsilon \), is included to ensure the existence of an optimal solution; see Tian et al. (2016). The estimated support set, \( \hat{E} = \text{Supp}(\hat{\Theta}) \), is treated as random, reflecting the fact that different samples of \( y \) and \( \omega \) will yield different active sets. Suppose we observe \( \hat{E}(y, \omega) = E \) on our specific data. We use \( X_E \in \mathbb{R}^{n \times |E|} \) to represent the restriction of \( X \) to the columns indexed by the set \( E \).

After observing the active set \( E \), we assume a linear model of the form \( y \sim N(X_E \beta_E, \sigma^2 I_n) \), where \( \sigma \) is fixed and \( I_n \) is the \( n \times n \) identity matrix. This is a simple way to specify a model based on the active set, but selective inference affords us the flexibility to choose from a range of specifications. For instance, the two-stage protocol we describe here can also be used for other linear models, such as those involving linear combinations of basis functions constructed from the active features.

In the second stage, we aim to construct \( 100(1 - \alpha) \% \) confidence intervals for the best linear coefficients,

\[
\beta_E = \arg\min_{\beta \in \mathbb{R}^{|E|}} \mathbb{E} \left[ ||y - X_E \beta||_2^2 \right],
\]

(3)

under the selected model. We note that some recent post-selection work by Hong et al. (2018); Wang et al. (2020); Panigrahi et al. (2022b) has focused on overfitting bias from model selection. However, the parameters in (3) are well-defined regardless of model misspecification, which could be either due to underfitting or overfitting bias.
A conditional likelihood for $\beta_E$ is obtained by conditioning upon the partition
\[
\mathcal{P}_E = \{ y \in \mathbb{R}^n, \omega \in \mathbb{R}^p : \hat{E}(y, \omega) = E \},
\]
that contains all instances leading us to observe the estimated support set $E$. Letting $\rho(x; \mu, \Omega)$ be the multivariate normal density function with mean vector $\mu$ and covariance $\Omega$, we can write this likelihood as
\[
y \mid \hat{E} = E \propto \frac{\rho(y; X_E \beta_E, \sigma^2 I_n)}{\int_{\mathcal{P}_E} \rho(\tilde{y}; X_E \beta_E, \sigma^2 I_n) \cdot \rho(\tilde{w}; 0, \Omega)d\tilde{w}d\tilde{y}}. \tag{4}
\]
The likelihood in (4) does not have a closed form since the normalizing constant is intractable. Instead, a tractable version of the likelihood function, called the selection-adjusted likelihood, is obtained by conditioning on a proper subset of the actual selection event:
\[
\left\{ \hat{S}(y, \omega) = S \right\} \subseteq \left\{ \hat{E}(y, \omega) = E \right\}.
\]
The refined conditioning event, on the left-hand side of the previous display, can be characterized by the polyhedral partition
\[
\{ y \in \mathbb{R}^n, \omega \in \mathbb{R}^p : Ay + B\omega \leq c \}
\]
for fixed matrices $A, B$ and $c$. For example, Lee et al. (2016) identify a polyhedral partition by conditioning further on the active signs of the LASSO solution, alongside the estimated support set $E$.

Approximate inference is then obtained by centering interval estimates around the maximum likelihood estimate (MLE) of the selection-adjusted likelihood, $\hat{\beta}_E^{\text{MLE}}$, and using the observed Fisher information matrix, $\hat{I}(\hat{\beta}_E^{\text{MLE}})$, to estimate the variance. This yields post-selection confidence intervals for $\beta_E$ of the form
\[
\hat{\beta}_{E,j}^{\text{MLE}} \pm z_{1-\alpha/2} \sqrt{\hat{I}_{jj}^{-1}(\hat{\beta}_E^{\text{MLE}})}, \quad j \in E;
\]
where $v_j$ is the $j$th entry of the vector $v$, $M_{ij}$ is the $(i, j)$th element of a matrix $M$, and $z_q$ is the $q$th quantile of the standard normal distribution.

### 2.2 An Objective Function with Shared Sparsity for MTL

We next set up the objective function in the multi-task setting. Suppose we have $K$ regression tasks, with $K$ distinct response variables and a common set of $p$ predictors. For $k \in [K]$, let $n_k$ denote the sample size available for task $k$, $y^{(k)} \in \mathbb{R}^{n_k}$ denote the response vector for the $k$th task, and $X^{(k)} \in \mathbb{R}^{n_k \times p}$ denote the corresponding predictor matrix. We assume that the predictors in each task have been centered and do not include intercept terms in the regression. Consistent with the two-stage procedure described in Section 2.1, we introduce a randomization variable for each task. Let $\omega^{(k)} \in \mathbb{R}^p$
denote a Gaussian randomization variable such that (i) \( \omega^{(k)} \sim N \left( 0, \Omega^{(k)} \right) \) for \( k \in [K] \); (ii) \( \omega^{(k)} \) is independent of \( \omega^{(k')} \) for all \( k' \neq k \); and (iii) \( \omega^{(k)} \) is independent of \( y^{(k')} \) for all \( k' \in [K] \).

For each task, we assume that the set of non-zero coefficients is sparse and use penalized multi-task regression to identify the relevant features. We impose a specific inter-task structure on the joint \( \Theta \) for all \( k \), namely, a common parameter that is shared between all tasks and a task-specific parameter that is unique to an individual task; namely,

\[
\Theta_{j}^{(k)} = \tau_{j} \gamma_{j}^{(k)}, \quad \text{for } j \in [p], k \in [K].
\]

A similar multiplicative parameterization has been used for joint estimation in a number of penalized MTL algorithms (Bi et al., 2008; Lozano and Swirszcz, 2012; Wang et al., 2016). To avoid a sign ambiguity, we take \( \tau_{j} \geq 0 \) for \( j \in [p] \). We do not impose any further constraints to ensure that the two components, \( \tau_{j} \) and \( \gamma_{j}^{(k)} \), are each identifiable since our interest lies in estimating the sparsity structure of their product. Note that \( \tau_{j} \) determines the sparsity at the global level as \( \tau_{j} = 0 \) implies that \( \Theta_{j}^{(k)} = 0 \) for all \( k \in [K] \). The parameter \( \gamma_{j}^{(k)} \) controls the task-specific sparsity, with \( \gamma_{j}^{(k)} = 0 \) indicating that \( \Theta_{j}^{(k)} = 0 \) for task \( k \).

We proceed to fit a sparse model by minimizing the penalized objective function

\[
\sum_{k=1}^{K} \mathcal{L}(\Theta^{(k)}; y^{(k)}, X^{(k)}, \omega^{(k)}) + \eta_{1} \sum_{j=1}^{p} \tau_{j} + \eta_{2} \sum_{j=1}^{p} \sum_{k=1}^{K} |\gamma_{j}^{(k)}|,
\]

subject to the constraint \( \tau_{j} \geq 0 \) for \( j \in [p] \). The minimizers \( \hat{\tau} \in \mathbb{R}^{p}, \hat{\gamma}^{(k)} \in \mathbb{R}^{p} \) will clearly both be sparse due to the \( \ell_1 \) penalties in the objective function. To optimize the objective (6), we note that its solution will yield coefficients \( \Theta^{(k)} \) for \( k \in [K] \) that could be obtained from an equivalent formulation

\[
\arg\min_{\{\Theta^{(k)}\}_{j \in [p], k \in [K]}} \sum_{k=1}^{K} \mathcal{L}(\Theta^{(k)}; y^{(k)}, X^{(k)}, \omega^{(k)}) + 2\lambda \sum_{j=1}^{p} \left( \sum_{k=1}^{K} |\Theta_{j}^{(k)}| \right)^{1/2}, \tag{7}
\]

where \( \lambda = \sqrt{\eta_{1} \eta_{2}} \). This equivalence was established in previous work; see, e.g., Guo et al. (2011) and Lozano and Swirszcz (2012).

Following the approach of Zou and Li (2008) and Guo et al. (2011), we use an iterative local linear approximation to the penalty term in (7), centered at the absolute value of the previous iterate:

\[
2 \left\{ \sum_{k=1}^{K} |\Theta_{j}^{(k)}| \right\}^{1/2} \sim c^{(t)} + \sum_{k=1}^{K} \frac{|\Theta_{j}^{(k)}|}{\sqrt{\sum_{k=1}^{K} |(\hat{\Theta}_{j}^{(k)})^{(t)}|}}.
\]
The constant $c(t)$ depends only on the previous iterate and can be ignored in the corresponding optimization problem. With this approximation, the multi-task objective conveniently decouples by task. The successive estimates of the regression coefficients can be computed by solving a LASSO problem separately for each of the $K$ tasks, dependent on the previous iterate only through the penalty weights

$$
\lambda_j^{(t+1)} = \min \left\{ \lambda_0, \lambda \cdot \left( \sum_{k=1}^{K} |(\hat{\Theta}_j^{(k)})^{(t)}| \right)^{-\frac{1}{2}} \right\}, \ j \in [p].
$$

Here, $\lambda_0$ is a pre-specified large positive constant used for numerical stability, and $\lambda$ is a tuning parameter. The iterative procedure is summarized in Algorithm 1.

**Algorithm 1 Estimating Shared Sparsity**

1: for $k=1,\ldots,K$ do
2: Initialize $(\hat{\Theta}^{(k)})^{(0)} = \arg\min_{\Theta^{(k)}} \left( \mathcal{L}(\Theta^{(k)}; y^{(k)}, X^{(k)}, \omega^{(k)}) + \lambda \sum_{j=1}^{p} |\Theta_j^{(k)}| \right)$
3: end for
4: procedure ITERATE UNTIL CONVERGENCE
5: Let $t = 0$
6: while convergence $< \text{tol}$ do
7: for $j = 1,\ldots,p$ do
8: $\lambda_j^{(t+1)} = \min \left\{ \lambda_0, \lambda \cdot \left( \sum_{k=1}^{K} |(\hat{\Theta}_j^{(k)})^{(t)}| \right)^{-\frac{1}{2}} \right\}$
9: end for
10: for $k = 1,\ldots,K$ do
11: LASSO:
12: Solve $(\hat{\Theta}^{(k)})^{(t+1)} = \arg\min_{\Theta^{(k)}} \left( \mathcal{L}(\Theta^{(k)}; y^{(k)}, X^{(k)}, \omega^{(k)}) + \sum_{j=1}^{p} \lambda_j^{(t+1)} |\Theta_j^{(k)}| \right)$
13: end for
14: $t = t + 1$
15: end while
16: end procedure

Upon convergence of Algorithm 1, we obtain

$$
\hat{E}_k = \text{Supp}(\hat{\Theta}^{(k)}) = E_k, \text{ for } k \in [K]. \quad (8)
$$

Let the cardinality of $E_k \subseteq [p]$ be equal to $\delta_k$, and let $\delta = \delta_1 + \cdots + \delta_K$. 
3 Maximum Likelihood Inference Post MTL

Next, we proceed to specify a model with the sparsity structure estimated through (6) and quantify uncertainty in the effects of selected predictors with respect to this model. We restrict our search to linear models of the form

\[ y^{(k)} = X_{E_k}^{(k)} \beta_{E_k}^{(k)} + \varepsilon_k \text{ for } k \in [K], \]  

where \( \varepsilon_k \) is a vector of errors with length \( n_k \), mean 0, and variance \( \sigma_k^2 \). We assume that the errors are independent across samples and across tasks. We fit this model by maximizing the corresponding normal likelihood, which can be viewed as a general \( M \)-estimation procedure. For the inference step, the post-selection confidence intervals for \( \beta_{E_k}^{(k)} \) will be based on the normal assumption for the error distribution.

3.1 Some preliminaries

We use boldface capital letters to denote stacked quantities across tasks. Let

\[ Y = \left( y^{(1)} \ldots y^{(K)} \right)', \quad E = \{ E_1, \ldots, E_K \}, \quad \beta_E = \left( \beta_{E_1}^{(1)} \ldots \beta_{E_k}^{(K)} \right)' \]

Without loss of generality, we can reorder each predictor matrix and randomization instance to have the active components precede the inactive components:

\[ X^{(k)} = \begin{bmatrix} X_{E_k}^{(k)} & X_{-E_k}^{(k)} \end{bmatrix}, \quad \omega^{(k)} = \begin{pmatrix} \omega_{E_k}^{(k)} \\ \omega_{-E_k}^{(k)} \end{pmatrix}, \]

where \( -A \) denotes the complement of set \( A \). For each predictor \( j \in [p] \), we use \( \tilde{j}_k \) to denote the index of this predictor in task \( k \) after the permutation. Let \( b^{(k)} \in \mathbb{R}^{\delta_k} \) denote the absolute values of the estimated non-zero multi-task regression coefficients for task \( k \in [K] \) under this permutation, i.e.,

\[ b_{j_k}^{(k)} = |\tilde{\Theta}_j^{(k)}| \text{ whenever } |\tilde{\Theta}_j^{(k)}| \neq 0, \]

and let

\[ B = \left( b^{(1)} \ldots b^{(K)} \right)' \]

We let \( s^{(k)} \in \mathbb{R}^{\delta_k} \) and \( u^{(k)} \in \mathbb{R}^{p-\delta_k} \) represent the active and inactive components of the subgradient vector of the \( \ell_1 \)-norm for the multi-task regression coefficients when evaluated at the solution, i.e.,

\[ \left( s^{(k)} u^{(k)} \right)' = D_\Theta \left\| \Theta_{E_k}^{(k)} \Theta_{-E_k}^{(k)} \right\|_1 \text{ for } k \in [K]. \]
Note that the vector $s^{(k)}$ gives the signs of the active multi-task coefficients for the $k$th task, and $u^{(k)}$ satisfies $\|u^{(k)}\|_\infty \leq 1$. To reference the active and inactive components, respectively, of all the evaluated $\ell_1$-norm sub-gradients, we define

$$S = \left(s^{(1)} \ldots s^{(K)}\right)' , \quad U = \left(u^{(1)}' \ldots u^{(K)}'\right)' .$$

We next introduce some notation to account for information shared between tasks. Suppose there are $r$ predictors, $j_1, \ldots, j_r$, that are active in one or more tasks. Define $\Gamma \in \mathbb{R}^r$ by

$$\Gamma = \left(\Gamma^{(j_1)} \ldots \Gamma^{(j_r)}\right)' ,$$

where

$$\Gamma^{(j)} = \sum_{k=1}^{K} |\hat{\Theta}_j^{(k)}| \quad \text{for} \; j \in \{j_1, \ldots, j_r\} .$$

Let the set of tasks where predictor $j$ is active be given by

$$\kappa(j) = \{k : |\hat{\Theta}_j^{(k)}| \neq 0\} ,$$

with $d_j = |\kappa(j)|$ and elements $\kappa_1 < \cdots < \kappa_{d_j}$ arranged in increasing order. For the active predictors $j_1, \ldots, j_r$, we define $v^{(j)}$ to be a vector representing the first $(d_j - 1)$ corresponding coefficients:

$$v^{(j)} = \left(|\hat{\Theta}_j^{(\kappa_1)}| \ldots |\hat{\Theta}_j^{(\kappa_{d_j-1})}|\right)' \quad \text{for} \; j \in \{j_1, \ldots, j_r\} .$$

The vector $v^{(j)}$ collects the absolute value of the non-zero coefficients for predictor $j$ across tasks, excluding the coefficient for the last task where predictor $j$ is non-zero. For most predictors in the active set, we expect that $d_j \geq 2$ due to the shared sparsity across tasks; however, if we estimate $d_j = 1$, then $v^{(j)}$ is empty and can be disregarded. After accounting for all but the last active coefficient corresponding to each predictor, we let

$$V = \left(v^{(j_1)' \ldots v^{(j_r)'}\right)' .$$

Note that there is a bijective mapping between $B$ and $(V, \Gamma)$. We introduce a matrix $D \in \mathbb{R}^{r \times (\delta - r)}$ to record which elements of $V$ correspond to each of the $r$ active predictors, with rows given by

$$D_i = \left(0_{(d_{j_1}-1)} \ldots 1_{(d_{j_i}-1)} \ldots 0_{(d_{j_r}-1)}\right) \quad \text{for} \; i \in [r].$$
For some permutation matrix $A \in \mathbb{R}^{\delta \times \delta}$, the relationship between $B$ and $(V, \Gamma)$ is given by

$$B = A \begin{pmatrix} V \\ \Gamma - DV \end{pmatrix}.$$

Let the matrix $H \in \mathbb{R}^{\delta \times (\delta - r)}$ and the vector $g \in \mathbb{R}^\delta$ be given by:

$$H = \begin{pmatrix} I_{(\delta - r)} \\ -D \end{pmatrix}, \quad g = \begin{pmatrix} 0_{(\delta - r)} \\ -\Gamma \end{pmatrix}.$$ 

To enforce the $\delta$ linear inequalities given by $Hv \geq g$, we use the following barrier function:

$$\phi_{H,g}(v) = \begin{cases} \sum_{j=1}^{\delta} \log \left( 1 + \frac{1}{H_j v - g_j} \right) & \text{if } Hv > g \\ \infty & \text{else}. \end{cases},$$

where $H_j$ is the $j^{th}$ row of $H$ and $g_j$ is the $j^{th}$ component of $g$.

### 3.2 Estimating Equations for Approximate MLE-Based Inference

A natural starting point for selective inference in the multi-task setting is the law for $Y$, conditioning on the event

$$\{ \hat{E} = E, \hat{S} = S \}.$$ 

This conditional prescription results in practical selective inference procedures for other $\ell_1$-regularized algorithms by constraining the response to fall within a polyhedral partition of the sample space. Consider, for example, the randomized LASSO procedure described in section 2.1. The event (10), characterized through the K.K.T. conditions, induces an affine map from the randomization variable $\omega$ to the absolute coefficients $b$ and subgradient $u$ at the optimal solution. Under this transformation, the Jacobian contributes only a proportionality constant to the conditional law of $(y, b, u)$ given (10), yielding a multivariate Gaussian distribution truncated to a polyhedral partition. The distribution for $y$ that results from conditioning upon $u$ and marginalizing over $b$ can be used to facilitate approximate MLE-based inference (Panigrahi and Taylor, 2022).

Unfortunately, this conditional prescription does not generalize well to the MTL setting. Note that the stationary map for the model selection procedure in Algorithm 1, given $E$ and $S$, induces a relationship between $W = \left( \omega^{(1)}', \ldots, \omega^{(K)}', \omega^{(1)}', \ldots, \omega^{(K)}', \omega^{(1)}', \ldots, \omega^{(K)}', \omega^{(1)}', \ldots, \omega^{(K)}' \right)'$, the collection of randomization variables, and $(B, U)$. This relationship implies a transformation of the form:

$$W = \begin{pmatrix} \pi^{(1)}(b^{(1)}, u^{(1)}) & \pi^{(2)}(b^{(2)}, u^{(2)}) & \cdots & \pi^{(K)}(b^{(K)}, u^{(K)}) \end{pmatrix},$$
where

\[
\pi^{(k)}(b^{(k)}, u^{(k)}) = -X^{(k)\prime}y^{(k)} + \left[ (X^{(k)}_{E_k})'X^{(k)}_{E_k} + \epsilon \cdot I_{\delta_k} \right] \text{Diag}(s^{(k)})b^{(k)} + \text{Diag}(\Lambda^{(k)}) \begin{pmatrix} s^{(k)} \\ u^{(k)} \end{pmatrix},
\]

\[
\Lambda^{(k)}_{jk} = \min \left\{ \lambda_0, \lambda \cdot \left( \sum_{k=1}^{K} |(\Theta_j^{(k)})|^{-\frac{1}{2}} \right) \right\}, \quad j \in [p].
\]

Observe that this transformation is non-affine since the penalty term is now related to the solution. The change-of-variables Jacobian, given in Proposition 1 of the Appendix, is a complicated function of \((B, U)\). Deriving estimating equations for maximum-likelihood inference using the law of \((Y, B, U)\) conditional upon \((10)\) would require closed-form expressions for the partial derivatives of the Jacobian, which are not available. For completeness sake, we provide the likelihood based on this conditional law in Proposition 2 under the Appendix.

Instead, we propose a different approach that can bypass the intractable Jacobian and avoid cumbersome numerical integrations to easily facilitate maximum likelihood inference. We will work with an exact selection-adjusted likelihood, derived by conditioning on the refined event

\[
\left\{ \hat{E} = E, \hat{S} = S, \hat{\Gamma} = \Gamma, \hat{U} = U \right\},
\]

which is a proper subset of the event in \((10)\). We form our estimating equations for maximum likelihood inference in terms of the least squares estimator based on the selected predictors for each task,

\[
\tilde{\beta}^{(k)}_{E_k} = (X^{(k)}_{E_k})'y^{(k)}.
\]

Note that the estimator in \((13)\) is the naive MLE that we would have used if the sets \(E_k\) were specified before looking at the data. Dependent on \(X^{(k)\prime}y^{(k)}\), the event of selection also relies on

\[
\tilde{\beta}^{(k)}_{\perp} = (X^{(k)})' \left( I_{nk} - X^{(k)}_{E_k} (X^{(k)}_{E_k})' \right) y^{(k)},
\]

the ancillary statistic we obtain through a projection of the response onto the subspace orthogonal to the span of the selected predictors \(X^{(k)}_{E_k}\).

Lemma 3.1 first identifies an equivalent representation for the refined conditioning event in terms of \(V, \Gamma\) and \(U\) that we observe after solving the MTL Algorithm 1; please see Section 3.1 for a complete
Lemma 3.1 Suppose $V, \Gamma$ are defined as above. Then the event (12) is equivalent to the event

$$\left\{ V > 0, \hat{\Gamma} = \Gamma, \Gamma - DV > 0, \hat{U} = U \right\}.$$ 

The proof can be found in the Appendix. Consider the bijective mapping $\Psi_A : \mathbb{R}^{Kp} \to \mathbb{R}^{Kp}$ such that

$$(B' \ U)' = \Psi_A \left( V' \Gamma \ U \right) = \text{Diag}(A, I_{(Kp - \delta)}) \left( V' (\Gamma - DV)' \ U \right)'.$$ 

Applying a change of variables via the composite mapping,

$$W \xrightarrow{(H_{X'V^o\Psi_A})^{-1}} \left( V' \Gamma' \ U \right),$$ 

we obtain an exact selection-adjusted likelihood in Theorem 3.2 after conditioning on the event in (12). The alternate characterization for the refined conditioning event in terms of the new variables $V, \Gamma$ and $U$ yields a selection-adjusted likelihood function that no longer involves the Jacobian, the term which previously hindered our attempts to solve the estimating equations. The normalizing constant for our refined conditioning event is simply a Gaussian integral over a support set that is characterized by exactly $\delta$ linear inequalities.

Theorem 3.2 Consider the model in (9). The likelihood obtained from the law of the least squares estimates based on $\left\{ X_{E_k}^{(k)}, y^{(k)} \right\}_{k=1}^K$ after conditioning upon the event in Lemma 3.1 is given by

$$\left( \int \rho \left( \tilde{\beta}; L\beta_E + m, \Sigma \right) \cdot \rho(\tilde{V}; P\tilde{\beta} + q, \Delta) \cdot 1(\tilde{H}\tilde{V} \geq g) \, d\tilde{V} \, d\tilde{\beta} \right)^{-1} \cdot \rho \left( \tilde{\beta}_E; L\beta_E + m, \Sigma \right).$$

Expressions for the matrices $L, m, \Sigma, P, q, \text{ and } \Delta$ are provided in the Appendix. To develop an easily solvable system of estimating equations for the MLE and the inverse observed Fisher information matrix, $\hat{I}^{-1}$, we bypass the integration in the normalizer, simply approximating it with the mode of the integrand in the selection region. That is,

$$\log \int \rho \left( \tilde{\beta}; L\beta_E + m, \Sigma \right) \cdot \rho(\tilde{V}; P\tilde{\beta} + q, \Delta) \cdot 1(\tilde{H}\tilde{V} \geq g) \, d\tilde{V} \, d\tilde{\beta} 
\approx -\inf_{\tilde{\beta}, \tilde{V}} \left\{ \frac{1}{2} (\tilde{\beta} - L\beta_E - m)' \Sigma^{-1} (\tilde{\beta} - L\beta_E - m) 
+ \frac{1}{2} (\tilde{V} - P\tilde{\beta} - q)' \Delta^{-1} (\tilde{V} - P\tilde{\beta} - q) + \phi_{H,g}(\tilde{V}) \right\}$$ 

(15)
ignoring an additive constant. The approximation in (15) then lends itself towards tractable equations for the selective MLE, $\hat{\beta}_{E}^{\text{MLE}}$, and the inverse observed Fisher information matrix, $\hat{I}^{-1}$, given in Theorem 3.3.

**Theorem 3.3** Under the modeling assumptions in Theorem 3.2, the approximate selective MLE and the observed information matrix satisfy the following system of estimating equations.

\[
\hat{\beta}_{E}^{\text{MLE}} = L^{-1} \hat{\beta}_{E} + L^{-1} \Sigma P' \Delta^{-1} (P \hat{\beta}_{E} + q - \hat{V}) - L^{-1} m,
\]

\[
\hat{I}^{-1} = L^{-1} \Sigma L'^{-1} + L^{-1} \Sigma \left( P' \Delta^{-1} P - P' \Delta^{-1} \left( \Delta^{-1} + \nabla^{2} \phi_{H,g}(\hat{V}) \right)^{-1} \Delta^{-1} P \right) \Sigma L'^{-1},
\]

where $\hat{V}$ is obtained from solving

\[
\hat{V} = \arg\min_{V} \frac{1}{2} (V - P \hat{\beta}_{E} - q)' \Delta^{-1} (V - P \hat{\beta}_{E} - q) + \phi_{H,g}(\hat{V}).
\] (16)

With these estimators for the approximate MLE and the inverse observed Fisher information matrix, it is now possible to use maximum likelihood inference to form $100 \times (1 - \alpha)$% confidence intervals for the parameters within the MTL model (9). Algorithm 2 summarizes our procedure for post-MTL inference by reusing the same data.

**Algorithm 2** Multi-Task Model Selection and Inference

17: Record the values for $\hat{E}$, $\hat{S}$, $\hat{\Gamma}$, and $\hat{U}$ at convergence of Algorithm 1
18: From the estimated sparsity structure, compute $\hat{\beta}_{E}^{(k)}$ and $\hat{\beta}_{\perp}^{(k)}$ for $k \in [K]$
19: Specify a significance level $\alpha$
20: Compute $P$, $q$, and $\Delta$
21: Optimize (16) with gradient descent to compute $\hat{V}$
22: Compute the matrices $L$, $m$, $\Sigma$
23: **procedure** MAXIMUM LIKELIHOOD INFERENCE
24: Find $\hat{\beta}_{E}^{\text{MLE}}$ and $\hat{I}^{-1}$ based on the estimating equations in Theorem 3.3
25: for $j \in \{1, \ldots, \delta\}$ do
26: Compute interval $\hat{\beta}_{E,j}^{\text{MLE}} \pm z_{1 - \alpha/2} \sqrt{\hat{I}_{jj}^{-1}(\hat{\beta}_{E}^{\text{MLE}})}$
27: end for
28: **end procedure**

4 Analysis of Neurocognitive Data from the ABCD Study

The Adolescent Brain Cognitive Development (ABCD) study, discussed in the introduction, is a large longitudinal study undertaken to characterize typical cognitive development in adolescence (Lu-
ciana et al., 2018; Volkow et al., 2018; Karcher and Barch, 2021). Researchers are interested in better understanding the organization of cognitive abilities during childhood development, i.e., whether there is a single, general ability or separable fluid and crystallized abilities. Existing work has relied almost exclusively on behavioral data and has reached equivocal findings. Some studies have found evidence for a single strong, general factor of cognitive ability in youth (Juan-Espinosa et al., 2000; Gignac, 2014). Other studies have concluded that a two-factor model of intelligence performs better than a single-factor model for children and adolescents, with fluid intelligence and crystallized intelligence becoming more differentiated with age (Hülür et al., 2011; Simpson-Kent et al., 2020).

Studies that leverage neurological data in studying the organization of cognitive abilities are extremely rare. Two exceptions are Tadayon et al. (2020) and Simpson-Kent et al. (2020), who examined cortical morphology and white matter, respectively. These structural modalities, however, tend to have much weaker associations with cognitive abilities than resting-state fMRI (Marek et al., 2022; Chen et al., 2022). In addition, these studies used traditional mass univariate approaches, treating all edge weights as a bag of features. We instead adopt a multi-task approach that, as we have argued, is better suited for the dual goals of prediction and cartographic mapping.

4.1 Multi-Task Framework for Studying ABCD Data

Applied to the ABCD data, our multi-task learning and selective inference procedure, MTL + SI, offers a novel approach to identifying the neurobiological underpinnings of fluid and crystallized intelligence in the developing brain. We have applied MTL + SI to the second ABCD release (the study is ongoing), using the same inclusion/exclusion criteria as Sripada et al. (2021). This leaves data for 5,937 subjects from 19 different research sites throughout the US. We limit the analysis to four of the tasks from the ABCD neurocognitive battery, two measuring fluid intelligence (the List Sorting Working Memory Test from the NIH Toolbox for the Assessment of Neurological and Behavioral Function and the Matrix Reasoning subtest from the Wechsler Intelligence Test for Children) and two measuring crystallized intelligence (the Picture Vocabulary Test and the Reading Comprehension Test from the same NIH Toolbox).

The predictors in our multi-task regression are neurological factors extracted from resting-state fMRI data by estimating the connections between 418 regions of interest (ROIs, or nodes) in the brain. These ROIs were identified based on the Gordon cortical parcellation (Gordon et al., 2016), augmented with additional subcortical and cerebellar atlases. The 418 ROIs are further classified into 15 functional groups, ranging in size from 4 to 54 nodes; see Table 3 for a complete list. These groups of ROIs, equivalent to communities in the statistical network analysis literature, are themselves called networks in neuroimaging, an unfortunate terminology overload.

To assess the strength of connection between each pair of ROIs, the Pearson correlation coefficient
is computed between the fMRI blood oxygen level dependent (BOLD) signals at those ROIs. We use the inter-node correlations from Sripada et al. (2021), computed after pre-processing the fMRI time series data to correct for nuisance covariates, such as physiological noise and head motion, through a standard pipeline that includes FreeSurfer normalization, ICA-AROMA denoising, CompCor correction, and omission of high-motion frames. With 418 nodes, each scan corresponds to $418 \times 417 / 2 = 87,153$ features. A standard approach to resting-state fMRI data is to replace the edge weights with top principal component scores (Cordes and Nandy, 2006), which both reduces dimensionality and helps with the low signal-to-noise ratio. We retain the first 500 principal component scores computed from the correlations to use as features in multi-task learning. Previous work has shown that a small number of components is typically sufficient to capture most inter-individual variation in functional connectivity (Sripada et al., 2019) and predict differences in cognition (Sripada et al., 2020).

After standardizing the response variables, we consider two different ways of identifying the principal components for each task through multi-task learning. One approach, which we refer to as joint, is to apply Algorithm 1 to all four tasks. This is the default approach to multi-task learning on both the simulated and real data when not otherwise specified. The alternative approach, which we refer to as pairwise, is to apply Algorithm 1 separately to the two tasks that measure fluid intelligence and the two tasks that measure crystallized intelligence. Comparing the two approaches allows us to assess how much joint learning of the fluid and crystallized tasks can improve detection of shared neurological structure. We proceed to fit a multi-task model on the selected principal components for each approach and construct confidence intervals for the best linear coefficients using Algorithm 2. A consistent plug-in estimator is used to approximate the noise level, following the recommendation of Tian and Taylor (2018). We use 80% of the original data for model selection and inference, hold out another 10% for selecting the tuning parameter for penalized multi-task regression, and use the last 10% as test data.

4.2 Validation of MTL+SI Through Simulation

We first use synthetic data with the same dimensions and estimated sparsity level as the fMRI features to investigate the efficacy of MTL + SI in recovering signals and estimating their strength. We generate data from the linear regression model in (9) with $K = 4$ and noise variance 1. Each predictor matrix is simulated by drawing 6000 samples from a multivariate Gaussian distribution of dimension $p = 500$, with all means equal to zero and covariance given by the identity matrix. Simulation results with non-Gaussian errors are reported in Section 3 of the Appendix.

The coefficients $\beta \in \mathbb{R}^{p \times K}$ in the MTL model are chosen based on two parameters, the global sparsity level $s_G$ and the task-specific sparsity level $s_T$, both numbers between 0 and 1. We define the global sparsity as the percentage of predictors that are zero for all of the tasks. The task sparsity, meanwhile, quantifies the average number of tasks that do not share any one of the global signals. Note
that a task-level sparsity value $s_T = 0$ would mean that all tasks have the same active predictors, and higher levels of $s_T$ indicate that the predictors used by each task are more heterogeneous. For given levels of $s_G$ and $s_T$, the entries of $\beta$ are determined as follows.

1. Select $[s_Gp]$ predictors to be globally null for every task, and set the corresponding rows of $\beta$ to zero.

2. For the remaining set of globally active predictors, set an average of $[s_TK]$ entries of each row of $\beta$ to zero.

3. For each predictor $j \in [p]$, specify the corresponding non-zero coefficients by sampling without replacement from an equally-spaced sequence of values covering the interval

$$[\sqrt{2 \log(p)}, \sqrt{6 \log(p)}].$$

Randomly assign each coefficient a sign drawn from $\{-1, 1\}$.

The task sparsity level is of primary scientific interest for its ability to measure the extent of feature-sharing between tasks, so we investigate the consequences of varying $s_T$ while fixing $s_G$. Our results from applying MTL+SI jointly to the four ABCD tasks, presented in the following sections, indicate that 77 of the top 500 principal components are significant for one or more tasks following selective inference, so we fix $s_G = 0.85$. We estimate that the task sparsity level of the real data is $s_T = 0.416$, but we consider three different task sparsity levels for our simulations: 0.25, 0.375, and 0.5. The simulations with a task sparsity level of 0.25 and 0.5 are designed so that each predictor in the globally active set is shared by three tasks and two tasks, respectively. For the simulations with a task sparsity level of 0.375, the globally active predictors are specified so that half are shared by three tasks and the remaining half are shared by two tasks. We conduct 100 simulation repetitions under each of the three data-generating models, corresponding to the three different sparsity levels. The tuning parameter is chosen to minimize the average MSE on the validation set across iterations, and the results are reported at that tuning parameter value for all 100 iterates.

To assess both multi-task learning and selective inference when there is shared structure across tasks, we compare MTL + SI against two alternative approaches, MTL with data splitting and single-task LASSO with selective inference. Each procedure is described below.

1. **MTL($v$) + SI**: Our proposed approach. Use Algorithm 1 to select an MTL model based on independent Gaussian randomization variables, $\omega^{(k)} \sim N(0, v^2 \sigma^2 \cdot I)$ for $k \in [K]$, and construct selective inference (SI) confidence intervals by Algorithm 2.
2. Data Splitting (DS($s$)): Divide the training data into two parts, using $[sn_k]$ samples from each task $k \in [K]$ for model selection with the usual multi-task algorithm, and reserve the rest for inference about the selected predictors as described in Cox (1975).

3. LASSO($v$) + SI: Apply the randomized LASSO separately to each task using independent randomization variables, $\omega^{(k)} \sim N(0, v^2 \sigma^2 \cdot I)$ for $k \in [K]$, and proceed with SI using the maximum likelihood approach of Panigrahi and Taylor (2022).

We consider two common choices for the data splitting parameter, $s = 0.5$ and $s = 0.67$. Panigrahi et al. (2021) show that there is a rough equivalence between data splitting and a randomization variable with variation parameter $v = \sqrt{(1 - s)/s}$ for a Gaussian response. We find that the variation parameter $v = 0.7$, corresponding to $s = 0.67$, strikes a slightly more optimal balance between the quality of model selection and the quality of inference than the variation parameter $v = 1.0$, corresponding to $s = 0.5$. Thus, we set $v = 0.7$ for all selective inference methods.

We use several metrics to evaluate the performance of each method. The empirical coverage rate (CR) of the confidence intervals is computed for the non-zero coefficients, defined by

$$CR = 1 - \frac{\{j \in \hat{E} : \beta_{E,j} \not\in C_{E,j}\}}{\max(|\hat{E}|, 1)}.$$ 

This rate is further averaged over replications. We assess the inferential power of each method by reporting the confidence interval lengths for the selected parameters. To measure the overall accuracy of model selection and subsequent inference, we compute the F1 score, defined as

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

For our purposes, precision is the proportion of truly active predictors among those that were both selected into the model and deemed significant, that is, the effects with confidence intervals that did not cover zero. Recall is the proportion of all truly active predictors that were both selected into the model and deemed significant after inference. Letting $E_0$ be the set of true active predictors,

$$\text{Precision} = \frac{|E_0 \cap \{j \in \hat{E} : 0 \not\in C_{E,j}\}|}{|\{j \in \hat{E} : 0 \not\in C_{E,j}\}|}; \quad \text{Recall} = \frac{|E_0 \cap \{j \in \hat{E} : 0 \not\in C_{E,j}\}|}{|E_0|}.$$ 

Figure 9 shows the distribution of coverage, interval length, and F1 score for MTL(0.7) + SI and the alternative methods. We observe that all methods achieve a nominal coverage level of 90%. In terms of interval length, MTL(0.7) + SI has a large advantage over DS(0.67), which asymptotically reserves
MTL(0.7) + SI also has similar or better performance than DS(0.5) in terms of interval length, even though the latter method asymptotically reserves more information for inference. Of the metrics we report, the F1 score provides the most direct comparison between the approaches, capturing both the validity of the chosen model and the significance of the results. We find that MTL(0.7) + SI achieves a higher median F1 score than all three alternatives for each task sparsity level. This indicates that similar tasks should be trained together whenever possible and that selective inference is a more optimal method for quantifying the uncertainty in models chosen through joint learning than sample splitting. Additional simulations which vary other parameters in the design—dimensions and global sparsity—are collected in the Appendix; please see Section 3.

**Figure 1:** Comparison of MTL(0.7) + SI against three alternative approaches: DS(0.67), DS(0.5), and LASSO(0.7) + SI, with 100 replications for each level of task sparsity $s_T$. Global sparsity is fixed at $s_G = 0.85$. The asterisk (*) denotes the task sparsity level closest to the fMRI data. Outlier points are not shown to improve readability. Left: coverage; Middle: interval length; Right: F1 score.

Finally, we compare the results of applying MTL(0.7) + SI jointly to four simulated tasks against the results of applying MTL(0.7) + SI separately to two pairs of simulated tasks when a) there is shared structure across all tasks and b) when the shared structure is only present within the two pairs. To test setup (a), we randomly assign each active predictor to two tasks, ensuring that any two of the four tasks have roughly the same number of common predictors. To test setup (b), we instead specify the active predictors so that the two related tasks share a common set of predictors and the unrelated pairs have no common predictors. Note that $s_T = 0.5$ in both setups, with only the relationship between tasks changing. Figure 2 indicates that the joint approach produces shorter intervals and a higher median F1 score under setup (a), while the joint and pairwise approaches have a similar median confidence interval length and F1 score under setup (b). These results confirm that the joint approach improves the quality of model selection and inference when there is common structure across tasks, without doing any real harm when there is no shared structure.
Figure 2: Comparison of the joint and pairwise approaches under setup (a), where any two tasks have roughly the same amount of common structure, and setup (b), where only the two pairs have common structure. Global and task sparsity are fixed at $s_G = 0.85$ and $s_T = 0.5$. Outliers are not shown. Left: coverage; Middle: interval length; Right: F1 score.

4.3 Assessing the Neurological Organization of Different Cognitive Abilities

We now apply our methods to the ABCD data, again comparing the joint and pairwise approaches to MTL(0.7) + SI. The left panel of Figure 3 shows the predictive performance of each method on validation data. Following the convention in the neuroimaging literature (Sripada et al., 2020, 2021), we measure the predictive performance of each method by the so-called predictive $r$, the correlation between predictions and observed responses on out-of-sample data. All tuning parameters are chosen to maximize the average predictive $r$ across tasks on the validation data. Table 1 reports the final performance of each method on the testing data. The joint approach generally maintains some advantage across all tasks, and the observed correlations are consistent with the expectations of domain experts and previous findings (Sripada et al., 2021).

The benefit of training all four tasks together is even more apparent when comparing the inferential power of the two methods. The right panel of Figure 3 shows the confidence interval lengths obtained under each approach. We observe that applying MTL(0.7) + SI jointly to all four tasks results in a substantially smaller median confidence interval length than applying MTL(0.7) + SI separately to the two crystallized and the two fluid intelligence tasks. Overall, the results indicate that the joint approach can better detect some neurological features while also matching or exceeding the predictive performance of the two separate multi-task models for fluid and crystallized intelligence.

Although the predictive advantage of the joint approach is slight, the chosen features are quite different from those recovered through the pairwise approach. Training all of the tasks together yields four very similar models, indicating that many of the same neurological features could underlie differ-
Table 1: Predictive correlations computed on the testing data for the joint and pairwise approaches, where MTL(0.7) + SI is applied, respectively, to all four tasks or separately to the pairs of fluid/crystallized tasks.

| Task          | Joint Predictive $r$ | Pairwise Predictive $r$ |
|---------------|----------------------|-------------------------|
| NIH Tlx RC    | 0.335                | 0.321                   |
| NIH Tlx PV    | 0.470                | 0.451                   |
| Matrix Reasoning | 0.283               | 0.273                   |
| NIH Tlx LS    | 0.354                | 0.317                   |

Figure 3: Left: The predictive $r$ for the joint and pairwise approaches to MTL(0.7) + SI. Right: The distribution of confidence interval lengths across tasks for the joint and pairwise approaches to MTL(0.7) + SI.

To quantify the structural overlap between any two tasks, we use the Jaccard index to measure the similarity of the significant features for those tasks, where significant features are the PCs with confidence intervals that do not contain zero. Figure 4 shows the Jaccard similarity of the four tasks under both the joint and the pairwise approaches to model selection and inference. When applied separately to the two pairs of tasks, MTL(0.7) + SI reveals a significant amount of shared structure between the two crystallized intelligence tasks and the two fluid intelligence tasks, as well as a lesser but still notable amount of shared structure across the two pairs of tasks. The joint approach to model selection and inference, however, reveals a significant amount of common structure between any two of the four tasks. Selective inference seems to confirm the existence of the shared signals identified through joint model selection, suggesting that there may be a common neurological basis for different cognitive abilities that can be best understood through joint multi-task learning.

We have also compared MTL(0.7) + SI to data splitting, applying DS(0.67) and DS(0.5) jointly to all four tasks. MTL(0.7) + SI seems to improve model selection relative to data splitting, with more of the selected features surviving inference. As shown in Figure 5, MTL(0.7) + SI also yields a shorter
Figure 4: Similarity of selected models for the four tasks. Left: Jaccard index between sets of significant PCs recovered through the joint approach. Right: Jaccard index between sets of significant PCs recovered through the pairwise approach.

median confidence interval length than data splitting, both overall and for predictors that are selected by both MTL(0.7) + SI and either DS(0.67) or DS(0.5). Note that some of the predictors vary between approaches since each method performs its own model selection; however, rough comparisons may still be possible when there is substantial overlap in the selected sets. Table 2 reports the average number of features selected and further deemed significant across the four tasks for each method, as well as the average number of shared features between selective inference and each data splitting procedure. The overlap is significant, making the comparison more meaningful.

| Method               | # selected | # significant |
|----------------------|------------|---------------|
| MTL (0.7) + SI       | 89.25      | 45.00         |
| DS (0.67)            | 114.25     | 39.75         |
| Common DS (0.67)     | 69.50      | 29.25         |
| DS (0.5)             | 78.75      | 35.50         |
| Common DS (0.5)      | 53.25      | 30.50         |

Table 2: Comparison of selected models, in average number of predictors per task

4.4 Neurological Interpretation of Selected Models

We now offer an interpretation of the shared structure identified through multi-task learning and validated through selective inference. Since experts are ultimately interested in understanding the effects of the connectome, we aim to translate the results back into the original 87,153-dimensional connectome feature space. Note that the mean structure for task $k$ in the original feature space of
Figure 5: Comparison of inferential power between MTL(0.7) + SI, DS(0.67), and DS(0.5) when applied jointly to all four tasks. Left: Boxplots of interval lengths. Right: Ratio of interval lengths for common parameters.

dimension $p$ can be represented as

$$X\beta^{(k)} = XM'\beta^{(k)} = Z\theta^{(k)},$$

where $M$ is the orthogonal matrix of eigenvectors for $X'X$, $Z$ is the matrix of principal component scores, and $\theta^{(k)} = M'\beta^{(k)} \in \mathbb{R}^p$. To recover $\beta^{(k)}$, a potential plug-in estimator is

$$\hat{\beta}^{(k)} = M\left(\hat{\theta}^{(k)}_{E_k} \quad 0\right) = M_{E_k}\hat{\theta}^{(k)}_{E_k},$$

where $\hat{\theta}^{(k)}_{E_k} \in \mathbb{R}^{\delta_k}$ is the MLE for task $k$ obtained from Algorithm 2. Although this estimator is only consistent for $\beta^{(k)}$ when $\delta_k = p$, we will use it to approximate $\beta^{(k)}$ as is typically done in ordinary principal components regression (Jolliffe, 2003).

Figure 6 shows the results from applying MTL(0.7) + SI jointly to all four tasks when projected back into the original feature space. There is notable similarity of connectivity patterns across all of the four tasks, providing some support for the general factor model. Key connectivity motifs include stronger positive and negative connections within the default mode network (DMN) and the visual network, as well as increased positive connectivity within the cerebellum. We also observe complex patterns of connectivity changes between regions/networks, especially between the frontoparietal network (FPN) and DMN; the auditory network and the somatomotor network; the cingulo-parietal network (CP) and the retrosplenial network (RSP); and RSP and DMN.

Our results show some agreement with a recent study that identified FPN, DMN, the dorsal attention network, and the visual network as influential in predicting general intelligence (Tong et al., 2022).
FPN is involved in flexible adaptive control (Cole et al., 2013), and a number of previous studies implicate it in executive functions and cognitive control (Cole and Schneider, 2007; Niendam et al., 2012), constructs closely related to the general factor of intelligence (Chen et al., 2019). DMN is involved in spontaneous thought (Andrews-Hanna et al., 2014) and semantic/conceptual representation (Wirth et al., 2011; Binder and Desai, 2011; Binder et al., 2009), capacities that likely facilitate abstraction and problem-solving. Cerebellum has been traditionally associated with coordination of movement (Spencer et al., 2005; Bastian, 2006), but there is growing recognition that it coordinates both external motor operations as well as internal mental operations, and thus it plays a critical role in supporting complex cognition (Schmahmann, 2019, 1996; Andreasen et al., 1999). CP and RSP are small networks that were identified relatively recently (Gordon et al., 2017), and their significance for higher cognitive functions requires further elucidation.

| Abbreviation | Full Region/Network Name        | Abbreviation | Full Region/Network Name        |
|--------------|---------------------------------|--------------|---------------------------------|
| SMH          | Somatomotor-Hand                 | SMM          | Somatomotor-Mouth               |
| CO           | Cingulo-Opercular                 | AUD          | Auditory                        |
| DMN          | Default                          | VIS          | Visual                          |
| FPN          | Frontoparietal                   | SAL          | Salience                        |
| SC           | Subcortical                      | VAN          | Ventral Attention               |
| DAN          | Dorsal Attention                 | CER          | Cerebellum                      |
| MONE         | Not Named                        | CP           | Cingulo-Parietal                |
| RST          | Retrosplenial Temporal           |              |                                 |

Table 3: Names and abbreviations for the 15 groups of ROIs, known as regions/networks

Overall, we find compelling evidence for shared connectivity patterns between the two fluid intelligence tasks and the two crystallized intelligence tasks. A joint approach to model selection recovers more shared structure across all four tasks than performing model selection separately, and selective inference provides additional statistical reassurance that the shared structure we identified through MTL reflects real patterns in the data. Our approach contributes a new perspective to the literature on theories of intelligence, using neurological data to discern the relationships between tasks rather than inferring them from behavioral data.

5 Discussion

In this paper, we address two key limitations of previous research on the role of the functional connectome in human cognition. Most previous studies have investigated the connectivity patterns associated with general cognitive ability, inferred from behavioral data, or the connectivity patterns associated with specific cognitive abilities like working memory. By contrast, we leverage a MTL approach to discover the relationship between cognitive domains entirely from shared patterns in the
Figure 6: Cartographic visualization showing standardized estimates of the original coefficients, grouped by region/network, that were obtained after applying MTL(0.7) + SI jointly to all four tasks. A threshold of 1.5 is used to improve readability, with smaller-magnitude coefficients set to zero.
brain-behavior data, avoiding the need to make any inferences about general ability from behavioral
data alone. Most prior studies on the role of the functional connectome in human cognition have
also focused on predictive power, neglecting statistical inference. We address this shortcoming by
developing selective inference procedures to measure the strength and certainty of each brain-behavior
relationship discovered from the data, offering improved interpretability.

By applying our selective inference procedures for MTL to two tasks from the ABCD study that
implicate fluid intelligence and two tasks from the ABCD study that implicate crystalized intelligence,
we uncover additional shared structure that cannot be detected by modeling the fluid and crystalized
tasks separately. This shared structure provides initial support for a general factor model of cognitive
abilities. After applying MTL+ SI to the data, we use cartographic mapping to visualize the brain
regions involved in each of the four tasks. Our results reveal that connections involving DMN, FPN,
and visual network generally have the most predictive power across tasks, showing some agreement
with previous studies.

6 Acknowledgments

We would like to thank Ji Zhu and Daniel Kessler for their helpful feedback throughout the project;
Qianhua Shan for guidance in accessing the ABCD dataset; Aman Taxali for generating the carto-
graphic maps and helping to run code in parallel on the University of Michigan’s high-performance
computing cluster; Michael Angstadt for creating the NDA study associated with this paper; and Tian
Xie and Qiang Chen for testing different simulations as part of their undergraduate research project.

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive De-
velopment (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite,
longitudinal study designed to recruit more than 10,000 children age 9-10 and
follow them over 10 years into early adulthood. The ABCD Study is supported by the National Insti-
tutes of Health and additional federal partners under award numbers U01DA041022, U01DA041028,
U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134,
U01DA041148, U01DA041156, U01DA041174, U24DA041123, and U24DA041147. A full list of
supporters is available at https://abcdstudy.org/nih-collaborators. A listing of par-
ticipating sites and a complete listing of the study investigators can be found at https://abcdstudy.
org/consortium_members/. ABCD consortium investigators designed and implemented the
study and/or provided data but did not necessarily participate in analysis or writing of this report. This
manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH
or ABCD consortium investigators. The ABCD data repository grows and changes over time. The
ABCD data used in this report came from NDA Study 721, 10.15154/1504041, which can be found at
https://nda.nih.gov/study.html?id=721. The specific NDA study associated with this
S. Panigrahi’s research is supported in part by NSF grants 1951980 and 2113342. N. Stewart is supported in part by NSF RTG grant 1646108 and a Rackham Science Award from the University of Michigan. E. Levina’s research is supported in part by NSF grants 1916222, 2052918, and 2210439 and NIH grant R01MH123458.

References

Adeli, E., Meng, Y., Li, G., Lin, W., and Shen, D. (2019). Multi-task prediction of infant cognitive scores from longitudinal incomplete neuroimaging data. *NeuroImage*, 185:783–792.

Anderson, E. D. and Barbey, A. K. (2022). Investigating cognitive neuroscience theories of human intelligence: A connectome-based predictive modeling approach. *Human Brain Mapping*.

Andreasen, N. C., Nopoulos, P., O’Leary, D. S., Miller, D. D., Wassink, T., and Flaum, M. (1999). Defining the phenotype of schizophrenia: cognitive dysmetria and its neural mechanisms. *Biological psychiatry*, 46(7):908–920.

Andrews-Hanna, J. R., Smallwood, J., and Spreng, R. N. (2014). The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the new York Academy of Sciences*, 1316(1):29–52.

Bastian, A. J. (2006). Learning to predict the future: the cerebellum adapts feedforward movement control. *Current opinion in neurobiology*, 16(6):645–649.

Bi, J., Xiong, T., Yu, S., Dundar, M., and Rao, R. (2008). An improved multi-task learning approach with applications in medical diagnosis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 117–132.

Binder, J. R. and Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in cognitive sciences*, 15(11):527–536.

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796.

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences*, 29(2):109–125.
Carroll, J. B. et al. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Number 1. Cambridge University Press.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological bulletin*, 40(3):153.

Chen, J., Tam, A., Kebets, V., Orban, C., Ooi, L. Q. R., Asplund, C. L., Marek, S., Dosenbach, N. U., Eickhoff, S. B., Bzdok, D., et al. (2022). Shared and unique brain network features predict cognitive, personality, and mental health scores in the abcd study. *Nature communications*, 13(1):2217.

Chen, Y., Spagna, A., Wu, T., Kim, T. H., Wu, Q., Chen, C., Wu, Y., and Fan, J. (2019). Testing a cognitive control model of human intelligence. *Scientific reports*, 9(1):1–17.

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., and Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature neuroscience*, 16(9):1348–1355.

Cole, M. W. and Schneider, W. (2007). The cognitive control network: Integrated cortical regions with dissociable functions. *Neuroimage*, 37(1):343–360.

Cordes, D. and Nandy, R. R. (2006). Estimation of the intrinsic dimensionality of fMRI data. *Neuroimage*, 29(1):145–154.

Cox, D. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., and Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671.

Fithian, W., Sun, D., and Taylor, J. (2017). Optimal inference after model selection.

Fraenz, C., Schlüter, C., Friedrich, P., Jung, R. E., Güntürkün, O., and Genç, E. (2021). Interindividual differences in matrix reasoning are linked to functional connectivity between brain regions nominated by parieto-frontal integration theory. *Intelligence*, 87:101545.

Gignac, G. E. (2014). Dynamic mutualism versus g factor theory: An empirical test. *Intelligence*, 42:89–97.

Gordon, E. M., Laumann, T. O., Adeyemo, B., Gilmore, A. W., Nelson, S. M., Dosenbach, N. U., and Petersen, S. E. (2017). Individual-specific features of brain systems identified with resting state functional correlations. *Neuroimage*, 146:918–939.
Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., and Petersen, S. E. (2016). Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral cortex, 26*(1):288–303.

Gray, J. R., Chabris, C. F., and Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature neuroscience, 6*(3):316–322.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika, 98*(1):1–15.

Hagler Jr, D. J., Hatton, S., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., Sutherland, M. T., Casey, B., Barch, D. M., Harms, M. P., et al. (2019). Image processing and analysis methods for the Adolescent Brain Cognitive Development study. *Neuroimage, 202*:116091.

Hearne, L. J., Mattingley, J. B., and Cocchi, L. (2016). Functional brain networks related to individual differences in human intelligence at rest. *Scientific reports, 6*(1):1–8.

Hong, L., Kuffner, T. A., and Martin, R. (2018). On overfitting and post-selection uncertainty assessments. *Biometrika, 105*(1):221–224.

Horn, J. L. and Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory.

Hüüür, G., Wilhelm, O., and Robitzsch, A. (2011). Intelligence differentiation in early childhood. *Journal of Individual Differences.

Humphreys, L. G. (1979). The construct of general intelligence.

Jolliffe, I. T. (2003). Principal Component Analysis. *Technometrics, 45*(3):276.

Juan-Espinosa, M., García, L. F., Colom, R., and Abad, F. J. (2000). Testing the age related differentiation hypothesis through the wechsler’s scales. *Personality and Individual differences, 29*(6):1069–1075.

Karcher, N. R. and Barch, D. M. (2021). The ABCD study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology, 46*(1):131–142.

Kivaranovic, D. and Leeb, H. (2018). Expected length of post-model-selection confidence intervals conditional on polyhedral constraints. *arXiv preprint arXiv:1803.01665.*

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference with the lasso. *The Annals of Statistics, 44*(3):907–927.
Liu, K., Markovic, J., and Tibshirani, R. (2018). More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*.

Lozano, A. and Swirszcz, G. (2012). Multi-level lasso for sparse multi-task regression. In *Proceedings of the 29th International Conference on Machine Learning*, pages 595–602.

Luciana, M., Bjork, J., Nagel, B., Barch, D., Gonzalez, R., Nixon, S., and Banich, M. (2018). Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (ABCD) baseline neurocognition battery. *Developmental cognitive neuroscience*, 32:67–79.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660.

Markett, S., Reuter, M., Heeren, B., Lachmann, B., Weber, B., and Montag, C. (2018). Working memory capacity and the functional connectome-insights from resting-state fmri and voxelwise centrality mapping. *Brain imaging and behavior*, 12(1):238–246.

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., and Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience*, 12(2):241–268.

Panigrahi, S. (2018). Carving model-free inference. *arXiv preprint arXiv:1811.03142*.

Panigrahi, S., Mohammed, S., Rao, A., and Baladandayuthapani, V. (2022a). Integrative bayesian models using post-selective inference: A case study in radiogenomics. *Biometrics*.

Panigrahi, S. and Taylor, J. (2022). Approximate selective inference via maximum likelihood. *Journal of the American Statistical Association*, pages 1–11.

Panigrahi, S., Taylor, J., et al. (2018). Scalable methods for bayesian selective inference. *Electronic Journal of Statistics*, 12(2):2355–2400.

Panigrahi, S., Taylor, J., and Weinstein, A. (2021). Integrative methods for post-selection inference under convex constraints. *Annals of Statistics*, 49(5):2803–2824.

Panigrahi, S., Wang, J., and He, X. (2022b). Treatment effect estimation with efficient data aggregation. *arXiv preprint arXiv:2203.12726*.
Schmahmann, J. D. (1996). From movement to thought: anatomic substrates of the cerebellar contribution to cognitive processing. *Human brain mapping*, 4(3):174–198.

Schmahmann, J. D. (2019). The cerebellum and cognition. *Neuroscience letters*, 688:62–75.

Simpson-Kent, I. L., Fuhrmann, D., Bathelt, J., Achterberg, J., Borgeest, G. S., and Kievit, R. A. (2020). Neurocognitive reorganization between crystallized intelligence, fluid intelligence and white matter microstructure in two age-heterogeneous developmental cohorts. *Developmental Cognitive Neuroscience*, 41:100743.

Spearman, C. (1961). "General Intelligence" Objectively Determined and Measured.

Spencer, R., Ivry, R. B., and Zelaznik, H. N. (2005). Role of the cerebellum in movements: control of timing or movement transitions? *Experimental brain research*, 161(3):383–396.

Sripada, C., Angstadt, M., Rutherford, S., Kessler, D., Kim, Y., Yee, M., and Levina, E. (2019). Basic units of inter-individual variation in resting state connectomes. *Scientific reports*, 9(1):1–12.

Sripada, C., Angstadt, M., Taxali, A., Clark, D. A., Greathouse, T., Rutherford, S., Dickens, J. R., Shedden, K., Gard, A. M., Hyde, L. W., et al. (2021). Brain-wide functional connectivity patterns support general cognitive ability and mediate effects of socioeconomic status in youth. *Translational psychiatry*, 11(1):1–8.

Sripada, C., Rutherford, S., Angstadt, M., Thompson, W. K., Luciana, M., Weigard, A., Hyde, L. H., and Heitzeg, M. (2020). Prediction of neurocognition in youth from resting state fMRI. *Molecular psychiatry*, 25(12):3413–3421.

Suzumura, S., Nakagawa, K., Umezu, Y., Tsuda, K., and Takeuchi, I. (2017). Selective inference for sparse high-order interaction models. In *International Conference on Machine Learning*, pages 3338–3347. PMLR.

Tadayon, E., Pascual-Leone, A., and Santarnecchi, E. (2020). Differential contribution of cortical thickness, surface area, and gyriﬁcation to fluid and crystallized intelligence. *Cerebral Cortex*, 30(1):215–225.

Tanizaki, K., Hashimoto, N., Inatsu, Y., Hontani, H., and Takeuchi, I. (2020). Computing valid p-values for image segmentation by selective inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9553–9562.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61.
Tian, X., Panigrahi, S., Markovic, J., Bi, N., and Taylor, J. (2016). Selective sampling after solving a convex problem. arXiv preprint arXiv:1609.05609.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. The Annals of Statistics, 46(2):679–710.

Tong, X., Xie, H., Carlisle, N., Fonzo, G. A., Oathes, D. J., Jiang, J., and Zhang, Y. (2022). Transdiagnostic connectome signatures from resting-state fmri predict individual-level intellectual capacity. Translational Psychiatry, 12(1):367.

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., Pérez-Stable, E. J., Riley, W. T., Bloch, M. H., Conway, K., et al. (2018). The conception of the ABCD study: From substance use to a broad NIH collaboration. Developmental cognitive neuroscience, 32:4–7.

Wang, J., He, X., and Xu, G. (2020). Debiased inference on treatment effect in a high-dimensional model. Journal of the American Statistical Association, 115(529):442–454.

Wang, X., Bi, J., Sun, J., and Song, M. (2016). Multiplicative multitask feature learning. The Journal of Machine Learning Research, 17(1):2820–2852.

Wirth, M., Jann, K., Dierks, T., Federspiel, A., Wiest, R., and Horn, H. (2011). Semantic memory involvement in the default mode network: a functional neuroimaging study using independent component analysis. Neuroimage, 54(4):3057–3066.

Zhang, Y. and Yang, Q. (2021). A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering.

Zhao, Q. and Panigrahi, S. (2019). Selective inference for effect modification: An empirical investigation. Observational Studies, 5(2):131–140.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. The Annals of Statistics, 36(4):1509–1533.
7 Appendix

A Proofs of Results in Section 3

Proof: Lemma 3.1. The proof follows immediately from the definition of $V$, $\Gamma$ and by noting that the constraints $b^{(k)} > 0$ for all $k \in [K]$ are equivalent to

$$V > 0, \Gamma - D V > 0.$$ 

\[\square\]

Proof: Theorem 3.2. Based on the independence between $b^{(k)}_{\beta}, b^{(k)}_{\perp}$ and $W$, the unconditional law for the corresponding variables is given by

$$\prod_{k=1}^{K} \rho \left( \left( \beta^{(k)}_{E_k} \beta^{(k)}_{E_k}, \sigma_{k}^{2} \cdot \left( X^{(k)}_{E_k} X^{(k)}_{E_k} \right)^{-1} \right) \rho \left( \beta^{(k)}_{\perp}, 0, \Sigma^{(k)}_{\perp} \right) \rho \left( \mu^{(k)}_{E_k}, 0, \Omega^{(k)} \right) \right),$$

treating $E_{k}$ for $k \in [K]$ as fixed sets. Consider the change of variables

$$W \xrightarrow{\left( \Pi_{X \cdot Y} \circ \Psi_{A} \right)^{-1}} \left( V \ \Gamma \ \U \right), \quad (17)$$

formed by inverting the composition of $\Pi_{X \cdot Y}$, defined in (11), and $\Psi_{A}$. Applying (17), we observe that that the law of the variables $\hat{\beta}_{E}, \hat{\beta}_{\perp}, V, \Gamma$ and $U$ is equal to

$$\left| J_{\left( \Pi_{X \cdot Y} \circ \Psi_{A} \right)} \right| \left\{ \prod_{k=1}^{K} \rho \left( \left( \beta^{(k)}_{E_k} \beta^{(k)}_{E_k}, \sigma_{k}^{2} \cdot \left( X^{(k)}_{E_k} X^{(k)}_{E_k} \right)^{-1} \right) \rho \left( \beta^{(k)}_{\perp}, 0, \Sigma^{(k)}_{\perp} \right) \right) \right\} \times \rho \left( \Pi_{X \cdot Y} \circ \Psi_{A} \left( V, \Gamma, U \right); 0, \Omega \right),$$

where we let $J_{\left( \Pi_{X \cdot Y} \circ \Psi_{A} \right)}$ be the Jacobian for the change of variables mapping and $\Omega \in \mathbb{R}^{p \times p}$ is the block diagonal matrix with the diagonal entries equal to $\Omega^{(k)} \in \mathbb{R}^{p \times p}$ for $k \in [K]$. To complete the details of the proof, we introduce some matrices. Let $A = \begin{bmatrix} A_{1} & A_{2} \end{bmatrix}$. Note that

$$(\Pi_{X \cdot Y} \circ \Psi_{A}) \left( V, \Gamma, U \right) = C_{1} \hat{\beta}_{E} + C_{2} V + f \left( \Gamma, U; \beta_{\perp} \right)$$

where

$C_{1} = - \text{Diag} \left( X^{(1)}_{E_{1}} X^{(1)}_{E_{1}}, \ldots, X^{(K)}_{E_{K}} X^{(K)}_{E_{K}} \right)$

$C_{0} = \left( - C_{1} + \epsilon \text{Diag} \left( (I_{\delta_{1}} \ 0)^{\prime}, \ldots, (I_{\delta_{K}} \ 0)^{\prime} \right) \right) \text{Diag} \left( S \right)$
\( C_2 = C_0 (A_1 - A_2 D), \) and
\[
f \left( \Gamma, U; \tilde{\beta}_\bot \right) = \text{Diag} (\Lambda) (S' \ U')' + C_0 A_2 \Gamma - \tilde{\beta}_\bot.
\]

Additionally, we define
\[
\Delta^{-1} = C_2' \Omega^{-1} C_2; \quad P = -\Delta C_2' \Omega^{-1} C_1; \quad q = -\Delta C_2' \Omega^{-1} f \left( \Gamma, U; \tilde{\beta}_\bot \right);
\]
\[
\Sigma^{-1} = \text{Diag} \left( \frac{1}{\sigma_1^2} X_{E_1}^{(1)}, \ldots, \frac{1}{\sigma_K^2} X_{E_K}^{(K)} \right) + C_1' \Omega^{-1} C_1 - P' \Delta^{-1} P;
\]
\[
L = \Sigma \text{Diag} \left( \frac{1}{\sigma_1^2} X_{E_1}^{(1)}, \ldots, \frac{1}{\sigma_K^2} X_{E_K}^{(K)} \right); \quad \text{and}
\]
\[
m = \Sigma P' \Delta^{-1} q - \Sigma C_1' \Omega^{-1} f \left( \Gamma, U; \tilde{\beta}_\bot \right).
\]

Conditioning upon \( \tilde{\beta}_\bot \) and the event in (12), which we note is equivalent to
\[
\{ V > 0, \Gamma - D V > 0, \tilde{\Gamma} = \Gamma, \tilde{U} = U \}
\]
based on Lemma 3.1, the law for \( \tilde{\beta}_E \) and \( V \) is given up to a constant by
\[
\left( \int \rho \left( \tilde{\beta}; L_{\beta_E} + m, \Sigma \right) \cdot \rho(\tilde{V}; P \tilde{\beta} + q, \Delta) \cdot 1 \left( H \tilde{V} \geq g \right) d\tilde{V} d\tilde{\beta} \right)^{-1} \cdot \rho \left( \tilde{\beta}_E; L_{\beta_E} + m, \Sigma \right) \cdot \rho(V; P \tilde{\beta}_E + q, \Delta) \cdot 1 \left( HV \geq g \right).
\]

Integrating over \( V \) in the last display and noting that
\[
\int_{\{ \tilde{V}; H \tilde{V} \geq g \}} \rho(\tilde{V}; P \tilde{\beta}_E + q, \Delta) d\tilde{V}
\]
is a constant free from the parameters \( \beta_E \) in our model, we complete the proof of the Theorem. \( \Box \)

**Proof:** Theorem 3.3. Observe that the maximum likelihood estimate is obtained by maximizing the following objective with respect to \( \beta_E \):
\[
\begin{align*}
\text{maximize}_{\beta_E} & \quad \tilde{\beta}_E' \Sigma^{-1} (L_{\beta_E} + m) + \inf_{\tilde{\beta}, \tilde{V}} \left\{ \frac{1}{2} \tilde{\beta}' \Sigma^{-1} \tilde{\beta} - \tilde{\beta}_E' \Sigma^{-1} (L_{\beta_E} + m) \\
& \quad + \frac{1}{2} (\tilde{V} - P \tilde{\beta} - q)' \Delta^{-1} (\tilde{V} - P \tilde{\beta} - q) + \phi_{H,g}(\tilde{V}) \right\}.
\end{align*}
\]
Let
\[ F^*(\zeta) = \sup_{\beta} \left\{ \tilde{\beta}' \zeta - \frac{1}{2} \tilde{\beta}' \Sigma^{-1} \tilde{\beta} - \inf_{\tilde{V}} \frac{1}{2} (\tilde{V} - P \tilde{\beta} - q)' \Delta^{-1} (\tilde{V} - P \tilde{\beta} - q) + \phi_{H,g}(\tilde{V}) \right\} \]
represent the convex conjugate for
\[ F(\beta) = \frac{1}{2} \beta' \Sigma^{-1} \beta + \inf_{\tilde{V}} \frac{1}{2} (\tilde{V} - P \tilde{\beta} - q)' \Delta^{-1} (\tilde{V} - P \tilde{\beta} - q) + \phi_{H,g}(\tilde{V}). \]

Then, solving (18) is equivalent to the problem
\[
\maximize_{\beta \in \mathcal{E}} \beta' E \Sigma^{-1} (L \beta + m) - \sup_{\beta} \left\{ \tilde{\beta}' \Sigma^{-1} (L \beta + m) - \frac{1}{2} \tilde{\beta}' \Sigma^{-1} \tilde{\beta} \right. \\
- \inf_{\tilde{V}} \frac{1}{2} (\tilde{V} - P \tilde{\beta} - q)' \Delta^{-1} (\tilde{V} - P \tilde{\beta} - q) + \phi_{H,g}(\tilde{V}) \right\} \\
= \maximize_{\beta \in \mathcal{E}} \tilde{\beta}' \Sigma^{-1} (L \beta + m) - F^*(\Sigma^{-1} (L \beta + m)).
\]

Using the identity \( (\nabla F^*)^{-1} = \nabla F \) from convex analysis, we have
\[
\Sigma^{-1} (L \beta_{E}^{MLE} + m) = (\nabla F^*)^{-1} (\hat{\beta}_E) \\
= \Sigma^{-1} \hat{\beta}_E + \left( \frac{\partial}{\partial \beta} V(\beta) \right)_{\beta = \hat{\beta}_E} (\beta) - P)' \Delta^{-1} (\hat{V} - P \hat{\beta}_E - q) \\
+ \left( \frac{\partial}{\partial \hat{\beta}_E} V(\beta) \right)_{\beta = \hat{\beta}_E} (\beta)' \nabla \phi_{H,g}(\hat{V}),
\]
where
\[
V(\beta) = \arg\min_{\tilde{V}} \frac{1}{2} (\tilde{V} - P \tilde{\beta} - q)' \Delta^{-1} (\tilde{V} - P \tilde{\beta} - q) + \phi_{H,g}(\tilde{V}) \tag{19}
\]
and \( V(\hat{\beta}_E) = \hat{V} \), the optimizer defined in (16). By the definition of \( \hat{V} \),
\[ \Delta^{-1} (\hat{V} - P \hat{\beta}_E - q) + \nabla \phi_{H,g}(\hat{V}) = 0, \]
and we derive the system of estimating equations
\[ \hat{\beta}_E^{MLE} = L^{-1} \hat{\beta}_E + L^{-1} \Sigma P' \Delta^{-1} (P \hat{\beta}_E + q - \hat{V}) - L^{-1} m. \]

Using the curvature of the smooth approximate likelihood, we observe that the expression for the
The observed Fisher information matrix is equal to
\[ L' \Sigma^{-1} \nabla^2 F^*(\zeta) \big|_{\zeta = \Sigma^{-1}(L \hat{\beta}_E^{MLE} + m)} \Sigma^{-1} L = L' \Sigma^{-1} \frac{\partial}{\partial \zeta} \beta^*(\zeta) \big|_{\zeta = \Sigma^{-1}(L \hat{\beta}_E^{MLE} + m)} \Sigma^{-1} L. \]

Here, \( \nabla^2 \) denotes the Hessian matrix, and
\[ \beta^*(\zeta) = \arg\sup_{\beta} \left\{ \beta' \zeta - \frac{1}{2} \beta' \Sigma^{-1} \beta - \inf_{\tilde{V}} \frac{1}{2} (\tilde{V} - P \tilde{\beta} - q)' \Delta^{-1} (\tilde{V} - P \tilde{\beta} - q) + \phi_{H,g}(\tilde{V}) \right\}. \quad (20) \]

From the K.K.T. conditions of optimality for (20),
\[ \zeta = (\Sigma^{-1} + P' \Delta^{-1} P) \beta^*(\zeta) - P' \Delta^{-1} (V(\beta^*(\zeta)) - q), \]
which leads to noting
\[ \frac{\partial}{\partial \beta} V(\beta) \big|_{\beta = \beta^*(\zeta)} = (\Delta^{-1} + \nabla^2 \phi_{H,g}(V(\beta^*(\zeta))))^{-1} \Delta^{-1} P, \]
and
\[ \frac{\partial}{\partial \zeta} \beta^*(\zeta) \big|_{\zeta = \zeta_0} = \left( \Sigma^{-1} + P' \Delta^{-1} P - P' \Delta^{-1} \frac{\partial}{\partial \beta} V(\beta) \big|_{\beta = \beta^*(\zeta)} \right)^{-1} \]
\[ = \left( \Sigma^{-1} + P' \Delta^{-1} P - P' \Delta^{-1} \left( \Delta^{-1} + \nabla^2 \phi_{H,g}(V(\beta^*(\zeta_0))) \right)^{-1} \Delta^{-1} P \right)^{-1}. \]

Using, once again, the K.K.T. conditions of optimality for (20), we observe that
\[ \beta^* \left( \Sigma^{-1}(L \hat{\beta}_E^{MLE} + m) \right) = \hat{\beta}_E. \]

Plugging
\[ \frac{\partial}{\partial \zeta} \beta^*(\zeta) \big|_{\zeta = \Sigma^{-1}(L \hat{\beta}_E^{MLE} + m)} \]
into the expression for the observed Fisher information, we obtain the expression for the inverse information matrix:
\[ \hat{\Gamma}^{-1} = L^{-1} \Sigma L'^{-1} + L^{-1} \Sigma \left( P' \Delta^{-1} P - P' \Delta^{-1} \left( \Delta^{-1} + \nabla^2 \phi_{H,g}(\tilde{V}) \right)^{-1} \Delta^{-1} P \right) \Sigma L^{-1}. \]
B Impracticality of Existing Conditional Prescription in Multi-Task Setting

In this section, we show in detail why the existing conditional prescription for other $\ell_1$-regularized algorithms after conditioning on the event in (10) does not lead to tractable inference. Given a stationary point $(B, U)$ defined through the K.K.T. mapping in (11), we immediately note the following equivalence

$$\hat{E} = E, \hat{S} = S \equiv b^{(k)} > 0, \|u^{(k)}\|_\infty \leq 1 \text{ for all } k \in [K].$$

(21)

The alternative characterization of the selection event through the K.K.T. mapping is directly analogous to the characterization used previously in conducting selective inference for the single-task LASSO. From here, the stationary mapping suggests a change of variables that may be useful in deriving a joint density function for the response and randomization variables. Proposition 1 provides us an expression for the Jacobian associated with the change of variables, and Proposition 2 describes the resulting selection-adjusted likelihood using the marginal law of the response vectors for the $K$ tasks.

Let $Q = \text{Diag}(Q^{(1,1)}, \ldots, Q^{(K,K)})$, where the block matrices $Q^{(k,k)} \in R^{p-\delta_k \times p-\delta_k}$ have columns given by

$$Q^{(k,k)}_{\cdot j_k} = \Lambda^{(k)}_{j_k} \cdot e_{\sim j_k} = \begin{cases} -\frac{1}{2} \lambda_{s, j_k}^{-1/2} S^{(k')}_{\sim j_k'} \cdot e_{\sim j_k'} & \text{if } j_k \in E_k \text{ and } \Lambda^{(k)}_{j_k} > \lambda_0 \\ 0 & \text{otherwise} \end{cases}$$

for $j_k \in E_k$. Here, $e_{\sim j_k} \in R^{-\delta_k}$ is the standard basis vector whose $j_k$th component is one, and all other components are zero. Let

$$R = \begin{pmatrix} R^{(1,1)} & \cdots & R^{(1,K)} \\ \vdots & \ddots & \vdots \\ R^{(K,1)} & \cdots & R^{(K,K)} \end{pmatrix},$$

where the block matrices $R^{(k,k')} \in R^{\delta_k \times \delta_{k'}}$ are defined column-wise as

$$R^{(k,k')}_{\cdot \sim j_k'} = \begin{cases} -\frac{1}{2} \lambda_{s, j_k}^{-1/2} S^{(k')}_{\sim j_k'} \cdot e_{\sim j_k'} & \text{if } j_k \in E_k \text{ and } \Lambda^{(k)}_{j_k} > \lambda_0 \\ 0 & \text{otherwise} \end{cases}$$

for $j_k \in E_{k'}$. Here, $e_{\sim j_k} \in R^{\delta_k}$ is the standard basis vector whose $j_k$th component is one, and all other components are zero. Finally, let

$$T = \text{Diag}\left((X_{E_1}^{(1)})'X_{E_1}^{(1)} + \epsilon \cdot I_{\delta_1}, (X_{E_2}^{(2)})'X_{E_2}^{(2)} + \epsilon \cdot I_{\delta_2}, \ldots, (X_{E_K}^{(K)})'X_{E_K}^{(K)} + \epsilon \cdot I_{\delta_K}\right) \cdot \text{Diag}(S).$$
Proposition 1 Consider a change of variables \( W^{\Pi_{X^\gamma}^1} (B, U) \) where

\[
W = \Pi_{X^\gamma}^1 (B, U) := (\pi^{(1)}(b^{(1)}, u^{(1)}), \pi^{(2)}(b^{(2)}, u^{(2)}), \ldots, \pi^{(K)}(b^{(K)}, u^{(K)}))
\]

and such that \( \pi^{(k)}(\cdot) \) is defined according to (11). The Jacobian determinant associated with the above change of variables is given by

\[
|J_{\Pi_{X^\gamma}^1} (B, U)| = \det (Q) \det (R + T).
\]

Proof: We note that the Jacobian for the change of variables mapping \( J_{\Pi_{X^\gamma}^1} \) is the determinant in absolute value for the following matrix

\[
\begin{bmatrix}
\frac{\partial \pi^{(1)}}{\partial b^{(1)}}, u^{(1)} & \frac{\partial \pi^{(1)}}{\partial b^{(2)}}, u^{(2)} & \cdots & \frac{\partial \pi^{(1)}}{\partial b^{(K)}}, u^{(K)} \\
\frac{\partial \pi^{(2)}}{\partial b^{(1)}}, u^{(1)} & \frac{\partial \pi^{(2)}}{\partial b^{(2)}}, u^{(2)} & \cdots & \frac{\partial \pi^{(2)}}{\partial b^{(K)}}, u^{(K)} \\
\vdots & \ddots & \ddots & \vdots \\
\frac{\partial \pi^{(K)}}{\partial b^{(1)}}, u^{(1)} & \frac{\partial \pi^{(K)}}{\partial b^{(2)}}, u^{(2)} & \cdots & \frac{\partial \pi^{(K)}}{\partial b^{(K)}}, u^{(K)}
\end{bmatrix}
\]

Noting in the above display \( \frac{\partial \omega_{E_k}}{\partial u^{(k)}} = 0 \) and \( \frac{\partial \omega_{E_k}}{\partial u^{(k)}} = 0 \) and \( \frac{\partial \omega_{-E_k}}{\partial u^{(k)}} = 0 \) for all \( k' \neq k \), the above Jacobian equals

\[
\begin{bmatrix}
\frac{\partial \omega_{E_1}}{\partial u^{(1)}} & 0 & \cdots & 0 \\
0 & \frac{\partial \omega_{E_1}}{\partial u^{(2)}} & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \frac{\partial \omega_{E_1}}{\partial u^{(K)}} \\
\frac{\partial \omega_{E_K}}{\partial u^{(1)}} & 0 & \cdots & 0 \\
0 & \frac{\partial \omega_{E_K}}{\partial u^{(2)}} & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \frac{\partial \omega_{E_K}}{\partial u^{(K)}}
\end{bmatrix}
\]

(22)
which further equals in absolute value the determinant of

\[
\begin{bmatrix}
M_1 & 0 \\
M_2 & Q
\end{bmatrix}
\]

where

\[
M_1 = \begin{bmatrix}
\frac{\partial \omega_1^{(1)}}{\partial b^{(1)}} & \frac{\partial \omega_1^{(1)}}{\partial b^{(2)}} & \cdots & \frac{\partial \omega_1^{(1)}}{\partial b^{(K)}} \\
\frac{\partial \omega_2^{(1)}}{\partial b^{(1)}} & \frac{\partial \omega_2^{(1)}}{\partial b^{(2)}} & \cdots & \frac{\partial \omega_2^{(1)}}{\partial b^{(K)}} \\
\vdots & \ddots & \ddots & \vdots \\
\frac{\partial \omega_K^{(1)}}{\partial b^{(1)}} & \frac{\partial \omega_K^{(1)}}{\partial b^{(2)}} & \cdots & \frac{\partial \omega_K^{(1)}}{\partial b^{(K)}}
\end{bmatrix},
M_2 = \begin{bmatrix}
\frac{\partial \omega_1^{(2)}}{\partial b^{(1)}} & \frac{\partial \omega_1^{(2)}}{\partial b^{(2)}} & \cdots & \frac{\partial \omega_1^{(2)}}{\partial b^{(K)}} \\
\frac{\partial \omega_2^{(2)}}{\partial b^{(1)}} & \frac{\partial \omega_2^{(2)}}{\partial b^{(2)}} & \cdots & \frac{\partial \omega_2^{(2)}}{\partial b^{(K)}} \\
\vdots & \ddots & \ddots & \vdots \\
\frac{\partial \omega_K^{(2)}}{\partial b^{(1)}} & \frac{\partial \omega_K^{(2)}}{\partial b^{(2)}} & \cdots & \frac{\partial \omega_K^{(2)}}{\partial b^{(K)}}
\end{bmatrix},
\]

and

\[
Q = \text{Diag} \left( \frac{\partial \omega_{E_1}^{(1)}}{\partial u^{(1)}}, \frac{\partial \omega_{E_2}^{(2)}}{\partial u^{(2)}}, \ldots, \frac{\partial \omega_{E_K}^{(K)}}{\partial u^{(K)}} \right).
\]

Observe \( M_1 = R + T \), because

\[
\frac{\partial \omega_{E_k}^{(k)}}{\partial b^{(k)}} = \lambda_{E_k}^{(k)} X_{E_k}^{(k)} + \epsilon \cdot I_{\delta_k} + R^{(k,k)}, \quad \frac{\partial \omega_{E_k}^{(k)}}{\partial b^{(k')}} = R^{(k,k')} \quad \text{for } k \neq k'.
\]

Using the block diagonal structure of (22), it follows that \( | J_{\Pi X \cdot Y} (B, U) | = \text{det} (Q) \text{det} (R + T). \)

**Proposition 2** Suppose we observe the event (10) after solving (7). Fixing the set

\[
O = \left\{ (\tilde{Y}, \tilde{B}, \tilde{U}) : \tilde{b}^{(k)} > 0, ||\tilde{u}^{(k)}||_\infty \leq 1 \text{ for all } k \in [K] \right\},
\]

we let

\[
\mathcal{N}_O (\beta_E) := \int_0 \left| J_{\Pi X \cdot \tilde{Y}} \left( \tilde{B}, \tilde{U} \right) \right| \prod_{k=1}^{K} \rho \left( \gamma^{(k)}; X_{E_k}^{(k)} \beta_{E_k}, \sigma_k^2 \cdot I_{n_k} \right)
\times \exp \left( -\frac{1}{2} \sum_{k=1}^{K} (\pi^{(k)}(\tilde{b}^{(k)}, \tilde{u}^{(k)}))' \Omega^{(k)}(\tilde{b}^{(k)}, \tilde{u}^{(k)}) \right) dB d\tilde{U} d\tilde{Y}
\]

Under the model in (9), the likelihood obtained from the law of \( Y \) conditional upon (10) is given up to a constant by

\[
(\mathcal{N}_O (\beta_E))^{-1} \prod_{k=1}^{K} \rho \left( y^{(k)}; X_{E_k}^{(k)} \beta_{E_k}, \sigma_k^2 \cdot I_{n_k} \right).
\]

**Proof:** Let \( \mathcal{L} \) be the set of realizations of the responses and randomization variables that result in observing (10), that is

\[
\mathcal{L} = \left\{ (Y, W) : \hat{E}(Y, W) = E, \hat{S}(Y, W) = S \right\}.
\]

39
We note that the joint law of $Y, W$ after conditioning out this observed event from (10) is obtained by truncating their unconditional Gaussian law and is given by:

$$\left\{ \prod_{k=1}^{K} \frac{\rho \left( y^{(k)}; X_{E_k}^{(k)} \beta_{E_k}^{(k)}, \sigma_k^2 \cdot I_{n_k} \right) \cdot \rho(\omega^{(k)}; \mathbf{0}, \Omega^{(k)})}{\int \left\{ \prod_{k=1}^{K} \rho \left( \tilde{y}^{(k)}; X_{E_k}^{(k)} \beta_{E_k}^{(k)}, \sigma_k^2 \cdot I_{n_k} \right) \cdot \rho(\tilde{\omega}^{(k)}; \mathbf{0}, \Omega^{(k)}) \right\}} \cdot 1_{\mathcal{L}(Y, W)} \right\}.$$ 

Using the change of variables mapping

$$W = \Pi_{X\cdot Y} (B \quad U) = \left( \pi^{(1)}(b^{(1)}, u^{(1)}) \quad \pi^{(2)}(b^{(2)}, u^{(2)}) \quad \ldots \quad \pi^{(K)}(b^{(K)}, u^{(K)}) \right)$$

defined in (11), the conditional law for $Y, B$ and $U$ is equal to

$$\mathcal{N}(\beta_E)^{-1} \left| J_{\Pi_{X\cdot Y}} (B, U) \right| \prod_{k=1}^{K} \frac{\rho \left( y^{(k)}; X_{E_k}^{(k)} \beta_{E_k}^{(k)}, \sigma_k^2 \cdot I_{n_k} \right)}{\int \left\{ \prod_{k=1}^{K} \rho \left( \tilde{y}^{(k)}; X_{E_k}^{(k)} \beta_{E_k}^{(k)}, \sigma_k^2 \cdot I_{n_k} \right) \cdot \rho(\tilde{\omega}^{(k)}; \mathbf{0}, \Omega^{(k)}) \right\}} \cdot 1_{\mathcal{L}(\tilde{Y}, \tilde{W})} dW dY,$$

where we have used the equivalence of

$$\widehat{E}(Y, W) = E, \quad \hat{S}(Y, W) = S$$

to the following constraints

$$b^{(k)} > \mathbf{0}, \|u^{(k)}\|_\infty < 1 \text{ for all } k \in \{1, 2, \ldots, K\}.$$

Ignoring constants in $\beta_E$ in the above expression, we deduce the claim in the Proposition.

The adjusted likelihood for this conditional prescription, seemingly a first-line recourse for post-selection inference, does not lend itself towards tractable estimating equations. Note that there is no clear method to compute the MLE given the following system of estimating equations:

$$\frac{\partial}{\partial \beta_{E_k}^{(k)}} \log \mathcal{N}(\beta_E) \bigg|_{\hat{\beta}_{E_k}^{MLE}} = \frac{1}{\sigma_k^2} X_{E_k}^{(k)'} \left( y^{(k)} - X_{E_k}^{(k)} \beta_{E_k}^{(k)} \right) \bigg|_{\hat{\beta}_{E_k}^{MLE}} \text{ for } k \in [K]$$

The absence of closed-form expressions for the partial derivatives of the Jacobian $J_{\Pi_{X\cdot Y}} (\cdot)$ and the normalizing constant $\mathcal{N}(\beta_E)$ poses a substantial obstacle in solving these equations. An attempt instead to numerically integrate out $Kp + N$ variables in the derivatives of the normalizing constant will turn out to be quite futile from an angle of computing efficiency.
C Simulated Experiments

C.1 Deviation from Gaussian distribution

We continue with the simulations in Section 4, and investigate the performance of our methods as the response variables deviate from a Gaussian distribution. Figure 7 shows the distribution of coverage rates for our method when the models errors are drawn from Exponential and Laplace distributions.

![Figure 7: Coverage rates for “MTL + SI” as a function of task sparsity $s_T$.](image)

We note that “MTL + SI” continues to attain nominal coverage levels on synthetic data matched to the dimensions and estimated sparsity level of the ABCD data even when the response variables deviate from the Gaussian distribution. One expects this behavior because the added Gaussian randomization ensures asymptotically-valid selective inference just as does data-splitting, thanks to a Central Limit Theorem proved in Panigrahi (2018).

C.2 Varying Sparsity and Dimensions

In this section, we design three sets of examples to study the effects of (1) varying the heterogeneity in predictors across tasks by varying task sparsity $s_T$ at a fixed value of global sparsity $s_G$; (2) varying...
global sparsity $s_G$ at a fixed value of task sparsity $s_T$; and (3) varying the number of predictors $p$, fixing $s_T$ and the total number of active predictors. In all of the examples, we follow the generative scheme in Section 4, and set $n = 500$, $K = 5$, $\rho = 0.3$, and $\alpha = 0.1$. For scenario (1), we additionally fix $p = 100$, $s_G = 0.9$, and vary task sparsity $s_T$ from 0 to 0.6. The signal-to-noise ratio per task, defined as

$$\text{SNR} = \frac{\beta^{(k)}' T(\rho) \beta^{(k)}}{n_k \sigma^2},$$

averages around 0.1 across tasks in the first set of experiments. For scenario (2), we fix $p = 100$, $s_T = 0.2$, and vary $s_G$ between 0.8 and 0.95. The average SNR per task is around 0.5 for global sparsity of 80 percent and decreases with increasing global sparsity. In both of the first two scenarios, we perform 100 replications. The optimal tuning parameter for each method is the value on the $\lambda$-path that yields the lowest average MSE on the validation set across iterations, and the results are reported at that tuning parameter value for all 100 iterates. For setting (3), we fix $t_S = 0.2$, set $s_G$ so that the expected number of active predictors per task is 10, and vary $p$ between 100 and 1000. The average SNR-per task increases to around 2.5 for the highest-dimensional setting to compensate for the added complexity. Due to the computational cost required to tune the model for larger $p$, we use 10 auxiliary iterations to choose tuning parameters and then use these parameters for the 100 iterations we report. The same metrics, CR, averaged lengths of confidence intervals, and F1, are used to evaluate performance.

We start from examining the results over the entire path of the tuning parameter $\lambda$. Figure 8 shows the coverage, interval length, and F1 score for a range of values of $\lambda$ for the setting $s_G = 0.9$, $s_T = 0.2$, and $p = 100$. The distribution of coverage for the “naive” method underscores the dangers of ignoring model selection bias, with much higher variability in coverage than the post-selection methods. Reassuringly, all the post-selection methods achieve the nominal coverage rate of 0.9, regardless of the choice of tuning parameter and subsequent quality of model selected. The interval lengths along the path of $\lambda$ show that the methods based on randomization consistently produce tighter confidence intervals than the equivalent data splitting approach. However, it is not possible to directly compare the accuracy of inference from each method at the same, fixed value of the tuning parameter since the quality of the model selected by each method varies considerably along with $\lambda$. The F1 score, for example, achieves its best value at different points along the path for different methods. The colored dots in the top panel of Figure 9 show the value of $\lambda$ that yielded the lowest mean squared error (MSE) on the holdout data set. From this point on, we will compare methods using the optimal value of the tuning parameter, chosen by minimizing the MSE on the validation data, for each method.

Figure 9 shows the distribution of coverage, interval length, and the F1 score as we vary the task sparsity $s_T$ from 0 to 0.6, holding $s_G$ and $p$ fixed. All methods consistently attain nominal or near-
nominal coverage, regardless of task sparsity. Selective inference (MTL + SI) yields consistently shorter intervals than data splitting (DS), illustrating the benefit of using the whole sample rather than splitting. At all levels of task sparsity, but especially when there is more homogeneity in active predictors across tasks, MTL methods with selective inference also achieve better accuracy than single-task LASSO with selective inference as measured by the F1 score.

Figure 10 shows the same patterns continue to hold when we vary global sparsity while keeping task sparsity fixed. We only vary global sparsity between 0.8 and 0.95 because the problem becomes very easy for all methods at lower levels of global sparsity. MTL+ SI again outperforms other methods on both interval length and accuracy at all levels of global sparsity. We note, however, that the variance increases for all methods at larger levels of $s_G$.

Finally, Figure 11 shows the results as a function of the number of predictors $p$. To control the difficulty of the problem as $p$ changes, we fix the average number of active predictors per task at 10, rather than specifying the fraction of active predictors. The average interval length for the selective inference methods grows slightly with $p$, reflecting the cost of conditioning upon additional information. Most overall patterns remain the same, with MTL + SI providing the best performance. The accuracy of LASSO + SI declines faster with $p$ than that of other methods, indicating the growing cost of disregarding shared information. In summary, our simulation results show that there is variability in performance of all methods, but multi-task learning with selective inference achieves the nominal coverage in all settings and provides the best performance, on average, compared to either data splitting or single-task prediction.
Figure 8: Coverage, length, and accuracy along the path of $\lambda$ for the setting $s_G = 0.9$, $s_T = 0.2$, and $p = 100$. The solid and dashed lines are used to group methods that use a comparable amount of information for model selection; randomization parameters $v = 0.70$ and $v = 1$ are comparable to data splitting with $s = 0.67$ and $s = 0.50$, respectively.
Figure 9: Results for 100 replications as a function of task sparsity $s_T$, fixing global sparsity $s_G = 0.9$ and $p = 100$. Top: coverage; middle: interval length; bottom: F1 score.

Figure 10: Results for 100 replications as a function of global sparsity $s_G$, fixing task sparsity $s_T = 0.2$ and $p = 100$. Left: coverage; middle: interval length; right: F1 score.
Figure 11: Results for 100 replications as a function of the number of predictors $p$, fixing task sparsity $s_T = 0$ and an average of 10 active predictors per task. Left: coverage; middle: interval length; right: F1 score.