Same Data Meta-Analysis Notes

Thomas E. Nichols, &co

May 22, 2023

Abstract

We propose different methods to analyze multiverse analyses, where multiple sets of results are obtained from running different pipelines on the same original data. The consensus method proposed in the NARPS paper is described and its limitations discussed. A more general framework that isolates pipeline bias, inter-pipeline correlation, and consistent signal is proposed.

Contents

1 Introduction

The purpose of this work is to present an approach to analyzing multiple sets of results produced by a multiverse analyses, where the same input data is analyzed under different pipelines. The resulting images would *ideally* be identical, but do differ but are dependent over pipelines (a violation of the independence assumption in a conventional

meta-analysis). Here we propose a unified framework where random sampling variation and inter-pipeline variation is integrated into a single model.

We identify 3 different types of multiverse outputs: Test statistics, parameter estimates paired with estimated standard deviation (standard errors) and parameter estimates only. For short we call these cases "Z" (though other statistics maybe produced), "PE+SE" and "PE". We anticipate that the primary application for multiverse analyses will be group models, where both Z and PE+SE will be available for any data type. Single subject analyses may produce PE, for structural MRI analyses, like with VBM, or with diffusion MRI. In the remainder of this work we only consider Z and PE+SE.

We identify two distinct goals with multiverse analyses. One is inferential, in trying to synthesize inferences made on different pipelines into a single inference. Another is qualitative, simply understanding the nature of multiverse structure, e.g. measuring the degree of disagreement between pipelines, or understanding patterns of similarity between different pipelines.

In the work below we propose a variety of models for Z and PE+SE data, propose how to fit them and how to extract meaningful insights. We will propose several different approaches to inference for Z data, and different ways of characterizing multiverse variation with PE+SE data.

2 Overview & Motivation

To set up the problem and motivate initial and further work, we first consider the specific case of test statistic images and few simple solutions and their limitations. To fix notation, let Y_{kj} be the test statistic for pipeline $k = 1, ..., K$ and voxel $j = 1, ..., J$, arranged into a $K \times J$ matrix $\mathbf{Y}$, with Y_k being the length-J row vector of voxels for pipeline k, and Y_j being the length-K column vector of different pipeline results.

2.1 Intuitive Solutions

For this simple model it is useful to explore how some simple solutions would perform. The most obvious approach to take would be to create an average map with value at

voxel j

$$\bar{Y}_j = \frac{1}{K} \sum_k Y_{kj}.$$

This would preserve the mean at each voxel and but would have reduced variance, as per

$$\mathrm{Var}(\bar{Y}_j) = \mathbf{1}_K^\top \mathbf{Q} \mathbf{1}_K / K^2,$$

where $\mathbf{1}_K$ is a column K-vector of ones (when the dimension is clear from context we will suppress the subscript on $\mathbf{1}$ going forward) and $\mathbf{Q}$ is the inter-pipeline correlation. If the K studies were independent, $\mathbf{Q} = \mathbf{I}$, and then the variance is $1/K$, showing the usual precision gain from independent sampling. If the K studies were identical, and $\mathbf{Q} = \mathbf{J}$ is a matrix of all 1's, then the variance is 1, intuitively showing *no* information gain as all inputs are equal. In general, $\mathbf{1}^\top \mathbf{Q} \mathbf{1}/K^2 = \frac{1}{K} + \frac{K-1}{K}\bar{q}$, where $\bar{q}$ is the average of all of the $K(K-1)/2$ possible correlations, clearly showing how correlation inflates the variance relative to the $1/K$ independence case. (Note that $\mathbf{1}^\top \mathbf{Q} \mathbf{1}/K^2 \leq 1$, meaning the average never has variance greater than the data, even when there are negative correlations; see Appendix B.1).

Hence, the simple approach of averaging statistic maps is unsatisfying because the resulting summary will have attenuated variance depending on inter-pipeline correlation $\mathbf{Q}$. We could address that, and in the spirit of the Stouffer's method (Stouffer et al., 1949) standardise by this attenuated variance:

$$\frac{\bar{Y}_j}{\sqrt{\mathrm{Var}(\bar{Y}_j)}} = \frac{\bar{Y}_j}{\sqrt{\mathbf{1}^\top \mathbf{Q} \mathbf{1}/K^2}},$$

However, while this approach standardises the variance it will distort the mean. For example, if the input maps had a average value of μ, the resulting consensus map will be $\mu / \sqrt{\mathbf{1}^\top \mathbf{Q} \mathbf{1}/K^2} \geq \mu$, exceeding μ depending K on the degree of variance attenuation. It was the shortcomings of these two approaches that motivated the NARPS consensus method.

3 Connecting Conventional Meta-Analysis to Same-Data Meta-Analysis

Here we consider the parallels between conventional meta-analysis in fixed and random effect forms to motivate different variants of same-data meta-analysis.

3.1 Conventional Meta-Analysis Models & Methods

In a conventional meta-analysis we have $J = 1$ outcome recorded from K studies, and so we suppress the j subscript for now.

3.1.1 Stouffer's Combining

The most rudimentary meta-analysis method is Stouffer's combining method, which assumes that Y_k, over $k = 1, ..., K$ different studies, are independent z-score test statistics,

$$Y \sim \mathcal{N}(\mu \mathbf{1}, \mathbf{I}),$$

where μ is a scalar and is the average non-centrality parameter (this assumes large degrees of freedom in each study; in practice a bias correction is recommended for small samples (Bossier et al., 2019; Hedges, 1981)). Stouffer's combining statistic is

$$\bar{Y}\sqrt{K},$$

where $\bar{Y} = \frac{1}{K}\sum_k Y_k$, is again a z-score and has mean zero and variance one under the null $H_0 : \mu = 0$, and magnified mean $\mu\sqrt{K}$ under the alternative.

Note that variants of Stouffer's method have been proposed that allow for arbitrary positive weights, $w_1, ..., w_K$, on each study (Zaykin, 2011). The assumed model is the same, but takes the form

$$\frac{\sum_{k=1}^K w_k Y_k}{\sqrt{\sum_{k=1}^K w_k^2}}$$

which is again a z-score, an idea we'll use in the NARPS consensus method. Note the weights are arbitrary up to a scale factor, though it is useful to constrain them so that $\sum_{k=1}^K w_k = K$ so that the mean is preserved, $\mathrm{E}(\frac{1}{K}\sum_{k=1}^K w_k Y_k) = \mu$.

Stouffer's method, however, is not considered good practice as it ignoring both sample size and any standard errors reported with each outcome.

3.1.2 Linear Mixed Effects Meta-Analysis

The linear mixed effects (LME) meta-analysis models outcome Y_k and reported standard errors σ_k, and additionally allows for a between-study variance that isn't captured by σ_k^2,

$$Y \sim \mathcal{N}(\mu \mathbf{1}, \mathbf{I}\tau^2 + \mathbf{D}),$$

where τ^2 is the inter-study variance and $\mathbf{D} = \text{diag}(\sigma_1^2, \ldots, \sigma_K^2)$. Restricting $\tau^2 = 0$ gives a fixed effects model, though current practice is to use a meta-analysis model by default.

3.2 Univariate Same-Data Meta-Analysis Models

Now we consider how these models might work for a same-data meta-analysis.

3.2.1 1 voxel SDMA Stouffer's Combining

For a Stouffer's like method, we could continue to accept that the inputs are z-scores, but we must accommodate an inter-pipeline correlation $\mathbf{Q}$, suggesting a model

$$Y \sim \mathcal{N}(\mu \mathbf{1}, \mathbf{Q})$$

and a SDMA Stouffer's z-score of

$$\frac{\bar{Y}}{\sqrt{\mathbf{1}^\top \mathbf{Q} \mathbf{1}/K^2}},$$

which again preserves the unit variance but will magnify the mean under the alternative. However, the practical problem remains how to estimate the $K \times K$ $\mathbf{Q}$ given just K observations.

For a weighted variant of Stouffer's in the SDMA setting, let $\mathbf{W} = \text{diag}(w_1^2, ..., w_K^2)$, then the weighted data takes the form

$$\mathbf{W}^{1/2}\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{W}^{1/2}\mathbf{1}, \mathbf{W}^{1/2}\mathbf{Q}\mathbf{W}^{1/2}).$$

If we write the weighted mean image as $\bar{Y}_{\text{w}} = \mathbf{1}^\top \mathbf{W}^{1/2}\mathbf{Y}/K$, then the weighted Stouffer's with dependence has a combining statistic of

$$\frac{\bar{Y}_{\text{w}}}{\sqrt{\mathbf{1}^\top \mathbf{W}^{1/2}\mathbf{Q}\mathbf{W}^{1/2}\mathbf{1}/K^2}}.$$

If it where the case that $\text{Var}(Y_j) = \mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}$, then choosing $\mathbf{W} = \mathbf{D}^{-1}$, would cause the weights and the variance to cancel, giving a Souffer's z-score of

$$\frac{\bar{Y}_{\text{w}}}{\sqrt{\mathbf{1}^\top\mathbf{Q}\mathbf{1}/K^2}}.$$

While it will have no impact on the final z-score, if we wanted preserve the units of the weighted average $\bar{Y}_{\text{w}}$, we should set $\mathbf{W} = c\mathbf{D}^{-1}$, where $c = (\frac{1}{K}\mathbf{1}^\top\mathbf{D}^{-1/2}\mathbf{1})^{-2} = (\frac{1}{K}\sum_k 1/\sigma_k)^{-2}$; this preserves the average mean:

$$\frac{1}{K}\mathbf{1}^\top(\mu\mathbf{W}^{1/2}\mathbf{1}) = \mu(\mathbf{1}^\top\mathbf{D}^{-1/2}\mathbf{1})^{-1}\mathbf{1}^\top\mathbf{D}^{-1/2}\mathbf{1} = \mu.$$

3.2.2 1 voxel SMDA Linear Mixed Effects Meta-Analysis

For a SMDA linear mixed effects meta-analysis model we propose it should take the form

$$Y \sim \mathcal{N}(\mu\mathbf{1}, \mathbf{Q}_{\text{II}}\tau^2 + \mathbf{D}^{1/2}\mathbf{Q}_{\text{I}}\mathbf{D}^{1/2}),$$

where $\mathbf{Q}_{\text{I}}$ is the inter-pipeline dependence that scales with the standard errors produced by each pipeline, while τ^2 is the excess variance that cannot be explained by $\mathbf{D}^{1/2}\mathbf{Q}_{\text{I}}\mathbf{D}^{1/2}$, that scales a distinct correlation $\mathbf{Q}_{\text{II}}$.

A practical simplification would be to assume inter-pipeline correlations are the same, $\mathbf{Q}_{\text{I}} = \mathbf{Q}_{\text{II}}$. Also, if these models were every deployed with subject-level data (e.g. for structural or diffusion MRI output) where no variance estimates σ_k^2 are available, single covariance term $\mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}$ would be used where both $\text{diag}(\mathbf{D})$ and $\mathbf{Q}$ would have to be estimated.

Yet in this univariate setting with only K observations, with either one or two correlation matrices each with $K(K-1)/2$ parameters, this model is hopelessly over-parameterised.

3.3 Multivariate Same-Data Meta-Analysis Models

The essential ingredient of SMDA is the use of images, with J outcomes. A matrix normal is a very convenient way to express a multivariate distribution for data organized as a matrix.

3.3.1 SDMA Stouffer's Combining

A Stouffer's-like method that takes z-scores for inputs would have a SDMA model of

$$\mathbf{Y} \sim \mathcal{MN}(\mathbf{1}\mu^\top, \mathbf{Q}, \mathbf{\Sigma})$$

where μ is now a length J vector of voxel-wise means, and $\mathbf{\Sigma}$ is the $J \times J$ spatial correlation matrix. To be absolutely explicit, this matrix normal specification implies that for data from pipeline k, voxel j, and pipeline k', voxel j', their joint distribution is

$$\begin{bmatrix} Y_{kj} \\ Y_{k'j'} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_j \\ \mu_{j'} \end{bmatrix}, \begin{bmatrix} \mathbf{Q}_{kk}\mathbf{\Sigma}_{jj} & \mathbf{Q}_{kk'}\mathbf{\Sigma}_{jj'} \\ \mathbf{Q}_{k'k}\mathbf{\Sigma}_{j'j} & \mathbf{Q}_{k'k'}\mathbf{\Sigma}_{j'j'} \end{bmatrix} \right).$$

Note that the variances are a product of diagonals of $\mathbf{Q}$ and $\mathbf{\Sigma}$; while here we are assuming unit variance, usually one chooses one term to be a correlation matrix and lets the other capture the variance. Also, in general we can never hope to estimate or really do anything with voxels by voxels $\mathbf{\Sigma}$, but it is important to represent it.

This basic model could be elaborated to allow heterogeneous mean or variance over studies, e.g.

$$\mathbf{Y} \sim \mathcal{MN}(\delta\mathbf{1}_J^\top + \mathbf{1}_K\mu^\top, \mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}, \mathbf{\Sigma})$$

where δ is a K-vector of pipeline biases about the overall mean μ and $\mathbf{D}$ is a diagonal matrix of pipeline variances; to be identifiable we constrain δ to sum to zero, $\delta^\top\mathbf{1} = 0$.

3.3.2 SMDA Linear Mixed Effects Meta-Analysis

The general mixed effects meta-analysis model could take the form

$$\mathbf{Y} \sim \mathcal{MN}(\delta\mathbf{1}_J^\top + \mathbf{1}_K\mu^\top, \mathbf{Q}_{\text{II}}\tau^2 + \mathbf{D}^{1/2}\mathbf{Q}_{\text{I}}\mathbf{D}^{1/2}, \mathbf{\Sigma}),$$

where μ is now a length J vector of voxel-wise means, however this is unrealistic in one key aspect: It assumes the variance is the same over the whole brain. We could address this by replacing $\mathbf{\Sigma}$ with, say, $\mathbf{V}^{1/2}\mathbf{\Sigma}\mathbf{V}^{1/2}$, $\mathbf{V}$ is a diagonal matrix of voxel-wise variance scaling factors. That is, $(\mathbf{V})_{jj}$ is not the variance at voxel j, but the scaling factor of the (global) pipeline variance terms τ^2 and $(\mathbf{D})_{kk}$. Like the previous Stouffer's model, this could also be extended with a pipeline bias term, giving matrix normal mean of $\delta\mathbf{1}_J^\top + \mathbf{1}_K\mu^\top$.

Alternatively, an elaboration would be to leave the separable matrix normal model and simply assert that each voxel j requires a full random effects meta-analysis model:

$$Y_j \sim \mathcal{N}(\mu_j \mathbf{1}_K, \mathbf{Q}_{\text{II},j}\tau_j^2 + \mathbf{D}_j^{1/2}\mathbf{Q}_{\text{I},j}\mathbf{D}_j^{1/2}).$$

As with the univariate model, this is overparameterised and impractical without other constraints.

In the remainder of the document we propose methods to find practical implementations of versions of these models.

4 Multiverse Analysis of Test Statistics: NARPS Consensus Analysis

The starting point for this work was a method developed for (Botvinik-Nezer et al., 2020), the output of the Neuroimaging Analysis Replication and Prediction Study (NARPS) study.

4.1 NARPS Consensus Analysis Objective

The goals of the NARPS consensus same-data meta-analysis method were to: (1) Generate a map such that, if each input map were mean zero and variance one, the output map would also be mean zero and variance one, and (2) Produce a map that was as similar as possible to the input maps.

4.1.1 Unweighted Stouffer's

We first describe a simplified Stouffer's combining method before giving the actual method used in the NARPS paper.

While several variants of Stouffer's model are described Section 3.3.1, the actual model used is slightly different. Let's start with a slightly simplified version

$$\mathbf{Y} \sim \mathcal{MN}(\mu \mathbf{1}_K \mathbf{1}_J^\top, \mathbf{Q}, \mathbf{\Sigma})$$

where μ is a scalar. This marks the (inadvertent) decision we made in NARPS to *not* model a spatially heterogeneous mean. Under a complete null of mean zero for every

pipeline, every voxel, this has no impact; but if there is a non-zero signal shared across pipelines, this will contribute to the inter-pipeline correlation $\mathbf{Q}$.

The conventional Stouffer's combining statistic at voxel j would be the z-score

$$\frac{\bar{Y}_j}{\sqrt{\mathbf{1}^\top \mathbf{Q}\mathbf{1}/K^2}} \sim \mathcal{N}\left(\frac{\mu}{\sqrt{\mathbf{1}^\top \mathbf{Q}\mathbf{1}/K^2}}, 1\right).$$

To 'calibrate' this and make a consensus method, we can give this image an arbitrary mean and variance, say μ_C and σ_C, producing a final consensus statistic

$$\left(\frac{\bar{Y}_j}{\sqrt{\mathbf{1}^\top \mathbf{Q}\mathbf{1}/K^2}} - \frac{\hat{\mu}}{\sqrt{\mathbf{1}^\top \mathbf{Q}\mathbf{1}/K^2}}\right)\sigma_C + \mu_C.$$

For now, don't specify μ_C and σ_C further, but just indicate that we can set the desired mean and variance. To estimate μ, the natural estimator is the image-wise mean of the mean image $\hat{\mu} = \frac{1}{J}\sum_j \bar{Y}_j$. However, note that centering the mean image can also be accomplished by first centering each image, i.e. this expression is the same as the previous

$$\frac{\frac{1}{K}\sum_k \left(Y_{jk} - \bar{Y}_k\right)}{\sqrt{\mathbf{1}^\top \mathbf{Q}\mathbf{1}/K^2}}\sigma_C + \mu_C.$$

where $\bar{Y}_k = \frac{1}{J}\sum_j Y_{jk}$ is the image-wise mean and $\{Y_{jk} - \bar{Y}_k\}_j$ is the centered image for pipeline k.

4.1.2 Weighted Stouffer's

If we make one final elaboration we have arrived at the actual NARPS consensus method. Having viewed histograms of the images, we were concerned about how each pipeline (team) had quite different mean and variance. So the actual model we used was

$$\mathbf{Y} \sim \mathcal{MN}(\delta \mathbf{1}_J + \mu \mathbf{1}_K \mathbf{1}_J^\top, \mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}, \mathbf{\Sigma}),$$

where δ is a sum-to-zero pipeline bias term, constrained to sum to zero, and $\mathbf{D} = \text{diag}(\sigma_1^2, ..., \sigma_K^2)$ are pipeline variances.

Hence the final method used constructed adjusted data where each pipeline's mean $\delta_k + \mu$ was removed, before using a weighted Stouffer's method. The adjusted data were

$$Y_{\text{a},jk} = Y_{jk} - (\hat{\delta}_k + \hat{\mu})$$

where we estimate $\hat{\delta}_k + \hat{\mu}$ together simply as the image-wise mean of map k. This was then was assumed to follow

$$Y_{\mathrm{a},j} \sim \mathcal{N}(0, \mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}).$$

Then the weighted Stouffer's method (see Section 3.2.1) can be computed using $\mathbf{W} = \mathbf{D}^{-1}$, adjusted mean $\bar{Y}_{\mathrm{aw},j} = \frac{1}{K}\sum_k Y_{\mathrm{a},jk}/\hat{\sigma}_k$, where σ_k^2 is the image-wise variance of map k. This gives a z-score of

$$\frac{\bar{Y}_{\mathrm{aw},j}}{\sqrt{\mathbf{1}^\top\mathbf{Q}\mathbf{1}/K^2}}.$$

which is mean 0 and variance 1 *even* if the null is not true, since we've substracted off the (image-wise) mean and bias term. This is then ready to be scaled and shifted with a desired consensus moments:

$$\frac{\bar{Y}_{\mathrm{aw},j}}{\sqrt{\mathbf{1}^\top\mathbf{Q}\mathbf{1}/K^2}}\sigma_C + \mu_C.$$

4.1.3 NARPS Consensus Method - Take 2

*** This is my first attempt at describing the NARPS method... leaving this in the document as it might actually be clearer than the previous description.

In summary, the approach of the NARPS consensus approach is to conduct a Stouffer's same-data meta-analysis method, accounting for inter-pipeline correlation, but then adjusting the resulting map so that it has the intervoxel mean and intervoxel variance equal to the inter-pipeline average of the intervoxel mean and variance of each pipeline. The one elaboration from a vanilla Stouffer's approach is that a weighted mean is computed, with weights $w_i = 1/\sigma_k$.

In outline steps:

1. Create standardized maps for each pipeline, by subtracting the image-wise mean and dividing by image-wise standard deviation,

2. Take the voxel-wise average of these standardized maps,

3. Te-standardized this average map accounting for inter-pipeline dependence, and

4. Restore a consensus image-wise mean and standard deviation, creating the final map.

We compute a standardized map Z_k^* for each pipeline k,

$$Z_{kj}^* = \frac{Z_{kj} - \bar{Z}_k}{\hat{\sigma}_k},$$

where $\hat{\sigma}_k^2$ is the image-wise variance for pipeline k,

$$\hat{\sigma}_k^2 \;=\; \frac{1}{J-1}\sum_j (Z_{kj} - \bar{Z}_k)^2.$$

These standardized maps are averaged over pipelines to create a map $\bar{Z}^*$,

$$\bar{Z}_j^* = \frac{1}{K}\sum_k Z_{kj}^*.$$

The image-wise mean of this average of standard maps is zero, and which is finally standardized, scaled and shifted to the consensus standard deviation and mean:

$$Z_{\mathrm{C},j}^* = \frac{\bar{Z}_j^*}{\sqrt{\frac{1}{K^2}\mathbf{1}^\top \mathbf{Q}\mathbf{1}}}\bar{\sigma}_\mathrm{C} + \bar{\mu}_\mathrm{C},$$

where $\frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}\mathbf{1}$ is the attenuated variance of the inter-pipeline mean, $\bar{\sigma}_\mathrm{C}^2 = \frac{1}{K}\sum_k \hat{\sigma}_k^2$ and $\bar{\mu}_\mathrm{C} = \frac{1}{K}\sum_k \bar{Z}_k$.

The inter-pipeline correlation matrix $\hat{\mathbf{Q}}$ was estimated with simple correlation of each pipeline pair

$$\hat{\mathbf{Q}}_{kk'} \;=\; \frac{\sum_j (Z_{k'j} - \bar{Z}_{k'})(Z_{k'j} - \bar{Z}_{k*})}{\sqrt{\left(\sum_j (Z_{k'j} - \bar{Z}_{k'})^2\right)\left(\sum_j (Z_{k'j} - \bar{Z}_{k'})^2\right)}}, \tag{1}$$

where $\bar{Z}_k = \frac{1}{J}Z_{kj}$ is the image-wise mean for pipeline k.

Note that the NARPS census analysis never uses the inter-pipeline variance. However, this was a measure computed as part of the paper's analysis, and was estimated as

$$\hat{\tau}_j^2 = \frac{1}{\mathrm{tr}(\mathbf{C}_K\hat{\mathbf{Q}})}\sum_k (Z_{kj} - \bar{Z}_j)^2,$$

where $\bar{Z}_j = \frac{1}{K}Y_{jk}$ is the inter-pipeline mean at voxel j, and we have carefully accounted for the impact of inter-pipeline correlation: note that $\mathrm{E}(\mathbf{C}_K Z_j) = \tau_j^2\,\mathrm{tr}(\mathbf{C}_K\mathbf{Q})$, where $\mathbf{C}_K$ is the centering matrix, an idempotent matrix $\mathbf{C}_K = \mathbf{I} - \mathbf{1}\mathbf{1}^\top/K$ such that $\mathbf{C}_K\mathbf{Z}$ demeans each column (and $\mathbf{Z}\mathbf{C}_J$ demeans each row).

4.1.4 NARPS Analysis - Discussion

The goal of the NARPS consensus analysis was obtain maps that had the same average characteristics as the input maps. Specifically, if the input maps were mean zero and variance one, so would the output consensus map.

The limitation of this approach is that it doesn't correspond to a single explicit model. For example, the presence of a mean signal (i.e. $\mu_j = \mathrm{E}(Z_{kj})$) isn't accounted for, nor is there really a fixed pipeline bias (say, $\delta_k = \mathrm{E}(Z_{kj})$). Any pipeline bias will contribute to variance τ_j^2, but then any such common variance will be removed in the consensus procedure; likewise, the strength of the correlations in $\mathbf{Q}$ will be influenced by the strength of a common signal $\mu = (\mu_j)$.

In this light, the carefully estimated inter-pipeline variance τ^2 has a very subtle interpretation: $\hat{\tau}^2$ is corrected to undo the effects of correlation and recover the *population* variance. On reflection, the correlation $\mathbf{Q}$ itself seems to possess much more valuable information on pipeline inter-dependence than τ^2.

In the remainder of this document I try to develop a set of same data meta-analysis procedures that start from a unified, coherent model for the matrix of $K \times J$ data that defines each multiverse analysis. I am motivated by the case of statistic images, and the occurrence of K different pipelines that all generate mean 0, variance 1 data but yet still do not produce identical output. This is a case where you *cannot* think of "variance inflation", because the variance is 1.0. However, I do acknowledge that for general data, e.g. %BOLD or anatomical features, it may be the case that a variance inflation model is more appropriate.

5 A Coherent Same Data Meta-Analysis Framework

5.1 SDMA Assumptions

Throughout we assume Normality, as we're working with test statistics and we're ultimately interested in producing valid inferences we would in any case require that assumption.

First a bit more notation. Let μ be the length-J row vector of true, noise free re-

sponses from an ideal pipeline, which we operationally define as simply the mean over the population of all possible pipelines of which we assume we have a representative sample. Image-wise dependence is described by a $J \times J$ correlation matrix $\mathbf{\Sigma}$ ($\text{diag}(\mathbf{\Sigma}) = \mathbf{I}$).

Let δ be the length-K column vector of true, image-wise biases induced by each pipeline. To preserve the interpretation of μ as the common signal mean, we require that δ is mean zero ($\mathbf{1}^\top \delta = 0$). We allow for correlated pipeline variation, described by $K \times K$ covariance matrix $\mathbf{Q}$ (i.e. $\text{diag}(\mathbf{Q})$ is not necessarily identity and can reflect varying variance by pipeline).

We start with an assumption of mean-separability, such that the (true) mean of a test statistic for pipeline k and voxel j is $\mu_j + \delta_k$. Very carefully: If we could repeatedly sample the data (e.g. N subjects from a population for a group analysis), and apply these particular K pipelines, the long-run average for pipeline k and voxel j will be $\mu_j + \delta_k$. In other words, the effect of pipeline k is to shift the mean of the data by a common value δ_k over all voxels.

While this is restrictive, note that we do allow the variance about the mean value $(\mu_j + \delta_k)$ to depend on pipeline. That is, the variance for pipeline k is Q_{kk} and isn't assumed to be the same for all pipelines. (Since we are starting with valid test statistics that have null variance of 1.0, we expect $Q_{kk} \geq 1$.)

These assumptions are compactly expressed by use of the Matrix Normal distribution:

$$\mathbf{Z} \sim \mathcal{MN}(\delta \mathbf{1} J^\top + \mathbf{1}\mu, \mathbf{Q}, \mathbf{\Sigma}).$$

Crucially, this is *not* a hierarchical specification, and we are jointly modeling in a single-level the random variation induced across pipelines and space.

Note that this is fundamentally different to the NARPS approach, in that the variance $\mathbf{Q}$ is the dispersion *about* the pipeline bias and common signal $\delta \mathbf{1} J^\top + \mathbf{1}\mu$. For NARPS we estimated the correlation (Eqn. (1)) with the raw data, i.e. any signal $\mu \neq 0$ structures each pipeline's image and contributes to the correlation. In short, in NARPS, the image signal and noise was both treated as random and captured by $\tau^2 \mathbf{Q}$, while here we're explicitly trying to capture deterministic pipeline signal (δ) and image signal (μ) separate from random pipeline ($\mathbf{Q}$) and image ($\mathbf{\Sigma}$) variance.

5.1.1 Results

From this framework we can immediately obtain the distribution of the inter-pipeline mean test statistic image $\bar{Z} = \mathbf{1}\mathbf{Z}/K$:

$$\bar{Z} \sim \mathcal{N}(\mu, \frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}\mathbf{1}\boldsymbol{\Sigma}),$$

where $\frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}\mathbf{1}$ is the scalar attenuating factor that accounts for averaging over K dependent pipelines.

5.2 Estimation

Because we assume that pipeline effects are centered about the true mean signal, the estimation of the mean test statistic is simple:

$$\hat{\mu} = \frac{1}{K}\mathbf{1}^\top\mathbf{Z}.$$

Estimation of pipeline bias is also just the mean of each image (after centering):

$$\hat{\delta} = \frac{1}{J}\mathbf{C}_K\mathbf{Z}\mathbf{1}J.$$

We do not attempt to estimate the (shared) spatial correlation $J \times J$ $\boldsymbol{\Sigma}$, and none of our results depend on it.

The inter-pipeline covariance is surprisingly tricky. The NARPS approach was just raw correlation, which we can write in matrix mode as $\frac{1}{J}(\mathbf{Z}\mathbf{C}_J)(\mathbf{Z}\mathbf{C}_J)^\top$, i.e. the outer product of the image-wise centered data matrix. However, in our framework this estimate is contaminated by the true signal μ.

Ideally, we would estimate $\mathbf{Q}$ from the true-mean centered data, $\mathbf{Z} - (\delta\mathbf{1}J^\top + \mathbf{1}\mu)$. If we simply plug in the sample means ($\hat{\delta}$ and $\hat{\mu}$) we can compute an estimate:

$$\begin{aligned}\tilde{\mathbf{Q}} &= \frac{1}{J}\left(\mathbf{Z} - (\hat{\delta}\mathbf{1}J^\top + \mathbf{1}\hat{\mu})\right)\left(\mathbf{Z} - (\hat{\delta}\mathbf{1}J^\top + \mathbf{1}\hat{\mu})\right)^\top \\ &= \frac{1}{J}\left(\mathbf{C}_K\mathbf{Z}\mathbf{C}_J\right)\left(\mathbf{C}_K\mathbf{Z}\mathbf{C}_J\right)^\top,\end{aligned}$$

where the second form shows how this estimate the outer product of the doubly-centered Z matrix.

However, there's a problem. Correlation is defined as an outer product of variable-centered data; additionally centering each measurement (here, voxels) induces a bias.

In particular, you can show that computing correlation from voxel-wise centered $\mathbf{C}_K\mathbf{Z}$ induces a bias, and $\tilde{\mathbf{Q}}$ instead estimates $\mathbf{C}_K\mathbf{Q}\mathbf{C}_K$. While for large K this is a minor effect, for small K it is profound.

Consider $K = 2$: When you have only 2 pipelines, centering voxelwise ensures that the sum of pipeline 1 and 2 will be zero at each voxel, i.e. that each will have equal absolute values differing only by a sign. This means that pipeline 2 is entirely predictable from pipeline 1 and vice versa, or, in other words they are perfectly anticorrelated. What I have just said in words looks like this in math mode:

$$\mathbf{C}_2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \mathbf{C}_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

That is, *no matter what* correlation you have when $K = 2$, $\mathbf{Q}$ estimated from voxel-wise centered data will give perfect anticorrelation, $\tilde{\mathbf{Q}}_{1,2} = -1.0$.

For $K \geq 3$ I have some ideas on how to possibly recover $\mathbf{Q}$ from $\tilde{\mathbf{Q}}$, but they involve making simplifying assumptions, like that $\text{diag}(\mathbf{Q})$ is constant (equal pipeline variance) or compound symmetry (inter-pipeline correlation all equal, $Q_{k'k*} = \rho$ for all k', $k*$).

6 Inference

In the current framework, the global null hypothesis $\mathcal{H}_0 : \mu = 0$ will not produce mean zero test statistics unless $\delta = 0$ as well. While this seems quite restrictive, in the context of group analyses, the sample variance is used to normalize and it doesn't seem so crazy that different pipelines can produce different test statistics *yet at the same time* each result is mean 0 and variance 1. Hence, in effect, the null hypothesis we operate on is $\mathcal{H}_0 : \mu = 0, \delta = 0$. Under this setting, at each voxel j $\bar{Z}_j$ has mean zero and variance $\frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}\mathbf{1}$.

Further, in a 'fixed effects' setting, we might want to consider a restriction where pipelines are *not* inducing any additional variation over-and-above that caused by the bias term δ. In that setting we assume $\text{diag}(\mathbf{Q}) = \mathbf{I}$ and we estimate $\mathbf{Q}$ as a correlation matrix. While this may seem arbitrary, it is very close in spirit to the Stouffer's fixed effects meta-analysis combining method, where the average Z value is scaled by square-root of the number of studies. This approach assumes that each Z score is $\mathcal{N}(0, 1)$ and

neglects any random effect.

7 Application

We propose that there are three different inferences that we may wish to conduct: A consensus analysis, where combined inference is as similar to each of the K pipelines as possible; a 'fixed effects' meta-analysis, which discount any random pipeline variation; a 'random effects' meta-analysis, where variance in excess of the expected unit variance is attributed to pipelines and used to down-weight or 'penalize' different results.

7.1 Same Data Consensus Meta-Analysis

This inference attempts to replicate the NARPS analysis within the current framework. Again, in this approach, we consider the image-wise mean and standard deviation as two fundamental summaries of statistic maps, and adjusts sample mean to have an image-wise mean and standard deviation that is as similar as possible to the K different maps as possible.

The focus is again on the inter-pipeline mean image $\bar{Z} = \frac{1}{K}\mathbf{1}^\top\mathbf{Z}$. We again standardise and rescale and shift :

$$\bar{Z}^*_{\mathrm{C},j} = \frac{\bar{Z}_j - \bar{\mu}_\mathrm{C}}{\sqrt{\bar{\sigma}^2_\mathrm{C}\frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}_0\mathbf{1}}}\bar{\sigma}_\mathrm{C} + \bar{\mu}_\mathrm{C} = \frac{\bar{Z}_j - \bar{\mu}_\mathrm{C}}{\sqrt{\frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}_0\mathbf{1}}} + \bar{\mu}_\mathrm{C}$$

where $\mathbf{Q}_0$ is the correlation corresponding to covariance $\mathbf{Q}$.

One difference from the original NARPS method is that the raw mean $\bar{Z}_j$ is used as the starting point, while NARPS first centered and scaled each map into $\bar{Z}^*_j$. I'm not sure how important that is... that ensures each pipeline contributes exactly equally, where here pipelines with larger variance could contribute more.

7.2 Same Data Fixed Effects Meta-Analysis

Now, we don't try to calibrate, and just compute a fixed effects combining inference statistic of

$$\bar{Z}^*_{\mathrm{F},j} = \frac{\bar{Z}_j}{\sqrt{\frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}_0\mathbf{1}}}.$$

This much simpler expression will still be mean zero and variance one if each input map Z_k is also mean zero and variance 1, but to the extent there is common signal across the pipelines, the signal will be magnified by the denominator.

This expression is 'fixed effect' in that by using $\mathbf{Q}_0$ instead of $\mathbf{Q}$, we are not letting inter-pipeline variance attenuate the strength of the effect.

7.3 Same Data Random Effects Meta-Analysis

By making one small alteration to the previous expression, we obtain a 'random effects' inference using $\mathbf{Q}$ which accounts for interpipeline variance in excess of the natural (test statistic) unit variance:

$$\bar{Z}^*_{\mathrm{R},j} = \frac{\bar{Z}_j}{\sqrt{\frac{1}{K^2}\mathbf{1}^\top \mathbf{Q}\mathbf{1}}}.$$

Of course, this is not the optimal random effects inference, as we are using an unweighted mean for $\bar{Z}_j$ when a weighted mean like

$$\sum_k \frac{(\mathbf{Q}_{kk})^{-1/2} Z_{kj}}{\sum_{k'} (\mathbf{Q}_{k'k'})^{-1/2}}$$

would be more optimal.

8 Multiverse Analysis of Arbitrary Values

When the object of interest is not a test statistic, we can no longer appeal to a null hypothesis and an expected mean 0, variance 1 behavior. This is in some ways an easier problem, that we don't need to carefully (awkwardly) construct a process that when feed $\mathcal{N}(0,1)$ data produces $\mathcal{N}(0,1)$ output. Rather, we can consider a natural hierarchical structure.

8.1 Model

For consistency, but at the risk of confusion, we continue to use $\mathbf{Z}$ as the data matrix even though now we do not consider the values to be Z-scores. For voxel j and multiverse pipeline k we observe

$$Z_{kj} = Y_j + M_{jk}$$

an additive superposition of a signal Y_j common to all pipelines and a pipeline-specific perturbation M_{jk}. Both of these components are random. The signal, over repeated sampling of the population (of single subjects, or sets of subjects) is distributed

$$Y \sim \mathcal{N}(\mu, \mathbf{\Sigma}_Y)$$

where Y is a length J row vector, μ is the mean $\mathbf{\Sigma}_Y$ the covariance. Independently, the pipeline perturbations are distributed

$$\mathbf{M} \sim \mathcal{MN}(\delta \mathbf{1} J^\top, \mathbf{Q}, \mathbf{\Sigma}_M).$$

where $\mathbf{M}$ is a $K \times J$ matrix.

As written this model is over parameterized (not identifiable). For starters, we assume that the inter-voxel average pipeline perturbations δ are mean zero, $\sum_k \delta_k = 0$. We probably have to put some constraints on some combination of $\mathbf{\Sigma}_Y$ and $\mathbf{\Sigma}_M$, and perhaps $\mathbf{Q}$, but we leave this for now.

This approach implies some conditional/hierarchical interpretations that may be useful. If we only ever see one pipeline k (i.e. the usual non-multiverse case) the data are distributed

$$Z_k | H_k \sim \mathcal{N}(\mu + H_k, \mathbf{\Sigma}_Y).$$

That is, the kth result is shifted by the random realization of the pipeline perturbation H_k.

While we can never, in practice, condition on the (unseen) common random signal Y, it is useful to record its conditional distribution:

$$\mathbf{Z} | Y \sim \mathcal{MN}(\delta \mathbf{1} J^\top + \mathbf{1} Y, \mathbf{Q}, \mathbf{\Sigma}_H).$$

Finally we can record the distribution of the inter-pipeline mean $\bar{Z} = \frac{1}{K} \mathbf{1}^\top \mathbf{Z}$,

$$\bar{Z} \sim \mathcal{N}(\mu, \mathbf{\Sigma}_Y + \tfrac{1}{K^2} \mathbf{1}^\top \mathbf{Q} \mathbf{1} \mathbf{\Sigma}_H).$$

Or, for voxel k

$$\bar{Z}_k \sim \mathcal{N}(\mu_k, \sigma^2_{Y,k} + \tfrac{1}{K^2} \mathbf{1}^\top \mathbf{Q} \mathbf{1} \sigma^2_{H,k}),$$

where $\sigma^2_{Y,k} = (\mathbf{\Sigma}_Y)_{k,k}$ and $\sigma^2_{H,k} = (\mathbf{\Sigma}_H)_{k,k}$, showing that inter-pipeline sample mean has two contributions to its variance: A portion independent of pipeline, $\sigma^2_{Y,k}$, that reflects

random variation in the population from the 'uncorrupted' Y, and a portion that is due to pipeline variance and which can be reduced by using more pipelines (because $1/K^2$ shrinks) as long as they're not too correlated (in the extreme, if $\mathbf{Q}$ is all 1's due to perfect correlation, $\frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}\mathbf{1} = 1$ and there is no reduction in variance from pipeline averaging).

A Illustration with Matlab

A central question is how to regard variance in the same-data, multiverse setting. Consider the following thought question:

> For a NARPS-like study, suppose that we have $K = 70$ teams and the organisers simulate $N = 100$ random values from a mean 0, variance 1 normal distribution. Each team is sent the 100 values and asked to compute a 1-sample t-test. The 70 values that are returned won't necessarily be identical – due to different operating systems and numerical libraries – but they would be very similar and have a sample variance far less than the true variance of 1.

But this is confusing sample variance with true variance, and neglecting the impact of correlation. Consider an alternate thought experiment:

> In the same setting, the organisers simulate $N = 100$ random $J = 100,000$-voxel *images*, each voxel independently sampled from a mean 0, variance 1 normal distribution. Each team is sent the 100 images and asked to compute a 1-sample t-test at each voxel. Even though a single fixed dataset was sent to all teams, the returned matrix of values, $70 \times 100,000$, clearly exhibits randomness: Each row will have values drawn from a t_{99} distribution, and each column will be very highly correlated if not identical.

The point here is that each column is in fact *random*, variance 1, but with very high correlation. In particular, as 99 is a quite large t degrees of freedom we could call the resulting matrix as normal, and specifically could describe it as following a "matrix normal" distribution, $\mathcal{MN}(0, \mathbf{Q}, \mathbf{I})$: The row covariance is the $K \times K$ covariance matrix

$\mathbf{Q}$ and the columns are independent. In this case, the off diagonals of $\mathbf{Q}$ will be nearly be 1.0, but we would *still* say that the variance of every element is 1.0!

This may seem implausible, but consider the following Matlab demonstration

```
% Simulate iid data
X=randn(70,100000);

% Compute sample variance either way, get same value around 1.0
var(X(:,1)) % variance of first column ~ 1.0
var(X(1,:)) % variance of first row    ~ 1.0

% Now simulate data with a very strong row correlation
Q=eye(70)*0.1+ones(70)*0.9};
X=mvnrnd(zeros(1,70),Q,100000)';

% Now you'll find very different variance
var(X(:,1))  % variance of first column ~ 0.3
var(X(1,:))  % variance of first row    ~ 1.0

% Now, using knowledge of true correlation, correct sample variance
C_K = (eye(70)-ones(70)/70);
var(X(:,1))*(70-1)/trace(C_K*Q) % var of first column ~ 1.0
```

Where, in the final expression we have canceled out the usual denominator of `70-1` and replaced it with the scaling factor $\mathrm{tr}(\mathbf{C}_K\mathbf{Q})$. This shows that accounting for the correlation lets you recover the true population variance.

B Linear Algebra Results

B.1 Sample mean attenuates variance

While sample mean of independent samples has a well known variance reduction, we show that this also applies to a correlated sample.

For any square matrix $\mathbf{A}$, the quadratic form is bounded

$$x^\top \mathbf{A} x \leq \lambda_1 x^\top x,$$

where λ_1 is the largest eigenvalue of $\mathbf{A}$. If we further assume $\mathbf{A}$ is positive definite then $\sum_{k=1}^{K} \lambda_k = \text{tr}(\mathbf{A})$ implies $\lambda_1 \leq \text{tr}(\mathbf{A})$, and for the case of a correlation matrix where $\text{tr}(\mathbf{A}) = K$, $\lambda_1 \leq K$

Then the variance of a mean of a random vector with homogeneous unit variance and correlation $\mathbf{Q}$, $\frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}\mathbf{1}$, can be at most one:

$$\frac{1}{K^2}\mathbf{1}^\top\mathbf{Q}\mathbf{1} \leq \frac{1}{K}\lambda_1 \leq 1$$

B.2 Expected value of a matrix normal product

A useful result for matrix normals due to Von rosen (1988).

References

Samuel A Stouffer, Edward A Suchman, Leland C Devinney, Shirley A Star, and Robin M Williams Jr. *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1.* Princeton Univ. Press, Oxford, England, 1949.

Han Bossier, Thomas E Nichols, and Beatrijs Moerkerke. Standardized Effect Sizes and Image-Based Meta-Analytical Approaches for fMRI Data. *bioRxiv*, pages 1–61, 2019. doi: 10.1101/865881. URL https://www.biorxiv.org/content/10.1101/865881v1.

Larry V. Hedges. Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, 6(2):107–128, 1981. URL http://www.jstor.org/stable/1164588.

Dmitri V Zaykin. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of evolutionary biology*, 24 (8):1836–41, aug 2011. ISSN 1420-9101. doi: 10.1111/j.1420-9101.2011.02297.x. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3135688{&}tool=pmcentrez{&}rendertype=abstract.

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M. W. J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolini, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Luca, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A. E. Golowin, Carlos González-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, João F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris

Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E. van 't Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, jun 2020. ISSN 0028-0836. doi: 10.1038/s41586-020-2314-9. URL http://www.nature.com/articles/s41586-020-2314-9.

Dietrich Von rosen. Moments for matrix normal variables. *Statistics*, 19(4):575–583, jan 1988. ISSN 0233-1888. doi: 10.1080/02331888808802132. URL http://www.tandfonline.com/doi/abs/10.1080/02331888808802132.