New methods for meta-analysis under a fixed effects framework, with frequentist and Bayesian estimation

Clara P. Domínguez Islas

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Kenneth M. Rice, Chair

Lurdes Y. T. Inoue

Noah Simon

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

New methods for meta-analysis under a fixed effects framework, with frequentist and Bayesian estimation

Clara P. Domínguez Islas

Chair of the Supervisory Committee:
Ph.D. Kenneth M. Rice
Biostatistics

Meta-analysis, as a pivotal component of systematic reviews, has been used extensively in recent years to synthesize the increasing amount of evidence produced in medical and health care research. The two main approaches to meta-analysis are based either on the assumption that all studies estimate a single common effect or on the assumption that the effects are sampled from an unknown distribution. In this dissertation, we propose and develop methods for an alternative approach to meta-analysis, based on the assumption that the effects estimated in the different studies are unknown, but fixed and not necessarily identical. In Chapter 1 we present a brief introduction and review of current methods for meta-analysis. In Chapter 2, we provide a novel yet simple justification for the precision weighted average estimator, commonly used in meta-analysis. Unlike standard arguments that require a homogeneity assumption on the study effects, our justification is based on an optimality property of this particular weighted average of the effect-size parameters, among the class of all affine combinations. We also propose a parameter to quantify the heterogeneity observed in the studies at hand, as an alternative to the classical between-studies variance in a random effects approach. We propose frequentist estimators of these parameters, illustrating their properties through an applied example and evaluating their performance in a simulation study. In Chapter 3, we propose Bayesian methods for estimation of the location and dispersion parameters. We discuss how different prior beliefs

like homogeneity, exchangeability or correlation, can be incorporated through a variety of prior distributions that may or may not involve the use of hyper parameters. Important properties of the estimators obtained from a conjugate prior are derived analytically and then illustrated in an example. In Chapter 4, we explore methods to improve on the estimation of the individual effects themselves, rather than a summary of them. We discuss some ideas on shrinkage estimation and convex clustering adapted to meta-analysis, with particular attention to penalized estimation. We propose a loss function which takes into account the precision of the effect estimates, and can be seen either as a weighted version of the Fused Lasso Signal Approximator (FLSA) or a specific case of a Convex Cluster criterion. We derive some important properties of the solutions to the loss function, which allow us to construct an efficient algorithm to obtain the complete solution path. We also propose a novel procedure based on the parametric bootstrap for the estimation of the tuning parameter, and present results from a simulation study evaluating the proposed shrinkage estimation method. Finally, we conclude with a discussion and conclusions in Chapter 5.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to express my gratitude to my dissertation adviser, friend and mentor Dr. Ken Rice. Thank you for all the time, guidance and thoughts, but above all, for the cookies and lemon bars. I would like to thank to the members of my reading committee and amazing faculty members of the department, Lurdes Inoue and Noah Simon, for their feedback, encouragement and patience.

I would like to express my sincere appreciation to the University of Washington, who has provided amazing support to many international students. To the department of Biostatistics, for giving me this great opportunity. To all the amazing faculty in the department, specially to Jennifer Nelson, Thomas Lumley, Scott Emerson, Barbara McKnight and Jon Wellner. Special thanks to Gitana Garofalo, Sheila Shapiro, Alex MacKenzie and Cathy Greenbaum, who have always been so nice and helpful to all the students.

Personally, I would like to thank Arthur Wongstchowski, for being the peace and calmness in my turbulent world. To my family, who have been terribly neglected for years, but have always been supportive and happy for my achievements. To my friends, the ones in Mexico that I once left to come to Seattle, and the ones in Seattle who I will be leaving soon. To Dulce Vargas, who was there for me in the time when I needed it the most. To Julio Amezcua, for his wise advise. To Socorro Romero, for being so caring in her own way. To Luz Hernandez and Daniel Ortiz, for their encouragement and support. To Leila Zelnick, in whom I found a second sister. To Do Peterson, for believing in me. To Elisa Sheng, with whom I have shared so much coffee, so many hours of work and not enough climbing time. To Navneet Hakhu, for all his unconditional support. To Arie Voorman, for the inspiring discussions at lunch time. To Roxana Chen, for a timely rescue. To Michael Karcher, for all the food and contra. To Jan Irvahn, for Dr. Seuss and Glivenko-Cantelli. To Dan Bohus, for all those tangos.

DEDICATION

to my dear mother, Beatriz, who I miss every single day

Chapter 1

INTRODUCTION

Meta-analyses are a pivotal component of systematic reviews [44], which have been extensively used in recent years to synthesize the increasing amount of evidence produced in health care research [15]. Here we refer to meta-analysis as "the use of statistical methods to summarize the results of independent studies" [29; 44].

In broad terms, the primary aim of most meta-analyses is to make inference on the size of effects, across several similar studies. We usually wish to summarize all the studies and make inference on the magnitude and direction of an average effect of some sort.

We can identify three different approaches to meta-analysis, based on the statistical assumptions that can be made about the true effect-size parameters underlying the studies. The first approach is the fixed-effect (singular) meta-analysis, also called the 'common-effect' meta-analysis [77]. This approach is based on the assumption of a single, common effect underlying all studies [37]. Under this simplifying assumption that all study effects are identical, the average effect is equivalent to the common effect size estimated in each study. Although commonly used, this method has often been judged inadequate in practice, as effects from different studies are expected to differ given the variability in study design, population, interventions, etc. [8; 28; 94].

A second approach is the fixed-effects (plural) meta-analysis, based on the assumption that the effects underlying the studies at hand are unknown, but fixed, and not necessarily identical [57; 72]. The fixed-effects approach is usually used to estimate the inverse-variance weighted average of the effect-sizes represented in the meta-analysis [77], although estimation of other weighted averages is also possible [57; 104]. As discussed by [77], this method typically estimates a reasonable and sensible parameter, even when the effect sizes are assumed to be different. On the other hand, even when statistically valid, inference on this parameter might not be a useful summary of the study effect sizes if they are too

heterogeneous [4].

The third approach is the random-effects meta-analysis, where the effect-size parameters are considered to be a random sample from a population, i.e. they follow a probability distribution [19]. By using random effects as a sampling model, this analysis allows the estimation of the average effect-size in the population of effect-sizes one might ever have observed [19]. This method not only takes into account the heterogeneity between studies, but also provides a natural way of quantifying it [46], making it a more attractive choice over the common- and fixed-effects approaches [8; 94]. On the other hand, as pointed out by Higgins et al. [46], this approach is based on a construct of an imaginary population that does not really exist, so the interpretation of the analysis is potentially unclear, and sometimes confusing. Also, the inference provided by this analysis, mainly focused on the hyper-parameters of this imaginary population, although valid, may not be of interest [73].

So far, we have described and discussed these three approaches from a frequentist framework, but analogs within the Bayesian paradigm are also commonly used. The random effects analysis is perhaps the most popular approach within the Bayesian framework and, as noted in Higgins et al. [46], exchangeability arguments provide further motivation for its use. In fact, it has been proposed to use a hierarchical Bayes linear model to integrate random-effects and common-effect models into a unified approach [22]. However, an analog to the precision weighted average under a fixed effects framework has not been yet proposed.

In each of the three approaches described, appeals to some form of frequentist optimality can be made. In the common-effect approach, when the study-specific standard errors are known precisely, the optimality is straightforward; without any distributional assumptions, the inverse-variance weighted estimator provides the best linear unbiased estimator (BLUE) of the common effect, or the unique minimum variance unbiased estimator (UMVUE) under the additional assumption of normality of effects estimates [37, Chapter 5]. When the standard errors must be estimated, a normal approximation based on the asymptotic distribution of the estimator is often used [40, Chapter 6]; in the common situation where all the studies are large, the standard errors are known with great accuracy, and any non-asymptotic inefficiency is minor.

For the fixed effects approach it was shown in Lin and Zeng [63] that the analysis

provides, in many situations, a statistically efficient estimate of the parameter that would be estimated, were it possible to pool the data across studies and to perform a single regression analysis that adjusts for study. However, this pooling is inherently somewhat hypothetical; were it possible to do it there would often be no motivation for use of meta-analysis, and so it may not always be obvious that this parameter is of direct interest.

The random-effects approach has perhaps the least direct connection to optimality; while likelihood-based and fully Bayesian methods in general have guarantees of good large-sample properties, under correct model assumptions, [20; 101], in finite samples there are no such guarantees. Indeed, the finite-sample sensitivity of Bayesian random-effects meta-analysis to choice of priors is well-documented [27; 30; 58; 84] and is a cause for concern in practice [89, Chapter 5].

One of our main objectives in this work is to motivate methods of meta-analysis in a new way, appealing directly to optimality. Specifically, we obtain data-adaptive estimators of an average that is 'best' amongst a broad class, in terms of being estimable with most precision.

1.1 Review of approaches to meta-analysis

Table 1.1 provides a summary of the three main statistical models used for meta-analysis and the estimators that are most typically used. Further details are presented in Sections 1.1.1 and 1.1.2 (common- and fixed-effects approaches), and Section 1.1.3 (random-effects approach).

Table 1.1: Statistical assumptions from three different approaches to meta-analysis of k studies, their target parameters for location summary and estimators.

Common assumption	$\hat{\beta}_i \sim N(\beta_i, \sigma_i^2)$, with σ_i^2 known, for $i = 1, 2, ..., k$		
Approach-specific	Common effect	Fixed Effects	Random Effects
assumption	$\beta_i = \beta_0 \ \forall i,\ \beta_0 \in \mathbb{R}$	$\boldsymbol{\beta} = (\beta_i, ..., \beta_k)^T \in \mathbb{R}^k$	$\beta_1, ..., \beta_k$ iid $f(\mu, \tau^2)$
Inference target	β_0	$\beta_F = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\beta_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}$	μ
Estimator	$\hat{\beta}_0 = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\hat{\beta}_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}$	$\hat{\beta}_F = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\hat{\beta}_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}$	$\hat{\mu} = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2+\hat{\tau}^2}\hat{\beta}_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2+\hat{\tau}^2}}$
Standard error	$\widehat{\mathrm{SE}}(\hat{\beta}_0) = \sqrt{\frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}}$	$\widehat{\mathrm{SE}}(\hat{\beta}_F) = \sqrt{\frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}}$	$\widehat{\mathrm{SE}}(\hat{\mu}) = \sqrt{\frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2+\hat{\tau}^2}}}$
Heterogeneity	Not present, by assumption	Hypothesis test based on $Q = \sum_{i=1}^k \frac{1}{\sigma^2}(\hat{\beta}_i - \hat{\beta}_F)^2$	$\hat{\tau}^2 = \max\left\{0, \frac{Q-(k-1)}{\sum \sigma_i^{-2} - \sum \sigma_i^{-4} / \sum \sigma_i^{-2}}\right\}$
Consistency	Not evaluated	$I^2 = \frac{Q-(k-1)}{Q}$	$I^2 = \frac{Q-(k-1)}{Q}$

1.1.1 The Precision Weighted Average

Let $\beta_1, \beta_2, ...\beta_k$ be the true effect-sizes from k different studies and let $\hat{\beta}_i$ be the estimate of the true effect β_i, with corresponding standard error σ_i, which we assume known for now. The inverse-variance (or precision) weighted average of the true effect-sizes is given by

$$\beta_F \quad = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2} \beta_i}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}, \tag{1.1}$$

which is a quantity of interest assuming either that the effect-sizes are identical (common-effect model) or that they are different and independent (fixed-effects model); under the common-effect model β_F reduces to β_0; under the fixed effects model, β_F is a weighted average of the effect-sizes β_i, where the weight is proportional to the precision with which each effect-size can be estimated, giving more weight to those that can be estimated more precisely. Rice, Higgins, and Lumley [77] discuss the equivalence of β_F to a population parameter and its interpretation under the fixed effects assumption.

If the standard errors σ_i are assumed to be known, a natural estimator of β_F is given by

$$\hat{\beta}_F = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2} \hat{\beta}_i}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}} \text{ , with } \mathrm{SE}(\hat{\beta}_F) = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}}. \tag{1.2}$$

Under the assumption of fixed-effects and known standard errors, $\hat{\beta}_F$ directly inherits any efficiency properties from the $\hat{\beta}_i$'s, as any other linear combination of the effect-size estimates. Therefore, $\hat{\beta}_F$ can be seen as an unbiased, efficient and/or normally distributed estimator of β_F, if within each study the estimator $\hat{\beta}_i$ can be assumed to be an unbiased, efficient and/or normally distributed estimator of β_i. Optimality of $\hat{\beta}_F$ under a common-effect approach has already been mentioned.

Confidence intervals for β_F are usually built based on a normal approximation, appealing to the large sample properties of the study estimators. Also, transformations of the outcome measure have been recommended, such as normalizations, log-transformations, bias corrections [40], and/or variance stabilizing transformations [25; 80]. The small sample properties of the normal approximation and sensitivity to the assumption of known variances have been studied through simulation studies [7; 40; 62], and some corrections and tests based on more robust test statistics have been proposed [7; 36].

1.1.2 Testing homogeneity and quantifying heterogeneity

In both fixed-effects and common-effect work, it is common to test for homogeneity of the study effects, that is, to test the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_k$. This is the statistical test of a key assumption in the common-effect analysis, while in the fixed-effects analysis the result of the test gives a simple summary of how much heterogeneity is present.

Homogeneity is typically tested using Cochran's chi-squared Q statistic,

$$Q = \sum_{i=1}^{k} \frac{1}{\sigma^2}(\hat{\beta}_i - \hat{\beta}_F)^2. \tag{1.3}$$

Under Normality of the effect estimates ($\hat{\beta}_i \sim N(\beta_i, \sigma_i^2)$), Q is distributed non-central chi-squared with $k-1$ degrees of freedom and non-centrality parameter λ given by

$$\lambda = \sum_{i=1}^{k} \frac{1}{\sigma^2}(\beta_i - \beta_F)^2, \tag{1.4}$$

with β_F as in (1.1). Under the null hypothesis that all the effects are identical, the Q statistic is distributed central χ^2 with $k-1$ degrees of freedom, thus providing a reference distribution to perform a test of homogeneity of effects.

It can be shown that the statistic Q is independent of the $\hat{\beta}_F$ statistic [77], allowing for straightforward control of type I error when testing for both location and heterogeneity is of interest. However, it also has been found that this test of homogeneity has low power when there are few studies [35] and is not adequate to summarize the extent of the heterogeneity present [43].

Other statistics have been proposed to not only test, but quantify the amount of the heterogeneity present, and thus provide a better measure of the consistency between trials [43]. Although these measures have been motivated and derived from a random effects framework, they still have valid interpretation under a fixed-effects framework [77]. The most-frequently used quantity is I^2, which is obtained from Cochran's Q statistic:

$$I^2 = \frac{Q-(k-1)}{Q}. \tag{1.5}$$

This quantity describes "the percentage of total variation across studies that is due to heterogeneity rather than chance" [45].

1.1.3 The random-effects approach

The random-effects approach is based on the assumption that the true study effects, denoted $\beta_1, \beta_2, ..., \beta_k$, are an independent and identically distributed sample from some distribution. The inference is then focused on the parameters of this distribution, like its mean (μ) and variance (τ^2).

With no further assumptions on the distribution of the random effects, an inverse-variance weighted average estimate of μ can be obtained [19; 46], along with an estimate of its standard error:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2+\hat{\tau}^2}\hat{\beta}_i}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2+\hat{\tau}^2}} \text{ , with } \widehat{\mathrm{SE}}(\hat{\mu}) = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2+\hat{\tau}^2}}}. \tag{1.6}$$

We notice that the weights involve the within study variance σ_i^2 and the heterogeneity (or between studies) variance τ^2, for which a moment-based estimator of τ^2 can be obtained:

$$\hat{\tau}^2 = \max\left\{0, \frac{Q-(k-1)}{\sum \sigma_i^{-2} - \sum \sigma_i^{-4} / \sum \sigma_i^{-2}}\right\}, \tag{1.7}$$

known as the DerSimonian-Laird estimator [19] and commonly used in random-effects meta-analysis. A similar moment-based estimator was proposed by [39], which although originally proposed for the meta-analysis of standardized mean differences, can be generalized for other types of outcomes [102].

Under the further assumption that the study effects follow a normal distribution, maximum likelihood (ML) [34; 75] and restricted maximum likelihood (REML) [38; 71] methods can be used to obtain estimates of τ^2 and μ. Although these methods are iterative and do not provide closed form estimates, it should be noticed that both the ML and REML estimators of μ take the same form as in (1.6). A simpler, non-iterative method for estimating τ^2 has been recently proposed [86], which is also based on the assumption of a normal distribution of the study effects. The performance of the different estimation methods has been evaluated and compared, in terms of bias and efficiency [102], as well as coverage probability [103].

A Bayesian approach can be also used for random-effects analysis, with the advantage that in a Bayesian framework the motivation for random effects comes more naturally, from

the exchangeability principle. As stated in [46], exchangeability refers to "a judgment that the treatment effects may be non-identical but their magnitudes cannot be differentiated a priori".

When implementing a Bayesian random-effects analysis, prior beliefs on the overall location and dispersion of the study effects are implemented as prior distributions for μ and τ^2, so that posterior distributions, conditional on the data, can be obtained. The challenge comes with the choice of a prior distribution for μ and τ^2. It is common practice to use vague or non-informative prior distributions, specially when there is a lack of strong beliefs or prior information. However, it has been found that results from random-effects meta-analysis can vary importantly when using different priors for the dispersion parameter, so performing a sensitivity analysis is highly encouraged [89, Chapter 5]. Specifically, it has been shown that there can be important variations in the point estimates for τ^2, as well as the precision of the estimates for μ, from the use of different prior distributions, specially when the number of studies is small [58].

Chapter 2

A NOVEL FIXED EFFECTS APPROACH TO META-ANALYSIS

As discussed in Chapter 1, using the fixed-effects assumption provides a framework that is less restrictive than common-effect, but still straightforwardly facilitates inference conditional on the studies at hand – unlike the random effects approach. However, the fixed-effects approach has been widely dismissed, ignored or misunderstood, and most of the literature focuses on common-effect and random-effects alone, and how to choose between them [8; 16; 37; 40]. We believe that part of the reason for the lack of popularity of the fixed-effects approach is that it does not make completely obvious which parameter is (or should be) targeted for inference. There are many weighted averages or linear combinations of the effects that can provide a summary of their overall location, and that would also make scientific sense [55; 57]. Among them, the precision weighted average (PWA) is only one. It has been shown that, in several useful cases, the PWA provides a statistically efficient estimate of the parameter that would be estimated, were it possible to pool the data across studies and to perform a single regression analysis that adjust for study [63]. However, such results do not hold uniformly [61]. In addition, this pooling is inherently somewhat hypothetical; were it possible to do it there would often be no motivation for use of meta-analysis, and so it may not always be obvious that this parameter is of direct interest.

2.1 An optimal location parameter by model-free criteria

Ideally, the parameters to which inference is targeted should be determined entirely by scientific criteria, i.e. by specific research goals. But realistically, these criteria may not be precise enough to determine a single parameter for inference. In the absence of such criteria, it makes sense to target our inference to a weighted average that is optimal in some way. We propose to estimate the affine combination that can be most precisely estimated,

or equivalently the parameter for which the data provides the most information. We start with a general result:

Lemma 2.1.1. *Let* $\{\boldsymbol{v}^T\boldsymbol{\beta} : \boldsymbol{v} \in \mathbb{R}^k, \boldsymbol{v}^T\mathbf{1}_k = 1\}$ *be the set of all possible affine combinations of the vector of effect-size parameters* $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)^T$ *and let* $\hat{\boldsymbol{\beta}}$ *be the vector of estimates* $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k)^T$ *with covariance matrix* $\boldsymbol{\Sigma}$*. Then the affine combination of the parameter vector* $(\boldsymbol{w}^T\boldsymbol{\beta})$ *for which the corresponding estimator* $(\boldsymbol{w}^T\hat{\boldsymbol{\beta}})$ *has the minimum variance is given by*

$$\boldsymbol{w} \quad = \underset{\boldsymbol{v}:\boldsymbol{v}^T\mathbf{1}_k=1}{\operatorname{argmin}} \left[\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}\right] = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_k}{\mathbf{1}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_k} \tag{2.1}$$

with $Var(\boldsymbol{w}^T\hat{\boldsymbol{\beta}}) = (\mathbf{1}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_k)^{-1}$.

A proof of Lemma 2.1.1 using Lagrange multipliers is provided in Appendix A.1. Under the fixed effects model, with the implicit assumption of conditional independence of the effects estimates ($\hat{\beta}_i \perp \hat{\beta}_k|\boldsymbol{\beta}$, for $i \neq k$), the covariance matrix of $\hat{\boldsymbol{\beta}}$ reduces to a diagonal matrix: $\boldsymbol{\Sigma}$ =diag$\{\sigma_i^2\}$. Using Lemma 2.1.1 and assuming that σ_i^2 is known exactly from each study, we have that the best affine combination of the effect-size parameters is

$$\left(\frac{\mathbf{1}_k^T\text{diag}\{\sigma_i^{-2}\}}{\mathbf{1}_k^T\text{diag}\{\sigma_i^{-2}\}\mathbf{1}_k}\right)\boldsymbol{\beta} = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\beta_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2}} = \beta_F, \tag{2.2}$$

the precision weighted average of the effect-size parameters. In other words, this result says that β_F is the affine combination of the vector parameter $\boldsymbol{\beta}$ for which the corresponding estimator has minimum variance, i.e. the one that can be more precisely estimated.

To show that β_F is a population parameter, we follow Rice, Higgins, and Lumley [77] and express σ_i^2 as $(n_i\phi_i)^{-1}$, where n_i and ϕ_i are the sample size and the Fisher information from each subject on β_i, respectively, in the ith study. Then we can write β_F as

$$\beta_F = \frac{\sum_{i=1}^k n_i\phi_i\beta_i}{\sum_{i=1}^k n_i\phi_i} = \frac{\sum_{i=1}^k \eta_i\phi_i\beta_i}{\sum_{i=1}^k \eta_i\phi_i}, \tag{2.3}$$

where $\eta_i = n_i/\sum n_{i'}$ is the sample size proportion from study i. In the asymptotic regime where $\eta_1, ..., \eta_k$ are fixed, (i.e. the same assumptions as in the earlier work of [63], and indeed most standard asymptotic settings) we can see that β_F is a population parameter: it

is an affine combination of $\boldsymbol{\beta}$ with the weight for each effect-size depending on the variability and relative size of the corresponding study. This particular affine combination is optimal, in the sense that it is the one that can be most precisely estimated.

If we further assume that the estimates of the effects are Normally distributed ($\hat{\boldsymbol{\beta}} \sim N_k(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ =diag$\{\sigma_i^2\}$), then any affine combination will also be distributed Normally ($\boldsymbol{v}^T\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{v}^T\boldsymbol{\beta},\ \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v})$), with the Fisher information for $\boldsymbol{v}^T\boldsymbol{\beta}$ given by $(\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v})^{-1}$, for any $\boldsymbol{v} \in \mathbb{R}^k$ [13]. This means that, from Lemma 2.1.1, the affine combination of the parameter vector $\boldsymbol{\beta}$ that maximizes the information (or equivalently, minimizes the variance) of the corresponding estimate is given by (2.1). In other words, β_F is the affine combination for which, under this distributional assumption, the data provide most information. Furthermore, we can write the inverse of the variance of $\hat{\beta}_F$ as $\mathbf{1}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_k = \sum_i \sigma_i^{-2} = \sum_i n_i\phi_i$, which means that the information that $\hat{\beta}_F$ provides on β_F under this model is exactly the sum of the total information that the estimates provide on the effect-size parameters.

It is important to notice that the result from Lemma 2.1.1 does not depend on any distributional assumptions, so β_F is the best affine combination given by a model-free criteria: the one that can be most precisely estimated. On the other hand, the assumption of Normality provides a framework that allows us to state this result in terms of the Fisher information, which facilitates the extension of this optimality property into a Bayesian framework.

2.2 A new parameter to quantify heterogeneity

Now we propose a parameter that quantifies the heterogeneity of a group of effect-size parameters, which is a natural extension of the location-summary β_F. We define:

$$\zeta^2 = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}(\beta_i - \beta_F)^2}{\sum_{i=1}^k \frac{1}{\sigma_i^2}} = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\beta_i^2}{\sum_{i=1}^k \frac{1}{\sigma_i^2}} - \beta_F^2 = \frac{\sum_{i=1}^k \eta_i\phi_i\beta_i^2}{\sum_{i=1}^k \eta_i\phi_i} - \beta_F^2 \tag{2.4}$$

where β_F is as in (2.2). For fixed sample size proportions $\eta_1, ..., \eta_k$, we can see that ζ^2 is also a population parameter, just like β_F. We can interpret ζ^2 as a weighted average of the squared deviations of each study effect-size from the weighted average effect β_F, where the weights are proportional to the precision (or the proportion of information) associated with each study effect. Consequently, deviations from more precisely estimable study effects

are more heavily weighted. This parameter ζ^2 is a weighted average squared deviation and quantifies the heterogeneity of the effect-sizes.

This parameter can be also motivated from the Cochran's Q statistic, the quantity upon which the classic test of homogeneity is based. Cochran's Q is defined as the numerator of ζ^2, in the first equation in (2.4). But, whereas Q is a test-statistic whose value increases with the sample size [43], ζ^2 is a parameter whose value does not change when the sample sizes increase at the same rate, that is, when $\eta_1, ..., \eta_k$ are held fixed.

2.3 Frequentist estimation

2.3.1 Estimating β_F and ζ_2 with the assumption of known variances

It is common practice in meta-analysis to assume that the sample size in each study is large enough for the variance of the effect estimate to be approximated with negligible error by its estimate. Is therefore common to assume that the variances of the effects estimates, $\sigma_1^2, ..., \sigma_k^2$ are known. Under this assumption, the natural estimator of β_F is given by

$$\hat{\beta}_F = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\hat{\beta}_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2}} \text{ , with } \mathrm{SE}(\hat{\beta}_F) = \sqrt{\frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}}. \tag{2.5}$$

Confidence intervals for β_F are usually built from a Normal approximation, appealing to the large sample properties of $\hat{\beta}_i$ [37]. The properties of (2.5) under the assumption of a common effect are well known, but they have not been as widely discussed under the assumption of fixed-effects. Here we introduce some of these properties.

If we assume that for all i in $1, ..., k$, $\hat{\beta}_i$ is a complete a sufficient statistic for β_i and that $E[\hat{\beta}_i] = \beta_i$, then by the Rao-Blackwell theorem [13, Chapter 7], $\hat{\beta}_F$ is the unique minimum variance unbiased estimator (UMVUE) of β_F. If we further assume that for each i, $\hat{\beta}_i$ is an efficient estimator of β_i, i.e. we can write $\sigma_i^2 = (n_i\phi_i)^{-1}$, where ϕ_i is the Fisher information for β_i provided by each of the n_i observations in the ith study, then it can be shown that (2.5) reaches the Cramer-Rao lower bound for the variance of any unbiased estimator of β_F, i.e. that β_F is an efficient estimator of β_F.

For the estimation of the heterogeneity parameter ζ^2, we assume that the variances are known and that the estimators $\hat{\beta}_i$ are efficient, so that $\sigma_i^2 = (n_i\phi_i)^{-1}$ for $i = 1, ..., k$ and we

also define $\Phi = \sum_{i=1}^{k} n_i \phi_i$ as the 'total information'. Under these assumptions and with no further distributional assumptions, a simple moment-based point estimate of ζ^2 is given by:

$$\hat{\zeta}^2 = \frac{\sum_{i=1}^{k} \sigma_i^{-2}(\hat{\beta}_i - \hat{\beta}_F)^2 - (k-1)}{\sum_{i=1}^{k} \sigma_i^{-2}} = \frac{Q-(k-1)}{\Phi} \tag{2.6}$$

A detailed derivation of 2.6 can be found in Appendix A.2. For strictly positive estimation of ζ^2, we can report:

$$\max\left(0, \frac{Q-(k-1)}{\Phi}\right) \tag{2.7}$$

To obtain confidence intervals around our point estimate, we add the assumption of normality of the effect-size estimators ($\hat{\beta}_i \sim N(\beta_i, \sigma_i^2)$), so that the Q statistic has non-central χ^2 distribution with $k-1$ degrees of freedom and non-centrality parameter λ given by

$$\lambda = \sum_{i=1}^{k} \frac{1}{\sigma_i^2}(\beta_i - \beta_F)^2 = \Phi\zeta^2. \tag{2.8}$$

Thus, (2.6) still holds and the variance of our proposed estimator $\hat{\zeta}^2$ is given by

$$\mathrm{Var}[\hat{\zeta}^2] = \frac{1}{\Phi^2}\left[4\sum_{i=1}^{k}\frac{1}{\sigma_i^2}(\beta_i - \beta_F)^2 + 2(k-1)\right]. \tag{2.9}$$

Given the convergence in distribution of a non-central χ^2 random variable to a Normal as either the degrees of freedom or the non-centrality parameter tend to infinity, then an approximate confidence interval for ζ^2 can be constructed using a plug-in estimate of (2.9). We note that the centrality parameter λ can be written as $\Phi\zeta^2 = (\sum_{i=1}^{k} n_i\phi_i)\zeta^2 = N(\sum_{i=1}^{k} \eta_i\phi_i)\zeta^2$, showing that a Normal approximation will be valid for a large enough total sample size N.

Alternatively, we can use methods for constructing exact confidence intervals for the non-centrality parameter of a χ^2 distribution that have been proposed and evaluated previously [53]. For example, a 'central' confidence interval for ζ^2 can be constructed by inverting a probability interval of the non-central χ^2 distribution of Q, this is:

$$\{\zeta^2 : \chi^2_{k-1,\alpha/2}(\Phi\zeta^2) \leq Q \leq \chi^2_{k-1,1-\alpha/2}(\Phi\zeta^2)\} \tag{2.10}$$

A solution specific to (2.10) can be obtained by using the more general methods described in [54].

2.3.2 Analytic estimation of β_F without the assumption known variances

When using the precision weighted average (2.5), it is common practice to assume that the sample size in each study is large enough for the variance of the effect estimate (σ_i^2) to be approximated with negligible error by its estimate (s_i^2) [16], basing test statistics and confidence intervals on the following plug-in estimator:

$$\hat{\hat{\beta}}_F = \frac{\sum_i^k \frac{1}{s_i^2}\hat{\beta}_i}{\sum_i^k \frac{1}{s_i^2}} \text{ , with } \widehat{\text{SE}}(\hat{\hat{\beta}}_F) = \sqrt{\frac{1}{\sum_{i=1}^k \frac{1}{s_i^2}}}. \tag{2.11}$$

The properties of (2.11) in small sample size settings have been studied via simulations, where inflated type I error rates were found for the test of the null hypothesis $H_0 : \beta_F = 0$, due to underestimation of the standard error of $\hat{\hat{\beta}}_F$ [7; 40; 62].

Corrected and alternative test statistics have been proposed [7; 36; 37], but all of them are based on the assumption of a common effect. To illustrate how the potential heterogeneity affects the estimation of $\text{Var}[\hat{\hat{\beta}}_F]$, consider the following expression [37]:

$$\begin{aligned}\text{Var}[\hat{\hat{\beta}}_F] &= \text{E}[\text{Var}(\hat{\hat{\beta}}_F|s_1^2, ..., s_k^2)] + \text{Var}[\text{E}(\hat{\hat{\beta}}_F|s_1^2, ..., s_k^2)] \\ &= \text{E}\left[\left(\sum_i^k \frac{\sigma_i^2}{s_i^4}\right)\Big/\left(\sum_i^k \frac{1}{s_i^2}\right)^2\right] + \text{Var}\left[\left(\sum_i^k \frac{1}{s_i^2}\beta_i\right)\Big/\left(\sum_i^k \frac{1}{s_i^2}\right)\right], \end{aligned} \tag{2.12}$$

under the assumption that $\hat{\beta}_i \perp \sigma_i^2$ for $i = 1, ..., k$. The second term in (2.12) is zero and can be ignored if a common effect is assumed ($\beta_i = \beta_0 \ \forall i$), but should be estimated in a fixed effects approach, where we allow for heterogeneity of the effects.

In the following sections we will propose methods for estimation and/or testing of β_F taking into account the uncertainty in the estimation of the study variances.

Large Sample Size Approximation (LSSA).

To obtain an estimator of $\text{Var}(\hat{\hat{\beta}}_F)$ based on asymptotic Normal distribution, we will first assume that within each study we can express the variance of the effect estimator $\hat{\beta}_i$ as $\text{Var}(\hat{\beta}_i) = \sigma_i^2 = (n_i\phi_i)^{-1}$, and that the estimator s_i^2 of σ_i^2 is of the form $(n_i\hat{\phi}_i)^{-1}$, so that

$$\hat{\hat{\beta}}_F = \frac{\sum_i^k \frac{1}{s_i^2}\hat{\beta}_i}{\sum_i^k \frac{1}{s_i^2}} = \frac{\sum_i^k n_i\hat{\phi}_i\hat{\beta}_i}{\sum_i^k n_i\hat{\phi}_i}, \tag{2.13}$$

and also that (2.12) can be expressed as:

$$\begin{aligned}\mathrm{Var}[\hat{\hat{\beta}}_F] &= \mathrm{E}[\mathrm{Var}(\hat{\hat{\beta}}_F|\hat{\phi}_i, ..., \hat{\phi}_k)] + \mathrm{Var}[\mathrm{E}(\hat{\hat{\beta}}_F|\hat{\phi}_i, ..., \hat{\phi}_k)] \\ &= \mathrm{E}\left[\frac{\sum_i^k (\eta_i\hat{\phi}_i)^2 (n_i\phi_i)^{-1}}{(\sum_i^k \eta_i\hat{\phi}_i)^2}\right] + \mathrm{Var}\left[\frac{\sum_{i=1}^k \eta_i\hat{\phi}_i\beta_i}{\sum_{i=1}^k \eta_i\hat{\phi}_i}\right], \end{aligned} \tag{2.14}$$

where the second term accounts for the uncertainty in the estimation of the variances. For estimators $\hat{\phi}_i$ satisfying:

$$\sqrt{n_i}(\hat{\phi}_i - \phi_i) \underset{d}{\rightarrow} N(0, f_i(\boldsymbol{\theta}_i)), \tag{2.15}$$

where $f_i(\boldsymbol{\theta}_i)$ is a function of the distributional moments of the population(s) in study i (see Appendix A.3 for exact definition and derivation), we have shown the following, by applying the Delta method (details in Appendix A.3):

$$\begin{aligned}\mathrm{E}\left[\frac{\sum_i^k (\eta_i\hat{\phi}_i)^2 (n_i\phi_i)^{-1}}{(\sum_i^k \eta_i\hat{\phi}_i)^2}\right] &\approx \frac{1}{N(\sum_i^k \eta_i\phi_i)} \\ \mathrm{Var}\left[\frac{\sum_{i=1}^k \eta_i\hat{\phi}_i\beta_i}{\sum_{i=1}^k \eta_i\hat{\phi}_i}\right] &\approx \frac{\sum_i^k \eta_i\left(\beta_i - \beta_F\right)^2 f_i(\boldsymbol{\theta}_i)}{N\left(\sum_i^k \eta_i\phi_i\right)^2} \end{aligned} \tag{2.16}$$

As expected for the variance of an estimator, both terms converge to zero as the total sample size N increases. More importantly, both terms converge at the same rate, meaning that the uncertainty produced by not knowing the variances can not be ignored, even with large sample sizes, and needs to be accounted for whenever heterogeneity is suspected. Figure 2.1 illustrates the inflation of asymptotic type I error rates of a test of hypothesis for β_F that would not take into account the uncertainty in the variances, as a function of the amount of heterogeneity, for a simple case (see details in Appendix A, Section A.3.1). This theoretical result has been confirmed empirically in a simulation study, which will be described in Section 2.4.

From (2.14) and (2.16) we propose a Large-Sample Size Approximation (LSSA) of the variance of $\hat{\hat{\beta}}_F$:

$$\mathrm{Var}[\hat{\hat{\beta}}_F] \approx \frac{1}{N\left(\sum_i^k \eta_i\phi_i\right)}\left[1 + \frac{\sum_i^k \eta_i\left(\beta_i - \beta_F\right)^2 f_i(\boldsymbol{\theta}_i)}{\sum_i^k \eta_i\phi_i}\right]. \tag{2.17}$$

Tests of hypothesis and confidence intervals can be based on a normal approximation using a plug-in estimator of (2.17) with the estimates of β_i, ϕ_i and $\boldsymbol{\theta}_i$ for $1 \leq i \leq k$ and β_F. Further

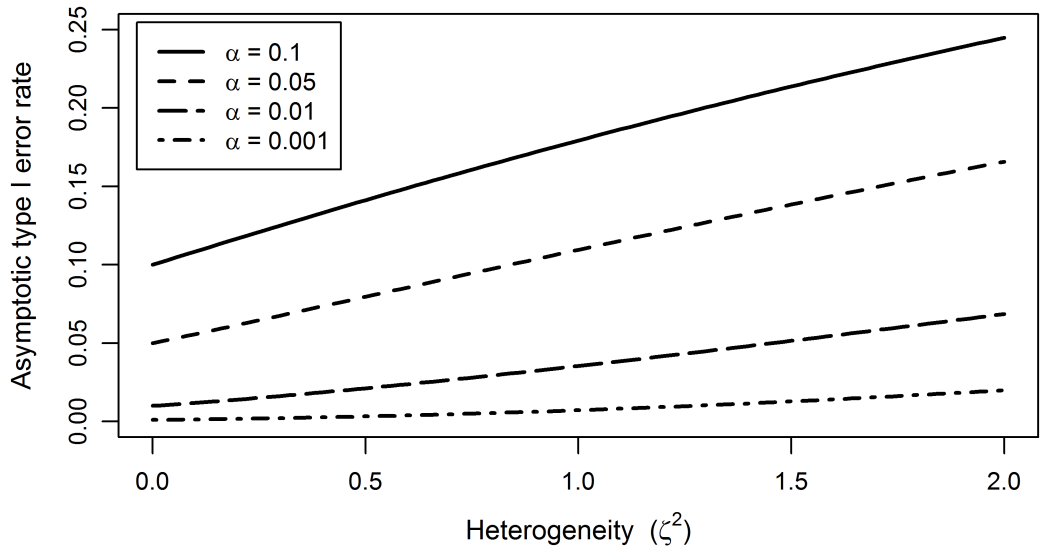

Figure 2.1: Inflated asymptotic type I error rate for the test of hypothesis $H_0 : \beta_F = 0$ in the presence of heterogeneity, when using a naïve estimator of the variance (2.11), according to (2.16) for a simple case of difference in means of continuous normal outcome with constant variance and balanced study designs (see details in Appendix A, Section A.3.1).

details on the specific form of $f_i(\boldsymbol{\theta}_i)$ in (2.17) for some common effect-size estimators are provided in Appendix A.3.

Quasi-F test statistic.

We propose constructing a test statistic for the null hypothesis $H_0 : \beta_F = 0$ similar to [36]. It is based on a 'quasi-F' test statistic, a statistic that approximates an F-distributed random variable [76].

For now we assume that the variances of the effect estimates ($\sigma_i^2 = (n_i\phi_i)^{-1}$) are known and fixed, and that the effect estimates are Normally distributed ($\hat{\beta}_i \sim N(\beta_i, \sigma_i^2)$, for $1 \leq i \leq k$). Under these assumptions, we have that

$$\frac{\hat{\beta}_F^2}{\mathrm{Var}[\hat{\beta}_F]} \sim \chi_1^2. \tag{2.18}$$

Under the same assumptions, the Q statistic is distributed non-central χ^2:

$$Q \sim \chi^2_{k-1}(\lambda), \text{ with } \lambda = \sum_{i=1}^{k} n_i \phi_i (\beta_i - \beta_F)^2 = \Phi\zeta^2 \tag{2.19}$$

A rough approximation to a non-central χ^2 distribution can be made by matching its moments to a scaled central χ^2 [41; 70]. Thus we can approximate the distribution of Q as a $\alpha-$scaled central χ^2 distribution with ν degrees of freedom ($\alpha\chi^2_\nu$), where

$$\begin{aligned} \alpha &= 1 + \frac{\lambda}{(k-1)+\lambda} = 1 + \frac{\Phi\zeta^2}{(k-1)+\Phi\zeta^2} \\ \nu &= (k-1) + \frac{\lambda^2}{(k-1)+2\lambda} = (k-1) + \frac{(\Phi\zeta^2)^2}{(k-1)+2\Phi\zeta^2} \end{aligned} \tag{2.20}$$

Under our stated assumptions, Q and $\hat{\beta}_F$ are independent [36; 77], so that the following quotient

$$\frac{\hat{\beta}_F^2 / \mathrm{Var}[\hat{\beta}_F]}{Q/\alpha\nu}, \tag{2.21}$$

has an approximate F^1_ν distribution, and its signed square root has an approximate Student-t distribution with ν degrees of freedom. Given that none of the quantities included in (2.21) are actually known, a 'quasi-F' statistic can be constructing by plugging-in estimators of all those quantities. Thus let $\hat{\hat{\beta}}$ be as in (2.11), $\widehat{\mathrm{Var}}[\hat{\hat{\beta}}_F]$ the large sample-size approximation given in (2.17), along with plug-in estimates of Q, ζ^2 and Φ, used in turn to estimate α and ν. Then, taking the square root, the following test statistic:

$$t = \sqrt{\frac{\hat{\alpha}\hat{\nu}\hat{\hat{\beta}}_F^2}{\widehat{\mathrm{Var}}[\hat{\hat{\beta}}_F]\hat{Q}}} \tag{2.22}$$

is expected to have an approximate Student-t distribution with $\hat{\nu}$ degrees of freedom under the null hypothesis $H_0 : \beta_F = 0$. This reference distribution would importantly differ from a standard Normal for small values of $\hat{\nu}$, which would be expected when the meta-analysis includes few studies (small k) and the total amount of information times the amount of heterogeneity is small (i.e. approaching the limit where $\Phi\zeta^2 \to 0$).

2.3.3 ***Parametric Bootstrap***

The alternative estimators described in Section 2.3.2, which take into account the potential heterogeneity of the effect-size parameters, are based on approximations that would be ex-

pected to work in large sample settings, but would probably perform poorly in settings with very small size samples. An alternative method, that could perform better in small sample size settings is bootstrap re-sampling, when individual-level observations are available. As this is not likely in practice, we consider the alternative of using parametric bootstrap sampling.

Estimates of the variance of $\hat{\hat{\beta}}_F$, as well as 95% confidence intervals and/or p-values for testing of hypothesis can all be obtained from parametric sampling, based on the estimates $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ and $s_1^2, s_2^2, ..., s_k^2$. Assuming a Normal distribution of the effect sizes estimates, a parametric bootstrap sample of size B for each of the effect-size parameters β_i can be obtained:

$$\hat{\beta}^*_{i[b]} \sim N(\hat{\beta}_i, s_i^2), \text{ for } i = 1, ..., k; b = 1, ..., B. \tag{2.23}$$

On the other hand, parametric sampling for the variances of the effect estimates would depend on the specific variance estimator used in each study. For example, for the variance of the difference in means of independent groups where equal variances are assumed, a bootstrap sample of $\hat{\sigma}_i^2$ can be obtained as:

$$\hat{\sigma}^{2*}_{i[b]} = \frac{\hat{\varsigma}^{*2}_{i[b]}}{n_i} \quad \text{with} \quad \hat{\varsigma}^{2*}_{i[b]} \sim \frac{\hat{\varsigma}_i^{\,2}}{n_i - 2}\chi^2_{n_i-2}, \text{ for } i = 1, ..., k; b = 1, ..., B, \tag{2.24}$$

where $\hat{\varsigma}_i^2$ is the pooled estimate of the common variance ς_i^2 [9]. More generally, for estimates from linear regression (where Normality and constant variance are assumed), the sampling can be done from a χ^2 distribution with $(n_i - p_i)$ degrees of freedom, where p_i denotes the number of predictors in the regression (including the intercept). In contrast, when β_i is estimated as the difference in means of independent groups with the variances not assumed to be equal, the parametric sampling of $\varsigma^2_{i,X}$ and $\varsigma^2_{i,Y}$ should be done separately and then combined to obtain the value of σ_i^2. Further details on the specific form of some of these estimators can found in Appendix C.

From the parametric bootstrap samples of effect-size and variance estimators, different estimates and/or test statistics can be obtained. We propose (and evaluate) the following:

1. A pivotal $(1 - \alpha)$% confidence interval based on a normal approximation and using

an estimate of the variance of $\hat{\hat{\beta}}_F$ from a bootstrap sample:

$$\hat{\hat{\beta}}_F \;\pm\; \xi_{1-\alpha/2}\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}^*_{F[b]} - \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^*_{F[b]}\right)^2},$$

$$\text{where } \hat{\beta}^*_{F[b]} = \frac{\sum_{i=1}^{k}\frac{1}{\hat{\sigma}^{2*}_{i[b]}}\hat{\beta}^*_{i[b]}}{\sum_{i=1}^{k}\frac{1}{\hat{\sigma}^{2*}_{i[b]}}} \tag{2.25}$$

2. A $(1-\alpha)\%$ confidence interval constructed from the percentiles of the empirical distribution of the bootstrap sample of $\hat{\beta}^*_{F[b]}$, as defined in (2.25)

$$\left(\hat{\beta}^*_{F(\alpha/2)}, \hat{\beta}^*_{F(1-\alpha/2)}\right)$$

3. A Bootstrap-t confidence interval, based on the percentiles from the distribution of a test statistic constructed using a 'naïve' estimator of the variance of $\hat{\hat{\beta}}_F$:

$$\left(\hat{\hat{\beta}}_F - t^*_{(1-\alpha/2)}\sqrt{\left(\sum_{i=1}^{k}\frac{1}{s_i^2}\right)^{-1}}, \hat{\hat{\beta}}_F - t^*_{(\alpha/2)}\sqrt{\left(\sum_{i=1}^{k}\frac{1}{s_i^2}\right)^{-1}}\right),$$

$$\text{where} \quad t^*_{[b]} = \frac{\hat{\beta}^*_{F[b]}}{\sqrt{\left(\sum_{i=1}^{k}\frac{1}{\hat{\sigma}^{2*}_{i[b]}}\right)^{-1}}}. \tag{2.26}$$

4. A Bootstrap-t confidence interval, based on the percentiles from the distribution of a test statistic constructed using the LSSA estimator of the variance of $\hat{\hat{\beta}}_F$, as given in (2.17):

$$\left(\hat{\hat{\beta}}_F - t^*_{(1-\alpha/2)}\sqrt{\widehat{\mathrm{Var}}[\hat{\hat{\beta}}_F]}, \hat{\hat{\beta}}_F - t^*_{(\alpha/2)}\sqrt{\widehat{\mathrm{Var}}[\hat{\hat{\beta}}_F]}\right),$$

$$\text{where} \quad t^*_{[b]} = \frac{\hat{\beta}^*_{F[b]}}{\sqrt{\mathrm{Var}^*_{[b]}[\hat{\hat{\beta}}_F]}} \tag{2.27}$$

Similar approaches are proposed for the heterogeneity parameter ζ^2, based on a bootstrap sample of the estimator proposed in (2.6):

$$\hat{\zeta}^{2*}_{[b]} = \frac{\sum_{i=1}^{k}\hat{\sigma}^{-2*}_{i[b]}(\hat{\beta}^*_{i[b]} - \hat{\beta}^*_{F[b]})^2 - (k-1)}{\sum_{i=1}^{k}\hat{\sigma}^{-2*}_{i[b]}}, \text{ for } b = 1, ..., B. \tag{2.28}$$

2.4 Simulation study

We conducted a simulation study to evaluate and compare the performance of the different estimators of β_F and ζ^2 proposed in sections 2.3.1-2.3.3. Our simulation settings consisted of meta-analysis with a small-to-medium number of studies ($k = 3, 5, 7, 15$) with fixed pre-specified effect-sizes ($\beta_1, ..., \beta_k$) that were evenly spread and centered around zero (so that $\beta_F = 0$) and with their absolute value determined by a pre-specified value of ζ^2 (0, 0.1, 0.4, 1, 2). All studies were set to have the same sample size ($n = 10, 20, 30, 40, 60, 80, 100, 500, 1000$) and the same population variance ($\varsigma_i^2 = 1$). A continuous Normal outcome was assumed, with the effect size given by the mean difference between two groups assuming equal variances and a balanced design. Under these assumptions, estimates of the effects were drawn from a Normal distribution

$$\hat{\beta}_i \sim N\left(\beta_i, \left(\frac{1}{n_i/2} + \frac{1}{n_i/2}\right)\varsigma_i^2\right) = N(\beta_i, 4/n_i), \text{ for } i = 1, ..., k;$$

while estimates of the variance where obtained from a scaled χ^2 distribution:

$$s_i^2 \sim \left(\frac{1}{n_i/2} + \frac{1}{n_i/2}\right)\left(\frac{\varsigma_i^2}{n_i - 2}\right)\chi^2_{n_i-2} = \frac{4}{n_i(n_i-2)}\chi^2_{n_i-2} \text{ for } i = 1, ..., k.$$

From each simulation we obtained the estimator $\hat{\hat{\beta}}_F$, as given in (2.11), along with 95% confidence intervals based on the following methods:

(a) A normal approximation using a 'naïve' estimator of $\widehat{SE}(\hat{\hat{\beta}}_F)$ (2.11) that does not account for the uncertainty in the estimation of the variances; also equivalent to assuming a common effect model.

(b) A normal approximation using the Large Sample Size Approximation (LSSA) for the variance of $\hat{\hat{\beta}}_F$ as given in (2.17), with the asymptotic variance of ϕ_i estimated by $(\kappa_i - 1)/4^2\hat{\varsigma}_i^4$ (see details in Appendix A.3.1) and assuming a normal outcome ($\kappa_i = 2$, for $i = 1, ..., k$).

(c) A Student-t approximation based on the signed square root of the quasi-F statistic (2.22), with the degrees of freedom estimated by $\hat{\nu}$ (2.20).

(d) Parametric bootstrap estimation methods, as described in 1-4 from Section 2.3.3, from 1000 bootstrap samples.

In addition to these, we also calculated the 95% confidence interval for the mean of the population of effects when using a random effects approach, based on the DerSimonian-Laird [19] estimator.

Estimates and 95% confidence intervals for ζ^2 were also computed from the simulated samples. Four different methods were implemented:

(a) A normal approximation using a plugin estimate of the variance of $\hat{\zeta}^2$ as given in (2.9), with s_i^2 replacing σ_i^2.

(b) Inverting the probability interval of a non-central χ^2 distribution as described in (2.10), with s_i^2 replacing σ_i^2.

(c) A normal approximation using an empirical estimate of the variance of $\hat{\zeta}^2$ from a parametric bootstrap sample of $\hat{\zeta}^2$ (2.28).

(d) A percentile based interval from the empirical distribution of the parametric bootstrap sample of $\hat{\zeta}^2$ (2.28).

The coverage probability of the estimators was calculated from a total of 10,000 simulations for each combination simulation settings. Results on the coverage probability for 95% confidence intervals for β_F are presented in Figures 2.2 and 2.3 while results on 95% confidence intervals for ζ^2 are presented in Figure 2.4. The estimated Monte Carlo error is represented in each graph as a gray bar around the nominal 0.95 coverage probability.

For the estimation of the location parameter β_F, we observed a better performance of parametric bootstrap methods over those based on asymptotic approximations, especially in small sample size settings. Among these, the confidence interval based on the percentiles of the empirical distribution of the parametric sample is recommended. This approach is simple and performed well, providing coverage close to nominal level. However, we also notice that the large sample size approximation (LSSA) method performed reasonably well for large

sample sizes (at least 60 subjects per study) and could be used if/when the parametric bootstrap could not be implemented.

In addition to evaluating the proposed estimation methods for β_F, we also compared their performance with methods typically used in meta-analysis, i.e. the common effect and random effects approaches. Although these approaches technically estimate different parameters, they all target a location parameter, which in our particular simulation setting has the same value in all three models, deliberately. As expected, the random effects approach (using the DerSimonian-Laird estimator of μ) provided over-conservative inference, as a result of wide confidence intervals that would account for the random sampling of effect sizes (that were actually set fixed in this simulation setting). On the other hand for the common effect estimator, which is equivalent to use a naïve estimate of the variance of β_F as given in (2.13), the coverage probability approaches the nominal level as the sample size increases but never reaching it in the presence of heterogeneity. The asymptotic coverage of this naïve estimator has been calculated using (2.17), and is shown as dotted horizontal lines in Figure 2.2.

For the heterogeneity parameter ζ^2, although all the proposed methods seemed to achieve nominal coverage probability asymptotically, none of them performed uniformly better for small sample size settings in all scenarios (Figure 2.4). The normal approximation with an estimate of the standard error showed both significant over-coverage and under-coverage in different scenarios. The normal approximation using a Bootstrap estimate of the standard error seemed to correct the under-coverage in some scenarios, but not when the number of studies was small ($k = 3$), while the bootstrap confidence intervals based on the percentiles showed important under-coverage for low values of heterogeneity and large number of studies ($k = 7, 15$). This result is consistent with a previous result, in which the consistency of bootstrap estimation is related to the asymptotic normality of the statistic [49; 67], while in our case the distribution of the statistic is far from normal (for small sample size and low level of heterogeneity). On the other hand, given the more consistent performance of the inverted probability interval from a non-central χ^2 distribution, we would recommend its use when the sample sizes are large enough (at least 40 observations per study) and the studies are not too heterogeneous.

2.5 Example

In this section we illustrate the estimation methods discussed in Section 2.3 applied on an example. Our example is from a systematic review of studies that evaluate the efficacy of zinc in reducing the incidence, severity and duration of common cold symptoms [87]. In this particular meta-analysis the authors included studies that compare zinc acetate lozenges with placebo, with the outcome being the duration of cold symptoms (in days) and the treatment effect measured by the mean difference (MD). A forest plot is shown in Figure 2.5.

In Table 2.1 we summarize the results of meta-analyses on the 6 studies comparing zinc lozenges to placebo, using three different approaches. We observe that the point estimates of β_0 and β_F from the common effect and fixed effects approaches, respectively, although numerically the same ($\hat{\beta}_0 = \hat{\beta}_F = -2.04$ days), estimate different parameters. The first estimates a common effect underlying all six studies, but given the evident heterogeneity between studies, this inference does not seem to be adequate, or even valid. On the other hand, $\hat{\beta}_F$ estimates a weighted average of the mean differences from the 6 studies, for which a significant amount of heterogeneity is observed, as reflected by the estimate of ζ^2. More specifically, $\hat{\beta}_F$ estimates the mean difference in duration of common cold averaged in a meta-population composed of the populations from which the samples of these 6 studies were drawn, in proportions given by $\sigma_i^{-2}/\sum_i^k \sigma_i^{-2}$. Similarly, ζ^2 can be thought as estimating how far apart the mean differences in two of these populations are, averaged over the same meta-population. We also observe that the results from different estimation methods, although not exactly the same, do not differ significantly, with a difference in length of 0.13 days between the 95% confidence intervals using the LSSA and the Parametric Bootstrap.

Finally, the random effects meta-analysis estimates the mean and variance of a population from which the mean differences from the 6 studies are thought to have been drawn (μ and τ^2). The inference now is not made for population of subjects (on whom we wish to estimate an average effect of a treatment) but for a population of potential average treatment effects. As shown in Table 2.1, different methods for estimating the between-studies variance give notably different results, with larger estimates of τ^2 yielding estimates of μ

that are closer to the un-weighted simple average of the study effects (-0.56). Moreover, the precision with which these parameters are estimated is much smaller when compared to the precision with which β_F and ζ^2 are estimated, even after taking into account the uncertainty in the estimation of the variances. This gain in precision, should be noted, is not a result of a particular choice of estimation method, but rather, it is the result of targeting our inference to a parameter that is easier to estimate, i.e. one for which the data provide most information.

Table 2.1: Three approaches to the meta-analyses on the effect of zinc acetate lozenges, estimated as the mean difference in the duration of symptoms of the common cold (in days) [87], with point estimates and 95% confidence intervals obtained from different methods of estimation.

	$\hat{\beta}_0$ (95% CI)	
Common effect approach	-2.04 (-2.45, -1.64)	
Fixed effects	$\hat{\beta}_F$ (95% CI)	$\hat{\zeta}^2$ (95% CI)
Large Sample Size Approximation (LSSA)	-2.04 (-2.48,-1.60)	
Quasi-F based Student-t	-2.04 (-2.50, -1.58)	
Non-Central χ^2 inverted PI		2.09 (1.09, 3.50)
Parametric Bootstrap (B=2000)	-2.04 (-2.54, -1.53)	2.09 (1.14, 3.78)
Random effects	$\hat{\mu}$ (95% CI)	$\hat{\tau}^2$ (95% CI)
DerSimonian-Laird	-1.21 (-2.69, 0.28)	2.81 (1.19, 46.0)
Maximum Likelihood	-1.21 (-2.69, 0.28)	2.79 (0, 6.53)
Restricted Maximum Likelihood	-1.13 (-2.83, 0.57)	3.78 (0, 9.25)
Sidik-Jonkman	-1.02 (-3.06, 1.01)	5.66 (2.20, 34.04)

2.6 Final remarks

We have provided a simple justification for the precision weighted average estimator in meta-analysis, without relying on the assumption of a common effect. We have shown that the PWA estimates the best affine combination of the true effect-sizes, i.e. the one that can be more precisely estimated. Also, to complete this approach under a fixed effects framework, we proposed the estimation of a parameter that quantifies the level of heterogeneity present among the effects from the different studies.

Frequentist methods for the estimation of both the location parameter β_F and the heterogeneity parameter ζ^2 were proposed, including corrected estimators that take into account the uncertainty in the estimation of the within study variances. Estimation methods based on asymptotic approximations, as well as methods based on parametric bootstrap were implemented and have been evaluated in a simulation study.

In the results of our simulation study, we observed a better performance of parametric bootstrap methods over those based on asymptotic approximations for the estimation of the location parameter β_F, specially in small sample size settings. Among these, the confidence interval based on the percentiles of the empirical distribution of the parametric sample would be recommended, because of its simplicity and good performance. However, we also notice that the large sample size approximation (LSSA) method performed reasonably well for large sample sizes ($n \geq 60$, per study) and could be used if the parametric bootstrap can not be implemented.

For the heterogeneity parameter ζ^2, although no method performed uniformly best, the construction of 95% confidence intervals by inverting the probability interval from a non-central χ^2 distribution seems to provide close to nominal coverage when the sample size is large enough (around 40 observations per study).

The main limitation in our simulation study is the fact that the proposed methods were implemented under the same distribution from which they data was simulated (Normal), and we did not explore model misspecification. Although the assumption of a Normal distribution is specially important for parametric bootstrap, we believe that, at least asymptotically, this assumption is fulfilled most of the times.

We also illustrated the results of different estimation methods, as well as different approaches, with a previously published meta-analysis. This example, along with the results of our simulation study, provide evidence that the fixed effects approach proposed in this paper is a valid alternative to the typically used common effect and random effect approaches. Our approach, based on the estimation of both a location and a heterogeneity parameter, is more flexible than the restrictive common effect approach while allowing inference on the population of interest. Our approach also makes it unnecessary to choose between statistical models based on their adequacy rather than the target of inference.

Finally, although we believe that estimation of both β_F and ζ^2 are useful for describing and combining in a meaningful way the effects of studies included in a meta-analysis, we propose their estimation only as part of a full battery of qualitative and quantitative tools that should be used to review, summarize and synthesize a group of studies. No one parameter or estimator can possibly summarize all there is to say in a systematic review of medical studies, and this limitation should be acknowledged in practice.

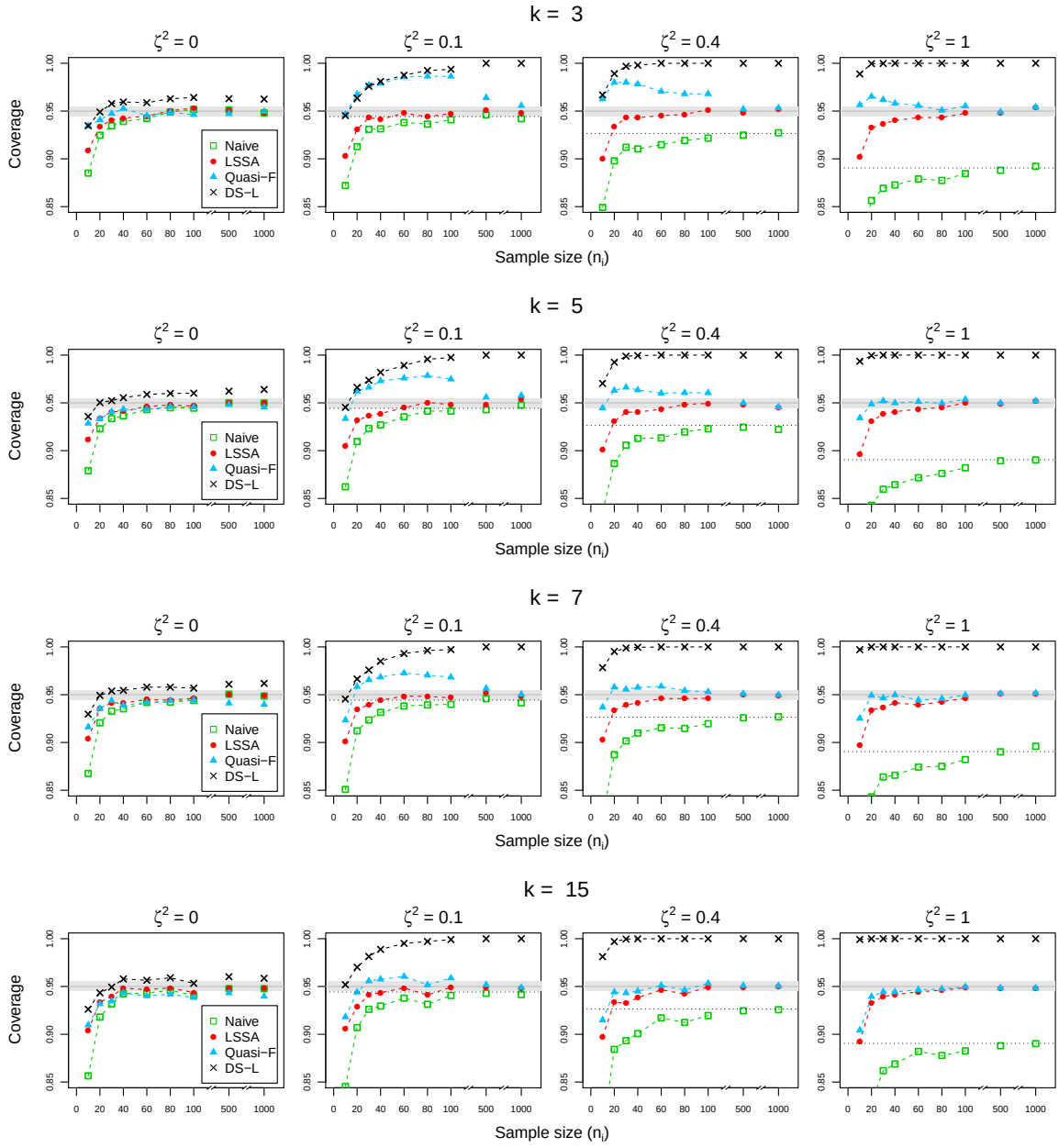

Figure 2.2: Coverage probability of 95% confidence intervals for β_F in a meta-analyses of k studies with various levels of heterogeneity of the fixed effects (ζ^2), using (a) a 'naïve' estimator of the standard error, (b) the large sample size approximation (LSSA) estimator of the standard error and (c) the t-statistic from the quasi-F approach, along with (d) the DerSimonian-Laird estimator for the mean of random effects. The dotted horizontal line represents the asymptotic coverage for the 'naïve' estimator, calculated analytically. These results were obtained from 10,000 simulations, with the gray bar reflecting the approximate Monte Carlo error.

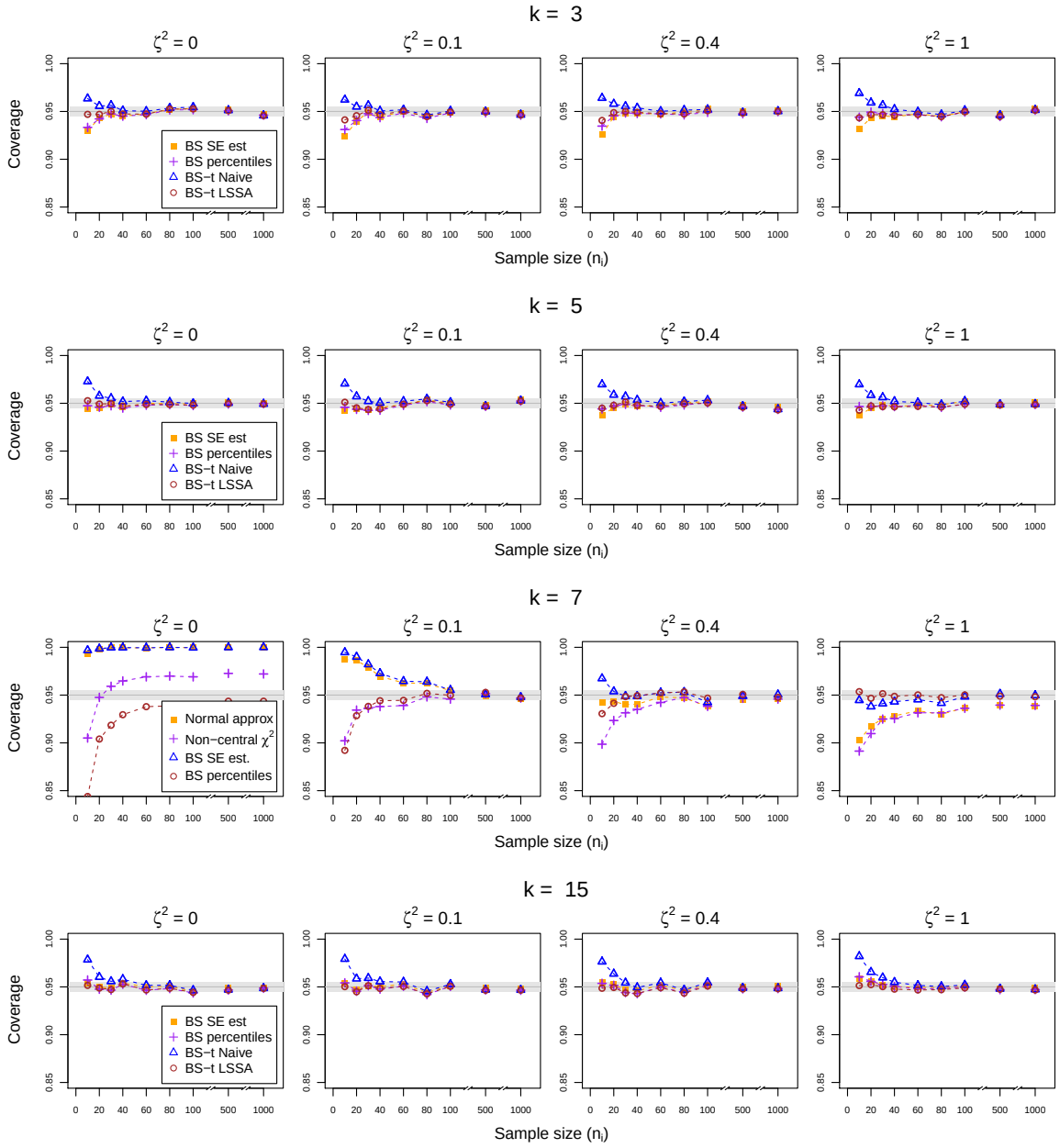

Figure 2.3: Coverage probability of 95% confidence intervals for β_F in meta-analyses of k studies with various levels of heterogeneity of the fixed effects (ζ^2), obtained from parametric bootstrap samples of size 1000: (a) a normal approximation using an empirical estimate of standard error, (b) the percentiles of the empirical distribution, (c) Bootstrap-t based on a t-statistic using a naïve estimation of the standard error and (d) Bootstrap-t based on a t-statistic using the LSSA estimation of the standard error. These results are from 10,000 simulations, with the gray bar reflecting the expected Monte Carlo error

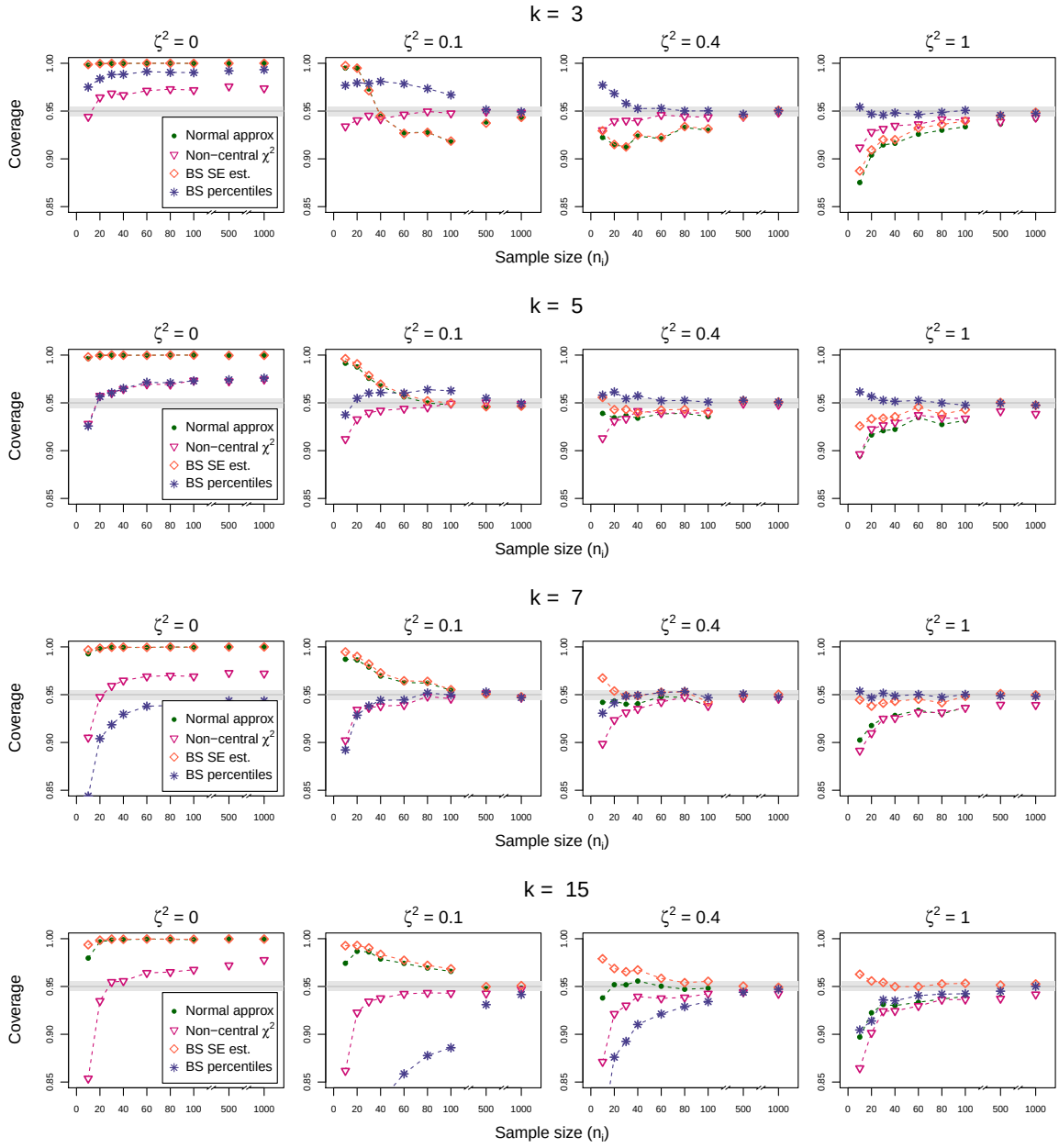

Figure 2.4: Coverage probability of 95% confidence intervals for ζ^2 in a meta-analyses of k studies with various levels of heterogeneity of the fixed effects (ζ^2), using: (a) a normal approximation using a plug-in estimate of the variance of ζ^2, (b) an inverted probability interval from a non-central χ^2 distribution, (c) a normal approximation using an empirical estimate of the standard error from a parametric Bootstrap sample of size 1,000 and (d) the percentiles of the empirical distribution from the same parametric Bootstrap sample. These results are from 10,000 simulations, with the gray bar reflecting the expected Monte Carlo error.

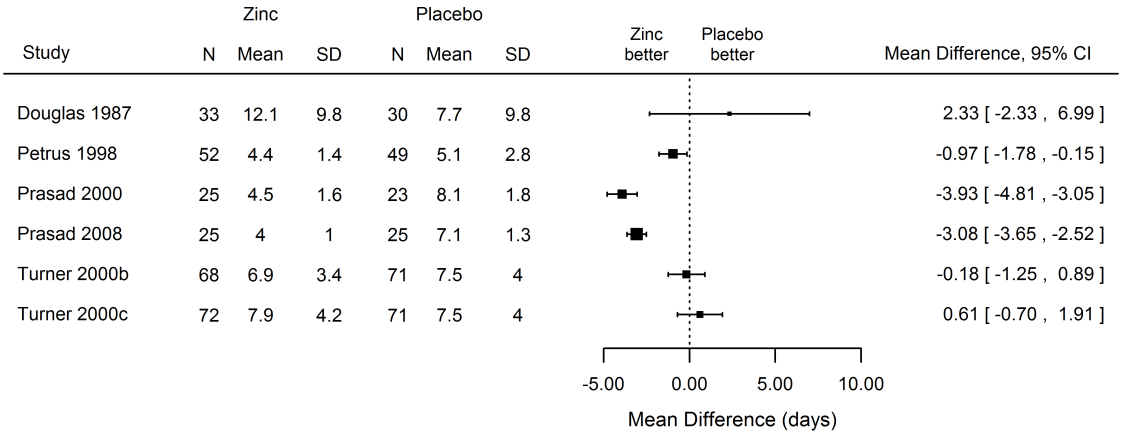

Figure 2.5: Meta-analysis on the efficacy of zinc acetate lozenges in reducing the duration of cold symptoms [87].

Chapter 3

A BAYESIAN FIXED EFFECTS APPROACH TO META-ANALYSIS

In the previous chapter, we proposed a fixed-effects approach to meta-analysis which does not rely on the homogeneity assumption. It consists of estimating a weighted average of the effect sizes to summarize their overall location (β_F), as well as a weighted average of the corresponding squared deviations to quantify the amount of heterogeneity (ζ^2). We have shown that the proposed parameters β_F and ζ^2 are 'population' parameters and that the inverse-variance weighted average is a statistically optimal choice. It is the affine combination of the effects that can be more precisely estimated, and the one for which the data provide the most information if a distributional assumption is made.

This same approach can be taken within the Bayesian paradigm, in which prior information or beliefs on the effect sizes from the studies are included in the meta-analysis, in the form of a multivariate prior distribution, and are updated using the observed data via regular Bayesian procedures. Weighted averages of these effect size parameters (or their deviations) can then be estimated from the updated information, to provide sensible summaries of the location and heterogeneity of the effect sizes included in the meta-analysis. And as stated before, an optimal weighted average could be the one for which the data provide the most information.

This chapter is structured as follows: in Section 3.1 we propose a Bayesian analog to the fixed-effects approach proposed in Chapter 2 using a family of conjugate multivariate priors and providing a closed form expression for the posterior distribution of the parameters of interest. In Section 3.2 we reconcile this approach with the use of a multilevel hierarchical model, recurrently used in the classic Bayesian random effects approach, thus allowing simultaneous estimation of our proposed parameters and the usual targeted hyper-parameters. Lastly, in Section 3.3 we apply the proposed approach to our example meta-analysis on the effect of zinc lozenges on the duration of cold symptoms, which allows us to illustrate and

contrast the properties of the posterior distribution of our proposed parameters with those of the hyper-parameters typically targeted in Bayesian random-effects meta-analysis.

3.1 A Bayesian fixed effects approach to meta-analysis

Using matrix notation, we can express the parameters β_F and ζ^2, respectively, as a linear combination and a quadratic form of the parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)$. Let $\boldsymbol{W} = (\mathbf{1}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{1}_k)^{-1} \boldsymbol{\Sigma}^{-1}$, with $\boldsymbol{\Sigma}$ =diag$\{\sigma_i^2\}$, then

$$\begin{aligned} \beta_F &= \mathbf{1}_k^T \boldsymbol{W} \boldsymbol{\beta}, \\ \zeta^2 &= \boldsymbol{\beta}^T \left(\boldsymbol{W} - \boldsymbol{W} 1_{kk} \boldsymbol{W} \right) \boldsymbol{\beta}. \end{aligned} \tag{3.1}$$

As this notation suggests, it makes sense to state both the model and the prior distribution in terms of the parameter vector $\boldsymbol{\beta}$. For the model, i.e. the assumed distribution of the effect-size estimates $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k)$, the choice of a normal distributions is sensible and close to the truth for large sample sizes. Under the fixed effects models, with the implicit assumption of conditional independence, this translates into a k-variate Normal distribution $(\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}))$, with a diagonal variance-covariance matrix $\boldsymbol{\Sigma}$ =diag$\{\sigma_i^2\}$.

On the other hand, prior beliefs can be incorporated as a multivariate probability distribution of the parameter vector $\boldsymbol{\beta}$. The use of a multivariate prior allows us to incorporate prior beliefs or information not only on each individual effect-size parameter, but also about how much information the effect-size from one study provides on the effect-size of other studies. A multivariate normal prior, for example, can be used to reflect beliefs on the particular location of each effect-size (the mean vector parameter), the uncertainty in these beliefs for each study (the variances) and on how related the effect-sizes from different studies are though to be (the pairwise correlation coefficients). Simply by varying the value of the correlation coefficient between the study effects from 0 to 1, we can specify very different scenarios, ranging from study effects believed to be completely unrelated (fixed effects) to study effects believed to be perfectly related or exactly the same (common effect).

When the belief of exchangeability of the effect-size parameters is reasonable, it can be easily reflected by the exchangeable correlation matrix in, for example, a multivariate

Normal or Student's t-distribution. On the other hand, considering that the study effects are exchangeable is, after de Finnetti [17], equivalent to assuming that the effects were independently drawn from a parametric distribution, with a prior distribution for the hyper-parameter. This leads to a hierarchical structure of the prior distribution, which should be noticed, is "a consequence of the belief in exchangeability rather than a physical randomization mechanism" [89, Chapter 3]. In Section 3.2 we illustrate and discuss the equivalence of a multivariate Normal prior and a hierarchical Normal prior.

Once a model and a prior distribution have been chosen, standard Bayesian methods can be used to obtain a posterior distribution for $\boldsymbol{\beta}$, and thus for β_F and ζ^2. Although the choice of conjugate priors can provide closed-form posterior distributions, computational methods, specifically Markov Chain Monte Carlo (MCMC) algorithms [18], have eliminated the need for conjugate priors. The availability of off-the-shelf software [66; 68] facilitates the implementation of MCMC methods and allows one to obtain results for practically any choice of prior distribution.

3.1.1 Multivariate Normal conjugate prior

In the following sections we present analytical results obtained from the use of a conjugate prior distribution and discuss some properties of the corresponding posterior distribution.

In a fixed effects approach and under the assumption of Normality of the effects estimates $\hat{\beta}_i$, the multivariate Normal constitutes a conjugate prior distribution for the parameter vector $\boldsymbol{\beta}$, with the posterior distribution of $\boldsymbol{\beta}$ conditional on the data $\hat{\boldsymbol{\beta}}$ also being a multivariate Normal [18]. Let the prior for $\boldsymbol{\beta}$ be k-variate normal with mean vector $\boldsymbol{\nu}$ and variance-covariance matrix $\boldsymbol{\Upsilon}$, then:

$$[\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}] \quad \sim \quad N_k\left((\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Upsilon}^{-1})^{-1}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}+\boldsymbol{\Upsilon}^{-1}\boldsymbol{\nu}\right),(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Upsilon}^{-1})^{-1}\right). \tag{3.2}$$

As a linear combination of a k-variate normal distribution, the posterior distribution of β_F is also normal:

$$[\beta_F|\hat{\boldsymbol{\beta}}] \sim N\left(\mathbf{1}_k^T\boldsymbol{W}(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Upsilon}^{-1})^{-1}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}+\boldsymbol{\Upsilon}^{-1}\boldsymbol{\nu}\right),\mathbf{1}_k^T\boldsymbol{W}(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Upsilon}^{-1})^{-1}\boldsymbol{W}\mathbf{1}_k\right). \tag{3.3}$$

The posterior distribution of ζ^2 is a quadratic form of the k-variate Normal vector $[\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}]$:

$$[\zeta^2|\hat{\boldsymbol{\beta}}] = [\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}]^T \left(\boldsymbol{W} - \boldsymbol{W} 1_{kk} \boldsymbol{W}\right) [\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}]. \tag{3.4}$$

Closed form expressions for the mean and variance of $[\zeta^2|\hat{\boldsymbol{\beta}}]$ can be obtained using known results for quadratic forms [83, Chapter 2], while quantiles of the distribution can be obtained by expressing (3.4) as a weighted sum of non-central chi-square random variables, as shown in Duchesne and de Micheaux [21]. The corresponding calculations can be implemented with the `R` package `CompQuadForm` [21].

3.2 Hierarchical prior distributions

A special case of a k-variate normal prior is one with an exchangeable correlation matrix $\boldsymbol{\Upsilon}_k(\rho) = (1-\rho)\boldsymbol{I}_k + \rho \mathbf{1}_{kk}$, where ρ denotes the correlation coefficient between any two effect-size parameters. This specific multivariate prior is equivalent to a hierarchical prior where the study effects β_i are i.i.d. Normal with mean μ and a Normal prior is assigned to the hyper-parameter μ. These two priors and their equivalent parametrizations are shown in Table 3.1.

Table 3.1: Exchangeable multivariate Normal prior for $\boldsymbol{\beta}$ and its equivalent parametrization as a hierarchical structured model

	Multivariate	**Hierarchical**
Prior	$\boldsymbol{\beta} \sim N_k(\nu \mathbf{1}_k, \xi^2 \boldsymbol{\Upsilon}_k(\rho))$	$\beta_i \mid \mu$ iid $N(\mu, \tau^2)$
	$\boldsymbol{\Upsilon}_k(\rho) = (1-\rho)\boldsymbol{I}_k + \rho \mathbf{1}_k \mathbf{1}_k^T$	$\mu \sim N(\nu, \psi^2)$
	$0 \leq \rho \leq 1,\ \xi^2 \geq 0$	$\tau^2 \geq 0, \psi^2 \geq 0$
Parametrization	$\xi^2 = \tau^2 + \psi^2$	$\tau^2 = (1-\rho)\xi^2$
	$\rho = \psi^2/(\tau^2 + \psi^2)$	$\psi^2 = \rho\xi^2$

We notice that both priors reflect the same belief of the study effects being exchangeable: explicitly in the multivariate model and implicitly in the hierarchical model. A careful comparison of both parametrizations allows us to better understand how prior beliefs can be incorporated. For example, prior beliefs on the average location of the study effects are

incorporated through the hyper-parameter ν in both models. Prior beliefs on how similar the study effects are (homogeneity) can be explicitly stated in the value the correlation parameter ρ, with 1 corresponding to a prior assumption of perfect homogeneity while values close to 0 correspond to a prior assumption of unrelated effects. The same correlation can be implicitly incorporated in the hierarchical model as $\psi^2/(\tau^2+\psi^2)$, so that high values of ψ^2 relative to τ^2 induce a high correlation of the study effects. Lastly, the parameter ξ^2 in the multivariate normal is equivalent to the sum of the parameters τ^2 and ψ^2 from the hierarchical model, so we can interpret ξ^2 as a total variance due both to the heterogeneity of the effects and the uncertainty on their overall location.

The posterior distribution of β_F can be obtained from (3.3). But furthermore, by noticing that the total information ($\Phi = \sum_i n_i\phi_i = \sum_i \sigma_i^{-2}$) can be expressed as $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$, so that $\boldsymbol{\Sigma}^{-1} = \Phi\boldsymbol{W}$, then the mean and variance of $[\beta_F|\hat{\boldsymbol{\beta}}]$ can be written as follows:

$$\begin{aligned}
\mathrm{Var}[\beta_F|\hat{\boldsymbol{\beta}}] &= \mathbf{1}_k^T\boldsymbol{W}\left(\Phi\boldsymbol{W}+\xi^{-2}\boldsymbol{\Upsilon}_k(\rho)^{-1}\right)^{-1}\boldsymbol{W}\mathbf{1}_k && (3.5)\\
\mathbb{E}[\beta_F|\hat{\boldsymbol{\beta}}] &= \mathbf{1}_k^T\boldsymbol{W}\left(\Phi\boldsymbol{W}+\xi^{-2}\boldsymbol{\Upsilon}_k(\rho)^{-1}\right)^{-1}\left(\Phi\boldsymbol{W}\hat{\boldsymbol{\beta}}+\xi^{-2}\boldsymbol{\Upsilon}_k(\rho)^{-1}\mathbf{1}_k\nu\right) && (3.6)
\end{aligned}$$

From these expressions we can see that the posterior distribution of β_F is a normal distribution with mean and variance that approach $\mathbf{1}_k^T\boldsymbol{W}\hat{\boldsymbol{\beta}} = \hat{\beta}_F$ and Φ^{-1}, respectively, as the total information ($\Phi = \sum n_i\phi_i$) increases or the total amount of variance due to heterogeneity and uncertainty ($\xi^2 = \tau^2+\psi^2$) decreases. This is, for large sample sizes, or in the absence of strong beliefs on the location of the study effects or their homogeneity, the Bayesian estimator of the precision weighted average reduces to the frequentist estimator.

It is of interest to compare the properties of the Bayesian estimator of β_F to those of the estimator of μ, as this is usually the parameter targeted in Bayesian random effects meta-analyses. The use of the hierarchical prior in Table 3.1 allows to obtain simultaneously the posterior distributions of both parameters, in terms of τ^2 and ψ^2. It can be shown that the posterior distribution of β_F is given by:

$$\begin{aligned}
[\beta_F|\hat{\boldsymbol{\beta}}] &\sim N\left(\sum_{i=1}^k\left(\frac{\sigma_i^{-2}(1-\lambda_i)}{\Phi}\right)\hat{\beta}_i+\left(\sum_{i=1}^k\frac{\sigma_i^{-2}\lambda_i}{\Phi}\right)\nu,\frac{1}{\Phi^2}\sum_{i=1}^k\sigma_i^{-2}(1-\lambda_i)\right)\\
&\text{with}\qquad \lambda_i = \left(\frac{1}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\left(\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right), \qquad (3.7)
\end{aligned}$$

(see Appendix B.2 for a detailed derivation), while the posterior distribution of μ is given by:

$$[\mu|\hat{\boldsymbol{\beta}}] \sim N\left(\left(\frac{1}{\psi^2}+\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}\right)^{-1}\left(\frac{\nu}{\psi^2}+\sum_{i=1}^{k}\frac{\hat{\beta}_i}{\sigma_i^2+\tau^2}\right)\ ,\ \left(\frac{1}{\psi^2}+\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}\right)^{-1}\right). \tag{3.8}$$

First we notice that the posterior mean of β_F can be expressed as a weighted average of the k effect estimates $(\hat{\beta}_1, ..., \hat{\beta}_k)$ and the mean of the prior (μ), meaning that it will never fall out of the range of these values. The same is true for the posterior mean of μ. Also, we notice that the posterior variance of β_F is always less than Φ^{-1}, that is, for any values of ψ^2 and τ^2 in our prior, the precision of the Bayesian estimate of β_F is at least the same precision we get for the frequentist estimator. This is not true for μ, for which large values of ψ^2 and τ^2 in the prior distribution will produce a large variance of the Bayesian estimate.

Table 3.2 shows the limit values of the posterior mean and variance of β_F and μ for some extreme values of ψ^2 and τ^2. The only case when the limit value of mean and variance is the same for both estimators is when using a prior reflecting almost perfect homogeneity $(\tau^2 \to 0)$. In this case, the posterior distribution of both β_F and μ is a weighted average of the inverse-variance weighted average of the estimates $(\hat{\beta}_F)$ and the prior mean (μ), with the weights depending on the information provided by the data (Φ) and the precision of the prior (ψ^2). In contrast, when using a prior that reflects large heterogeneity $(\tau^2 \to \infty)$, the posterior distribution of μ approaches it's prior distribution, with little or no influence from the data, while the posterior distribution of β_F approaches that of the corresponding frequentist estimator.

Table 3.2: Limit values of the mean and variance for the posterior normal distributions of μ and β_F, when using a hierarchical normal prior distribution as in Table 3.1.

Prior distribution			Posterior distribution	
Belief	Value of hyper-parameter	Estimator	Mean	Variance
Homogeneity of the study effects	$\tau^2 \to 0$	$\beta_F \vert \hat{\boldsymbol{\beta}}$	$\frac{\Phi\hat{\beta}_F+\psi^{-2}\nu}{\Phi+\psi^{-2}}$	$\Phi^{-1}\left(\frac{\Phi}{\Phi+\psi^{-2}}\right)$
		$\mu \vert \hat{\boldsymbol{\beta}}$	$\frac{\Phi\hat{\beta}_F+\psi^{-2}\nu}{\Phi+\psi^{-2}}$	$\Phi^{-1}\left(\frac{\Phi}{\Phi+\psi^{-2}}\right)$
Large heterogeneity of the study effects	$\tau^2 \to \infty$	$\beta_F \vert \hat{\boldsymbol{\beta}}$	$\hat{\beta}_F$	Φ^{-1}
		$\mu \vert \hat{\boldsymbol{\beta}}$	ν	ψ^2
Informative prior for the location of study effects	$\psi^2 \to 0$	$\beta_F \vert \hat{\boldsymbol{\beta}}$	$\frac{1}{\Phi}\sum_i \sigma_i^{-2}\left(\frac{\tau^2\hat{\beta}_i+\sigma_i^2\nu}{\sigma_i^2+\tau^2}\right)$	$\frac{1}{\Phi^2}\sum \sigma_i^{-2}\left(\frac{\tau^2}{\sigma_i^2+\tau^2}\right)$
		$\mu \vert \hat{\boldsymbol{\beta}}$	ν	0
Vague prior for the location of study effects	$\psi^2 \to \infty$	$\beta_F \vert \hat{\boldsymbol{\beta}}$	$\hat{\beta}_F$	Φ^{-1}
		$\mu \vert \hat{\boldsymbol{\beta}}$	$\frac{\sum(\sigma_i^2+\tau^2)^{-1}\hat{\beta}_i}{\sum(\sigma_i^2+\tau^2)^{-1}}$	$\frac{1}{\sum(\sigma_i^2+\tau^2)^{-1}}$

When using a diffuse prior for the overall location of the effects ($\psi^2 \to \infty$) the posterior distribution of μ approaches that of the frequentist estimator of μ under random effects model with known between-studies variance. Its variance is always greater than Φ^{-1} and increases with τ^2. For a very small value of τ^2 the mean and variance of $(\mu|\hat{\boldsymbol{\beta}})$ approximate $\hat{\beta}_F$ and Φ^{-1}, respectively, while for a value of τ^2 sufficiently large relative to the study variances (σ_i^2), the mean and variance of $(\mu|\hat{\boldsymbol{\beta}})$ are approximately equal to $\sum_i \hat{\beta}_i/k$ and τ^2/k, respectively. That is, a diffuse prior for μ along with moderate to large heterogeneity will produce a Bayesian estimate of μ that approaches the unweighted average of the effect estimates with its precision directly depending on the number of studies and completely independent of the size or precision of such studies. This is not the case for β_F, for which the posterior distribution under a vague prior reduces again to that of the frequentist estimator, with the precision increasing with the total amount of information ($\Phi = \sum_i \sigma_i^{-2} = N \sum_i \eta_i \phi_i$). In other words, in the absence of strong prior beliefs, more precise estimates of the individual effect-size parameters will produce a more precise Bayesian estimation of β_F, but not of μ.

Finally, when using a very precise or informative prior ($\psi^2 \to 0$), the posterior distribution of β_F approaches that of a precision weighted average of the effect-size parameters, after each one being 'corrected' or 'shrunk' towards a Normal prior with mean ν and variance τ^2. Each correction will depend on the precision of the corresponding estimate (σ_i^2) relative to the precision of that prior (τ^2). We notice that the gain in precision of this estimate relative to the frequentist estimator depends on τ^2, the degree of homogeneity induced by the prior; greater gain will be obtained when τ^2 is small, that is, when the effect sizes are thought to be similar and are then allowed to 'borrow strength' from each other. In contrast, the posterior distribution of μ will again be close to the prior distribution when $\psi^2 \to 0$, with little or no influence of the data.

To better understand the difference between the posterior distributions of β_F and μ, we look further at their means. From (3.8) we can see that the weight for ν in $\mathrm{E}(\mu|\hat{\boldsymbol{\beta}})$ is ψ^{-2} and the weight for each $\hat{\beta}_i$ is $(\sigma_i^2 + \tau^2)^{-1}$. On the other hand, from (3.7), the weights for ν and each of the β_i in $\mathrm{E}(\beta_F|\hat{\boldsymbol{\beta}})$ can be expressed as being proportional to ψ^{-2} and $(\sigma_i^2 + \tau^2)^{-1}\{1 + \tau^2\sigma_i^{-2}[1 + \psi^{-2}(\sum_i \frac{1}{\sigma_i^2+\tau^2})^{-1}]\}$, respectively. We notice then that more weight is given to the effect size estimates $\hat{\beta}_i$ in the posterior mean of β_F than in the

posterior mean of μ. In this sense, we can say that $(\beta_F|\hat{\boldsymbol{\beta}})$ is 'closer' to the data than $(\mu|\hat{\boldsymbol{\beta}})$.

In summary, we can say that priors that reflect certainty and homogeneity have some influence in the posterior distribution of β_F, which translates into a gain in precision (relative to the estimation obtained without any prior information, i.e. frequentist estimator), while priors that reflect uncertainty or heterogeneity have little influence, reducing the posterior distribution to that of the frequentist estimator but with no harm in its precision. On the other hand, priors reflecting certainty and homogeneity will have a much greater influence in the posterior distribution of μ, allowing little or none contribution from the data, while priors reflecting uncertainty and heterogeneity can produce estimates with very poor precision.

3.2.1 Prior beliefs on the between-studies variance

So far, we have considered hierarchical models with the hyper-parameter τ^2 fixed, which induce a k-variate normal prior distribution for the parameter vector $\boldsymbol{\beta}$. However, as it is the standard practice in Bayesian random effects meta-analysis, a prior distribution can be used for τ^2. A fixed effects approach to meta-analysis would not exclude the use of these types of hierarchical priors, whenever the effect-sizes are considered to be exchangeable. However, if we keep in mind that in the fixed effects approach the inference is targeted to functions of the vector $\boldsymbol{\beta}$, then it is evident that we need to understand and illustrate the prior (multivariate) distribution of $\boldsymbol{\beta}$ that is induced by the hierarchical structure that includes priors for both μ and τ^2. Consider the following hierarchical model:

$$\begin{aligned} \beta_1, ..., \beta_k|\mu, \tau^2 &\quad \text{iid} \quad N(\mu, \tau^2) \\ \mu|\nu, \psi^2 &\quad \sim \quad N(\nu, \psi^2) \\ \tau^2|\alpha &\quad \sim \quad U(0, \alpha) \end{aligned} \tag{3.9}$$

Figure 3.2.1 shows the prior distribution of the parameter vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ that is induced by different combinations of Normal priors for μ and Uniform priors for τ^2 as in (3.9). We observe that the marginal distribution of the individual β_i is a bell shaped, with their variance given by $\psi^2 + E(\tau^2)$. However, as it can be observed in the contour plots of the bivariate joint prior of any pair (β_i, β_j), the distribution shows a 'ridge' along the $\beta_i = \beta_j$ line. This is also evident in the 'pointy' marginal distribution of the difference

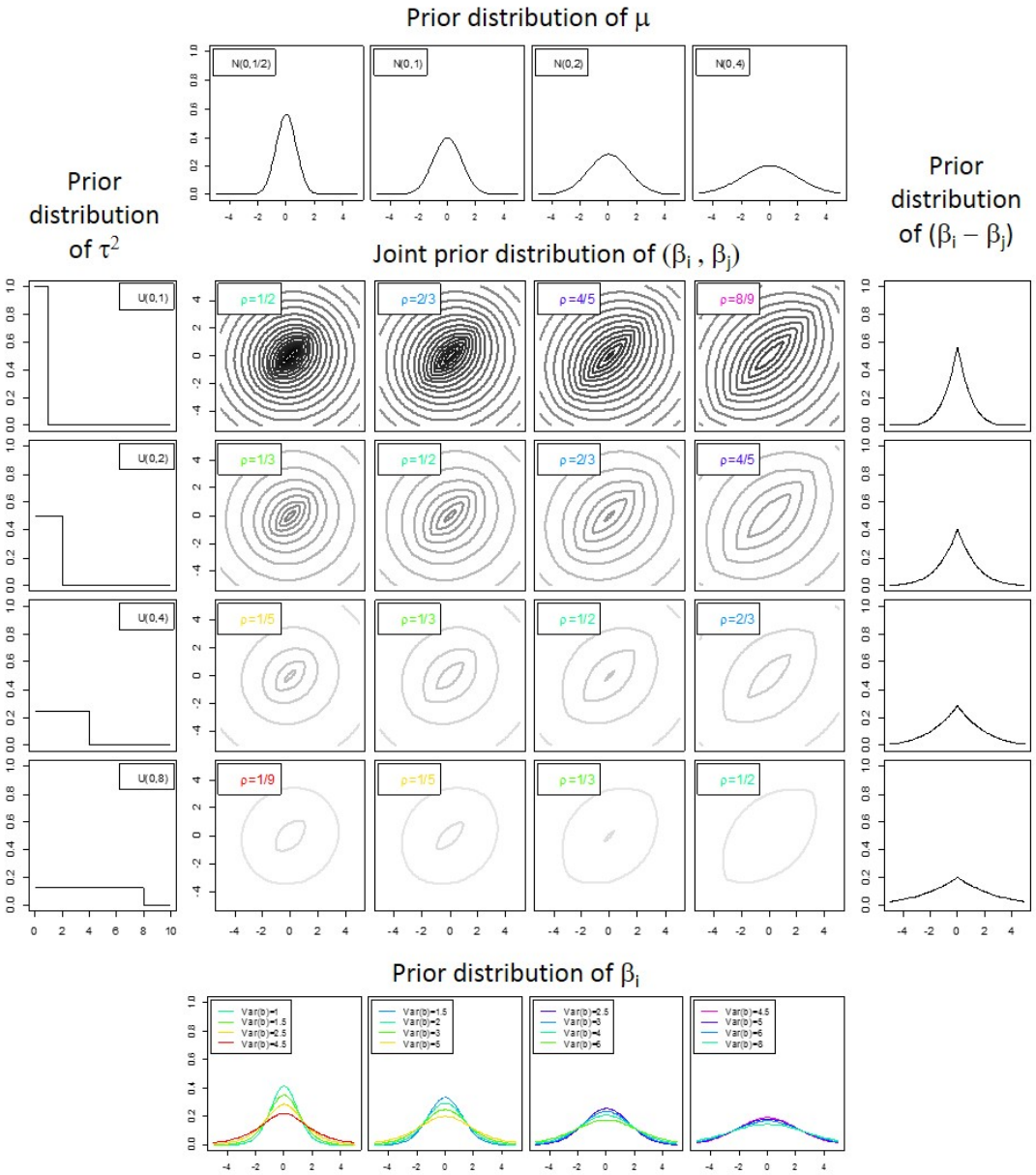

Figure 3.1: Marginal distributions of the effect size parameters $(\beta_1, ..., \beta_k)$ induced by different combinations of priors in μ (top) and τ^2 (left) in a hierarchical model of the form $\beta_i|\mu, \tau^2 \sim N(\mu, \tau^2)$ for $i = 1, ..., k$: the bivariate joint distribution of any pair (β_i, β_j) (middle), the marginal distribution of β_i (bottom) and the marginal distribution of $(\beta_i - \beta_j)$ (right).

$(\beta_i - \beta_j)$. This reveals a prior that more strongly suggests homogeneity of the effect-size parameters, than say a multivariate Normal prior with the same variance and correlation. Similar 'ridge-like' joint distributions of $\boldsymbol{\beta}$ are obtained when using families of distributions other than Uniform as priors for τ^2 (results not shown). And even 'sharper' distributions are obtained when 'vague' or 'diffuse' priors are used for μ, as is commonly done, which induce almost flat priors with very little information on the location of the parameters, but with a very strong suggestion of homogeneity, which in turn produce the shrinkage observed in the updated distributions of the effect sizes.

The use of MCMC sampling methods for the estimation of the posterior distributions for all parameters is well known [18], and will not be discussed here. We just note that the posterior distribution of the parameters β_F and ζ^2, as functions of the parameter vector $\boldsymbol{\beta}$, can be then easily obtained from a random sample of the joint posterior distribution of $\boldsymbol{\beta}$. Sample code (using `R` package `R2WinBUGS`) is shown in Appendix B.3.

A well known characteristic of hierarchical models like the one in (3.9) is the sensitivity of the resulting inference, both for the location and heterogeneity parameters μ and τ^2, to the choice of prior for τ^2 [58]. Because of this, sensitivity analyses are recommended as a routine practice [89]. In the following sections, we use our meta-analysis example to study the sensitivity of the proposed parameters β_F and ζ^2 to the choice of prior.

3.3 Example

In this section we illustrate the Bayesian estimation methods discussed in previous Sections of this Chapter, applied to the example meta-analysis on the efficacy of zinc lozenges in reducing the duration of common cold symptoms (see Figure 2.5).

In Figure 3.3 we present results from frequentist and Bayesian estimation of the proposed parameters β_F and ζ^2 from a fixed effects approach, as well as the parameters μ and τ^2 from a random effects approach. Results of Bayesian analyses include those from selected priors of the family of conjugate Normal distributions in Table 3.1, as well hierarchically structured priors compiled in the paper by Lambert et al. [58]. For the former, the mean and variance of the posterior Normal distribution of β_F and μ were obtained using equations (3.3) and (3.8), respectively, while the mean and variance of the posterior distribution of

ζ^2 were obtained from equations in Appendix B.1, and selected percentiles were obtained using the `R` package `CompQuadForm`.

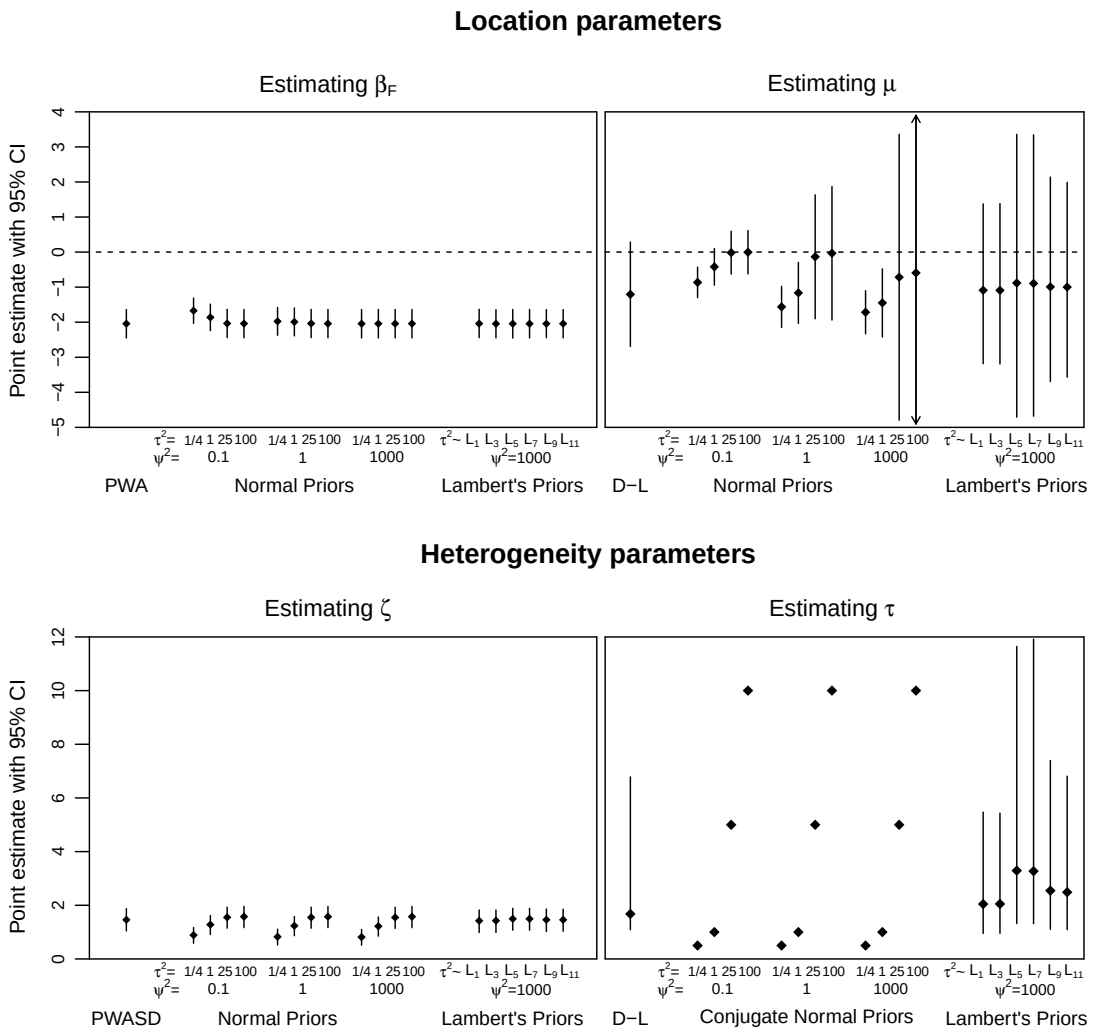

Figure 3.2: Posterior distribution (mean with 95% probability interval) of the location parameters μ and β_F (top) and posterior distribution (median with 95% probability interval) of the heterogeneity parameters $\tau = \sqrt{\tau^2}$ and $\zeta = \sqrt{\zeta^2}$ (bottom), along with frequentist estimates (the FE precision weighted average (PWA, $\hat{\beta}_F$), the precision weighted average squared deviation (PWASD, $\hat{\zeta}^2$), and the RE DerSimonian-Laird (D-L) estimators). Results from hierarchical Normal prior distributions ($\beta_i|\mu \sim N(\mu, \tau^2)$, for $i = 1, ..., k$; $\mu \sim N(0, \psi^2)$)), with a fixed value for the between-study heterogeneity or a diffuse prior distribution taken from Lambert et al. [58] (L1: $\tau^{-2} \sim$ Gamma$(0.001, 0.001)$; L3: $\log(\tau^2) \sim$ Uniform$(-10, 10)$; L5: $\tau^{-2} \sim$ Uniform$(1/1000, 1000)$; L7: $\tau^{-2} \sim$Pareto$(1, 0.001)$; L9: $\tau \sim$Uniform$(0, 100)$; L11:$\tau \sim N(0, 100)$ for $\tau > 0$).

These results illustrate how the point estimate and precision of μ are very sensitive to the choice of prior distribution, and specifically to the level of heterogeneity in the prior. For example, when using a very flat prior for μ ($N(0, 1000)$) with a homogeneous prior for the study effects ($\beta_i \sim N(\mu, 0.25)$) then $[\mu|\hat{\boldsymbol{\beta}}] \sim N(-1.72, 0.31^2)$, while the same vague prior for μ with a more heterogeneous prior for the study effects ($\beta_i \sim N(\mu, 25)$) results in a very different posterior: $[\mu|\hat{\boldsymbol{\beta}}] \sim N(-0.72, 2.08^2)$. This is not the case for β_F, for which the posterior distribution, in both cases, approximates that of the frequentist estimator.

Further results of Bayesian analyses using a hierarchical model with Normal prior for μ and a fixed value of τ^2 (equivalent to a multivariate Normal prior for $\boldsymbol{\beta}$) are shown in Figure 3.3, while results from hierarchical models with a Uniform prior distribution for τ^2 are shown in Figure 3.3. In these plots we show the posterior mean and 95% credible interval of the location parameters μ and β_F, for a range of values of τ^2 (or the hyper-parameter θ that determines its distribution). These correspond to priors that go from close to homogeneous ($\tau^2 \to 0$) to more heterogeneous ($\tau^2 \to \infty$). We also use selected values of ψ^2, which correspond to priors that go from very precise ($\psi^2 = 0.01$) to very flat or vague ($\psi^2 = 1000$). It is evident again that the estimation of μ is more sensitive to the choice of prior than the estimation β_F. The example also illustrates the behavior of the estimates in the limit cases described in Table 3.2. For example, the posterior distribution of μ approaches the distribution of β_F as the heterogeneity decreases, while it approaches either the prior distribution (when ψ^2 is small) or the distribution of the un-weighted average (when ψ^2 is large) as the heterogeneity increases. On the other hand, the posterior distribution of β_F is more stable, i.e. more robust to vagueness and/or heterogeneity in the prior, approaching in such cases the distribution of the frequentist estimator.

As for the quantification of heterogeneity, we can see that the posterior distribution of ζ^2 is, as expected, influenced by prior beliefs on the heterogeneity of the study effects. However, this influence is limited to a range of values τ^2, with the posterior distribution of ζ^2 'stabilizing', as consequence of the posterior distribution of the individual study effects stabilizing around their frequentist estimates. Although a similar 'stabilizing' behavior is observed in the median of the posterior distribution of τ^2, the precision (as reflected by the credible intervals) is importantly sensitive to the choice of prior (See Figure 3.3).

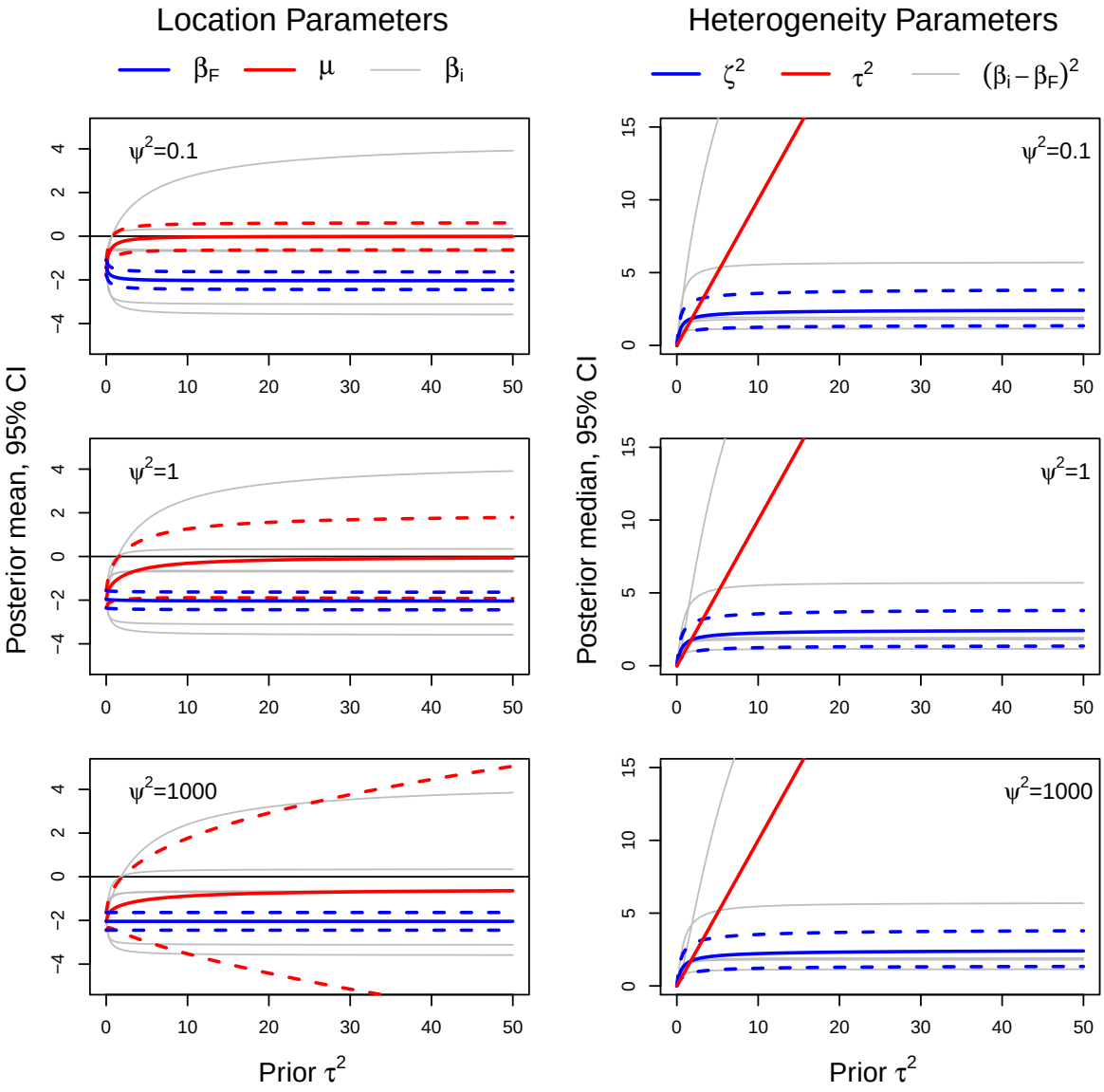

Figure 3.3: Mean and 95% credible interval from posterior distribution of the location parameters β_F and μ (left), and median with 95% credible interval from posterior distribution of heterogeneity parameter ζ^2 along with value of τ^2 (right), as function of the hyper-parameters ψ^2 and τ^2 from the hierarchical Normal prior distribution ($\beta_i|\mu \sim N(\mu, \tau^2)$, for $i = 1, ..., k$; $\mu \sim N(0, \psi^2)$).

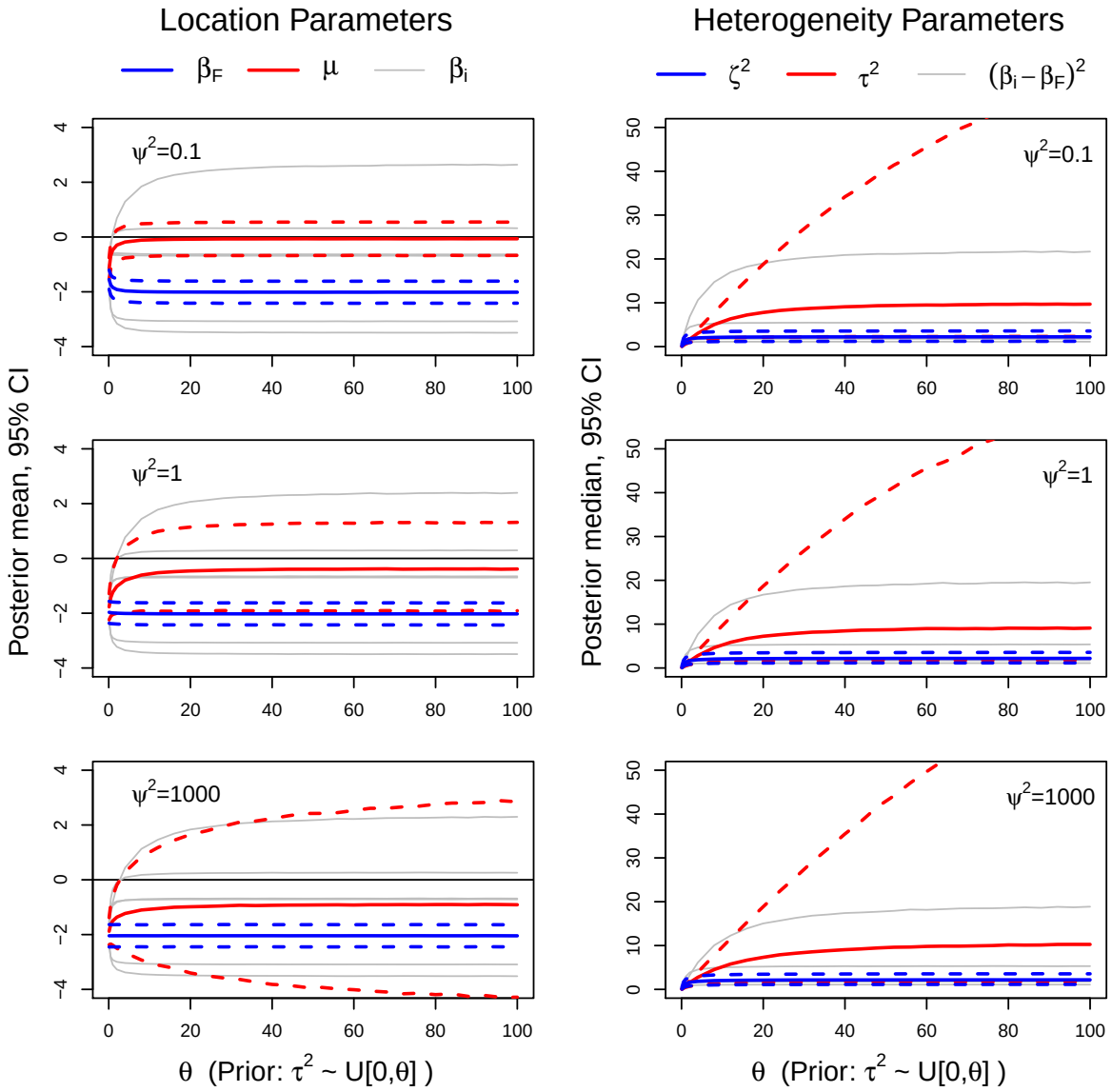

Figure 3.4: Mean and 95% credible interval from MCMC samples of posterior distributions of the location parameters β_F and μ (left), and median and 95% credible interval from posterior distribution of heterogeneity parameter ζ^2 along with value of τ^2 (right). Results are presented for different values of hyper-parameters ψ^2 and θ from the hierarchical model: $\beta_i|\mu \sim N(\mu, \tau^2)$, for $i = 1, ..., k$; $\mu \sim N(0, \psi^2)$); $\tau^2 \sim U(0, \theta)$

3.4 Final remarks

In this chapter, we have proposed and implemented a Bayesian fixed effects approach to meta-analysis, based on the estimation of a precision weighted average, to describe the overall location of the effect-size parameters, along with the estimation of a precision weighted average of their squared deviations, to describe their heterogeneity. The choice of this particular weighted average is also justified within a Bayesian framework by their statistical optimallity among all affine combinations.

Estimation of both the location parameter β_F and the location parameter ζ^2 have been discussed. We also derived and discussed properties of these estimators, not only showing that the fixed effects approach is a valid alternative to the random effects approach in the presence of heterogeneity, but furthermore, that estimation of the proposed parameters can be more precise and stable, and less sensitive to the choice of prior. Such benefits are more evident in meta-analyses with few heterogeneous studies, like in the example included here. We emphasize that both the precision and stability observed do not come from any particular choice of method or prior distribution, but from targeting our inference to a parameter that is 'easier' to estimate, a parameter for which the data provide more information.

Chapter 4

SHRINKAGE AND CLUSTERING IN FIXED EFFECTS META-ANALYSIS, VIA PENALIZED ESTIMATION

In previous chapters we have proposed summarizing the overall location and spread of the effects from the studies included in a meta-analysis through data-adaptive weighed averages of the estimates and their squared deviations. We have shown that targeting our inference to these parameters results in more precise and robust estimation, as these are the parameters for which the data provide more information. Therefore, we have recommended this approach as an alternative to traditional methods used in meta-analysis, whenever there is interest in summarizing the results from all the different studies.

However, when combining results from different studies in a meta-analysis, it can also be of interest to obtain better estimates of all the effect-sizes, improving on the estimation that was obtained in each individual study. Even when the true effect sizes are assumed as fixed, unknown and independent quantities, it is quite possible that combining information from all the studies might lead to better estimators than the ones independently provided. As shown by James and Stein in a seminal paper [50], better predictive accuracy is achieved, on average, when sample means are pooled together (i.e. shrunk), even when they estimate independent and unrelated quantities. Similarly, methods of penalized estimation in linear regression which also produce shrunken estimates, like Ridge Regression or the 'least absolute shrinkage and selection operator' (lasso) [98] have been successfully applied in areas of 'big data' like genomics and proteomics.

In the context of meta-analysis, full Bayesian and empirical Bayesian methods naturally provide shrunken estimates of the individual effects. The updated estimates are said to 'borrow' strength from each other, as this method shrinks the estimates towards the mean of a population of effects, whose distribution is updated using the whole vector of parameters. However, these approaches rely on incorporating prior information and/or a random effects

model. There has not been proposed (to our knowledge) a frequentist approach for shrinkage estimation in meta-analysis, and furthermore, one based on a fixed effects model.

On the other hand, efforts have been made to develop methods in meta-analysis that perform some form of grouping or clustering of effects. This is often motivated by the belief that there is a small number of subgroups of studies with effects that are the same (or very similar) within each subgroup, but different between subgroups. This idea has a special and important application in transethnic meta-analysis of GWAS, where more closely related populations are expected to have similar allelic effects than those from diverse ethnic groups. This is taken into account in the methodology proposed in Morris [69], where some subset structure is incorporated as prior information into a Bayesian analysis. A similar idea is explored in Bhattacharjee et al. [6], where a method is proposed that "exhaustively explores subsets of studies for the presence of true association signals that are in either the same direction or possibly opposite directions".

In this chapter we propose a method for shrinkage estimation in meta-analysis that shares some traits of the continuous shrinkage seen, for example, in Bayesian estimation or ridge regression methods, with grouping or clustering techniques. For this, we propose adapting penalized shrinkage estimation methods, which provide estimates that are continuously shrunk to be more similar (perhaps even identical) to each other, as determined by a tuning parameter. Thus, different degrees of shrinkage will produce estimation from a k-dimensional 'unshrunk' statement of all the $\hat{\beta}_i$, through to a 'fully-shrunk' univariate average effect, given by the precision weighted average $\hat{\beta}_F$. Along this trajectory, it will be also possible to identify different subgroups among the studies, as the various $\hat{\beta}_i$ merge together as the degree of shrinkage increases.

The outline of the chapter is as follows: in the next section we provide a literature review of shrinkage methods, as well as clustering methods. In Section 4.2 we apply and adapt loss functions for penalized estimation to meta-analysis, achieving continuous shrinkage in the form of a solution path, along a range of values of a tuning parameter. We also derive the properties of these solution paths and illustrate them on a meta-analysis example. In Section 4.3 we discuss a method for the estimation of the tuning parameter, which is suitable in the context of meta-analysis. Finally in Section 4.4 we present the results of a simulation

study, where we evaluate the performance of our proposed method, followed by a few final remarks in Section 4.5.

4.1 A brief literature review on shrinkage estimation

4.1.1 The James-Stein shrinkage estimator and Empirical Bayes estimation

Following an important result showing the inadmissibility of the sample mean for multivariate normal distributions with dimensionality $k \geq 3$ [90], James and Stein proposed an estimator for the vector of means from independent normal distributions ($\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)^T$), which had uniformly lower mean squared error than the vector of sample means [50]. Assuming the following simple setting

$$X_i|\theta_i \sim N(\theta_i, 1), \text{ for } i = 1, ..., k \geq 3, \tag{4.1}$$

an improved version of the James-Stein shrinkage estimator (known as the 'positive part James-Stein estimator') is given by:

$$\hat{\theta}_i = \mu_i + \left(\frac{1 - (k-2)}{\sum_{j=1}^{k} (X_j - \mu_j)^2} \right)^+ (X_i - \mu_i),$$

with $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_k)^T$ as an initial guess for $\boldsymbol{\theta}$. In later work, the James-Stein estimator was derived in an empirical Bayes context [23; 64; 91], based on a hierarchical model. By setting a normal prior for $\theta_1, ..., \theta_k$:

$$\theta_i \sim N(\mu, \tau^2), \text{ for } i = 1, ..., k,$$

the posterior means are given by

$$E[\theta_i|X_i] = \mu_i + \left(1 - \frac{1}{1+\tau^2}\right)(X_i - \mu_i), \tag{4.2}$$

which depend on the unknown hyper-parameter τ^2. The James-Stein estimator results when the unbiased estimate of τ^2 given by $(k-2)/S$, is substituted in (4.2) [24]. Stemming from this result, other empirical Bayes shrinkage estimators have been proposed, which use different estimates of τ^2. For example, a maximum likelihood (ML) estimator (with a slight modification) is proposed in [23] for the heteroskedastic case (i.e. unequal variances

of $X_1, ..., X_k$); a simpler ML estimator and an iterative method of moments are used in [11], where also a 'non-parametric' empirical Bayesian approach is proposed. As pointed out in [11], these methods are closely related to the random effects model often used in meta-analysis.

More recent work by Xie et al. [106] proposes a class of shrinkage estimators for heteroscedastic hierarchical models. Their work is based on an important result known as the Stein's unbiased risk estimate (SURE) [92], which has been very influential on shrinkage estimation: when the amount of shrinkage is determined by a 'tuning' parameter (like τ^2 in (4.2)) an optimal value can be selected such that it minimizes the SURE. Using this idea, various classes of estimators are derived, called SURE shrinkage estimators [106], which enjoy good asymptotic optimality properties. The SURE shrinkage estimators include estimators based on a hierarchical model, as well as some more robust non-parametric estimators, that were shown to perform better than the James-Stein and empirical Bayes estimators.

4.1.2 Shrinkage through penalized estimation

In linear model regression, ridge regression is a method used to produce shrunken estimates. The method consists of estimating the regression coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ by minimizing:

$$\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} \beta_i^2, \tag{4.3}$$

with $\boldsymbol{Y} \in \mathbb{R}^n$ denoting the response vector and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the matrix of predictors. By varying the tuning parameter λ, the method provides a continuous path of solutions that shrink from the Ordinary Least Squares (OLS) estimator to zero (when covariates have been centered).

In 1996, Tibshirani introduced the least absolute shrinkage and selection operator (LASSO) method of estimation for linear models [98], in which the sum of the absolute value of the coefficients (rather than their square) is penalized. It has been shown that it is also equivalent to a special form of ridge regression, called adaptive ridge regression [12; 32]. This method tends to produce some coefficients equal to 0, and thus it "tries to retain the good features

of both subset selection and ridge regression" [98]. The LASSO consists of minimizing:

$$\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} |\beta_i|. \tag{4.4}$$

An important extension of this method is the Fused LASSO [99], in which a natural order of the coefficients is assumed, so that by penalizing the neighboring distances it induces shrinkage of neighboring coefficients towards each other:

$$\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_i - \beta_{i+1}|. \tag{4.5}$$

A particular case of this method, which assumes $\boldsymbol{X} = \boldsymbol{I}$, is called the Fused Lasso Signal Approximator (FLSA), and it is widely used, especially in image reconstruction [26]. For one dimensional problems the loss function is:

$$L(\boldsymbol{Y}, \boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}|. \tag{4.6}$$

In higher dimensions, coefficients may correspond to nodes in a graph ($G = (V, E)$), and distances among nodes that are connected in the graph are penalized. This is called the general Fused Lasso Signal Approximator [26; 48].

$$L(\boldsymbol{Y}, \boldsymbol{\beta}) = \frac{1}{2} \sum_{s}^{n} (y_s - \beta_s)^2 + \lambda_1 \sum_{s}^{n} \beta_s^2 + \lambda_2 \sum_{(s,t) \in E, s<t} |\beta_s - \beta_t|. \tag{4.7}$$

Similar loss functions have been proposed as an alternative for performing hierarchical clustering. Recent papers [47; 65] propose formulating the clustering task as a convex relaxation of an optimization problem. As described in [14], given n points $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ in $\mathbb{R}^p$, the aim is to minimize the following 'relaxed' convex criterion:

$$F_\gamma(\boldsymbol{U}) = \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij} \|\boldsymbol{u}_i - \boldsymbol{u}_j\|_p \tag{4.8}$$

where γ is a positive tuning constant, w_{ij} is a non negative weight, and the ith column $\boldsymbol{u}_i$ of the matrix $\boldsymbol{U}$ is the cluster center attached to point $\boldsymbol{x}_i$. A positive weight $w_{ij} > 0$ denotes a connection between nodes i and j. Different norm penalties (ℓ_1,ℓ_2,ℓ_∞) are considered by the different authors on the differences $\boldsymbol{u}_i - \boldsymbol{u}_j$.

4.1.3 Penalizations inducing grouping

Inducing Grouping through a truncated L_1 Norm. As pointed out in Shen and Huang [85], the penalty involving all pairwise differences "is not adaptive for discriminating large from small pairwise differences. As a result, over-penalizing large differences due to shrinking small differences towards zero impedes predictive performance" [85]. To deal with this, they propose the following non convex criteria:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda_1 \sum_{j<j'} G(\beta_j - \beta_{j'}) \\ &\text{where} \qquad G(z) = \begin{cases} \lambda_2 \text{ if } |z| > \lambda_2 \\ |z| \text{ otherwise.} \end{cases} \end{aligned}$$

This penalization involves a new thresholding parameter, λ_2, which helps discriminate the pairs of parameters to be shrunk towards a common value from those that are too far apart. The adaptive grouping is then achieved by the combination of the thresholding and shrinkage parameters. The penalization function G, although non-convex, has some nice computational and statistical properties, including piece-wise linear solution paths.

Inducing Grouping by partially penalizing pairwise differences. Ke et al. [52] have recently introduced the Clustering Algorithm in Regression via Data-driven Segmentation (CARDS). The first two steps of the algorithm consist of finding a preliminary ranking of the coefficients and constructing an ordered segmentation (with segments denoted as B_l for $l = 1, ..., L$) which is determined by a second tuning parameter. They proceed to minimize the following criterion:

$$\begin{aligned} Q_n(\boldsymbol{\beta}) &= \frac{1}{2n}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^2 + P_{\gamma,\lambda_1,\lambda_2}(\boldsymbol{\beta}) \\ \text{with } P_{\gamma,\lambda_1,\lambda_2}(\boldsymbol{\beta}) &= \sum_{i=1}\sum_{i\in B_l, j\in B_{l+1}} p_{\lambda_1}(\beta_i - \beta_j) + \sum_{l=1}^{L}\sum_{i,j\in B_l} p_{\lambda_2}(|\beta_i - \beta_j|) \end{aligned} \tag{4.9}$$

This 'hybrid' penalty function separately penalizes pairwise differences of coefficients between neighboring segments B_l and B_{l+1}, and coefficients within segments. The formulation provides 'intermediate versions' between the fused penalty (when $L = p$) and the exhaustive pairwise penalty (when $L = 1$). By including 'just enough' pairwise differences, this

penalization aims to use enough information to recover the intrinsic grouping (homogeneity) of the coefficients, while avoiding the over-shrinkage produced by including redundant information with the total variation penalty.

4.1.4 Selection of the tuning parameter(s)

In most settings, when shrinkage is performed, either to achieve variable selection or grouping (sparsity in the model, in some form) the main objective is to improve prediction. As such, cross-validation is the most used technique to select the best model. To implement cross-validation, splitting of the sample into various sets (for training, validation and testing) is necessary. The process often involves either minimizing some measure of accuracy, like the mean squared error (MSE), the prediction error [98; 99] or Stein's unbiased risk estimate [11], or maximizing measures of goodness of fit, like the Bayesian Information Criteria (BIC) [52].

4.1.5 Subset methods

So far we have explored shrinkage estimation methods, some of which might be useful in producing some form of hierarchical clustering of the different effect-sizes included in a meta-analysis. At the other end of the spectrum, we find specific methods that perform sub-setting or clustering of the effects in a more 'discrete' way (as opposed to the 'continuous' behavior of shrinkage methos), often involving exhaustive exploration of all possible arrangements. For example in Bhattacharjee et al. [6], an agnostic approach is proposed which allows some of the studies to have no effect (or null effect), and thus identifying the subset of studies with 'non-null' effects in the same direction (one sided) or different subsets of studies with 'non-null' effects in opposite directions (two sided). The method "explores all possible subsets of 'non-null' studies to identify the strongest association signal", and then evaluates the significance of the signal accounting for the multiple tests performed by the subset search. The approach can be seen as discrete in the sense that the number of subsets is pre-specified, and the challenge is then to find the best arrangement of the effect sizes into such subsets. The method has been shown to achieve important gains in power for detecting

association between SNP markers and multiple traits in genome-wide association studies (GWAs), relative to the methods that 'pool' together all the effects.

A similar approach is followed in Morris [69], where the aim is to take into account "the expected similarity in allelic effects between the most closely related populations". The authors use a Bayesian partition model to cluster the effects of similar populations, "in terms of relatedness (i.e. shared ancestry)". By introducing a prior distribution on the number of clusters and assuming that populations in the same cluster have the same underlying allelic effects, the method shows improvement in detecting associations and localizing causal variants, relative to classical meta-analysis approaches.

4.2 Penalized estimation for continuous shrinkage in Meta-analysis

In this section we propose some penalized loss functions for shrinkage and clustering in meta-analysis. Keeping our notation in the meta-analysis context, we will use $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_k)^T$ to denote the vector of estimates of the true effects $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_3)^T$ from the k different studies, with variances given by σ_i^2, for $i = 1, ..., k$, which for now we will assume to be known. In addition, we will use $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, ..., \tilde{\beta}_k)^T$ to denote the minimizer of our loss function, and more specifically $\tilde{\beta}_i(\lambda)$ to denote the solution for a given value of the tuning parameter λ, which will constitute a shrunken estimate of the effect size parameter β_i.

4.2.1 Fused LASSO Signal Approximator (FLSA)

A natural first step is to apply off-the-shelf methods, with slight adaptations to fit the meta-analysis framework. We first consider the loss function used for the one-dimensional fused LASSO signal approximator (FLSA), penalizing only distances between neighboring estimates. We respect the ordering of the effect estimates $\hat{\beta}_i$, thus implicitly assuming that it reflects the natural order the true effects. Given that within the context of meta-analysis there is no particular interest in shrinking estimates towards zero, we exclude penalties on the absolute value of the coefficients ($\lambda_1 = 0$ in equation (4.6)). Hence we obtain:

$$L(\boldsymbol{\beta}, \lambda) = \frac{1}{2} \sum_{i=1}^{k} (\hat{\beta}_i - \beta_i)^2 + \lambda \sum_{i=1}^{k} |\beta_i - \beta_{i+1}|. \tag{4.10}$$

When the tuning parameter λ is zero, no shrinkage is done and $\tilde{\beta}_i = \hat{\beta}_i$ for $1 \leq i \leq k$, while for λ large enough L_λ is minimized at $\tilde{\beta}_i = \frac{1}{k}\sum \hat{\beta}_i = \bar{\beta}$.

By exhaustively penalizing all pairwise distances, we obtain a loss function similar to that of the general FLSA (4.7), which is also the loss function proposed for convex clustering (4.8), using an L-1 norm and all weights equal to 1:

$$L_g(\boldsymbol{\beta}, \lambda) = \frac{1}{2}\sum_{i=1}^{k}(\hat{\beta}_i - \beta_i)^2 + \lambda \sum_{i<j} |\beta_i - \beta_j|. \tag{4.11}$$

To illustrate and compare the solutions obtained from minimizing these loss functions, we use our example of a meta-analysis of clinical trials comparing efficacy of zinc lozenges, compared to placebo, in reducing duration of cold symptoms. To obtain the solutions over a range of values of the tuning parameter, we used the package for convex optimization `SCS` [**?**] and `Convex` [100] in **Julia** [5]. In Figures 4.2.1(a) and 4.2.1(b), we show the solution paths of these two convex criteria applied to the estimates from the example meta-analysis, over a range of values of λ. We note two important properties of the solution paths, which have been studied and exploited for constructing fast algorithms to obtain the solution. One is the piece-wise linearity of the solution, this is, that the rate of shrinkage as function of the tuning λ is piece-wise constant. The second one is that when two or more solutions are equal for a given value of the tuning parameter ($\tilde{\beta}_i(\lambda^0) = \tilde{\beta}_j(\lambda^0)$), then the solutions will remain equal for larger values of the tuning parameter ($\tilde{\beta}_i(\lambda) = \tilde{\beta}_j(\lambda)$ for $\lambda \geq \lambda^0$). Expressed in terms of the solution path, once the trajectories of two estimates are fused together, they will remain fused. Proof of these properties of the FLSA can be found in Friedman et al. [26] and Hoefling [48].

We notice that the general FLSA (Figure 4.2.1(b)), which penalizes all pairwise distances, produces a solution path that suggests a preferable clustering of the effects, as similar estimates are fused together at low values of the tuning parameter. This is not necessarily the case for the one-dimensional FLSA (Figure 4.2.1(a)), which penalizes only distances between neighbors, as we observe that no shrinkage occurs for some estimates at low values of λ, but instead coefficients seem to shrink sequentially, 'from the outside to the inside'.

However, we also note that the solution to the general FLSA, when applied to our example meta-analysis, has two main drawbacks. The first one is that in the limit as λ increases, all the shrunken estimates are equal to the un-weighted mean. As we have shown in previous chapters, the inverse variance weighted average is an optimal summary, and thus a criteria that shrinks all estimates towards this summary would be desirable. The second drawback is that the rate of shrinkage of the estimates depends solely on their relative position (this will be shown more clearly in Section 4.2.3, where we derive the exact rate of shrinkage as a function λ). In a meta-analysis, where we seek to 'borrow strength' across studies in proportion to the amount of information each study provides, estimates from larger studies that are more precise should shrink less, while more variable estimates should be shrunk the most.

4.2.2 Weighted versions of the FLSA

To produce shrinkage towards the precision weighted average (or any other weighted average), weights must be included in the the loss function. Let $v_i = \sigma_i^{-2}/\sum_i^k \sigma_i^{-2}$, with $\sum_i^k v_i = 1$. Then the following convex criterion

$$L_{\boldsymbol{v}}(\boldsymbol{\beta}, \lambda) = \frac{1}{2}\sum_{i=1}^{k} v_i(\hat{\beta}_i - \beta_i)^2 + \lambda \sum_{i<j} |\beta_i - \beta_j| \tag{4.12}$$

will produce estimates that shrink to the precision weighted average, $\hat{\beta}_F = \sum \sigma_i^{-2}\hat{\beta}_i / \sum \sigma_i^{-2} = \sum v_i \hat{\beta}_i$, for large values of λ, as shown in Figure 4.2.1. Also, we can see that less precise estimates seem to be shrunk more, compared to precise estimates, which are shrunk less. However, the introduction of a weighted sum of squares changes some of the properties of the solution path. In particular, estimates that have been fused together do not necessarily remain fused for larger values of λ. We can see this in the solution path shown in Figure 4.2.1(c), where the trajectories of the fifth and sixth estimates (in blue and purple, respectively) split up after remaining fused for a range of values of λ.

We next consider the following convex criteria, which introduces weights on the penalized sum of the pairwise distances between coefficients:

$$L_{\boldsymbol{v},\boldsymbol{W}}(\boldsymbol{\beta}, \lambda) = \frac{1}{2}\sum_{i=1}^{k} v_i(\hat{\beta}_i - \beta_i)^2 + \lambda \sum_{i<j} w_{ij}|\beta_i - \beta_j|. \tag{4.13}$$

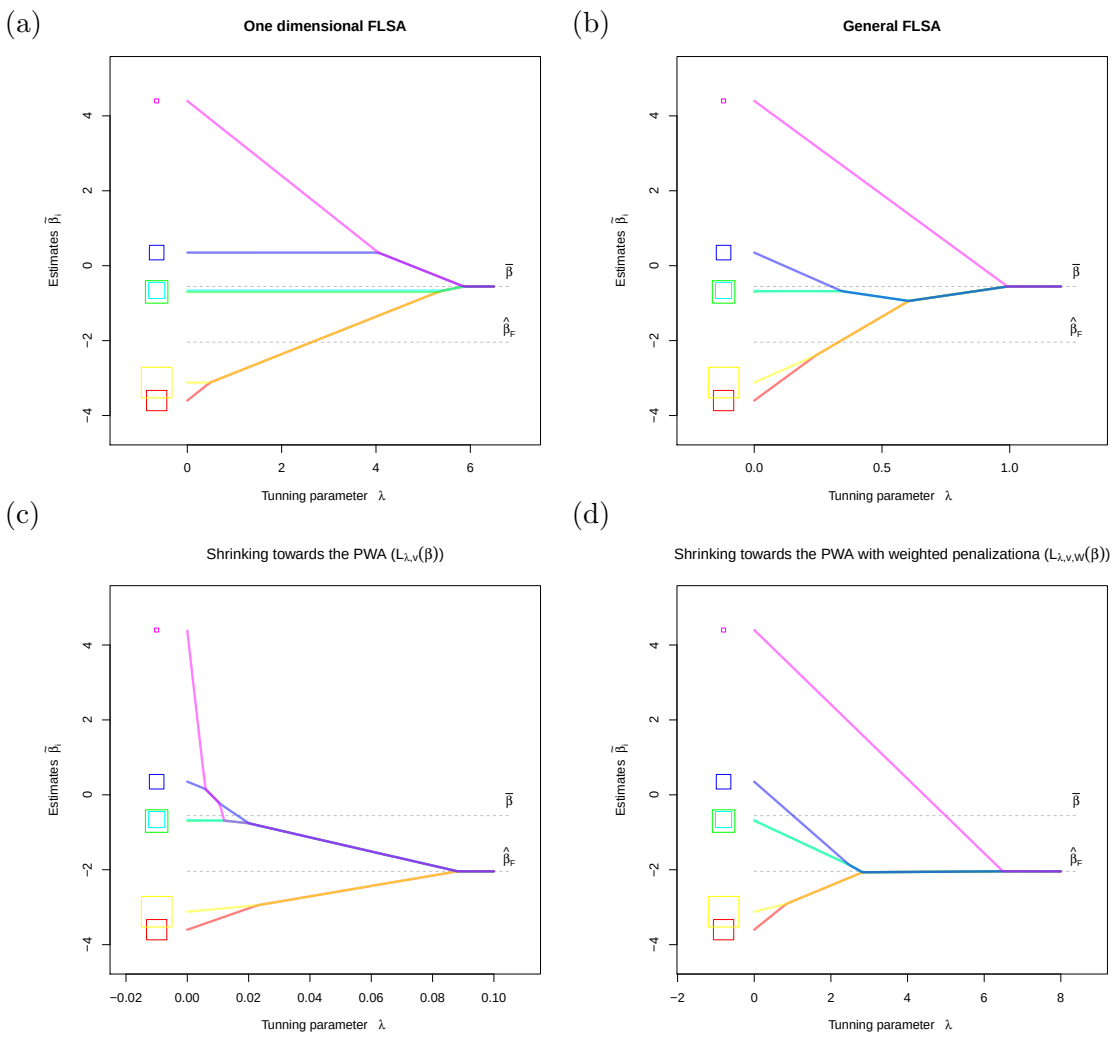

Figure 4.1: Solution path from different convex criteria applied to our meta-analysis example: (a) the one-dimensional fused Lasso signal approximation (FLSA, $L(\lambda, \boldsymbol{\beta})$), which penalizes neighboring distances; (b) the general FLSA, ($L_g(\lambda, \boldsymbol{\beta})$) which penalizes all pairwise distances, (c) the weighted version of the general FLSA, which shrinks towards the precision weighted average ($L_{\boldsymbol{v}}(\lambda, \boldsymbol{\beta})$) and (d) the weighted version of FLSA with weighted penalties ($L_{\boldsymbol{v},\boldsymbol{W}}(\lambda, \boldsymbol{\beta})$). Solutions were obtained using the convex optimization packages `SCS` and `Convex` in `Julia`.

For the particular choice of $\boldsymbol{W} = \boldsymbol{v}\boldsymbol{v}^T$ (i.e. $w_{ij} = v_i v_j$ for $i, j = 1, ..., k$), this loss function can be seen as a simplification of (4.11) applied to individual level data from the k studies, with the study relative sample sizes given by v_i. To illustrate this, assume for now that

$\sigma_i^2 = (n_i\phi)^{-1}$, for $i = 1, ..., k$, so that $v_i = n_i/N$. Then pre-clustering subjects within studies and using $\hat{\beta}_i$ as the estimated effect for all n_i subjects in study i, we get the following loss function:

$$\begin{aligned}
& \frac{1}{2}\sum_{i=1}^{k}\sum_{l}^{n_i}(\hat{\beta}_i - \beta_i)^2 + \lambda\sum_{i<j}\sum_{l}^{n_i}\sum_{l'}^{n_j}|\beta_i - \beta_j| \\
= \; & \frac{1}{2}\sum_{i=1}^{k} n_i(\hat{\beta}_i - \beta_i)^2 + \lambda\sum_{i<j} n_i n_j|\beta_i - \beta_j| \\
= \; & N\left[\frac{1}{2}\sum_{i=1}^{k} v_i(\hat{\beta}_i - \beta_i)^2 + \frac{\lambda}{N}\sum_{i<j} v_i v_j|\beta_i - \beta_j|\right] = L_{\boldsymbol{v},\boldsymbol{v}\boldsymbol{v}^T}(\boldsymbol{\beta}, \lambda').
\end{aligned}$$

In general, v_i can be seen as the proportion of the total information contributed by observations in study i, even when the information per observation ϕ_i is not constant across studies. It will be shown in the following section that for this particular choice of penalization weights, fused coefficients remain fused in the solution path, making it an attractive alternative. However, the amount of shrinkage does not directly depend on the precision of the estimates, as would be desirable, so it is not guaranteed that more variable estimates will shrink the most (this is illustrated in Figure 4.2.1(d)).

4.2.3 Properties of the solution path

In this section we derive some important properties of the solution path for the convex criteria that have been explored in the previous section. These results will be used later on to develop an algorithm for constructing the solution path of our proposed loss functions. We will focus on the generalized weighted version of the FLSA as given in (4.13), from which the loss functions in (4.10)-(4.12) follow as particular cases.

Without loss of generality we assume that $\hat{\beta}_1$, $\hat{\beta}_2$,..., $\hat{\beta}_k$ are ordered, so that $\hat{\beta}_i \leq \hat{\beta}_j$ for $1 \leq i < j \leq k$. Quoting from [48], we introduce the following formal definition for sets of fused solutions:

Definition 4.2.1. *Let $F_s, s = 1, ...n_F(\lambda)$ be the set of estimates at λ that are considered to be fused, where $n_F(\lambda)$ is the number of such sets. In order for these sets to be valid, every*

set F_s has to be of the form $F_s = \{i | l_s \leq i \leq u_s\}$ and the following statements have to hold as well:

- $\cup_{s=1}^{n_F(\lambda)} F_s = \{1, ..., k\}$
- $F_s \cap F_t = \emptyset$, *for* $s \neq t$
- *Assuming the F_s are ordered, for every $i, j \in F_s$ we have $\tilde{\beta}_i(\lambda) = \tilde{\beta}_j(\lambda)$, and for $i \in F_s, j \in F_{s+1}$, then it holds that $\tilde{\beta}_i(\lambda) \neq \tilde{\beta}_j(\lambda)$.*

Let β_{F_s} denote the value of the solution for the fused set F_s, so that $\beta_{F_s} = \beta_i$ for all $i \in F_s$. For given sets of fused coefficients, the value of the solution from minimizing (4.13) is locally a linear function of λ, for as long as the sets of fused coefficients remain unchanged, i.e. no fusion or splitting of coefficients occur. This is stated in the following lemma, which also provides the slope or rate of shrinkage for small increases in λ.

Lemma 4.2.2. *Let F_s, for $s = 1, ..., n_F$ defined as in 4.2.1, and $\beta_{F_s}(\lambda)$ the common value of the fused solutions in F_s from minimizing (4.13). Then for sufficiently small increments of λ,*

$$\frac{\partial \beta_{F_s}(\lambda)}{\partial \lambda} = -\frac{\sum_{t \neq s} \left[\left(\sum_{i \in F_s} \sum_{j \in F_t} w_{ij} \right) sign(\beta_{F_s} - \beta_{F_t}) \right]}{\sum_{i \in F_s} v_i} \tag{4.14}$$

A proof of this lemma is provided in Appendix C.1. Using this result, we can calculate the rates of shrinkage, as functions of the tuning parameters, of the solution paths for the loss functions discussed in Sections 4.2.1 and 4.2.2.

We start with the one-dimensional FLSA (4.10), where $v_i = 1$ for all $i = 1, ..., k$ and $w_{ij} = 1$ if $j = i + 1$, $i = 1, ..., k - 1$ or $w_{ij} = 0$ otherwise. At $\lambda = 0$, $\tilde{\beta}_i(\lambda) = \hat{\beta}_i$, and we have that for $1 < i < k$:

$$\frac{\partial \beta_i(\lambda)}{\partial \lambda} = \begin{cases} -\text{sign}(\beta_1 - \beta_2) = 1 & \text{for } i = 1 \\ -\text{sign}(\beta_{i-1} - \beta_i) - \text{sign}(\beta_i - \beta_{i+1}) = 0 & \text{for } 1 < i < k \\ -\text{sign}(\beta_{k-1} - \beta_k) = -1 & \text{for } i = k, \end{cases} \tag{4.15}$$

which corresponds with what we observed in Figure 4.2.1(a), where the estimates in the middle do not get shrunk at all until they are fused with the outermost estimates. Also,

as more and more estimates are fused, the rate of shrinkage decreases, as given by the denominator $\sum_{i \in F_s} v_i = |F_s|$ in (4.14).

For the general FLSA (4.11), where $v_i = 1, w_{ij} = 1$ for all i, j in $1, ..., k$, we have the following at $\lambda = 0$:

$$\frac{\partial \beta_i(\lambda)}{\partial \lambda} = \sum_{j<i} \text{sign}(\beta_i - \beta_j) + \sum_{j>i} \text{sign}(\beta_i - \beta_j) = -(k-i) + (i-1) = 2i - k - 1. \quad (4.16)$$

This shows that the initial shrinkage of the estimates is given by their relative position in their ordering, with outermost estimates shrinking more than estimates in the middle (Figure 4.2.1(b)). It is the difference in the rate of shrinkage that produces the fusing of estimates, providing the clustering property that we look for. However, as discussed before in Section 4.2.1, this criterion does not take into account the precision of the estimates.

Our proposed weighted version of the general FLSA that shrinks towards the precision weighted average (4.12), with $w_{ij} = 1$, gives the following at $\lambda = 0$:

$$\frac{\partial \beta_i(\lambda)}{\partial \lambda} = \frac{1}{v_i} \left(\sum_{j<i} \text{sign}(\beta_i - \beta_j) + \sum_{j>i} \text{sign}(\beta_i - \beta_j) \right) = \frac{2i - k - 1}{v_i}, \quad (4.17)$$

which corroborates our observations that the initial rate of shrinkage is influenced by the precision of the estimates, with less precise estimates shrinking more than more precise estimates (Figure 4.2.1(b)).

Lastly, the weighted version of the FLSA that also includes weighted penalizations, with $w_{ij} = v_i v_j$ (4.13) gives the following:

$$\frac{\partial \beta_i(\lambda)}{\partial \lambda} = \frac{1}{v_i} \left(\sum_{j<i} v_i v_j \text{sign}(\beta_i - \beta_j) + \sum_{j>i} v_i v_j \text{sign}(\beta_i - \beta_j) \right) = -\sum_{j<i} v_j + \sum_{j>i} v_j, \quad (4.18)$$

which means that the rate of shrinkage depends on a relative 'weighted' position of the particular estimate. For example, if estimates that are below $\hat{\beta}_i$ provide more total information than the estimates above, then $\tilde{\beta}_i$ will be shrunk towards smaller values; on the other hand, if estimates above and below $\hat{\beta}_i$ provide about the same total information then $\tilde{\beta}_i$ will not be shrunk (regardless of its own precision). Under this reasoning, the previous loss function would be more attractive, but as has been illustrated by our example, it is not guaranteed

that fused estimates will remain fused. Next, we present a result that will help determine the form of the loss function for which no splits are produced in the solution path.

Theorem 4.2.3. *Let $\tilde{\beta}_i(\lambda)$ be the optimal solution to the loss function in (4.13), with the particular choice of $\boldsymbol{W} = \boldsymbol{v}\boldsymbol{v}^T$. Then if for some value λ^0 it holds that $\tilde{\beta}_i(\lambda^0) = \tilde{\beta}_{i+1}(\lambda^0)$, then for any $\lambda > \lambda^0$ it holds that $\tilde{\beta}_i(\lambda) = \tilde{\beta}_{i+1}(\lambda)$.*

A proof by contradiction is provided in Appendix C.2.

4.2.4 An algorithm for the weighted FLSA

The results in the previous section are now used to construct a piece-wise linear algorithm to recover the solution path, instead of using more general convex optimization algorithms. This particular algorithm does not account for potential splits of fused sets of estimates, and therefore it will not exactly reproduce the solution path in cases that contain splits.

The details of the algorithm are shown in Appendix C.3. The main output of the algorithm includes the values of λ at which neighboring estimates fuse together, and the value of the vector $\tilde{\boldsymbol{\beta}}$ at these points. In a nutshell, the algorithm does the following: initializing $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ at $\lambda = 0$, the algorithm starts by calculating slopes for each of the estimates' trajectories using (4.14); we denote these slopes as $\boldsymbol{\alpha}$. The value of λ at which the first fusion of coefficients occurs is then obtained by calculating and comparing the intersection points of the trajectories of adjacent coefficients. Here the algorithm exploits the fact that fusion can only occur among neighboring coefficients. Shrunken estimates are obtained for this value of λ and stored and new slopes are calculated. The algorithm is iterated until all fusions have occurred (a maximum of k times).

In Figure 4.2.4 we show some examples of the results obtained from the piece-wise linear algorithm, directly comparing with the solution path obtained using convex optimization algorithms. Discrepancies are only observed when the path contains splitting trajectories.

4.3 Estimating the tuning parameter via Parametric Bootstrap

Traditionally, cross-validation techniques are used to estimate tuning parameters for shrinkage methods via penalized estimation. In the context of meta-analysis, where individual

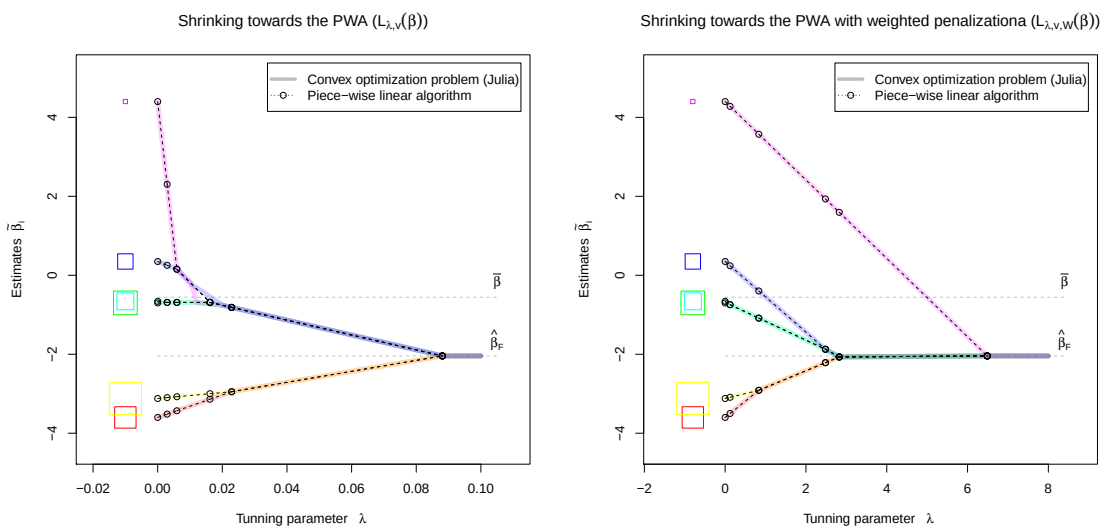

Figure 4.2: Comparing the solution path given by the piece-wise linear algorithm with the solution path resulting from solving the convex optimization problem (using packages `SCS` and `Convex` in `Julia`).

level data is not available, these methods do not seem adequate. Instead, we propose to use a parametric bootstrap procedure to select the value of the tuning parameter λ.

The parametric bootstrap was successfully used to obtain confidence intervals for the estimates of the precision weighted average and the precision weighted average of squared deviations in Chapter 2. Here, we propose using the bootstrap to sample estimates and select the value of the tuning parameter that minimizes a measure of discrepancy with respect to the original estimates, like the Mean Squared Error (MSE) or a weighted version of the MSE. A major advantage of this procedure is that it can not only take into account the uncertainty around the estimates of the effect size parameters ($\hat{\beta}_i$), but also the uncertainty on the estimated variances ($\hat{\sigma}_i^2$).

We describe the proposed procedure in detail. First, we use the parametric bootstrap to sample the estimates $\hat{\beta}_i$ and their variances $\hat{\sigma}_i^2$. In the particular meta-analysis used as an example here, the effect size is estimated as the mean difference of two independent groups, and its variance is estimated not assuming constant variances. A parametric bootstrap

sample of size B would be given by:

$$\begin{aligned}
\hat{\beta}^*_{i[b]} &\sim N(\hat{\beta}_i, s_i^2), \text{ for } i = 1, ..., k; b = 1, ..., B. \\
\hat{\sigma}^{2*}_{i[b]} &= \frac{\hat{\varsigma}^{*2}_{i1[b]}}{n_{i1}} + \frac{\hat{\varsigma}^{*2}_{i2[b]}}{n_{i2}} \text{ , for } i = 1, ..., k; b = 1, ..., B, \\
&\text{with } \hat{\varsigma}^{2*}_{ig[b]} \sim \frac{\hat{\varsigma_{ig}}^2}{n_{ig} - 1}\chi^2_{n_{ig}-1}, \text{ for } g = 1, 2. \qquad (4.19)
\end{aligned}$$

A solution path for a particular loss function can be then constructed for each bootstrap sampled vector of the estimates $(\hat{\beta}^*_{1[b]}, ..., \hat{\beta}^*_{k[b]})$ with variances $(\hat{\sigma}^2_{1[b]}, ..., \hat{\sigma}^2_{k[b]})$. For any given value of λ, shrunken estimates of $\boldsymbol{\beta}$ can be obtained from all the solution paths of the bootstrap sample($\tilde{\beta}^{*[b]}_i(\lambda)$ for $i = 1, ..., k$ and $b = 1, ..., B$), so that a measure of accuracy, relative to the observed data $\hat{\boldsymbol{\beta}}$, can be calculated. We propose both an un-weighted and a weighted version of the MSE:

$$\begin{aligned}
MSE(\lambda) &= \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{k}\sum_{i=1}^{k}(\tilde{\beta}^{*[b]}_i(\lambda) - \hat{\beta}_i)^2\right] \\
&= \frac{1}{k}\sum_{i=1}^{k}\left[\frac{1}{B}\sum_{b=1}^{B}(\tilde{\beta}^{*[b]}_i(\lambda) - \hat{\beta}_i)^2\right] \qquad (4.20) \\
wMSE(\lambda) &= \frac{1}{B}\sum_{b=1}^{B}\left[\frac{\sum_{i=1}^{k}\hat{\sigma}_i^{-2}(\hat{\beta}^{*[b]}_i(\lambda) - \hat{\beta}_i)^2}{\sum_{i=1}^{k}\hat{\sigma}_i^{-2}}\right] \\
&= \sum_{i=1}^{k}\frac{\hat{\sigma}_i^{-2}}{\sum_{i'=1}^{k}\hat{\sigma}_{i'}^{-2}}\left[\frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}^{*[b]}_i(\lambda) - \hat{\beta}_i)^2\right] \qquad (4.21)
\end{aligned}$$

Then we can select the value of λ that minimizes $MSE(\lambda)$ or $wMSE(\lambda)$. In Figure 4.3 we illustrate the process of estimation for three versions of our example meta-analysis of studies on efficacy of zinc in reducing duration of cold symptoms. The first row corresponds to the original data; in the second and third row we show the same data artificially shrunk to achieve smaller heterogeneity ($\zeta^2 = 0.23$ and $\zeta^4 = 0.023$, a tenth and hundredth of the value of ζ^2 in the original data), but keeping the same variances and sample sizes. In the first column we present the solution path for 100 of the Bootstrap samples of the estimates and their variances, along with the solution path of the observed data. In the second column, the individual MSE for each effect estimated is plotted, along with the un-weighted (4.20) and weighted total MSE (4.21). We can see that the former is somewhat driven by the estimate

that is farther from the rest, even when it is estimated with very low precision, while the latter is less influenced by that same estimate. We also observe that as the heterogeneity decreases, the MSE is minimized at larger values of lambda (relative to the 'top' of the solution path), producing a solution with fewer clusters, as shown in the third column of the panel Figure 4.3.

4.4 Simulation Study

A simulation study was conducted to evaluate the performance of the weighted versions of the FLSA for shrinkage and clustering estimation in meta-analysis. We applied the proposed bootstrap technique for estimation of the tuning parameter on three of the convex criteria presented in Section 4.2. These are the general FLSA in (4.11, $L_g(\lambda, \boldsymbol{\beta})$) and the two weighted versions of the general FLSA (4.12, $L_{\boldsymbol{v}}(\lambda, \boldsymbol{\beta})$) and (4.13, $L_{\boldsymbol{v},\boldsymbol{W}}(\lambda, \boldsymbol{\beta})$). In our simulations we implement the piece-wise linear algorithm presented in Section 4.2.4.

Other meta-analysis summaries were calculated and compared in the simulation study, including: (i) a naïve approach of reporting all six estimates with no shrinkage, (ii) reporting a single summary measure, like the simple mean ($\bar{\beta}$), the precision weighted average $\hat{\beta}_F$ or the estimate of the mean of the random effects $\hat{\mu}$, using the DerSimonian-Laird estimate of the between-studies variance [19]; (iii) reporting shrunken estimates as given by Empirical Bayes methods (4.2), using the DerSimonian-Laird estimates.

We use a simple simulation setting. Six studies are simulated, sampling effect estimates from normal distributions, with variances sampled from χ^2 distributions. The same population variance (equal to 1) and the same sample size is assumed in all the studies. We consider three different scenarios with the true effect size parameters being the same within clusters, with (a) two clusters of three studies, (b) three clusters of two studies, evenly spaced and (c) all six studies with different effects, evenly spaced. The values of the effect size parameters are set so that $\beta_F = 0$, with various values of ζ^2 going from 0 to 0.4. We considered a small sample size setting (10 observations per arm in each study) and a large sample size setting (100 observations per arm in each study).

In Figure 4.4 we present the results, in terms of the MSE of the shrunken estimates for three different loss functions and two criteria for optimizing the tuning parameter λ,

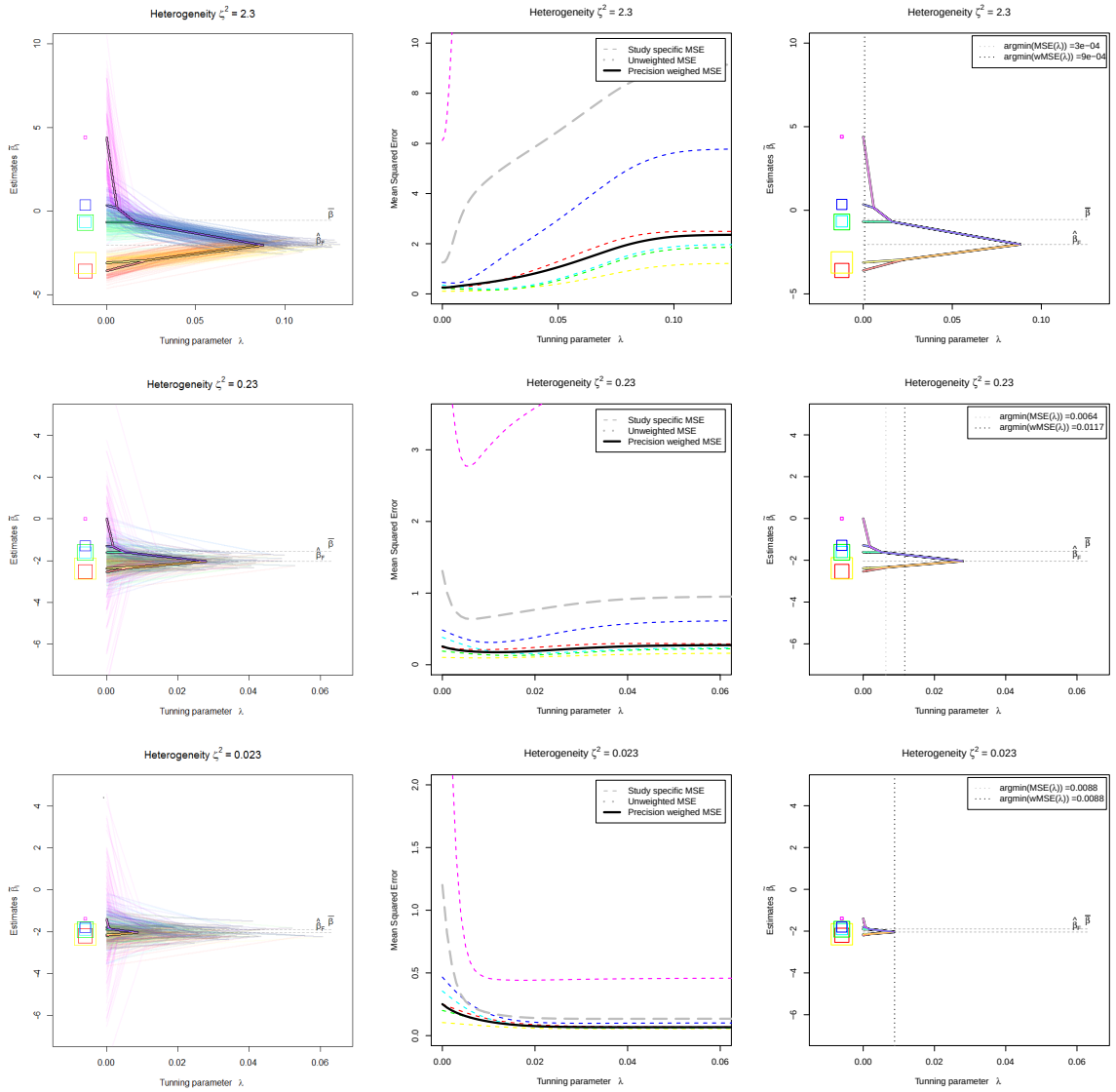

Figure 4.3: Estimation of tuning parameter via bootstrap, on our example data set (first row) and two artificially shrunk versions (second and third rows, note different scales in the axes). First column: solution paths from 100 bootstrap samples of the effect size estimates and their variances. Second column: mean squared error of the individual effects and their un-weighed and weighted average, MSE and $uMSE$. Third column: values of the tuning parameter that minimize $MSE(\lambda)$ and $uMSE(\lambda)$.

$MSE(\lambda)$ and $wMSE(\lambda)$. We can see that for these simple simulation settings, the method performs very similarly regardless of the loss function or the optimization criteria, with slightly worse performance of $L_{\boldsymbol{v}}(\lambda, \boldsymbol{\beta})$.

In Figure 4.4 we compare the results of our method, using the weighted loss function $L_{\boldsymbol{v},\boldsymbol{W}}(\lambda, \boldsymbol{\beta})$, to other possible summaries from the meta-analysis. For the small sample size setting, we see that our method performs as well as the single summary $\hat{\beta}_F$ for low values of ζ^2, i.e. when the true effects are completely homogeneous ($\zeta^2 = 0$) or very close to each other. But while the MSE of $\hat{\beta}_F$ increases linearly with ζ^2, the MSE of $\tilde{\boldsymbol{\beta}}$ plateaus at lager values of ζ^2, performing slightly worse than the MSE of the simple un-shrunk vector of estimates $\hat{\boldsymbol{\beta}}$. A similar behavior is observed for the empirical Bayes estimators, although our method performs slightly better. Note that other single summaries like $\hat{\mu}$ and $\bar{\beta}$ have smaller MSE than $\hat{\beta}_F$, which is expected in this simple setting, where the true precision weights are exactly $1/k$. For the large sample settings we observe similar results, with the MSE low when the effects are homogeneous and increasing to be similar to the MSE of $\hat{\boldsymbol{\beta}}$ when the effects are more heterogeneous. But in addition, we also observe that for settings with the true effects clustered in two or three homogeneous groups ($c = 2$, $c = 3$) our method can perform slightly better or as well as the naïve approach of reporting $\hat{\boldsymbol{\beta}}$; this is not observed when the true effects are not clustered, but evenly spaced ($c = 6$), which suggests that the improvement in accuracy is due to gains in precision from fusing/clustering some of the estimates.

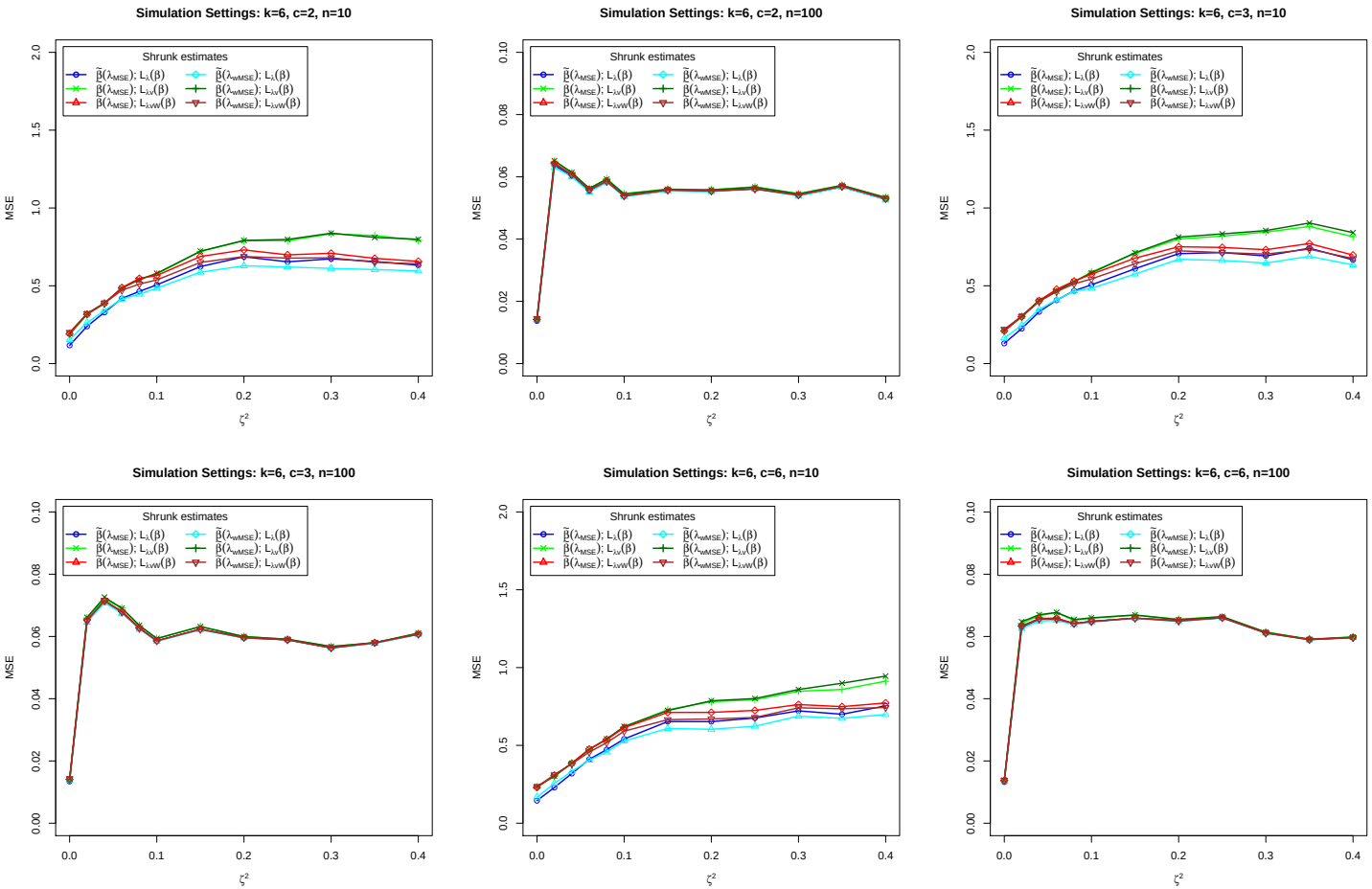

Figure 4.4: Results from simulation study (500 simulations): MSE of shrunken estimates, using three different loss functions ($L_g(\lambda, \boldsymbol{\beta})$, $L_{\boldsymbol{v}}(\lambda, \boldsymbol{\beta})$, $L_{\boldsymbol{v},\boldsymbol{W}}(\lambda, \boldsymbol{\beta})$) and two optimization criteria for estimating λ ($MSE(\lambda)$ and $wMSE(\lambda)$) for six different simulation settings varying the number of clusters (c=2,3,6) and sample size (n=10,100).

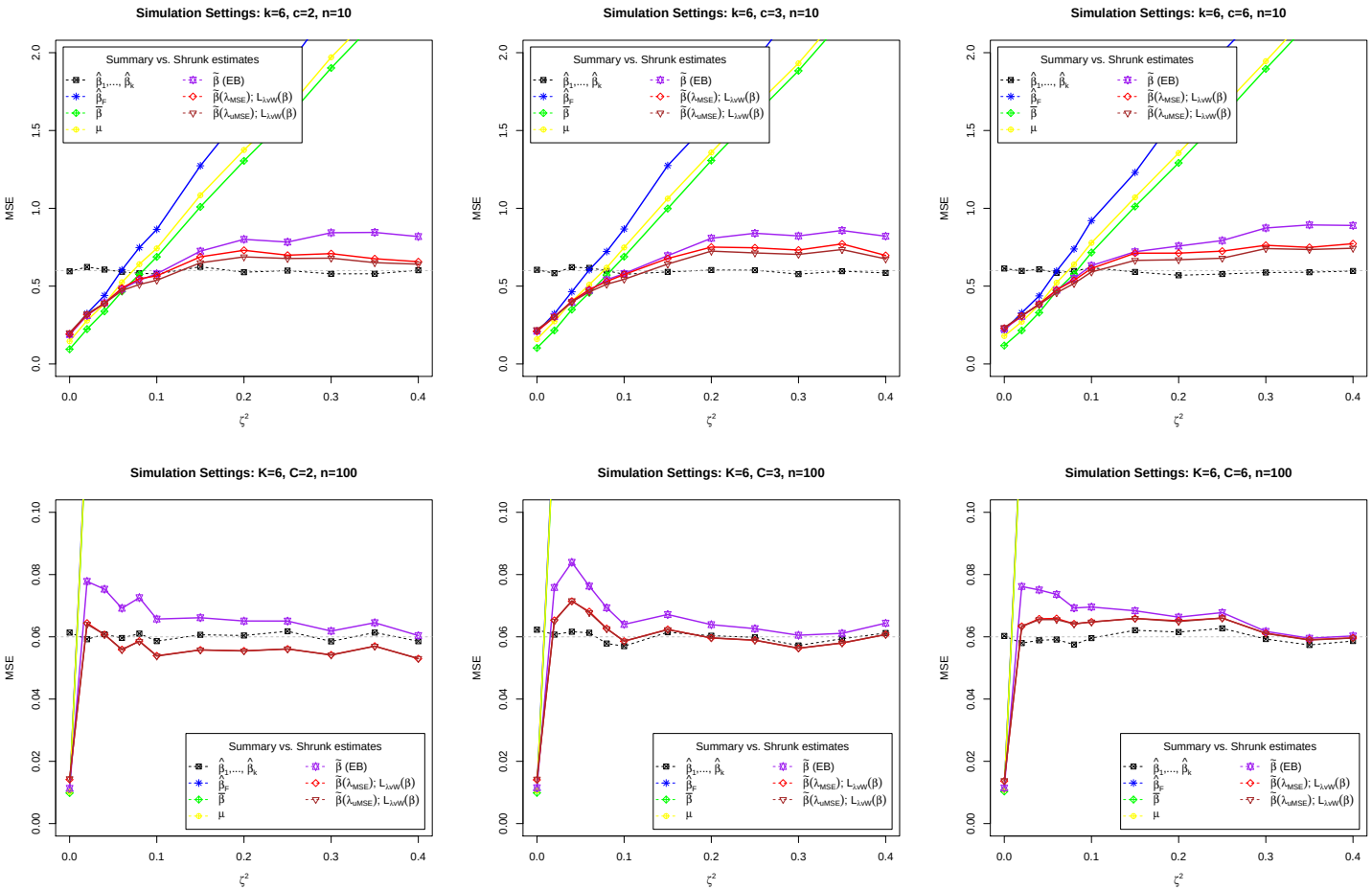

Figure 4.5: Results from simulation study (500 simulations). MSE of shrunken estimates using loss functions $L_{\boldsymbol{v},\boldsymbol{W}}(\lambda, \boldsymbol{\beta})$ compared to other approaches: (1) un-shrunk estimates $\hat{\boldsymbol{\beta}}$, (2) single summaries $\bar{\beta}, \hat{\beta}_F, \hat{\mu}$ and (3) Empirical Bayes shrunken estimates $\tilde{\boldsymbol{\beta}}$ (EB).

To get an insight into how our method clusters (fuses) estimates together, we identify the number of resulting clusters (fused estimates) in each of the shrunk solutions of the 500 simulations. Results are shown in Figure 4.4. In general, the solutions have more clusters than the true underlying setting, except in the cases when the true effects are homogeneous ($\zeta^2 = 0$), for which most of the simulations produced a single cluster. It is possible that for large values of ζ^2, where the two or three clusters of effects are sufficiently far from each other, larger values of λ that would produce further shrinkage and clustering would also produce estimates that are too biased, and thus not minimizing the MSE or wMSE. It must be mentioned here, that although the clustering/grouping of some estimates is a desirable property of our approach, our main goal is to provide improved estimates of the effects, in terms of accuracy. The correct identification of subgroups of homogeneous effects can be seen more as a matter of model selection, and some work specific to the LASSO indicates that certain condition is necessary for consistency of model selection [109].

4.5 Final remarks

In this chapter we proposed a shrinkage estimation method specific for meta-analysis. The method consists on minimizing a weighted version of the fused lasso signal approximator (FLSA) loss function, of which we explored some particular forms. For increasing values of the tuning parameter, the solutions follow trajectories that fuse with each other, while shrinking towards the precision weighted average. We also provided a simple algorithm to calculate the solution to the minimization problem, which takes advantage of the piece-wise linearity of the solution path. Lastly, a parametric bootstrap method, suitable in context of meta-analysis, has been proposed for the estimation of the tuning parameter.

In a simulation study, the shrunken estimates obtained by minimizing our proposed loss function seem to have better accuracy, in terms of MSE, than other alternatives like the empirical Bayes estimator. In settings with important heterogeneity among the true effect size parameters, our method did not perform better than the naïve approach of reporting the un-shrunk vector of estimates, but also it did not perform much worse, even with small sample sizes.

As implemented here, our method for selecting the tuning parameter uses a piece-wise

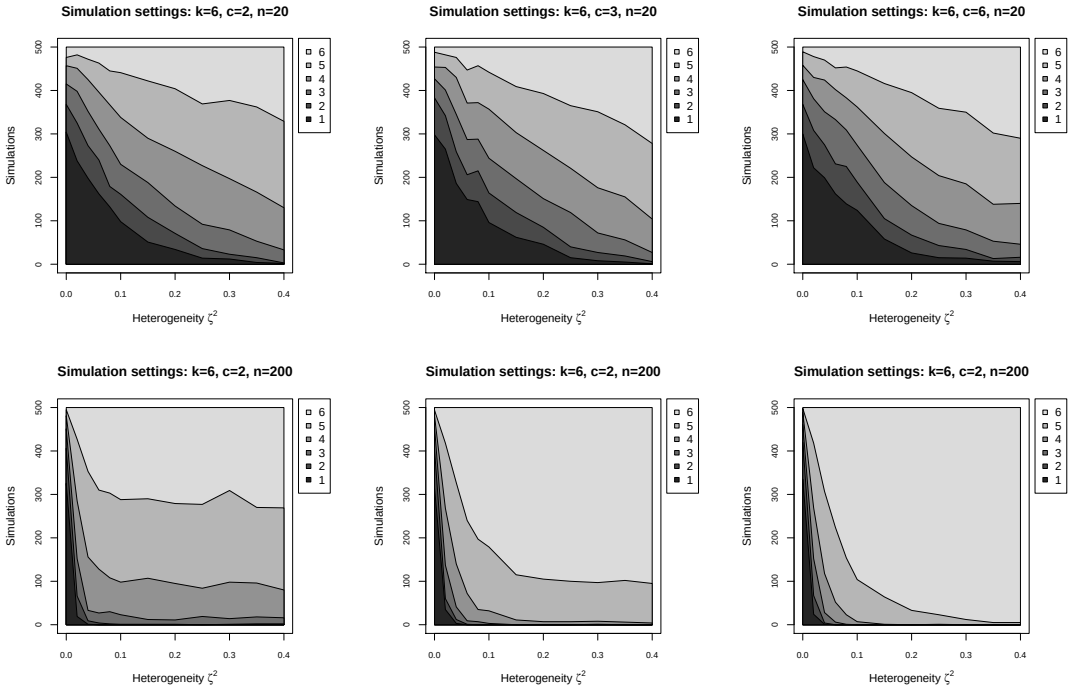

Figure 4.6: Results from simulation study (500 simulations). Number of clusters (fused estimates) identified in the solution vector, for the loss function $L_{\boldsymbol{v},\boldsymbol{W}}(\lambda, \boldsymbol{\beta})$ using the $uMSE$ accuracy criteria to select λ.

linear algorithm instead of standard algorithms for solving the convex optimization problem. This is because, as the method is based on bootstrap sampling, fast computations are needed. However we should note that the algorithm does not account for splits in the solution path, and therefore our method is actually an approximation. We could argue that the approximate solution through the piece-wise algorithm has better properties, and therefore is preferable, even if it would not exactly correspond to the true solution of the convex criterion. On the other hand, it could be argued that splits in the solution path are actually a desirable characteristic, as they might be a consequence of the criterion somehow 'correcting' an erroneous ordering of the estimates (a similar argument is presented in [85]). In general, we believe that a split of two fused estimates is produced by a combination of extreme distances and extreme difference in weights, as it was the case in the example shown here, and therefore we hypothesize that they are not very common. However, further

research to either evaluate this hypothesis, to improve the algorithm or to speed up the required computations would be needed.

Chapter 5

DISCUSSION AND CONCLUSIONS

In the preceding chapters we have proposed a novel approach to meta-analysis which, within a fixed effects framework, allows the estimation of a statistically optimal parameter and provides inference on a mixture of the populations from which the samples in the different studies are drawn. The proposal of a new parameter to quantify heterogeneity completes the approach, by providing a measure of dispersion of the effects that are different, but not assumed to be a random sample from any particular distribution.

Methods of estimation, both from the frequentist and Bayesian perspectives, have been proposed and evaluated. In addition, shrinkage estimation methods, based on penalized estimation, have been explored and adapted to meta-analysis, providing more accurate estimation of all the individual effect-size at the same time. This approach combines the continuous shrinkage of ridge regression with clusterings identification of a small number of mutually-exclusive subsets. Thus, we have provided methods that combine two of the most practically desirable characteristics in multi-dimensional work.

In this concluding discussion, we consider ways in which these methods may be useful, and areas of research left open. Our initial motivation was that, despite the large literature on the classic fixed (i.e. common) effect and random effects approaches, when one tries to reconcile the fields current recommendations, emphasizing the adequacy of the models, with the type of inference they provide, a gap is evident. Specifically, despite the assumption of a single common effect being widely-regarded as unrealistic in many practical situations [8; 94], prominent authors [16; 37] imply that this assumption is necessary for fixed effects approaches to be valid. Furthermore, as soon as sample sizes are sufficient to reject the common effect model (as a straw-man null hypothesis) the alternative approach they offer is to use a random effects analysis, meaning – as seen in Chapter 2 – a change on the location parameter that is the target of inference, and also adding hyper-parameters to the analysis

that were not previously of interest.

This disconnect in the literature – as also discussed in Chapter 2 – is known, albeit perhaps not widely. For example, Hedges 1998 discusses how, without further assumptions, unconditional inference can be obtained using a random effects approach, and inference that is conditional on the studies at hand (and their populations), can be obtained using a fixed effects approach [42]. The difficulties of motivating random effects analysis in the frequentist framework have been acknowledged more recently, [46]. These difficulties do not affect fixed effects analysis, as one only requires that one recognize that the effect estimates come from different studies which, although aiming to estimate the same quantity, have different protocols, are conducted in different ways and sample from different populations. While in each study, it is reasonable to expect variation of the estimate of the effect due to sampling variation within that studys population, and to assume a random distribution to deal with this. But this does not motivate assumptions that the true effects are a sample of independent identically distributed random effects, particularly given that results from the earliest studies are used to plan and more recent studies, also present in the same meta-analysis.

Most of these concerns are not new; the precision weighted average, as a summary of the overall location of non-homogeneous effects, was emphasized by [72; 73] in early development of methods for systematic reviews [108], and clear distinctions between fixed effect (singular) and fixed effects (plural) assumptions have been made by various authors [57; 89]. The validity of the fixed effects approach, even in the presence of heterogeneity, has been satisfactorily discussed in various papers [42; 77; 104]. But despite these solid arguments in favor of the fixed effects model, the approach has not become standard. We suggest that this may be because it lacks a place in a more complete approach, in which users can also address issues of heterogeneity of the effects, or improved estimation of them through Bayesian or shrinkage methods. We here identify some specific ways in which our work has contributed to taking the fixed effects towards a complete, flexible, analytic approach:

(a) By showing the optimality of the precision weighted average estimator, we have provided a stronger justification to using this particular data-adaptive estimator even in the presence of heterogeneity.

(b) By proposing a new parameter to measure heterogeneity, which does not rely on a hierarchical structured model or estimation of hyper-parameters, we have provided alternative ways to describe both the location and dispersion in the observed data, while keeping the inference conditional on the studies at hand.

(c) By proposing and evaluating different methods of estimation, and in particular methods that take into account the uncertainty in variance estimates, we have ensured that the proposed parameters can be estimated, and that the estimators have desirable statistical properties.

(d) By using multivariate exchangeable priors for Bayesian estimation, we have eliminated the need for hierarchical models and the consequent shifting of the inference target to the hyper-parameters of the model. The use of multivariate priors has also provided a different perspective on how prior information can be incorporated.

(e) By adapting penalized estimation to meta-analysis, we have offered an alternative method of shrinkage estimation that provides improved estimation of the individual effects, avoiding again the need of relying on a hierarchical models.

There are, of course, limitations to this approach. The inference is data driven, in that an estimate is provided of the mean effect in a population consisting of a mix of the populations from the different studies, with those proportions being determined by the observed data, i.e. the studies selected for the meta-analysis. In some situations, it may not be of interest to generalize to this particular population. Therefore, the appropriateness of this particular parameter should be considered alongside other scientific goals.

Also, as discussed in [77], it is not possible to make predictions about new or future studies in general when using a fixed-effects approach. In order to be able to predict a new effect β_i, it is necessary to assume some connection between the new β_i and the existing β_i, and standard fixed-effects analyses do not provide this. (In the special case of the constant effect model, in which we also assume that all the effects are exactly equal, in studies included in the meta-analysis and in future studies, the connection is trivial and prediction is possible, but not interesting). But in general in the fixed effects approach "no assumptions or models are available to connect these (studies) to each other or to future

studies" [77].

We also note that using a fixed effects approach does not prevent one from estimating other parameters of interest. In particular, assuming a particular distribution and estimating the mean and variance from a random effects model, can and should be done when predictive inference is desired. And in general, we are not advocating exclusive use of a fixed effects approaches to meta-analysis, but for consideration of it as at least a default. We do not propose that it replace current methods, but complement them, as it can provides different information that may enhance practical inference.

Further work is still needed. In this work we have focused mainly on continuous outcomes, or more generally, in scenarios where a Normal approximation for the distribution of each effect estimate is reasonable and linear combinations of all the estimates provides a meaningful summary. More work is needed on developing an analog approach for binary or discrete outcomes, including characterizing the population parameter that such approach would estimate, and utilizing any available information on the mean-variance relationship and its impact on the relationship between the effect estimate and its estimated standard error. In terms of estimation, improvements might be achieved for the case of unknown variances, especially in the Bayesian framework, where some prior information on the similarity of the populations variances might be incorporated. And even though Chapter 4 is a useful first effort exploring alternative methods of shrinkage estimation specific for meta-analysis, there is clearly more work that needs to be done in this area, including further exploration of other convex and non-convex penalized functions, as well as different criteria for parameter selection.

Beyond the methodology, further work is also needed to get and to provide better understanding on the applicability, importance and impact of the proposed approach. By applying our methods to different and more varied examples of meta-analyses, we can better characterize and contrast its strengths and weaknesses, the type of inference that it provides and the type of scientific questions that it can help answer. This would also help reassure practitioners of its value when used with traditional approaches.

BIBLIOGRAPHY

[1] Allen, R. C. and J. H. Stone (2001). The Gauss Markov theorem: A pedagogical note. *The American Economist 45*(1), 92–94.

[2] Baujat, B., C. Mahé, J.-P. Pignon, and C. Hill (2002). A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Statistics in medicine 21*(18), 2641–2652.

[3] Berlin, J. A. (1995). Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies. *American journal of epidemiology 142*(4), 383–387.

[4] Berrington de Gonzlez, A. and D. R. Cox (2007, 12). Interpretation of interaction: A review. *Ann. Appl. Stat. 1*(2), 371–385.

[5] Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah (2014, November). Julia: A fresh approach to numerical computing.

[6] Bhattacharjee, S., P. Rajaraman, K. B. Jacobs, W. A. Wheeler, B. S. Melin, P. Hartge, M. Yeager, C. C. Chung, S. J. Chanock, N. Chatterjee, et al. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics 90*(5), 821–835.

[7] Böckenhoff, A. and J. Hartung (1998). Some corrections of the significance level in meta-analysis. Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund 1998, 23.

[8] Borenstein, M., L. V. Hedges, J. P. Higgins, and H. R. Rothstein (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods 1*(2), 97–111.

[9] Borensteing, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, and J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. New York: Russel Sage Foundation.

[10] Bowden, J., S. Seaman, X. Huang, and I. White (2014). Weighted log-rank tests for the analysis of late stage cancer trials with substantial treatment switching. In *WNAR/IMS Conference*.

[11] Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies. *The Annals of Applied Statistics 2*(1), pp. 113–152.

[12] Canu, S. and Y. Grandvalet (1999). Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. *Advances in neural information processing systems*, 445.

[13] Casella, G. and R. L. Berger (1990). *Statistical inference*, Volume 70. Duxbury Press Belmont, CA.

[14] Chi, E. C. and K. Lange (2013). Splitting methods for convex clustering. *arXiv preprint arXiv:1304.0499*.

[15] Cochrane Collaboration (1996). *The Cochrane Library*. Update Software Ltd.

[16] Cooper, H., L. Hedges, and J. Valentine (2009). *The handbook of research synthesis and meta-analysis*. New York: Russel Sage Foundation.

[17] De Finetti, B. (1931). Funzione Caratteristica Di un Fenomeno Aleatorio. 6. Memorie, pp. 251–299. Academia Nazionale del Linceo.

[18] DeGroot, M. H. (2005). *Optimal statistical decisions*, Volume 82. John Wiley & Sons.

[19] DerSimonian, R. and N. Laird (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials 7*(3), 177 – 188.

[20] Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, 23–27.

[21] Duchesne, P. and P. L. de Micheaux (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics and Data Analysis 54*, 858–862.

[22] DuMouchel, W. and S.-L. Normand (2000). Computer-modeling and graphical strategies for meta-analysis. In D. Stangl and B. D.A. (Eds.), *Meta-analysis in medicine and health policy*, pp. 127–178. Marcel Dekker: New York.

[23] Efron, B. and C. Morris (1973). Stein's estimation rule and its competitorsan empirical bayes approach. *Journal of the American Statistical Association 68*(341), 117–130.

[24] Efron, B. and C. Morris (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association 70*(350), 311–319.

[25] Fisher, R. A. et al. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron 1*, 3–32.

[26] Friedman, J., T. Hastie, H. Hfling, and R. Tibshirani (2007, 12). Pathwise coordinate optimization. *Ann. Appl. Stat. 1*(2), 302–332.

[27] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis 1*(3), 515–534.

[28] Gelman, A. and K. O'Rourke (2014). Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics 15*(1), 18–23.

[29] Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher 5*(10), 3–8.

[30] Goel, P. (1983). Information measures and bayesian hierarchical models. *Journal of the American Statistical Association*, 408–410.

[31] Goodman, S. N. (1989). Meta-analysis and evidence. *Controlled Clinical Trials 10*(2), 188–204.

[32] Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. In *ICANN 98*, pp. 201–206. Springer.

[33] Graybill, F. A. (1976). *Theory and applications of the linear model.* Duxbury.

[34] Hardy, R. J. and S. G. Thompson (1996). A likelihood approach to meta-analysis with random effects. *Statistics in medicine 15*(6), 619–629.

[35] Hardy, R. J. and S. G. Thompson (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine 17*(8), 841–856.

[36] Hartung, J. and G. Knapp (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in medicine 20*(12), 1771–1782.

[37] Hartung, J., G. Knapp, and B. K. Sinha (2008). *Statistical meta-analysis with applications.* John Wiley & Sons.

[38] Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*(358), 320–338.

[39] Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin 93*(2), 388.

[40] Hedges, L. V. and I. Olkin (1985). *Statistical methods for meta-analysis.* Academic Press New York.

[41] Hedges, L. V. and T. D. Pigott (2001). The power of statistical tests in meta-analysis. *Psychological methods 6*(3), 203.

[42] Hedges, L. V. and J. L. Vevea (1998). Fixed-and random-effects models in meta-analysis. *Psychological methods 3*(4), 486.

[43] Higgins, J. and S. G. Thompson (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine 21*(11), 1539–1558.

[44] Higgins, J. P. and S. Green (2011). *Cochrane handbook for systematic reviews of interventions Version 5.1.0 [updated March 2011]*, Volume 5. The Cochrane Collaboration. Available from http://www.cochrane-handbook.org.

[45] Higgins, J. P., S. G. Thompson, J. J. Deeks, and D. G. Altman (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal 327*(7414), 557.

[46] Higgins, J. P. T., S. G. Thompson, and D. J. Spiegelhalter (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(1), 137–159.

[47] Hocking, T. D., A. Joulin, F. Bach, J.-P. Vert, et al. (2011). Clusterpath an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*.

[48] Hoefling, H. (2010). A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics 19*(4), 984–1006.

[49] Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman and E. Leamer (Eds.), *Handbook of econometrics*, Volume 5. Elsevier.

[50] James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 361–379.

[51] Kaufman, L. and P. J. Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis*, Volume 344. John Wiley & Sons.

[52] Ke, T., J. Fan, and Y. Wu (2013). Homogeneity pursuit. Technical report, University of Chicago.

[53] Kent, J. T. and T. J. Hainsworth (1995). Confidence intervals for the noncentral chi-squared distribution. *Journal of Statistical Planning and Inference 46*(2), 147 – 159.

[54] Kent, J. T. and T. J. Hainsworth (2005). Calculations for the noncentral chi distribution. Technical report. Availble at http://www1.maths.leeds.ac.uk/ john/software/.

[55] Kulinskaya, E., S. Morgenthaler, and R. G. Staudte (2008). *Meta analysis: a guide to calibrating and combining statistical evidence*, Volume 756. Wiley. com.

[56] L'abbé, K. A., A. S. Detsky, and K. O'Rourke (1987). Meta-analysis in clinical research. *Annals of internal medicine 107*(2), 224–233.

[57] Laird, N. M. and F. Mosteller (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care 6*(1), 5–30.

[58] Lambert, P., A. Sutton, P. Burton, K. Abrams, and D. Jones (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in medicine 24*(15), 2401–2428.

[59] Lee, S., M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. C. Christiani, M. M. Wurfel, and X. Lin (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics 91*(2), 224–237.

[60] Lee, S., M. C. Wu, and X. Lin (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics 13*(4), 762–775.

[61] Li, S., B. Mukherjee, J. M. G. Taylor, K. M. Rice, X. Wen, J. D. Rice, H. M. Stringham, and M. Boehnke (2014). The role of environmental heterogeneity in meta-analysis of gene-environment interactions with quantitative traits. *Genetic Epidemiology 38*(5), 416–429.

[62] Li, Y., L. Shi, and H. Daniel Roth (1994). The bias of the commonly-used estimate of variance in meta-analysis. *Communications in Statistics - Theory and Methods 23*(4), 1063–1085.

[63] Lin, D. and D. Zeng (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika 97*(2), 321–332.

[64] Lindley, D. (1962). Discussion of the paper by stein. *Journal of the Royal Statistical Society. Series B (Methodological) 24*, 265–296.

[65] Lindsten, F., H. Ohlsson, and L. Ljung (2011). Just relax and come clustering!: A convexification of k-means clustering. Technical Report. Available at: http://www.diva-portal.org/smash/record.jsf?pid=diva2:650707.

[66] Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing 10*(4), 325–337.

[67] Mammen, E. (1992). *When does bootstrap work?* Springer.

[68] Martin, A. D., K. M. Quinn, and J. H. Park (2011). Mcmcpack: Markov chain monte carlo in r. *Journal of Statistical Software 42*(9), 1–21.

[69] Morris, A. P. (2011). Transethnic meta-analysis of genomewide association studies. *Genetic epidemiology 35*(8), 809–822.

[70] Patnaik, P. B. (1949). The non-central chi^2 - and f-distributions and their applications. *Biometrika 36*(1-2), 202–232.

[71] Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*(3), 545–554.

[72] Peto, R. (1987). Why do we need systematic overviews of randomized trials? (transcript of an oral presentation, modified by the editors). *Statistics in Medicine 6*(3), 233–240.

[73] Peto, R. (1987). Discussion. *Statistics in Medicine 6*(3), 241–244.

[74] R Core Team (2013). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

[75] Rao, P. S., J. Kaplan, and W. G. Cochran (1981). Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association 76*(373), 89–97.

[76] Raudenbush, S. W. (2009). Analyzing effect sizes: random-effects model. In H. Cooper, L. V. Hedges, and J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis.* New York: Russel Sage Foundation.

[77] Rice, K., J. Higgins, and T. Lumley. A re-evaluation of fixed-effect(s) meta-analysis. *Submitted*.

[78] Roberts, M., K. Tchanturia, D. Stahl, L. Southgate, J. Treasure, et al. (2007). A systematic review and meta-analysis of set-shifting ability in eating disorders. *Psychological Medicine-London- 37*(8), 1075.

[79] Rothman, K. J., S. Greenland, and T. L. Lash (2008). *Modern epidemiology*. Lippincott Williams & Wilkins.

[80] Rücker, G., G. Schwarzer, J. Carpenter, and I. Olkin (2009). Why add anything to nothing? the arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in medicine 28*(5), 721–738.

[81] Schmid, C. H., P. C. Stark, J. A. Berlin, P. Landais, and J. Lau (2004). Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *Journal of clinical epidemiology 57*(7), 683–697.

[82] Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika 68*(1), 316–319.

[83] Searle, S. R. (1971). *Linear Models*. John Wiley & Sons.

[84] Senn, S. (2007). Trying to be precise about vagueness. *Statistics in medicine 26*(7), 1417–1430.

[85] Shen, X. and H.-C. Huang (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association 105*(490).

[86] Sidik, K. and J. N. Jonkman (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(2), 367–384.

[87] Singh, M. and R. R. Das (2013). Zinc for the common cold. *Cochrane Database of Systematic Reviews 6*.

[88] Song, F. (1999). Exploring heterogeneity in meta-analysis: is the l'abbe plot useful? *Journal of clinical epidemiology 52*(8), 725–730.

[89] Spiegelhalter, D. J., K. R. Abrams, and J. P. Myles (2004). *Bayesian approaches to clinical trials and health-care evaluation.* John Wiley & Sons.

[90] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 197–206.

[91] Stein, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological) 24*(2), pp. 265–296.

[92] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*(6), pp. 1135–1151.

[93] Tak, L. M., A. Meijer, A. Manoharan, P. de Jonge, and J. G. Rosmalen (2010). More than the sum of its parts: meta-analysis and its potential to discover sources of heterogeneity in psychosomatic medicine. *Psychosomatic medicine 72*(3), 253–265.

[94] Thompson, S. and S. Pocock (1991). Can meta-analyses be trusted? *The Lancet 338*(8775), 1127 – 1130.

[95] Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *Britsh Medical Journal 309*(6965), 1351–1355.

[96] Thompson, S. G. and J. Higgins (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in medicine 21*(11), 1559–1573.

[97] Thompson, S. G. and S. J. Sharp (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in medicine 18*(20), 2693–2708.

[98] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

[99] Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

[100] Udell, M., K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd (2014). Convex optimization in julia. In *HPTCDL'14 Proceedings of the 1st Workshop on High Performance Technical Computing in Dynamic Languages*, New York, pp. 18–28. ACM.

[101] Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

[102] Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics 30*(3), 261–293.

[103] Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in medicine 26*(1), 37–52.

[104] Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software 36*(3), 1–48.

[105] Wasserman, L. (2006). *All of nonparametric statistics*. Springer.

[106] Xie, X., S. C. Kou, and L. D. Brown (2012). Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association 107*(500), 1465–1479.

[107] Xu, R. and I. Wunsch, D. (2005, May). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on 16*(3), 645–678.

[108] Yusuf, S., R. Peto, J. Lewis, R. Collins, and P. Sleight (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in cardiovascular diseases 27*(5), 335–371.

[109] Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Appendix A

SUPPLEMENTAL MATERIAL FOR CHAPTER 2

A.1 Proof of Lemma 2.1.1

Here we provide the proof for Lemma 2.1.1, which sets the basis for the estimation of the optimal weighted average proposed in the paper.

Proof. . Let $\boldsymbol{v}^T = (v_1, v_2, ..., v_k)$ be a vector of arbitrary weights with $\sum_{i=1}^{k} v_i = 1$, and let $\boldsymbol{v}^T\hat{\boldsymbol{\beta}} = \sum v_i\hat{\beta}_i$ be the estimator of $\boldsymbol{v}^T\boldsymbol{\beta} = \sum v_i\beta_i$, with $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}$, then

$$\mathrm{Var}(\boldsymbol{v}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}$$

To minimize this expression, we use Lagrange multipliers:

$$\begin{aligned}
&\frac{d}{d\boldsymbol{v}}\left[\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v} + \lambda(1 - \boldsymbol{v}^T\mathbf{1})\right] = 2\boldsymbol{\Sigma}\boldsymbol{v} - \lambda\mathbf{1} \\
&\frac{d}{d\lambda}\left[\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v} + \lambda(1 - \boldsymbol{v}^T\mathbf{1})\right] = 1 - \boldsymbol{v}^T\mathbf{1} \\
&\Rightarrow \boldsymbol{v} = \frac{\lambda}{2}\boldsymbol{\Sigma}^{-1}\mathbf{1}; \quad \lambda = \frac{2}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}
\end{aligned}$$

so then

$$\boldsymbol{w} = \underset{\boldsymbol{v}:\boldsymbol{v}^T\mathbf{1}=1}{\mathrm{argmin}}\left[\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}\right] = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_k}{\mathbf{1}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_k}$$

with

$$\mathrm{Var}(\boldsymbol{w}^T\hat{\boldsymbol{\beta}}) = \frac{\mathbf{1}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\mathbf{1}_k}{(\mathbf{1}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_k)^2} = \frac{1}{\mathbf{1}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_k}.$$

□

A.2 Moment based estimator of ζ^2

Assuming known variances $\sigma_1^2, ..., \sigma_k^2$, with $\sigma_i^2 = (n_i\phi_i)^{-1}$ for $i = 1, ..., k$, we start out calculating the expected value of a plug-in estimate of ζ^2:

$$\begin{aligned}
\mathrm{E}\left[\frac{\sum_{i=1}^k n_i\phi_i(\hat{\beta}_i - \hat{\beta}_F)^2}{\sum_{i=1}^k n_i\phi_i}\right] &= \mathrm{E}\left[\sum_{i=1}^k \frac{n_i\phi_i}{\Phi}\hat{\beta}_i^2 - \hat{\beta}_F^2\right] \\
&= \sum_{i=1}^k \frac{n_i\phi_i}{\Phi}\left(\mathrm{Var}[\hat{\beta}_i] + (\mathrm{E}[\hat{\beta}_i])^2\right) - \left(\mathrm{Var}[\hat{\beta}_F] + (\mathrm{E}[\hat{\beta}_F])^2\right) \\
&= \left(\sum_{i=1}^k \frac{n_i\phi_i}{\Phi}\beta_i^2 - \beta_F^2\right) + \left(\frac{1}{\Phi}\sum_{i=1}^k n_i\phi_i\sigma_i^2 - \frac{1}{\Phi}\right) \\
&= \sum_{i=1}^k \frac{n_i\phi_i}{\Phi}(\beta_i - \beta_F)^2 + \frac{k-1}{\Phi} \\
&= \zeta^2 + \frac{k-1}{\Phi}.
\end{aligned}$$

So an unbiased estimator of ζ^2 is then given by

$$\hat{\zeta}^2 = \frac{\sum_{i=1}^k \sigma_i^{-2}(\hat{\beta}_i - \hat{\beta}_F)^2 - (k-1)}{\sum_{i=1}^k \sigma_i^{-2}} = \frac{Q - (k-1)}{\Phi}$$

A.3 Derivation of the Large Sample Size Approximation (LSSA) for the variance of $\hat{\hat{\beta}}_F$

In this appendix we provide more details on the derivation of the general form of our LLSA estimator of the variance of $\hat{\hat{\beta}}_F$ as given in equation (2.17) of our paper. We also derive the specific form of this variance estimator for some common effects estimates of continuous outcomes in subsequent sub-sections.

A.3.1 General form of the LSSA

We recall that the goal is to find a large sample size approximation for the variance of $\hat{\hat{\beta}}_F$ (19), which has been expressed as:

$$\mathrm{Var}[\hat{\hat{\beta}}_F] = \mathrm{E}\left[\frac{\sum_i^k(\eta_i\hat{\phi}_i)^2(n_i\phi_i)^{-1}}{(\sum_i^k \eta_i\hat{\phi}_i)^2}\right] + \mathrm{Var}\left[\frac{\sum_{i=1}^k \eta_i\hat{\phi}_i\beta_i}{\sum_{i=1}^k \eta_i\hat{\phi}_i}\right]. \tag{A.1}$$

We also write $s_i^2 = (n_i\phi_i)^{-1}$ as the estimator of $\sigma_i^2 = (n_i\phi_i)^{-1}$, the variance $\hat{\beta}_i$ in the ith study. We start with the assumption of a Normal asymptotic distribution for $\hat{\phi}_i$:

$$\sqrt{n_i}(\hat{\phi}_i - \phi_i) \underset{d}{\rightarrow} N(0, f_i(\boldsymbol{\theta}_i)), \tag{A.2}$$

where $f_i(\boldsymbol{\theta}_i)$ is some function of the distributional moments of the population(s) in study i. For a meta-analysis of studies with different sample sizes, (A.2) can be written as:

$$\sqrt{N}(\hat{\phi}_i - \phi_i) \underset{d}{\rightarrow} N\left(0, \frac{f_i(\boldsymbol{\theta}_i)}{(n_i/N)}\right) \tag{A.3}$$

where $N = \sum_i^k n_i$. Setting $\eta_i = n_i/N$ fixed for $i = 1, ..., k$, we can write:

$$\sqrt{N}\left(\begin{pmatrix} \hat{\phi}_1^2 \\ \vdots \\ \hat{\phi}_k^2 \end{pmatrix} - \begin{pmatrix} \phi_1^2 \\ \vdots \\ \phi_k^2 \end{pmatrix}\right) \underset{d}{\rightarrow} N_k(\mathbf{0}_k, \mathrm{Diag}\{f_i(\boldsymbol{\theta}_i)/\eta_i\}) \tag{A.4}$$

Then, as long as $f_i(\boldsymbol{\theta}_i) < \infty$ for $i = 1, ..., k$, we can apply the Delta method to obtain an approximation for the two terms in (A.1). For large enough N:

$$\begin{aligned} \mathrm{E}\left[\frac{\sum_i^k (\eta_i\hat{\phi}_i)^2 (n_i\phi_i)^{-1}}{(\sum_i^k \eta_i\hat{\phi}_i)^2}\right] &= \frac{1}{N}\mathrm{E}\left[\frac{\sum_i^k (\eta_i\hat{\phi}_i)^2 (\eta_i\phi_i)^{-1}}{(\sum_i^k \eta_i\hat{\phi}_i)^2}\right] \\ &\approx \frac{1}{N(\sum_i^k \eta_i\phi_i)} \end{aligned} \tag{A.5}$$

While for the second term in (A.1), we take derivative

$$\frac{d}{d\phi_i}\left(\frac{\sum_i^k \eta_i\phi_i\beta_i}{\sum_i^k \eta_i\phi_i}\right) = \frac{\left(\sum_{i=1}^k \eta_i\phi_i\right)\eta_i\beta_i - \left(\sum_{i=1}^k \eta_i\phi_i\beta_i\right)\eta_i}{\left(\sum_{i=1}^k \eta_i\phi_i\right)^2} = \frac{\eta_i(\beta_i - \beta_F)}{\sum_{i=1}^k \eta_i\phi_i},$$

and apply the Delta Method

$$\sqrt{N}\left[\left(\frac{\sum_i^k \eta_i\hat{\phi}_i\beta_i}{\sum_i^k \eta_i\hat{\phi}_i}\right) - \left(\frac{\sum_i^k \eta_i\phi_i\beta_i}{\sum_i^k \eta_i\phi_i}\right)\right] \underset{d}{\rightarrow} N\left(0, \frac{\sum_i^k \eta_i\left(\beta_i - \beta_F\right)^2 f_i(\boldsymbol{\theta}_i)}{\left(\sum_i^k \eta_i\phi_i\right)^2}\right).$$

So that, for sufficiently large N:

$$\mathrm{Var}\left[\frac{\sum_{i=1}^k \eta_i\hat{\phi}_i\beta_i}{\sum_{i=1}^k \eta_i\hat{\phi}_i}\right] \approx \frac{\sum_i^k \eta_i\left(\beta_i - \beta_F\right)^2 f_i(\boldsymbol{\theta}_i)}{N\left(\sum_i^k \eta_i\phi_i\right)^2}. \tag{A.6}$$

Finally, combining (A.5) and (A.6) and factorizing, we get our Large-Sample Size Approximation (LSSA) of the variance of $\hat{\hat{\beta}}_F$, expressed as the product of the variance of $\hat{\beta}_F$ multiplied by an inflation factor, which is greater than 1 when the effects are heterogeneous.

$$\mathrm{Var}[\hat{\hat{\beta}}_F] \approx \frac{1}{N\left(\sum_i^k \eta_i \phi_i\right)} \left[1 + \frac{\sum_i^k \eta_i \left(\beta_i - \beta_F\right)^2 f_i(\boldsymbol{\theta}_i)}{\sum_i^k \eta_i \phi_i}\right]. \tag{A.7}$$

Here we notice that for some special cases when $f_i(\boldsymbol{\theta}_i)/\phi_i = c$, a constant, for $i = 1, ..., k$, this expression can be factorized out and the inflation factor can be then expressed as $(1 + c\zeta^2)$, a function of the amount of heterogeneity. The specific form of $f_i(\boldsymbol{\theta}_i)$ for some common effect estimators of continuous outcomes is given in the following sections.

A.3.2 LSSA approximation for the variance of the mean difference of independent groups

Following Borenstein [9], we first look at meta-analyses of studies that compare the means of two independent groups, when an assumption of equal variances is made. Here, the effect size in the ith study, $\beta_i = \Delta_i = \mu_{i,X} - \mu_{i,Y}$, is estimated by $\hat{\beta}_i = \bar{X}_i - \bar{Y}_i$, with $\mathrm{Var}(\hat{\beta}_i) = \sigma_i^2 = (1/n_{i,X} + 1/n_{i,Y})\varsigma_i^2$, where ς_i^2 is the population variance, assumed to be the same for the two groups in study i, and $n_{i,X}$ and $n_{i,Y}$ are the respective sample sizes (with $n_i = n_{i,X} + n_{i,Y}$). Here we can write $\sigma_i^2 = (n_i\phi_i)^{-1}$ and $s_i^2 = (n_i\hat{\phi}_i)^{-1}$, with

$$\phi_i \quad = \quad g(\varsigma_i^2) = \left[\left(\frac{1}{n_{i,X}/n_i} + \frac{1}{n_{i,Y}/n_i}\right)\varsigma_i^2\right]^{-1} \tag{A.8}$$

and $\hat{\phi} = g(\hat{\varsigma}_i^2)$, where $\hat{\varsigma}_i^2$ is the pooled estimator of the variance, given by

$$\hat{\varsigma}_i^2 = \frac{\sum_{j=1}^{n_{i,X}}(X_{ij} - \bar{X}_i)^2 + \sum_{j=1}^{n_{i,Y}}(Y_{ij} - \bar{Y}_i)^2}{n_i - 2} = \frac{(n_{i,X} - 1)\hat{\varsigma}_{i,X}^2 + (n_{i,Y} - 1)\hat{\varsigma}_{i,Y}^2}{n_{i,X} + n_{i,Y} - 2}, \tag{A.9}$$

and $\hat{\varsigma}_{i,X}^2$ and $\hat{\varsigma}_{i,Y}^2$ are the sample variances of the two groups in study i. To obtain an asymptotic distribution for $\hat{\phi}_i$, we start from a standard result for the sample variance:

$$\sqrt{n_i}(\hat{\varsigma}_i^2 - \varsigma_i^2) \underset{d}{\rightarrow} N(0, (\kappa_i - 1)\varsigma_i^4), \tag{A.10}$$

where κ_i denotes to the kurtosis in the population distribution (which we will also assume to be the same in the two groups for now). So then, keeping the sample size proportions ($n_{i,X}/n_i$ and $n_{i,Y}/n_i$) fixed, the derivative of (A.8) is

$$\frac{d}{d\varsigma_i^2} g(\varsigma_i^2) = -\left(\frac{1}{n_{i,X}/n_i} + \frac{1}{n_{i,Y}/n_i}\right)^{-1} (\varsigma_i^2)^{-2}. \tag{A.11}$$

Applying the Delta method on (A.10) we conclude that

$$\sqrt{n_i}(\hat{\phi}_i - \phi_i) \underset{d}{\rightarrow} N\left(0, \frac{\kappa_i - 1}{\left(\frac{1}{n_{i,X}/n_i} + \frac{1}{n_{i,Y}/n_i}\right)^2 \varsigma_i^4}\right) \tag{A.12}$$

We notice that for studies with balanced design ($n_{i,X} = n_{i,Y}$) the information $\phi_i = 1/4\varsigma_i^2$ and the asymptotic variance in the last expression reduces to $(\kappa_i - 1)/4^2\varsigma_i^4$. If all studies in a meta-analysis are balanced and the population variance and kurtosis can be assumed constant across all studies, then the inflation factor in (2.17) is given by $(1 + \frac{\kappa-1}{4\varsigma^2}\zeta^2)$.

Now, when the assumption of equal variances is not made, the variance of $\hat{\beta}_i = \bar{X} - \bar{Y}$ is given by $\text{Var}(\hat{\beta}_i) = \sigma_i^2 = \varsigma_{i,X}^2/n_{iX} + \varsigma_{i,Y}^2/n_{i,Y}$, where $\varsigma_{i,X}^2$ and $\varsigma_{i,Y}^2$ are the population variances of the two groups in the ith study [9]. Similar to the case of equal variances, here we can also write $\sigma_i^2 = (n_i\phi_i)^{-1}$ and $s_i^2 = (n_i\hat{\phi}_i)^{-1}$, now with

$$\phi_i \quad = \quad g(\varsigma_{i,X}^2, \varsigma_{i,X}^2) = \left(\frac{\varsigma_{i,X}^2}{n_{i,X}/n_i} + \frac{\varsigma_{i,Y}^2}{n_{i,Y}/n_i}\right)^{-1} \quad , \tag{A.13}$$

and $\hat{\phi}_i = g(\hat{\varsigma}_{i,X}^2, \hat{\varsigma}_{i,X}^2)$. The asymptotic distribution for $\hat{\varsigma}_{i,X}^2$:

$$\sqrt{n_{i,X}}(\hat{\varsigma}_{i,X}^2 - \varsigma_{i,X}^2) \underset{d}{\rightarrow} N(0, (\kappa_{i,X} - 1)\varsigma_{i,X}^4), \tag{A.14}$$

can be expressed as:

$$\sqrt{n_i}(\hat{\varsigma}_{i,X}^2 - \varsigma_{i,X}^2) \underset{d}{\rightarrow} N\left(0, \frac{(\kappa_{i,X} - 1)\varsigma_{i,X}^4}{n_{i,X}/n_i}\right), \tag{A.15}$$

and similarly for $\sqrt{n_i}(\hat{\varsigma}_{i,Y}^2 - \varsigma_{i,Y}^2)$. Keeping the sample size proportions within each study ($n_{i,X}/n_i$ and $n_{i,Y}/n_i$) fixed, we can write:

$$\sqrt{n_i}\left[\begin{pmatrix} \hat{\varsigma}_{i,X}^2 \\ \hat{\varsigma}_{i,Y}^2 \end{pmatrix} - \begin{pmatrix} \varsigma_{i,X}^2 \\ \varsigma_{i,Y}^2 \end{pmatrix}\right] \underset{d}{\rightarrow} N_2\left[\mathbf{0}_2, \begin{pmatrix} \frac{(\kappa_{i,X}-1)\varsigma_{i,X}^4}{n_{i,X}/n_i} & 0 \\ 0 & \frac{(\kappa_{i,Y}-1)\varsigma_{i,Y}^4}{n_{i,Y}/n_i} \end{pmatrix}\right]. \tag{A.16}$$

Taking derivatives of (A.13) we find that

$$\nabla g(\varsigma_{i,X}^2, \varsigma_{i,X}^2) = -\left(\frac{\varsigma_{i,X}^2}{n_{i,X}/n_i} + \frac{\varsigma_{i,Y}^2}{n_{i,Y}/n_i}\right)^{-2} \left(\frac{1}{n_{i,X}/n_i}, \frac{1}{n_{i,Y}/n_i}\right)^T, \tag{A.17}$$

and applying the Delta method we obtain:

$$\sqrt{n_i}(\hat{\phi}_i - \phi_i) \underset{d}{\rightarrow} N\left[0, \frac{\frac{(\kappa_{i,X}-1)\varsigma_{i,X}^4}{(n_{i,X}/n_i)^3} + \frac{(\kappa_{i,Y}-1)\varsigma_{i,Y}^4}{(n_{i,Y}/n_i)^3}}{\left(\frac{\varsigma_{i,X}^2}{n_{i,X}/n_i} + \frac{\varsigma_{i,Y}^2}{n_{i,Y}/n_i}\right)^4}\right] \tag{A.18}$$

A.3.3 LSSA for the mean difference between two matched samples

When the effect size in each study is the mean difference between two matched samples, we have that $\hat{\beta}_i = \bar{X}_i - \bar{Y}_i$, with $\mathrm{Var}(\hat{\beta}_i) = \sigma_i^2 = (\varsigma_{i,X}^2 + \varsigma_{i,Y}^2 - 2\rho\varsigma_{i,X}\varsigma_{i,Y})/m_i$, where ρ denotes the population correlation between any two matched observations X_{ij} and Y_{ij} in study i and $m_i = n_i/2$ is the number of the paired observations [9]. Assuming a bivariate Normal distribution for the observations (X_{ij}, Y_{ij}), the following asymptotic distribution can be obtained for the sample variances ($\hat{\varsigma}_{i,X}^2$,$\hat{\varsigma}_{i,X}^2$) and sample covariance ($\hat{\varsigma}_{i,XY} = \frac{1}{m_i}\sum_{j=1}^{m_i}(X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_i)$):

$$\sqrt{m_i}\left[\begin{pmatrix} \hat{\varsigma}_{i,X}^2 \\ \hat{\varsigma}_{i,Y}^2 \\ \hat{\varsigma}_{i,XY} \end{pmatrix} - \begin{pmatrix} \varsigma_{i,X}^2 \\ \varsigma_{i,Y}^2 \\ \varsigma_{i,XY} \end{pmatrix}\right] \underset{d}{\rightarrow} N_3\left[\mathbf{0}_3, \begin{pmatrix} 2\varsigma_{i,X}^4 & 2\rho^2\varsigma_{i,X}^2\varsigma_{i,Y}^2 & 2\rho\varsigma_{i,X}^3\varsigma_{i,Y} \\ 2\rho^2\varsigma_{i,X}^2\varsigma_{i,Y}^2 & 2\varsigma_{i,Y}^4 & 2\rho\varsigma_{i,X}\varsigma_{i,Y}^3 \\ 2\rho\varsigma_{i,X}^3\varsigma_{i,Y} & 2\rho\varsigma_{i,X}\varsigma_{i,Y}^3 & (1+\rho^2)\varsigma_{i,X}^2\varsigma_{i,Y}^2 \end{pmatrix}\right]. \tag{A.19}$$

Let $\sigma_i^2 = (n_i\phi_i)^{-1} = 2(\varsigma_{i,X}^2 + \varsigma_{i,Y}^2 - 2\varsigma_{i,XY})/n_i$ and

$$\phi_i = g(\varsigma_{i,X}^2, \varsigma_{i,Y}^2, \varsigma_{i,XY}^2) = [2(\varsigma_{i,X}^2 + \varsigma_{i,Y}^2 - 2\varsigma_{i,XY}^2)]^{-1}, \tag{A.20}$$

with $\hat{\phi} = g(\hat{\varsigma}_{i,X}^2, \hat{\varsigma}_{i,Y}^2, \hat{\varsigma}_{i,XY}^2)$. Taking derivatives:

$$\nabla g(\varsigma_{i,X}^2, \varsigma_{i,Y}^2, \varsigma_{i,XY}^2) = -\frac{1}{2}(\varsigma_{i,X}^2 + \varsigma_{i,Y}^2 - 2\varsigma_{i,XY}^2)^{-2}(1, 1, -2)^T \tag{A.21}$$

so we have, by the Delta Method:

$$\sqrt{m_i}(\hat{\phi}_i - \phi) \underset{d}{\rightarrow} N\left(0, \frac{2[\varsigma_{i,X}^4 - 8\rho\varsigma_{i,X}^3\varsigma_{i,Y} + 2(1+2\rho^2)\varsigma_{i,X}^2\varsigma_{i,Y}^2 - 8\rho\varsigma_{i,X}\varsigma_{i,Y}^3 + \varsigma_{i,Y}^4]}{4(\varsigma_{i,X}^2 + \varsigma_{i,Y}^2 - 2\rho\varsigma_{i,X}\varsigma_{i,Y})^4}\right). \tag{A.22}$$

Appendix B

SUPPLEMENTAL MATERIAL FOR CHAPTER 3

B.1 Posterior distribution of ζ^2 when using a conjugate multivariate Normal prior

The following equations have been adapted from Duchesne and de Micheaux [21]. Let the matrix $\boldsymbol{C}$ be the Cholesky decomposition of $\text{Cov}[\beta_F|\hat{\boldsymbol{\beta}}]$, so that $\boldsymbol{C}^T\boldsymbol{C} = \text{Cov}[\beta_F|\hat{\boldsymbol{\beta}}]$, and let $\boldsymbol{P}$, with $\boldsymbol{P}\boldsymbol{P}^T$, such that it diagonalizes the the matrix $\boldsymbol{C}(\boldsymbol{W} - \boldsymbol{W}\boldsymbol{1}_{kk}\boldsymbol{W})\boldsymbol{C}^T$, so that

$$\boldsymbol{P}\boldsymbol{C}(\boldsymbol{W} - \boldsymbol{W}\boldsymbol{1}_{kk}\boldsymbol{W})\boldsymbol{C}^T\boldsymbol{P}^T = \boldsymbol{D} = \text{diag}\{\lambda_1, \lambda_2, ..., \lambda_k\}$$

Given that the rank of $(\boldsymbol{W} - \boldsymbol{W}\boldsymbol{1}_{kk}\boldsymbol{W})$ is $k-1$, we have that $\lambda_1 \geq ... \geq \lambda_{k-1} \geq 0$ and $\lambda_k = 0$. Then, letting $\boldsymbol{Y} = \boldsymbol{P}(\boldsymbol{C}^T)^{-1}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}})$, with $\boldsymbol{Y} \sim N_k(\boldsymbol{P}(\boldsymbol{C}^T)^{-1}\mathbb{E}[\beta_F|\hat{\boldsymbol{\beta}}], I_k)$, we get that

$$\begin{aligned}(\zeta^2|\hat{\boldsymbol{\beta}}) &= (\boldsymbol{\beta}|\hat{\boldsymbol{\beta}})^T\left(\boldsymbol{W} - \boldsymbol{W}1_{kk}\boldsymbol{W}\right)(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{Y}^T\boldsymbol{D}\boldsymbol{Y} = \sum_{i=1}^{k-1}\lambda_i\chi^2(\delta_i)\end{aligned}$$

where δ_i equal to the square of the ith element of $\boldsymbol{P}(\boldsymbol{C}^T)^{-1}\mathbb{E}[\beta_F|\hat{\boldsymbol{\beta}}]$. From this equation, the expected value and variance of $(\zeta^2|\boldsymbol{\beta})$ can be easily evaluated:

$$\begin{aligned}\mathbb{E}[\zeta^2|\hat{\boldsymbol{\beta}}] &= \sum_{i=1}^{k-1}\lambda_i(1+\delta_i) \\ \text{Var}[\zeta^2|\hat{\boldsymbol{\beta}}] &= \sum_{i=1}^{k-1}\lambda_i^2(2+4\delta_i)\end{aligned}$$

Also, Duchesne and de Micheaux [21] have implemented some algorithms to evaluate probabilities of a sum of chi-square random variables into the `R` package `CompQuadForm`, which can be used to obtain selected percentiles of this posterior distribution.

B.2 Posterior distribution of β_F when using a hierarchical prior

The posterior variance and mean of $\beta_F = \mathbf{1}_k^T \boldsymbol{W}\boldsymbol{\beta}$ are given by (3.6) and (3.5). Reparametrizing in terms of τ^2 and ψ^2 we obtain the following identities:

$$\boldsymbol{\Upsilon}^{-1} = \left(\tau^2 \boldsymbol{I}_i + \psi^2 \mathbf{1}_{kk}\right)^{-1} = \tau^{-2}\boldsymbol{I}_k - \left(\frac{\psi^2/\tau^2}{\tau^2 + k\psi^2}\right)\mathbf{1}_{kk}$$

$$(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Upsilon}^{-1})^{-1} = \tau^2 \mathrm{Diag}\left\{\frac{\sigma^2}{\sigma^2+\tau^2}\right\} + \left(\frac{\psi^2}{1+\psi^2\sum_i \frac{1}{\sigma_i^2+\tau^2}}\right)\mathrm{Diag}\left\{\frac{\sigma^2}{\sigma^2+\tau^2}\right\}\mathbf{1}_{kk}\mathrm{Diag}\left\{\frac{\sigma^2}{\sigma^2+\tau^2}\right\}$$

Then, for the posterior variance of β_F we obtain:

$$\begin{aligned}
\mathrm{Var}[\beta_F|\hat{\boldsymbol{\beta}}] =& \frac{1}{\Phi^2}\mathbf{1}_k^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Upsilon}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{1}_k \\
=& \frac{1}{\Phi^2}\mathbf{1}_k^t\left[\tau^2\mathrm{Diag}\{\sigma_i^{-2}\}\mathrm{Diag}\left\{\frac{\sigma^2}{\sigma^2+\tau^2}\right\}\mathrm{Diag}\{\sigma_i^{-2}\}\right]\mathbf{1}_k \\
&+ \frac{1}{\Phi^2}\mathbf{1}_k^t\left[\left(\frac{\psi^2}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\mathrm{Diag}\left\{\frac{1}{\sigma^2+\tau^2}\right\}\mathbf{1}_k\mathbf{1}_k^t\mathrm{Diag}\left\{\frac{1}{\sigma^2+\tau^2}\right\}\right]\mathbf{1}_k \\
=& \frac{1}{\Phi^2}\left[\sum_{i=1}^k\left(\frac{1}{\sigma^2}\right)\left(\frac{\tau^2}{\sigma^2+\tau^2}\right) + \left(\frac{\psi^2}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\left(\sum_{i=1}^k\frac{1}{\sigma^2+\tau^2}\right)^2\right] \\
=& \frac{1}{\Phi^2}\sum_{i=1}^k\frac{1}{\sigma_i^2}\left[\left(1-\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right) + \left(\frac{\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right] \\
=& \frac{1}{\Phi^2}\sum_{i=1}^k\frac{1}{\sigma_i^2}\left[1-\left(\frac{1}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right].
\end{aligned}$$

And for the posterior mean:

$$\mathrm{E}[\beta_F|\hat{\boldsymbol{\beta}}] = \frac{1}{\Phi}\mathbf{1}_k^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Upsilon}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}} + \frac{1}{\Phi}\mathbf{1}_k^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Upsilon}^{-1})^{-1}\boldsymbol{\Upsilon}^{-1}\mathbf{1}_k\nu,$$

where the first term reduces to:

$$
\begin{aligned}
&\frac{1}{\Phi^2}\mathbf{1}_k^t\left[\tau^2\mathrm{Diag}\{\sigma_i^{-2}\}\mathrm{Diag}\left\{\frac{\sigma^2}{\sigma^2+\tau^2}\right\}\mathrm{Diag}\{\sigma_i^{-2}\}\right]\hat{\boldsymbol{\beta}}\\
&+\frac{1}{\Phi^2}\mathbf{1}_k^t\left[\left(\frac{\psi^2}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\mathrm{Diag}\left\{\frac{1}{\sigma^2+\tau^2}\right\}\mathbf{1}_k\mathbf{1}_k^t\mathrm{Diag}\left\{\frac{1}{\sigma^2+\tau^2}\right\}\right]\hat{\boldsymbol{\beta}}\\
=&\frac{1}{\Phi}\sum_{i=1}^{k}\left[\left(\frac{1}{\sigma_i^2}\right)\left(\frac{\tau^2}{\sigma_i^2+\tau^2}\right)+\left(\frac{\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\left(\frac{1}{\sigma_i^2+\tau^2}\right)\right]\hat{\beta}_i\\
=&\frac{1}{\Phi}\sum_{i=1}^{k}\left(\frac{1}{\sigma_i^2}\right)\left[\left(1-\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right)+\left(\frac{\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\left(\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right)\right]\hat{\beta}_i\\
=&\frac{1}{\Phi}\sum_{i=1}^{k}\left(\frac{1}{\sigma_i^2}\right)\left[1-\left(\frac{1}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\left(\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right)\right]\hat{\beta}_i,
\end{aligned}
$$

and within the second term we have that:

$$
\begin{aligned}
&\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Upsilon}^{-1})^{-1}\boldsymbol{\Upsilon}^{-1}\\
=\;&\mathrm{Vec}\left\{\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right\}+\left(\frac{\psi^2\tau^{-2}\sum_i\frac{\sigma_i^2}{\sigma_i^2+\tau^2}}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\mathrm{Vec}\left\{\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right\}-\left(\frac{k\psi^2}{\tau^2+k\psi^2}\right)\mathrm{Vec}\left\{\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right\}\\
&-\left(\frac{k\psi^2}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\left(\frac{\psi^2\tau^{-2}}{\tau^2+k\psi^2}\right)\left(\sum_i\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right)\mathrm{Vec}\left\{\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right\}\\
=\;&\left[1-\frac{k\psi^2}{\tau^2+k\psi^2}+\left(\frac{\psi^2\tau^{-2}}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\left(\sum_i\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right)\left(1-\frac{k\psi^2}{\tau^2+k\psi^2}\right)\right]\mathrm{Vec}\left\{\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right\}\\
=\;&\left[\frac{\tau^2}{\tau^2+k\psi^2}\left(1+\frac{k\psi^2\tau^{-2}-\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\right]\mathrm{Vec}\left\{\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right\}\\
=\;&\left(\frac{1}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\mathrm{Vec}\left\{\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right\}.
\end{aligned}
$$

so that

$$
\begin{aligned}
\mathrm{E}[\beta_F|\hat{\boldsymbol{\beta}}]=&\frac{1}{\Phi}\sum_{i=1}^{k}\left(\frac{1}{\sigma_i^2}\right)\left\{\left[1-\left(\frac{1}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\left(\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right)\right]\hat{\beta}_i\right.\\
&\left.+\left[\left(\frac{1}{1+\psi^2\sum_i\frac{1}{\sigma_i^2+\tau^2}}\right)\left(\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right)\right]\nu\right\}
\end{aligned}
$$

B.3 Sample code for MCMC sampling

The following is a sample code in **R** (as an interface to WinBUGS [66]) used for sampling of the posterior distribution of the parameters of interest, through MCMC methods.

```
library(R2WinBUGS)

## Setting hierarchical model
hmodel <- function()
 {
  for (i in 1:k)
  {
    w[i] <- pow(se[i], -2)
    hatbeta[i] ~ dnorm(beta[i], w[i])
    theta[i] ~ dnorm(mu,prec)
    num[i] <- w[i]*beta[i]
    qua[i] <- w[i]*pow(beta[i],2)
  }
 mu ~ dnorm(0,0.001)
 tau ~ dunif(0,100)
 tau2 <- pow(tau,2)
 prec <- 1/tau2
 betaF <- sum(num[])/sum(w[])
 zeta2 <- sum(qua[])/sum(w[]) - pow(betaF,2)
}

## Saving to file
myfilename <- file.path(mydirectory, "hier_model.bug")
write.model(hmodel, myfilename)

## Initial Values (3 chains)
myinitial   <- list( list(beta=rep(0,k),mu=0, tau=0.1),
list(beta=rep(0,k),mu=0, tau=0.9),
list(beta=rep(0,k),mu=0, tau=0.4) )

## Data
mydata <- list(hatbeta=y,se=yse,k=6,psi2=100)

## Parameters of interest
myparameters <- c("beta","mu","tau2","betaF","zeta2")

## Running chains
output <- bugs(data=mydata, inits=initial, parameters=myparameters,
model.file=myfilename, n.chains=3, n.iter=200000,
 n.thin=10, bugs.directory=bugs.dir, debug=TRUE)

## Displaying results
summary(output)
```

B.4 Bayesian estimation for the example meta-analysis

Table B.1: Meta-analyses on the effect of zinc acetate lozenges on the duration of common cold (in days). Bayesian estimation includes a family of conjugate normal priors parametrized as a hierarchical model with fixed value of τ^2 and priors induced by a vague prior on μ and vague priors on τ^2 as used in Lambert et al. [58].

Bayesian Estimation for Hierarchical Model: $\beta_1, ..., \beta_k$ **iid** $N(\mu, \tau^2)$**,** $\mu \sim N(0, \psi^2)$

Fixed value for τ^2 (equivalent to a multivariate conjugate prior)

ψ^2	τ^2	E[$\beta_F \vert \boldsymbol{\beta}$] (95% CI)	E[$\zeta^2 \vert \boldsymbol{\beta}$] (95% CI)	E[$\mu \vert \boldsymbol{\beta}$] (95% CI)
0.1	0.25	-1.67 (-2.03, -1.31)	0.79 (0.34, 1.37)	-0.87 (-1.3, -0.43)
0.1	1	-1.86 (-2.24, -1.48)	1.63 (0.83, 2.62)	-0.42 (-0.94, 0.1)
0.1	25	-2.04 (-2.44, -1.63)	2.4 (1.31, 3.73)	-0.02 (-0.63, 0.6)
0.1	100	-2.04 (-2.45, -1.64)	2.48 (1.36, 3.84)	0.0 (-0.62, 0.61)
1	0.25	-1.98 (-2.37, -1.58)	0.68 (0.27, 1.23)	-1.56 (-2.15, -0.98)
1	1	-1.99 (-2.39, -1.59)	1.53 (0.75, 2.49)	-1.16 (-2.03, -0.29)
1	25	-2.04 (-2.44, -1.63)	2.4 (1.31, 3.72)	-0.13 (-1.9, 1.63)
1	100	-2.04 (-2.45, -1.64)	2.48 (1.36, 3.84)	-0.03 (-1.94, 1.87)
1000	0.25	-2.04 (-2.45, -1.64)	0.66 (0.26, 1.2)	-1.72 (-2.33, -1.1)
1000	1	-2.04 (-2.45, -1.64)	1.49 (0.72, 2.45)	-1.45 (-2.42, -0.48)
1000	25	-2.04 (-2.45, -1.64)	2.39 (1.3, 3.71)	-0.72 (-4.8, 3.36)
1000	100	-2.04 (-2.45, -1.64)	2.48 (1.36, 3.83)	-0.59 (-8.57, 7.39)

Priors for τ^2 from Lambert et al. [58] ($\psi^2 = 1000$)

	E[$\beta_F \vert \boldsymbol{\beta}$] (95% CI)	E[$\zeta^2 \vert \boldsymbol{\beta}$] (95% CI)	E[$\mu \vert \boldsymbol{\beta}$] (95% CI)	τ^2, Median(95% CI)
$\tau^{-2} \sim \Gamma(0.001, 0.001)$	-2.04 (-2.44, -1.63)	2.02 (0.97, 3.33)	-1.16 (-3.18, 1.37)	4.18 (0.9, 29.95)
$\log(\tau^2) \sim$U$(-10, 10)$	-2.05 (-2.45, -1.64)	2.03 (0.98, 3.32)	-1.16 (-3.19, 1.38)	4.2 (0.91, 29.54)
$\tau^{-2} \sim$U$(0.001, 1000)$	-2.04 (-2.45, -1.64)	2.23 (1.14, 3.56)	-0.98 (-4.71, 3.36)	10.84 (1.74, 135.6)
$\tau^{-2} \sim$Par$(1, 0.001)$	-2.04 (-2.45, -1.63)	2.23 (1.14, 3.58)	-0.98 (-4.68, 3.35)	10.7 (1.72, 142.1)
$\tau \sim$U$(0, 100)$	-2.04 (-2.44, -1.64)	2.13 (1.04, 3.46)	-1.07 (-3.69, 2.14)	6.47 (1.22, 54.71)
$\tau \sim N(0, 100)$, $\tau > 0$	-2.04 (-2.45, -1.64)	2.13 (1.06, 3.44)	-1.08 (-3.57, 1.99)	6.18 (1.19, 46.4)

Appendix C

SUPPLEMENTAL MATERIAL FOR CHAPTER 4

C.1 Proof of Lemma 4.2.2

We follow Hoefling [48], and re-write the loss function (4.13) in terms of the fused coefficients:

$$L_{\lambda,\boldsymbol{v},\boldsymbol{W}}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{s=1}^{n_F(\lambda)}\left(\sum_{i\in F_s} v_i(\hat{\beta}_i - \beta_{F_s})^2\right) + \lambda\sum_{t<s}\left[\left(\sum_{i\in F_s}\sum_{j\in F_t} w_{ij}\right)|\beta_{F_s} - \beta_{F_t}|\right]$$

by definition, $\beta_{F_s}(\lambda) \neq \beta_{F_t}(\lambda)$ for $s \neq t$, except for a finite number of λ for which sets are fused or split) and therefore the loss function is differentiable with respect to β_{F_s} at the solution $\beta_{F_s}(\lambda)$. We then have the following:

$$\frac{\partial L_{\lambda,\boldsymbol{v},\boldsymbol{W}}(\boldsymbol{\beta})}{\partial \beta_{F_i} i} = -\sum_{i\in F_s} v_i\hat{\beta}_i + \beta_{F_s}\sum_{i\in F_s} v_i + \lambda\sum_{t\neq s}\left[\left(\sum_{i\in F_s}\sum_{j\in F_t} w_{ij}\right)\text{sign}(\beta_{F_s} - \beta_{F_t})\right] = 0.$$

"By taking the derivative w.r.t. λ and noting that for small changes of λ, the sign of $\beta_{F_s} - \beta_{F_t}$ does not change, it is possible to determine $\partial\beta_{F_s}/\partial\lambda$", as

$$\frac{\partial\beta_{F_s}(\lambda)}{\partial\lambda} = -\frac{\left[\left(\sum_{i\in F_s}\sum_{j\in F_t} w_{ij}\right)\text{sign}(\beta_{F_s} - \beta_{F_t})\right]}{\sum_{i\in F_s} v_i,}$$

which is constant as long as the F_i do not change. Therefore, the solution $\beta_{F_s}(\lambda)$ is a piecewise linear function. At the breakpoints of the solution path, the sets of fused variables change. □

C.2 Proof of Theorem 4.2.3

The following proof has been adapted from [47], where a proof by contradiction is provided for the particular case of $v_i = 1$ and $w_{ij} = 1$ for all i, j. We begin by assuming that $F_s(\lambda^0)$ splits into two sets, $F_{s(1)}(\lambda)$ and $F_{s(2)}(\lambda)$ for some $\lambda > \lambda^0$, with $\beta_{s(1)}(\lambda) < \beta_{s(2)}(\lambda)$. By

Lemma 4.2.2, this should be an optimal solution, with slopes given by:

$$\frac{\partial \beta_{F_{s(1)}}(\lambda)}{\partial \lambda} = \frac{1}{\sum_{i \in F_{s(1)}} v_i} \left[\sum_{t>s} \left(\sum_{i \in F_{s(1)}} \sum_{j \in F_t} w_{ij} \right) + \left(\sum_{i \in F_{s(1)}} \sum_{j \in F_{s(2)}} w_{ij} \right) - \sum_{t<s} \left(\sum_{i \in F_{s(1)}} \sum_{j \in F_t} w_{ij} \right) \right]$$

$$\frac{\partial \beta_{F_{s(2)}}(\lambda)}{\partial \lambda} = \frac{1}{\sum_{i \in F_{s(2)}} v_i} \left[\sum_{t>s} \left(\sum_{i \in F_{s(2)}} \sum_{j \in F_t} w_{ij} \right) - \left(\sum_{i \in F_{s(2)}} \sum_{j \in F_{s(2)}} w_{ij} \right) - \sum_{t<s} \left(\sum_{i \in F_{s(2)}} \sum_{j \in F_t} w_{ij} \right) \right]$$

For penalization weights given by $w_{ij} = v_i v_j$ for all i, j, these reduce to:

$$\begin{aligned} \frac{\partial \beta_{F_{s(1)}}(\lambda)}{\partial \lambda} &= \sum_{t>s} \sum_{j \in F_t} v_j + \sum_{j \in F_{s(2)}} v_j - \sum_{t<s} \sum_{j \in F_t} v_j \\ \frac{\partial \beta_{F_{s(2)}}(\lambda)}{\partial \lambda} &= \sum_{t>s} \sum_{j \in F_t} v_j - \sum_{j \in F_{s(1)}} v_j - \sum_{t<s} \sum_{j \in F_t} v_j, \end{aligned}$$

which means that

$$\frac{\partial \beta_{F_{s(1)}}(\lambda)}{\partial \lambda} > \frac{\partial \beta_{F_{s(2)}}(\lambda)}{\partial \lambda}$$

which contradicts our assumption that $\beta_{s(1)}(\lambda) < \beta_{s(2)}(\lambda)$, thus proving that no split is possible for the particular set of penalization weights given by $w_{ij} = v_i v_j$. $\square$

C.3 Algorithm for the weighted version of fused LASSO signal approximator (FLSA)

Algorithm for the generalized weighted version of the FLSA

initialize

$\lambda^{(0)} = 0$;

$\tilde{\beta}_i^{(0)} = \hat{\beta}_i$ for $i = 1, ..., k$;

$\boldsymbol{\alpha}^{(0)} = (\alpha_1, ..., \alpha_k)$ vector of slopes, as given by (4.14);

end

for j in $0 : k-1$

Calculate the increment in λ to the intersection of all pairs of neighboring estimates, by solving:

$$\tilde{\beta}_i^{(j)} + \Delta\lambda_i \alpha_i^{(j)} = \tilde{\beta}_{i+1}^{(j)} + \Delta\lambda_i \alpha_{i+1}^{(j)}; \text{ for } i = 1, ..., k-1;$$

Identify the increment of lambda to the next fusion:

$$\Delta\lambda^{(j)} = \min\{\Delta\lambda_i\}$$

Update λ and identify the index of the 'left most' estimate to be fused next:

$$\lambda^{(j+1)} = \lambda^{(j)} + \Delta\lambda^{(j)}, \text{ and } f = i | \Delta\lambda_i = \Delta\lambda^{(j)};$$

Update $\tilde{\boldsymbol{\beta}}$:

$$\tilde{\beta}_i^{(j+1)} = \tilde{\beta}_i^{(j)} + \lambda\alpha_i^{(j)}$$

Identify indexes of all estimates that have been fused:

$$\boldsymbol{f} = \{i || \tilde{\beta}_i^{(j+1)} - \tilde{\beta}_f^{(j+1)}| < \epsilon\}$$

Update $\boldsymbol{\alpha}$:

$$\alpha_i^{(j+1)} = (\tilde{\beta}_i - \hat{\beta}_{\boldsymbol{f}})/\lambda^{(j+1)} \text{ for } i \in \boldsymbol{f}, \text{ where } \hat{\beta}_{\boldsymbol{f}} = \textstyle\sum_{i \in \boldsymbol{f}} v_i \hat{\beta}_i$$

$$\alpha_i^{(j+1)} = \alpha_i^{(j)} \text{ for } i \notin \boldsymbol{f}$$

end

output: $\lambda^{(j)}, \tilde{\boldsymbol{\beta}}^{(j)}$ for $j = 0, ..., k-1$.

VITA

Clara Penélope Domínguez Islas was born in México City in 1979 and raised in Colima, México. She received a Bachelor degree in Sciences with specialty in Mathematics from the Universidad de Colima in 2000, and a Masters in Applied Statistics from the Instituto Tecnólogico y de Estudios Superiores de Monterrey in 2004. She worked as a statistician at the Instituto Nacional de Salud Pública de México from 2005 to 2009. She entered the Biostatistics Doctoral program in the University of Washington in September of 2009 and starting in August of 2015, she will be a post-doctoral statistical researcher at the Biostatistics Unit of the Medical Research Council in Cambridge, U.K.