The Annals of Applied Statistics
2024, Vol. 18, No. 1, 445–467
https://doi.org/10.1214/23-AOAS1796

SELECTIVE INFERENCE FOR SPARSE MULTITASK REGRESSION WITH APPLICATIONS IN NEUROIMAGING

BY SNIGDHA PANIGRAHI[1,a], NATASHA STEWART[1,b], CHANDRA SRIPADA[2,d] AND ELIZAVETA LEVINA[1,c]

[1]*Department of Statistics, University of Michigan,* [a]*psnigdha@umich.edu,* [b]*nstew@umich.edu,* [c]*elevina@umich.edu*
[2]*Department of Philosophy, University of Michigan,* [d]*sripada@umich.edu*

Multitask 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 multitask 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 multitask 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 predictors 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 multitask learning with selective inference can more accurately recover true signals than single-task methods.

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 (Blair (2006), Gray, Chabris and Braver (2003)). 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 (Carroll et al. (1993), Humphreys (1979), Spearman (1961)). 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 noninvasively investigate the neurological organization of cognitive abilities, avoiding the need to rely solely on behavioral data.

Received June 2022; revised April 2023.
Key words and phrases. Multitask learning, multilevel lasso, joint sparsity, postselection inference, selective inference, neuroimaging, fMRI 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 multitask 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 multitask 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 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 (Anderson and Barbey (2023), Finn et al. (2015), Hearne, Mattingley and Cocchi (2016), Sripada et al. (2021), 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 (2023), 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, Sun and Taylor (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 (Liu, Markovic and Tibshirani (2018), Suzumura et al. (2017), Tanizaki et al. (2020), Taylor and Tibshirani (2018), Zhao and Panigrahi (2019), 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, Sun and Taylor (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. (2023), Panigrahi and Taylor (2018), Panigrahi, Taylor and Weinstein (2021) 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 multitask 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 multitask 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. Multitask 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 μ unknown and Σ known, and an independent randomization variable $\omega \sim N(0, \Omega) \in \mathbb{R}^p$ with known Ω. 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

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L}(\Theta; y, X, \omega) + \|\Lambda\Theta\|_1, \tag{1}$$

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

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

We use M' to denote the transpose of a matrix M. The ridge term in the loss function, with a small ϵ, is included to ensure the existence of an optimal solution; see Tian et al. (2016). The estimated support set, $\widehat{E} = \operatorname{Supp}(\widehat{\Theta})$, is treated as random, reflecting the fact that different samples of y and ω will yield different active sets. Suppose we observe $\widehat{E}(y, \omega) = E$ 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 $y \sim N(X_E\beta_E, \sigma^2\mathrm{I_n})$, where σ is fixed and $\mathrm{I_n}$ 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,

$$\beta_E = \underset{\beta\in\mathbb{R}^{|E|}}{\operatorname{argmin}}\, \mathbb{E}\big[\|y - X_E\beta\|_2^2\big], \tag{3}$$

under the selected model. We note that some recent postselection work by Hong, Kuffner and Martin (2018), Panigrahi, Wang and He (2022), Wang, He and Xu (2020) 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 β_E is obtained by conditioning upon the partition

$$\mathcal{P}_E = \big\{y \in \mathbb{R}^n, \omega \in \mathbb{R}^p : \widehat{E}(y, \omega) = E\big\}$$

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 μ and covariance Ω, we can write this likelihood as

$$y|\widehat{E} = E \propto \frac{\rho(y; X_E\beta_E, \sigma^2\mathrm{I_n})}{\int_{\mathcal{P}_E} \rho(\widetilde{y}; X_E\beta_E, \sigma^2\mathrm{I_n}) \cdot \rho(\widetilde{w}; 0, \Omega)\, d\widetilde{w}\, d\widetilde{y}}. \tag{4}$$

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,

$$\big\{\widehat{S}(y, \omega) = S\big\} \subseteq \big\{\widehat{E}(y, \omega) = E\big\}.$$

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, $\widehat{\beta}_E^{\mathrm{MLE}}$, and using the observed Fisher information matrix, $\widehat{I}(\widehat{\beta}_E^{\mathrm{MLE}})$, to estimate the variance. This yields postselection confidence intervals for β_E of the form

$$\widehat{\beta}_{E,j}^{\mathrm{MLE}} \pm z_{1-\alpha/2}\sqrt{\widehat{I}_{jj}^{-1}(\widehat{\beta}_E^{\mathrm{MLE}})}, \quad j \in E,$$

where v_j is the jth entry of the vector v, M_{ij} is the (i, j)th element of a matrix M, and z_q is the qth 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 multitask 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 kth 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) $\omega^{(k)} \sim \mathrm{N}(0, \Omega^{(k)})$ 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 nonzero coefficients is sparse and use penalized multitask regression to identify the relevant features. We impose a specific intertask structure on the joint regression by assuming that the coefficients $\Theta^{(1)}, \ldots, \Theta^{(K)} \in \mathbb{R}^p$ can be represented as the product of a common parameter that is shared between all tasks and a task-specific parameter that is unique to an individual task; namely,

$$\Theta_j^{(k)} = \tau_j \, \gamma_j^{(k)}, \quad \text{for } j \in [p], k \in [K]. \tag{5}$$

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, τ_j and $\gamma_j^{(k)}$, are each identifiable since our interest lies in estimating the sparsity structure of their product. Note that τ_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

$$\sum_{k=1}^{K} \mathcal{L}\big(\Theta^{(k)}; y^{(k)}, X^{(k)}, \omega^{(k)}\big) + \eta_1 \sum_{j=1}^{p} \tau_j + \eta_2 \sum_{j=1}^{p} \sum_{k=1}^{K} |\gamma_j^{(k)}|, \tag{6}$$

subject to the constraint $\tau_j \geq 0$ for $j \in [p]$. The minimizers $\widehat{\tau} \in \mathbb{R}^p$, $\widehat{\gamma}^{(k)} \in \mathbb{R}^p$ will clearly both be sparse due to the ℓ_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

$$\underset{\{\Theta_j^{(k)}\}_{j\in[p],k\in[K]}}{\operatorname{argmin}} \sum_{k=1}^{K} \mathcal{L}\big(\Theta^{(k)}; y^{(k)}, X^{(k)}, \omega^{(k)}\big) + 2\lambda \sum_{j=1}^{p} \left\{ \sum_{k=1}^{K} |\Theta_j^{(k)}| \right\}^{1/2}, \tag{7}$$

where $\lambda = \sqrt{\eta_1 \eta_2}$. This equivalence was established in previous work; see, for example, 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,

$$2\left\{ \sum_{k=1}^{K} |\Theta_j^{(k)}| \right\}^{1/2} \sim c^{(t)} + \sum_{k=1}^{K} \frac{|\Theta_j^{(k)}|}{\sqrt{\sum_{k=1}^{K} |(\widehat{\Theta}_j^{(k)})^{(t)}|}}.$$

The constant $c^{(t)}$ depends only on the previous iterate and can be ignored in the corresponding optimization problem. With this approximation the multitask 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

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

Here λ_0 is a prespecified large positive constant used for numerical stability, and λ is a tuning parameter. The iterative procedure is summarized in Algorithm 1.

Upon convergence of Algorithm 1, we obtain

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

Let the cardinality of $E_k \subseteq [p]$ be equal to δ_k, and let $\delta = \delta_1 + \cdots + \delta_K$.

Algorithm 1 Estimating Shared Sparsity

for k = 1,...,K **do**
  Initialize $(\widehat{\Theta}^{(k)})^{(0)} = \operatorname{argmin}_{\Theta^{(k)}}(\mathcal{L}(\Theta^{(k)}; y^{(k)}, X^{(k)}, \omega^{(k)}) + \lambda \sum_{j=1}^{p} |\Theta_j^{(k)}|)$
end for
procedure ITERATE UNTIL CONVERGENCE
  Let $t = 0$
  while convergence $<$ tol **do**
    for $j = 1, \ldots, p$ **do**
      $\lambda_j^{(t+1)} = \min\{\lambda_0,\ \lambda \cdot (\sum_{k=1}^{K} |(\widehat{\Theta}_j^{(k)})^{(t)}|)^{-\frac{1}{2}}\}$
    end for
    for $k = 1, \ldots, K$ **do**
      LASSO:
      Solve $(\widehat{\Theta}^{(k)})^{(t+1)} = \operatorname{argmin}_{\Theta^{(k)}}(\mathcal{L}(\Theta^{(k)}; y^{(k)}, X^{(k)}, \omega^{(k)}) + \sum_{j=1}^{p} \lambda_j^{(t+1)} |\Theta_j^{(k)}|)$
    end for
    $t = t + 1$
  end while
end procedure

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

$$y^{(k)} = X^{(k)}_{E_k} \beta^{(k)}_{E_k} + \varepsilon_k, \quad \text{for } k \in [K], \tag{9}$$

where ε_k is a vector of errors with length n_k, mean 0, and variance σ_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 postselection confidence intervals for $\beta^{(k)}_{E_k}$ 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

$$\mathbf{Y} = \big(y^{(1)\prime} \dots y^{(K)\prime}\big)', \qquad \mathbf{E} = \{E_1, \dots, E_k\}, \qquad \boldsymbol{\beta}_{\mathbf{E}} = \big(\beta^{(1)\prime}_{E_1} \dots \beta^{(K)\prime}_{E_k}\big)'.$$

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

$$X^{(k)} = \begin{bmatrix} X^{(k)}_{E_k} & X^{(k)}_{-E_k} \end{bmatrix}, \qquad \omega^{(k)} = \begin{pmatrix} \omega^{(k)}_{E_k} \\ \omega^{(k)}_{-E_k} \end{pmatrix},$$

where $-A$ denotes the complement of set A. For each predictor $j \in [p]$, we use $\widetilde{j}_k$ 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 nonzero multitask regression coefficients for task $k \in [K]$ under this permutation, that is,

$$b^{(k)}_{\widetilde{j}_k} = |\widehat{\Theta}^{(k)}_j| \quad \text{whenever } |\widehat{\Theta}^{(k)}_j| \neq 0,$$

and let

$$\mathbf{B} = \big(b^{(1)\prime} \dots b^{(K)\prime}\big)'.$$

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 ℓ_1-norm for the multitask regression coefficients when evaluated at the solution, that is,

$$\big(s^{(k)\prime} u^{(k)\prime}\big)' = \mathcal{D}_{\widehat{\Theta}} \big\| \big(\Theta^{(k)\prime}_{E_k} \Theta^{(k)\prime}_{-E_k}\big)' \big\|_1 \quad \text{for } k \in [K].$$

Note that the vector $s^{(k)}$ gives the signs of the active multitask coefficients for the kth task, and $u^{(k)}$ satisfies $\|u^{(k)}\|_\infty \leq 1$. To reference the active and inactive components, respectively, of all the evaluated ℓ_1-norm subgradients, we define

$$\mathbf{S} = \big(s^{(1)\prime} \dots s^{(K)\prime}\big)', \qquad \mathbf{U} = \big(u^{(1)\prime} \dots u^{(K)\prime}\big)'.$$

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

$$\boldsymbol{\Gamma} = \big(\Gamma^{(j_1)} \dots \Gamma^{(j_r)}\big)',$$

where

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

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

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

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,

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

The vector $v^{(j)}$ collects the absolute value of the nonzero coefficients for predictor j across tasks, excluding the coefficient for the last task where predictor j is nonzero. 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

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

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

$$D_i = \left(\mathbf{0}'_{(d_{j_1}-1)} \quad \cdots \quad \mathbf{1}'_{(d_{j_i}-1)} \quad \cdots \quad \mathbf{0}'_{(d_{j_r}-1)}\right) \quad \text{for } i \in [r].$$

For some permutation matrix $\mathcal{A} \in \mathbb{R}^{\delta \times \delta}$, the relationship between $\mathbf{B}$ and $(\mathbf{V}, \mathbf{\Gamma})$ is given by

$$\mathbf{B} = \mathcal{A}\begin{pmatrix} \mathbf{V} \\ \mathbf{\Gamma} - D\mathbf{V} \end{pmatrix}.$$

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

$$H = \begin{pmatrix} \mathrm{I}_{(\delta - \mathrm{r})} \\ -D \end{pmatrix}, \qquad g = \begin{pmatrix} \mathbf{0}_{(\delta - r)} \\ -\mathbf{\Gamma} \end{pmatrix}.$$

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

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

where H_j is the jth row of H and g_j is the jth component of g.

3.2. *Estimating equations for approximate MLE-based inference.* A natural starting point for selective inference in the multitask setting is the law for $\mathbf{Y}$, conditioning on the event

$$\{\widehat{\mathbf{E}} = \mathbf{E}, \widehat{\mathbf{S}} = \mathbf{S}\}. \tag{10}$$

This conditional prescription results in practical selective inference procedures for other ℓ_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 ω 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 $\mathbf{E}$ and $\mathbf{S}$, induces a relationship between $\mathbf{W} = (\omega^{(1)'} \dots \omega^{(K)'})'$, the collection of randomization variables, and $(\mathbf{B}, \mathbf{U})$. This relationship implies a transformation of the form

$$\mathbf{W} = \left(\pi^{(1)}\big(b^{(1)}, u^{(1)}\big) \quad \pi^{(2)}\big(b^{(2)}, u^{(2)}\big) \quad \cdots \quad \pi^{(K)}\big(b^{(K)}, u^{(K)}\big)\right),$$

where

$$\begin{aligned}\pi^{(k)}\big(b^{(k)}, u^{(k)}\big) = -X^{(k)'} y^{(k)} + \begin{bmatrix} \big(X^{(k)}_{E_k}\big)' X^{(k)}_{E_k} + \epsilon \cdot \mathrm{I}_{\delta_k} \\ \big(X^{(k)}_{-E_k}\big)' X^{(k)}_{E_k} \end{bmatrix} \mathrm{Diag}\big(s^{(k)}\big) b^{(k)} \\ + \mathrm{Diag}\big(\Lambda^{(k)}\big) \begin{pmatrix} s^{(k)} \\ u^{(k)} \end{pmatrix}, \end{aligned} \tag{11}$$

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

Observe that this transformation is nonaffine since the penalty term is now related to the solution. The change-of-variables Jacobian, given in Proposition 1 of the Supplementary Material (Panigrahi et al. (2024)), is a complicated function of $(\mathbf{B}, \mathbf{U})$. Deriving estimating equations for maximum-likelihood inference using the law of $(\mathbf{Y}, \mathbf{B}, \mathbf{U})$, 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 Supplementary Material.

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

$$\{\widehat{\mathbf{E}} = \mathbf{E}, \widehat{\mathbf{S}} = \mathbf{S}, \widehat{\mathbf{\Gamma}} = \mathbf{\Gamma}, \widehat{\mathbf{U}} = \mathbf{U}\}, \tag{12}$$

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,

$$\widehat{\beta}^{(k)}_{E_k} = \big(X^{(k)}_{E_k}\big)^{\dagger} y^{(k)}. \tag{13}$$

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 $X^{(k)'} y^{(k)}$, the event of selection also relies on

$$\widehat{\beta}^{(k)}_{\perp} = \big(X^{(k)}\big)' \big(\mathrm{I}_{\mathrm{n_k}} - X^{(k)}_{E_k} \big(X^{(k)}_{E_k}\big)^{\dagger}\big) y^{(k)},$$

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 $\mathbf{V}$, $\mathbf{\Gamma}$, and $\mathbf{U}$ that we observe after solving the MTL Algorithm 1; please see Section 3.1 for a complete list of definitions.

LEMMA 3.1. *Suppose* $\mathbf{V}, \mathbf{\Gamma}$ *are defined as above. Then the event* (12) *is equivalent to the event*

$$\{\mathbf{V} > \mathbf{0},\ \widehat{\mathbf{\Gamma}} = \mathbf{\Gamma},\ \mathbf{\Gamma} - D\mathbf{V} > \mathbf{0},\ \widehat{\mathbf{U}} = \mathbf{U}\}.$$

The proof can be found in the Supplementary Material. Consider the bijective mapping $\Psi_{\mathcal{A}} : \mathbb{R}^{K \cdot p} \to \mathbb{R}^{K \cdot p}$ such that

$$\begin{pmatrix}\mathbf{B}' & \mathbf{U}'\end{pmatrix}' = \Psi_{\mathcal{A}}\begin{pmatrix}\mathbf{V} & \mathbf{\Gamma} & \mathbf{U}\end{pmatrix} = \mathrm{Diag}(\mathcal{A}, \mathrm{I}_{(\mathrm{Kp}\,-\,\delta)})\begin{pmatrix}\mathbf{V}' & (\mathbf{\Gamma} - D\mathbf{V})' & \mathbf{U}'\end{pmatrix}'.$$

Applying a change of variables via the composite mapping,

$$\mathbf{W} \xrightarrow{(\Pi_{\mathbf{X}'\mathbf{Y}} \circ \Psi_A)^{-1}} \begin{pmatrix}\mathbf{V} & \mathbf{\Gamma} & \mathbf{U}\end{pmatrix}, \tag{14}$$

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 $\mathbf{V}$, $\mathbf{\Gamma}$, and $\mathbf{U}$ 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 δ linear inequalities.

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

$$\left(\int \rho(\widetilde{\beta}; L\boldsymbol{\beta}_{\mathbf{E}} + m, \Sigma) \cdot \rho(\widetilde{V}; P\widetilde{\beta} + q, \Delta) \cdot 1(H\widetilde{V} \geq g)\, d\widetilde{V}\, d\widetilde{\beta}\right)^{-1} \cdot \rho(\widehat{\boldsymbol{\beta}}_{\mathbf{E}}; L\boldsymbol{\beta}_{\mathbf{E}} + m, \Sigma).$$

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

$$\begin{aligned}&\log \int \rho(\widetilde{\beta}; L\boldsymbol{\beta}_{\mathbf{E}} + m, \Sigma) \cdot \rho(\widetilde{V}; P\widetilde{\beta} + q, \Delta) \cdot 1(H\widetilde{V} \geq g)\, d\widetilde{V}\, d\widetilde{\beta} \\ &\quad\approx -\inf_{\widetilde{\beta}, \widetilde{V}} \Big\{ \frac{1}{2}(\widetilde{\beta} - L\boldsymbol{\beta}_{\mathbf{E}} - m)' \Sigma^{-1} (\widetilde{\beta} - L\boldsymbol{\beta}_{\mathbf{E}} - m) \\ &\qquad + \frac{1}{2}(\widetilde{V} - P\widetilde{\beta} - q)' \Delta^{-1} (\widetilde{V} - P\widetilde{\beta} - q) + \phi_{H,g}(\widetilde{V}) \Big\}\end{aligned} \tag{15}$$

ignoring an additive constant. The approximation in (15) then lends itself toward tractable equations for the selective MLE, $\widehat{\boldsymbol{\beta}}_{\mathbf{E}}^{\mathrm{MLE}}$, and the inverse observed Fisher information matrix, $\hat{\mathrm{I}}^{-1}$, 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*:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathbf{E}}^{MLE} &= L^{-1}\widehat{\boldsymbol{\beta}}_{\mathbf{E}} + L^{-1}\Sigma P' \Delta^{-1}(P\widehat{\boldsymbol{\beta}}_{\mathbf{E}} + q - \widehat{\mathbf{V}}) - L^{-1}m, \\ \widehat{\mathbf{I}}^{-1} &= L^{-1}\Sigma L'^{-1} + L^{-1}\Sigma\big(P'\Delta^{-1}P - P'\Delta^{-1}\big(\Delta^{-1} + \nabla^2\phi_{H,g}(\widehat{\mathbf{V}})\big)^{-1}\Delta^{-1}P\big)\Sigma L'^{-1},\end{aligned}$$

where $\widehat{\mathbf{V}}$ *is obtained from solving*

$$\widehat{\mathbf{V}} = \operatorname*{argmin}_{\widetilde{V}} \frac{1}{2}(\widetilde{V} - P\widehat{\boldsymbol{\beta}}_{\mathbf{E}} - q)' \Delta^{-1} (\widetilde{V} - P\widehat{\boldsymbol{\beta}}_{\mathbf{E}} - q) + \phi_{H,g}(\widetilde{V}). \tag{16}$$

Algorithm 2 Multitask model selection and inference

Record the values for $\widehat{\mathbf{E}}, \widehat{\mathbf{S}}, \widehat{\mathbf{\Gamma}}$, and $\widehat{\mathbf{U}}$ at convergence of Algorithm 1
From the estimated sparsity structure, compute $\widehat{\beta}_{E_k}^{(k)}$ and $\widehat{\beta}_{\perp}^{(k)}$ for $k \in [K]$
Specify a significance level α
Compute P, q, and Δ
Optimize (16) with gradient descent to compute $\widehat{\mathbf{V}}$
Compute the matrices L, m, Σ
procedure MAXIMUM LIKELIHOOD INFERENCE
    Find $\widehat{\boldsymbol{\beta}}_{\mathbf{E}}^{\mathrm{MLE}}$ and $\widehat{\mathbf{I}}^{-1}$ based on the estimating equations in Theorem 3.3
    for $j \in \{1, \ldots, \delta\}$ **do**
        Compute interval $\widehat{\boldsymbol{\beta}}_{\mathbf{E},j}^{\mathrm{MLE}} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathbf{I}}_{jj}^{-1}(\widehat{\boldsymbol{\beta}}_{\mathbf{E}}^{\mathrm{MLE}})}$
    end for
end procedure

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 $100 \cdot (1-\alpha)\%$ confidence intervals for the parameters within the MTL model (9). Algorithm 2 summarizes our procedure for post-MTL inference by reusing the same data.

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 (Karcher and Barch (2021), Luciana et al. (2018), Volkow et al. (2018)). Researchers are interested in better understanding the organization of cognitive abilities during childhood development, that is, 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 (Gignac (2014), Juan-Espinosa et al. (2000)). 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 (Simpson-Kent et al. (2020)).

Studies that leverage neurological data in studying the organization of cognitive abilities are extremely rare. Two exceptions are Tadayon, Pascual-Leone and Santarnecchi (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 (Chen et al. (2022), Marek 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 multitask approach that, as we have argued, is better suited for the dual goals of prediction and cartographic mapping.

4.1. *Multitask framework for studying ABCD data.* Applied to the ABCD data, our multitask 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 5937 subjects from 19 different research sites throughout the U.S. 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 multitask 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 four 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 internode 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 multitask learning. Previous work has shown that a small number of components is typically sufficient to capture most interindividual 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 multitask learning. One approach, which we refer to as joint, is to apply Algorithm 1 to all four tasks. This is the default approach to multitask 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 multitask 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 multitask 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 Supplementary Material.

The coefficients $\boldsymbol{\beta} \in \mathbb{R}^{p \times K}$ 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 $\boldsymbol{\beta}$ are determined as follows:

1. Select $[s_G p]$ predictors to be globally null for every task, and set the corresponding rows of $\boldsymbol{\beta}$ to zero.
2. For the remaining set of globally active predictors, set an average of $[s_T K]$ entries of each row of $\boldsymbol{\beta}$ to zero.
3. For each predictor $j \in [p]$, specify the corresponding nonzero 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 multitask 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, $\omega^{(k)} \sim N(0, v^2\sigma^2 \cdot \mathrm{I})$ 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 multitask 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, $\omega^{(k)} \sim N(0, v^2\sigma^2 \cdot \mathrm{I})$ 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, Taylor and Weinstein (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 nonzero coefficients, defined

by

$$\mathrm{CR} = 1 - \frac{|\{j \in \widehat{\mathbf{E}} : \boldsymbol{\beta}_{\mathbf{E},j} \notin \boldsymbol{C}_{\widehat{\mathbf{E}},j}\}|}{\max(|\widehat{\mathbf{E}}|, 1)}.$$

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

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

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 $\mathbf{E}_0$ be the set of true active predictors,

$$\mathrm{Precision} = \frac{|\mathbf{E}_0 \cap \{j \in \widehat{\mathbf{E}} : 0 \notin \boldsymbol{C}_{\widehat{\mathbf{E}},j}\}|}{|\{j \in \widehat{\mathbf{E}} : 0 \notin \boldsymbol{C}_{\widehat{\mathbf{E}},j}\}|}; \qquad \mathrm{Recall} = \frac{|\mathbf{E}_0 \cap \{j \in \widehat{\mathbf{E}} : 0 \notin \boldsymbol{C}_{\widehat{\mathbf{E}},j}\}|}{|\mathbf{E}_0|}.$$

Figure 1 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 a similar amount of information for inference. 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 Supplementary Material; please see Section 3.

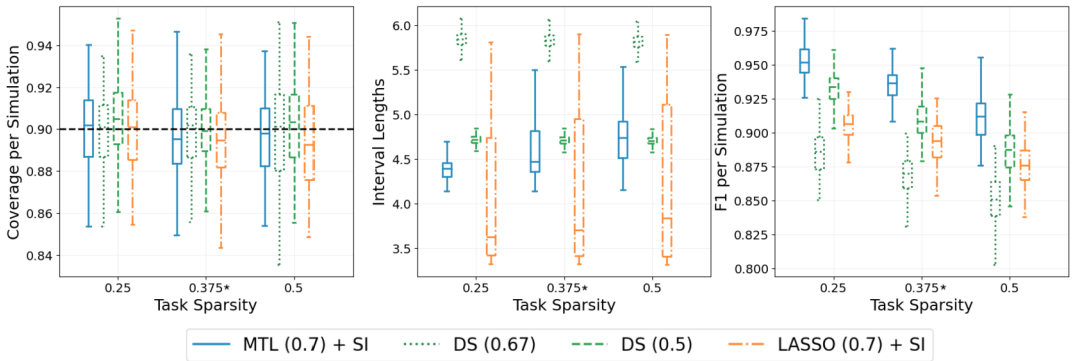

FIG. 1. *Comparison of MTL*(0.7) + *SI against three alternative approaches*: *DS*(0.67), *DS*(0.5), *and LASSO*(0.7) + *SI*, *with* 100 *replications for each level of task sparsity* s_T. *Global sparsity is fixed at* $s_G = 0.85$. *The asterisk* (*) *denotes the task sparsity level closest to the fMRI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

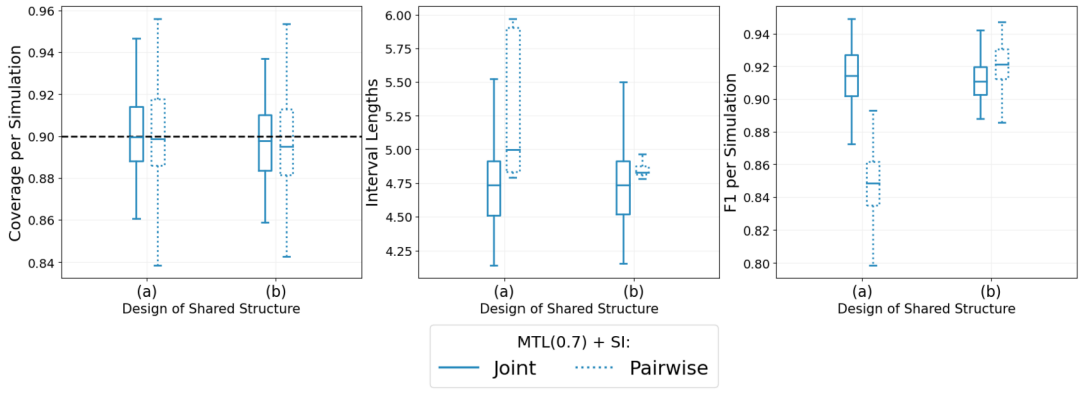

FIG. 2. *Comparison of the joint and pairwise approaches under setup* (*a*), *where any two tasks have roughly the same amount of common structure, and setup* (*b*), *where only the two pairs have common structure. Global and task sparsity are fixed at* $s_G = 0.85$ *and* $s_T = 0.5$. *Outliers are not shown. Left*: *Coverage*; *Middle*: *Interval length*; *Right*: *F*1 *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.

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

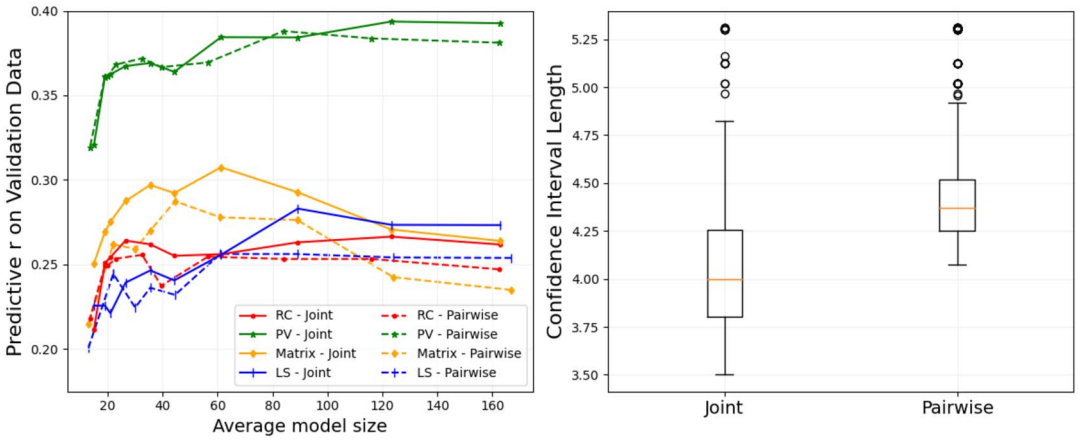

FIG. 3. *Left*: *The predictive r for the joint and pairwise approaches to MTL*(0.7) + *SI. Right*: *The distribution of confidence interval lengths across tasks for the joint and pairwise approaches to MTL*(0.7) + *SI.*

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

	Joint predictive r	Pairwise predictive r
NIH Tlbx RC	0.335	0.321
NIH Tlbx PV	0.470	0.451
Matrix Reasoning	0.283	0.273
NIH Tlbx LS	0.354	0.317

on validation data. Following the convention in the neuroimaging literature (Sripada et al. (2021), Sripada et al. (2020)), 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 multitask 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 different cognitive abilities. 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

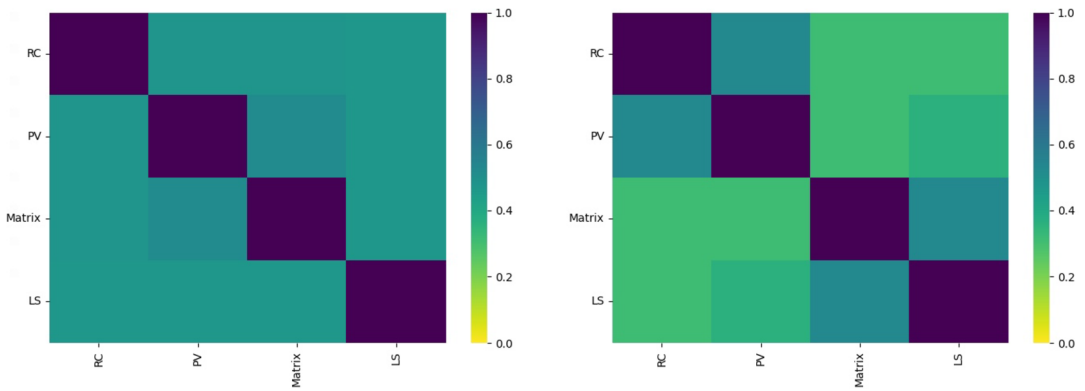

FIG. 4. *Similarity of selected models for the four tasks. Left: Jaccard index between sets of significant PCs recovered through the joint approach. Right: Jaccard index between sets of significant PCs recovered through the pairwise approach.*

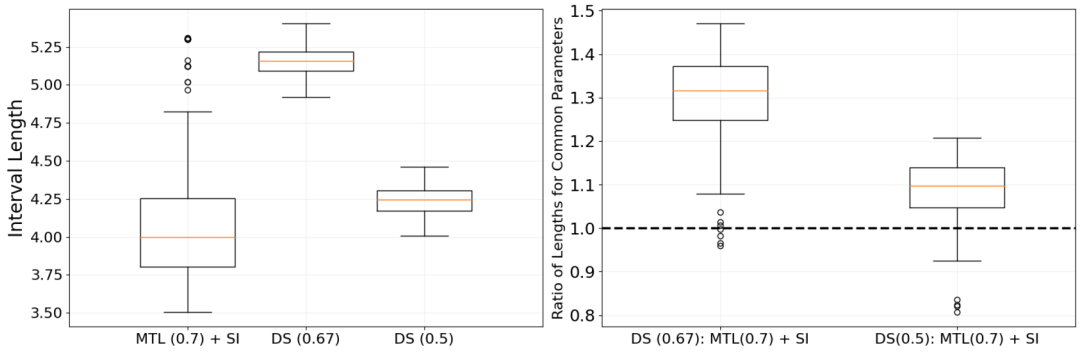

FIG. 5. *Comparison of inferential power between MTL*(0.7) + *SI*, *DS*(0.67) *and DS*(0.5) *when applied jointly to all four tasks*. *Left*: *Boxplots of interval lengths*. *Right*: *Ratio of interval lengths for common parameters*.

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 multitask 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 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.

4.4. *Neurological interpretation of selected models*. We now offer an interpretation of the shared structure identified through multitask 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

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

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

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

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
NONE	Not Named	CP	Cingulo-Parietal
RST	Retrosplenial Temporal		

feature space. Note that the mean structure for task k in the original feature space of dimension p can be represented as

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

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

$$\hat{\beta}^{(k)} = M\begin{pmatrix}\hat{\theta}^{(k)}_{E_k} \\ 0\end{pmatrix} = M_{E_k}\hat{\theta}^{(k)}_{E_k},$$

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, Smallwood and Spreng (2014)) and semantic/conceptual representation (Binder and Desai (2011), Binder et al. (2009), Wirth et al. (2011)), capacities that likely facilitate abstraction and problem-solving. Cerebellum has been traditionally associated with coordination of movement (Bastian (2006), Spencer, Ivry and Zelaznik (2005)), 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 (Andreasen et al. (1999), Schmahmann (1996), Schmahmann (2019)). 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.

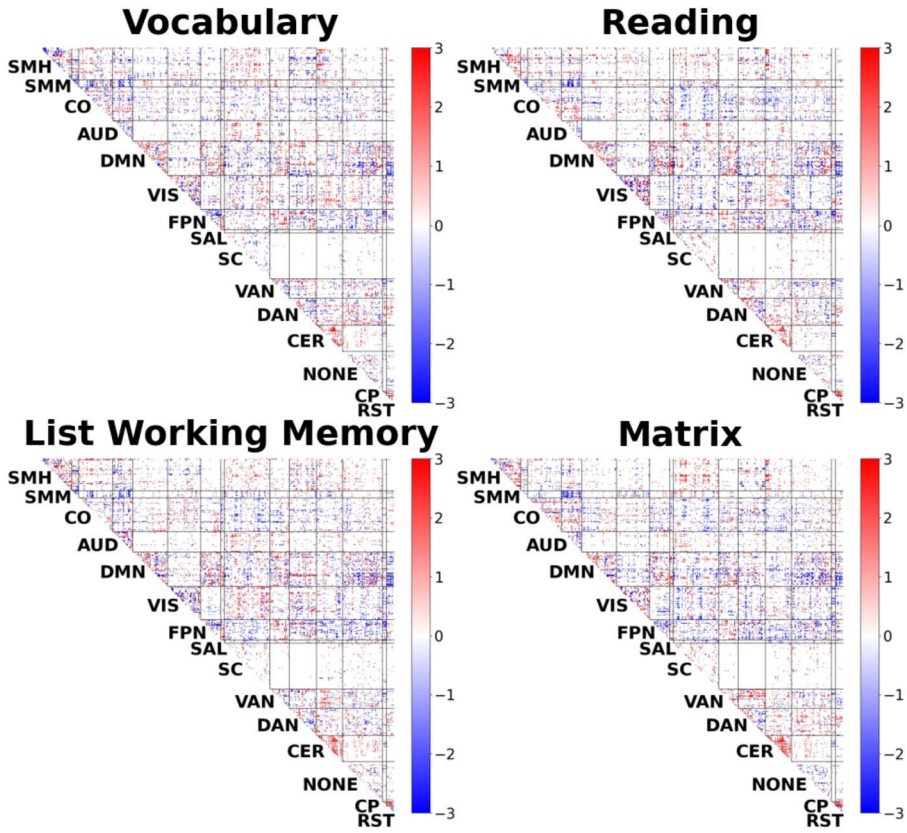

FIG. 6. *Cartographic visualization showing standardized estimates of the original coefficients, grouped by region/network, that were obtained after applying MTL*(0.7) + *SI jointly to all four tasks. A threshold of* 1.5 *is used to improve readability, with smaller-magnitude coefficients set to zero.*

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 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.

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 cartographic 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 Development (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, aged 9–10, and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes 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 participating 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 report is NDA Study 1689, 10.15154/1527789.

Funding. 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.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/23-AOAS1796SUPP; .pdf). Proofs for our main results and additional simulations are collected in the Supplementary Material.

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. (2023). Investigating cognitive neuroscience theories of human intelligence: A connectome-based predictive modeling approach. *Hum. Brain Mapp.* **44** 1647–1665. https://doi.org/10.1002/hbm.26164

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. *Biol. Psychiatry* **46** 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. *Ann. N.Y. Acad. Sci.* **1316** 29–52. https://doi.org/10.1111/nyas.12360

BASTIAN, A. J. (2006). Learning to predict the future: The cerebellum adapts feedforward movement control. *Curr. Opin. Neurobiol.* **16** 645–649. https://doi.org/10.1016/j.conb.2006.08.016

BI, J., XIONG, T., YU, S., DUNDAR, M. and RAO, R. B. (2021). An improved multi-task learning approach with applications in medical diagnosis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 117–132.

BINDER, J. R. and DESAI, R. H. (2011). The neurobiology of semantic memory. *Trends Cogn. Sci.* **15** 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. *Cereb. Cortex* **19** 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. *Behav. Brain Sci.* **29** 109–125.

CARROLL, J. B. et al. (1993). *Human Cognitive Abilities*: *A Survey of Factor-Analytic Studies* 1. Cambridge Univ. Press, Cambridge.

CATTELL, R. B. (1943). The measurement of adult intelligence. *Psychol. Bull.* **40** 153.

CHEN, J., TAM, A., KEBETS, V., ORBAN, C., OOI, L. Q. R., ASPLUND, C. L., MAREK, S., DOSENBACH, N. U. F., 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. *Nat. Commun.* **13** 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. *Sci. Rep.* **9** 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. *Nat. Neurosci.* **16** 1348–1355. https://doi.org/10.1038/nn.3470

COLE, M. W. and SCHNEIDER, W. (2007). The cognitive control network: Integrated cortical regions with dissociable functions. *NeuroImage* **37** 343–360. https://doi.org/10.1016/j.neuroimage.2007.03.071

CORDES, D. and NANDY, R. R. (2006). Estimation of the intrinsic dimensionality of fMRI data. *NeuroImage* **29** 145–154. https://doi.org/10.1016/j.neuroimage.2005.07.054

COX, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189 https://doi.org/10.1093/biomet/62.2.441

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. *Nat. Neurosci.* **18** 1664–1671. https://doi.org/10.1038/nn.4135

FITHIAN, W., SUN, D. and TAYLOR, J. (2017). Optimal inference after model selection. Available at arXiv:1410.2597.

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. F. 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. *Cereb. Cortex* **26** 288–303. https://doi.org/10.1093/cercor/bhu239

GRAY, J. R., CHABRIS, C. F. and BRAVER, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* **6** 316–322. https://doi.org/10.1038/nn1014

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15. MR2804206 https://doi.org/10.1093/biomet/asq060

HAGLER JR., D. J., HATTON, S., CORNEJO, M. D., MAKOWSKI, C., FAIR, D. A., DICK, A. S., SUTHERLAND, M. T., CASEY, B. J., BARCH, D. M. 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. *Sci. Rep.* **6** 1–8.

HONG, L., KUFFNER, T. A. and MARTIN, R. (2018). On overfitting and post-selection uncertainty assessments. *Biometrika* **105** 221–224. MR3768876 https://doi.org/10.1093/biomet/asx083

HORN, J. L. and NOLL, J. (1997). *Human cognitive capabilities*: *Gf-Ge theory*. Guilford, New York.

HÜLÜR, G., WILHELM, O. and ROBITZSCH, A. (2011). Intelligence differentiation in early childhood. *J. Individ. Differ.*

HUMPHREYS, L. G. (1979). The construct of general intelligence. *Intelligence* **3** 105–120.
JOLLIFFE, I. T. (2003). Principal component analysis. *Technometrics* **45** 276.
JUAN-ESPINOSA, M., GARCÍA LUÍS, F., COLOM, R. and ABAD, F. J. (2000). Testing the age related differentiation hypothesis through the Wechsler's scales. *Pers. Individ. Differ.* **29** 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** 131–142. https://doi.org/10.1038/s41386-020-0736-6
KIVARANOVIC, D. and LEEB, H. (2018). Expected length of post-model-selection confidence intervals conditional on polyhedral constraints. Preprint. Available at arXiv:1803.01665.
LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371
LIU, K., MARKOVIC, J. and TIBSHIRANI, R. (2018). More powerful post-selection inference. with application to the Lasso. Preprint. Available at arXiv:1801.09037.
LOZANO, A. C. and SWIRSZCZ, G. (2012). Multi-level lasso for sparse multi-task regression. In *Proceedings of the* 29*th International Conference on Machine Learning* 595–602.
LUCIANA, M., BJORK, J. M., NAGEL, B. J., BARCH, D. M., GONZALEZ, R., NIXON, S. J. and BANICH, M. T. (2018). Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (ABCD) baseline neurocognition battery. *Dev. Cogn. Neurosci.* **32** 67–79. https://doi.org/10.1016/j.dcn.2018.02.006
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. et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature* **603** 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 Behav.* **12** 238–246. https://doi.org/10.1007/s11682-017-9688-9
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. *Cogn. Affect. Behav. Neurosci.* **12** 241–268.
PANIGRAHI, S. (2018). Carving model-free inference. Preprint. Available at arXiv:1811.03142.
PANIGRAHI, S., MOHAMMED, S., RAO, A. and BALADANDAYUTHAPANI, V. (2023). Integrative Bayesian models using post-selective inference: A case study in radiogenomics. *Biometrics* **79** 1801–1813. https://doi.org/10.1111/biom.13740
PANIGRAHI, S., STEWART, N., SRIPADA, C. and LEVINA, E. (2024). Supplement to "Selective inference for sparse multitask regression with applications in neuroimaging." https://doi.org/10.1214/23-AOAS1796SUPP
PANIGRAHI, S. and TAYLOR, J. (2018). Scalable methods for Bayesian selective inference. *Electron. J. Stat.* **12** 2355–2400. MR3832095 https://doi.org/10.1214/18-EJS1452
PANIGRAHI, S. and TAYLOR, J. (2022). Approximate selective inference via maximum likelihood. *J. Amer. Statist. Assoc.* Forthcoming.
PANIGRAHI, S., TAYLOR, J. and WEINSTEIN, A. (2021). Integrative methods for post-selection inference under convex constraints. *Ann. Statist.* **49** 2803–2824. MR4338384 https://doi.org/10.1214/21-aos2057
PANIGRAHI, S., WANG, J. and HE, X. (2022). Treatment Effect Estimation with Efficient Data Aggregation. Preprint. Available at arXiv:2203.12726.
SCHMAHMANN, J. D. (1996). From movement to thought: Anatomic substrates of the cerebellar contribution to cognitive processing. *Hum. Brain Mapp.* **4** 174–198.
SCHMAHMANN, J. D. (2019). The cerebellum and cognition. *Neurosci. Lett.* **688** 62–75. https://doi.org/10.1016/j.neulet.2018.07.005
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. *Dev. Cogn. Neurosci.* **41** 100743.
SPEARMAN, C. (1961). "General Intelligence" Objectively Determined and Measured.
SPENCER, R. M. C., IVRY, R. B. and ZELAZNIK, H. N. (2005). Role of the cerebellum in movements: Control of timing or movement transitions? *Exp. Brain Res.* **161** 383–396. https://doi.org/10.1007/s00221-004-2088-6
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. *Sci. Rep.* **9** 1–12.
SRIPADA, C., ANGSTADT, M., TAXALI, A., CLARK, D. A., GREATHOUSE, T., RUTHERFORD, S., DICKENS, J. R., SHEDDEN, K., GARD, A. M. et al. (2021). Brain-wide functional connectivity patterns support general cognitive ability and mediate effects of socioeconomic status in youth. *Transl. Psychiatry* **11** 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. *Mol. Psychiatry* **25** 3413–3421. https://doi.org/10.1038/s41380-019-0481-6

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* 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. *Cereb. Cortex* **30** 215–225. https://doi.org/10.1093/cercor/bhz082

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* 9553–9562.

TAYLOR, J. and TIBSHIRANI, R. (2018). Post-selection inference for ℓ_1-penalized likelihood models. *Canad. J. Statist.* **46** 41–61. MR3767165 https://doi.org/10.1002/cjs.11313

TIAN, X., PANIGRAHI, S. MARKOVIC, J., BI, N. and TAYLOR, J. (2016). Selective sampling after solving a convex problem. Preprint. Available at arXiv:1609.05609.

TIAN, X. and TAYLOR, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46** 679–710. MR3782381 https://doi.org/10.1214/17-AOS1564

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. *Transl. Psychiatry* **12** 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. *Dev. Cogn. Neurosci.* **32** 4–7.

WANG, J., HE, X. and XU, G. (2020). Debiased inference on treatment effect in a high-dimensional model. *J. Amer. Statist. Assoc.* **115** 442–454. MR4078474 https://doi.org/10.1080/01621459.2018.1558062

WANG, X., BI, J., YU, S., SUN, J. and SONG, M. (2016). Multiplicative multitask feature learning. *J. Mach. Learn. Res.* **17** Paper No. 80, 33. MR3517103

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** 3057–3066. https://doi.org/10.1016/j.neuroimage.2010.10.039

ZHANG, Y. and YANG, Q. (2021). A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.*

ZHAO, Q. and PANIGRAHI, S. (2019). Selective inference for effect modification: An empirical investigation. *Obs. Stud.* **5** 131–140.

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802