Spring 5-13-2023

Development of Bayesian Hierarchical Methods involving Meta-Analysis

Jackson Barth
Southern Methodist University, jbarth216@gmail.com

DEVELOPMENT OF BAYESIAN HIERARCHICAL METHODS

INVOLVING META-ANALYSIS

Approved by:

Dr. Xinlei Wang
Professor, UTA

Dr. Jing Cao
Professor, SMU

Dr. Dan Heitjan
Professor, SMU

Dr. Raanju Sundararajan
Assistant Professor, SMU

Dr. Guanghua Xiao
Professor, UTSW

DEVELOPMENT OF BAYESIAN HIERARCHICAL METHODS INVOLVING META-ANALYSIS

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Statistical Science

by

Jackson Barth

B.S., Commerce and Business Administration, University of Alabama
M.S., Economics, University of Alabama

May 13, 2023

Acknowledgments

I am deeply grateful to my advisor, Dr. Sherry Wang, whose guidance, patience, and expertise I have greatly benefited from. I would also like to acknowledge the Statistical Science faculty at SMU and my dissertation committee. Finally, I would not be here without my friends and family, particularly my wife, Claire. I cannot thank you enough for your love and support through this process.

Barth, Jackson B.S., Commerce and Business Administration, University of Alabama
M.S., Economics, University of Alabama

Development of Bayesian Hierarchical Methods
involving Meta-Analysis

Advisor: Dr. Xinlei Wang
Doctor of Philosophy degree conferred May 13, 2023
Dissertation completed April 28, 2023

When conducting statistical analysis in the Bayesian paradigm, the most critical decision made by the researcher is the identification of a prior distribution for a parameter. Despite the mathematical soundness of the Bayesian approach, a wrongly specified prior can lead to biased and incorrect results. To avoid this, prior distributions should be based on real data, which are easily accessible in the "big data" era. This dissertation explores two applications of Bayesian hierarchical modelling that incorporate information obtained from a meta-analysis. The first of these applications is in the normalization of genomics data, specifically for nanostring nCounter datasets. A meta-analysis of 13 nCounter datasets were used to identify informative prior distributions, which were then incorporated into RCRnorm, a leading normalization procedure for nCounter data that utilizes a Bayesian hierarchical model. With the new prior and other structural changes applied to the underlying model, the new normalization approach "MetaNorm", improves on its predecessor with faster speed, better convergence and stabilized estimation, even when normalizing lower-quality datasets. The second application covers a novel sample-size determination method for one and two-sample t-tests. This novel methodology uses an empirical Bayes approach to construct a posterior predictive distribution for the variance estimate, based on data from previous studies. Simulations and empirical studies demonstrate that this methodology outperforms other aggregate approaches (simple average, weighted average, median) in variance estimation

for SSD, especially in meta-analyses with large disparities in sample size and variance. Thus, it offers a robust and practical solution for sample size determination in t-tests.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

This work is dedicated to Claire and Oliver, whom I love dearly.

CHAPTER 1

MetaNorm: Incorporating Meta-analytic Priors into Normalization of NanoString nCounter Data

1.1. Introduction

1.1.1. Normalization methods for nCounter gene expression data The medium-throughput platform NanoString nCounter has quickly become one of the most popular and efficient ways to analyze complex mRNA transcripts (Geiss et al., 2008). One reason for this is its ability to process formalin fixed, parrafin-embedded (FFPE) tissue samples effectively. While freshly frozen (FF) samples experience less degradation in storage, FFPE samples require less stringent storage requirements and are therefore cheaper and easier to retain (Perlmutter et al., 2004). This has led to a ubiquity of FFPE samples and has made the nCounter system an invaluable resource in medical research. Among other areas, the nCounter framework has been used to analyze FFPE samples in studies of colon, breast, and lung cancer (Chen et al., 2016a; Lim et al., 2020; Walter et al., 2016). However, a significant downside to using FFPE samples is the level of mRNA "modification" during the preservation and storage of the sample (Masuda et al., 1999), causing higher levels of variability and uncertainty in the readings. Thus, it is critical that the gene read counts undergo an efficient normalization procedure before being formally analyzed. In addition to degradation in FFPE samples, normalization can also help to remove background noise or lane-by-lane variation, which is common in FF samples as well. Perhaps the simplest approach to normalization is NanoStringNorm, an R package that implements NanoString guidelines and relies on summary statistics from the positive, negative and housekeeping probes to account for separate types of variation (Waggott et al., 2012). Other normalization methods,

such as NAPPA[1] and NanoStringDiff (Wang et al., 2016), use model-based approaches that can better account for the complexities in the data. Despite the improvements of NAPPA and NanoStringDiff, all three of these approaches use each type of probe in the same way, in that positive probes only remove lane-by-lane variation, negative probes only remove background noise, and housekeeping genes only remove variation in the amount of the input sample material. Among the latest normalization protocols to be developed is RCRnorm, which stands for **r**andom **c**oefficient hierarchical **r**egression **norm**alization (Jia et al., 2019). RCRnorm is unique in that (1) it is an integrated system of hierarchical models for the different types of probes, (2) it is designed specifically to normalize FFPE data (but can be used with other types of samples), (3) it allows for sample-to-sample variation in the housekeeping genes (Eisenberg and Levanon, 2013), and (4) it uses a Bayesian framework rather than a frequentist framework for normalization, thus allowing statistical inference about key model parameters and uncertainty quantification for estimates while leading to better interpretability. Essentially, the RCRnorm model assumes that all probes in a given sample have a shared intercept and slope on a linear regression of $\log_{10}$ (gene read count) and $\log_{10}$ (RNA amount). While other probe-specific effects are included, this allows all types of probes to influence the normalization process, yielding more robust estimates (technical details of RCRnorm are outlined in Section 1.1.3). The model is implemented with a Gibbs sampler, which is a type of Markov chain Monte Carlo (MCMC) algorithm where parameters are individually resampled until reaching a stationary distribution. In general, RCRnorm is much more flexible than its frequentist counterparts, with the MCMC algorithm able to explore the extremely high-dimensional parameter space and land on the targeted posterior distribution. Jia et al. (2019) also shows a significant improvement in the correlation between FFPE samples and their corresponding FF samples, which have gene read counts that are thought to be more accurate to the ground truth than the FFPE counts. They show that RCRnorm has superior correlation, both at the gene level and patient level.

[1]R Package: http://CRAN.R-project.org/package=NAPPA

Additionally, in a comparison of different normalization procedures, Bhattacharya et al. (2021) find that RCRnorm removes more technical variation than its counterparts, including NanoStringNorm and NanoStringDiff.

1.1.2. Motivation to improve RCRnorm

Despite its impressive performance, RCRnorm is not without its drawbacks. The weakest point is its high computational cost, which is well-documented in Bhattacharya et al. (2021). Large datasets will require a significant amount of time and working memory to produce results, making RCRnorm an occasionally inconvenient choice for normalization. While computational inefficiency is a common criticism of MCMC algorithms, often times this can be alleviated with optimized code and special software packages. Another area of concern is the high number of parameters, though it is typical in Bayesian hierarchical models. It is common that Bayesian approaches rely on priors to regularize these parameters. Still, in a highly complex system, an abundance of parameters combined with an absence of accurate prior information can cause weak or fake convergence. Further, a concern closer to the crux of the RCRnorm model is the construction of the prior distributions for the hyper-parameters μ_a, μ_b. These hyper-parameters represent the means of the sample-specific intercept and slope terms (respectively), which should reflect how the nCounter system inherently operates rather than the characteristics of an individual dataset. They have a significant impact on model estimates and so the results of RCRnorm are sensitive to the choice of prior distribution placed on these hyper-parameters. As described in Jia et al. (2019), RCRnorm uses a jackknife approach on the positive probe data to estimate the prior mean and variance for μ_a, μ_b (a normal distribution is assumed), so that the priors vary across datasets. This makes the results of RCRnorm more sensitive to the size and quality of the dataset being normalized, in the sense that any data-based prior has no ability to mitigate potential bias that can arise from sparse or unreliable data. For these reasons,

identifying and implementing new priors for these quantities will improve model estimation, especially for messy or sparse datasets.

Since μ_a, μ_b are important global parameters in the RCRnorm model, replacing the jackknife priors with non-informative priors would not effectively mitigate the potential bias and noise stemming from low quality data. Rather, μ_a, μ_b reflects the underlying mechanism of the nCounter system and constructing an informative *a priori* distribution based on largely existing data of the same type will help avoid excessive data-dependence and provide external calibration to low quality data. In this big-data era, accessing data from multiple independent studies has become increasingly easy, due to federal regulations and substantial efforts made in data sharing. Thus, leveraging such information to construct informative yet objective priors becomes feasible.

The goal of this study is to create these prior distributions via a rigorous meta-analysis of public FFPE gene expression datasets from independent studies using NanoString nCounter platform. To do so, we devise a Bayesian hierarchical model, which adopts the linear regression setup with random coefficients from RCRnorm. However, it is distinct from the RCRnorm model in several aspects. First of all, it deals with data from multiple studies and includes a layer to account for study-specific effects. Secondly, for reasons detailed in the beginning of Section 1.2, this new model focuses on modeling data from positive control probes only. Thus, it is able to allow probe-specific variance terms, removing the simplifying assumption made by RCRnorm (i.e., constant variance across all positive controls). Thirdly, unlike RCRnorm, the new model, with much more data available in a meta-analysis, is able to account for the dependence between the sample-specific intercepts and slopes.

In addition to the new prior from the Bayesian meta-analysis, we have also made several important changes to RCRnorm. We refer to this new algorithm as MetaNorm. Among these enhancements are (1) optimized code and data structure, (2) implementation of constraints on the positive and negative probe residuals, and (3) a simplified sampling approach to

normalized RNA expression estimates for housekeeping and regular genes. We show that these adjustments, along with the informative meta-analysis prior, improve computational cost, model convergence, prediction error and performance with low-quality data. In the remainder of this section, we outline in detail the RCRnorm model from Jia et al. (2019). In Section 1.2, we provide an overview of our meta-analysis and a detailed summary of other improvements made to RCRnorm. In Section 1.3, we compare the performance of MetaNorm with RCRnorm on four real-world datasets. Finally, we provide a discussion of our findings in Section 2.5.

1.1.3. Overview of the RCRnorm model

We conclude the Introduction with an overview of the notation used in this paper and a description of the RCRnorm model. Each FFPE dataset has I patient samples (indexed by i) with P, N, H, R positive probes, negative probes, housekeeping genes and regular genes (indexed by p, n, h, r), respectively. The datasets have a factorial structure, with each sample and gene being balanced with respect to the number of repetitions in the overall dataset ($I \cdot (P + N + H + R)$ total observations). Typically, there are 6 positive probes ($P = 6$), a relatively small number of negative probes and housekeeping genes (i.e., $N < 10$ and $H < 20$) and a high number of regular genes ($R > 80$). Each of the six positive probes contains a known amount of mRNA, while the amount of mRNA in a regular or housekeeping gene is unknown and varies across samples. These mRNA levels must be estimated using the read counts. For all negative probes, target transcripts are absent and so their mRNA levels should be zero ideally. The aim of RCRnorm is to create consistent, unbiased estimates for this mRNA amount for every gene/sample combination that adjust for unwanted biological

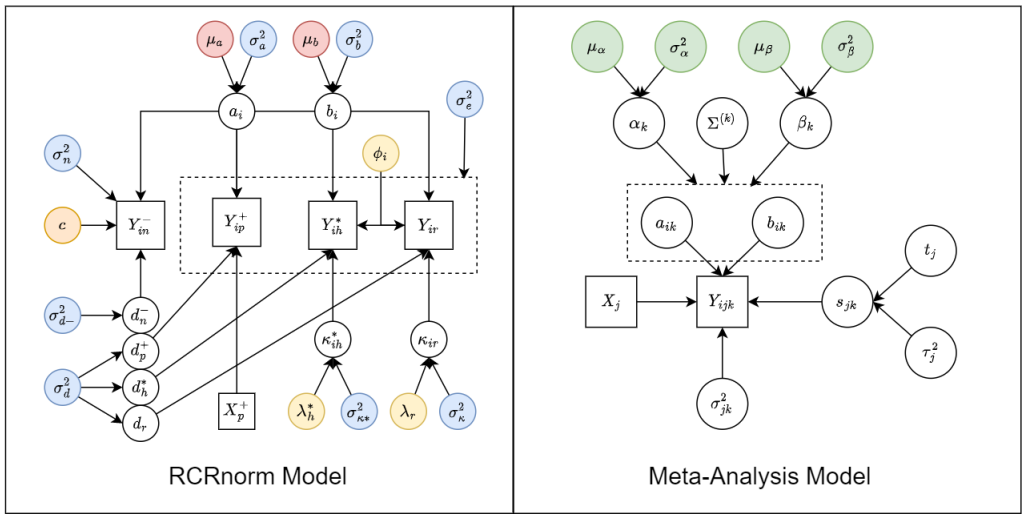

Figure 1.1: Graphic representation of the original RCRnorm model (left) and the meta-analysis model (right). Squares denote observed or fixed values while circles represent model parameters (random variables). Parameters and hyperparameters with blue color-coding represent diffuse inverse gamma priors, orange and yellow have "safe range" uniform priors, with yellow having ranges determined by positive probe data. Red hyperparameters have priors created by a jackknife analysis of the positive probe data. On the right, green parameters represent those being estimated by the meta-analysis.

and technical effects. The remainder of this section summarizes the outline and technical details of the model presented in Jia et al. (2019).

The RCRnorm model is based on a series of integrated linear regression models with random coefficients. At the lowest level, the $\log_{10}$ gene read count Y for each of the four probe types is summarized in the below equations:

$$Y_{ip}^{+} \sim N\left(a_i + b_i X_p^{+} + d_p^{+}, \sigma_e^2\right) \tag{1.1}$$

$$Y_{in}^{-} \sim N\left(a_i + b_i c + d_n^{-}, \sigma_n^2\right) \tag{1.2}$$

$$Y_{ih}^{*} \sim N\left(a_i + b_i X_{ih}^{*} + d_h^{*}, \sigma_e^2\right), \qquad \text{where } X_{ih}^{*} = \phi_i + \kappa_{ih}^{*} \tag{1.3}$$

$$Y_{ir} \sim N\left(a_i + b_i X_{ir} + d_r, \sigma_e^2\right), \qquad \text{where } X_{ir} = \phi_i + \kappa_{ir} \tag{1.4}$$

where the superscripts $(+, -, *)$ indicate positive probes, negative probes, and housekeeping genes, respectively (no superscript in gene specific variables indicates regular genes); the

$\log_{10}$ RNA levels are known and fixed for each positive probe (X_p^+) in all samples and are unknown but differ for all housekeeping and regular gene observations (X_{ih}^*, X_{ir}); the unknown scalar c for negative probes can be interpreted as the mean non-specific binding level due to background noise. The intercept and slope terms (a_i, b_i) reflect how the log read count Y relates to the log RNA level X in each sample, and are assumed to be independent and identically distributed normal variables: $a_i \overset{iid}{\sim} N(\mu_a, \sigma_a^2)$ and $b_i \overset{iid}{\sim} N(\mu_b, \sigma_b^2)$. Note that for a given sample, these terms are shared by all the four categories and so occur in all four equations. Whereas ϕ_i represents the sample specific degradation level (with the constraint $\sum_i^I \phi_i = 0$), κ_{ir} and κ_{ih}^* signify sample/gene specific mRNA levels before degradation (these parameters can be thought of as the model output), with $\kappa_{ir} \sim N(\lambda_r, \sigma_\kappa^2)$ and $\kappa_{ih}^* \sim N(\lambda_h^*, \sigma_{\kappa*}^2)$. The variables $(d_p^+, d_n^-, d_h^*, d_r)$ represent deviations from the general linear pattern and can be thought of as probe-specific residuals, with $d_p^+, d_h^*, d_r \sim N(0, \sigma_d^2)$ and $d_n^- \sim N(0, \sigma_{d-}^2)$. Jia et al. (2019) suggests that all probe types have similar variability except for the negative probes, which tend to have higher variances, hence σ_d^2 vs. σ_{d-}^2 and similarly, σ_e^2 vs. σ_n^2 in (1.1)–(1.4). The left panel of Figure 1.1 summarizes the hierarchical structure of the RCRnorm model. Note that the hyper-parameters μ_a and μ_b have data-based jackknife priors, which we seek to improve with the results of our meta-analysis. See sections 3 and 4 of Jia et al. (2019) for more details of RCRnorm.

1.2. Methods

The first step in improving RCRnorm is to identify realistic, informative prior distributions to replace the current priors for μ_a and μ_b. We maintain the assumption that the priors are independent normal distributions (the conjugate priors for μ_a and μ_b), but rather than relying on the data that is currently being analyzed, the MetaNorm prior comes from a meta-analysis. We identified multiple independent NanoString nCounter gene expression datasets of FFPE samples from past studies and designed a Bayesian hierarchical model for meta-analysis to attain parameter estimates for the prior mean and variance of μ_a and μ_b

irrespective of a specific dataset. For positive control probes, both Y (the log transformed count) and X (input RNA amount) are known so that we have strong information from the data about the intercept and slope terms. In contrast, for housekeeping and regular genes, only Y is known and X is latent with a complex underlying structure, so such information is much weaker. Since our main interest is on the intercept and slope terms, the meta-analysis only considers these positive controls (6 per sample) to find the posterior distributions for μ_a and μ_b. As will be shown later, this would also allow us to remove simplifying assumptions made by RCRnorm for the purpose of estimation stability. The remainder of this section outlines the design, data and results of this meta-analysis.

1.2.1. Datasets

We identified 17 potential datasets containing mRNA FFPE samples from human subjects, which are listed in detail in Table reftab:Meta-Analysis-Datasets of the supplementary material, including GSE numbers for locating the data[2]. Each of these datasets are publicly available and can be located at ncbi.nlm.nih.gov. Two of these datasets (id 4 and 5) were used for testing the RCRnorm model, and were removed from the meta-analysis pool in an attempt to be completely independent from the original study. Two other datasets were removed for having low data quality, with one having a high number of low-quality samples by the NanoString quality control guidelines[3] and the other having a high positive correlation between a_i and b_i estimates, contrary to the rest of the pool. This left us with $K = 13$ datasets, which ranges in size from $n_{13} = 8$ up to $n_9 = 1,950$ samples. All 13 datasets had 6 positive probes, each having the same control RNA levels ($128, 32, 8, 2, .5, .125$ fM for probes 1-6, respectively).

[2]https://www.ncbi.nlm.nih.gov/gds

[3]https://nanostring.com/wp-content/uploads/Gene_Expression_Data_Analysis_Guidelines.pdf

1.2.2. Design

This section outlines the hierarchical model used to create our meta-analysis prior. We assume that the $\log_{10}$ count of each sample/probe/study combination Y_{ijk} is characterized by

$$Y_{ijk} = a_{ik} + b_{ik}X_j^+ + r_{ijk} \tag{1.5}$$

where X_j^+ is the $\log_{10}$ RNA level for each probe j, which is fixed across all samples and studies at $\log_{10}(128, 32, 8, 2, .5, .125)$, a_{ik} and b_{ik} are the slope and intercept terms for sample i of study k, and r_{ijk} is the residual term reflecting the remaining variability of Y_{ijk} after taking into account the linear trend, for $i = 1, \dots I_k$, $j = 1, \dots 6$, and $k = 1, \dots K$ (I_k is the number of samples in study k and $K = 13$). This structure maintains the original design of RCRnorm while accounting for multiple data sources.

First, we focus on modeling the random regression coefficients a_{ik} and b_{ik} in (1.5). Unlike RCRnorm, the meta-analysis includes samples from a variety of datasets, meaning heterogeneity between data sources is a potential factor. To examine this, we first calculated empirical estimates for all a_{ik}, b_{ik} by fitting individual linear regressions of the form

$$E[Y_{ijk}|a_{ik}, b_{ik}] = a_{ik} + b_{ik}X_j^+ \tag{1.6}$$

where each $(\hat{a}_{ik}, \hat{b}_{ik})$ is calculated with 6 positive control datapoints of sample i ($j = 1, ..., 6$). Figure 1.2(a) and (b) summarize the estimates obtained from this analysis. There is clear heterogeneity in the distributions, both in the center and variance, when broken down by dataset. Therefore it is reasonable to assume that parameters associated with a_{ik} and b_{ik} are diverse for different datasets. This leads us to a separate bivariate normal distribution

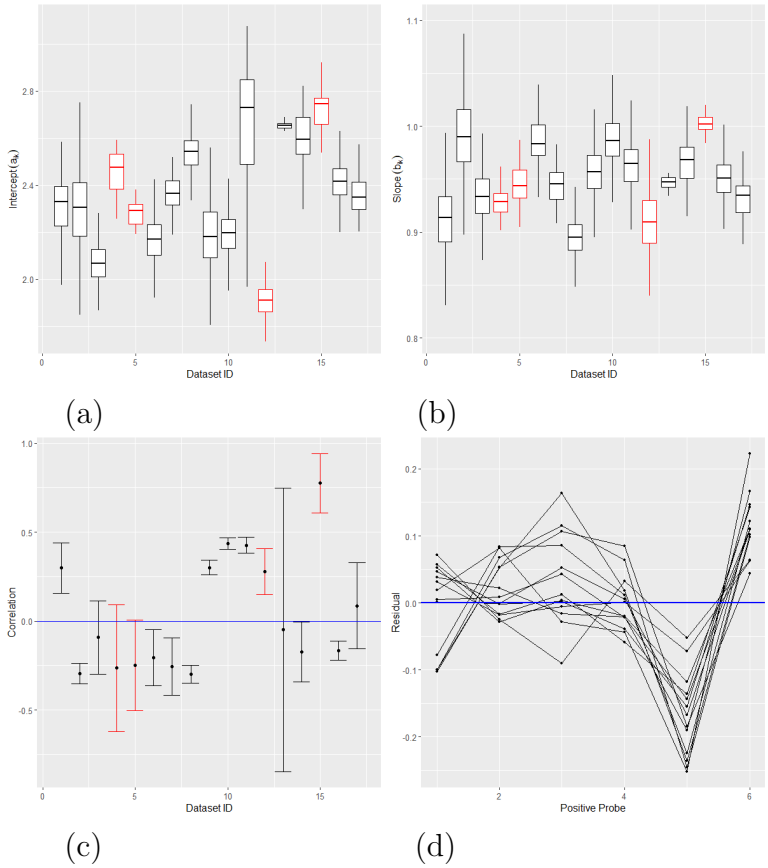

Figure 1.2: Meta-analysis: (a) and (b) show boxplots without outliers for empirical a_{ik} (left) and b_{ik} (right) estimates by dataset; (c) shows error bar plots of (a_{ik}, b_{ik}) correlation by dataset; (d) shows average empirical residual for positive probes by dataset. Red color represents datasets that were excluded from the meta-analysis (id 4, 5, 12, and 15). In plot (c), the dots represent the Pearson correlation between empirical estimates of a_{ik}, b_{ik} while the bars reflect 95% confidence bounds. In plot (d), points on the same line come from the same dataset, and this plot only shows the datasets included in the meta-analysis.

in (1.7) for each study k, where the study-specific means α_k and β_k are equivalent to μ_a, μ_b in the RCRnorm model.

The next step is to determine the nature of the covariance matrix $\Sigma^{(k)}$, that is, whether or not a_{ik} is independent from b_{ik}. Figure 1.2(c) shows a breakdown of correlation between the empirical estimates $\hat{a}_{ik}$ and $\hat{b}_{ik}$ by dataset. Several datasets show positive correlation while others have a negative correlation, providing justification for modeling the study-specific correlation rather than assuming a common correlation value across all studies. Despite the fact that a few datasets would not have enough evidence to reject $r = 0$

at the $\alpha = .05$ level, 12 of the 17 datasets have 95% confidence bounds strictly above or below 0, meaning we cannot assume that a_{ik} and b_{ik} are independent of each other when coming from the same dataset k. The implication of this is that $\Sigma^{(k)}$ must be modeled with a multivariate distribution rather than independent univariate distributions. Several candidate distributions were considered, but a thorough simulation, which will be detailed in Section 1.2.4, showed little difference in performance between prior distributions. We therefore chose the path of least resistance by implementing a classic inverse-Wishart distribution in (1.8), which is the conjugate prior for the covariance matrix of a multivariate normal distribution and allows for both simplicity and computational ease. Finally, the mean parameters α_k and β_k are modeled independently, since the empirical estimates have a correlation near 0. The hyperparameters μ_α, μ_β have "safe range" uniform priors while $\sigma^2_\alpha, \sigma^2_\beta$ have diffuse inverse gamma priors. Safe range prior distributions are those that follow a uniform distribution with a range that comfortably covers all plausible values for the parameters (usually it has a half-width of 3 or 5 standard deviations).

The full hierarchical structure of the slope and intercept terms is summarized below,

$$\begin{pmatrix} a_{ik} \\ b_{ik} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \alpha_k \\ \beta_k \end{pmatrix}, \Sigma^{(k)} \right) \tag{1.7}$$

$$\Sigma^{(k)} \sim IW_3\left(I_2\right), \qquad \text{where } I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{1.8}$$

$$\alpha_k \sim N\left(\mu_\alpha, \sigma^2_\alpha\right) \qquad \beta_k \sim N\left(\mu_\beta, \sigma^2_\beta\right)$$

$$\mu_\alpha \sim Uniform\left(L_\alpha, U_\alpha\right) \qquad \mu_\beta \sim Uniform\left(L_\beta, U_\beta\right), \qquad \sigma^2_\alpha, \sigma^2_\beta \sim IG(\epsilon, \epsilon)$$

where IW_3 is an inverse-Wishart distribution with three degrees of freedom, IG is an inverse-gamma distribution, and ϵ is a placeholder for a small value (i.e., .01) so that the prior is diffuse.

We are now left with the residual term r_{ijk} from (1.5), which is modeled by the following hierarchical structure:

$$r_{ijk} = s_{jk} + \delta_{ijk} \tag{1.9}$$

$$s_{jk} \sim N(t_j, \tau^2) \qquad \text{where } \sum_{j=1}^{6} s_{jk} = 0 \text{ and } \sum_{j=1}^{6} s_{jk} X_j^+ = 0 \tag{1.10}$$

$$\delta_{ijk} \sim N\left(0, \sigma_{jk}^2\right)$$

$$t_j \sim Uniform\left(L_j, U_j\right) \qquad \text{where } \sum_{j=1}^{6} t_j = 0 \text{ and } \sum_{j=1}^{6} t_j X_j^+ = 0,$$

$$\tau^2, \sigma_{jk}^2 \sim IG(\epsilon, \epsilon).$$

The above structure was carefully chosen to reflect what we observed from the data and to achieve computing efficiency and stability. Figure 1.2(d) shows the mean residual for each corresponding probe by dataset. While some probes seem to be centered at 0, there is a clear pattern in the residuals especially for 5 and 6. This indicates that the model for r_{ijk} should have probe specific terms. Since there is significant heterogeneity in the probe effect between different datasets, the parameters are specific to both the probe and dataset (i.e., indexed by j, k). Therefore the residual term can be broken down into two pieces, shown in (1.9). The parameters s_{jk} are assumed to have a probe specific mean t_j and a shared variance τ^2, as the points by probe in Figure 1.2(d) have centers clearly distinct from one another but have similar variability. We further enforce the constraints shown in (1.10), to avoid issues with identifiability, making s_{jk} dependent on other probe effects from the same

dataset. We include these constraints to stabilize our distribution in the absence of strong prior information, relying on the fact that the log-transformed positive probe data have a strong linear relationship with the log-transformed control RNA expression levels in each study alone. Since these terms represent probe specific residuals, we can reasonably expect them to follow the same rules as residuals, namely that they sum to 0 with and without multiplication by the factor-effects. While this is a frequentist technique, there is significant value in using both frequentist and bayesian approaches together, despite the tendency to think of them as separate, binary categories (Bayarri and Berger, 2004). In this case, our analysis benefits from the constraints by reaching quick convergence without using artificially informative prior distributions. A "safe range" uniform prior is applied to t_j (including the same constraints used for s_{jk}) while a non-informative IG prior is used for τ^2. Finally, δ_{ijk} acts as our pure residual term that has no pattern remaining, which we center at 0. Given the diversity shown in this section by probe and by dataset, we assume that the variance of this residual differs by these factors. We use a diffuse IG prior for the σ^2_{jk} terms.

The right panel of Figure 1.1 gives a concise summary of variable relationships in the meta-analysis. The values we want to identify are represented by the green parameters in the figure. The full probability model is shown below:

$$
\begin{aligned}
P(\boldsymbol{Y}, \Theta) \propto & \prod_{k=1}^{K} \left(\prod_{i=1}^{n_k} \left(\prod_{j=1}^{6} N\left(Y_{ijk} | a_{ik} + b_{ik} X_j^+ + s_{jk}, \sigma^2_{jk}\right) \right) \right) \\
& \cdot \prod_{k=1}^{K} \prod_{i=1}^{n_k} \left(\mathcal{N}\left(a_{ik}, b_{ik} \middle| \begin{pmatrix} \alpha_k \\ \beta_k \end{pmatrix}, \Sigma^{(k)} \right) \right) \\
& \cdot \prod_{k=1}^{K} \left(N\left(\alpha_k \middle| \mu_\alpha, \sigma^2_\alpha\right) \cdot N\left(\beta_k \middle| \mu_\beta, \sigma^2_\beta\right) \cdot IW_3\left(\Sigma^{(k)} \middle| I_2\right) \right) \cdot \pi(\mu_\alpha) \cdot \pi(\mu_\beta) \cdot \pi(\sigma^2_\alpha) \cdot \pi(\sigma^2_\beta) \\
& \cdot \prod_{k=1}^{K} \prod_{j=1}^{6} \left(N\left(s_{jk} \middle| t_j, \tau^2\right) \cdot \pi(\sigma^2_{jk}) \right) \cdot \prod_{j=1}^{6} \left(\pi\left(t_j\right) \right) \cdot \pi(\tau^2)
\end{aligned}
$$

For derivation of the full conditional distributions, see Section S1 of the supplementary material.

1.2.3. Diagnostics and Results

The MCMC algorithm for posterior sampling was written in R (R core team, 2019) and implemented via a Gibbs sampler using the package rcpp, which greatly reduced the computational cost. The meta-analysis model was run via 4 parallel chains of 12,000 samples, with each chain needing no more than 5,000 samples to converge. Figure S1 in the supplementary material shows trace plots for our parameters of interest $(\mu_\alpha, \mu_\beta, \sigma^2_\alpha, \sigma^2_\beta)$.

For each parameter, the trace plots show that all four chains converge to the same distribution, which is further supported by the Gelman-Rubin diagnostics (median and 95% potential scale reduction factor at 1.00 for all four global parameters). Table 1.1 shows summary statistics for the posterior draws across all chains. Since we observed relatively low correlation among sequential draws, with all variables below 5% correlation for a lag of 2 (Figure S2), we thin the chains after burn-in by 1 draw, so every other sample is used to calculate the summary statistics (a total of 14,000 samples). Therefore, we accept our meta analysis prior for μ_a and μ_b to be

$$\mu_a \sim N\left(2.357, .041\right) \qquad \mu_b \sim N\left(.952, .003\right)$$

based on posterior means for μ parameters and posterior medians for σ^2 parameters.

1.2.4. Sensitivity Analysis

To confirm our choice of the prior distribution for the covariance matrix $\Sigma^{(k)}$, we conducted a simulation study and evaluated model performance in estimating the key global parameters

Parameter	1st Quartile	Median	3rd Quartile	Mean	Standard Deviation
μ_α	2.318	2.357	2.395	2.357	0.060
μ_β	0.941	0.952	0.962	0.952	0.016
σ^2_α	0.031	0.041	0.056	0.047	0.024
σ^2_β	0.002	0.003	0.004	0.003	0.002

Table 1.1: Summary statistics for global parameters in our Bayesian meta-analysis after a burn-in of 5,000 and thinning every other draw

μ_α and μ_β across various objective priors for the bivariate covariance in the literature, as summarized in Berger and Sun (2008). Seven simulation settings were included in our study, each of which contains 50 synthetic datasets. For each dataset, we estimated μ_α and μ_β using nine different covariance prior setups: Inverse-Wishart, Scaling, Half-t distribution, Prior J (Jeffery's prior), Prior IJ (Independence Jeffery's prior), Prior $R\rho$, Prior $R\sigma$, Prior $\hat{R}\sigma$, and Prior S. Details about each of these priors can be found in Berger and Sun (2008). Below we list the details of each simulation setting:

- Base: The empirical estimates of the model parameters are used as the truth.
- High Variance (HV): Parameters from the Base setting were used with the magnitudes of σ^2_α and σ^2_β being multiplied by 2.
- Low Variance (LV): Parameters from the Base setting were used with the magnitudes of σ^2_α and σ^2_β being divided by 2.
- Large Meta-Analysis (LMA, i.e., a large number of studies K): Parameters from the Base setting were used. However, the number of studies is doubled.
- Small Meta-Analysis (SMA, i.e., a small number of studies K): Parameters from the Base setting were used. However, the number of studies is halved.
- Large Sample (LS): Parameters from the Base setting were used. However, the number of patients within each study is doubled.

- Small Sample (SS): Parameters from the Base setting were used. However, the number of patients within each study is halved.

To add some more variation into our synthetic datasets, for each setting, we simulated the $\Sigma^{(k)}$'s. Specifically, the variance components of $\Sigma^{(k)}$'s were generated from Inverse-Gamma distributions whose parameters were estimated using moment matching. To simulate the correlation components, we first generated samples from a normal distribution with mean and variance based on observed correlations after Fisher's z-transformation. The variables were then transformed back to the (-1,1) scale. Additionally, we generated the variance for each probe within each study from an Inverse-Gamma distribution where the parameters were estimated using moment matching. Finally, for each study, the number of samples (n_i) was drawn with replacement from the original numbers of patients.

For each covariance setup, we ran an MCMC chain of length 2,500 and burnt in the first half of the chain. To measure the estimation performance of each method, we calculated the MSEs of the estimated posterior means of μ_α and μ_β. The Scaling method took too long to generate posterior sample, so we excluded it from the comparison. The results, scaled by a factor 1000, are summarized in Table 1.2, which shows that in each simulation setting, the MSE's under the different priors are quite close to each other. Furthermore, there is no clear-cut winner among the priors. For instance, the inverse-Wishart approach has the highest MSE in the base scenario for μ_α but has the lowest MSE in the same scenario for μ_β. These results show that the meta-analysis performance has relatively low sensitivity to the prior choice for $\Sigma^{(k)}$.

	Estimation of μ_α							Estimation of μ_β						
	Base	HV	LV	LMA	SMA	LS	SS	Base	HV	LV	LMA	SMA	LS	SS
IW	2.522	4.392	1.259	1.162	7.941	3.09	4.072	7.741	11.08	4.206	3.347	12.11	6.288	7.489
Half-t	2.456	4.411	1.253	1.171	8.187	3.067	4.097	7.989	10.63	3.898	3.296	12.06	5.712	7.208
Prior J	2.423	4.534	1.289	1.172	8.141	3.053	4.019	7.758	10.29	3.989	3.355	12.27	5.663	7.222
Prior IJ	2.408	4.473	1.278	1.15	8.171	3.087	4.024	7.975	10.01	4.014	3.193	12.41	5.776	7.437
Prior $R\rho$	2.483	4.482	1.316	1.177	8.077	3.039	4.071	7.984	10.36	3.884	3.257	12.69	5.791	6.982
Prior $R\sigma$	2.403	4.44	1.281	1.154	8.085	3.058	4.039	7.876	10.27	4.013	3.268	12.15	5.983	7.489
Prior $\hat{R}\sigma$	2.443	4.447	1.286	1.177	8.113	3.081	4.064	8.029	10.62	4.056	3.277	12.47	5.869	7.429
Prior S	2.396	4.498	1.307	1.175	8.148	3.068	4.075	7.872	10.26	4.022	3.224	11.98	5.865	7.3

Table 1.2: Analysis of sensitivity to various prior choices of the study-specific covariance matrix Σ^k: 1,000×MSE is reported to evaluate the performance of estimating μ_α and μ_β. LV stands for large variance, SV for small variance, LMA for large meta-analysis, SMA for small meta-analysis, LS for large sample, and SS for small sample.

1.2.5. Additional Enhancements

This section outlines the additional enhancements made to RCRnorm to improve model efficiency, convergence and stability. In addition to the meta-analysis prior, three changes were made to the model:

1. Implementing the constraints $\sum_{p=1}^{P} d_p = 0$, $\sum_{p=1}^{P} d_p X_p^+ = 0$, $\sum_{n=1}^{N} d_n = 0$ in the positive and negative probe equations

2. Updating λ_r, λ_h and ϕ_i with fixed calculations (i.e., no random draws) in each iteration

3. Refurbishing R code with more efficient data structures and procedures to improve computational cost

One other notable update is that MetaNorm does not allow non-randomized starting points. The remainder of this section gives details and justification for the changes listed above.

1.2.5.1. Constraints on d_p, d_n

In the presence of a high number of local parameters, implementing reasonable constraints can help to stabilize a Gibbs sampler. This is the case with RCRnorm, where random intercepts and slopes (a_i, b_i) are assumed for each sample along with probe-specific effects. In a frequentist analysis of the positive probes, two constraints would be required to achieve a unique solution. Without these constraints, there are an infinite number of solutions, reflecting the fact that predictive power could be distributed in different ways to the probe or sample effects. While this is less severe in an MCMC algorithm, where we rely on thousands of "solutions" (draws) to form a posterior distribution, convergence can sometimes suffer in such a highly complex system as described in (1.1)–(1.4), especially when the sample size I is not large. This effect can be seen in the trace plots of d_6^+ from the lung cancer FFPE dataset ($I = 28$) used for testing in Jia et al. (2019), shown in Figure 1.3. The lefthand plot shows a standard run of RCRnorm. While eventually all chains reach a similar range, it takes more than 10,000 draws for this to occur. It is also clear that there is a significant amount of autocorrelation and variability between chains, causing issues of replicability when chains are not long enough. The righthand plot shows 5 chains of RCRnorm with the constraints added. Clearly all chains are quickly converging to the same distribution, with minimal autocorrelation present. These constraints not only help to stabilize d_p^+, but also with other key model parameters including κ_{ir}.

1.2.5.2. Updating λ_h, λ_r and ϕ_i

Now we turn our attention to the housekeeping and regular genes. Currently, RCRnorm relies on "safe range" uniform prior distribution for λ_r, λ_h (the mean parameters for $\kappa_{ir}, \kappa_{ih}^*$ respectively) and ϕ_i, which is generated based on empirical estimates of a_i and b_i. Because there are a relatively large number of parameters for the information provided by the housekeeping and regular gene read counts, the samples for these parameters can oscillate

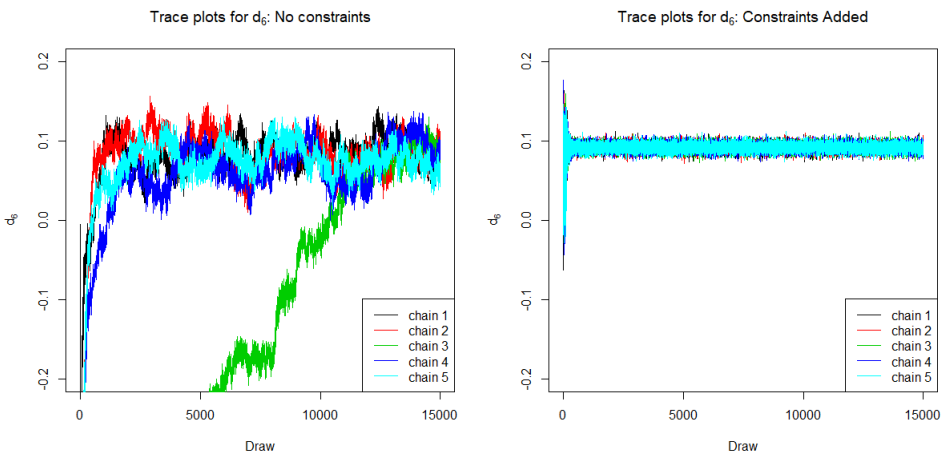

Figure 1.3: Trace plots of d_6^+ for RCRnorm with and without constraints (right and left, respectively) from the lung cancer FFPE dataset ($I = 28$) used for testing in Jia et al. (2019), showing that the convergence was greatly facilitated by the added constraints.

within the pre-defined range, failing to converge to a much narrower distribution as we may expect. Figure S3 in the supplementary material shows an example of this behavior using the ϕ_i trace plots from dataset 13 ($I = 8$).

Since the model structure is well-justified in Jia et al. (2019), we propose updating these parameters with a fixed calculation to stabilize the ϕ and λ terms, while allowing the κ terms to continue to be updated with its conditional distribution. First, we start by defining

$$\tilde{X}_{ih}^{(t)} = \frac{Y_{ih} - a_i^{(t)}}{b_i^{(t)}} \qquad \tilde{X}_{ir}^{(t)} = \frac{Y_{ir} - a_i^{(t)}}{b_i^{(t)}}$$

which is an estimate of the $\log_{10}$ RNA amount for housekeeping and regular genes given the t^{th} draw of a_i, b_i. Let j index both housekeeping and regular genes such that $j \in (1, ..., H, H+1, ..., H+R = J)$ so that $\tilde{X}_{ij}$ represents $\log_{10}$ RNA from both types of genes. For notational simplicity, we ignore the superscript t here. We propose that the linear fixed-effects model $\tilde{X}_{ij} \sim \phi_i + \lambda_j$ with the constraint $\sum_{i=1}^{I} \phi_i = 0$ be estimated in every loop of the MCMC, with the parameter estimates $\hat{\phi}_i$ and $\hat{\lambda}_j$ used as the draws for ϕ_i, λ_j. While this might seem computationally expensive, the factorial design of the data allows us to take a

simple shortcut, shown below.

$$\hat{\phi}_i = \frac{1}{J}\sum_{j=1}^{J}\tilde{X}_{ij} - \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\tilde{X}_{ij} \qquad \hat{\lambda}_j = \frac{1}{I}\sum_{i=1}^{I}\tilde{X}_{ij}$$

This method allows us to separate the effect of a_i and b_i (driven by the positive probes) so that ϕ_i can focus on the much more subtle sample effects in the housekeeping and regular gene counts. Similarly, this will allow us to obtain stable estimates of λ_h, λ_r which will in-turn stabilize the κ parameters. This will improve model convergence and allow the MetaNorm output to be significantly more precise (as shown in section 1.3), increasing its value and reliability to researchers.

1.3. Application and Comparison to RCRnorm

We compare the performance and diagnostics of MetaNorm to that of RCRnorm in 5 different areas, including computation time, convergence, stability of estimates, model bias and performance with low-quality datasets. We will use four different datasets of varying quality, three that were excluded from the meta-analysis and one with a low number of samples. These datasets are summarized in Table 1.3. Dataset 4 is the lung cancer FFPE dataset used in Jia et al. (2019) and contained in the RCRnorm R package. It contains 28 samples and 104 genes after cleaning, and is thought to be a high-quality dataset. Dataset 12 was excluded from the meta-analysis due to a high number of low-quality samples (low R-squared in the positive probe data; see footnote 3) and dataset 15 was excluded for having a very high positive correlation between a_i and b_i (see Figure 1.2(c)). Despite being included in the meta-analysis, Dataset 13 is also used for testing due to its low number of samples and high number of genes.

Dataset ID	Description	# of Samples	P	N	H	R	Total
4	Lung cancer FFPE	28	6	8	7	83	2,912
12	Triple negative breast cancer mRNA signatures	253	6	8	5	231	63,250
13	Merkel cell carcinoma samples	8	6	8	40	730	6,272
15	Metastatic Melanoma FFPE samples	24	6	6	12	204	5,472

Table 1.3: Characteristics of testing datasets

1.3.0.1. Computation Time

One of the most significant drawbacks to RCRnorm is the computation time. In a comparison study observing NanoString normalization techniques, Bhattacharya et al. (2021) cite this high computation cost, noting that RCRnorm could not be used to normalize a dataset with $> 1,000$ samples efficiently. After restructuring the algorithm and implementing more efficient data-structures, MetaNorm significantly improves the computation time. Table 1.4 shows a breakdown of computation time for different datasets and samples in R-studio using a laptop with an i7 core and 16 GB of RAM. The results show that MetaNorm outperforms RCRnorm in computational efficiency, with a minimum improvement of 6-fold. But the effects of MetaNorm is seen most clearly when the number of samples is large ($I > 20$), with RCRnorm taking 25 to 40 times longer to produce the same amount of samples. In addition to this, MetaNorm tends to reach convergence quickly with little autocorrelation, usually needing under 500 draws to reach a stationary distribution and leading to improvement of another 5-fold at least. So MetaNorm improves upon RCRnorm by being faster per-draw but also by requiring fewer draws, leading to an increase an efficiency by anywhere from 30-fold for low-sample datasets to 200-fold for large-sample datasets.

1.3.0.2. Convergence

Part of assessing the performance of a stochastic process such as an MCMC is observing how well it converges to a stationary distribution when initiated from diverse starting points. When starting from reliable estimates from data, RCRnorm shows quick and stable

Dataset ID	Average Computation Time (seconds)		MetaNorm
	RCRnorm	MetaNorm	Improvement
4	96.80	2.29	42.2
12	1340.60	39.39	34.03
13	55.86	9.26	6.03
15	90.88	1.78	25.60

Table 1.4: Comparison between RCRnorm and MetaNorm on computation time using four real datasets (1,000 draws per run)

convergence of high-level parameters such as μ_a, μ_b. However when the starting points are randomized, convergence weakens, and occasionally a chain will fail to stabilize altogether. This section demonstrates that MetaNorm resolves this issue, including cases where RCRnorm produces a "faux" convergence. All datasets were run with 5 chains of 15,000 draws each except for dataset 12, which we limited to 5 chains of 5,000 draws each due to the size of the dataset.

Figure S4 in the supplementary material shows boxplots of post burn-in draws of μ_a broken down by chain. The posterior distributions for both datasets 4 and 15 differ by chain for RCRnorm (not much but visibly), whereas the posterior distributions for MetaNorm are practically identical. This is confirmed by the convergence diagnostics shown in panel (a) of Table 1.5, where RCRnorm has decent convergence for dataset 4 and (to a lesser extent) dataset 15 but MetaNorm produces better convergence in either scenario. The corresponding trace plots for this example are shown in figure S5 of the supplementary material, showing that MetaNorm converges much faster than RCRnorm with merely a few hundreds of draws. Panel (a) of Table 1.5 also shows that datasets 12 and 13 achieve good convergence for global parameters regardless of which model is used. However the next level of the hierarchy tells a different story. Figure S6 in the supplementary material shows trace plots of a_1, a_2, a_3 from the first three patients for the RCRnorm (top) and MetaNorm (bottom) chains. Clearly, RCRnorm is not converging to the same distribution in every chain, which is a significant

	(a) Gelman-Rubin Diagnostics for μ_a				(b) Median Standard Deviation of κ_{ir}s	
Dataset ID	RCRnorm		MetaNorm		RCRnorm	MetaNorm
	Median	Upper C.I. (95%)	Median	Upper C.I. (95%)		
4	1.05	1.12	1.00	1.00	0.037	0.009
12	1.00	1.00	1.00	1.00	0.205	0.003
13	1.02	1.06	1.00	1.00	0.961	0.015
15	1.22	1.52	1.00	1.00	0.092	0.007

Table 1.5: Comparison between RCRnorm and MetaNorm on (a) convergency using Gelman-Rubin diagnostics for μ_a and (b) median standard deviation of κ_{ir}s.

issue. MetaNorm resolves this issue, with all 8 samples converge in their a_i estimates. This problem also exists in dataset 12 (results not reported due to the space limit). As a result, the normalized gene read counts vary significantly between runs with different starting points, casting doubt on our results (see panel (b) of Table 1.5). The issues raised in this section indicate that RCRnorm suffers from a kind of "faux" convergence, where good convergence around global parameters masks weak or absent convergence in lower-level parameters. It should be noted that this does not occur in all datasets when using RCRnorm - such as dataset 4 - but (to our knowledge) does not occur at all when using MetaNorm.

1.3.0.3. Stability of κ_{ir} Estimates

Perhaps the most critical piece of both the RCRnorm and MetaNorm models are the κ estimates for housekeeping and regular genes. Representing the RNA amount after accounting for sample degradation, the κ parameters are the true model output and any value that the models offer lies in these estimates. With this in mind, if we run separate chains on the same dataset, ideally we will observe very similar estimates for these parameters. We will once again show that MetaNorm improves upon RCRnorm in this way, providing more stable estimates for these parameters. Panel (b) of Table 1.5 shows the median standard deviation between chains among all κ_{ir} estimates. For datasets 12, 13 and 15, which do not have all chains of sample-specific parameters converging to the same distribution, the stark

difference in standard deviation should not be surprising; if the a_i and b_i terms are different, of course the κ_{ir} terms would be wildly different. But even when the data is good-quality and convergence is somewhat stable (as with dataset 4), there is still significant reduction in standard error in MetaNorm. The left panel of Figure 1.4 shows the densities of these standard errors.

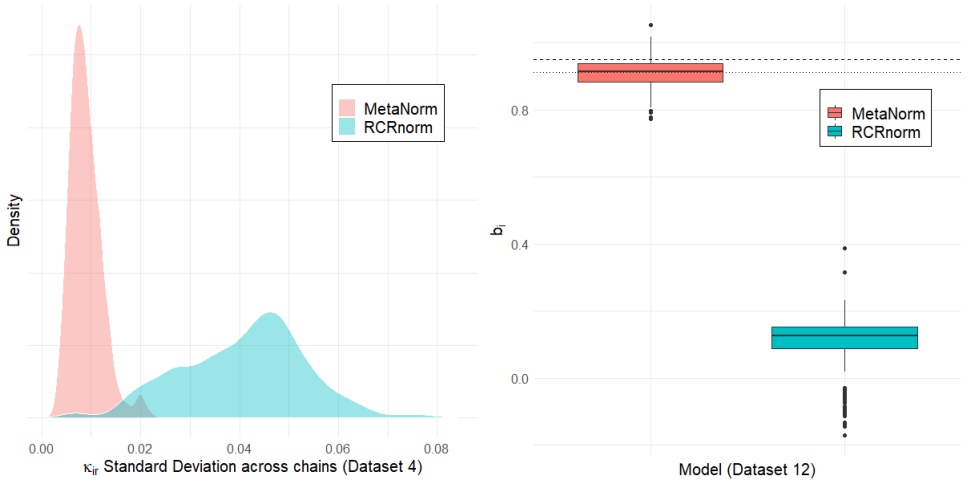

Figure 1.4: Comparison between RCRnorm and MetaNorm: the left panel shows the density of κ_{ir}'s standard errors for dataset 4; the right panel shows boxplots of b_i estimates for dataset 12, where the dashed line represents μ_β, the meta-analysis estimate for the prior mean of μ_b, and the dotted line represents the empirical estimate of μ_b, using the positive probe data.

1.3.0.4. Bias of κ_{ir} Estimates

A small simulation study was performed to compare model bias of the κ_{ir} estimates between RCRnorm and MetaNorm. Data for the simulation was generated based on dataset 4 (lung cancer FFPE), under constraints $\sum_{p=1}^{P} d_p = 0$, $\sum_{p=1}^{P} d_p X_p^+ = 0$, $\sum_{n=1}^{N} d_n = 0$. Bias was estimated using average mean difference between actual κ_{ir} values that the data was generated from and κ_{ir} estimates from the normalization procedures. In 100 runs of RCRnorm, we found that the normalization produced output that was off by an average of -.0651 per κ_{ir}. This translates to an average 13.9% reduction from actual to estimated

normalized gene read counts. In 100 runs of MetaNorm, we found that the output differed by an average of .0018 per κ_{ir}. This translates to an average increase of 0.4% from actual to estimated normalized gene read counts. While this does show that MetaNorm has superior bias to RCRnorm, this result is not as significant as the reduction in prediction variance (especially as it pertains to mean-squared error contribution).

1.3.0.5. Performance with Messy Data

We conclude the discussion of model improvements with a specific example of MetaNorm producing more intuitive and interpretable estimates for data with a large number of low quality samples. Dataset 12 contains 61 (24%) samples which have an $R^2 < .95$ from a linear regression of positive probe data. Since the nCounter system is designed to maintain very strong linear relationship between the log-transformed counts and log-transformed control RNA expression levels, NanoString's recommendation is to discard any samples with $R^2 < .95$. In this section, we included all 253 samples for testing. Three chains of 10,000 draws were run on each model, with the first 5,000 discarded as burn-in. The estimates for b_i (the sample-specific slope terms) were drastically different, as the right panel of Figure 1.4 shows, despite the fact that the jackknife prior implemented in RCRnorm keeps the μ_b samples from deviating off the empirical estimate. While the MetaNorm estimates are within a reasonable neighborhood of μ_b, the RCRnorm estimates are much lower, with a significant portion of them (˜34%) less than 0. As a point of comparison, empirical estimates for b_i from dataset 12 have a minimum of .829. While some variation from the empirical estimates is expected, the amount seen from the RCRnorm is counter-intuitive both to the original data and to the nCounter system in general. Alternatively, MetaNorm is able to overcome the pitfalls of messy data and produces reasonably intuitive estimates for b_i. These estimates have a downstream effect, causing the estimates for κ_{ir} to become too large when b_i approaches 0 and become too small when b_i dips below 0 (see figure S7 in supplementary material). For

this reason, the κ_{ir} estimates from MetaNorm are more plausible than those from RCRnorm for this dataset.

1.4. Discussion

When conducting statistical analysis in the Bayesian framework, many researchers rely on non-informative or diffuse prior distributions to maintain objectivity. However, when meaningful prior information exists and can be identified, using an informative prior distribution to accurately reflect current knowledge may lead to superior outcomes and great efficiency. This paper demonstrates a practical example of achieving improvement by incorporating such prior information via meta-analysis into data normalization.

RCRnorm is a groundbreaking normalization tool that improved upon existing methodology by leveraging the complexities of nCounter data. While other methods rely on frequentist techniques, the Bayesian framework allows RCRnorm to thoroughly explore the parameter space, identifying hidden information in the data, casting aside implausible assumptions, and increasing model transparency and interpretability. But despite its improvements, RCRnorm struggles to provide reliable estimates when the quality of data is in question. One reason for this is the "non-informative" data-based prior implemented for two of the most important global parameters μ_a and μ_b, potentially reinforcing bias stemming from the data. The goal of this study was to enhance the RCRnorm model by (1) identifying prior distributions for μ_a and μ_b based on a comprehensive meta-analysis of FFPE datasets and (2) implementing additional algorithmic enhancements to improve computational cost, model convergence and stability of estimates. This new Gibbs sampler is called MetaNorm.

The meta-analysis employs a Bayesian hierarchical regression model of similar structure to RCRnorm. Since almost all of the information about a_i and b_i come from the positive probe read counts, the meta-analysis only considers these 6 observations per sample (i.e., negative probe and gene data are ignored). Conjugate hierarchical structures are used

throughout the meta-analysis, allowing for straightforward conditional distributions. To ensure quick model convergence, constraints are implemented on the probe/study effects s_{jk} as well as the means of these parameters t_j. The estimates obtained from the meta-analysis are used as informative priors in MetaNorm, which serve as guide-rails for μ_a and μ_b. In addition to the implementation of the new priors, we made 3 other changes to the model. The first, was the enforcement of constraints on d_p^+ and d_n^-. These constraints mitigate the identifiability concerns while also implementing assumptions we associate with regression residuals (i.e., completely separating the probe effect from the additive sample effect and the mRNA sample effect). For similar reasons, ϕ_i, λ_r,and λ_h are updated using empirical estimates in MetaNorm, based on updated samples of a_i and b_i. This helps to separate information into sample and gene-specific buckets provided by the regular and housekeeping gene read counts, mitigating the effect of having more parameters than actual observations for these genes. Finally, we restructured the code and used more efficient functions and processes to improve computation, which led to a significant reduction in run time. Additionally, all datasets tested on MetaNorm for this manuscript converged in less than 1,000 draws, suggesting that smaller chains can be used for MetaNorm. We also showed that MetaNorm leads to more consistent convergence in global parameters while ensuring that sample-specific parameters (a_i, b_i) reliably converge to the same distribution from diverse starting points. We also observed reduced bias and variance in normalized κ_{ir} estimates then those produced by RCRnorm. Finally, a dataset of triple negative breast cancer mRNA signatures shows that MetaNorm continues to produce realistic and intuitive parameter estimates when data-quality is a concern, while RCRnorm falters.

One area where both RCRnorm and MetaNorm can be improved is the computational load. The downside to MCMC approaches in general is the amount of additional computation time needed to individually sample different parameters one after another. However beyond that is the amount of RAM (random access memory) needed to store thousands of samples for (potentially) 100,000+ parameters. Bhattacharya et al. (2021) describes this issue when

trying to test RCRnorm against other normalization approaches. When the amount of samples and/or genes is high (1,278 samples, in their case), there is simply not enough RAM in standard computers to store the chains for all parameters. A potential remedy for this is to add options to the algorithm to not store and track all draws for local parameters, but instead to keep the most recent draw as well as calculating posterior means in a piecemeal fashion after burn-in. This would keep the algorithm from inefficiently storing hundreds of thousands of chains. We hope to implement such a change to MetaNorm in the near future. Additionally, estimating the posterior distribution with variational inference rather than MCMC would further improve computational efficiency and is a topic of future research (Blei et al., 2017).

Chapter 2

A Meta-analysis based Hierarchical Variance Model for Powering One and Two-sample t-tests

2.1. Introduction

One of the most important steps in designing a clinical trial is determining the number of patients (n) to recruit for the study. Underestimating n can result in an underpowered study where any effect, no matter how clinically relevant, is unlikely to yield a statistically significant result, leading to financial and ethical consequences (Halpern et al., 2002). However, being overly conservative and recruiting more patients than needed comes with its own financial and ethical costs, wasting capital and subjecting an excessive number of patients to (potentially risky) experimental treatments (Hochster, 2008). Moreover, the importance of the targeted effect size should not be overlooked either, as small effect sizes may be practically insignificant yet require larger sample sizes to detect. Therefore, it is crucial to determine an accurate sample size for a clinically relevant effect to mitigate these issues and put researchers in the best position to succeed. Power analysis provides a mechanism for researchers to justify and calculate n based on a set of specific criteria.

For a one or two-sample t-test, traditional power analysis is determined solely by point estimates of the anticipated variance and mean (or mean difference). While some methodology exists to help researchers identify reasonable values for variance and effect size based on historical data (Lenth, 2001; Liu, 2010; Santis, 2007), the resulting estimates are, at best, educated guesses based on previous work or estimates from relevant literature. While this is a reasonable approach when done responsibly, a lack of scrutiny given to sample size justification gives way to "power-hacking", a practice where researchers start with a

specific value for n and then hunt for effect sizes and variance estimates in the literature to retroactively justify it. An expensive alternative to this is to conduct a pilot study, using the data collected to estimate the variance and effect size. Despite the general benefits of pilot studies when planned correctly (Lancaster et al., 2004; Moore et al., 2011), the risks of using these values as direct stand-ins for a power calculation while ignoring uncertainty is well-documented (Kraemer et al., 2006). More complex procedures exist, including both Bayesian and frequentist approaches to incorporate uncertainty in the variance estimate from a pilot study (Browne, 1995; Sims et al., 2007; Shieh, 2017).

Approaches that account for uncertainty in effect size or variance are not limited to pilot studies, where most leverage the Bayesian paradigm to take in a distribution of plausible values rather than singular estimates. While most of these methods are designed to power Bayesian analyses, typically by controlling the width/coverage rate of credibility intervals (Cao et al. 2009 and references therein) or the type II error of a Bayes factor test (Weiss, 1997; Santis, 2004), the Bayesian framework is well-suited for power analysis of frequentist studies (Spiegelhalter and Freedman, 1986; Joseph et al., 1997). Rather than clashing ideals, this approach allows researchers to use both methods in harmony (Bayarri and Berger, 2004), leveraging the flexibility of Bayesian methods to account for the significant levels of uncertainty encountered in sample size determination, while using the more traditional and (in most cases) conservative frequentist methodology for the actual analysis. Perhaps the best example of this is the "what-if" prior discussed in Gelfand and Wang (2002), a simulation-based approach where different prior distributions are used for sample size determination (reflecting what we expect/hope) and the actual analysis (reflecting what we actually know). Thus, if researchers have an idea of where a parameter might lie, the Bayesian approach can be implemented with a prior that reflects this idea to design the study, while the actual analysis makes no such assumptions to remain unbiased. Several other approaches exist as well, including leveraging historical data, or even meta-analysis, to determine prior distributions for effect sizes (Du and Wang, 2016). While information

obtained from meta-analysis should yield more accurate estimation, these approaches lead to a different issue: if the effect size distribution crosses into "clinically irrelevant" territory, the target sample size will then increase to account for these smaller effect sizes, inflating the cost of the study. In this paper, we propose a new methodology for sample size determination that leverages historical information from multiple studies to construct a distribution for the variance based on an empirical Bayes (EB) approach while keeping the unstandardized effect size fixed to ensure practically significant results. This methodology is based on the Bayesian paradigm and uses a hierarchical gamma-inverse gamma model to allow the variance to vary. To the knowledge of the authors, this is the first outlined methodology that uses multiple studies from the literature to model the variability of the variance estimate alone, reducing the bias of sample size determination from one particular study while guaranteeing practical significance. Compared to approaches that recommend pilot studies, it requires much less time and fewer resources.

In Section 2.2, we outline our method in detail, exploring both the hierarchy and the simulation approach to calculating the target sample size. In Section 2.3, we evaluate the performance of our method against alternative approaches using a simulation study. In Section 2.4 we demonstrate our method using real data examples and provide discussion and future directions in Section 2.5. Code used to implement the algorithms described in this paper can be found at https://github.com/jbarth216/Bayesian-Meta-SSD.

2.2. Methodology

In this section, we introduce our innovative sample size determination approach and explain the underlying theory and rationale. Our goal is to develop a method that can effectively synthesize and utilize information from multiple studies instead of relying solely on aggregate point estimates for the variance. When variances from different studies are exchangeable, our proposed Bayesian method, detailed in this section, is able to capture more information about the variability than a simple point estimate. To calculate the power

of a study with a given sample size, we employ a simulation-based procedure based on the variance distribution with parameters estimated using a formal EB approach, which can later be simplified to a discretized probability mass function to facilitate computation.

2.2.1. One-sample t-tests

Assume we have $X_1, ...X_n$ from a normal population $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Furthermore, assume that there are k studies that are related to the study of interest, with each study having an observed sample variance $s_1^2 ... s_i^2 ... s_k^2$ which, for the sake of notational simplicity, we will call $y_1, ..., y_i, ...y_k$. Note that the studies included in the meta-analysis should be carefully chosen based on the response variable and the specific population being studied. Since each of these studies assume a normal distribution, classical theories teach that for study i,

$$y_i|\theta_i \sim \text{Gamma}(\frac{n_i - 1}{2}, \frac{n_i - 1}{2\theta_i})$$

where $\theta_i \equiv \sigma_i^2$, and the first parameter of the Gamma distribution is the shape parameter and the second is the rate parameter. We assume *a priori* that

$$\theta_1, ..., \theta_k, \tilde{\theta} \sim \text{IG}(\alpha, \beta),$$

where the variance of the study of interest, $\tilde{\theta}$, is exchangeable with respect to those of the meta-analysis, and the inverse-gamma (IG) distribution with shape parameter α and scale parameter β is used to achieve conjugacy. Let $\boldsymbol{\theta} = (\theta_i)_{i=1}^k$ and $\boldsymbol{y} = (y_i)_{i=1}^k$. Obtaining the distribution of $\tilde{\theta}|\boldsymbol{y}$ entails averaging $p(\tilde{\theta}|\alpha, \beta)$ over all possible values of (α, β) according to

their posterior distribution $p(\alpha, \beta|\boldsymbol{y})$, namely

$$\begin{aligned} p(\tilde{\theta}|\boldsymbol{y}) &= \int\int p(\tilde{\theta}|\alpha,\beta)p(\alpha,\beta|\boldsymbol{y})d\alpha d\beta \\ &= \int\int\int p(\tilde{\theta}|\alpha,\beta)p(\boldsymbol{\theta},\alpha,\beta|\boldsymbol{y})d\boldsymbol{\theta}d\alpha d\beta \\ &\propto \int\int\int p(\tilde{\theta}|\alpha,\beta)p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha,\beta)p(\alpha,\beta)d\boldsymbol{\theta}d\alpha d\beta, \end{aligned}$$

where $p(\alpha, \beta)$ is the hyper-prior distribution on (α, β). Although the distribution of $\boldsymbol{\theta}|\boldsymbol{y}, \alpha, \beta$ is known under the conjugate structure, $p(\tilde{\theta}|\boldsymbol{y})$ is not analytically tractable. Thus, we employ an empirical Bayes approach; that is, we find the estimates of (α, β) using the observed data, say $(\hat{\alpha}, \hat{\beta})$, by maximizing the marginal likelihood (MML) $\ell(\alpha, \beta; \boldsymbol{y}) \equiv p(\boldsymbol{y}|\alpha, \beta)$, and then use $p(\tilde{\theta}|\hat{\alpha}, \hat{\beta})$ instead. In this way, we also avoid the need for specifying $p(\alpha, \beta)$, which is a nontrivial task.

We first identify the joint distribution $\boldsymbol{\theta}, \boldsymbol{y}|\alpha, \beta$ and then integrate out $\boldsymbol{\theta}$. We have

$$\begin{aligned} p(\boldsymbol{y}|\alpha,\beta) &= \int_{\boldsymbol{\theta}} p(\boldsymbol{y},\boldsymbol{\theta}|\alpha,\beta)d\boldsymbol{\theta} \\ &= \int_{\theta}\prod_{i=1}^{k}\left[\frac{(\frac{n_i-1}{2\theta_i})^{\frac{n_i-1}{2}}}{\Gamma(\frac{n_i-1}{2})}y_i^{\frac{n_i-3}{2}}e^{-y_i\frac{n_i-1}{2\theta_i}}\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\theta_i^{-\alpha-1}e^{\frac{-\beta}{\theta_i}}\right]d\boldsymbol{\theta} \end{aligned}$$

which reduces to

$$p(\boldsymbol{y}|\alpha,\beta) = \prod_{i=1}^{k}\left[h_i\cdot\frac{\Gamma(\alpha+\frac{n_i-1}{2})}{(\beta+\frac{y_i(n_i-1)}{2})^{\alpha+\frac{n_i-1}{2}}}\frac{\beta^\alpha}{\Gamma(\alpha)}\right], \qquad h_i = \frac{(\frac{n_i-1}{2})^{\frac{n_i-1}{2}}}{\Gamma(\frac{n_i-1}{2})}y_i^{\frac{n_i-3}{2}}.$$

Now we can solve for the MML estimates of α, β in terms of our observed data $\boldsymbol{y}$. The log likelihood $\ell\ell(\alpha, \beta; \boldsymbol{y})$ is shown in (2.1) and its partial derivatives are shown in (2.2) and (2.3).

$$\ell\ell(\alpha, \beta; \boldsymbol{y}) = \sum_{i=1}^{k} [\log(h_i) + \alpha \log(\beta) - \log(\Gamma(\alpha)) + \log(\Gamma(\alpha + \frac{n_i - 1}{2})) - (\alpha + \frac{n_i - 1}{2}) \log(\beta + \frac{y_i(n_i - 1)}{2})] \tag{2.1}$$

$$\frac{d\ell\ell}{d\alpha} = k(\log(\beta) - \psi(\alpha)) + \sum_{i=1}^{k} [\psi(\alpha + \frac{n_i - 1}{2}) - \log(\beta + \frac{y_i(n_i - 1)}{2})] \tag{2.2}$$

$$\frac{d\ell\ell}{d\beta} = k(\frac{\alpha}{\beta}) - \sum_{i=1}^{k} \frac{2\alpha + n_i - 1}{2\beta + y_i(n_i - 1)} \tag{2.3}$$

Setting the derivatives to zero, we get the following set of equations, shown in (2.4) and (2.5).

$$k(\log(\hat{\beta}) - \psi(\hat{\alpha})) + \sum_{i=1}^{k} [\psi(\hat{\alpha} + \frac{n_i - 1}{2}) - \log(\hat{\beta} + \frac{y_i(n_i - 1)}{2})] = 0 \tag{2.4}$$

$$\frac{k\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^{k} \frac{2\hat{\alpha} + n_i - 1}{2\hat{\beta} + y_i(n_i - 1)} = 0. \tag{2.5}$$

Then 2.5 can be re-written as

$$\frac{k\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^{k} \frac{2\hat{\alpha}}{2\hat{\beta} + y_i(n_i - 1)} - \sum_{i=1}^{k} \frac{n_i - 1}{2\hat{\beta} + y_i(n_i - 1)} = 0 \rightarrow \hat{\alpha} = \frac{\sum_{i=1}^{k} \frac{n_i - 1}{2\hat{\beta} + y_i(n_i - 1)}}{\frac{k}{\hat{\beta}} - 2\sum_{i=1}^{k} \frac{1}{2\hat{\beta} + y_i(n_i - 1)}}$$

which can be substituted into (2.4) for $\hat{\alpha}$, and then a simple numerical root finder can do the heavy lifting.

Using the above MML approach, we can sample the variance of our prospective study $\tilde{\theta} = \sigma^2$ from $IG(\hat{\alpha}, \hat{\beta})$. For one-sample, one-sided t-tests, the hypotheses of interest are of

the form

$$H_0 : \mu = \mu_0, \qquad H_a : \mu > \mu_0.$$

Without loss of generality, we will for now assume that $\mu_0 = 0$. Because σ^2 is unknown under H_0, the well-known t-test statistic is

$$T = \frac{\bar{x}}{\sqrt{s^2/n}} \sim t_{n-1} \tag{2.6}$$

while under the alternative hypothesis, this statistic follows a non-central t distribution with $n-1$ degrees of freedom and non-centrality parameter $\lambda = \sqrt{\frac{n}{\tilde{\theta}}}\mu$.

In calculating the power, we assume a specific value of μ, representing a minimum practically significant result, while the variance σ^2 follows the distribution determined by the meta-analysis. By definition, if a random variable z follows a standard normal distribution and (independently) v follows a chi-squared distribution with n_v degrees of freedom, then the random variable

$$T = \frac{z + \lambda}{\sqrt{v/n_v}}$$

follows a non-central t distribution with n_v degrees of freedom and non-centrality parameter λ. When this is applied to a sample from a normal distribution $X_1, ..., X_n$ with mean μ and variance σ^2 (as is done in this setting), the components of this non-central t are

$$v = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n_v=n-1}, \qquad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \to \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} + \lambda}{\sqrt{\frac{s^2}{\sigma^2}(n-1)/(n-1)}} = \frac{\bar{x} - \mu + \lambda\frac{\sigma}{\sqrt{n}}}{\sqrt{s^2/n}}$$

This becomes 2.6 when $\lambda = \mu\frac{\sqrt{n}}{\sigma}$. Note that this is the conditional distribution of T, since the variance is generally assumed to be fixed; that is, $T|\tilde{\theta} = \sigma^2 \sim t_{n-1}(\lambda = \mu\frac{\sqrt{n}}{\sigma})$. Finding the marginal distribution gets messy very quickly, since the distributions of both $\bar{x}$ and s^2 depend on σ^2. However, since software packages can generate random variables from both

Gamma and non-central t distributions, a simulation approach based on sequential sampling can be easily implemented to obtain the marginal distribution of T, outlined below.

2.2.1.1. SSD algorithm

The general framework for the algorithm to search for the minimum sample size to achieve a certain power $1-\beta^t$ given α^t for a one-sided test is shown below. Note that α^t and β^t represent the desired level of Type I and II error for a t-test under study, respectively, while $\hat{\alpha}$ and $\hat{\beta}$ refer to the MME estimator for the shape and scale parameter discussed previously. When considering a two-sided test, the same algorithm can be used with α^t as half of the Type I error rate. This does ignore a miniscule amount of probability on the opposite side of the fixed effect size, as commonly done in power analysis.

1. Generate M draws $(\tilde{\theta}_j)_{j=1}^{M}$ from $\tilde{\theta} \sim IG(\hat{\alpha}, \hat{\beta})$

2. For a hypothetical sample size $n^{(i)}$(i denotes the iteration number, so $i = 1$ in the first iteration), calculate $R^{(i)} = F^{(-1)}_{n^{(i)}-1}(1-\alpha^t)$, where $F^{(-1)}_{n^{(i)}-1}$ is the inverse CDF of a t-distribution with $n^{(i)}-1$ degrees of freedom. This is the edge of the rejection region for the test.

3. Estimate the power, $P = Pr(T \geq R^{(i)})$, recognizing that

$$\begin{aligned} Pr(T \geq R^{(i)}) &= \int_{\tilde{\theta}} \left[Pr(T \geq R^{(i)} | \tilde{\theta}) \cdot p(\tilde{\theta} | \hat{\alpha}, \hat{\beta}) \right] d\tilde{\theta} \\ &\approx \frac{1}{M} \sum_{j=1}^{M} Pr(T \geq R^{(i)} | \tilde{\theta}_j) = \frac{1}{M} \sum_{j=1}^{M} \left[1 - F_{T|\tilde{\theta}_j}(R^{(i)}) \right] \end{aligned} \tag{2.7}$$

   where $F_{T|\tilde{\theta}_j}$ is the CDF of a non-central t-distribution with $n^{(i)}-1$ degrees of freedom and non-centrality parameter $\lambda = \mu\sqrt{\frac{n^{(i)}}{\tilde{\theta}_j}}$. Note that μ is the pre-specified clinically relevant effect size.

4. If $P \geq 1 - \beta^t$, stop, and accept $n^{(i)}$ as the target sample size. Otherwise, let $n^{(i+1)} = n^{(i)} + 1$ and return to step 2.

An initial value $n^{(1)}$ must be specified. At minimum this can be 2, but higher starting values will lead to quicker computation. An alternate approach to this is to set $n^{(1)}$ to the sample size using an aggregate approach to variance estimation (i.e., the weighted variance approach explained in Section 2.3.2), and then iterate n up or down depending on the level of power it achieves. The approximation used in step 3 of the algorithm becomes more accurate as M gets larger, but the computation cost also increases greatly as M gets larger. In Section 2.3.1 we explore the accuracy of P for different levels of M and evaluate different methods of sampling $\tilde{\theta}$. As outlined in Section 2.3.1, the recommended methodology is remarkably accurate, consistent and fast.

2.2.2. Two-sample t-tests with homogeneity of variance

We now extend the methodology to the much more common two-sample case. Assume that we have normal data from two groups for comparison, $W_1, ..., W_n \sim N(\mu_w, \sigma_w^2)$ and $V_1, ...V_m \sim N(\mu_v, \sigma_v^2)$. If it can be assumed that $\sigma_w^2 = \sigma_v^2 = \sigma^2$, then this reduces almost entirely to the one-sample case, with a slightly different set of constraints . When testing the hypothesis $H_0 : \mu_w - \mu_v = 0$ vs. $H_A : \mu_w - \mu_v > 0$, the test-statistic used is given below:

$$T = \frac{\bar{w} - \bar{v}}{\sqrt{\tilde{y}(\frac{1}{n} + \frac{1}{m})}}$$

Under the null hypothesis, this follows a t-distribution with $m + n - 2$ degrees of freedom. In reality, this follows a non-central t with the same degrees of freedom and non-centrality parameter $\lambda = (\mu_w - \mu_v)/\sqrt{\tilde{\theta}(\frac{1}{n} + \frac{1}{m})}$.

2.2.2.1. *SSD Algorithm*

The algorithm used here is very similar to the one-sample case, since only one variance needs to be estimated and the test statistic follows a similar distribution. Since there are two samples, we will assume a fixed allocation ratio for all two-sample problems, so $m = \zeta n$.

1. Generate M draws $(\tilde{\theta}_j)_{j=1}^{M}$ from $\tilde{\theta} \sim IG(\hat{\alpha}, \hat{\beta})$

2. For a hypothetical sample size $n^{(i)}$(and $m^{(i)} = \zeta n^{(i)}$), calculate $R^{(i)} = F_{df^{(i)}}^{(-1)}(1-\alpha^t)$, where $F_{df^{(i)}}^{(-1)}$ is the inverse CDF of a t-distribution with $df^{(i)} = n^{(i)} + m^{(i)} - 2$. This is the rejection region for the test.

3. Estimate the power, $P = Pr(T \geq R^{(i)})$ using ((2.7)) in Section (2.2.1), where $F_{T|\tilde{\theta}_j}$ is the CDF of a non-central t-distribution with $n^{(i)}+m^{(i)}-2$ degrees of freedom and non-centrality parameter $\lambda = d/\sqrt{\tilde{\theta}_j(\frac{1}{n^{(i)}} + \frac{1}{m^{(i)}})}$. Note that $d = \mu_w - \mu_v$ is the pre-specified clinically relevant difference.

4. If $P \geq 1-\beta^t$, stop, and accept $n^{(i)}, m^{(i)}$ as the target sample sizes. Otherwise, let $n^{(i+1)} = n^{(i)} + 1$ and return to step 2.

2.2.3. Two-sample t-test with heterogeneous variance

Let $\boldsymbol{\sigma}_i^2 = (\sigma_{w.i}^2, \sigma_{v.i}^2)$, where $\sigma_{w.i}^2$ and $\sigma_{v.i}^2$ denote the variance of each group in study i in the meta-analysis, for $i = 1, ..., k$. Recall that in the one-sample case, we assume σ_i^2s are exchangeable. If $\sigma_{w.i}^2 = \sigma_{v.i}^2$ cannot be assumed, there are three alternative approaches that can be taken to deal with the variances:

1. Assume that all the variances including both $\sigma_{w.i}^2$'s and $\sigma_{v.i}^2$'s are exchangeable.

2. Assume that $\sigma_{w.i}^2$'s are exchangeable and $\sigma_{v.i}^2$'s are exchangeable. Further, for each study i, $\sigma_{w.i}^2$ and $\sigma_{v.i}^2$ are independent, and are generated from different (IG) distributions

3. Assume that $\boldsymbol{\sigma}_i^2$'s are exchangeable. Further, for each study i, $\sigma_{w.i}^2$ and $\sigma_{v.i}^2$ have pair-wise correlation, and are generated from a multivariate distribution with $\rho > 0$, typically.

It should be noted that (1) and (2) reduce to the same variance case for identifying the distribution of $\tilde{y}$, the main difference being that we perform this twice in (2), using the meta-analysis results from the one-sample case. (3) is a much more complicated problem, and will not be covered in this manuscript. It should be noted, however, that in cases where the sample sizes are equal between two groups, the pooled-variance two-sample t-test maintains strong power even when the true variances differ (Moser and Stevens, 1992).

2.2.4. Stratified sampling with discretization

The SSD algorithm described in Section 2.2.1 produces consistent sample size estimates when M, the number of draws from $IG(\hat{\alpha}, \hat{\beta})$, is large. But when the effect size is small (and the corresponding sample size is larger), the computation time increases exponentially as M moves from 1,000 to 10,000 to 100,000 etc. While this is a trade-off frequently seen in computational statistics, it is a major obstacle to our approach, as more common SSD methods can be quickly and systematically calculated. One potential remedy to this trade-off is to consider a stratified sampling approach, which is able to better capture the overall variability of a distribution with a smaller number of samples. Suppose one wishes to draw 100 samples of a random variable X with CDF $F_X(x)$ and support $(0, \infty)$. Rather than allowing each and every draw to come from any part of the distribution of X, we instead draw each sample i from the i^{th} 1% of the distribution. In other words, the first draw is forced to be between $\big(0, F_X^{-1}(.01)\big)$, the second between $\big(F_X^{-1}(.01), F_X^{-1}(.02)\big)$, and so on until the last is between $\big(F_X^{-1}(.99), \infty\big)$. This guarantees that each draw comes from a different sub-range. Since power analysis is less concerned with the randomness properties and more concerned with a consistent and representative sample, we take this a step further

by accepting the median of each range (i.e., the first draw of X would be $F_X^{-1}(.005)$). In this way, we represent the distribution of $\tilde{y}$ as an approximate discrete distribution rather than a continuous distribution. In Section 2.3.1 we show that this method has the speed of a low-M completely random simulation with the accuracy and stability of a high-M simulation.

As to the SSD algorithm, we only need to replace Step 1 with a step of stratified sampling with discretization: compute $\tilde{\boldsymbol{\theta}} \equiv (\tilde{\theta}_i)_{i=1}^M = \left\{F_\theta^{-1}(p) : p = v_i\right\}_{i=1}^M$, where F_θ^{-1} is the inverse CDF of $\theta|\hat{\alpha}, \hat{\beta}$ and $v_i = \frac{i}{M} - \frac{1}{2M}$ for $i = 1, ..., M$. All other steps remain the same as before.

2.3. Numerical Experiments

Three numerical studies were conducted to evaluate our methodology. The first uses simulated data to explore the performance of the discretized sampling approach outlined in Section 2.2.4; the second conducts an empirical study to evaluate the accuracy of our SSD approach, where meta-analyses of different sizes were simulated from real data; and the last explores the circumstances under which our method differs from other aggregate approaches.

2.3.1. A simulation study for comparing sampling strategies

To test the performance of the discretized sampling (DS) algorithm, we compared the simple random sampling (SRS) approach with 10,000 samples against the DS approach with 1,000 samples. For each setting, the SRS approach was run 100 times, with the mean and variance of n, the recommended sample size, recorded for each run. The DS approach is designed to produce the same sample size every time, so there was no need to repeat it for any individual setting. To be as conservative as possible, we assumed that with $M = 10,000$, the SRS approach is virtually unbiased, and that the observed $\bar{n}$ is the "true" sample size. For specific settings, we used $\alpha = \{0.5, 1, 3, 5, 10, 20, 50, 100\}$, $\beta = \{10, 100, 1000, 10000\}$, power$1 - \beta^t = \{.8, .9\}$ and standardized effect sizes of $\{.25, .5, 1\}$ (192 unique settings in total). The Type I error rate α^t was fixed at 5% for two-sided tests. Here, to compare the

Table 2.1: A simulation study for comparing two sampling strategies: simple random sampling (SRS) vs. discretized sampling (DS) in terms of estimation performance and computation time

				Median Computation Time (seconds)		
α	SRS $M = 10k$ (Variance)	SRS $M = 100k$ (Variance)	DS (Bias squared)	SRS $M = 10k$	SRS $M = 100k$	DS
0.5	7886.792	367.326	57.595	1.324	6.1565	0.525
1	55.843	9.097	0.764	0.312	1.2460	0.135
3	1.183	0.162	0.058	0.147	0.597	0.060
5	0.493	0.070	0.043	0.132	0.554	0.055
10	0.215	0.022	0.023	0.123	0.482	0.050
20	0.092	0.038	0.035	0.117	0.477	0.050
50	0.026	0	0.001	0.108	0.422	0.045
100	0.053	0.015	0.025	0.108	0.439	0.050

performance of DS vs. SRS, there is no need to estimate α and β via meta-analysis and thus we take samples from $IG(\alpha, \beta)$ directly using the different strategies in the first step of the SSD algorithm.

The results of this simulation are shown in Table 2.1, aggregated by the value of α. We find that the DS method produced estimates at (or in some cases very close to) our best estimate from the SRS approach, while maintaining computational efficiency in nearly all settings considered. Perhaps the only cause for concern is the first row for $\alpha = 0.5$ where the DS approach appears to have relatively high (squared) bias. Partly to blame for this is the fact that the SRS approach has an extremely high variance, and 100 runs is not enough for $\bar{n}$ to be a consistent predictor of the true sample size. However we can also point to $\alpha \leq 1$ as a possible culprit as well, since this implies that the mean parameter of the inverse-gamma distribution is infinite and that the studies collected are inconsistent (see Section 2.5; we do not recommend any version of our approach when $\alpha \leq 1$).

2.3.2. Empirical Studies for performance evaluation

To conduct an empirical comparison of the proposed SSD method versus other existing alternatives, we identified two large meta-analyses in the literature (Spooner et al., 2000;

M.D. et al., 2017) having 17 and 19 studies, respectively. In our numerical analysis, one study is excluded from M.D. et al. (2017), since the response variable is defined slightly differently than the rest of the meta-analysis, so 18 studies were included. For each, we performed the following steps:

1. Let $\mathcal{P}$ be the set of all possible subsets of size k from a large meta-analysis of N datasets.

2. For a subset $A \in \mathcal{P}$:

   (a) For a study $a \notin A$, perform SSD using the following five methods:

      i. Calculate a target sample size using the reported variance from a (the ground truth).

      ii. Calculate a target sample size using A with our method.

      iii. Calculate a target sample size with a weighted average variance from A, weighted by the sample sizes of individual studies.

      iv. Calculate a target sample size with a simple (non-weighted) average variance from A.

      v. Calculate a target sample size using the median variance from A.

   (b) Repeat for all $a \notin A$, for a total of $N - k$ per A.

3. Repeat for all possible $A \in \mathcal{P}$, for a total of $\begin{pmatrix} N \\ k \end{pmatrix}$.

In this way, we treat the sample size calculated using the actual reported variance as the ground truth. This allows us to objectively evaluate our method against common approaches for approximating power analysis. With the exception of our method, all sample sizes (and corresponding power) were calculated using power.t.test() in R. Here, we set the Type I error rate $\alpha^t = 5\%$ and power $1 - \beta^t = 90\%$ for two-sided tests, and a standardized effect size of

Table 2.2: An empirical study for performance evaluation on SSD using 19 datasets from Nedocromil Sodium meta-analysis

	MSE			Median Power Lost			Median Excess n		
Method	$k=3$	$k=4$	$k=10$	$k=3$	$k=4$	$k=10$	$k=3$	$k=4$	$k=10$
Our Method	**734.78**	**647.59**	524.98	**13.5%**	**12.8%**	11.3%	**15**	**15**	15
Weighted Average	756.08	669.10	541.93	13.8%	13.1%	**9.7%**	**15**	16	18
SimpleAverage	756.61	664.21	531.86	13.7%	**12.8%**	12.2%	16	**15**	**14**
Median	833.36	698.83	**502.87**	16.5%	14.1%	11.7%	18	16	**14**

Table 2.3: An empirical study for performance evaluation on SSD using 17 datasets from blood pressure meta-analysis

	MSE			Median Power Lost			Median Excess n		
Method	$k=3$	$k=4$	$k=10$	$k=3$	$k=4$	$k=10$	$k=3$	$k=4$	$k=10$
Our Method	**143.58**	**135.66**	**124.21**	**5.6%**	5.4%	**3.9%**	**6**	**6**	6
Weighted Average	154.61	146.79	130.41	6.2%	6.2%	4.2%	**6**	**6**	**5**
Simple Average	162.23	151.33	131.00	**5.6%**	**5.3%**	6.0%	8	7	6
Median	161.26	147.23	126.88	6.2%	5.4%	4.4%	7	7	6

$\frac{1}{2}$ as the clinically relevant effect. Let n_a^* be the "ground truth" sample size for study a, and $n_a^{(A)}$ be the sample size determined by one of the other methods ((b)-(e) above) for study a using the subset A. Then we approximate a mean squared error for this method with the following formula,

$$MSE = \frac{1}{M} \sum_{A \in \mathcal{P}} \sum_{a \notin A} (n_a - n_a^{(A)})^2$$

where $M = (N-k) \begin{pmatrix} N \\ \\ k \end{pmatrix}$. The methods were also evaluated by calculating the median difference between the power at the ground truth sample size (i.e., 90%) and power from a (b)-(e) method given that the study is underpowered (i.e., power $< 90\%$), and the median difference between the ground truth sample size and a (ii)-(v) method given that the study is overpowered (i.e., power $> 90\%$). We performed the simulation for each meta-analysis 3 times, varying k in $\{3, 4, 10\}$. The results are shown below: Across both datasets, our method clearly seems to have the best overall performance based on MSE, but the gap

between it and the other simpler methods shrinks as k increases. A similar trend emerges when looking at median power lost, a measure of how much a study is underpowered given that the study is underpowered, and median excess n, a measure of how many extra subjects are identified when the study is overpowered. Our method tends to perform the strongest when the sample size is small, but as it increases the other methods tend to catch up. It should be noted that the results here do depend on the data in question: these meta-analyses were chosen for testing because the variances fit the exchangeability assumption well. In cases where the variances (σ^2) could plausibly be the same among all studies, then we would expect the weighted average method to perform the best, as this would use the MLE of σ^2 for estimation.

2.3.3. Cases where our method differs from others

Inspecting the results from Section 2.3.2 can help identify specific circumstances when our method provides sample size recommendations different from the weighted average method, which, among the three alternatives, tends to perform well consistently. The first (and perhaps obvious) result to note is the case when $\hat{\alpha}$ and $\hat{\beta}$ are very large, drifting to infinity with a stable ratio. In this case, the empirical variance distribution $IG(\hat{\alpha}, \hat{\beta})$ is quite trivial, largely reduced to a point mass at the mode $\hat{\beta}/(\hat{\alpha}+1)$,which is very close to the weighted average variance. When this happens, the estimated sample size with our method is very similar if not equal to that of the weighted average method, as in the left panel of Figure 2.1. Theoretically, this occurs when the sample variances ($\boldsymbol{y}$) could plausibly share a true variance (θ) and there is no need for heterogeneity in the distribution. The right panel of Figure 2.1 shows that, in general, the studies with sample variances that are similar to the rest tend to produce more trivial distributions, but the criterion for similarity relaxes as the sample size decreases. So even if a study has a sample variance inconsistent with other studies, if the sample size is small enough it can still lead to a trivial distribution. In general, non-

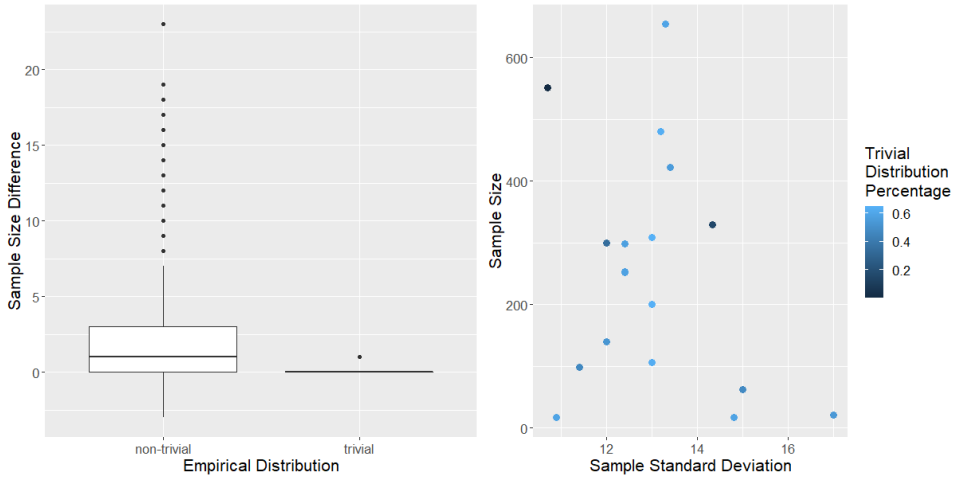

Figure 2.1: The left panel shows the difference in sample size of our method vs. the weighted average method by triviality of the empirical variance distribution (i.e., $\hat{\alpha} \to +\infty$). This covers all possible datasets of size 4 from the blood pressure meta-analysis. The right panel shows a scatter plot of sample standard deviation vs. sample size of datasets from the blood pressure meta-analysis. Each point represents a different study, and the color of each dot corresponds to how often it led to a trivial distribution (via simulation results from all combinations of 4 datasets).

trivial distributions tend to arise when the sample variances are different enough to suggest heterogeneity in distribution with somewhat large sample sizes.

As the lefthand boxplot in Figure 2.1 shows, often it is the case that sample size recommendations from trivial empirical distributions lead to very similar results as the weighted average method. Inspecting the empirical distributions from our simulations suggest that this occurs when these distributions have higher skewness and kurtosis, which are driven solely by the shape parameter estimate $\hat{\alpha}$. Inspecting the sample size calculation results from Section 2.3.2 shows that this occurs more often when the meta-analyses has greater diversity in the sample variances and when the sample sizes are diverse as well (see Figure 2.2). Note that the structure of our method considers not only the relative size of each study but also the actual size. For example, in a two-study meta-analysis, the weighted average approach will produce the same variance estimate (and therefore sample size) as long as the ratio between the two sample sizes stays the same; in other words, using sample sizes of 10

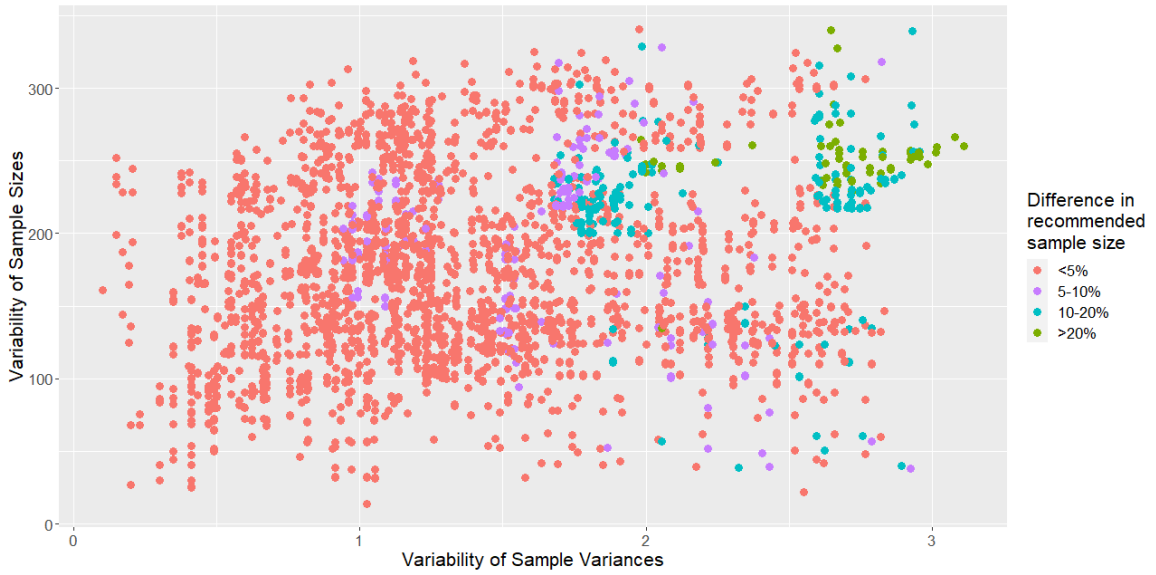

Figure 2.2: The scatterplot shows variation in sample variance vs. variation in sample size of meta-analytic studies for all possible combinations of 4 studies from the numerical analysis in Section 2.3.2. The color represents the difference between the sample size recommended by our method and that of the weighted average approach.

and 100 produce the exact same result as sample sizes of 50 and 500. Our method avoids this issue, by recognizing that larger sample sizes in general produce more stable variance estimates, and the reliability of a sample variance from 10 observations is much weaker than that of size 50. So with our method, the study with $n = 50$ will have more influence than the study with $n = 10$, despite having equal weight with the other method.

2.4. Real Data Applications

To illustrate our approach in practice, we now look at two examples of meta-analyses where our approach can be applied.

2.4.1. UPDRS Data

The Unified Parkinson's Disease Rating Scale (UPDRS) is a questionnaire designed to measure the severity of symptoms experienced by patients diagnosed with Parkinson's disease. The questionnaire looks at a wide range of symptoms, covering everything from

Table 2.4: UPDRS study overview

Reference	Intervention	Reported Sample Size	Reported Variance (SD)
Holloway, et. al. (2000)	Pramipexole	151	161.29 (12.7)
	Levodopa	150	116.64 (10.8)
Palhagen, et. al. (2006)P	Selegiline	35	134.56 (11.6)
	Placebo	41	158.76 (12.6)
Shoulson (2007)	CEP-1347 (25mg BID)	157	105.88 (10.3)
	Placebo	145	82.63 (9.1)

mental/emotional health to the ability of patients to complete normal physical activities (walking, eating food without assistance, etc.). For example, under tremors (probably the most well-known symptom of Parkinson's) the patient scores a 0 on this section if they have experienced no tremors, but would score a 4 if they have severe tremors that interfere with most daily activities. Patients with higher overall scores (up to 199) have experienced significant debilitation, while those with lower scores (at or just above 0) have endured very little disruption to their lives. Because it is a slow-developing disease, reducing the rate at which a patient UPDRS score increases overtime can have drastic effects on patients' life quality and even their long-term survival rate. Our first meta-analysis looks at 3 different studies of pharmaceutical interventions meant to slow the increase of UPDRS score overtime. These were chosen as comparable studies by researchers investigating the effects of Isradipine on UPDRS in a recent clinical trial and were used to justify the sample size, according to the protocol[1]. The data for these studies are shown in Table 2.4.

For each of these studies, the response variable measures the change in UPDRS score from baseline over a period of 36 months (or closest time period). Using the two-sample, same variance methodology outlined in Section 2, we find that $\hat{\alpha} = 33.397$ and $\hat{\beta} = 4034.366$, with the large $\hat{\alpha}$ suggesting that the data fits the IG distribution well (see Figure 2.3). At 80% power for a two-sided test, our method recommends a sample size of 123 participants per group, which is then increased to 145 to account for an anticipated dropout rate of 15%.

[1] https://clinicaltrials.gov/ct2/show/NCT02168842?term=STEADY-PD+III&draw=2&rank=1

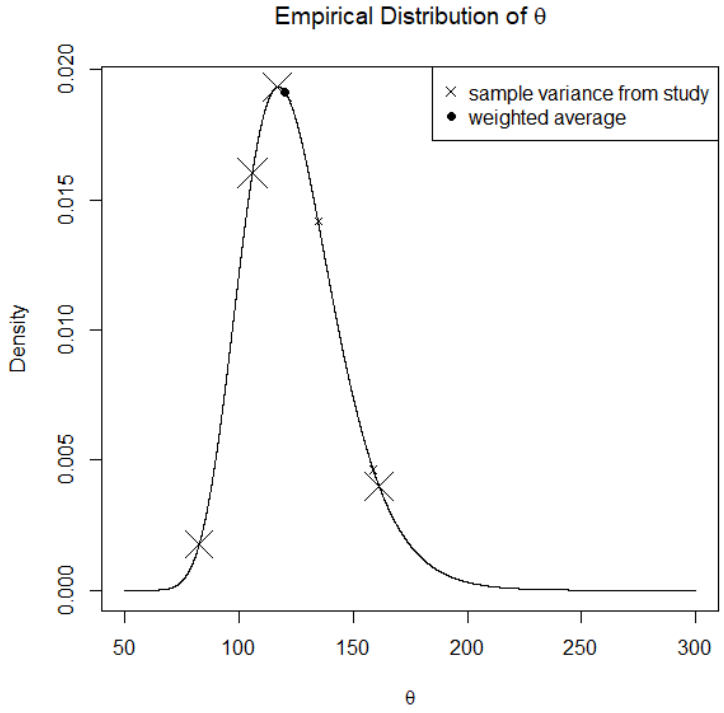

Figure 2.3: UPDRS study: empirical distribution for θ (inverse gamma with $\alpha = 33.397, \beta = 4034.366$), with markers at sample variances from each of the meta-analysis studies.

The weighted average yields a sample size of 140 per group (after accounting for dropout). The protocol cites that a standard deviation of 12 (the 82nd percentile of our fitted inverse gamma distribution) is supported by the studies, yielding a target sample size of 168 per group .

2.4.2. CBT data

Cognitive Behavior Therapy (CBT) is a form of psycho-social therapy used to treat various mental illnesses, including depression and anxiety. Segool and Carlson (2008) sought to evaluate and compare CBT with pharmacological treatments on children suffering from social anxiety using a meta-analysis of small CBT studies. The data for these studies are shown in Table 2.5 below (note that the response variables for these studies are various forms of social anxiety scores).

Table 2.5: CBT study overview

Study (CBT treatment)	Reported Sample Size	Reported Variance
Albano (1995)	5	4.251
Gallagher (2004)	12	1.031
Hayward (2000)	11	1.133
Lumpkin (2002)	4	0.961
Masia (2001)	6	2.126
Shortt (2001)	10	3.252
Spence (2000) NPI	19	0.951
Spence (2000) PI	7	0.464

Since each study has such a small sample size, we can expect the sample variances to differ significantly from each other, making this a prime example for our approach. The weighted average of these variances yields $\tilde{\theta} = 1.5745$. With this variance estimate and an unstandardized effect of 0.5, a one-sample, two-sided test with a type I error of 5% rejects the null hypothesis with 90% probability when 69 patients are studied. With our method, the number of patients needed for these specifications increases to 74. However, it should be noted that the distance between our method and the weighted average approach depends on the targeted amount of power, as shown in Figure 2.4. Interestingly, our method recommends smaller sample sizes for lower power levels compared to the weighted average method. An intuition-based explanation is that for small power values, a larger proportion of the total power can be obtained from the smaller variances in the distribution; as the power level increases, more power needs to be borrowed from the larger variances, leading to a need for larger sample sizes. Or in other words, as the power increases, our method has to consider the possibility of larger variances in the tail, driving up the sample size. When the power is low, our method can ignore the probability of "drawing" a large variance, bringing the overall sample size down.

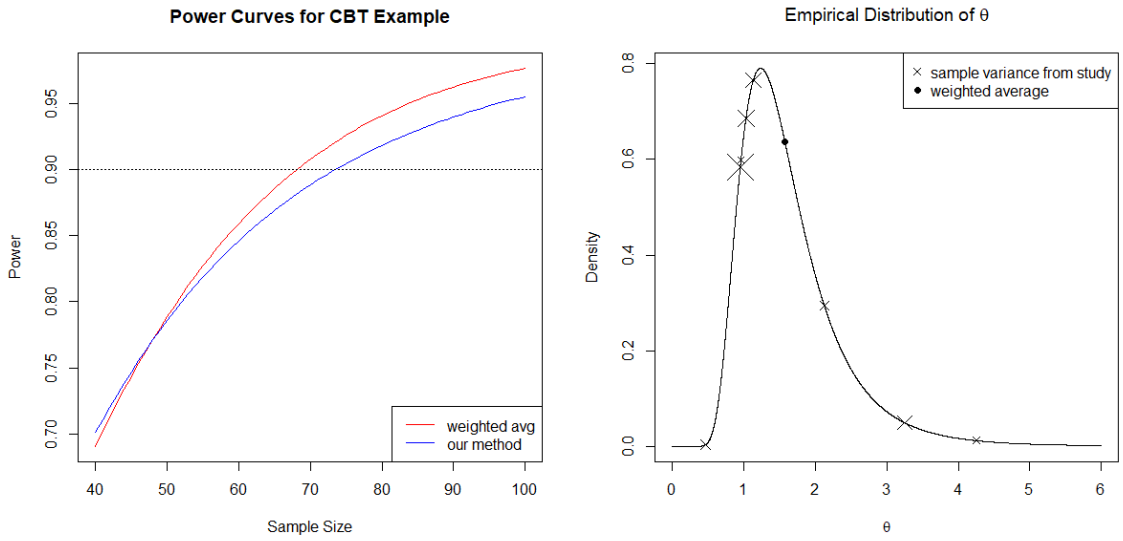

Figure 2.4: CBT study: the left panel shows power curves for both approaches. The right panel shows the empirical distribution for θ (inverse gamma with $\alpha = 7.011, \beta = 9.909$), with markers at sample variances from each of the meta-analysis studies.

2.5. Discussion

We introduced a new methodology for powering one and two-sample t-tests using a set of related studies to construct an empirical distribution for the variance. The method assumes that the true variances from each of these studies are exchangeable with respect to an inverse gamma distribution with shape and scale parameters α, β. In some cases, the marginal MLEs for these parameters lie on the boundary, where all mass of the fitted inverse gamma distribution lies at one point. In these cases, our method reduces to a simple weighted mean approach. The actual sample sizes are calculated via a discretized grid of possible values for σ^2, which is shown to have similar accuracy to an SRS approach with a high number of draws while significantly reducing computation time and prediction variance. Our method demonstrates excellent performance in terms of mean squared error (MSE) when the studies used to estimate the variance distribution are similar to the study of interest, satisfying the assumption of exchangeability. Specifically, it outperforms other methods when the number of studies in the meta-analysis is less than or equal to 5, which is true of roughly

75% of meta-analyses (Rhodes et al., 2015). Our method leverages all information provided by the relevant studies, considering a wide distribution of plausible variances rather than condensing all information into one single estimate. We do not recommend proceeding with this approach if $\hat{\alpha} < 1$, as this suggests that the distribution has an infinite mean and likely does not meet the exchangeability assumption. If $1 < \hat{\alpha} < 2$, we recommend proceeding with caution and comparing the results to other methods.

There are several directions in which we hope to expand this research. The first is to expand on the two-sample, unequal variance case, particularly when the variances have inter-study correlation. This will lead to further exploration of multivariate distributions for the two variances from each of the component studies. Another area of interest is exploring the use of other prior distributions for σ^2 outside of an inverse gamma approach. Additionally, we would like to explore a fully Bayes approach that would allow for more flexible modeling of parameters, as well as cases where the effect size is allowed to vary while assuring it stays in a clinically relevant range. The last, and perhaps most obvious, is to apply this methodology to other types of hypothesis testing, including tests of binary events and analysis of variance.

Appendix S

Supplementary Material for Chapter 1

S1. Full Conditional Distributions for Meta-Analysis

Let $\boldsymbol{\theta_{ik}} = \begin{pmatrix} a_{ik} \\ \\ b_{ik} \end{pmatrix}, \boldsymbol{X_{ik}} = \begin{bmatrix} 1 & X_{i1k} \\ \\ ... & ... \\ \\ 1 & X_{i6k} \end{bmatrix}, \boldsymbol{m_k} = \begin{pmatrix} \alpha_k \\ \\ \beta_k \end{pmatrix}$, while $\boldsymbol{Y_{ik}}, \boldsymbol{s_k}$ are vectors over all 6 Y_{ijk}, s_{jk} and $\text{diag}(\sigma^2_{jk})$ is a diagonal matrix with $\sigma^2_{1k}...\sigma^2_{6k}$ in the diagonal. $I_{(L,U)}$ is the indicator function, where $I_{(L,U)} = 1$ when $L < x < U$ and 0 otherwise. For all Inverse-Gamma (IG) distributions, $\epsilon = .01$.

$$\boldsymbol{\theta_{ik}}|\cdots, \boldsymbol{Y} \sim N\left(\mu_{\alpha\beta}, \boldsymbol{\Sigma}^{(k)}_{\alpha\beta}\right) \text{ where}$$

$$\mu_{\alpha\beta} = \boldsymbol{\Sigma}^{(k)}_{\alpha\beta}\left(\boldsymbol{\Sigma}^{(k)^{-1}}\boldsymbol{m_k} + \boldsymbol{X^T_{ik}}\text{diag}(\sigma^2_{jk})^{-1}(\boldsymbol{Y_{ik}} - \boldsymbol{s_k})\right) \text{ and}$$

$$\boldsymbol{\Sigma}^{(k)}_{\alpha\beta} = \left(\boldsymbol{\Sigma}^{(k)^{-1}} + \boldsymbol{X^T_{ik}}\text{diag}(\sigma^2_{jk})^{-1}\boldsymbol{X_{ik}}\right)^{-1}$$

$$\boldsymbol{\Sigma}^{(k)}\Big|\cdots, \boldsymbol{Y} \sim IW_{n_k+3}\left(I + \sum_{i=1}^{n_k}\begin{bmatrix} (a_{ik}-\alpha_k)^2 & (a_{ik}-\alpha_k)(b_{ik}-\beta_k) \\ \\ (a_{ik}-\alpha_k)(b_{ik}-\beta_k) & (b_{ik}-\beta_k)^2 \end{bmatrix}\right)$$

$$\boldsymbol{m}_k|\cdots, \boldsymbol{Y} \sim BVN\left(\left(\boldsymbol{\Sigma}^{-1}_m + n_k\boldsymbol{\Sigma}^{(k)^{-1}}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}_m\boldsymbol{\mu} + \sum_{i=1}^{n_k}\left(\boldsymbol{\Sigma}^{(k)^{-1}}\boldsymbol{\theta}_{ik}\right)\right), \left(\boldsymbol{\Sigma}^{-1}_m + n_k\boldsymbol{\Sigma}^{(k)^{-1}}\right)^{-1}\right)$$

$$\mu_\alpha|\cdots, \boldsymbol{Y} \sim I_{(L_\alpha, U_\alpha)} \cdot N\left(\frac{1}{K}\sum_{k=1}^{K}\alpha_k, \frac{\sigma^2_\alpha}{K}\right)$$

$$\mu_\beta|\cdots,\boldsymbol{Y}\sim I_{(L_\beta,U_\beta)}\cdot N\left(\frac{1}{K}\sum_{k=1}^{K}\beta_k,\frac{\sigma_\beta^2}{K}\right)$$

$$\sigma_\alpha^2\big|\cdots,\boldsymbol{Y}\sim IG\left(\epsilon+\frac{K}{2},\epsilon+\frac{1}{2}\sum_{k=1}^{K}(\alpha_k-\mu_\alpha)^2\right)$$

$$\sigma_\beta^2\big|\cdots,\boldsymbol{Y}\sim IG\left(\epsilon+\frac{K}{2},\epsilon+\frac{1}{2}\sum_{k=1}^{K}(\beta_k-\mu_\beta)^2\right)$$

For s_{jk}, we apply two constraints and update the parameter vectors s_{5k} and s_{6k} based on $s_{1k},s_{2k},s_{3k},s_{4k}$. Let $h_j=\frac{X_6-X_j}{X_5-X_6}$, $r_{ijk}=Y_{ijk}-a_{ik}-b_{ik}X_j$, $S_{1:4,-j}=\left(\sum_{l=1}^{4}s_{lk}\right)-s_{jk}$, $(XS)_{1:4,-j}=\left(\sum_{l=1}^{4}X_l s_{lk}\right)-X_j s_{jk}$.

$$s_{jk}|\cdots,\boldsymbol{Y}\sim N\left(\mu_{jk}^{(s)},\sigma_{jk}^{2(s)}\right)\text{ where }j=1,2,3,4$$

$$\sigma_{jk}^{2(s)}=\left(\frac{1}{\tau^2}\left(1+h_j^2+(1+h_j)^2\right)+\frac{n_k}{\sigma_{jk}^2}+\frac{n_k}{\sigma_{5k}^2}h_j^2+\frac{n_k}{\sigma_{6k}^2}(1+h_j)^2\right)^{-1}$$

$$\begin{aligned}\mu_{jk}^{(s)}=\sigma_{jk}^{2(s)}&\left(\frac{t_j}{\tau^2}+\frac{(t_5-\frac{X_6S_{1:4,-j}-(XS)_{1:4,-j}}{X_5-X_6})}{\tau^2}h_j-\frac{t_6+S_{1:4,-j}+\frac{X_6S_{1:4,-j}-(XS)_{1:4,-j}}{X_5-X_6}}{\tau_6^2}(1+h_j)\right)\\&+\sigma_{jk}^{2(s)}\left(+\frac{1}{\sigma_{jk}^2}\sum_{i=1}^{n_k}r_{ijk}+\frac{1}{\sigma_{5k}^2}\sum_{i=1}^{n_k}\left(r_{i5k}-\frac{X_6S_{1:4,-j}-(XS)_{1:4,-j}}{X_5-X_6}\right)h_j\right)\\&-\sigma_{jk}^{2(s)}\left(\frac{1}{\sigma_{6k}^2}\sum_{i=1}^{n_k}\left(r_{i6k}+S_{1:4,-j}+\frac{X_6S_{1:4,-j}-(XS)_{1:4,-j}}{X_5-X_6}\right)(1+h_j)\right)\end{aligned}$$

$$s_{5k}=\frac{X_6\sum_{j=1}^{4}s_{jk}-\sum_{j=1}^{4}X_j s_{jk}}{X_5-X_6},\qquad s_{6k}=-\left(\sum_{j=1}^{5}s_{jk}\right)$$

For t_j, we apply two constraints and update the parameters t_5 and t_6 based on t_1, t_2, t_3, t_4. Note that $T_{1:4,-j}$ and $(XT)_{1:4,-j}$ are defined similarly to $S_{1:4,-j}$ and $(XS)_{1:4,-j}$ above.

$$t_j | \cdots, \boldsymbol{Y} \sim I_{(L_t, U_t)} \cdot N\left(\mu_j^{(t)}, \sigma_j^{2(t)}\right) \text{ where } t = 1, 2, 3, 4$$

$$\sigma_j^{2(t)} = \left(\frac{K}{\tau^2}\left(1 + h_j^2 + (1+h_j)^2\right)\right)^{-1}$$

$$\mu_j^{(t)} = \sigma_j^{2(t)}\left[\frac{1}{\tau^2}\left(\sum_{k=1}^{K} s_{jk} + h_j \sum_{k=1}^{K}\left(s_{5k} - \frac{X_6 T_{1:4,-j} - (XT)_{1:4,-j}}{X_5 - X_6}\right) - (1+h_j)\sum_{k=1}^{K}\left(s_{6k} + T_{1:4,-j} + \frac{X_6 T_{1:4,-j} - (XT)_{1:4,-j}}{X_5 - X_6}\right)\right)\right]$$

$$\tau^2 \big| \cdots, \boldsymbol{Y} \sim IG\left(\epsilon + 3K, \epsilon + \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{6}(s_{jk} - t_j)^2\right)$$

$$\sigma_{jk}^2 | \cdots, \boldsymbol{Y} \sim IG\left(\epsilon + \frac{n_k}{2}, \epsilon + \frac{1}{2}\sum_{i=1}^{n_k}(Y_{ijk} - (a_{ik} + b_{ik}X_{ijk} + s_{jk}))^2\right)$$

S2. Additional Tables and Figures

Table S1: Meta-Analysis Datasets

ID	Description of samples	# of samples	Reference	GSE ID or Link
1	Lung adenocarcinoma (LUAD) patients	162	Molania et al. (2019)	N/A[1]
2	Inflammatory bowel disease patients	989	Molania et al. (2019)	GSE73094
3	Colon cancer patients	96	Chen et al. (2016b)	GSE62932
4	Lung cancer patients	28	Jia et al. (2019)	N/A[2]
5	Colorectal cancer (CRC) patients	54	Jia et al. (2019)	GSE86561
6	Early stage CRC patient tumors	144	Low et al. (2017)	GSE81983
7	Early stage CRC patient tumors	131	Low et al. (2017)	GSE81985
8	Breast tumor samples	1321	Liu et al. (2016)	GSE74821
9	Patients stimulated with anti-CD3/CD28	1950	Molania et al. (2019)	GSE60341
10	Healthy individuals (stimulated and controls)	2441	Molania et al. (2019)	GSE53165
11	Carolina breast cancer study (CBCS) tumors	1278	Patel et al. (2022)	GSE148418
12	Survivors of triple negative breast cancer	254	Cascione et al. (2013)	GSE45498
13	Merkel-cell carcinoma patients	8	Gravemeyer et al. (2021)	GSE159662
14	T-cell lymphoma or dermititis (plus controls)	128	Nielsen et al. (2019)	GSE143382
15	Metastatic melanoma patients	24	DeVito et al. (2021)	GSE165745
16	Breast cancer patients	1253	Pu et al. (2019)	GSE147126
17	Squamous cell carcinoma patients	67	Meehan et al. (2020)	GSE148944

[1] https://github.com/RMolania/NanostringNormalization/tree/master/Example%201_LungCancerStudy

[2] Data can be accessed in the RCRnorm R package

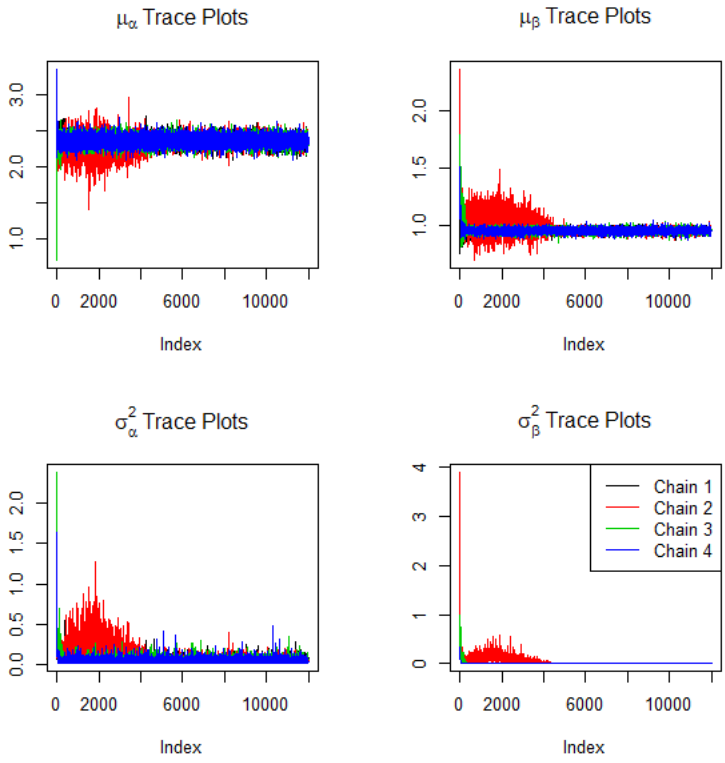

Figure S1: Trace plots (left) and Gelman-Rubin plots for global parameters in in our Bayesian meta-analysis

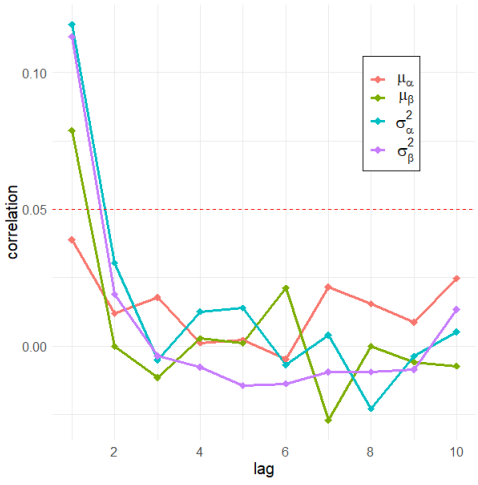

Figure S2: Autocorrelation for global parameters in our Bayesian meta-analysis. Since all chains produced similar results, only autocorrelation from chain 1 is shown.

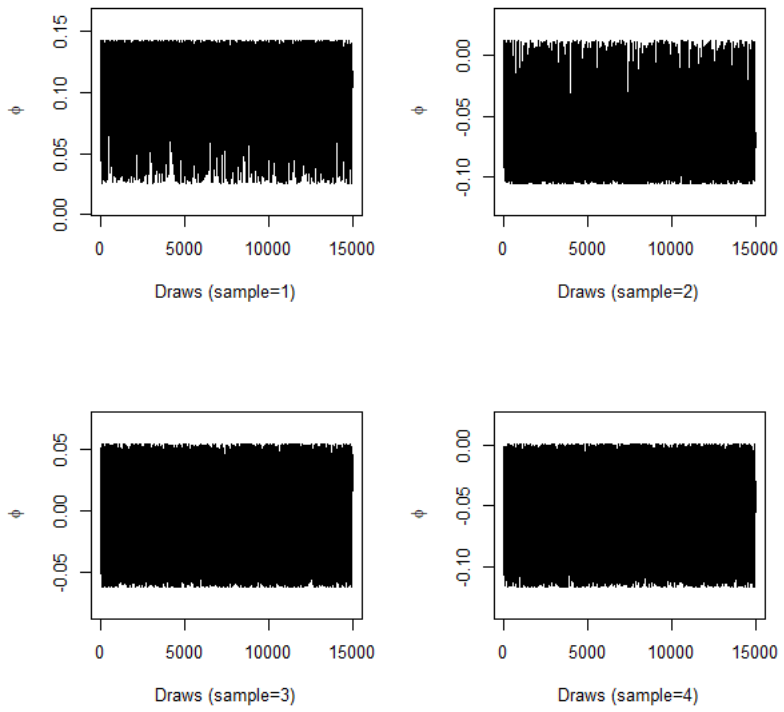

Figure S3: Trace plots for $\phi_1 - \phi_4$ (dataset 13).

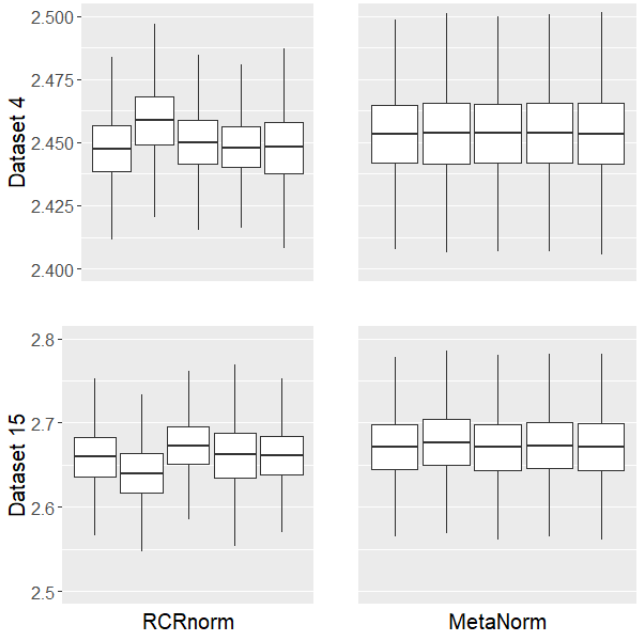

Figure S4: Posterior distribution (by chain) of μ_a for RCRnorm and MetaNorm normalization of datasets 4 and 15

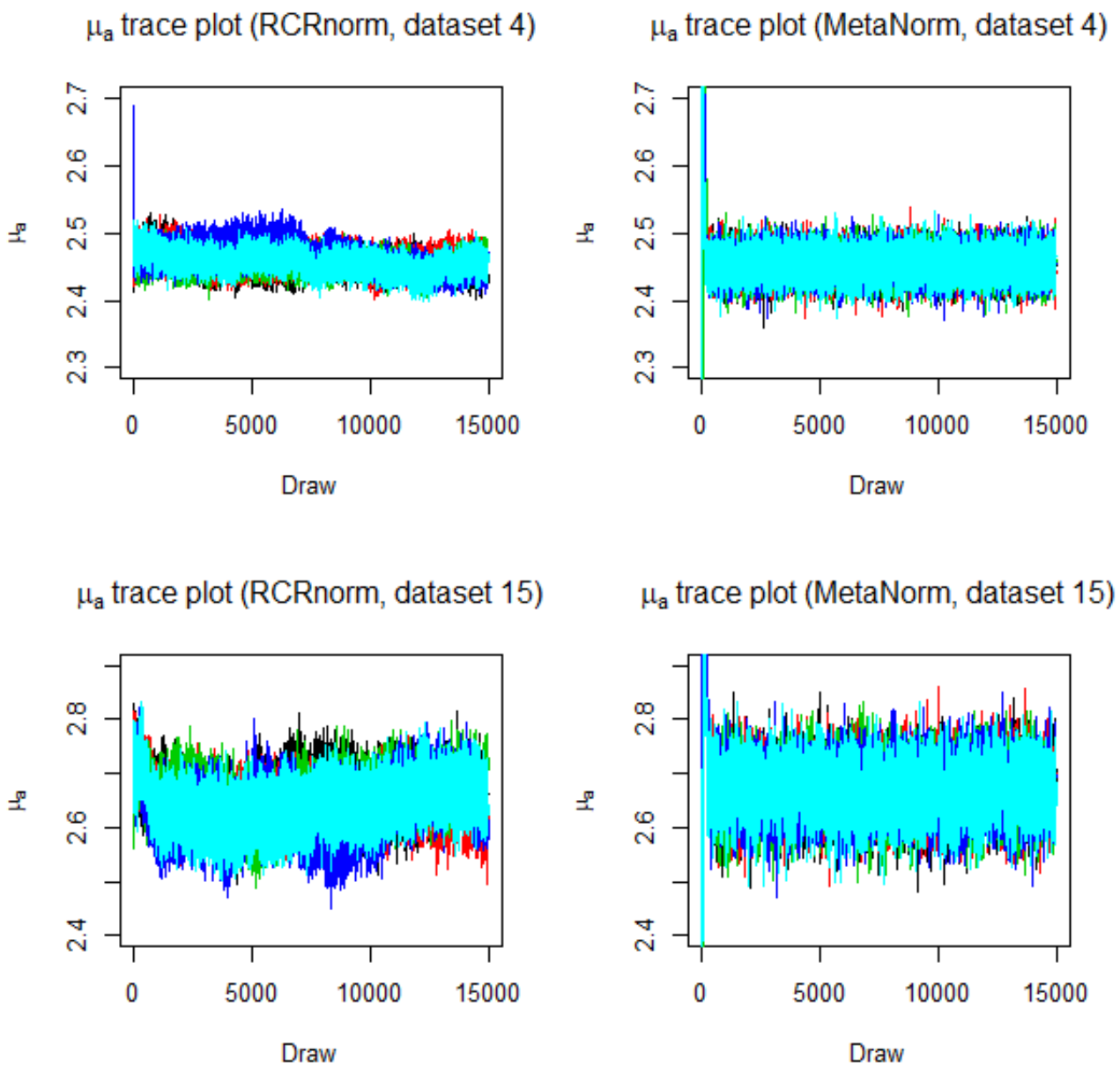

Figure S5: Comparison between RCRnorm and MetaNorm on convergency using trace plots for μ_a

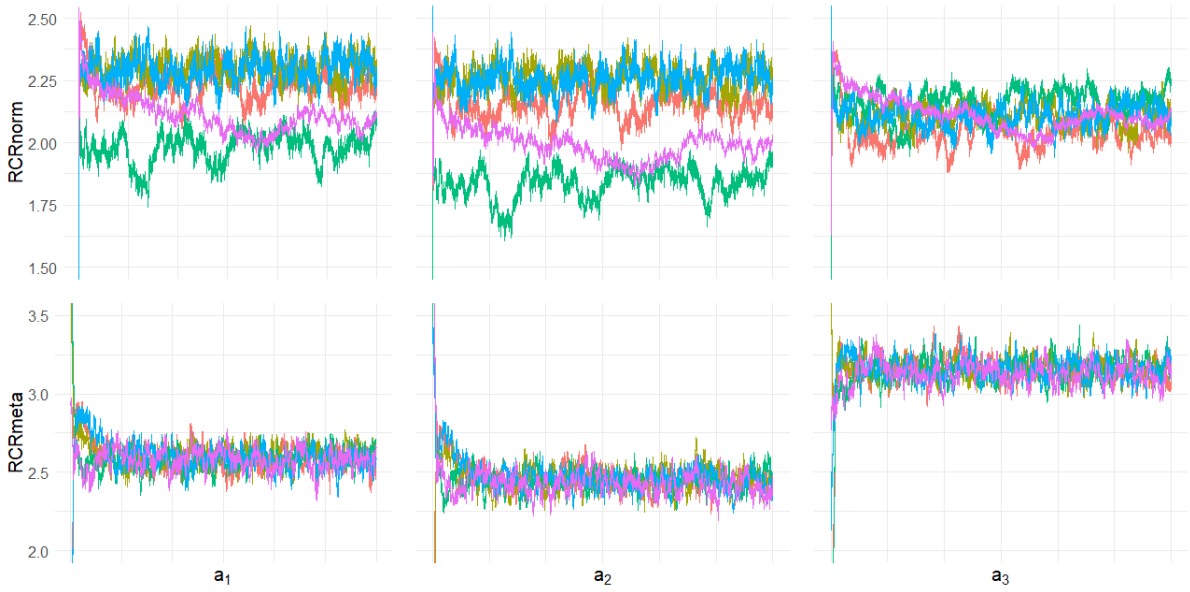

Figure S6: Comparison between RCRnorm and MetaNorm using traceplots of sample-specific intercepts $a_1 - a_3$ (Dataset 13).

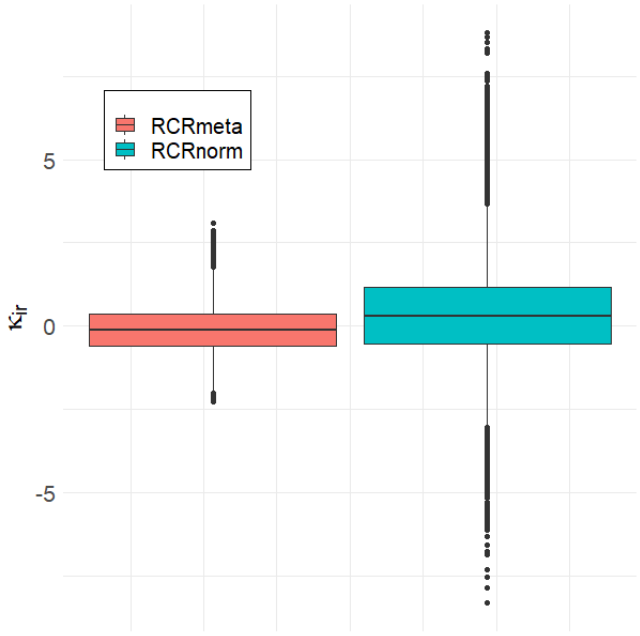

Figure S7: Comparison of κ_{ir} (normalized $\log_{10}$ mRNA expression levels) for dataset 12

BIBLIOGRAPHY

Bayarri, M. J. and Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical science*, 19(1):58–80. 13, 30

Berger, J. O. and Sun, D. (2008). Objective priors for the bivariate normal model. *The Annals of Statistics*, 36(2):963–982. 15

Bhattacharya, A., Hamilton, A. M., Furberg, H., Pietzak, E., Purdue, M. P., Troester, M. A., Hoadley, K. A., and Love, M. I. (2021). An approach for normalization and quality control for nanostring rna expression data. *Briefings in bioinformatics*, 22(3). 3, 21, 27

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877. 28

Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in medicine*, 14(17):1933–1940. 30

Cao, J., Lee, J. J., and Alber, S. (2009). Comparison of bayesian sample size criteria: Acc, alc, and woc. *Journal of statistical planning and inference*, 139(12):4111–4122. 30

Cascione, L., Gasparini, P., Lovat, F., Carasi, S., Pulvirenti, A., Ferro, A., Alder, H., He, G., Vecchione, A., Croce, C. M., Shapiro, C. L., and Huebner, K. (2013). Integrated microrna and mrna signatures associated with survival in triple negative breast cancer. *PloS one*, 8(2):e55910. 55

Chen, X., Deane, N. G., Lewis, K. B., Jiang, L., Zhu, J., Washington, M. K., and Beauchamp, R. D. (2016a). Comparison of nanostring ncounter® data on ffpe colon cancer samples and affymetrix microarray data on matched frozen tissues. *PLoS One*, 11(5):e0153784. 1

Chen, X., Deane, N. G., Lewis, K. B., Li, J., Zhu, J., Washington, M. K., and Beauchamp, R. D. (2016b). Comparison of nanostring ncounter.sup.[r] data on ffpe colon cancer samples and affymetrix microarray data on matched frozen tissues. *PloS one*, 11(5). 55

DeVito, N. C., Sturdivant, M., Thievanthiran, B., Xiao, C., Plebanek, M. P., Salama, A. K. S., Beasley, G. M., Holtzhausen, A., Novotny-Diermayr, V., Strickler, J. H., and Hanks, B. A. (2021). Pharmacological wnt ligand inhibition overcomes key tumor-mediated resistance pathways to anti-pd-1 immunotherapy. *Cell reports (Cambridge)*, 35(5):109071. 55

Du, H. and Wang, L. (2016). A bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate behavioral research*, 51(5):589–605. 30

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574. 2

Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., Grogan, T., James, J. J., Maysuria, M., Mitton, J. D., Oliveri, P., Osborn, J. L., Peng, T., Ratcliffe, A. L., Webster, P. J., Davidson, E. H., Hood, L., and Dimitrov, K. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology*, 26(3):317–325. 1

Gelfand, A. E. and Wang, F. (2002). A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical science*, 17(2):193–208. 30

Gravemeyer, J., Lange, A., Ritter, C., Spassova, I., Song, L., Picard, D., Remke, M., Horny, K., Sriram, A., Gambichler, T., Schadendorf, D., Hoffmann, D., and Becker, J. C. (2021). Classical and variant merkel cell carcinoma cell lines display different degrees of neuroendocrine differentiation and epithelial-mesenchymal transition. *Journal of investigative dermatology*, 141(7):1675–1686.e4. 55

Halpern, S. D., Karlawish, J. H. T., and Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *JAMA : the journal of the American Medical Association*, 288(3):358–362. 29

Hochster, H. S. (2008). The power of "p": on overpowered clinical trials and "positive" results. *Gastrointestinal cancer research*, 2(2):108–109. 29

Jia, G., Wang, X., Li, Q., Lu, W., Tang, X., Wistuba, I., and Xie, Y. (2019). Rcrnorm: An integrated system of random-coefficient hierarchical regression models for normalizing nanostring ncounter data. *The annals of applied statistics*, 13(3):1617–47. ix, 2, 3, 5, 6, 7, 18, 19, 20, 55

Joseph, L., Berger, R. D., and Belisle, P. (1997). Bayesian and mixed bayesian/likelihood criteria for sample size determination. *Statistics in medicine*, 16(7):769–781. 30

Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., and Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, 63(5):484–489. 30

Lancaster, G. A., Dodd, S., and Williamson, P. R. (2004). Design and analysis of pilot studies: recommendations for good practice. *Journal of evaluation in clinical practice*, 10(2):307–312. 30

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American statistician*, 55(3):187–193. 29

Lim, G. B., Kim, Y.-A., Seo, J.-H., Lee, H. J., Gong, G., and Park, S. H. (2020). Prediction of prognostic signatures in triple-negative breast cancer based on the differential expression analysis via nanostring ncounter immune panel. *BMC Cancer*, 20(1). 1

Liu, F. (2010). An extension of bayesian expected power and its application in decision making. *Journal of Biopharmaceutical Statistics*, 20(5):941–953. 29

Liu, M. C., Pitcher, B. N., Mardis, E. R., Davies, S. R., Friedman, P. N., Snider, J. E., Vickery, T. L., Reed, J. P., DeSchryver, K., Singh, B., Gradishar, W. J., Perez, E. A., Martino, S., Citron, M. L., Norton, L., Winer, E. P., Hudis, C. A., Carey, L. A., Bernard, P. S., Nielsen, T. O., Perou, C. M., Ellis, M. J., and Barry, W. T. (2016). Pam50 gene signatures and breast cancer prognosis with adjuvant anthracycline- and taxane-based chemotherapy: correlative analysis of c9741 (alliance). *NPJ breast cancer*, 2(1):15023. 55

Low, Y. S., Blöcker, C., McPherson, J. R., Tang, S. A., Cheng, Y. Y., Wong, J. Y. S., Chua, C., Lim, T. K. H., Tang, C. L., Chew, M. H., Tan, P., Tan, I. B., Rozen, S. G., and Cheah, P. Y. (2017). A formalin-fixed paraffin-embedded (ffpe)-based prognostic signature to predict metastasis in clinically low risk stage i/ii microsatellite stable colorectal cancer. *Cancer letters*, 403:13–20. 55

Masuda, N., Ohnishi, T., Kawamoto, S., Monden, M., and Okubo, K. (1999). Analysis of chemical modification of rna from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic acids research*, 27(22):4436–4443. 1

M.D., J. J., M.D.PhD., M. R., M.D.M.P.H., S. H., and M.D., C. E. (2017). Are automated blood pressure monitors comparable to ambulatory blood pressure monitors? a systematic review and meta-analysis. *Canadian journal of cardiology*, 33(5):644–652. 42

Meehan, K., Leslie, C., Lucas, M., Jacques, A., Mirzai, B., Lim, J., Bulsara, M., Khan, Y., Wong, N. C., Solomon, B., Sader, C., Friedland, P., Arnau, G. M., Semple, T., and Lim, A. M. (2020). Characterization of the immune profile of oral tongue squamous cell carcinomas with advancing disease. *Cancer medicine (Malden, MA)*, 9(13):4791–4807. 55

Molania, R., Gagnon-Bartsch, J., Dobrovic, A., and Speed, T. P. (2019). A new normalization for nanostring ncounter gene expression data. *Nucleic acids research*, 47(12):6073–6083. 55

Moore, C. G., Carter, R. E., Nietert, P. J., and Stewart, P. W. (2011). Recommendations for planning pilot studies in clinical and translational research. *Clinical and translational science*, 4(5):332–337. 30

Moser, B. K. and Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *The American statistician*, 46(1):19–21. 39

Nielsen, P. R., Eriksen, J. O., Lindahl, L. M., Wehkamp, U., Andersen, G. R., Bzorek, M., Litman, T., ødum, N., and Gjerdrum, L. M. R. (2019). 029 - a diagnostic two gene classifier in patients with mycosis fungoides: a retrospective multicenter study. *European journal of cancer (1990)*, 119:S11. 55

Patel, A., García-Closas, M., Olshan, A. F., Perou, C. M., Troester, M. A., Love, M. I., and Bhattacharya, A. (2022). Gene-level germline contributions to clinical risk of recurrence scores in black and white patients with breast cancer. *Cancer research (Chicago, Ill.)*, 82(1):25–35. 55

Perlmutter, M. A., Best, C. J. M., Gillespie, J. W., Gathright, Y., González, S., Velasco, A., Linehan, W. M., Emmert-Buck, M., and Chuaqui, R. F. (2004). Comparison of snap freezing versus ethanol fixation for gene expression profiling of tissue specimens. *The Journal of Molecular Diagnostics*, 6(4):371–377. 1

Pu, M., Messer, K., Davies, S. R., Vickery, T. L., Pittman, E., Parker, B. A., Ellis, M. J., Flatt, S. W., Marinac, C. R., Nelson, S. H., Mardis, E. R., Pierce, J. P., and Natarajan, L. (2019). Research-based pam50 signature and long-term breast cancer survival. *Breast cancer research and treatment*, 179(1):197–206. 55

Rhodes, K. M., Turner, R. M., and Higgins, J. P. T. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of clinical epidemiology*, 68(1):52–60. 51

Santis, F. D. (2004). Statistical evidence and sample size determination for bayesian hypothesis testing. *Journal of statistical planning and inference*, 124(1):121–144. 30

Santis, F. D. (2007). Using historical data for bayesian sample size determination. *Journal of the Royal Statistical Society.Series A, Statistics in society*, 170(1):95–113. 29

Segool, N. K. and Carlson, J. S. (2008). Efficacy of cognitive-behavioral and pharmacological treatments for children with social anxiety. *Depression and anxiety*, 25(7):620–631. 48

Shieh, G. (2017). The equivalence of two approaches to incorporating variance uncertainty in sample size calculations for linear statistical models. *Journal of applied statistics*, 44(1):40–56. 30

Sims, M., Elston, D. A., Harris, M. P., and Wanless, S. (2007). Incorporating variance uncertainty into a power analysis of monitoring designs. *Journal of agricultural, biological, and environmental statistics*, 12(2):236–249. 30

Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in medicine*, 5(1):1–13. 30

Spooner, C., Rowe, B. H., and Saunders, L. D. (2000). Nedocromil sodium in the treatment of exercise-induced asthma: a meta-analysis. *The European respiratory journal*, 16(1):30–37. 41

Waggott, D., Chu, K., Yin, S., Wouters, B. G., Liu, F.-F., and Boutros, P. C. (2012). Nanostringnorm: an extensible r package for the pre-processing of nanostring mrna and mirna data. *Bioinformatics*, 28(11):1546–1548. 1

Walter, R. F. H., Werner, R., Vollbrecht, C., Hager, T., Flom, E., Christoph, D. C., Schmeller, J., Schmid, K. W., Wohlschlaeger, J., and Mairinger, F. D. (2016). Actb, cdkn1b, gapdh, grb2, rhoa and sdcbp were identified as reference genes in neuroendocrine lung cancer via the ncounter technology. *PLoS ONE*, 11:e0165181. 1

Wang, H., Horbinski, C., Wu, H., Liu, Y., Sheng, S., Liu, J., Weiss, H., Stromberg, A. J., and Wang, C. (2016). Nanostringdiff: a novel statistical method for differential expression analysis based on nanostring ncounter data. *Nucleic acids research*, 44(20):e151. 2

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society.Series D (The Statistician)*, 46(2):185–191. 30