A sensitivity analysis to quantify the impact of neuroimaging preprocessing strategies on subsequent statistical analyses

Brice Ozenne[1,2][0000−0001−9694−2956], Martin Nørgaard[3,4][1111−2222−3333−4444], Cyril Pernet[1][2222−−3333−4444−5555], and Melanie Ganz[1,4][2222−−3333−4444−5555]

[1] Neurobiology Research Unit, Rigshospitalet
[2] Section of Biostatistics, University of Copenhagen
[3] Molecular Imaging Branch, National Institute of Mental Health (NIMH)
[4] Department of Computer Science, University of Copenhagen
brice-ozenne@nru.dk

Abstract. Even though novel imaging techniques have been successful in studying brain structure and function, the measured biological signals are often contaminated by multiple sources of noise, arising due to e.g. head movements of the individual being scanned, limited spatial/temporal resolution, or other issues specific to each imaging technology. Data preprocessing (e.g. denoising) is therefore critical. Preprocessing pipelines have become increasingly complex over the years, but also more flexible, and this flexibility can have a significant impact on the final results and conclusions of a given study. This large parameter space is often referred to as multiverse analyses. Here, we provide conceptual and practical tools for statistical analyses that can aggregate multiple pipeline results along with a new sensitivity analysis testing for hypotheses across pipelines such as "no effect across all pipelines" or "at least one pipeline with no effect". The proposed framework is generic and can be applied to any multiverse scenario, but we illustrate its use based on positron emission tomography data.

Keywords: neuroimaging · preprocessing · multiverse analyses

1 Introduction

Modern neuroimaging techniques have provided unique opportunities to measure complex signaling pathways in the living human brain with the goal of identifying reliable biomarkers of disease states and treatment outcomes [10]. Data arising from state-of-the-art neuroimaging techniques are, however, often contaminated with noise confounds such as motion-related artefacts, affecting both the spatial and temporal correlation structure of the data [9]. Carefully designed preprocessing steps have been developed to remove unwanted noise sources, but in the absence of a "ground truth" it remains a major challenge to evaluate the impact of preprocessing choices on subsequent statistical analyses and results. Over time, preprocessing pipelines (i.e. a set of preprocessing steps)

have become more complex and flexible, and this increase in researcher degrees of freedom (termed multiverse analyses) has consistently been shown to affect the outcomes of neuroimaging studies [2, 3, 8]. The most common approach in the neuroimaging field is, to date, to use a single pipeline and ignore the heterogeneity of preprocessing choices. This approach not only makes abstraction of the multitude of possible results but is likely also sub-optimal because the best pipeline is more often than not, study, population or even subject dependent [4, 8]. More concerning, neuroscientists might be tempted to "tune" the pipeline in order to obtain the most satisfying results. This will generally lead to spurious and non-reproducible results since the variability induced by the choice of pipeline is not independent from the results. However, since it is neither realistic nor optimal to move toward a single unified preprocessing pipeline, there is an urgent need for a statistical framework allowing to explore results among many preprocessing pipelines in a principled way.

The aim of this work is thus to provide a statistical framework that can aggregate the evidence from multiverse analyses to produce conclusions robust to the choice of the pipeline. More specifically, the present paper proposes a statistical sensitivity analysis providing:

(i) visualizations of the heterogeneity of several preprocessing pipelines
(ii) estimation of a global effect across all preprocessing pipelines
(iii) quantification of the proportion of pipelines with evidence for an effect
(iv) a statistical framework for testing hypotheses across pipelines such as "no effect across all pipelines", "at least one pipeline with no effect"

The corresponding software can be found at *anonymous*, including all code to reproduce all simulations and figures.

2 Materials and experimental settings

2.1 Data

We use two different data sources in our analyses: in sillico and real data. For the in sillico data, different noise structures were chosen to reflect different configurations of pipelines and the sample size was varied to encompass small to larger scale clinical studies. In the real data analysis, we mimic how real neuroimaging studies compare an intervention to a reference measurement. Pipelines were selected independently of the intervention data using the healthy/placebo arm of [6] using results from [8]. The proposed sensitivity analysis was illustrated on the intervention arm of the study where the follow-up value was compared to a reference value taken from a normative serotonergic atlas [1].

To simulate in sillico data, we consider the simple case of a single brain measurement ($R = 1$), with a single binary exposure ($P = 1$) following a Bernoulli distribution with parameter $\pi = 0.5$ (i.e. two balanced groups) and no covariates ($C = 0$). Latent Y values for the brain measurement are simulated using a normal distribution with variance 1 and mean β times the exposure X values where

$\beta = 0$ (null hypothesis) or $\beta = 0.5$ (alternative hypothesis). The observed Y is simulated for each pipeline adding pipeline specific noise to the latent Y. This noise is simulated using a multivariate normal distribution with mean 0 and variance $\Sigma_{\text{scenario 1}}$, $\Sigma_{\text{scenario 2}}$, and $\Sigma_{\text{scenario 3}}$ depending on the scenario. In scenario 1, we simulated many pipelines ($J = 20$) with correlated homoscedastic noise, in scenario 2 a few pipelines ($J = 6$) with uncorrelated heteroscedastic noise, and in scenario 3 many pipelines ($J = 20$) with correlated heteroscedastic noise. Scenario 2 and 3 included one pipeline with high signal to noise ratio (SNR), i.e. low variance, and many pipelines with low SNR, i.e., high variance. The sample size was varied from $n = 10$ to $n = 500$ in each group, such that the smallest sample size was well below the number of parameters ($2 \times J$ mean parameters, J^2 variance-covariance parameters) in scenario 1 and 3 and the asymptotic regime was reached for the largest sample size. We generate 10,000 datasets per scenario and sample size - this provides sufficient precision about the mean, standard deviation, and rejection rate to neglect the Monte Carlo uncertainty. This data will be used to assess the large sample size properties of the procedure (bias, relative efficacy, type 1 error control) in finite samples.

To illustrate the use of the proposed sensitivity analysis on real data, we utilize neuroimaging results from a placebo-controlled, double-blinded, clinical study [6]. The study was registered and approved by the ethics committee for the capital region of Copenhagen (protocol-ID: H-2-2010-108) and registered as a clinical trial: `www.clinicaltrials.gov` under the trial ID NCT02661789. All subjects provided written informed consent prior to participation, in accordance with The Declaration of Helsinki II. The aim of the study was to assess the association between the emergence of depressive symptoms and change in cerebral serotonin transporter (SERT) availability following a hormonal treatment ($p = 1$). Data is available from the CIMBI database (`www.cimbi.dk`) upon request. It consists of structural Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) imaging data for 60 healthy females who underwent a baseline scan, received either Placebo ($n = 30$) or a GnRHa implant intervention ($n = 30$), and participated in a follow-up scan. SERT availability estimates were extracted for each subject for 28 subcortical and cortical regions, and averaged across hemispheres, producing a final sample of $R = 14$ regions per subject and pipeline. These regions (amygdala, thalamus, putamen, caudate, anterior cingulate cortex, hippocampus, orbital frontal cortex, superior frontal cortex, occipital cortex, superior temporal gyrus, insula, inferior temporal gyrus, parietal cortex, and entorhinal cortex) were chosen because they cover the entire brain, and many are target regions in published serotonin transporter (SERT) PET studies. No covariates where consider and ($C = 0$).

2.2 Data preprocessing

Five preprocessing steps were used to curate the data and estimate the SERT availability (outcome measure). These steps include, motion correction (with/without),

co-registration (4 options), delineation of volumes of interest (3 options), partial volume correction (4 options), and kinetic modeling for quantification of SERT availability (MRTM, SRTM, Non-invasive Logan and MRTM2). More information about the preprocessing choices can be found in [8]. The combination of individual preprocessing steps leads to a number of $J = 2 \times 3 \times 4^3 = 384$ possible combinations.

2.3 Notation and assumptions

In the following section, we use generic notations to describe the proposed sensitivity analysis. We are interested in relating R brain measurements ($\boldsymbol{Y} = (Y_1, \ldots, Y_R)$) to P exposures or treatments ($\boldsymbol{X} = (X_1, \ldots, X_P)$) accounting for C covariates ($\boldsymbol{W} = (W_1, \ldots, W_C)$). We consider a set of J pipelines used to preprocess the neuroimaging data. For a given pipeline $j \in \{1, \ldots, J\}$ we fit a statistical model with parameters θ_j that relates $\boldsymbol{Y}$ processed by pipeline j, $\boldsymbol{X}$, and $\boldsymbol{W}$. We then obtain from this model an estimate $\widehat{\psi}_k$ of the effect of interest (denoted ψ). In our real life example we use, for each pipeline, a paired t-test to compare the observed change in SERT availability to an atlas value so for $j \in \{1, \ldots, J\}$, $\widehat{\theta}_j = (\widehat{\psi}_j, \widehat{\sigma}_j^2)$ where $\widehat{\psi}_j$ is the empirical mean and σ_j^2 the empirical variance of the change in SERT availability (processed with pipeline j) between baseline of follow-up.

We make the following working assumptions: first the observed data $(\mathcal{O}_i)_{i\in\{1,\ldots\}} = (\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{w}_i)_{i\in\{1,\ldots,n\}}$ correspond to independent and identically distributed replicates of $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W})$. Second we have chosen a set of reasonable pipelines, meaning that the estimated effects $\widehat{\psi}_1, \ldots, \widehat{\psi}_J$ found in the follow-up statistical analysis will converge to the right value ψ as the sample size increases. This set can include pipelines distorting the signal $\boldsymbol{Y}$ (e.g. adding a fixed value) if that has no consequence, asymptotically, on the estimated effect (the mean change is not biased as the added value cancels out when substracting the baseline value to the follow-up value). Finally, when considering asymptotic results we will consider a fixed number of pipelines and let the sample size n increase to infinity.

3 Proposed sensitivity analysis

To be able to draw conclusions across pipelines, we not only need the result of each pipeline but also some information about how they relate. If all pipelines were equally reliable and equally different, we would weight each pipeline equally. If there exists one independent pipeline and a block of correlated pipelines, all equally reliable, then we would weight the independent pipeline more compared with other pipelines. By treating pipelines as black boxes, we can investigate their relation in terms of each observations influence on the estimated effects across pipelines, $\widehat{\boldsymbol{\psi}} = \left(\widehat{\psi}_1, \ldots, \widehat{\psi}_J\right)$. This relation is fully characterized by the joint distribution of the effects. Once estimated, we can extract summaries of

this distribution, e.g. an average value, and carry out statistical tests, e.g. testing the compatibility between the observations and the joint distribution that would have been observed under a specific hypothesis.

3.1 Estimating the joint distribution across pipelines

The joint distribution could be obtained using a multivariate model, e.g. modelling data from all pipelines at once using a mixed model. Because of the complexity of the dependency among pipelines, this is, however, rarely feasible with the available sample size. Instead, and this matches common practice, we perform the same analysis separately for each pipeline and obtain a vector of estimated associations $\widehat{\boldsymbol{\psi}}$ with their standard errors $\boldsymbol{\sigma}_{\widehat{\psi}} = \left(\sigma_{\widehat{\psi}_1}, \ldots, \sigma_{\widehat{\psi}_J}\right)$. Using tools from the semi-parametric theory (see [7] and [12] for more details), we can approximate the influence of each observation on the estimate by a random variable called the influence function, denoted $\varphi_{\widehat{\psi}_j}$ for pipeline j, and satisfying:

$$\sqrt{n}\left(\widehat{\psi}_j - \psi\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{\widehat{\psi}_j}(\mathcal{O}_i) + o_p(1)$$

where $o_p(1)$ denotes a residual term that convergences toward zero in probability as the sample size tends to infinity. As shown in appendix A, $\varphi_{\widehat{\psi}_j}$ has a simple expression when $\widehat{\psi}_j$ is the empirical mean or an element of a maximum likelihood (ML) estimator. Since this decomposition applies to all pipelines, we get from the multivariate central limit theorem that the joint distribution of the estimates is asymptotically multivariate normal. It has mean ψ and its variance-covariance, denoted $\Sigma_{\widehat{\boldsymbol{\psi}}}$, is the same as the one of $\boldsymbol{\varphi}_{\widehat{\boldsymbol{\psi}}} = \left(\varphi_{\widehat{\psi}_1}, \ldots, \varphi_{\widehat{\psi}_J}\right)$ divided by n. Note that with limited number of observations, typically when $n < J$, the estimated variance-covariance $\widehat{\Sigma}_{\widehat{\boldsymbol{\psi}}}$ based on the estimated influence function is not guaranteed to be positive definite.

3.2 Testing hypotheses across pipelines

The global null hypothesis "no effect across all pipelines" can be tested using a max-test approach: more extreme realizations would correspond to larger values of the maximum statistic $\widehat{\boldsymbol{t}}_{\max} = \max(|\widehat{t}_1|, .., |\widehat{t}_J|)$ where $|.|$ denotes the absolute value and $\widehat{t}_j = \frac{\widehat{\psi}_j}{\sigma_{\widehat{\psi}_j}}$. A p-value may therefore be computed by integrating the joint density under the null hypothesis outside of the domain $\mathcal{D}(\widehat{\boldsymbol{t}}_{\max}) = [-\widehat{\boldsymbol{t}}_{max}, \widehat{\boldsymbol{t}}_{max}]^{\otimes J}$ (see figure A in appendix B). Here, we use the notation $[a, b]^{\otimes J} = \prod_{j=1}^{J} [a, b]$ that represents the Cartesian product between J intervals $[a, b]$. The value t_c, such that the integral outside $\mathcal{D}(t_c)$ equals α, provides a critical threshold for the estimated test statistics $\left(|\widehat{t}_1|, .., |\widehat{t}_J|\right)$. This threshold can also be used to derive confidence intervals.

The null hypothesis "at least one pipeline with no effect" is an intersection union test. As such it can be rejected if and only if all the un-adjusted p-values relative to each pipeline are below α or equivalently if the largest un-adjusted p-value is below α.

The proportion of pipelines where there is evidence for an effect η can be estimated as $\tilde{\eta} = \frac{1}{J}\sum_{j=1}^{J} \mathbb{1}_{|\widehat{t}_j|<t_c}$ where $\mathbb{1}_{\cdot}$ denotes the indicator function. One drawback with this non-parametric estimator is that it is a non-smooth function of $\widehat{t}_j$, making the associated uncertainty difficult to evaluate. Instead, one can use a parametric approach, assuming normally distributed test statistics:

$$\widehat{\eta} = \frac{1}{J}\sum_{j=1}^{J} 1 - \mathbb{P}\left[-t_c < \widehat{t}_j < t_c\right] = 1 - \frac{1}{J}\sum_{j=1}^{J} \Phi\left(t_c - \widehat{t}_j\right) - \Phi\left(-t_c - \widehat{t}_j\right)$$

which is a smooth (but complex) function of the model parameters. Here Φ refers to the cumulative distribution function of a standard normal distribution. The uncertainty about the estimator can therefore be derived using a non-parametric bootstrap or a delta method where $Var(\widehat{\eta}) = \frac{\partial \widehat{\eta}}{\partial \Theta} \Sigma_{\widehat{\Theta}} \frac{\partial \widehat{\eta}}{\partial \Theta}^{\intercal}$ where $\Theta = (\theta_j)_{j\in\{1,\dots,J\}}$ is the set of parameters of the statistical model across pipelines.

3.3 Visualizing the heterogeneity across pipelines

We have derived that the estimated associations are, asymptotically, normally distributed. They can therefore be summarized by their expectation and variance-covariance matrix (i.e. standard errors and correlation matrix). We suggest two graphical displays to visualize the heterogeneity of the results across pipelines:

- A heatmap of the estimated correlation among estimates, obtained by converting $\widehat{\Sigma}_{\widehat{\boldsymbol{\psi}}}$ into a correlation matrix. [5]
- A forest plot displaying $\left(\widehat{\psi}_1, \dots, \widehat{\psi}_J\right)$ and $\left(\widehat{\sigma}_{\widehat{\psi}_1}, \dots, \widehat{\sigma}_{\widehat{\psi}_J}\right)$ through the confidence intervals, possibly using the previously established re-ordering.

3.4 Estimating a global effect across pipelines

Several methods can be used for estimating a global effect across pipelines. A naive method would be to compute the mean of the estimated associations:

$$\widehat{\Psi}_{\text{average}} = \frac{1}{J}\sum_{j=1}^{J} \widehat{\psi}_j$$

This estimator will, however, not be efficient if some pipelines lead to more precise estimates, i.e. $\left(\sigma_{\widehat{\psi}_1}, \dots, \sigma_{\widehat{\psi}_J}\right)$ are not equal. Intuitively, we would like

[5] Re-ordering the pipelines may be useful to better visualize blocks of pipelines that are especially correlated. This can for instance be performed by converting R to a dissimilarity matrix and then use hierarchical clustering.

to pool the estimates with weights inversely proportional to the standard errors such that we put more weight on precise estimates:

$$\widehat{\Psi}_{\text{pool-se}} = \sum_{j=1}^{J} w_j^{se} \widehat{\psi}_j \text{ where } w_j^{se} = \frac{1/\sigma^2_{\widehat{\psi}_j}}{\sum_{j=1}^{J} 1/\sigma^2_{\widehat{\psi}_j}}$$

However, we also need to take into account the correlation between estimates. Indeed, perfectly correlated estimates should weight as if there was only one estimate. To do so, we can use the following GLS estimator of the global effect:

$$\widehat{\Psi}_{\text{GLS}} = \left(\mathbf{1}^\intercal \widehat{\Sigma}^{-1}_{\widehat{\Psi}} \mathbf{1}\right)^{-1} \mathbf{1}^\intercal \widehat{\Sigma}^{-1}_{\widehat{\Psi}} \widehat{\boldsymbol{\psi}} = \sum_{j=1}^{J} w_j^{GLS} \widehat{\psi}_j \tag{1}$$

where $\mathbf{1}$ is a column vector filled with ones. Equation 1 can be shown to be equivalent to performing a spectral decomposition of $\widehat{\Sigma}_{\widehat{\boldsymbol{\psi}}}$, and use the eigenvectors to combine estimates into independent components that can be pooled according to weights proportional to the eigenvalues (appendix C.1). This is used when $\Sigma_{\widehat{\boldsymbol{\psi}}}$ is singular to compute $\widehat{\Psi}_{\text{GLS}}$, by restricting the spectral decomposition to the eigenvalues above a given threshold ($\epsilon = 10^{-10}$). In the simple case where $R = 1$, $p = 1$, $C = 0$, brain measurements are jointly normally distributed, X is binary, and there is no missing data, the GLS estimator can be shown to be asymptotically efficient (appendix C.2).

Cox and colleagues [5] studied a similar estimator when $J = 2$ and found that, under unequal variance, the global estimate can be outside of the range of the (pipeline specific) estimates. As a remedy to this unpleasant behavior, we propose a constrained GLS estimator, denoted $\widehat{\Psi}_{\text{constrained GLS}}$, which constrains the weight of each estimate to be at most 1 in absolute value and ensures that the sum of the weights is 1:

$$w_j^{\text{constrained GLS}} = \frac{w_j^{GLS}}{\kappa + \max_{j\in\{1,\dots,J\}} |w_j^{GLS}|} + \frac{1}{J}\left(1 - \frac{1}{\kappa + \max_{j\in\{1,\dots,J\}} |w_j^{GLS}|}\right)$$

where κ is chosen to satisfy the constraints. Note that if some of the pipelines may induce some bias, the previous approaches will propagate this bias and therefore be unsatisfactory. Systematic differences between pipelines can be investigated by comparing the estimates between pipelines, e.g. $\widehat{\psi}_j - \widehat{\psi}_{j'}$ and using $\widehat{\Sigma}_{\widehat{\boldsymbol{\psi}}}$ to obtain the corresponding uncertainty $\mathbb{V}ar\left[\widehat{\psi}_j - \widehat{\psi}_{j'}\right] = \mathbb{V}ar\left[\widehat{\psi}_j\right] + Var\left[\widehat{\psi}_{j'}\right] - 2\mathbb{C}ov\left(\widehat{\psi}_j, \widehat{\psi}_{j'}\right)$.

4 Results

4.1 Simulation results - pooled estimator

We compare the performance of the proposed estimators when using unbiased pipelines on simulated data. For each dataset $\Sigma_{\text{scenario 1}}$, $\Sigma_{\text{scenario 2}}$, and

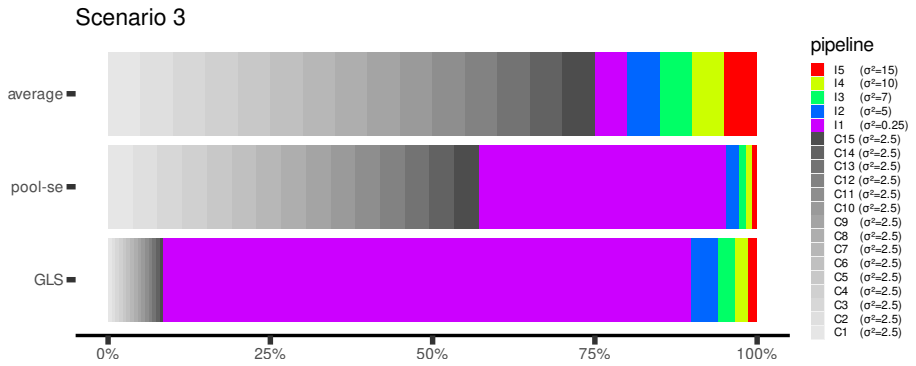

Fig. 1. Large sample weights used by $\widehat{\Psi}_{\text{average}}$, $\widehat{\Psi}_{\text{pool-se}}$, and the GLS estimators ($\widehat{\Psi}_{\text{GLS}}$ or $\widehat{\Psi}_{\text{constrained GLS}}$) to combine the pipeline specific estimates in scenario 3. Weights relative to the correlated pipeline are shown in shades of gray (first fifteen blocks); weights relative to the independent pipeline are shown using rainbow colors (last five blocks). In the legend the variance for a given pipeline effect is indicated in parenthesis.

$\Sigma_{\text{scenario 3}}$, we computed the four previously described estimators of the global effect, $\widehat{\Psi}_{\text{average}}$, $\widehat{\Psi}_{\text{pool-se}}$, $\widehat{\Psi}_{\text{GLS}}$, and $\widehat{\Psi}_{\text{constrained GLS}}$. P-values for $\widehat{\Psi}_{\text{pool-se}}$, $\widehat{\Psi}_{\text{GLS}}$, and $\widehat{\Psi}_{\text{constrained GLS}}$ were computed neglecting the uncertainty of the weights (w_j^{se}, w_j^{GLS}, $w_j^{\text{constrained GLS}}$).

Weights: Based on $\Sigma_{\text{scenario 1}}$, $\Sigma_{\text{scenario 2}}$, and $\Sigma_{\text{scenario 3}}$, we can compute how $\widehat{\Psi}_{\text{average}}$, $\widehat{\Psi}_{\text{pool-se}}$, and $\widehat{\Psi}_{\text{GLS}}$ (or $\widehat{\Psi}_{\text{constrained GLS}}$) would weight the results from each pipeline if the variance-covariance matrix of the pipeline estimates was known (Figure 1). In scenario 1, $\widehat{\Psi}_{\text{average}}$ and $\widehat{\Psi}_{\text{pool-se}}$ would provide equal weight to all pipelines with a weight of 5%, while $\widehat{\Psi}_{\text{GLS}}$ would weight by 2% the correlated pipelines and by 14% the uncorrelated pipelines (in this paragraph all weights are rounded for readability). In scenario 2, $\widehat{\Psi}_{\text{average}}$ equally weights all pipelines by 17% while $\widehat{\Psi}_{\text{pool-se}}$ and $\widehat{\Psi}_{\text{GLS}}$ would use the following weights: 8%, 82%, 4%, 3%, 2%, 1%, favoring the high SNR pipeline. In scenario 3, $\widehat{\Psi}_{\text{average}}$ would equally weight all pipelines by 5%, $\widehat{\Psi}_{\text{pool-se}}$ would weight each correlated pipeline by 3.8% and the remaining pipelines by 38% (high SNR pipeline), 2%, 1.3%, 1%, 0.6%, while $\widehat{\Psi}_{\text{GLS}}$ would weight by 0.6% the correlated pipelines and by 81% (high SNR pipeline), 4%, 3%, 2%, 1% the remaining pipelines.

Estimate and bias: No significant bias was found. Under the null, the proportion of pipelines with evidence for an effect ranged between 5% and 7% and was relatively stable with sample size. Under the alternative, the proportion of pipelines increased from 6.3% in scenario 1 to 98% in scenario 1 when the sample size increased from $n = 20$ to $n = 500$. A similar behavior was observed in the other scenarios (up to 63% in scenario 2 and up to 79% in scenario 3).
TODO: add random intercept model

Efficiency: Similar results were obtained for $\beta = 0$ and $\beta = 0.5$, so we will only discuss the former case (Figure 2). In scenario 1, $\widehat{\Psi}_{\text{average}}$ and $\widehat{\Psi}_{\text{pool-se}}$ showed similar empirical variance, whereas $\widehat{\Psi}_{\text{GLS}}$ showed higher empirical variance in very small samples (+159% for $n = 10$, +13% for $n = 25$) and lower empirical variance at larger sample sizes (e.g. -12% for $n = 500$). In scenario 2, $\widehat{\Psi}_{\text{pool-se}}$ and $\widehat{\Psi}_{\text{GLS}}$ showed similar empirical variance (slightly higher for $\widehat{\Psi}_{\text{GLS}}$ in low sample sizes and slightly lower afterwards), both smaller compared to $\widehat{\Psi}_{\text{average}}$, e.g. 10% for $n = 10$ and -24% for $n = 500$ for the GLS estimator. Results in scenario 3 were similar to scenario 1 up to a larger decrease in variance in large samples for GLS estimators (-28% for n=500). $\widehat{\Psi}_{\text{constrained GLS}}$ was similar to $\widehat{\Psi}_{\text{GLS}}$ but with better small sample properties (at most +4% in standard error compared to $\widehat{\Psi}_{\text{average}}$).

Type 1 error: Across all scenarios, the type 1 errors for $\widehat{\Psi}_{\text{average}}$ and $\widehat{\Psi}_{\text{pool-se}}$ were well controlled except in very small samples where small deviations from the nominal level were observed (maximum of 7% for $\widehat{\Psi}_{\text{average}}$ and 8% for $\widehat{\Psi}_{\text{pool-se}}$). When neglecting the uncertainty about the weights, the type 1 error control of $\widehat{\Psi}_{\text{GLS}}$ and $\widehat{\Psi}_{\text{constrained GLS}}$ were only controlled for large samples, i.e. for $n = 500$; large type 1 error rates were found in very small samples (>10%).

4.2 Real-world application

We illustrate the statistical sensitivity analysis with data described in section 2.1 in which the SERT availability was assessed after a drug intervention and compared to normative values. We start by studying the behavior of the four statistical estimators (pooled GLS, pooled constrained GLS, pooled average and pooled SE) to estimate a common effect across pipelines for the given null hypothesis, reported as a forest plot in Figure 3 (left panel). The dashed vertical line in Figure 3 represents the normative value, and all horizontal error bars represent the estimated effect (mean and 95% CI) for a given pipeline and estimator. Across a reasonable set of preprocessing pipelines, 3 of the 8 selected reject the null hypothesis (as indicated by the non-overlapping CI with the normative value) with estimated percent differences between groups ranging between -9% (pipeline 6) and 2.5% (pipeline 1). The pooled constrained GLS, pooled SE and pooled average, all fail to reject a common effect across pipelines, whereas the GLS estimator rejects the null hypothesis. However, when inspecting the pipeline weights for the GLS estimator for this data, it assigned a very high weight to four pipelines (i.e. weight above 1 in absolute values), leading to an unreliable estimate as illustrated by the large standard deviation found in the simulation study for low sample sizes (Figure 2). The constrained GLS estimator did not exhibit this problem and had weights between -0.79 and 1.

Figure 3 (right panel) shows a heatmap for the estimated correlations across preprocessing pipelines, ranging from 0.68 to 1. Pipelines 1-4 show a very high

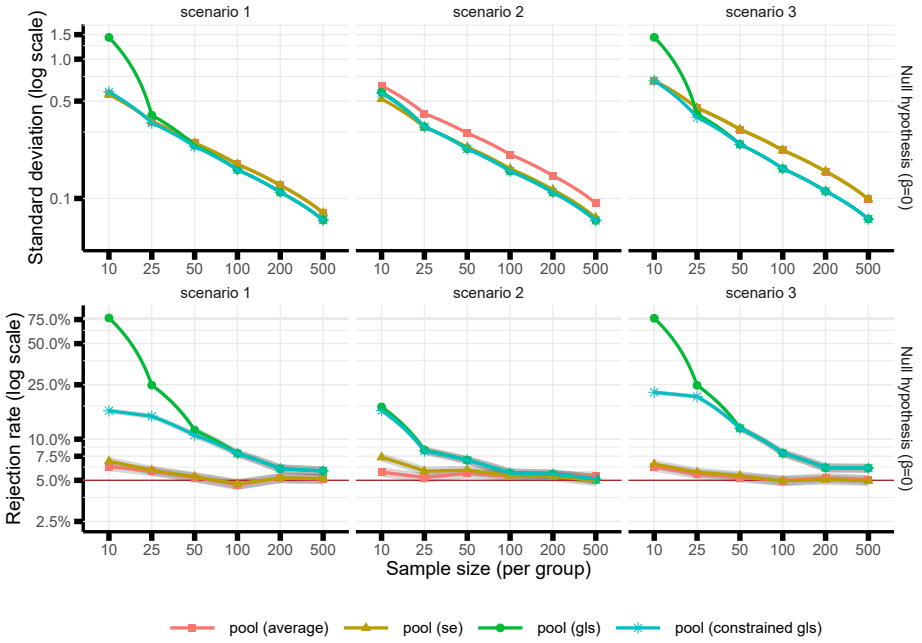

Fig. 2. Upper panel: empirical standard deviation of the estimated common effect for each estimator, scenario, and sample size under the null hypothesis. For large samples the green line (GLS) is covered by the blue line (constrained GLS) in all scenarios. In scenario 1 and 3, the red line (pool-average) is covered by the yellow line (pool se). In scenario 1, the GLS estimators have the same or higher standard deviation as the average for small samples and lower standard deviation (approx. 12% lower) for large samples.

Lower panel: rejection rate under the null hypothesis (i.e. type 1 error). Shaded area represents the Monte Carlo uncertainty.

correlation with each other (only varying the registration choices), whereas pipelines 5, 6 and 8 are equally correlated with each other (different kinetic models), and pipeline 7 is somewhat in between (no motion correction). The heatmap captures important differences in the correlation structure between pipelines, suggesting that not all pipelines perform similarly with moderate levels of unexplained variance. Pipelines 4 to 8 exhibit numerically smaller correlation compared to pipelines 1 to 4 (and similar variance), which explains why GLS estimators produce estimates with lower numerical values compared to pooled estimators ignoring the correlation, as GLS estimators assign more weight to the last four pipelines.

Notably, no correction for multiple comparisons was carried out across regions and pipelines at this point. The rationale for not including this (as should otherwise always be carried out) is that we wanted to make our analysis as com-

parable as possible to the Neuroimaging Analysis Replication and Prediction Study [2], where each participating institution analyzed the data using their own established pipeline and tested only a single region in a hypothesis-driven fashion.

5 Discussion

Looking first at the simulated data results, we observe that, asymptotically, $\widehat{\Psi}_{\text{GLS}}$ is the best estimator, leading to more precise estimates than $\widehat{\Psi}_{\text{average}}$ or $\widehat{\Psi}_{\text{pool-se}}$. It is, however, not suited for small sample sizes or large numbers of pipelines, as it exhibits large variance. We proposed an alternative estimator $\widehat{\Psi}_{\text{constrained GLS}}$ which can be seen as regularizing the GLS estimator toward the empirical average in small samples. Even though this considerably improves

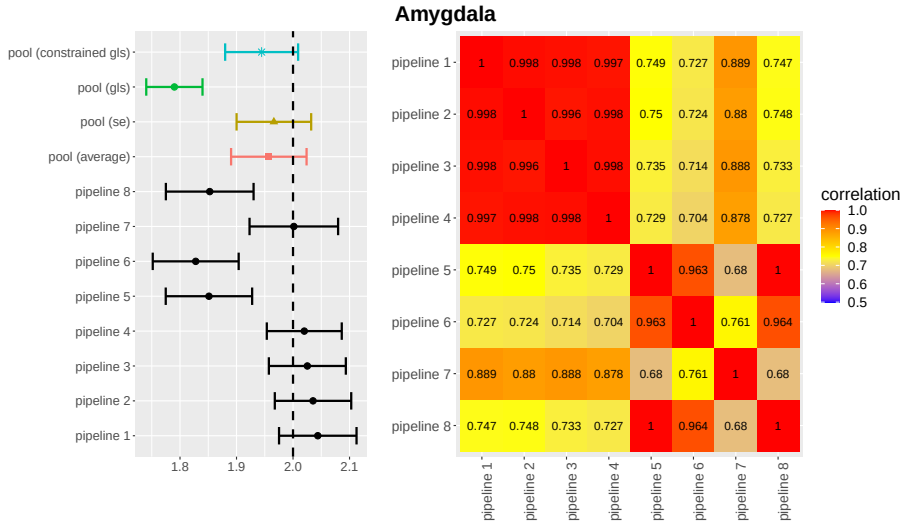

Fig. 3. Left panel: forest plot of the estimated SERT availability in the amygdala for the intervention group (point and full line) vs. the normative values (dashed line) for each pipeline and the four proposed pooled estimators.

Right panel: correlation of the estimated SERT availability between pipelines.
Pipeline 1: with motion correction (MC), boundary based registration (BBR) using the time-weighted average PET image (twa), and MRTM2 as kinetic modeling choice.
Pipeline 2: MC, normalized mutual information registration (NMI) using twa, MRTM2.
Pipeline 3: MC, BBR using the average PET image, and MRTM2.
Pipeline 4: MC, NMI using the average PET image, and MRTM2.
Pipeline 5: MC, BBR_twa, and MRTM.
Pipeline 6: no motion correction (nMC), BBR_twa, and MRTM.
Pipeline 7: nMC, BBR_twa, and MRTM2.
Pipeline 8: MC, BBR_twa, and SRTM.

the small sample performance of the estimator, other regularization approaches may lead to further gain. An alternative approach could be to regularize the estimated variance-covariance matrix between pipeline specific estimates, e.g. using graphical lasso. However, this is challenging since the correlation structure among pipelines is typically complex and non-sparse. The proposed method to quantify the uncertainty of $\widehat{\Psi}_{\text{constrained GLS}}$ is numerically fast, but unreliable in small samples or with a large number of pipelines. However, these latter points are beyond the scope of this paper, and is left for future work. In the real-world application, we observed that all estimators could be readily applied, and three of them performed as expected based on the results from individual pipelines. Only the $\widehat{\Psi}_{\text{GLS}}$ estimator performed differently and was the only one rejecting the null hypothesis hinting at an effect across pipelines. This is though, due to the estimator not being able to fit the weights properly in small sample sizes. Since PET neuroimaging studies are rarely beyond sample sizes of n>50, other estimators should be used. We recommend using $\widehat{\Psi}_{\text{pool-se}}$ or $\widehat{\Psi}_{\text{constrained GLS}}$: the latter when pipelines are not similarly related and the sample size is moderate to large, otherwise the former.

The framework that we are proposing is not without limitations. The underlying assumption of combining results across pipelines in our analysis is that all pipelines are unbiased. This can however not always be guaranteed. Alternative approaches (e.g. STAPLE [14]) could be used to reduce the bias of the pooled estimators by assuming a majority of unbiased pipelines or identifying clusters of pipelines and pooling pipeline-specific estimates within clusters.

6 Conclusion

In this work, we have developed a statistical sensitivity analysis that can quantify the impact of different preprocessing choices on subsequent statistical analyses. As has been reported in previous studies, we observe that the influence of preprocessing pipelines on subsequent statistical analysis can be quite large. Hence, we provide tools for statistical analyses that can aggregate multiple analyses of the same data. We introduce four statistical estimators, $\widehat{\Psi}_{\text{average}}$, $\widehat{\Psi}_{\text{pool-se}}$, $\widehat{\Psi}_{\text{GLS}}$, and $\widehat{\Psi}_{\text{constrained GLS}}$ to combine the pipeline specific estimates. This enables testing hypotheses across pipelines, such as "no effect across all pipelines" or "at least one pipeline with no effect". The proposed framework is generic and can be applied to any imaging modality.

Bibliography

[1] Beliveau, V., et al.: A high-resolution in vivo atlas of the human brain's serotonin system. Journal of Neuroscience **37**(1), 120–128 (2017)
[2] Botvinik-Nezer, R., et al.: Variability in the analysis of a single neuroimaging dataset by many teams. Nature **582**(7810), 84–88 (2020)

[3] Carp, J.: On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. Frontiers in neuroscience **6**, 149 (2012)
[4] Churchill, N.W., et al.: An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional MRI. PLoS ONE **10**(7), 1–25 (2015)
[5] Cox, M., Eiø, C., Mana, G., Pennecchi, F.: The generalized weighted mean of correlated quantities. Metrologia **43**(4), S268 (2006)
[6] Frokjaer, V.G., et al.: Role of serotonin transporter changes in depressive responses to sex-steroid hormone manipulation: a positron emission tomography study. Biological psychiatry **78**(8), 534–543 (2015)
[7] Kennedy, E.H.: Semiparametric theory. arXiv:1709.06418 (2017)
[8] Nørgaard, M., , et al.: Optimization of preprocessing strategies in positron emission tomography (PET) neuroimaging: a [11C] DASB PET study. Neuroimage **199**, 466–479 (2019)
[9] Nørgaard, M., Ozenne, B., Svarer, C., Frokjaer, V.G., Schain, M., Strother, S.C., Ganz, M.: Preprocessing, prediction and significance: Framework and application to brain imaging. International Conference on Medical Image Computing and Computer-Assisted Intervention pp. 196–204 (2019)
[10] Poldrack, R.A., et al.: Establishment of Best Practices for Evidence for Prediction: A Review. JAMA Psychiatry **77**(5), 534–540 (2020)
[11] Stenbæk, D.S., Fisher, P.M., Ozenne, B., Andersen, E., Hjordt, L.V., McMahon, B., Hasselbalch, S.G., Frokjaer, V.G., Knudsen, G.M.: Brain serotonin 4 receptor binding is inversely associated with verbal memory recall. Brain and behavior **7**(4), e00674 (2017)
[12] Tsiatis, A.A.: Semiparametric theory and missing data. Springer (2006)
[13] Van der Vaart, A.W.: Asymptotic statistics, vol. 3. Cambridge university press (2000)
[14] Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**(7), 903–921 (2004). https://doi.org/10.1109/TMI.2004.828354

Appendix A Influence function

In the univariate case (i.e. $R = P = 1$) without covariate ($C = 0$), the effect of interest ψ may be the Pearson's correlation coefficient:

$$\psi = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - E[X]^2}\sqrt{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2}}$$

when X is continuous or the mean difference when X is binary.

$$\psi = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0]$$

Here $\mathbb{E}[.]$ denotes the expectation and $\mathbb{E}[.|.]$ the conditional expectation. The later case leads to simple expression: $\widehat{\psi}_j$ is the empirical mean difference between the groups

$$\widehat{\psi}_j = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{ij}x_i}{\pi} - \frac{y_{ij}(1-x_i)}{1-\pi}\right)$$

where π denotes the proportion of observations with $X = 1$, y_{ij} the brain signal for individual i processed by pipeline j, and x_i the exposure value for individual i. The previous expression is equivalent to:

$$\sqrt{n}\left(\widehat{\psi}_j - \psi_j\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{\widehat{\psi}_j}(\mathcal{O}_i)$$

where $\varphi_{\widehat{\psi}_j}(\mathcal{O}_i) = \frac{y_{ij}x_i}{\pi} - \frac{y_{ij}(1-x_i)}{1-\pi} - \psi_j$, $\mathcal{O}_j = (y_{i1}, \ldots, y_{iJ}, x_i)$, and ψ_j is the large sample value of $\widehat{\psi}_j$. Denoting Y_j the random variable representing the brain signal processed by pipeline j, the estimator can be seen as the empirical average of independent realizations of a new random variable $\frac{Y_jX}{\pi} - \frac{Y_j(1-X)}{1-\pi} - \psi_j$, called the influence function of $\widehat{\psi}_j$. Thus from the multivariate central limit theorem we get that the joint distribution of the estimates is asymptotically multivariate normal. It has mean ψ (under our assumption that $\forall j \in \{1, \ldots, J\}, \psi = \psi_j$) and its variance-covariance, denoted $\Sigma_{\widehat{\boldsymbol{\psi}}}$, is the same as the one of $\boldsymbol{\varphi}_{\widehat{\boldsymbol{\psi}}} = \left(\varphi_{\widehat{\psi}_1}, \ldots, \varphi_{\widehat{\psi}_J}\right)$ divided by n. Because $\boldsymbol{\varphi}_{\widehat{\boldsymbol{\psi}}}$ involves some unknown parameters like π and ψ_j we do not observe it and cannot directly estimate $\Sigma_{\widehat{\boldsymbol{\psi}}}$. However, by plugging our estimates of these unknown parameters we can approximate $\varphi_{\widehat{\psi}_j}$ as $\widehat{\varphi}_{\widehat{\psi}_j}(\mathcal{O}_{ij}) = \frac{y_{ij}x_i}{\frac{1}{n}\sum_{i=1}^{n}x_i} - \frac{y_{ij}(1-x_i)}{1-\frac{1}{n}\sum_{i=1}^{n}x_i} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{ij}x_i}{\pi} - \frac{y_{ij}(1-x_i)}{1-\pi}\right)$ and approximate $\Sigma_{\widehat{\boldsymbol{\psi}}}$.

In a more general case, we would define a statistical model $\mathcal{M}(\Theta)$ relating $\boldsymbol{X}$ and $\boldsymbol{Y}$ via a parameter ψ. ψ may be an element of Θ, the set of model parameters, or a function of elements of Θ. For instance one could use a latent variable model (LVM) with two latent variables, one summarizing the brain measurements and another summarizing the exposure variables. ψ is then the

coefficient relating the two latent variables. See Figure 1 of [11] for a graphical representation of a LVM - in this example the latent variable "LVu" represents the PET measurement and the latent variable "LVpos" the memory relative to positive word. In this more general case the previous decomposition does not hold exactly but up to a residual term $o_p(1)$ which converges to 0 as the sample approaches infinity:

$$\sqrt{n}\left(\widehat{\psi}_j - \psi_j\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{\widehat{\psi}_j}(\mathcal{O}_i) + o_p(1)$$

as indicated in the main text of this article. This decomposition exists for any estimator $\widehat{\psi}$ derived from an M-estimator [13] (section 5.3), including likelihood-based estimators. Denote by $\widehat{\theta}_j$ the ML estimator and $\widehat{\Psi}_j = c^\intercal\widehat{\theta}_j$ the parameter of interest (c may be a vector starting by 1 and followed by 0's, i.e. selects the first element of $\widehat{\theta}_j$. The corresponding influence function only involves the first two derivatives of the log-likelihood (formula 3.6 in [12]):

$$\varphi_{\widehat{\psi}_j}(\mathcal{O}_i) = -c^\intercal\mathbb{E}\left[\frac{\partial\mathcal{S}_j(\mathcal{O}_i,\theta_j)}{\partial\theta_j}\right]^{-1}\mathcal{S}_j(\mathcal{O}_i,\theta_j)$$

with $\mathcal{S}_j(\mathcal{O}_i,\theta_j)$ being the score for individual i when considering pipeline j, i.e. vector containing the first derivatives of the log-likelihood contribution of individual i. Once the influence function has been estimated for each individual, we can use it to obtain a consistent estimator of $\Sigma_{\widehat{\boldsymbol{\psi}}}$:

$$\widehat{\Sigma}_{\widehat{\boldsymbol{\psi}}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}_{\widehat{\boldsymbol{\psi}}}(\mathcal{O}_i)^\intercal\boldsymbol{\varphi}_{\widehat{\boldsymbol{\psi}}}(\mathcal{O}_i)$$

where $\boldsymbol{\varphi}_{\widehat{\boldsymbol{\psi}}}(\mathcal{O}_i)^\intercal$ denotes the transpose of the J-dimensional vector of influence functions relative to individual i. Practically speaking the above allows us assess the variance-covariance matrix of effects across pipelines, e.g., if pipelines are completely independent in terms of their estimated effect, then $\widehat{\Sigma}_{\widehat{\boldsymbol{\psi}}}$ would be a diagonal matrix with the uncertainty of the estimated effect per pipeline in the diagonal. However, if the estimated effects across pipelines are correlated, the matrix would not be sparse.

Appendix B Integration of the Gaussian density - bivariate case

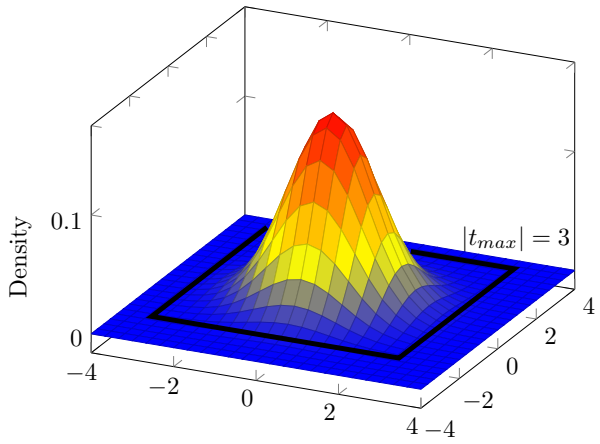

Fig. A. Density of the two dimensional standard normal distribution (colored surface). The black line delimits the domain $\mathcal{D}(\widehat{\boldsymbol{t}}_{\max})$ where $J = 2$, $\widehat{t}_1 = 1.5$, $\widehat{t}_2 = 3$, and $\Sigma_{\widehat{\psi}}$ is the identity matrix with two rows and two columns. The area under the blue surface external to the black line corresponds to the p-value relative to the first test, adjusted for two tests.

Appendix C Reformulation of the GLS estimator

C.1 Via a spectral decomposition

$\widehat{\Sigma}_{\widehat{\psi}}$ being a symmetric semi-positive definite matrix, it admits the decomposition $\widehat{\Sigma}_{\widehat{\psi}} = QDQ^\intercal$ where Q is an orthogonal matrix (i.e. $QQ^\intercal = I_J$, the identity matrix) and D is a diagonal matrix with non negative values $(\lambda_1, \ldots, \lambda_J)$. Thus:

$$\widehat{\boldsymbol{\psi}}_{\text{GLS}} = \left(\mathbf{1}^\intercal \widehat{\Sigma}_{\widehat{\Psi}}^{-1} \mathbf{1}\right)^{-1} \mathbf{1}^\intercal \widehat{\Sigma}_{\widehat{\Psi}}^{-1} \widehat{\boldsymbol{\psi}} = \left(\mathbf{1}^\intercal Q D^{-1} Q^\intercal \mathbf{1}\right)^{-1} \mathbf{1}^\intercal Q D^{-1} Q^\intercal \widehat{\boldsymbol{\psi}}$$

We first note that $\bar{Q} = \mathbf{1}^\intercal Q$ is a vector with element q equal to the sum (column-wise) of the eigenvectors. Then:

$$\mathbf{1}^\intercal Q D^{-1} Q^\intercal \mathbf{1} = \sum_{j=1}^{J} \bar{q}_j \lambda^{-1} \bar{q}_j = \sum_{j=1}^{J} w_j$$

Where $w_j = {\bar{q}_j}^{-2} / \lambda_j$. Moreover $\mathbf{1}^\intercal Q D^{-1} Q^\intercal$ is a vector with k-th elements $\sum_{j=1}^{J} \bar{q}_j \lambda_j^{-1} q_{kj} = \sum_{j=1}^{J} w_j q^*_{kj}$ where $q^*_{kj} = q_{kj}/\bar{q}_j$

Therefore

$$\widehat{\Psi}_{\text{GLS}} = \frac{1}{\sum_{j=1}^{J} w_j} \sum_{j=1}^{J} w_j {q_j^*}^\intercal \widehat{\boldsymbol{\psi}}$$

C.2 Via joint modeling

Consider the simple case of a single continuous brain measurement ($R = 1$), a single binary exposure ($P = 1$) with equal probability of being 0 and 1, no covariate ($C = 0$), and no missing value. We can use a joint linear model:

$$Y_{ij} = \alpha_j + \beta X_i + \varepsilon_{ij} \text{ where } (\varepsilon_{i1}, \ldots, \varepsilon_{iJ}) \sim \mathcal{N}(0, \Sigma_\varepsilon)$$

Denote by $\widehat{\boldsymbol{\mu}}_g = (\widehat{\mu}_{g1}, \ldots, \widehat{\mu}_{gJ})$ the vector empirical mean in each group (i.e. one relative to $X = 1$ another to $X = 0$) of sample size $\frac{n}{2}$. Since the mean and variance are sufficient statistics in Gaussian models, the joint linear model is equivalent to:

$$\widehat{\mu}_{gj} = \alpha_j + \beta X_g + e_{gj} \text{ where } (e_{g1}, \ldots, e_{gJ}) \sim \mathcal{N}(0, 2\Sigma_\varepsilon/n)$$

Denote by $\Delta\widehat{\boldsymbol{\mu}} = (\Delta\widehat{\mu}_1, \ldots, \Delta\widehat{\mu}_J)$ the vector of difference in mean, the previous model implies:

$$\Delta\widehat{\mu}_j = \beta + \epsilon_j \text{ where } (\epsilon_1, \ldots, \epsilon_J) \sim \mathcal{N}(0, 4\Sigma_\varepsilon/n)$$

whose maximum likelihood solution is $\widehat{\beta} = \left(\mathbf{1}^\intercal \widehat{\Sigma}_\epsilon^{-1} \mathbf{1}\right)^{-1} \mathbf{1}^\intercal \widehat{\Sigma}_\epsilon^{-1} \Delta\widehat{\boldsymbol{\mu}}$.

Now consider the GLS estimator of the common exposure effect $\widehat{\boldsymbol{\psi}}_{\text{GLS}} = \left(\mathbf{1}^\intercal \widehat{\Sigma}_{\widehat{\Psi}}^{-1} \mathbf{1}\right)^{-1} \mathbf{1}^\intercal \widehat{\Sigma}_{\widehat{\Psi}}^{-1} \widehat{\boldsymbol{\psi}}$ based on the pipeline specific Ordinary Least Squares (OLS) estimators of the exposure effect $\widehat{\boldsymbol{\psi}} = \left(\widehat{\psi}_1, \ldots, \widehat{\psi}_J\right)$. Denote by $\Theta_j = (\alpha_j, \psi_j)$ the mean parameters of each pipeline specific model and σ_j^2 the residual variance parameter. We have:

- $\widehat{\boldsymbol{\psi}} = (X^\intercal X)^{-1} X^\intercal Y_{\cdot j}$ where $Y_{\cdot j}$ denotes the brain measurements across individuals relative to the j-th pipeline.
- $\widehat{\Sigma}_{\widehat{\Psi}}$ has elements $\sum_{i=1}^n (Y_{ij} - X_i\Theta_j)^\intercal (Y_{ij} - X_i\Theta_j) \left(X_i (X^\intercal X)^{-1}\right)^2 c^\intercal$ where $c = (0, 1)$. This follows from the fact the the score relative to $\widehat{\psi}_j$ is $\frac{1}{\sigma_j^2} X^\intercal (Y_j - X\Theta_j)\, c$ and the variance covariance $\sigma_j^2 (X^\intercal X)^{-1}$.

With a single binary covariate (and an intercept), $X^\intercal X = \begin{bmatrix} n & n/2 \\ n/2 & n \end{bmatrix}$ whose inverse is $\begin{bmatrix} 2/n & -2/n \\ -2/n & 4/n \end{bmatrix}$ Hence the second element of $X_i(X^\intercal X)^{-1}$ is either $-2/n$ or $2/n$ so $\left(X_i(X^\intercal X)^{-1}\right) c^\intercal = 4/n$. Therefore, assuming that the observations are sorted by group:

$$\widehat{\psi}_j = \frac{n}{2} \sum_{i=1}^{n/2} Y_{ij} - \frac{n}{2} \sum_{i=n/2+1}^{n} Y_{ij} = \Delta \widehat{\mu}_j$$

$$\widehat{\Sigma}_{\widehat{\psi}} = \frac{4}{n^2} \sum_{i=1}^{n} (Y_{i\cdot} - X\Theta)^\intercal (Y_{i\cdot} - X\Theta) = \frac{4}{n} \widehat{\Sigma}_\varepsilon$$

So $\widehat{\boldsymbol{\psi}}_{\text{GLS}} = \left(\mathbf{1}^\intercal \widehat{\Sigma}_\varepsilon^{-1} \mathbf{1}\right)^{-1} \mathbf{1}^\intercal \widehat{\Sigma}_\varepsilon^{-1} \Delta \widehat{\boldsymbol{\mu}}$. If the common effect models holds (i.e. $\widehat{\psi}_1 = \ldots = \widehat{\psi}_J = \beta$) then $\widehat{\Sigma}_\varepsilon$ is an unbiased estimate of Σ_ε so $\widehat{\boldsymbol{\psi}}_{\text{GLS}}$ is asymptotically equivalent to the ML estimator $\widehat{\beta}$ and thus asymptotically efficient.