Fall 12-19-2020

Bayesian Semi-supervised Keyphrase Extraction and Jackknife Empirical Likelihood for Assessing Heterogeneity in Meta-analysis

GUANSHEN WANG
Southern Methodist University, guanshenw@smu.edu

BAYESIAN SEMI-SUPERVISED KEYPHRASE EXTRACTION
AND JACKKNIFE EMPIRICAL LIKELIHOOD FOR
ASSESSING HETEROGENEITY IN META-ANALYSIS

Approved by:

Dr. Xinlei (Sherry) Wang
Professor in Department of Statistical Science, SMU

Dr. Yichen Cheng
Assistant Professor of Analytics in Institute for Insight, Robinson College of Business, GSU

Dr. Daniel F. Heitjan
Professor in Department of Statistical Science, SMU & Population & Data Sciences, UTSW

Dr. Chul Moon
Assistant Professor in Department of Statistical Science, SMU

BAYESIAN SEMI-SUPERVISED KEYPHRASE EXTRACTION AND JACKKNIFE EMPIRICAL LIKELIHOOD FOR ASSESSING HETEROGENEITY IN META-ANALYSIS

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Statistical Science

by

Guanshen Wang

B.S., Mathematics, Shanghai University
M.S., Statistics, University of Illinois, Urbana-Champaign

December 19, 2020

ACKNOWLEDGMENTS

I would like to thank my Ph.D. advisor Dr. Xinlei (Sherry) Wang. It has been an honor to be her Ph.D. student. Dr. Wang gave me endless support like nobody else did. Her immerse knowledge and enthusiasm was so contagious for me in the time of my Ph.D. years. In our very first meeting, she taught me that great researchers should be better in every single day, which was one of the most important lesson I have even had. Her advice and help motivated me not only to be a qualified researcher, but to be a great person. I could not ask for a better mentor for my Ph.D. study.

Secondly, I would like to appreciate huge support and help from Dr. Yichen Cheng. She has provided me with great research ideas and insights. It would be much more difficult to finish my Ph.D. study without her guidances and suggestions. I also thanks our department chair Dr. Daniel Heitjan for his excellent teaching during my Ph.D. years and great comments on my dissertation. Also, I am grateful to Dr. Chul Moon for his time and interest to serve on my dissertation committee.

In addition, thank all fellow graduate students, I really enjoy the time that we studied and worked together. Thank all friends that I met in the SMU. It is a long list but each of you made my fantastic four years.

Last but not least, I would like to thank my family for their unconditional love and support. Thank my dad, Rong Wang and my mom, Yibo Shen for their huge efforts on raising and caring me. They sacrificed so much for me and I loved them so much. Thank my lovely wife, Binbin Weng for loving me since high school. She kept pushing and supporting me when times were rough. Thank my big family for their help and suggestion. I could not go this far without them.

Wang, Guanshen B.S., Mathematics, Shanghai University
M.S., Statistics, University of Illinois, Urbana-Champaign

Bayesian Semi-supervised Keyphrase Extraction
and Jackknife Empirical Likelihood for
Assessing Heterogeneity in Meta-analysis

Advisor: Dr. Xinlei (Sherry) Wang
Doctor of Philosophy degree conferred December 19, 2020
Dissertation completed October 23, 2020

This dissertation investigates: (1) A Bayesian Semi-supervised Approach to Keyphrase Extraction with Only Positive and Unlabeled Data, (2) Jackknife Empirical Likelihood Confidence Intervals for Assessing Heterogeneity in Meta-analysis of Rare Binary Events.

In the big data era, people are blessed with a huge amount of information. However, the availability of information may also pose great challenges. One big challenge is how to extract useful yet succinct information in an automated fashion. As one of the first few efforts, keyphrase extraction methods summarize an article by identifying a list of keyphrases. Many existing keyphrase extraction methods focus on the unsupervised setting, with all keyphrases assumed unknown. In reality, a (small) subset of the keyphrases may be available for an article. To utilize such information, we propose a probability model based on a semi-supervised setup. Our method incorporates the graph-based information of an article into a Bayesian framework so that our model facilitates statistical inference, which is often absent in the existing methods. To overcome the difficulty arising from high-dimensional posterior sampling, we develop two Markov chain Monte Carlo algorithms based on Gibbs samplers, and compare their performance using benchmark data. We further propose a false discovery rate (FDR) based approach for selecting the number of keyphrases, while the existing methods use ad-hoc threshold values. Our numerical results show that the proposed method compared favorably with state-of-the-art methods for keyphrase extraction.

In meta-analysis, the extent to which effect sizes vary across component studies is called heterogeneity. Typically, it is reflected by a variance parameter in a widely used random-effects (Re) model. In the literature, methods for constructing confidence intervals (CIs) for the parameter often assume that study-level effect sizes be normally distributed. However, this assumption may be violated in practice, especially in meta-analysis of rare binary events. We propose to use jackknife empirical likelihood (JEL), a nonparametric approach that uses jackknife pseudo-values, to construct CIs for the heterogeneity parameter, which lifts the requirement of normality in the Re model. To compute jackknife pseudo-values, we employ a moment-based estimator and consider two commonly used weighing schemes (i.e., equal and inverse variance weights). We prove that with each scheme, the resulting log empirical likelihood ratio follows a chi-square distribution asymptotically. We further examine the performance of the proposed JEL methods and compare them with existing CIs through simulation studies and data examples that focus on data of rare binary events. Our numerical results suggest that the JEL method with equal weights compares favorably with other alternatives, especially when (observed) effect sizes are non-normal and the number of component studies is large. Thus, it is worth serious consideration in statistical inference.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I dedicate this dissertation to my family.

CHAPTER 1

Bayesian Semi-supervised Learning for Keyphrase Extraction

1.1. Introduction

With the explosion of science and technology innovations, a huge amount of new text information is being generated daily. The extensive knowledge and innovations have benefited people tremendously. However, they can sometimes be overwhelming. Therefore, how to effectively process and make use of the available text information becomes a big challenge. Keyphrases, defined as a set of phrases or words that summarizes an article, provide a promising solution to this question. An effective keyphrase extraction method can compress the original documents into a concise form. For example, researchers can grasp the gist of a paper by just reading its keyphrases; a short list of keyphrases from product reviews on Amazon can help customers to determine if this product is worth buying; readers can learn what happens today quickly by just reading a highlight of the daily news.

Most popular keyphrase extraction methods are unsupervised, since it is usually hard to obtain label information for a large set of training data without extensive background knowledge or tremendous human effort. However, it is possible to easily obtain part of the keyphrases for an article. For instance, research articles usually list a collection of keyphrases provided by the authors; the title of a news article is usually informative. Thus, in this paper, we propose a Bayesian method in the semi-supervised setting which makes use of the partially observed keyphrases. Unlike existing methods, our method is model based, allowing us to gauge the uncertainties about model parameters and the (binary) decision easily. Also, it

does not need to predetermine the number of keyphrases, which is unknown and varies from article to article. Our numerical experiments show that the proposed method compared favorably with existing unsupervised and semi-supervised approaches.

The rest of the chapter is organized as follows. State-of-the-art approaches to keyphrase extraction is reviewed in Section 1.2. Section 1.3 describes our proposed Bayesian model. Section 1.4 presents the Bayesian posterior computation and introduces a way to select the number of keyphrases. Sections 1.5 and 1.6 evaluate the performance of the proposed Bayesian method using benchmark data and further illustrate it using data examples from various fields without known ground truth. Section 1.7 concludes the paper with a brief discussion.

1.2. Review of Related Work

There exist three categories of methods in the literature of keyphrase extraction: supervised, unsupervised, and semi-supervised. Most supervised methods first generate a set of features (e.g., phrase frequencies) from articles and then rely on a large amount of high-quality labeled data to train classification algorithms for identifying keyprhases (e.g., [11, 20, 30, 55, 61]). We focus on a common scenario where such labeled data are not available in this paper and review unsupervised and semi-supervised methods below.

1.2.1. Unsupervised learning: from PageRank to TextRank

Brin and Page [10] developed an algorithm called PageRank, which is the first and best-known algorithm used by Google Search to rank web pages in its search engine results. The PageRank algorithm uses a graph to represent a hyperlinked set of documents from the World Wide Web, with each web page involved being a node in the graph. If page i has a hyperlink to page j, then there is a directed edge from page i to page j. Based on the assumption

that more important web pages are likely to receive more links from other pages, PageRank works by finding the stationary distribution for the following stochastic process. Suppose a surfer is browsing a web page at random. Each time, the surfer can either randomly click on one of the hyperlinks available from the current web page or get bored and start on another random page or stop browsing. The probability that the surfer continues clicking is called the damping factor d. Once the surfer determines to continue, the probability that he clicks on a specific hyperlink would be $1/L$, where L is the total outbound edges that the current page has. PageRank computes an important score for each individual web page, reflecting how likely the random surfer would visit it among all given pages. Then PageRank ranks the web pages by these scores. High-ranked pages can be thought of as important nodes in the graph.

Motivated by the idea of PageRank, Mihalcea and Tarau [43] considered keyphrase extraction as a phrase search process that is similar to the web search process, and developed a graph-based ranking model for identifying keyphrases. In their original paper, each article is represented by a graph $< V, E >$, with V denoting the set of vertices and E denoting the set of edges (either directed or undirected). Each vertex is a candidate phrase selected from the article to best define the task at hand. Two vertices are connected by an edge if they follow certain rules (e.g., phrases v_i and v_j are connected if they appear sufficiently close to each other in an article). Mihalcea and Tarau [43] proposed TextRank (TR) to find the importance scores $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T$ for a list of candidate phrases by solving the following equation:

$$\boldsymbol{\theta} = (1-d)\mathbf{1}_n + d\mathbf{G}^T\boldsymbol{\theta}, \tag{1.1}$$

where d is the damping factor (usually set to be 0.85), $\mathbf{1}_n = (1, 1, \cdots, 1)^T$ is a vector of n 1's, n is the number of candidate phrases, $\mathbf{G} = \mathbf{D}^{-1}\mathbf{A}$ is a $n \times n$ matrix representing the normalized graph of the article, $\mathbf{A}$ is the graph represented by a weight matrix, and $\mathbf{D}$ is a diagonal matrix with the diagonal elements $D_{ii} = \sum_{j=1}^{n} A_{ij}$. Those phrases with

high importance scores will be selected as keyphrases. Note that PageRank and TextRank are highly similar: both obtain importance scores for the units (web pages or phrases) by solving $\boldsymbol{\theta}$ from (1.1). However, in Brin and Page [10], PageRank represents each edge using a binary value (1 if present, 0 if absent) while in Mihalcea and Tarau [43], TextRank assigns non-negative weights A_{ij}'s to edges, in order to measure how "strong" the connections are.

Figure 1.1 shows an example of directed graph, nodes of the graph are a subset of words from the article example in 1.5.2. In this graph, there is an edge from word A to word B if and only if word B appears after word A within a window size of 2 in the article, weighted by the frequency of the appearance. Thus, connections are not necessarily mutual. Example: "loss" has an outbound edge to "accident" but "accident" does not have an outbound edge to "loss" because in the article, "accident" appears after "loss" within a size of 2 twice but "loss" never occur after "accident" within the same size. In addition, even there are mutual connections, they can have different weights. Example: the weight of connection from "lifetime" to "independent" is 2 but the weight from "independent" to "lifetime" is 1. Figure 1.2 shows a undirected graph with the same article example. In this undirected graph, two nodes are connected by the frequency of co-occurrence within a window size of 2. Examples: "age" and "loss" are connected with a weight 12 as they co-occurred within a window size of 2 for 12 times in the article; "age" is not connected to "accident" because they never co-occur in the article. In this paper, we mainly focus on the undirected graph.

1.2.2. Semi-supervised learning

In practice, it is possible to know some keyphrases without reading the entire article. For example, many academic journals require authors to specify up to a maximum number of keyphrases (typically 3-5); a domain expert may instantly identify a few keyphrases by reading an article title only. Motivated by the availability of such (partial) label information, semi-supervised learning may be adopted to help improve the detection of keyphrases.

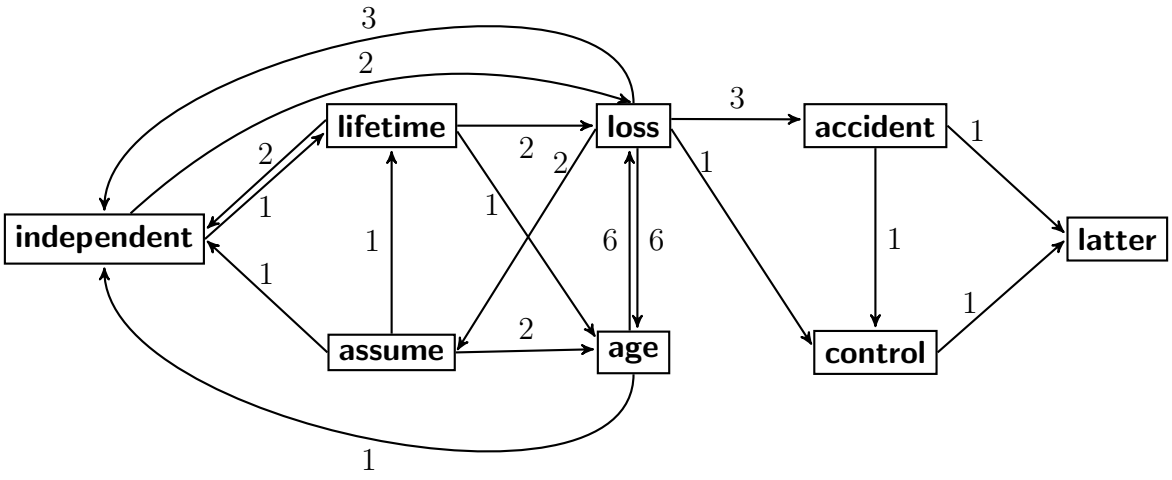


Figure 1.1. A simple example for directed graph. Nodes of the graph are a subset of words from the article example in 1.5.2.

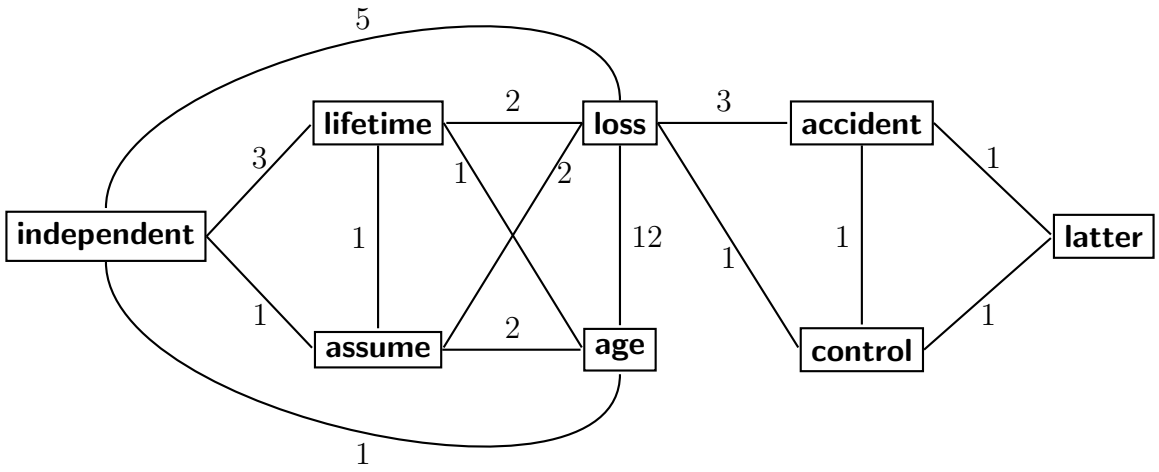


Figure 1.2. A simple example for undirected graph. Nodes of the graph are a subset of words from the article example in 1.5.2.

Existing literature on semi-supervised keyphrase extraction is sparse. Li et al. [38] proposed a semi-supervised method, labeled SS, which aims to preserve the so-called "local consistency" of a graph by solving the following equation with respect to the important scores $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = (1-d)\mathbf{y} + d\mathbf{G}^{*T}\boldsymbol{\theta}, \tag{1.2}$$

where $\mathbf{y}$ is a vector of the observed labels (1 if known to be a keyphrase, 0 otherwise), and $\mathbf{G}^* = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ represents a graph structure that is similar to the $\mathbf{G}$ matrix described in the unsupervised setting. The idea of "local consistency", first proposed by Zhou et al. [64] in a general semi-supervised learning setting, argues that similar nodes should have similar labels. Additionally, Zhou et al. [64] showed that finding the solution to (1.2) is equivalent to minimizing the following cost function:

$$Q(\boldsymbol{\theta}) = \sum_{(i,j)\in E} A_{ij}\left(\frac{1}{\sqrt{D_{ii}}}\theta_i - \frac{1}{\sqrt{D_{jj}}}\theta_j\right)^2 + \mu \parallel \boldsymbol{\theta} - \mathbf{y} \parallel^2,$$

where the regularization parameter $\mu = (1-d)/d$. The first term in the cost function encourages "local consistency" and the second term penalizes the distance between the observed data and the importance scores. Note that $\boldsymbol{\theta}$ from (1.2) is not between 0 and 1 even though it should be close to $\mathbf{y}$. In machine learning, this problem is also referred to as learning using only positive and unlabeled data [17], since all the labels observed are positive.

1.2.3. Constructing a graph from an article

The above algorithms for keyphrase identification are graph based. Typically, such algorithms require a user to first identify a set of candidate units from an article, add them as vertices in a graph, and then define relations that connect such units to draw edges between vertices in the graph. Edges can be directed or undirected, weighted or unweighted.

In the past, various approaches have been proposed to construct a graph from an article. Mihalcea and Tarau [43] recommended to use frequencies of co-occurrence within a given

window size as weights. Wan and Xiao [58] used a larger graph containing words from a group of similar articles rather than those from one single article, in which each edge is weighted using a weighted average of the numbers of co-occurrences among the group of articles. Bougouin et al. [8] used topics instead of phrases as vertices of a graph. Wang et al. [59] used a word embedding technique so that each phrase is represented by a numeric vector and the weight for an edge is constructed based on the co-occurrence relation of the two phrases as well as the distance between the corresponding two numeric vectors. Florescu and Caragea [19] incorporated position information (i.e., phrases that appear earlier in an article are likely to be more important) so that the importance scores are also adjusted to reflect their location in the article. As an alternative to graph-based methods, Liu et al. [40] proposed a cluster-based approach, where phrases are clustered according to their relevance.

1.2.4. Overview

Most of the existing approaches identify keyphrases in an article by first calculating the importance score for each candidate phrase, and then identifying all the phrases with scores greater than a certain threshold as keyphrases. However, such approaches have two limitations. Firstly, even though their rankings reflect the relative importance of candidate phrases, those importance scores cannot be used to gauge the uncertainty about the 0/1 decision (i.e., how likely an individual unit is a keyphrase in the presence of information from the article). Secondly, the threshold value is usually chosen *ad-hoc*. Many existing methods choose a threshold such that 1/3 of the phrases in an article will be selected as keyphrases. Such methods inherently assume that every document has the same proportion of keyphrases. However, in practice, different articles may have different proportions of keyphrases.

We propose a graph-based Bayesian semi-supervised (BSS) learning model, where we incorporate the graph information into a prior distribution and use the observed vector $\mathbf{y}$ to construct the likelihood. Using the BSS model, for each candidate phrase, we can estimate

its probability of being a keyphrase based on the posterior distribution. We further propose a false discover rate (FDR) based criterion to select the number of keyphrases. To evaluate the performance of the proposed method, we apply our method to a popular benchmark data set in Hulth [30]. The results show that overall, our method performs better when compared with baseline approaches in terms of FDR, precision, recall and F-1 measure. To further illustrate potential applications, we apply BSS to a well-known statistical paper and an Amazon review example.

1.3. Model

We describe the proposed BSS learning method for keyphrase extraction, assuming a (small) subset of keyphrases is known. We assign two labels to each candidate phrase i: the observed label y_i and the actual label y_i^*, where y_i indicates whether phrase i is observed to be a keyphrase and y_i^* indicates whether it is indeed a keyphrase (1 for keyphrase, 0 otherwise). Note that only positive labels can be observed. Thus, if a phrase is observed to be a keyphrase (i.e., $y_i = 1$), then it must be a keyphrase (i.e., $y_i^* = 1$). On the other hand, for a phrase with $y_i = 0$, it can be either a keyphrase or a non-keyphrase (i.e., $y_i^* = 0$ or 1). We refer to $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ as the "true labels", which are unobserved to us unless the corresponding observed label is 1.

1.3.1. The likelihood and prior elicitation

Let π_i denote the probability that phrase i is a keyphrase, and α_i denote the probability that a keyphrase is not observed, namely, $\pi_i \equiv Pr(y_i^* = 1)$ and $\alpha_i \equiv Pr(y_i = 0 | y_i^* = 1)$. The conditional probability of y_i given the parameters π_i and α_i can be calculated by integrating out the hidden variable y_i^* in the full likelihood function:

$$
\begin{aligned}
Pr(y_i|\pi_i,\alpha_i) &= \sum_{y_i^*} Pr(y_i, y_i^*|\pi_i,\alpha_i) \\
&= Pr(y_i, y_i^*=1|\pi_i,\alpha_i) + Pr(y_i, y_i^*=0|\pi_i,\alpha_i) \\
&= Pr(y_i|y_i^*=1,\pi_i,\alpha_i)Pr(y_i^*=1|\pi_i,\alpha_i) + Pr(y_i|y_i^*=0,\pi_i,\alpha_i)Pr(y_i^*=0|\pi_i,\alpha_i)
\end{aligned}
$$

Since all the observed positives are true positives, we have $Pr(y_i^*=1|y_i=1)=1$. This also implies that if a phrase is not a keyphrase, it cannot be observed as a keyphrase; that is, $Pr(y_i=0|y_i^*=0)=1$. It follows that

$$
Pr(y_i|\pi_i,\alpha_i) = \begin{cases} (1-\alpha_i)\pi_i & y_i=1 \\ \alpha_i\pi_i + 1 - \pi_i & y_i=0 \end{cases}
$$

For simplicity, we assume α_i's are the same across all phrases: $\alpha_i=\alpha$. Then the likelihood function is

$$
Pr(\mathbf{y}|\boldsymbol{\pi},\alpha) = \prod_{i=1}^{n}[(1-\alpha)\pi_i]^{y_i}[1-\pi_i+\alpha\pi_i]^{1-y_i},
$$

where n is the total number of candidate phrases.

Under a Bayesian framework, the unsupervised importance scores $\boldsymbol{\theta}$ obtained from TextRank, when linked to $\boldsymbol{\pi}$, can be used to form a prior distribution, to incorporate the information from the graph constructed from an article under consideration. For the ith phrase, π_i has a probability scale $(0,1)$ while an importance score θ_i can be any real number. Thus, we use the logit function to link π_i and θ_i: $\theta_i = \text{logit}(\pi_i)$.

Next, we propose a multivariate normal prior on a linear transformation of $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_n)^T$:

$$
\boldsymbol{\theta} - [(1-d)\mathbf{1}_n + d\mathbf{G}^T\boldsymbol{\theta}] \sim N(0,\sigma^2\mathbf{I}),
$$

This prior implies that without the label information, we would like to set $\boldsymbol{\theta}$ to be centered at the solution to (1.1). The prior on $\boldsymbol{\theta}$ can be rewritten as:

$$\pi(\boldsymbol{\theta}) \sim N(\boldsymbol{\theta}_0, \mathbf{B}^{-1}(\mathbf{B}^{-1})^T\sigma^2),$$

where $\boldsymbol{\theta}_0 = \mathbf{B}^{-1}(1-d)\mathbf{1}_n$, and $\mathbf{B} = \mathbf{I} - d\mathbf{G}^T$. One typical choice of the prior distribution for σ^2 is $\pi(\sigma^2) \sim 1/\sigma^2$. However, it is an improper prior, which may lead to an improper posterior. Thus, we use an inverse-gamma distribution $IG(0.001, 0.001)$ instead to approximate $1/\sigma^2$. A uniform prior is used for α: $\pi(\alpha) = 1, \alpha \in (0, 1)$.

1.3.2. The full probability model

With the prior distributions specified in the previous subsection, the full probability model is given by

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}, \alpha, \sigma^2) &= Pr(\mathbf{y}|\boldsymbol{\theta}, \alpha)\pi(\boldsymbol{\theta}|\sigma^2)\pi(\alpha)\pi(\sigma^2) \\ &= \prod_{i=1}^{n} \left[\frac{(1-\alpha)e^{\theta_i}}{1+e^{\theta_i}}\right]^{y_i} \left[1 - \frac{(1-\alpha)e^{\theta_i}}{1+e^{\theta_i}}\right]^{1-y_i} \\ &\cdot \quad N\left(\boldsymbol{\theta}; \boldsymbol{\theta}_0, \mathbf{B}^{-1}(\mathbf{B}^{-1})^T\sigma^2\right) \cdot IG(\sigma^2; 0.001, 0.001). \end{aligned} \tag{1.3}$$

Then the joint posterior distribution is $p(\boldsymbol{\theta}, \alpha, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}, \boldsymbol{\theta}, \alpha, \sigma^2)$. The main parameters of interest are the θ_i's while σ^2 and α are nuisance parameters. In the next section, we will discuss how to sample from the joint posterior distribution and identify keyphrases.

1.4. Posterior Computation and Keyphrase Identification

We propose two MCMC algorithms for posterior sampling. In the first algorithm, we use a Gibbs sampler to iteratively sample σ^2, $\boldsymbol{\theta}$, and α from $p(\boldsymbol{\theta}, \alpha, \sigma^2|\mathbf{y})$, and when sampling $\boldsymbol{\theta}$ within each iteration, we use a Metropolis-Hastings (MH) algorithm. The MH method usually works well when the dimension of $\boldsymbol{\theta}$ is not too high. However, when the dimension of $\boldsymbol{\theta}$ gets higher, this algorithm is likely to get trapped at local maxima. To overcome this issue, we consider a commonly used strategy for high-dimensional MH sampling, which resorts to a tiny scale in the proposal distribution to help the samples gradually escape from the "unhealthy" neighborhood of a local mode over time. The spirit is to accumulate many tiny moves, which are much easier to be accepted than a big move, so as to travel around the entire posterior space over time. In the second algorithm, we first integrate σ^2 out of $p(\boldsymbol{\theta}, \alpha, \sigma^2|\mathbf{y})$, and then use a Gibbs sampler to iteratively sample $\boldsymbol{\theta}$ and α, where a component-wise adaptive MH algorithm is employed to sample $\boldsymbol{\theta}$ at each iteration. The spirit is to move one dimension only each time so that accepting a proposed move is not unlikely. We implement both algorithms and compare their performance for high-dimensional posterior sampling in the context of keyphrase identification.

1.4.1. MH within Gibbs with tiny moves

Using a Gibbs sampler, we iteratively sample from the three posterior conditionals: $p(\sigma^2|\mathbf{y}, \boldsymbol{\theta}, \alpha)$, $p(\alpha|\mathbf{y}, \boldsymbol{\theta}, \sigma^2)$, and $p(\boldsymbol{\theta}|\mathbf{y}, \sigma^2, \alpha)$. It follows from the full probability model (1.3) that the conditional posterior distribution of σ^2 given $\boldsymbol{\theta}$ and α is $IG(\frac{n}{2} + 0.001, \frac{C}{2} + 0.001)$, where $C = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{B}^T \mathbf{B} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. The conditional posterior distribution of $\boldsymbol{\theta}$ given σ^2 and α is

$$p(\boldsymbol{\theta}|\mathbf{y}, \sigma^2, \alpha) \propto \prod_{i=1}^{n} \left[\frac{(1-\alpha)\exp(\theta_i)}{1+\exp(\theta_i)} \right]^{y_i} \left[1 - \frac{(1-\alpha)\exp(\theta_i)}{1+\exp(\theta_i)} \right]^{1-y_i} \exp\left\{ -\frac{C}{2\sigma^2} \right\}.$$

Here, we use an MH algorithm to sample $\boldsymbol{\theta}$, with the proposal distribution set to be $N(\boldsymbol{\theta}^{(t-1)}, \mathbf{B}^{-1}(\mathbf{B}^{-1})^T\sigma^{2(t)} \cdot b/n)$, where $\boldsymbol{\theta}^{(t-1)}$ is the sampled $\boldsymbol{\theta}$ obtained at step $t-1$, $\sigma^{2(t)}$ is the sampled σ^2 at step t, b is a pre-determined constant, and $1/n$ is used to adjust for the dimension of θ in order to make tiny moves. Note that a large n yields a small variance of the proposal distribution so that the proposed move cannot deviate much from the previous draw $\boldsymbol{\theta}^{(t-1)}$. We adjust the value of b to ensure that the acceptance rate of the algorithm is between 0.2 and 0.4, as recommended by Gelman et al. [21, Chap. 11]. Lastly, under the uniform prior, we have $p(\alpha|\mathbf{y}, \boldsymbol{\theta}, \sigma^2) \propto Pr(\mathbf{y}|\boldsymbol{\theta}, \alpha)$. In order to improve the sampling performance and to achieve faster convergence, we follow a procedure similar to empirical Bayes, which sets $\alpha^{(t)}$ to be the value that maximizes $p(\alpha|\mathbf{y}, \boldsymbol{\theta}^{(t)}, \sigma^{2(t)})$, instead of sampling from its conditional posterior.

We refer to the above algorithm as tMH within Gibbs, where tMH stands for MH with tiny moves. The main steps of the algorithm are described in Algorithm 1.

Algorithm 1 tMH within Gibbs

Generate graph, compute graph related terms $\mathbf{B}$, $\boldsymbol{\theta_0}$ and C.
Initialize the starting point of $\boldsymbol{\theta}^{(0)}, \alpha^{(0)}$
for t in $(0 < t \leq T)$: **do**
  Update $\sigma^{2(t)}$ by $IG(\frac{n}{2} + 0.001, \frac{C}{2} + 0.001)$
  Generate $\boldsymbol{\theta}^* \sim N(\boldsymbol{\theta}^{(t-1)}, \mathbf{B}^{-1}(\mathbf{B}^{-1})^T\sigma^{2(t)} \cdot b/n)$.
  Acceptance probability: $A(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = \min(1, \frac{p(\boldsymbol{\theta}^*|\mathbf{y}, \sigma^{2(t)}, \alpha^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y}, \sigma^{2(t)}, \alpha^{(t-1)})})$
  $\mu \sim U(0, 1)$
  if $\mu < A(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$ **then**
    $\boldsymbol{\theta}^{(t)}$=$\boldsymbol{\theta}^*$
  else
    $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$
  end if
  Update $\alpha^{(t)}$ by maximizing $p(\alpha|\boldsymbol{\theta}^{(t)}, \mathbf{y}, \sigma^{2(t)})$ over a grid on the interval (0,1)
end for

1.4.2. Adaptive component-wise MH within Gibbs

Since we adopt a conditional conjugate prior for the nuisance parameter σ^2, we can integrate σ^2 out:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{y},\alpha) \propto & p(\mathbf{y},\boldsymbol{\theta},\alpha) \\
&= \int Pr(\mathbf{y}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta}|\sigma^2)\pi(\alpha)\pi(\sigma^2)d\sigma^2 \\
&\propto \prod_{i=1}^{n}\left[\frac{(1-\alpha)e^{\theta_i}}{1+e^{\theta_i}}\right]^{y_i}\left[1-\frac{(1-\alpha)e^{\theta_i}}{1+e^{\theta_i}}\right]^{1-y_i}\int(\sigma^2)^{-(\frac{n}{2}+1.001)}\exp\left\{-\left(\frac{C+0.002}{2\sigma^2}\right)\right\}d\sigma^2 \\
&\propto \prod_{i=1}^{n}\left[\frac{(1-\alpha)e^{\theta_i}}{1+e^{\theta_i}}\right]^{y_i}\left[1-\frac{(1-\alpha)e^{\theta_i}}{1+e^{\theta_i}}\right]^{1-y_i}\left(\frac{C}{2}+0.001\right)^{-(\frac{n}{2}+0.001)}.
\end{aligned}
$$

Thus, we only need to sample from the posterior conditionals of α and $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{y},\alpha)$ and $p(\alpha|\mathbf{y},\boldsymbol{\theta})$, iteratively. In addition, because α and σ^2 are conditionally independent, $p(\alpha|\mathbf{y},\boldsymbol{\theta})$ remains the same as in Section 1.4.1. Note that for the tMH within Gibbs introduced in Section 1.4.1, we cannot integrate σ^2 out, because it is part of the covariance matrix of the proposal distribution for $\boldsymbol{\theta}$.

To mitigate the inefficiency of high-dimension MH sampling, we consider an adaptive component-wise MH method [23]. That is, instead of sampling $\boldsymbol{\theta}$ as a whole at each iteration, we sample one component of $\boldsymbol{\theta}$ at a time. At iteration t, the proposal for $\theta_i^{(t)}$ is set to an univariate normal distribution $N(\theta_i^{(t-1)}, v_i^{(t)})$, where $v_i^{(t)}$ is an adaptive variance calculated using samples obtained from previous iterations. We refer to this algorithm as acMH within Gibbs, where acMH stands for adaptive component-wise MH. The main steps of this algorithm is described in Algorithm 2.

Algorithm 2 acMH within Gibbs

Generate graph, compute graph related terms $\mathbf{B}$, $\boldsymbol{\theta}_0$ and C.

Initialize the starting point of $\boldsymbol{\theta}^{(0)}, \alpha^{(0)}$

for t in $(0 < t \leq T)$: **do**

for i in $(1 \leq i \leq n)$ **do**

Generate $\theta_i^* \sim N(\theta_i^{(t-1)}, v_i^{(t)})$.

Acceptance probability: $A(\theta_i^*|\theta_i^{(t-1)}) = \min(1, \frac{p(\theta_1^{(t)},...,\theta_{i-1}^{(t)},\theta_i^*,\theta_{i+1}^{(t-1)},...,\theta_n^{(t-1)}|\mathbf{y},\alpha^{(t-1)})}{p(\theta_1^{(t)},...,\theta_{i-1}^{(t)},\theta_i^{(t-1)},\theta_{i+1}^{(t-1)},...,\theta_n^{(t-1)}|\mathbf{y},\alpha^{(t-1)})}$

$\mu \sim U(0,1)$

if $\mu < A(\theta_i^*|\theta_i^{(t-1)})$ **then**

$\theta_i^{(t)}=\theta_i^*$

else

$\theta_i^{(t)} = \theta_i^{(t-1)}$

end if

end for

Update $\alpha^{(t)}$ by maximizing $p(\alpha|\boldsymbol{\theta}^{(t)}, \mathbf{y})$ over a grid on the interval (0,1)

end for

1.4.3. An FDR-based approach for keyphrase detection

After we obtain multiple posterior draws of $\boldsymbol{\theta}$ via MCMC, we can transform $\boldsymbol{\theta}$ back to the probability scale to obtain $\boldsymbol{\pi}$. To determine whether a candidate phrase is a keyphrase, a threshold is needed. We employ a FDR control method [44] using the Bayesian estimator $\hat{\boldsymbol{\pi}}$, obtained by averaging the posterior draws of $\boldsymbol{\pi}$. For a given probability cutoff h, the FDR can be estimated by $\widehat{FDR}(h) = \{\sum_{i=1}^n (1-\hat{\pi}_i) I(\hat{\pi}_i \geq h)\}/\{\sum_{i=1}^n I(\hat{\pi}_i \geq h)\}$, where $I(\cdot)$ is the indicator function. In practice, h is the largest number such that $\widehat{FDR}(h) \leq \gamma$, where γ is a pre-specified value such as 0.1, 0.2 or 0.3. Candidate phrases with $\hat{\pi}_i \geq h$ are identified as keyphrases. The larger γ is, the more keyphrases are identified. Evidently, this FDR control method allows different articles to have different probability cutoffs for keyphrase selection.

1.5. Evaluation using benchmark data

1.5.1. Hulth abstract data

1.5.1.1. Data and preprocessing

To evaluate the performance of the proposed BSS method and compare it with existing methods, we use the Hulth data set [30], which is regarded as the standard benchmark data in the keyphrase extraction literature. The data set consists of journal abstracts from Computer Science and Information Technology fields. Keywords for each article in the Hulth data set have been hand labeled. For simplicity, we consider keyword extraction instead of keyphrase extraction as keyphrases can be assembled from adjacent keywords in a post-processing step.

We preprocess all documents using standard natural language processing (NLP) steps before constructing a graph for each. These preprocessing steps include tokenization and part-of-speech tagging (POS-tagging), which help reduce the number of total words used for graph building. Tokenization is to split a sentence into a list of words, such that the unit for analysis is a word rather than a sentence. POS-tagging helps categorize each word into a word class such as noun, adverb and so on. As a result, words that do not carry much information, such as conjunction, preposition, can be removed from the candidate word list. After these two steps, each remaining word is used as a vertex in the graph. The edge weight between any two words is calculated as the number of co-occurrences in a window of two words, as suggested by Mihalcea and Tarau [43].

For each article, we randomly select five words from the keyword list and treat them as observed keywords. We evaluate the number of remaining keywords correctly identified by each method. Articles with less than ten keywords are excluded in our experiment, such that for each article, at least 50% of the keywords are unknown. This leaves us with 216

documents with around 20 keywords per article on average. A summary of the number of keywords is given in Table 1.1.

Table 1.1. Hulth abstract data: summary statistics for the number of keywords in 216 selected documents.

Min	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**
11	14	19	19.83	24	42

To apply the proposed method, we use both tMH within Gibbs (Algorithm 1) and acMH within Gibbs (Algorithm 2) to draw samples from the posterior distribution. For tMH within Gibbs, we set $b = 4$ so that the acceptance rate of the MCMC samples is between 20% and 40%, as mentioned in Section 1.4.1. For acMH within Gibbs, $v_i^{(t)}$ is set to be 1 in the first 10 iterations, and $2.4[Var(\theta_i^0, \theta_i^1, ..., \theta_i^{t-1}) + \varepsilon]$ for the remaining iterations as suggested by Haario et al. [23], where ε is a small positive constant so that the variance in the proposal distribution is always positive. For all of our experiments, we set ε to be 0.01. For each article, we run 50,000 MCMC iterations with the first 2,000 as burn-in samples and use the posterior mean of the remaining iterations for Bayesian inference.

We compare the performance of BSS with those of TR and SS, the two state-of-art methods for keyphrase extraction reviewed in Section 1.2. To select keywords for BSS, we use the FDR control method described in Section 1.4.3, with γ ranging from 0.05 to 0.2 for these short abstracts. To make a fair comparison, we control the total number of identified keywords to be the same across different methods. To be specific, we calculate the ratio between the number of identified keywords (by BSS) and the total number of candidate words among the 216 articles (say, $r\%$). Then for TR and SS, we select the top $r\%$ words (ranked by importance scores) as keywords for each article. As semi-supervised learning methods, both BSS and SS methods ensure that observed keywords are identified

as keywords. However, TR, as an unsupervised method, cannot guarantee this as it does not utilize information of observed labels. This would place TR at a disadvantage in the comparison. To avoid this, we force all the observed keywords to be keywords for TR as well.

1.5.1.2. FDR comparison

The actual FDR rates for BSS, TR and SS are reported in Table 1.2 for different FDR cutoffs ($\gamma = 0.05$, 0.10, 0.15, and 0.2). We report the results for both tMH and acMH within Gibbs and observe that they both have better performance than the two competitors. In general, tMH appears to identify more words as keywords while the actual FDRs for acMH are closer to the nominal values. However, the acMH algorithm requires $n \times T$ iterations, which greatly increases the computation time. Thus, for all the following analyses in this paper, we use tMH only.

In Table 1.2, we highlight the smallest actual FDR in bold for each FDR cutoff. We observe that our algorithms produced lowest FDRs in all the situations, suggesting that the proposed BSS can identify more true keywords. Between TR and SS, TR performs better than SS when the FDR threshold is small ($\gamma = 0.05$ or $\gamma = 0.1$), but worse when γ is larger.

1.5.1.3. Precision, recall and F-measure

In the literature of keyphrase extraction, precision, recall and F-measure are usually used as the performance measures. Precision is defined as

$$\text{Precision} = \frac{\text{the number of correctly identified words (True Positives)}}{\text{total number of identified words (True Positives+False Positives)}};$$

Table 1.2. Hulth abstract data: FDR comparison for BSS, TR and SS. Total no. of positives represents the total number of words identified as keywords. The actual FDRs for TR and SS are calculated using the following steps. Taking $\gamma = 0.1$ as an example, we have 2028 words identified as keywords from tMH with Gibbs, which is 18.0% of 11243 candidate words in the 216 documents. Then 18.0% is used as the cutoff for both TR and SS, that is, we select the top 18.0% of the candidate words with the highest importance scores to be keywords for each article. Then the corresponding actual FDRs across all the documents are computed.

	FDR Control	Total No. of Positives	Actual FDR		
			BSS	TR	SS
tMH within Gibbs	γ= 0.05	1438	**0.056**	0.086	0.086
	γ= 0.1	2028	**0.144**	0.177	0.175
	γ= 0.15	2899	**0.254**	0.273	0.282
	γ= 0.2	4229	**0.368**	0.391	0.383
acMH within Gibbs	γ= 0.05	1404	**0.059**	0.084	0.086
	γ= 0.1	1937	**0.137**	0.165	0.164
	γ= 0.15	2678	**0.241**	0.250	0.260
	γ= 0.2	3780	**0.344**	0.354	0.349

Recall is defined as

$$\text{Recall} = \frac{\text{the number of correctly identified words (True Positives)}}{\text{total number of keywords words (True Positives+False Negatives)}};$$

F-measure is the harmonic mean of precision and recall:

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}};$$

The performance measured based on the three criteria is summarized in Table 1.3. For each criterion and FDR cutoff, we highlight the largest number (i.e., best performance) in bold. Similar to the previous subsection, we observe that BSS performs the best, regardless of the cutoff value and measure. TR generally performs the worst because it cannot utilize the label information, even though all observed keywords are forced to be keywords for TR.

Table 1.3. Hulth abstract data: comparison of precision, recall and F-measure using different FDR control cutoff values.

FDR Control	Precision			Recall			F-measure		
	BSS	TR	SS	BSS	TR	SS	BSS	TR	SS
γ= 0.05	**0.944**	0.915	0.915	**0.317**	0.307	0.307	**0.475**	0.460	0.460
γ= 0.1	**0.856**	0.822	0.824	**0.405**	0.390	0.391	**0.550**	0.529	0.530
γ= 0.15	**0.745**	0.728	0.719	**0.504**	0.492	0.486	**0.601**	0.587	0.580
γ= 0.2	**0.632**	0.610	0.618	**0.624**	0.602	0.610	**0.628**	0.606	0.614

1.5.1.4. Factors affecting the performance

In order to better understand the behavior of the proposed method, we evaluate the performance based on two key factors: the number of keywords and the proportion of keywords in an article. Based on these two factors, we divide the articles into four groups of equal size using quartiles and evaluate the performance within each quartile. For each group, we count the overall number of correctly identified words (true positives), overall number of identified words (positives) and overall number of keywords based on $\gamma = 0.15$ so as to calculate an overall F-measure.

We show the overall F-measure vs. number of keywords in Figure 1.3. Group A represents articles with the smallest numbers of keywords (less or equal to the first quartile) and group D contains those with the largest numbers of keywords (larger or equal than the third quartile). Since we assume each article contains 5 observed keywords regardless of the actual number, it is not surprising that the F-measure decreases as the number of actual keywords increases. From Figure 1.3, we observe that BSS generally performs the best in terms of F-measure except for group D, where TR slightly outperforms SS and BSS. This implies that semi-supervised methods (BSS and SS) may not perform better than the unsupervised method

(TR) if the label information is sparse (i.e. observed keyphrases are just a very small portion of the true ones).

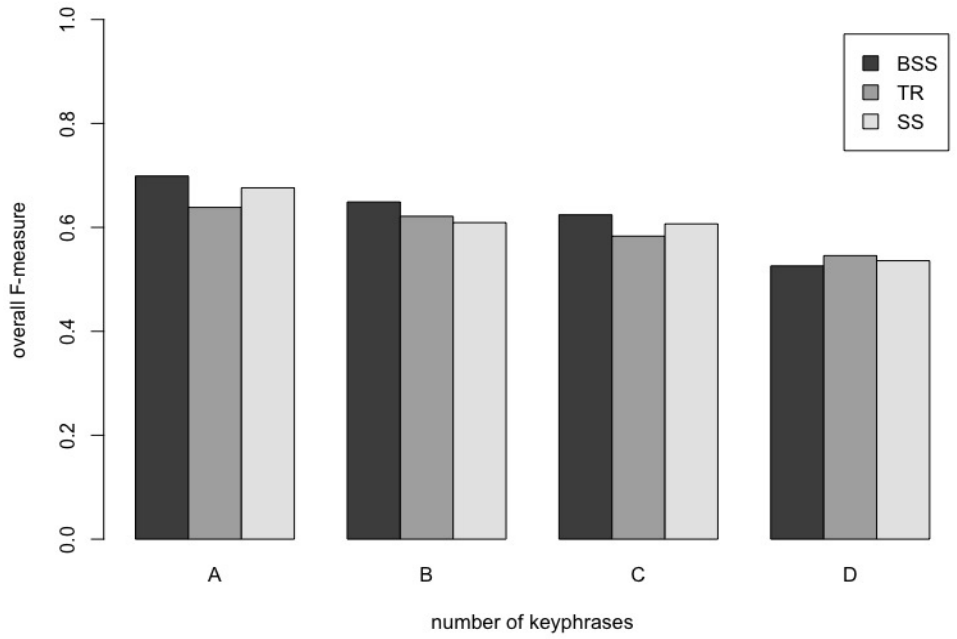


Figure 1.3. Hulth abstract data: bar chart of overall F-measure (calculated based on $\gamma = 0.15$) vs. number of keywords (from smallest to largest for the four groups of articles A-D).

Table 1.4. Hulth abstract data: performance comparison for different article groups by keyword proportion based on $\gamma = 0.15$. No. (proportion) of positives is the number (proportion) of words identified as keywords. The positive proportion of baseline methods (TR/SS) is 26.2% for all article groups. No. of true positives is the number of actual keywords identified.

Article Group	No. (Prop.) of Positives		No. of True Positives			F-measure			Total No. (Prop.) of Keywords
	BSS	TR/SS	BSS	TR	SS	BSS	TR	SS	
A	734 (0.223)	863 (0.262)	470	501	528	0.602	0.593	**0.624**	828 (0.251)
B	736 (0.249)	770 (0.261)	547	560	549	0.610	**0.612**	0.600	1060 (0.359)
C	812 (0.278)	763 (0.261)	632	613	581	**0.600**	0.596	0.565	1293 (0.443)
D	617 (0.327)	499 (0.264)	511	434	424	**0.594**	0.542	0.530	1102 (0.584)

In Table 1.4, we show the results by grouping articles according to the keyword proportion. The keyword proportion for an article is defined as the number of keywords divided by the total number of candidate words in the article. Articles are divided into four groups of equal size: A to D for articles with the smallest to largest proportions of keywords, respectively. For each method, we report the number and proportion (in bracket) of words identified as keywords, the number of actual keywords identified and F-measure. We also report the total number of keywords within each article group for reference. It is interesting to observe that as the keyword proportion increases, our method captures such information and selects more words (in proportion) to be keywords (proportion of positives increases from 22.6% in group A to 31.8% in group D), while the proportions for TR and SS do not show such a pattern (26.0% for all groups). F-measure comparison indicates that the advantage of BSS over the other methods increases when the keyword proportion increases. While the F-measure of BSS is around 0.6 for all the article groups, the F-measures for TR and SS appear to decrease as the keyword proportion increases. This is deemed to be a merit of the proposed BSS, showing that when an article has a relatively large proportion of keywords, BSS is able to detect that by returning a larger proportion of positives. By contrast, both TR and SS fail to reflect this underlying characteristic.

1.5.2. A long article in computer science

Unlike abstracts, a complete article may contain hundreds or even thousands of unique words. Using the whole set of words leads to a big graph. On the other hand, a longer article may still have a small number of keywords. In order to reduce the dimensionality, we apply stemming to reduce the number of unique words and a frequency-based filter to remove words that only appear once or twice in an document. Stemming is a process of reducing words to their root forms. For example, stemming can reduce three different words "compute", "computing" and "computer" into their common root "comput".

We apply our method to a long article with 586 candidate words and 27 keywords. After stemming and removing words that appear no more than twice in the article, we have 149 words left. Note that such preprocessing steps do not remove any keyword from the list of candidate words. Observed keywords are generated from the title, leading to 8 observed keywords, as listed in Table 1.5.

We use tMH within Gibbs and set $b = 5$ and $\gamma = 0.25$ to identify keywords. As in Section 1.5.1, we assemble identified keywords into keyphrases if they are next to each other in the document, and report them for each method in Table 1.5. We use () to denote the keywords that are not identified. For example, "(hydrocarbon) reservoir model" means the method finds keywords "reservoir" and "model" but does not find "hydrocarbon". All three methods identify the exactly same set of 12 keyphrases but BSS and SS identify one more keyword than TR. As mentioned in Section 1.5.1.1, all observed keywords are enforced to be keywords as an extra step for TR. By contrast, both semi-supervised methods identify the observed keywords automatically without the extra enforcement step, suggesting their efficient use of the observed label information.

1.6. Real Data Examples without Ground Truth

1.6.1. A statistical paper

We also apply our BSS method to a well known statistical paper [36], titled "Nonparametric Estimation from Incomplete Observations" to see if BSS can grasp the gist of the paper. Unlike all the previous examples, we don't have a list of hand labeled keyphrases for this paper. Fortunately, people who are familiar with Statistics know that this paper proposed a nonparametric method for estimating the survival function using lifetime (time-to-event) data. With all the preprocessing techniques applied and removal of mathematical formulas, we have 840 candidate words. After stemming and applying the frequency-based

Table 1.5. A long article in computer science: comparison of keyphrases identified by different methods under two scenarios. All identified keywords are assembled to keyphrases and we use () to denote the keywords that are not identified. For example: "ensemble (kalman) filter" means "ensemble" and "filter" were identified as keywords while "kalman" is not identified.

Observed keyphrases	BSS keyphrases	TR keyphrases	SS keyphrases
ensemble kalman filter, data assimilation methodology, reservoir, grid-enabling	ensemble kalman filter, data assimilation methodology, (hydrocarbon) reservoir model, grid-enabling, TIGRE grid computing environment, (pooling) license, grid computing, gridway metascheduler, reservoir simulation, (strategic) application area, EnKF, TIGRE	ensemble kalman filter, data assimilation methodology, (hydrocarbon) reservoir model, grid-enabling, TIGRE grid computing environment, (pooling) license, grid computing, gridway metascheduler, reservoir simulation, (strategic) application (area), EnKF, TIGRE	ensemble kalman filter, data assimilation methodology, (hydrocarbon) reservoir model, grid-enabling, TIGRE grid computing environment, (pooling) license, grid computing, gridway metascheduler, reservoir simulation, (strategic) application area, EnKF, TIGRE

filter, we end up with 180 candidate words. Furthermore, we treat "nonparametric estimation", "incomplete observations" as observed keyphrases, which is directly extracted from the title.

In this example, we again use the faster algorithm, tMH within Gibbs, and set $b = 5$ and $\gamma = 0.3$. We show keywords identify by the three methods, BSS, TR and SS, and grouped adjacent ones into keyphrases in Table 1.6. All three methods identified "product-limit estimation", the main method introduced in the paper; "p(t)", referring to the survival function; "reduced-sample estimates",which is another main method discussed in this paper; "death", "age", and "integer numbers", together indicating that the paper is about time-to-event data. Besides, BSS found "actuarial estimates", which is an important term in the paper, while the other two methods did not. From all keywords found by BSS, we can conclude that the paper focuses on different estimation methods for lifetime data.

Table 1.6. A statistical paper: comparison of identified keyphrases/keywords identified by different methods.

Observed keyphrases	BSS keyphrases/keywords	TR keyphrases/keywords	SS keyphrases/keywords
nonparametric estimation, incomplete observations	nonparemetric estimation, incompleted observation, product-limit estimation, p(t), reduced-sample estimates, actuarial estimates, item, age, death, integer number, important, not, case, results.	nonparametric estimation, incompleted observation, product-limit estimation, p(t), reduced-sample estimates, death, loss, item, integer number, age, limit, variance, not, case, value, n	nonparametric estimation, incompleted observation, product-limit estimation, p(t), reduced-sample estimates, death, item, loss, age, integer number, variance, not, case, online

1.6.2. An Amazon product review example

We apply keyphrase extraction methods to an Amazon product review example, available from McAuley and Leskovec [42]. Beside the actual content, the review contains a brief summary, as shown in Table 1.7. We treat the summary "oatmeal for oatmeal lovers" as the observed keyphrase, so "oatmeal" and "lovers" are observed keywords. After applying the preprocessing steps described in Section 1.5.2 to the review content, we end up with 83 candidate words. The tMH within Gibbs algorithm with $b = 5$ is applied, and to identify keywords, we set $\gamma = 0.25$. The keywords identified by the different methods BSS, TR and SS are listed in Table 1.8. BSS returns quite a few very informative words, including: "mccann", "oatmeal", "personally", "like", "well", "loved/lovers". By combining words that are adjacent in the original review, we have the following keyphrases: "mccann oatmeal", "personally like", "well loved oatmeal", which are sufficient for people to interpret that the reviewer has a quite positive feedback on this product. In contrast, keywords found by TR and SS contain some noninformative words such as "cook", "water","eater", "single", as listed in Table 1.8. Those words may bring some difficulty to interpret this review. In

addition, they are isolated from each other in the initial review so it is hard to combine those into keyphrases.

Table 1.7. A review example of Amazon Fine Food review data from McAuley and Leskovec [42].

Summary: Oatmeal for oatmeal lovers
McCann's makes oatmeal for every oatmeal connoisseur, whether one likes it from the raw pellet state that cooks for half an hour, to the sloth addled instant, which can be done in the microwave for under three minutes. It's all good, that's for sure, and the beauty of the instant variety is that it is available in different flavors as well as regular. This variety pack allows different tastes to be explored, as well as giving you a chance to experience the difference between McCann's and other well-known oatmeals. What I personally like about McCann's is that it cooks up thicker and with more body than the top brand here in America. The Apples & Cinnamon, though, tends to be a little liquidy so you may want to experiment with the amount of water you add. In my 1300watt microwave the oatmeal cooks up in about one minute and twenty-seven seconds, so you should also watch that to get a handle on how much time and water to use. The only bad thing -- if you can consider it a bad thing -- about this offering is that you have to buy in lot so you'll end up with six ten-count boxes. This is good if you have a whole family of oatmeal-eaters, but if you're a single person alone -- well, love oatmeal.

Table 1.8. Keywords identified by BSS, TR and SS. NOTE: "lovers/loved" means it can be either "lover" or "loved" depending on contexts as they have the same root in the stemming step.

Observed keyphrases	BSS keywords	TR keywords	SS keywords
oatmeal, lovers	mccann, oatmeal, lovers/loved, personally, every, like, well, known, alone	mccann, oatmeal lovers/loved, well, cook, different, water, thing, good	mccann, oatmeal, lovers/loved, personally, every, well, alone, single, eater

1.7. Discussion

We propose a novel Bayesian method for semi-supervised keyphrase extraction in situations when a (small) subset of keyphrases is known. We use an informative prior to incorporate the graph-based information about the document structure. We propose two MCMC algorithms, tMH within Gibbs and acMH within Gibbs, to sample from the high-dimensional posterior distribution. Both algorithms provide favorable results compared to the two state-of-the-art methods TR and SS. We recommend using tMH within Gibbs due to its computational advantage.

Apart from existing methods, where keywords are selected based on importance scores, our method produces the posterior probability of each word being a keyword. Thus, we can use an FDR-based method to select the number of keywords. In practice, one has to predetermine the FDR threshold γ. We select γ using the following approach: firstly categorize articles into document pools by lengths and numbers of observed keyphrases; then for short articles, γ can be set at 0.2 or lower since we do not expect a lot of words/phrases identified; for long articles with lots of keyphrases observed (for examples, articles with long titles), we can select γ at 0.25; for long articles with limited information observed, we can pick a large γ such as 0.3 so as to identify more keyphrases/keywords. As shown in our examples, this approach works reasonably well. We note that after fixing γ for a given document pool, our method still has the ability to select different proportions of words as keywords for different articles. In contrast, all existing methods select a fixed proportion of words as keywords. Our experiment indicates that the proposed FDR-based method tends to identify more when the keyword proportion is higher. We demonstrate the performance of the proposed methods using a collection of different documents, including abstracts, articles from different subjects, and an Amazon product review. We find that our BSS method has better or competitive performance in all cases.

CHAPTER 2

Jackknife Empirical Likelihood for Assessing Heterogeneity in Meta-analysis

2.1. Introduction

Meta-analysis is a statistical procedure that combines results from multiple independent studies to achieve a reliable conclusion. Since its introduction, it has been widely used in fields such as biology, psychology, medical and social sciences [5, 60, 54, 45]. Meta-analysis can be extremely useful when different studies addressing the same research question have inconsistent results, perhaps due to reasons such as small sample sizes, sparse data, different experimental conditions, and heterogeneous population subtypes, etc. For instance, Goyal et al. [22] conducted a literature search to examine if obesity has an effect on the outcomes of spinal surgeries. Several studies showed statistically significant evidence of higher perioperative morbidity for patients with obesity while others did not find any significant difference. By conducting a meta-analysis, they were able to draw an overall conclusion that the difference was not statistically significant among patients with a minimally invasive surgery, but significant for people who underwent open surgeries.

In meta-analysis, the random-effects (Re) model [7, 15, 24, 6] has been commonly used to model observed effect sizes from component studies. Suppose that there are K independent studies involved, each reporting Y_i, the estimates of the effect sizes, and s_i^2, their within-study variances. The Re model assumes that for study i, $i = 1, 2, ..., K$,

$$Y_i = \theta_i + \epsilon_i, \qquad \theta_i = \theta + \delta_i, \tag{2.1}$$

where all δ_i's and ϵ_i's are independent, θ_i is the true effect size of study i that may vary across studies, ϵ_i is the experimental error of Y_i, θ is the mean effect size, and δ_i is the random error of θ_i. Further, the Re model assumes $\epsilon_i \sim N(0, \sigma_i^2)$, where σ_i^2 denotes the within-study variance, which is usually replaced by its estimate s_i^2 and then treated as known for convenience, and $\delta_i \sim N(0, \tau^2)$, where τ^2 is the variance parameter that accounts for the between-study heterogeneity. In a special case, when $\tau^2 = 0$, all studies share a common effect size θ (i.e., $\theta_i \equiv \theta$), and consequently the Re model reduces to the fixed-effect (Fe) model [41, 6, 7]. Many early studies in meta-analysis focused on estimating and testing the overall treatment effect θ [41, 13]. Later on, people started to study the heterogeneity parameter τ^2, because a poorly estimated τ^2 can lead to inaccurate estimation and inference of θ. In this paper, we aim to construct nonparametric confidence intervals (CIs) for τ^2, which are especially useful in meta-analysis of rare binary events data, where the normality assumption in the Re model is often unmet.

Cochran [12] proposed a Q-statistic, which was later used by others to test if different experiments have variation in effect sizes in addition to their own experimental errors, namely

$$Q = \sum_{i=1}^{K} w_i (Y_i - \hat{\theta})^2, \tag{2.2}$$

where $\hat{\theta} = \sum_i w_i Y_i / \sum_i w_i$ is the estimated overall effect size and $w_i = 1/s_i^2$ is the weight assigned to study i. Under the Fe model with the normality assumption, Q approximately follows a chi-square distribution with $K-1$ degrees of freedom, denoted χ^2_{K-1}, given the sample size of each study is large. This asymptotic property has been widely used in random-effects meta-analysis to test the null hypothesis $H_0 : \tau^2 = 0$, and this test has been known as the standard Q test. However, in meta-analysis with non-normal Y_i's, the distribution of

Q is complex in general, which may not be simply approximated by the χ^2_{K-1} distribution [28].

To construct CIs for the heterogeneity parameter τ^2, there are three major approaches in the literature [63]. The first, also the most common approach is to consider the distribution of Cochran Q-statistic in (2.2) or a modified Q-statistic (e.g., Q with a different weighing scheme) under the *Re* model as a function of τ^2, and then obtain the intervals via a test-inverting process. Methods using this approach include (i) QP, the Q-profile method considered in both Knapp et al. [37] and Viechtbauer [56], (ii) MQP, the modified Q-profile method, proposed in Knapp et al. [37], (iii) BT, proposed in Biggerstaff and Tweedie [3], (iv) BJ, proposed in Biggerstaff and Jackson [4], (v) J, proposed in Jackson [31], (vi) AJ, the approximate Jackson method proposed in Jackson et al. [34], and (vii) QP_{UT}, the unequal-tail Q profile method proposed in Jackson and Bowden [32]. The second approach is to construct profile likelihood CIs under the *Re* model based on maximum likelihood (ML) estimation or restricted maximum likelihood (REML) estimation, denoted by PL_{ML} [24] and PL_{REML} [56], respectively. The third is to construct Wald confidence intervals based on the ML or REML estimation, denoted by Wald_{ML} and $\text{Wald}_{\text{REML}}$[3], respectively. Other approaches include CIs based on bootstrapping [16], the Sidik and Jonkman (SJ) estimator of τ^2 that is derived from the weighted residual sum of squares in the framework of a linear regression model [51], and an improved SJ estimator τ^2 [52], denoted by BS, SJ, and SJ_{HO}, respectively. We refer readers to Zhang et al. [63] for a detailed review of these CI methods.

The above methods typically assume that the *Re* model (2.1) has normally distributed ϵ_i's and δ_i's. This yields $Y_i|\theta_i \sim N(\theta_i, \sigma_i^2)$ and $\theta_i \sim N(\theta, \tau^2)$, which Jackson and White [33] referred to as the within-study and between-study distributions. The normality of the within-study distributions is often justified by the central limit theorem that states the sampling distribution of an estimated effect size is approximately normal when the sample size is large. However, it would be invalid for data with a small to moderate sample size or data with severe skewness or heavy tails. The normality of the between-study distribution is usually

assumed based on mathematical convenience, and it is hard to defend why the underlying true study effects should be always normal. Hardy and Thompson [25] provided practical strategies to examine this assumption, including informal inspection of normal Quartile-Quartile plots and formal normality testing (e.g., the Shapiro-Wilk test [50]). In the past, concerns on these normality assumptions have been raised, and it has been further pointed out that statistical methods that make fewer normality assumptions should be considered more often in practice [29, 33]. We propose to use jackknife empirical likelihood (JEL) for constructing CIs for the heterogeneity parameter, which does not require any distributional assumption on effect sizes.

Empirical likelihood (EL) is an effective nonparametric tool for various statistical inferences [46, 47]. One nice feature of EL is that it only requires independent and identically distributed (i.i.d) samples with no extra assumption. The major idea of EL is to maximize the profile empirical likelihood function with constraints from parameters of interest. However, computation would be extensive if such constraints are nonlinear. Jackknife empirical likelihood (JEL), introduced by Jing et al. [35], constructs jackknife pseudo-values and then treats them as i.i.d observations in the empirical likelihood function so that the constraints are always linear, greatly facilitating the computation involved in many statistical problems (e.g., [1, 62, 49]).

The rest of the chapter is organized as follows. In Section 2.2, we first review the preliminaries on EL and JEL, and then apply JEL to interval estimation of the heterogeneity parameter τ^2 in random-effects meta-analysis. We prove that the resulting jackknife empirical likelihood ratio approximately follows a χ^2_1 distribution or a scaled χ^2_1 distribution from which CIs can be constructed. In Sections 2.3 and 2.4, using simulated data and real examples for rare binary events, we examine the performance of the proposed JEL methods and compare them with existing methods in situations when non-normality occurs. We conclude the paper with a brief discussion in Section 2.5.

2.2. Methodology

2.2.1. Review of EL and JEL

Let $\mathbf{X} = (X_1, ..., X_n)$ denote i.i.d random variables from an unknown distribution F and $\boldsymbol{\theta}$ be a $p \times 1$ parameter vector of interest with conditions $E\{\mathbf{m}(X_1, \boldsymbol{\theta})\} = 0$, where $\mathbf{m}$ is a $d \times 1$ vector-valued function. Let $\mathbf{x} = (x_1, ..., x_n)$ be the observed samples and suppose $p_i \geq 0$ and $\sum_{j:x_j=x_i} p_j = F(x_i) - F(x_i-)$ for $i = 1, 2, ..., n$. Then the empirical likelihood of p_i's is defined as

$$L(p_1, ..., p_n) = L(p_1, ...p_n; \mathbf{x}) = \prod_{i=1}^{n} p_i, \tag{2.3}$$

and it reaches the maximum when $p_i = 1/n$. Consider the following hypothesis test

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Let $\tilde{p}_1, ..., \tilde{p}_n$ be the maximum empirical likelihood estimators (MELEs) under H_0 that maximize (2.3) with the following constraints:

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \cdot \mathbf{m}(x_i, \boldsymbol{\theta}_0) = \mathbf{0}. \tag{2.4}$$

We denote $(\tilde{p}_1, ..., \tilde{p}_n)$ by $\{\tilde{p}_1(\boldsymbol{\theta}_0), ..., \tilde{p}_n(\boldsymbol{\theta}_0)\}$. Then $L(\tilde{p}_1, ..., \tilde{p}_n) = L\{\tilde{p}_1(\boldsymbol{\theta}_0), ..., \tilde{p}_n(\boldsymbol{\theta}_0)\} = L(\boldsymbol{\theta}_0)$ can be considered as a profile empirical likelihood function for $\boldsymbol{\theta}_0$. As a result, a likelihood ratio test can be derived:

$$R(\boldsymbol{\theta}_0) = \frac{\max L(p_1, ..., p_n)}{\max_{H_0} L(p_1, ..., p_n)} = \frac{(n^{-1})^n}{L(\tilde{p}_1, ..., \tilde{p}_n)} = \frac{(n^{-1})^n}{L(\boldsymbol{\theta}_0)},$$

where the null hypothesis H_0 is rejected when $R(\boldsymbol{\theta}_0)$ has large values. Owen [47] proved that $-2\log\{R(\boldsymbol{\theta})\}$ has a limiting χ^2_d distribution under mild conditions, thus a confidence region of $\boldsymbol{\theta}$ can be constructed by inverting the likelihood ratio test.

A simple application of empirical likelihood is for estimating the distribution mean $\theta = E(X_1)$. The moment condition on the parameter is $E\{m(x,\theta)\} = 0$, where $m(x,\theta) = x - \theta$. Under the null hypothesis $H_0 : \theta = \theta_0$, (2.4) can be simplified to

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i x_i = \theta_0. \tag{2.5}$$

Owen [46] showed that $L(\theta_0)$ can be calculated in a closed form under (2.5) using the Lagrange multiplier, namely

$$L(\theta_0) = \prod_{i=1}^{n} \frac{1}{n\{1 + \lambda(x_i - \theta_0)\}}, \tag{2.6}$$

where λ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \theta_0}{1 + \lambda(x_i - \theta_0)} = 0. \tag{2.7}$$

This yields

$$-2\log\{R(\theta_0)\} = 2\sum_{i=1}^{n}\{1 + \lambda(x_i - \theta_0)\} \xrightarrow{d} \chi^2_1.$$

However, when conditions are generalized from $E\{\mathbf{m}(X_1, \boldsymbol{\theta})\} = 0$ to $E\{\mathbf{m}(\mathbf{X}, \boldsymbol{\theta})\} = 0$, the use of empirical likelihood may not be appealing. Such generalization can add nonlinear constraints of p_i to the profile empirical likelihood and make it computationally difficult when finding the solution. Jing et al. [35] used U-statistics to illustrate the problem. Let the parameter of interest $\theta = E[h(X_1, ..., X_r)]$, where $2 \leq r \leq n$ and $h(\cdot)$

is a symmetric kernel function of θ. Then the one-sample U-statistic of θ is given by $U_n = \binom{n}{r}^{-1} \sum_{1\le i_1\le\ldots\le i_r\le n} h(X_{i_1},...,X_{i_r})$, and the constraints in (2.4) can be written as

$$p_i \ge 0, \sum_{i=1}^{n} p_i = 1, \tilde{\theta}(\mathbf{X}, p_1, ..., p_n) = \theta_0,$$

where

$$\tilde{\theta}(\mathbf{X}, p_1, ..., p_n) = \binom{n}{r}^{-1} \sum_{1\le i_1\le\ldots\le i_r\le n} n^r p_{i_1}...p_{i_r} h(X_{i_1}, ..., X_{i_r}).$$

It is easy to see that the constraints are nonlinear in p_i's. To solve such a problem, they proposed a jackknife empirical likelihood approach that creates jackknife pseudo-values as follows:

$$\hat{V}_i = nT_n - (n-1)T_{n-1}^{(-i)}, \tag{2.8}$$

where T_n is an unbiased estimator of the parameter using all n samples and $T_{n-1}^{(-i)}$ is the same estimator but with the ith sample removed. Those pseudo-values are treated as observed values in the empirical likelihood so that the constraints become

$$p_i \ge 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \hat{V}_i = \theta_0,$$

which can be solved easily using (2.6) and (2.7) with x_i's replaced by $\hat{V}_i$'s. Jing et al. [35] further showed that for the U-statistics, $-2\log\{R(\theta)\} = 2\sum_{i=1}^{n}\{1+\lambda(\hat{V}_i-\theta)\} \xrightarrow{d} \chi_1^2$ under mild conditions. The main advantage of JEL is that it can greatly reduce the computational complexity by turning the statistic of interest into a sample mean based on the jackknife pseudo-values.

2.2.2. JEL CIs for the heterogeneity parameter in meta-analysis

Consider the *Re* model (2.1) without the normality assumptions, where we only assume that the first two moments satisfy $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_i^2$, $E(\delta_i) = 0$, and $Var(\delta_i) = \tau^2$. We start with finding an unbiased estimator of the between-study heterogeneity parameter τ^2 in a meta-analysis of K studies in total. Dersimonian and Kacker [14] showed that

$$E(Q) = \tau^2 \left(\sum_{i=1}^{K} w_i - \frac{\sum_{i=1}^{K} w_i^2}{\sum_{i=1}^{K} w_i} \right) + \left(\sum_{i=1}^{K} w_i \sigma_i^2 - \frac{\sum_{i=1}^{K} w_i^2 \sigma_i^2}{\sum_{i=1}^{K} w_i} \right),$$

where σ_i^2's are assumed to be known and will be replaced by s_i^2's (i.e., σ_i^2 and s_i^2 are treated equivalently), w_i can be any positive weight for the ith study in (2.2). Using the method of moments, an unbiased estimator of τ^2 is then given by

$$\hat{\tau}^2 = \frac{Q - \left(\sum_{i=1}^{K} w_i \sigma_i^2 - \frac{\sum_{i=1}^{K} w_i^2 \sigma_i^2}{\sum_{i=1}^{K} w_i} \right)}{\sum_{i=1}^{K} w_i - \frac{\sum_{i=1}^{K} w_i^2}{\sum_{i=1}^{K} w_i}}. \tag{2.9}$$

Once each study has been assigned a weight, we can construct confidence intervals for τ^2 by finding jackknife pseudo-values via (2.9) and (2.8). In particular, we consider two sets of weights: (i) equal weights $w_i = 1/K$, (ii) inverse-variance weights $w_i = 1/s_i^2$.

2.2.2.1. JEL with equal weights

For equal weights, the estimator of τ^2 in (2.9) can be written as

$$T_K = \frac{\sum_{i=1}^{K} (Y_i - \bar{Y})^2}{K - 1} - \frac{\sum_{i=1}^{K} s_i^2}{K}, \tag{2.10}$$

and jackknife pseudo-values $\hat{V}_i$'s can be obtained via $\hat{V}_i = KT_K - (K-1)T_{K-1}^{(-i)}$, where $T_{K-1}^{(-i)}$ is defined similarly as in (2.8) (i.e., recomputing the estimator by leaving study i out). It

is straightforward to show that all pseudo-values are also unbiased estimates of τ^2. Further, the profile empirical likelihood can be expressed by

$$L(\tau^2) = \max\{\prod_{i=1}^{K} p_i : p_i \geq 0, \sum_{i=1}^{K} p_i = 1, \sum_{i=1}^{K} p_i \hat{V}_i = \tau^2\},$$

and the log empirical likelihood ratio is then given by

$$-2\log\left\{R(\tau^2)\right\} = 2\sum_{i=1}^{K}\left\{1 + \lambda(\hat{V}_i - \tau^2)\right\}, \tag{2.11}$$

where λ is the solution of

$$\frac{1}{K}\sum_{i=1}^{K}\frac{\hat{V}_i - \tau^2}{1 + \lambda(\hat{V}_i - \tau^2)} = 0. \tag{2.12}$$

Define $\delta_i + \epsilon_i \triangleq \mu_i$ and $Var(\mu_i^2) \triangleq S_i^2$. The following theorem states that the log empirical likelihood ratio in (2.11) converges to χ_1^2 as the number of studies K increases.

Theorem 1. *For the τ^2 estimator with equal weights $w_i = 1/K$, under assumptions that $\sigma_i^2 > 0$ and $S_i^2 < \infty$ for $i = 1, 2, \ldots, K$, we have*

$$l(\tau^2) = -2\log\{R(\tau^2)\} \xrightarrow{d} \chi_1^2 \text{ as } K \to \infty.$$

The proof is given in Section A1.1 of Appendix A. As a result, an asymptotic $100(1-\alpha)\%$ CI for τ^2 using JEL with equal weights can be constructed as

$$\{\tau^2 : -2\log\{R(\tau^2)\} < \chi^2_{1,1-\alpha}\}. \tag{2.13}$$

We refer to this CI method as JEL with equal weights, labeled JEL_{EQ}.

2.2.2.2. JEL with inverse-variance weights

For inverse-variance weights $w_i = 1/s_i^2$, the estimator of τ^2 in (2.9) can be written as

$$T_K = \frac{Q - (K-1)}{A_0}, \tag{2.14}$$

where $A_0 = \sum_i w_i - (\sum_i w_i^2)/(\sum_i w_i)$. Then jackknife pseudo-values and the log empirical likelihood ratio can be calculated in the same way as the equal-weight case. With the same definition of S_i^2 in Section 2.2.2.1, we establish the following theorem.

Theorem 2. *For the τ^2 estimator with inverse-variance weights $w_i = 1/s_i^2$, assuming that $\sigma_i^2 > 0$, $\epsilon < \max_i s_i^2/\min_i s_i^2 < 1/\epsilon$ for some constant $\epsilon > 0$, and $S_i^2 < \infty$ for $i = 1, 2, \ldots, K$, we have*

$$C \cdot l(\tau^2) = -2C \log\{R(\tau^2)\} \xrightarrow{d} \chi_1^2 \text{ as } K \to \infty,$$

where C is a constant scale factor.

The definition of C and the proof of Theorem 2 can be found in Section A1.2 of Appendix A. Similarly, an asymptotic $100(1-\alpha)\%$ CI for τ^2 can be constructed using (2.13), which we refer to as JEL with inverse-variance weights, labeled JEL_{IV}.

2.2.2.3. Ending remarks

With equal and inverse-variance weights, the estimator in (2.9) becomes the Hedges and Olkin (HO) estimator Hedges and Olkin [27] and the Dersimonian and Laird (DL) estimator Dersimonian and Laird [15], respectively, but without truncation. As $\tau^2 \geq 0$, the HO and DL estimators both truncate negative values to zero, which leads to biased estimation. Thus, when constructing jackknife pseudo-values to obtain JEL CIs, we should not apply truncation, in order to ensure unbiasedness of the τ^2 estimates. However, if an upper/lower

bound of a JEL CI is negative, we should replace it with zero; that is, for a JEL CI that covers zero, the lower bound is automatically reset to zero; in the extreme case when a JEL CI is entirely below zero, it is reset to $\{0\}$.

2.3. Simulation focusing on rare binary events

2.3.1. Simulation set-up

We conduct simulation in the context of meta-analysis of rare binary events, to examine the performance of the proposed JEL methods on interval estimation of the heterogeneity parameter τ^2 and to compare it with existing CI methods. In a meta-analysis of size K, each of the K studies is represented by a 2×2 contingency table with elements $(n_i^C, n_i^T, x_i^C, x_i^T)$, satisfying $x_i^C \sim \text{Binomial}(n_i^C, p_i^C)$ and $x_i^T \sim \text{Binomial}(n_i^T, p_i^T)$, where $p_i^C(p_i^T)$ is the event probability, $n_i^C(n_i^T)$ and $x_i^C(x_i^T)$ are the number of subjects and events in the control (treatment) group, respectively. To generate event probabilities, Li and Wang [39] proposed a flexible binomial-normal simulation model, which allows treatment and control groups to have different within-group variability on the logit scale, namely

$$\text{logit}(p_i^C) = \mu_i - \omega\theta_i, \qquad \text{logit}(p_i^T) = \mu_i + (1-\omega)\theta_i, \tag{2.15}$$

where $\mu_i \sim N(\mu, \sigma^2)$, $\theta_i \sim N(\theta, \tau^2)$, $\mu_i's$ and $\theta_i's$ are independent, and ω is a constant between 0 and 1. Bhaumik et al. [2] employed the model with $\omega = 0$, where the treatment group is assumed to have larger variability than the control group. When $\omega = 0.5$, (2.15) is reduced to a commonly used model in Smith et al. [53] where both groups are assumed to have equal variability. In our simulation, we set $\omega \in \{0, 0.5, 1\}$, representing smaller/equal/larger variability in the control group, compared with the treatment group. Further, to understand how the proposed nonparametric methods perform under different distributions of treatment effects, we adapt the model (2.15) by simulating θ_i's from (i) $\theta_i \sim N(\theta, \tau^2)$; (ii) $\theta_i = \theta +$

$\tau t_i/\sqrt{3}$, where $t_i \sim T_3$ (a standard t distribution with three degrees of freedom); (iii) $\theta_i = \theta + \tau(e_i - 1)$, where $e_i \sim \exp(1)$. Note that the non-normal distributions also have mean θ and variance τ^2, but (ii) represents heavy-tailed distributions and (iii) represents skewed distributions. Since the parameter of our interest is τ^2 instead of the overall effect θ, we set $\theta = 0$, $\sigma^2 = 0.5$ but vary τ^2 from zero to one with step size 0.1. Further, to estimate the treatment effect θ_i measured by log odds ratio (LOR) in study i, we add 0.5 to each cell count in the ith contingency table, as suggested by Walter and Cook [57], to reduce bias,

$$Y_i = \hat{\theta}_i = \log \frac{x_i^C + 0.5}{n_k^C - x_i^C + 0.5} - \log \frac{x_i^T + 0.5}{n_k^T - x_i^T + 0.5}.$$

We further estimate the within-study variance by

$$s_i^2 = \frac{1}{x_i^C + 0.5} + \frac{1}{n_i^C - x_i^C + 0.5} + \frac{1}{x_i^T + 0.5} + \frac{1}{n_i^T - x_i^T + 0.5}.$$

Other parameters are set up as follows. The number of studies K is set to be 20, 50, and 80. As in Li and Wang [39], to allow varying allocation ratios across studies, we set $n_i^T = R_i n_i^C$, where $\log_2 R_i \sim N(\log_2 R, \sigma_R^2)$, $R = 1$, $\sigma_R^2 = 0.5$, and n_i^C's are randomly generated from uniform$[2000, 3000]$ for large-sample (LS) cases and from uniform$[20, 1000]$ for small-sample (SS) cases. To reflect rare and very rare event rates, we set $\mu \in \{-2.5, -5\}$, which is equivalent to $\{0.076, 0.0067\}$ in the probability scale, respectively. All those choices of K, μ, ω, along with two scenarios of sample sizes (LS vs. SS) and three distributions of θ_i's, result in 108 combinations in our simulation for different τ^2 values in the set $\{0, 0.1, ..., 1\}$. For each unique setting, 1000 replicate datasets are generated.

To benchmark the performance of our proposed JEL methods on interval estimation, we calculate CIs using JEL_{EQ} and JEL_{IV} as well as a comprehensive list of existing methods. The list includes fourteen methods mentioned in the introduction: QP, MQP, QP_{UT}, BT, BJ, J, AJ, SJ, SJ_{HO}, PL_{ML}, PL_{REML}, Wald_{ML}, $\text{Wald}_{\text{REML}}$, and BS. We implement BS in a nonparametric manner: first randomly draw $B = 200$ samples of size K with replacement

from a simulated dataset with K component studies; next, compute the DL estimator $\hat{\tau}^2_{DL}$ [15] for each of the B samples and obtain the empirical distribution of $\hat{\tau}^2_{DL}$; last, construct the 95% CI using the 2.5th and the 97.5th percentiles of the empirical distribution. Thus, all these methods except for BS, JEL_{EQ}, and JEL_{IV} assume the normality for valid inference. The performance metric used in our simulation is the empirical coverage probability, defined as the proportion of computed CIs that cover the true value of τ^2. For each method, 95% CI's are computed so that the nominal level of the coverage probability is 0.95.

2.3.2. JEL_{EQ} is superior to JEL_{IV}

	Normal		T_3		Exponential	
τ^2	JEL_{IV}	JEL_{EQ}	JEL_{IV}	JEL_{EQ}	JEL_{IV}	JEL_{EQ}
0	0.975	**0.971**	0.979	**0.973**	0.980	**0.977**
0.1	0.896	**0.930**	0.688	**0.737**	0.796	**0.853**
0.2	0.876	**0.911**	0.669	**0.733**	0.776	**0.851**
0.3	0.886	**0.937**	0.631	**0.740**	0.787	**0.852**
0.4	0.869	**0.928**	0.619	**0.747**	0.757	**0.841**
0.5	0.868	**0.930**	0.579	**0.727**	0.723	**0.850**
0.6	0.863	**0.942**	0.564	**0.739**	0.715	**0.853**
0.7	0.847	**0.930**	0.493	**0.725**	0.698	**0.848**
0.8	0.830	**0.922**	0.503	**0.750**	0.630	**0.843**
0.9	0.809	**0.919**	0.459	**0.733**	0.639	**0.854**
1	0.790	**0.929**	0.374	**0.706**	0.577	**0.850**

Table 2.1. Empirical coverage probabilities of 95% CIs constructed using JEL_{EQ} and JEL_{IV} for large-sample settings with $K = 50$, $\mu = -2.5$, and $\omega = 0.5$. Note that for $\tau^2 = 0$, the distribution of treatment effects is irrelevant and so the three settings for the different distributions are merely replicates.

We begin with the comparison between JEL_{EQ} and JEL_{IV}, the two JEL methods with different weighing schemes. Table 2.1 shows empirical coverage probabilities of the two methods for large-sample settings with $K = 50$, $\mu = -2.5$, and $\omega = 0.5$. While the two

methods have similar performance at $\tau^2 = 0$, JEL_{IV} provides consistently lower coverage than JEL_{EQ} for all positive τ^2 values. As τ^2 gets larger, the performance of JEL_{IV} becomes worse but JEL_{EQ} is relatively robust to the change of τ^2 so that the advantage of JEL_{EQ} over JEL_{IV} gets larger, regardless of the type of distribution. This advantage prevails in all other settings as well (results are omitted for brevity).

Recall that JEL_{IV} uses $1/\sigma_i^2$ as weights when estimating τ^2 and further assumes these variances be known. But in practice, they have to be estimated by $1/s_i^2$. By contrast, JEL_{EQ} uses constant weights $1/K$, and so avoids errors in estimating such weights. Due to the superior performance of JEL_{EQ} as well as its simplicity, we prefer JEL_{EQ} to JEL_{IV}. In what follows, we report results from the different methods but excluding JEL_{IV}.

2.3.3. When between-study heterogeneity does not exist ($\tau^2 = 0$)

Method	Large Sample					Small Sample				
K	JEL_{EQ}	QP	MQP	QP_{UT}	BT	JEL_{EQ}	QP	MQP	QP_{UT}	BT
20	0.974	0.974	0.974	0.990	1	0.983	0.985	0.986	0.997	1
50	0.975	0.975	0.975	0.987	1	0.977	0.988	0.988	1	1
80	0.970	0.983	0.983	0.990	1	0.992	0.991	0.992	0.998	1
K	BJ	J	AJ	SJ	SJ_{HO}	BJ	J	AJ	SJ	SJ_{HO}
20	1	0.976	0.989	0	0	1	0.994	0.994	0	0
50	1	0.972	0.977	0	0	1	0.993	0.994	0	0
80	1	0.974	0.981	0	0	1	0.998	0.981	0	0
K	PL_{ML}	PL_{REML}	Wald_{ML}	$\text{Wald}_{\text{REML}}$	BS	PL_{ML}	PL_{REML}	Wald_{ML}	$\text{Wald}_{\text{REML}}$	BS
20	0.994	0.986	1	1	0.992	0.993	0.986	1	1	0.996
50	0.976	0.971	0.999	0.999	0.984	N/A	N/A	1	1	0.996
80	0.983	0.980	0.998	0.997	0.990	N/A	N/A	1	1	0.992

Table 2.2. Empirical coverage probabilities of 95% CIs constructed using different methods for settings with $\tau^2 = 0$ and $\mu = -2.5$. Note that when $K = 50$ or 80, PL_{ML} and PL_{REML} failed to construct CIs for small-sample settings, due to the convergence issue.

When the heterogeneity does not exist, the treatment effects are constant across component studies so that the distribution of θ_i's and the value of ω become irrelevant. Table 2.2

shows empirical coverage probabilities for 15 different methods of computing 95% CIs for τ^2 by the number of studies K in LS and SS settings with $\mu = -2.5$, respectively. Most methods including JEL_{EQ} have coverage higher than the nominal level 0.95 and some methods such as BT, BJ, Wald_{ML} and $\text{Wald}_{\text{REML}}$ even have have a (virtually) 100% coverage. By contrast, SJ and SJ_{HO} have zero coverage, due to the fact that they always produce positive intervals. Similar patterns can be found in Table 2.3 when event rates go lower ($\mu = -5$). Except for SJ and SJ_{HO}, the coverage for every method is always high and gets even closer to 100% when the event rates or sample sizes become smaller. Also, the coverage appears not to change drastically with different K values.

	Large Sample					Small Sample				
K	JEL_{EQ}	QP	MQP	QP_{UT}	BT	JEL_{EQ}	QP	MQP	QP_{UT}	BT
20	0.985	0.989	0.989	0.997	1	1	1	1	1	1
50	0.983	0.995	0.995	0.999	1	1	1	1	1	1
80	0.990	0.999	0.999	1	1	1	1	1	1	1
K	BJ	J	AJ	SJ	SJ_{HO}	BJ	J	AJ	SJ	SJ_{HO}
20	1	0.993	0.997	0	0	1	1	1	0	0
50	1	0.999	0.999	0	0	1	1	1	0	0
80	1	1	1	0	0	1	1	1	0	0
K	PL_{ML}	PL_{REML}	Wald_{ML}	$\text{Wald}_{\text{REML}}$	BS	PL_{ML}	PL_{REML}	Wald_{ML}	$\text{Wald}_{\text{REML}}$	BS
20	0.996	0.993	1	1	0.995	0.999	0.998	1	1	1
50	0.997	0.992	1	1	0.997	0.999	0.999	1	1	1
80	0.994	0.993	1	1	0.998	1	1	1	1	1

Table 2.3. Empirical coverage probabilities of 95% CIs constructed using different methods for settings with $\tau^2 = 0$ and $\mu = -5$.

2.3.4. When between-study heterogeneity exists ($\tau^2 > 0$)

We compare the performance of the methods under different distributions of θ_i's. For each type of distribution, we report results for three cases: (i) $\mu = -2.5$ and LS; (ii) $\mu = -2.5$ and SS; (iii) $\mu = -5$ and LS. We omit the most difficult case: $\mu = -5$ and SS, where all methods do not perform adequately and the coverage of 95% CIs can be often below 50%.

For case (i), the results for all fifteen methods are reported. For cases (ii) and (iii), we only report results for eleven methods because the likelihood-based methods, PL_{ML}, PL_{REML}, Wald_{ML}, and $\text{Wald}_{\text{REML}}$, frequently fail due to the convergence issue. Also, the performance of QP and MQP is very similar and so their coverage curves overlap in many cases.

Results for treatment effects from heavy-tailed distributions: Figure 2.1 presents empirical coverage probabilities of different 95% CIs for settings with $\omega = 0.5$ and $\theta_i's$ generated from T_3 distributions with much heavier tails than normal distributions. In fact, T_3 has the heaviest tails among T distributions with finite variances. The rows of the figure correspond to cases (i), (ii) and (iii), from top to bottom, respectively; and the columns represent $K = 20$, 50 and 80, from left to right, respectively. We can observe that all methods have coverage lower than the nominal level 0.95, indicated by the horizontal line at the top. However, for K =50 and 80, JEL_{EQ} is a clear winner and provides higher coverage than all the other methods. Even for $K = 20$, it is among the top-performance group, though BT becomes the best in all settings of $K = 20$ except for cases (ii) and (iii) with small τ^2, where QP and MQP have the best performance instead. We note that BT can perform badly elsewhere (e.g., settings with $K = 80$ and large τ^2). It appears that the performance of JEL_{EQ} is not sensitive to the change of τ^2; however, a larger K would improve its coverage and lift its gain over the other methods. Similar patterns can be observed in Figures A1 and A2 in Appendix A, which present coverage results for settings with $\omega = 1$ and 0, respectively. It seems that the change of ω does not make a big difference for heavy-tailed distributions as long as they are symmetric.

Results for treatment effects from skewed distributions: We find that ω has an impact on the relative performance of the different methods when treatment effects are exponentially distributed. Figures 2.2—2.4 show results for $\omega = 1$, 0.5 and 0, respectively. As we can observe from Figures 2.2 and 2.3, JEL_{EQ} outperforms the other methods in nearly all settings with $K = 50$ and 80; for $K = 20$, QP and MQP offer the highest (or close to

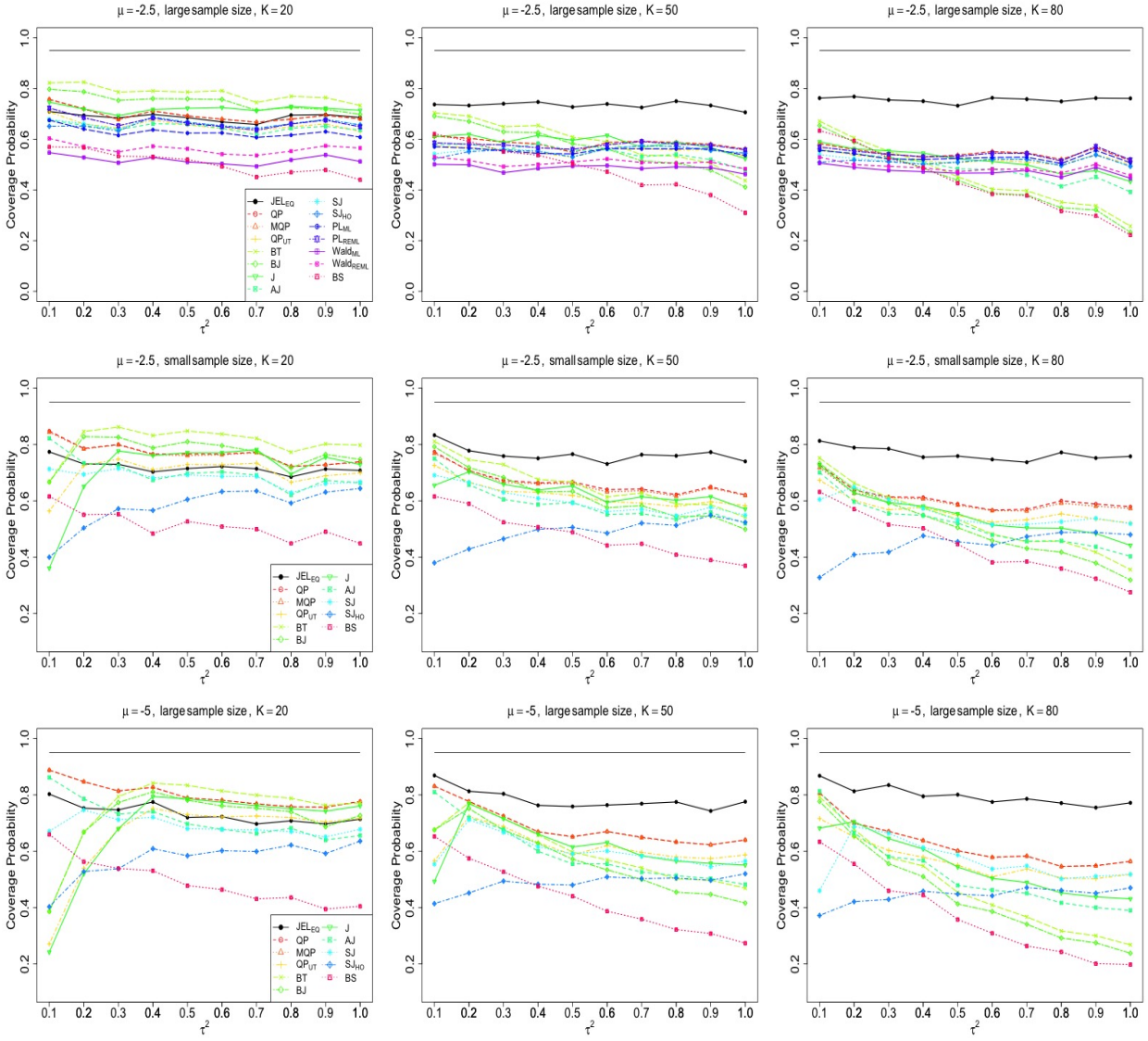


Figure 2.1. Empirical coverage probabilities of 95% CIs constructed using different methods for settings with $\omega = 0.5$ (i.e., equal variability in treatment and control groups) and effect sizes θ_i's from T_3 distributions.

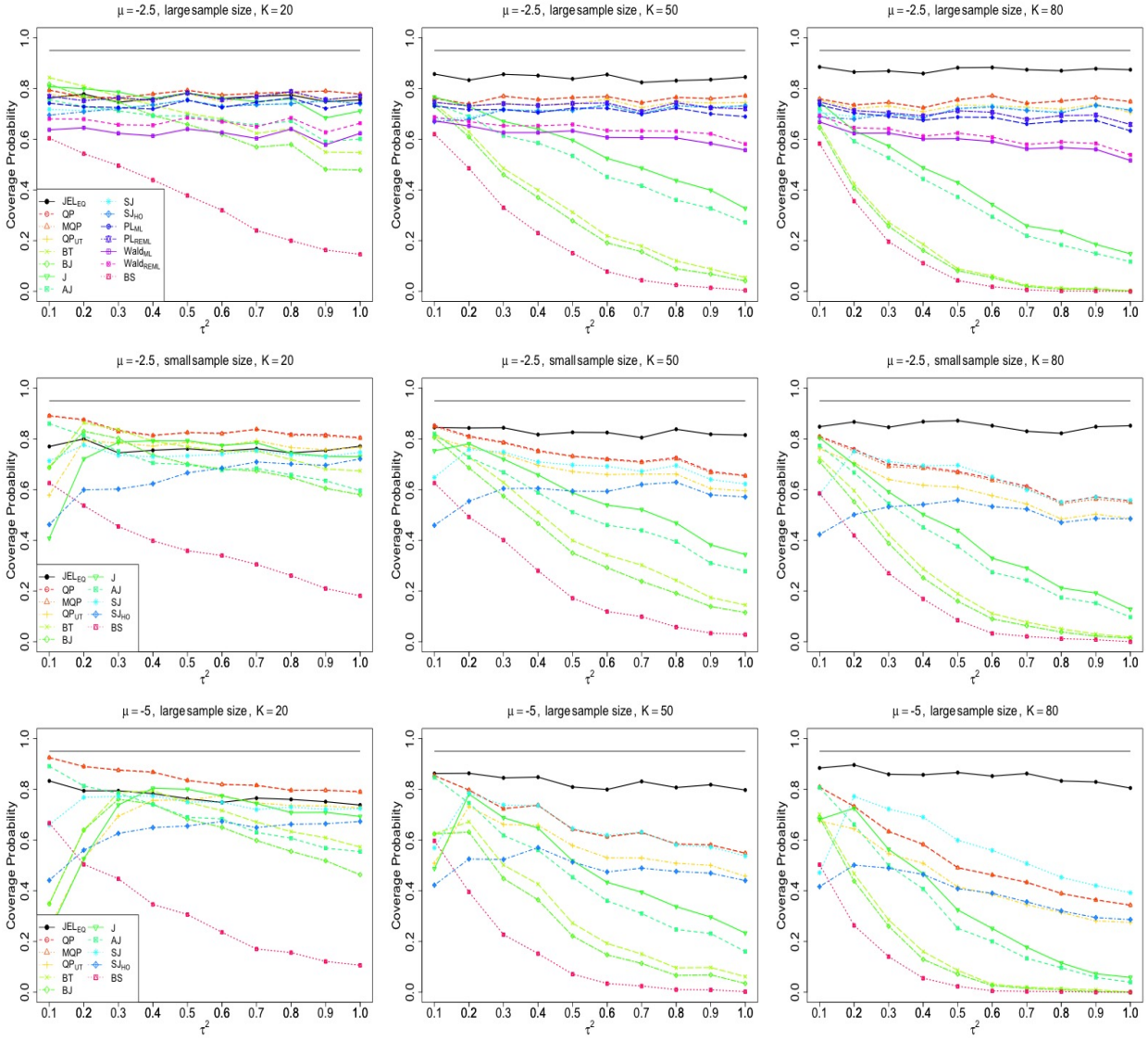


Figure 2.2. Empirical coverage probabilities of 95% CIs constructed using different methods for settings with $\omega = 1$ (i.e., smaller variability in the treatment than in the control) and effect sizes θ_i's from exponential distributions.

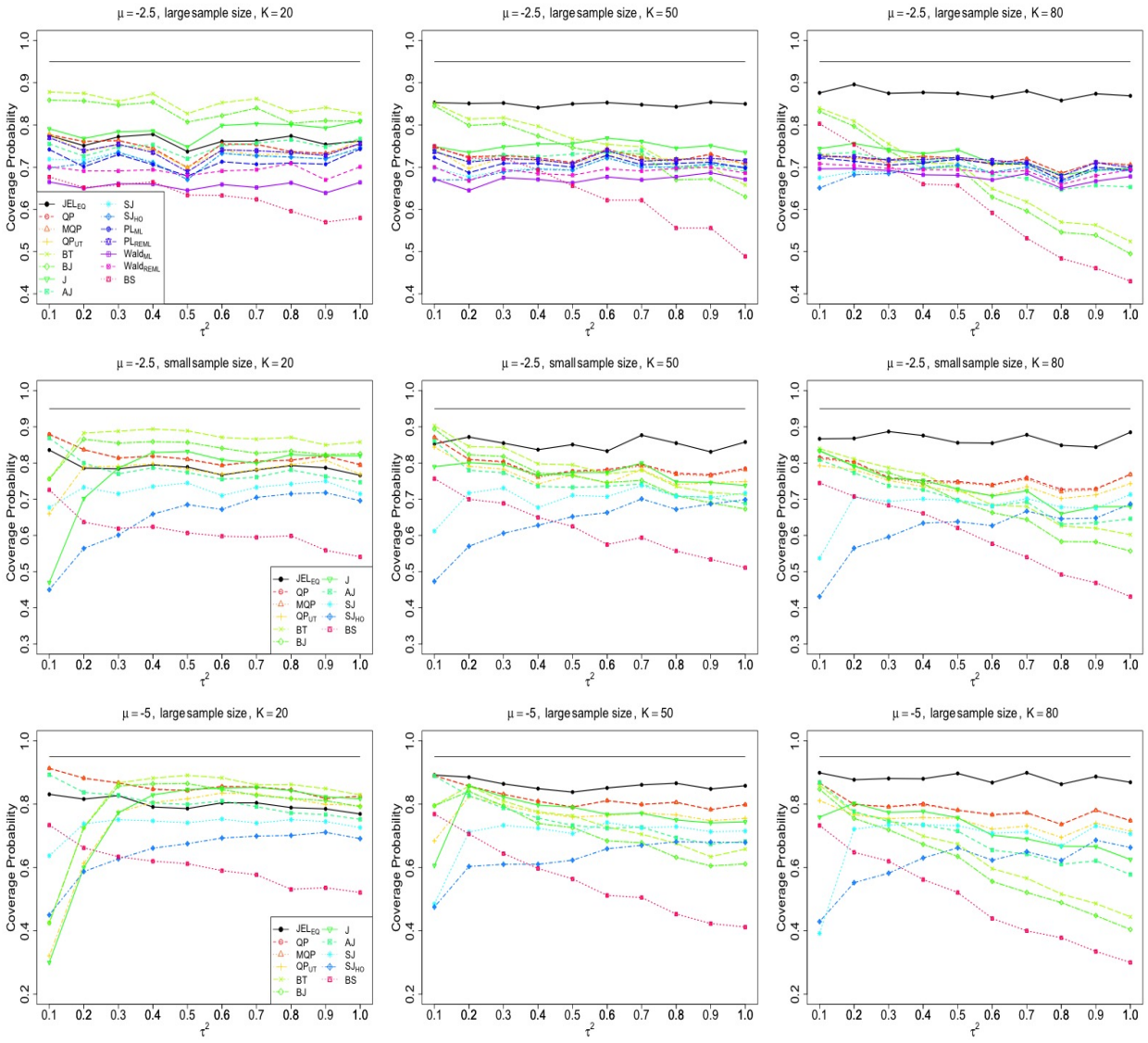


Figure 2.3. Empirical coverage probabilities of 95% CIs constructed using different methods for settings with $\omega = 0.5$ (i.e., equal variability in treatment and control groups) and effect sizes θ_i's from exponential distributions.

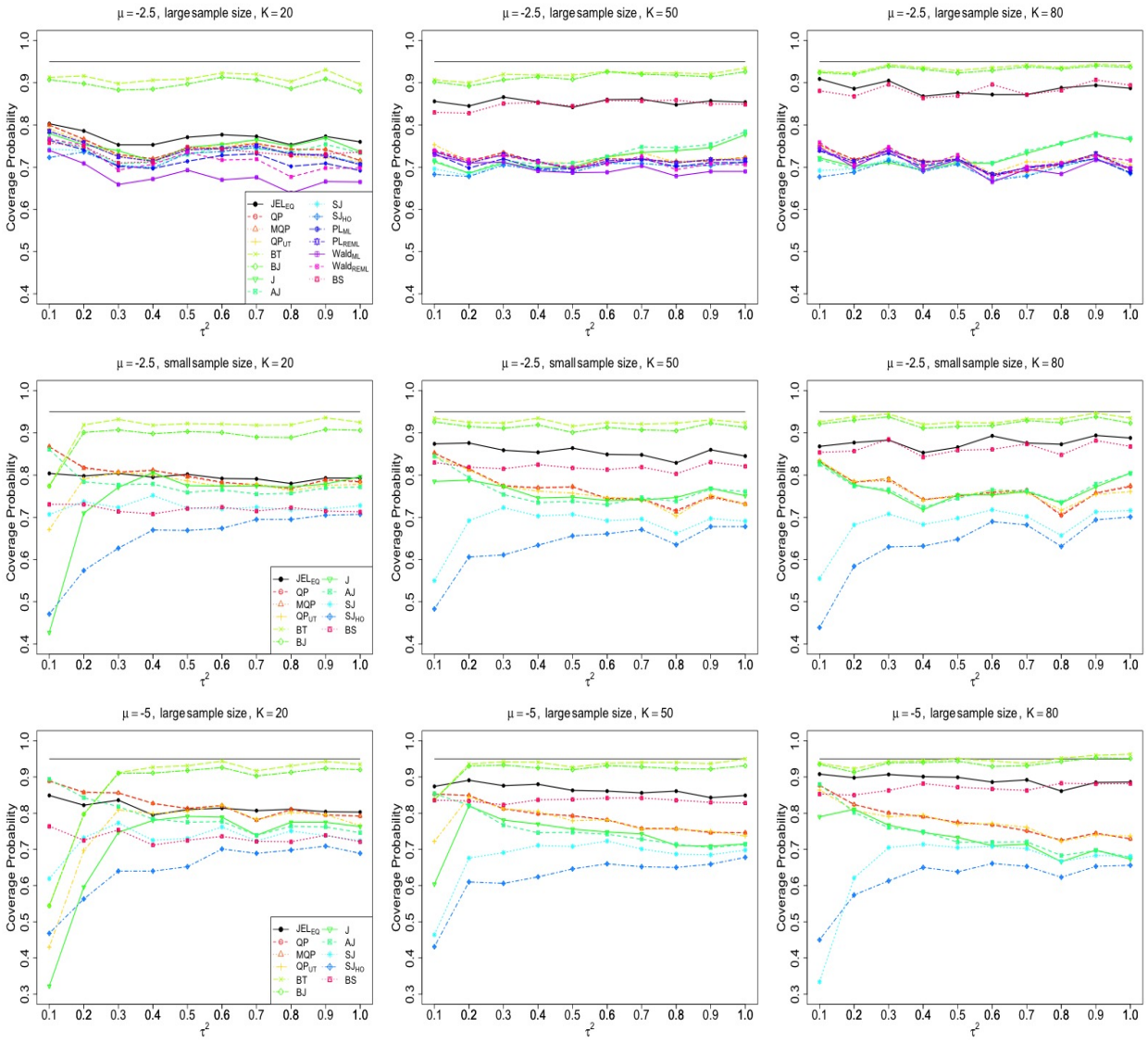


Figure 2.4. Empirical coverage probabilities of 95% CIs constructed using different methods for settings with $\omega = 0$ (i.e., larger variability in the treatment than in the control) and effect sizes θ_i's from exponential distributions.

highest) coverage when $\omega = 1$ while BT is often the best when $\omega = 0.5$. Figure 2.4 shows that for $\omega = 0$ (i.e. larger variability in the treatment than in the control), JEL_{EQ} is still among the top-performance group, but usually beaten by BT and BJ. Also, as in previous cases under heavy-tailed distributions, the performance of JEL_{EQ} seems to be quite robust to the change of τ^2 and ω (while many other methods are not), and the larger K is, the better it performs.

Figure 2.5 compares widths of the different CIs for settings with $K = 50, 80$ and $\omega = 0.5$, in which JEL_{EQ} has the best coverage; Figure 2.6 compares the widths for settings with $K = 50, 80$ and $\omega = 0$, in which BT and BJ have the best coverage instead. In each of the figures, the first (second) row corresponds to $K = 50$ (80), respectively. Evidently, the CI width of each method increases (roughly linearly) as τ^2 gets larger, meanwhile the variability between the methods also becomes larger. More importantly, we observe from the two figures that the method offering the highest coverage also has the largest width. We note that when CIs have quite different coverage probabilities and widths, it is actually a choice between methods with low bias and high variability and methods with high bias and low variability. In practice, people generally prefer the method with the least bias. Narrower intervals are only useful when they are able to provide adequate coverage. Thus, for different types of CIs, we should focus on the comparison on their coverage probabilities. Only for those that can offer comparable (adequate) coverage, we may further compare their widths and the method offering the shortest intervals would win.

Results for normal treatment effects: We proceed to examine the performance of JEL_{EQ} when the normality assumption holds. Here, we do not expect that JEL_{EQ}, as a nonparametric method, outperforms the other methods that utilize the normality to construct CIs. Figure 2.7 shows coverage probabilities of different CIs for settings with $\omega = 0.5$ and $\theta_i's$ generated from normal distributions. QP and MQP perform well, whose coverage is always close to the nominal level 0.95. Other methods like BT, BJ, J, SJ and SJ_{HO} have high

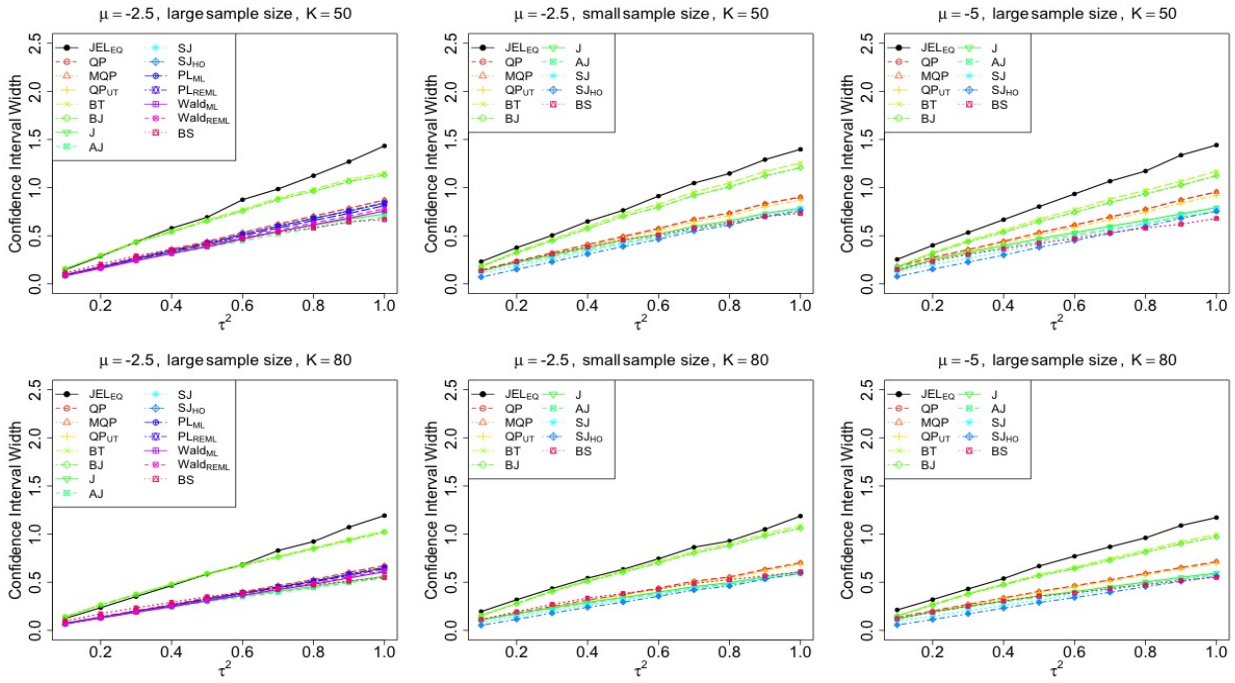


Figure 2.5. Width curves of 95% CIs constructed using different methods for settings with $\omega = 0.5$ (i.e., equal variability in treatment and control groups) and effect sizes θ_i's from exponential distributions.

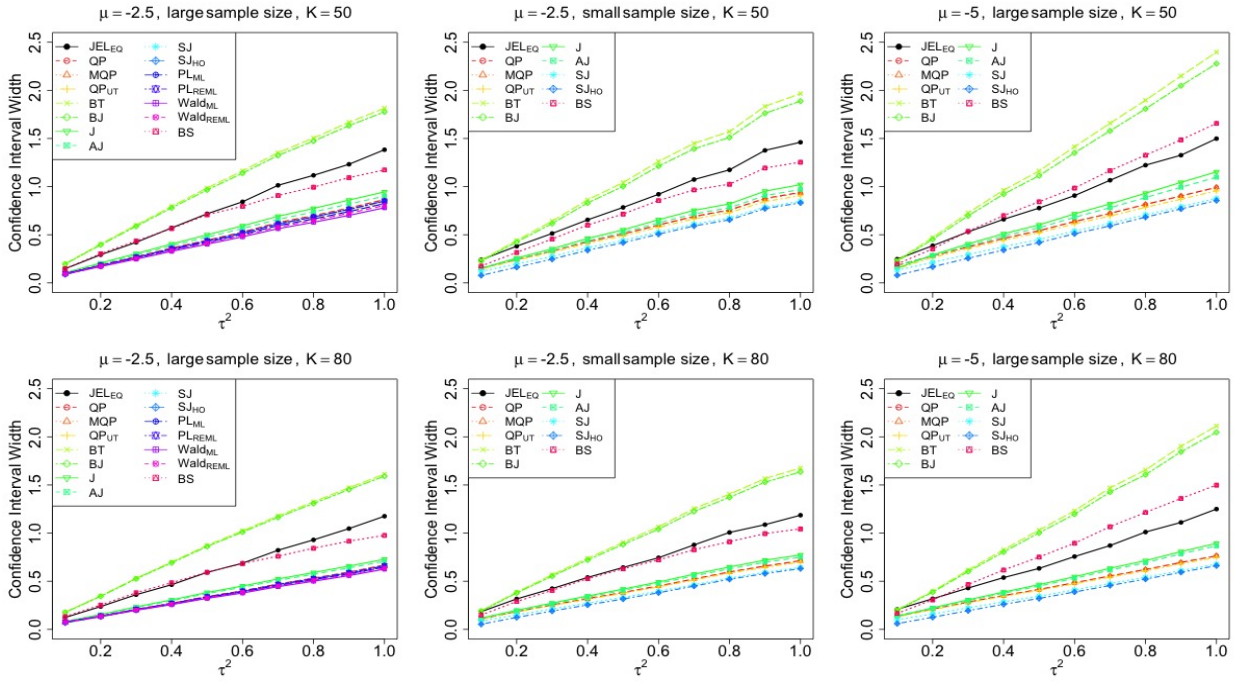


Figure 2.6. Width curves of 95% CIs constructed using different methods for settings with $\omega = 0$ (i.e., larger variability in the treatment than in the control) and effect sizes θ_i's from exponential distributions.

coverage as well except that their performance dips in cases (ii) and (iii) with small τ^2. BS, as another nonparametric method, has the lowest coverage in most of the cases. In contrast, JEL_{EQ} works reasonably well though it does not make any use of the normality assumption – it has decent coverage probabilities (around 0.90) for $K = 20$, and has competitive performance (mostly 0.92-0.93) for K =50 and 80. Again, Figure 2.7 shows that the performance of JEL_{EQ} is quite stable as τ^2 changes. Figures A3 and A4 in Appendix A further report results for settings with $\omega = 1$ and $\omega = 0$, which suggest that ω does not have a significant effect on the coverage, due to the symmetry of normal distributions.

2.3.5. Summary

For normal treatment effects, most methods can provide high coverage. However, when the treatment effects are non-normal, none of the CIs reaches the nominal level, and the difference in coverage can be substantial. The performance of likelihood-based methods (PL_{ML}, PL_{REML}, Wald_{ML}, $\text{Wald}_{\text{REML}}$) is generally not good, perhaps because they are more sensitive to the violation of the normality assumption. Another issue for the likelihood-based CIs is the computational cost. It takes much longer time to construct these CIs, compared with other methods, and sometimes they do not converge. Further, the performance of some methods is sensitive to the value of ω for skewed distributions. For example, BT and BJ work well when $\omega = 0$ but may work poorly when $\omega = 0.5$ or 1 (e.g., settings with large τ^2 and K values). In contrast, JEL_{EQ} has relatively steady performance across different ω values. Further, as the number of studies K increases, JEL_{EQ} has generally better coverage, while some other methods such as QP, MQP, J and AJ tend to have worse coverage.

Among all, the winner can be any of JEL_{EQ}, BT, QP, and MQP, depending on the type of distribution, the values of ω and τ^2, sample size, and event rate. JEL_{EQ} is clearly the best in many non-normal settings with large K. Overall, the performance of JEL_{EQ} seems to be robust to different values of μ, ω, τ^2 and the type of distribution. Thus, even when it is not

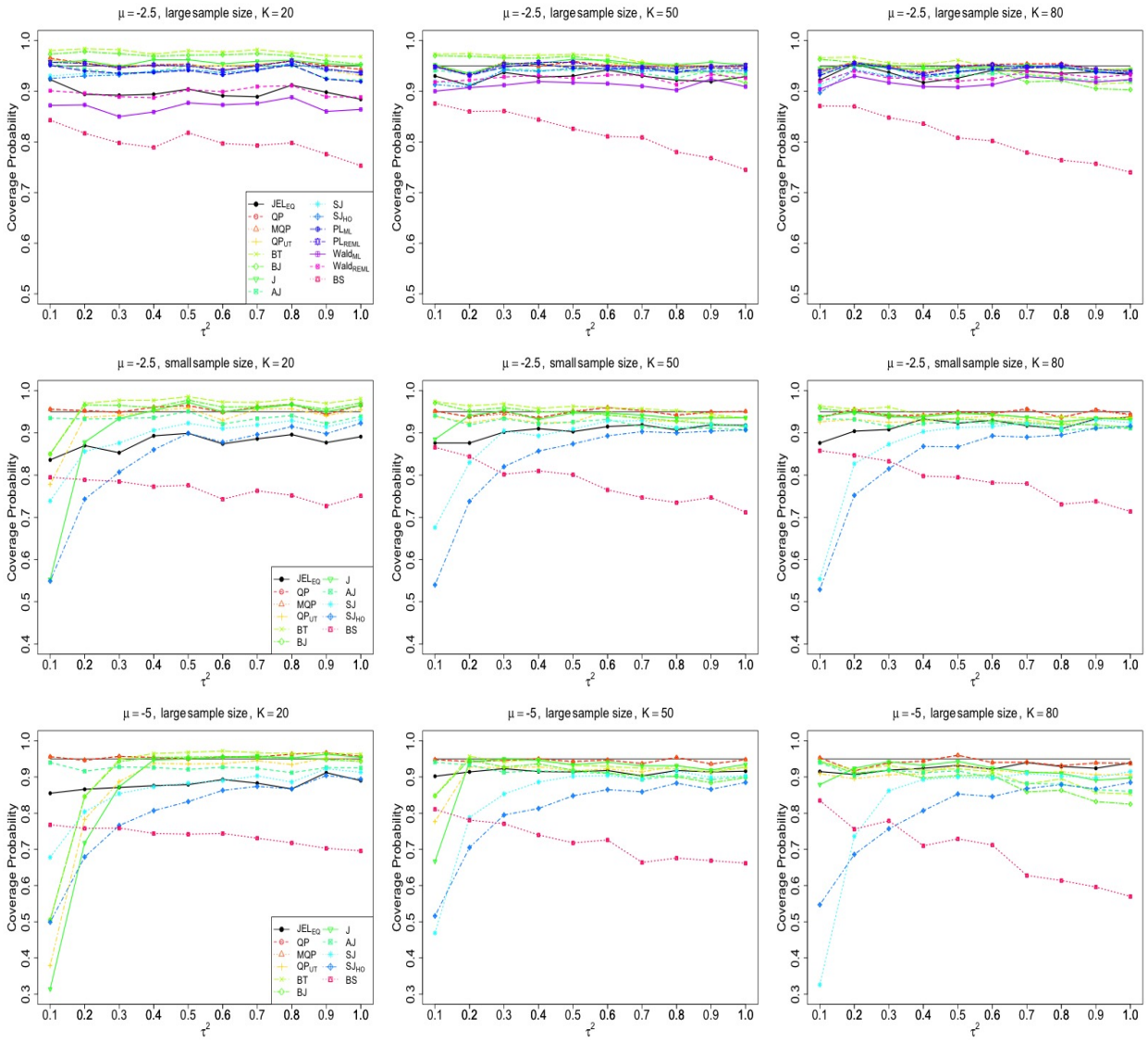
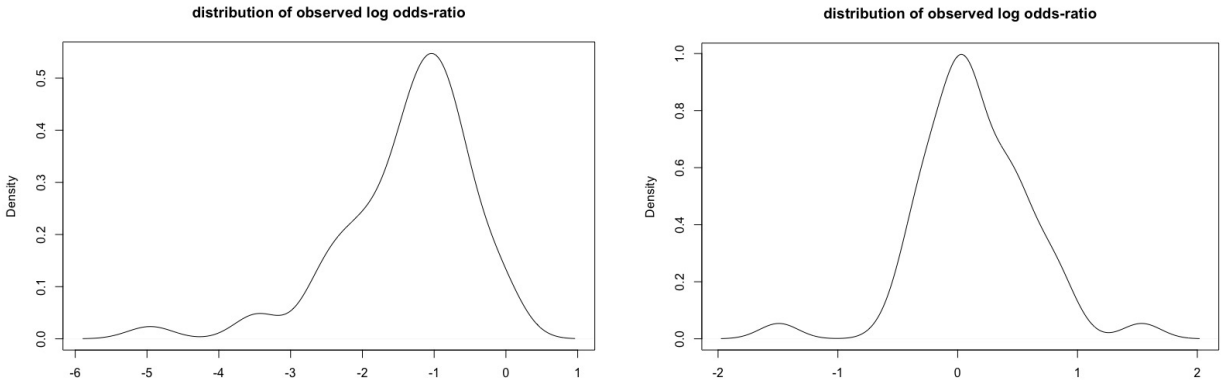


Figure 2.7. Empirical coverage probabilities of 95% CIs constructed using different methods for settings with $\omega = 0.5$ (i.e., equal variability in treatment and control groups) and effect sizes θ_i's from normal distributions.

the best, JEL_{EQ} is still in the top-performance group. Unlike JEL_{EQ}, the other three (BT, QP and MQP) can perform poorly in some settings.

2.4. Data Examples

2.4.1. Handedness and Eye-dominance



(a) (b)

Figure 2.8. Density plots of observed effect sizes (measured by LOR) from (a) handedness and eye-dominance data; (b) GSTP1 and lung cancer data.

People have preferences for use of a hand, called the dominant hand, to do major activities. Such preferences are also known as handedness. Most people prefer to use their right hands. According to Hardyck and Petrinovich [26], only one in ten people is left-handed. Likewise, people also have their dominant eyes and about one third people are left-eyed [48]. Bourassa et al. [9] conducted a meta-analysis to investigate the association between handedness and eye-dominance. The meta-analysis includes 54,087 subjects from 54 independent studies, each summarized by a 2×2 table recording counts of being "left-handed, left-eyed", "left-handed, right-eyed", "right-handed, left-eyed", and "right-handed, right-eyed". We divide the subjects into two groups based on their dominant eyes, and consider "left-handed"

as the event of interest (see Table A1 in Appendix A). The overall event rate is 10.1% and there are small-sized studies (e.g., 10+ studies have sample sizes smaller than 100). Figure 2.8(a) presents the density of observed LORs from the individual studies, which clearly has a long left tail, indicating a left-skewed distribution. Furthermore, the Shapiro-Wilk test yields a p-value smaller than 0.001, showing very strong statistical evidence for violation of the normality in this dataset. The HO and DL estimates of τ^2 are given by 0.531 and 0.303, respectively. Also, the estimated variances of the observed LORs in the treatment and control groups are 1.088 and 0.553, respectively, suggesting that the value of ω is close to zero. Thus, this meta-analysis is somewhat similar to our simulation settings of $K = 50$, $\mu = -2.5$, SS, $\omega = 1$ for right skewed distributions with the middle-range values of τ^2. Based on the fifth subplot in Figure 2.2, we may expect that JEL_{EQ} has the best coverage, followed by QP and MQP, while BT, BJ and BS are the worst three.

Table 2.4 shows the CIs of τ^2 obtained by the different methods. Besides SJ and SJ_{HO} that produce only positive intervals, all the other methods except for BT and BJ exclude zero in their CIs. Combined with our simulation results, this seems to indicate (i) BT and BJ perform poorly, and (ii) the between-study heterogeneity occur in the data. Compared to JEL_{EQ}, the other intervals are shorter; however, for settings as in this example, simulation suggests that they provide poorer coverage. In order to cover the true value of τ^2 with a high confidence, a longer interval may be needed and JEL_{EQ} does so automatically.

Method	JEL_{EQ}	QP	MQP	Q_{UT}	BT
CI	(0.203, 1.227)	(0.253, 0.816)	(0.250, 0.816)	(0.227, 0.765)	[0, 0.973)
Method	BJ	J	AJ	SJ	SJ_{HO}
CI	[0, 0.786)	(0.225, 0.749)	(0.209, 0.713)	(0.381, 0.823)	(0.332, 0.717)
Method	PL_{ML}	PL_{REML}	Wald_{ML}	$\text{Wald}_{\text{REML}}$	BS
CI	(0.122, 0.523)	(0.245, 0.730)	(0.252, 0.753)	(0.192, 0.650)	(0.196, 0.670)

Table 2.4. Data example of handiness and eye-dominance: 95% CIs of the between-study heterogeneity τ^2 constructed using different methods.

2.4.2. GSTP1 Gene and Lung cancer

Feng et al. [18] conducted a meta-analysis to investigate the association between the glutathione S-transferase P1 (GSTP1) gene and lung cancer. The event of interest is the GG genotype of GSTP1 and the overall event rate is around 10.5%. Subjects are split into two groups based on whether they had lung cancer or not. The meta-analysis includes 50 studies and 44 of them have non-missing counts for the GSTP1/GG genotype. Table A2 in Appendix A shows actual data for these 44 studies, and again, there exist small-sized studies (e.g., the smallest sample size is 35). Figure 2.8(b) shows that the density of the observed LORs is roughly symmetric, but with heavy tails on both sides. The Shapiro-Wilk test confirms that the evidence for non-normality in this example is statistically significant at the significance level $\alpha = 0.05$ (p-value is 0.034). The HO and DL estimates of τ^2 are given by 0 and 0.006, respectively. Also, the estimated variances of the observed LORs in the treatment and control groups are 0.762 and 0.869, respectively, suggesting that the value of ω is perhaps about 0.5. Thus, this meta-analysis is similar to our simulation settings of $K = 50$, $\mu = -2.5$, SS, $\omega = 0.5$ for heavy-tailed distributions with τ^2 close or equal to zero. Based on the fifth subplot in Figure 2.1, we may expect that again, JEL_{EQ} has the best coverage, followed by BT.

Table 2.5 shows the CIs of τ^2 obtained by the different methods. Except for the positive intervals SJ and SJ_{HO}, all other intervals include zero, among which QP, MQP, BJ, BT, and J even produce [0, 0]. Overall, the results indicate that that heterogeneity might not exist among these lung cancer studies.

Method	JEL_{EQ}	QP	MQP	Q_{UT}	BT
CI	[0 0.075)	[0, 0]	[0, 0]	[0, 0]	[0, 0]
Method	BJ	J	AJ	SJ	SJ_{HO}
CI	[0, 0]	[0,0]	[0, 0.076)	(0.006, 0.141)	(0.006, 0.015)
Method	PL_{ML}	PL_{REML}	Wald_{ML}	$\text{Wald}_{\text{REML}}$	BS
CI	[0, 0.037)	[0, 0.053)	[0, 0.060)	[0, 0.029)	[0, 0.035)

Table 2.5. Data example of GSTP1 and lung cancer: 95% CIs of the between-study heterogeneity τ^2 constructed using different methods.

2.5. Discussion

We propose to use jackknife empirical likelihood, a nonparametric approach, to construct CIs for the between-study heterogeneity parameter τ^2 in a relaxed random-effects model that lifts the normality assumptions for meta-analysis. Here, to obtain jackknife pseudo-values, we use an unbiased estimator of τ^2 based on the method of moments and consider two commonly used weighing schemes (i.e., equal and inverse variance weights); we show that with each scheme, the resulting log empirical likelihood ratio follows a (scaled) χ^2_1 distribution asymptotically. We further construct CIs, namely, JEL_{EQ} and JEL_{IV} (based on equal and inverse variance weights, respectively) by inverting the likelihood ratio test according to this asymptotic distribution. Our simulation shows that JEL_{EQ} is consistently better than JEL_{IV}; also, it often has better performance over existing methods when effect sizes follow non-normal distributions and the number of studies K is large. When K is relatively small, there is no uniform winner but JEL_{EQ} is always one of the top performers. When the normality is satisfied, JEL_{EQ} still has reasonable performance even though the other methods may perform better, due to their utilization of the normality when constructing the CIs.

As mentioned above, we have considered two different weighing schemes when computing jackknife pseudo-values. Dersimonian and Kacker [14] discussed other weight options, which

could be potentially considered as alternatives in the future work. Most methods for computing CIs of τ^2 assume that the within-study variances σ_i^2's are known, including our JEL methods. However, uncertainty arises with replacing σ_i^2 by its estimate s_i^2, which should be accounted for especially when component studies are small-sized. One should further consider the correlation between s_i^2 and Y_i when making inference when possible [33].

We have focused on meta-analysis of rare binary events in our numerical evaluation, due to its practical importance and wide range of applications where the normality assumption may be most vulnerable. However, the proposed JEL approach is general-purposed and is not restricted to (rare) binary events data. How JEL would perform in other scenarios (e.g. continuous outcomes with an interest on the mean difference) would be an interesting question. Furthermore, Zhang et al. [63] reported that none of existing CI methods work well when events are very rare (less than 1%) and studies have small sample sizes under the normality. It was further verified by our simulation that all methods including JEL_{EQ} fail to work adequately when the normality assumption is violated. Thus, the case of very rare events coupled with small-sized individual studies would be another direction that requires future research work, where Bayesian approaches can play an important role by incorporating prior knowledge (such as the information about event rates).

APPENDIX A

APPENDIX of CHAPTER 2

A.1. Technical detail

A.1.1. Proof of Theorem 1

To prove Theorem 1, we first introduce notations and establish Lemmas 3 and 4. We consider the *Re* model without the normality assumptions, and so we have model (2.1) with $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_i^2$, $E(\delta_i) = 0$ and $Var(\delta_i) = \tau^2$. Here, σ_i^2's are assumed to be known and can be replaced by their estimates s_i^2's (i.e., s_i^2 and σ_i^2 are exchangeable in our proof).

For the JEL method with equal weights $w_i = 1/K$, we define $\delta_i + \epsilon_i \equiv \mu_i$, $\sum_i \mu_i^2 \triangleq a$ and, $\sum_i \mu_i \triangleq b$. So the Q-statistic in (2.2) is given by

$$Q_{EQ} = \frac{\sum_i (Y_i - \bar{Y})^2}{K} = \frac{\sum_i (\mu_i - \frac{\sum_i \mu_i}{K})^2}{K} = \frac{a}{K} - \frac{b^2}{K^2}.$$

Lemma 3. *Under assumptions in Theorem 1, as $K \to \infty$, we have*

$$K \left\{ \frac{1}{K} \sum_i \hat{V}_i - \tau^2 \right\} \xrightarrow{d} N(0, \sum_i S_i^2)$$

where $S_i^2 \equiv Var(\mu_i^2)$.

Proof. It is easy to show that $E(a) = Var(b) = \sum_i \sigma_i^2 + K\tau^2$, $E(b) = 0$, and $Var(a) = \sum_i S_i^2$. By Lyapunov's CLT, we have

$$\sqrt{K}\frac{b}{K} = \sqrt{K}\frac{\sum_i \mu_i}{K} \xrightarrow{d} N(0, v), \text{ as } K \to \infty \tag{A.1}$$

where $v = (\sum_i \sigma_i^2 + K\tau^2)/K$, yielding $\frac{b^2}{Kv} \xrightarrow{d} \chi_1^2$, thus $E(b^2) = Kv = \sum_i \sigma_i^2 + K\tau^2$, and $Var(b^2) = 2(\sum_i \sigma_i^2 + K\tau^2)^2$. As a result, we have

$$E(Q_{EQ}) = \frac{\sum_i \sigma_i^2 + K\tau^2}{K} - \frac{\sum_i \sigma_i^2 + K\tau^2}{K^2} = \frac{(K-1)(\sum_i \sigma_i^2 + K\tau^2)}{K^2}.$$

In addition, as $\mu_i \perp \mu_j$ for $i \neq j$, we have

$$Cov(a, b^2) = Cov(\sum_i \mu_i^2, (\sum_i \mu_i)^2) = Cov(\sum_i \mu_i^2, \sum_i \mu_i^2) = Var(a),$$

which implies

$$\begin{aligned} Var(Q_{EQ}) =& Var\left(\frac{a}{K} - \frac{b^2}{K^2}\right) = Var\left(\frac{a}{K}\right) + Var\left(\frac{b^2}{K^2}\right) - 2Cov\left(\frac{a}{K}, \frac{b^2}{K^2}\right) \\ =& \frac{\sum_i S_i^2}{K^2} + \frac{(\sum_i \sigma_i^2 + K\tau^2)^2}{K^4} - \frac{2\sum_i S_i^2}{K^3} \\ =& \frac{\sum_i S_i^2}{K^2} + O(K^{-2}). \end{aligned}$$

Thus, Q can be approximated by a normal distribution

$$Q_{EQ} \sim N\left(\frac{(K-1)(\sum_i \sigma_i^2 + K\tau^2)}{K^2}, \frac{\sum_i S_i^2}{K^2}\right).$$

From (2.10), it is easy to see that

$$T_K = \frac{K \cdot Q_{EQ}}{K-1} - \frac{\sum_i s_i^2}{K},$$

which, combined with the expression for $E(Q_{EQ})$, leads to that T_k is an unbiased estimator of τ^2, where the estimates s_i^2 is treated the same as σ_i^2, as mentioned before. Some simple algebra yields

$$KT_K \sim N(K\tau^2, \sum_i S_i^2) \text{ as } K \to \infty.$$

From (2.8), the jackknife pseudo-value can be written as

$$\begin{aligned}
\hat{V}_i =& \frac{K}{K-1}a - \frac{b^2}{K-1} - \sum_i \sigma_i^2 - \left\{ \frac{K-1}{K-2}(a-\mu_i^2) - \frac{(b^2 - 2\mu_i b + \mu_i^2)}{K-2} - \sum_{j\neq i} \sigma_j^2 \right\} \\
=& \left(\frac{K}{K-1} - \frac{K-1}{K-2} \right) a + \frac{K-1}{K-2}\mu_i^2 - \left(\frac{1}{K-1} - \frac{1}{K-2} \right) b^2 - \frac{2\mu_i b}{K-2} + \frac{\mu_i^2}{K-2} - \sigma_i^2 \\
=& \frac{-(a-b^2)}{(K-1)(K-2)} + \frac{K}{K-2}\mu_i^2 - \frac{2\mu_i b}{K-2} - \sigma_i^2 \\
\triangleq& d_i - \sigma_i^2. \qquad \text{(A.2)}
\end{aligned}$$

Therefore, the average of pseudo-values is given by

$$\begin{aligned}
\frac{1}{K}\sum_i \hat{V}_i =& \frac{-(a-b^2)}{(K-1)(K-2)} + \frac{1}{K-2}a - \frac{2b^2}{K(K-2)} - \frac{1}{K}\sum_i \sigma_i^2 \\
=& \frac{1}{K-1}a - \frac{1}{K(K-1)}b^2 - \frac{1}{K}\sum_i \sigma_i^2 = T_k,
\end{aligned}$$

which completes the proof. □

Lemma 4. *Under assumptions in Theorem 1, as $K \to \infty$, we have*

$$\mathcal{S}_K \triangleq \frac{1}{K}\sum_i (\hat{V}_i - \tau^2)^2 \xrightarrow{p} \frac{1}{K}\sum_i S_i^2.$$

Proof. First note that

$$E(\hat{V}_i - \tau^2)^2 = E(\hat{V}_i^2) - 2E(\hat{V}_i)\tau^2 + \tau^4. \tag{A.3}$$

Using the fact that

$$\begin{aligned}
E(d_i) &= O(K^{-2}) + \frac{K}{K-2}E(\mu_i^2) + O(K^{-1}) = E(\mu_i^2) + O(K^{-1}) \\
E(\mu_i b) &= E(\mu_i^2 + \mu_i \sum_{j\neq i} \mu_j) = E(\mu_i^2) \\
E(\mu_i^3 b) &= E(\mu_i^4 + \mu_i^3 \sum_{j\neq i} \mu_j) = E(\mu_i^4) = (\sigma_i^2 + \tau^2)^2 + S_i^2,
\end{aligned}$$

the first term of (A.3) can be written as

$$\begin{aligned}
E\left(\hat{V}_i^2\right) &= \left(\frac{K}{K-2}\right)^2 E(\mu_i^4) + \sigma_i^4 - E(2\sigma_i^2 d_i) - \frac{E(4K\mu_i^3 b)}{(K-2)^2} + O(K^{-2}) \\
&= \left(\frac{K}{K-2}\right)^2 E(\mu_i^4) + \sigma_i^4 - 2\sigma_i^2 E(\mu_i^2) + O(K^{-1}).
\end{aligned}$$

Similarly, the second term is

$$-2E(\hat{V}_i)\tau^2 = -2\tau^2\{E(\mu_i^2) - \sigma_i^2\} + O(K^{-1}).$$

So as $K \to \infty$,

$$\begin{aligned} E(\hat{V}_i - \tau^2)^2 = &\left(\frac{K}{K-2}\right)^2 E(\mu_i^4) + \sigma_i^4 - 2\sigma_i^2 E(\mu_i^2) - 2\tau^2\{E(\mu_i^2) - \sigma_i^2\} + \tau^4 + O(K^{-1}) \\ \to &E(\mu_i^4) + \sigma_i^4 - 2\sigma_i^2 E(\mu_i^2) - 2\tau^2\{E(\mu_i^2) - \sigma_i^2\} + \tau^4 \\ = &(\sigma_i^2 + \tau^2)^2 + S_i^2 + \sigma_i^4 - 2\sigma_i^2(\sigma_i^2 + \tau^2) - 2\tau^2\{\tau^2\} + \tau^4 \\ = &S_i^2, \end{aligned}$$

which completes the proof. □

We now proceed to prove Theorem 1. From (2.12), we have

$$\begin{aligned} 0 = &\frac{1}{K}|\sum_i (\hat{V}_i - \tau^2) - \lambda \sum_i \frac{(\hat{V}_i - \tau^2)^2}{1 + \lambda(\hat{V}_i - \tau^2)}| \\ \geq &|\frac{\lambda}{K} \sum_i \frac{(\hat{V}_i - \tau^2)^2}{1 + \lambda(\hat{V}_i - \tau^2)}| \\ \geq &\frac{|\lambda| \mathcal{S}_K}{1 + |\lambda| W_K} - |\frac{1}{K} \sum_i (\hat{V}_i - \tau^2)|, \end{aligned} \tag{A.4}$$

where $W_K = \max_{1 \leq i \leq K} |\hat{V}_i - \tau^2|$. By Lemmas 3 and 4, one have $\mathcal{S}_K$ is $O_p(1)$ and the second term of (A.4) is $O_p(K^{-1/2})$. Also, since S_i^2 is finite, we have $W_K = o_p(K^{1/2})$. Therefore, one has $\lambda = O_p(K^{-1/2})$. Let $\gamma_i = \lambda(\hat{V}_i - \tau^2)$, then

$$\underset{1 \leq i \leq K}{max} |\gamma_i| = O_p(K^{-1/2}) o_p(K^{1/2}) = o_p(1).$$

Applying the Taylor expansion to (2.12), we have

$$\begin{aligned}
0 =& \frac{1}{K}\sum_i (\hat{V}_i - \tau^2)\{1 - \gamma_i + \frac{\gamma_i^2}{1+\gamma_i}\} \\
=& \frac{1}{K}\sum_i (\hat{V}_i - \tau^2) - \mathcal{S}_K\lambda + \frac{1}{K}\sum_i \frac{(\hat{V}_i - \sigma_0^2)\gamma_i^2}{1+\gamma_i} \\
=& \frac{1}{K}\sum_i (\hat{V}_i - \tau^2) - \mathcal{S}_K\lambda + O_p(K^{-1}).
\end{aligned}$$

This gives

$$\lambda = \mathcal{S}_K^{-1}\frac{1}{K}\sum_i (\hat{V}_i - \tau^2) + O_p(K^{-1}).$$

As a result, (2.11) can be written as

$$\begin{aligned}
-2\log\{R(\tau^2)\} =& 2\sum_i \log(1+\gamma_i) \\
=& 2\sum_i \gamma_i - \sum_i \gamma_i^2 + o_p(1) \\
=& 2K\lambda\sum_i (\hat{V}_i - \tau^2) - K\mathcal{S}_K\lambda^2 + o_p(1) \\
=& \frac{K(\frac{1}{K}\sum_i \hat{V}_i - \tau^2)^2}{\mathcal{S}_K} + o_p(1).
\end{aligned}$$

Combining the above equality with Lemmas 3 and 4, we have $-2\log\{R(\tau^2)\} \to \chi_1^2$ based on Slutsky's theorem.

A.1.2. Proof of Theorem 2

The proof of Theorem 2 under the inverse-variance weights can be shown similarly, thus we only outline the sketch of the proof. We denote the Q-statistic in (2.2) by Q_{IV} when it has inverse-variance weights $w_i = 1/s_i^2$, which can be obtained from (2.2):

$$Q_{IV} = \sum_i w_i \mu_i^2 - \frac{(\sum_i w_i \mu_i)^2}{\sum_i w_i} = a_0 - \frac{b_0^2}{C_0}.$$

Only for the proof, we assume that the sum of w_i is bounded, that is, there exists a constant C_2 such that $\sum w_i < C_2$. In fact, this assumption is very mild as we can always achieve it by a simple rescaling of w_i: $w_i^* = w_i / \sum w_i$, and using the new weights w_i^* will not affect our estimation. We use the following notations:

$$\sum_i w_i - (\sum_i w_i^2)/(\sum_i w_i) \triangleq A_0, \quad \sum_{j \neq i} w_i - (\sum_{j \neq i} w_i^2)/(\sum_{j \neq i} w_i) \triangleq A_i,$$
$$\sum_i w_i \mu_i^2 \triangleq a_0, \quad \sum_{j \neq i} w_j \mu_j^2 \triangleq a_i,$$
$$\sum_i w_i \mu_i \triangleq b_0, \quad \sum_{j \neq i} w_j \mu_j \triangleq b_i,$$
$$\sum_i w_i \triangleq C_0, \quad \sum_{j \neq i} w_j \triangleq C_{0i},$$
$$\sum_i \frac{1}{A_i} \triangleq C_1, \quad \sum_i \frac{1}{A_i^2} \triangleq C_2.$$

Next, we introduce Lemmas 5 and 6 that show properties of $\sum_i \hat{V}_i$ and $\mathcal{S}_K$. Note that from the condition $\epsilon < \max_i s_i^2 / \min_i s_i^2 < 1/\epsilon$ for some positive constant ϵ, we have $Var(a_0) = O_p(K^{-1})$. Thus, the variances of $\sum_i \hat{V}_i$ and $\mathcal{S}_K$ obtained from Lemmas 5 and 6 are of the same order as those obtained from Lemmas 3 and 4, respectively.

Lemma 5. *Under assumptions in Theorem 2, we have*

$$K \left\{ \frac{1}{K} \sum_i \hat{V}_i - \tau^2 \right\} \xrightarrow{d} N(0, B_1 Var(a_0))$$

as $K \to \infty$, *where*

$$B_1 = \left(\frac{K}{A_0}\right)^2 + \left(\frac{K-1}{K}\right)^2 \left(C_1 - \frac{1}{A_0}\right)^2 - \left(C_1 - \frac{1}{A_0}\right) \frac{2(K-1)}{A_0}.$$

Proof. From (2.14), we have

$$T_K = \frac{a_0 - b_0^2/C_0 - (K-1)}{A_0},$$

and the pseudo-values are given by

$$\begin{aligned}\hat{V}_i =& KT_K - (K-1)T_{K-1}^{(-i)} \\ =& \left(\frac{K}{A_0}\right)\left(a_0 - \frac{b_0^2}{C_0}\right) - \frac{K(K-1)}{A_0} - \frac{K-1}{A_i}\left(a_i - \frac{b_i^2}{C_{0i}}\right) + \frac{(K-1)(K-2)}{A_i}.\end{aligned}$$

The variance of the mean pseudo-value can be derived:

$$Var\left(\frac{1}{K}\sum_i \hat{V}_i\right) = \left(\frac{K}{A_0}\right)^2 Var(a_0) + \left(\frac{K-1}{K}\right)^2 Var\left(\sum_i \frac{a_i}{A_i}\right) - \frac{2(K-1)}{A_0} Cov\left(a_0, \frac{\sum_i a_i}{A_i}\right). \tag{A.5}$$

After some tedious but straightforward calculation, the variance and covariance from the second and third terms in (A.5) can be approximated by

$$\begin{aligned} Var\left(\sum_i \frac{a_i}{A_i}\right) &= \sum_i \left(C_1 - \frac{1}{A_0}\right)^2 w_i^2 s_i^2 \approx \left(C_1 - \frac{1}{A_0}\right)^2 Var(a_o), \\ Cov\left(\sum_i \frac{a_i}{A_i}, a_0\right) &= C_1 \cdot Var(a_0) - \sum_i \frac{w_i^2 s_i^2}{A_i} \approx \left(C_1 - \frac{1}{A_0}\right) Var(a_0), \end{aligned}$$

which completes the proof. □

Lemma 6. *Under assumptions in Theorem 2, as* $K \to \infty$, *we have*

$$\mathcal{S}_K \triangleq \frac{1}{K}\sum_i (\hat{V}_i - \tau^2)^2 \xrightarrow{p} B_2 Var(a_0),$$

where

$$B_2 = \left(\frac{K}{A_0}\right)^2 + \frac{(K-1)^2}{K}\left(C_2 - \frac{1}{A_0^2}\right) - \frac{2(K-1)}{A_0}\left(C_1 - \frac{1}{A_0}\right).$$

Proof. Simply noting

$$\begin{aligned}
E(\mathcal{S}_K) =& \frac{1}{K}\sum_i Var(\hat{V}_i) \\
=& \left(\frac{K}{A_0}\right)^2 Var(a_0) + \frac{(K-1)^2}{K}\left(\sum_i \frac{1}{A_i^2}\right) Var(a_0) - \frac{(K-1)^2}{K}\sum_i \frac{w_i^2 s_i^2}{A_i^2} \\
& - \frac{2(K-1)}{A_0}\left(\sum_i \frac{1}{A_0}\right) Var(a_0) + \frac{2(K-1)}{A_0}\sum_i \frac{w_i^2 s_i^2}{A_i^2} \\
\approx& \left\{\left(\frac{K}{A_0}\right)^2 + \frac{(K-1)^2}{K}\left(\sum_i \frac{1}{A_i^2}\right) - \frac{(K-1)^2}{KA_0} - \frac{2(K-1)}{A_0}\sum_i \frac{1}{A_i} + \frac{2(K-1)}{A_0^2}\right\} Var(a_0)
\end{aligned}$$

completes the proof.

We proceed to prove Theorem 2 and define $C \triangleq B_2/(KB_1)$. Note that using Lemmas 5 and 6, Theorem 2 can be proved similarly as in Section A.1.1.

First, using a similar argument, we can get $\lambda = O_p(K^{-1/2})$. As a result, we have:

$$-2\log\{R(\tau^2)\} = \frac{K(\frac{1}{K}\sum_i \hat{V}_i - \tau^2)^2}{\mathcal{S}_K} + o_p(1).$$

Thus, by Lemmas 5 and 6, we have $-2C\log\{R(\tau^2)\} \xrightarrow{d} \chi_1^2$ as $K \to \infty$. □

A.2. Additional simulation results

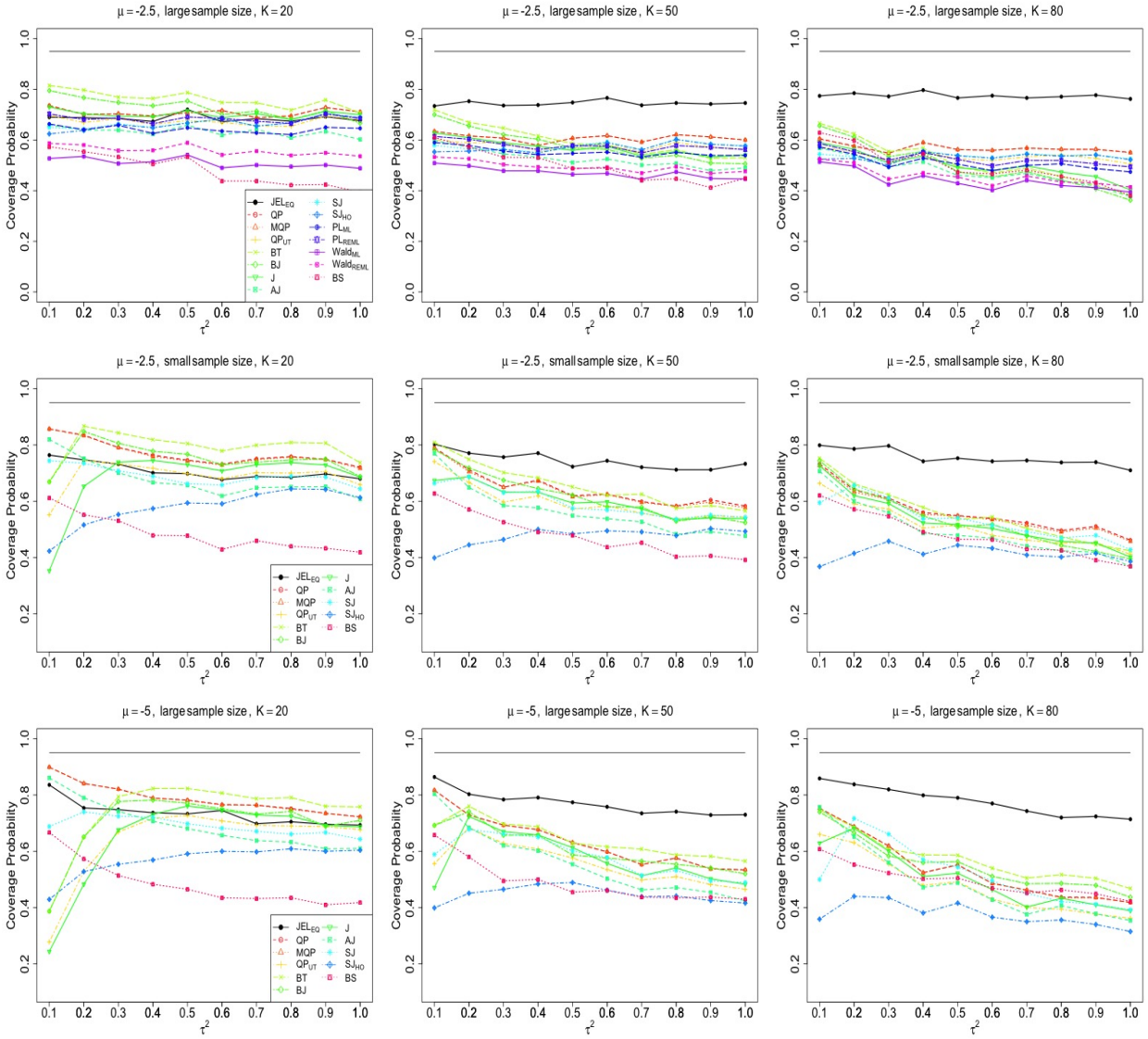


Figure A.1. Empirical coverage probabilities of 95% CIs of the between-study heterogeneity τ^2 constructed using different methods for settings with $\omega = 1$ (i.e., smaller variability in the treatment than in the control) and effect sizes θ_i's from T_3 distributions.

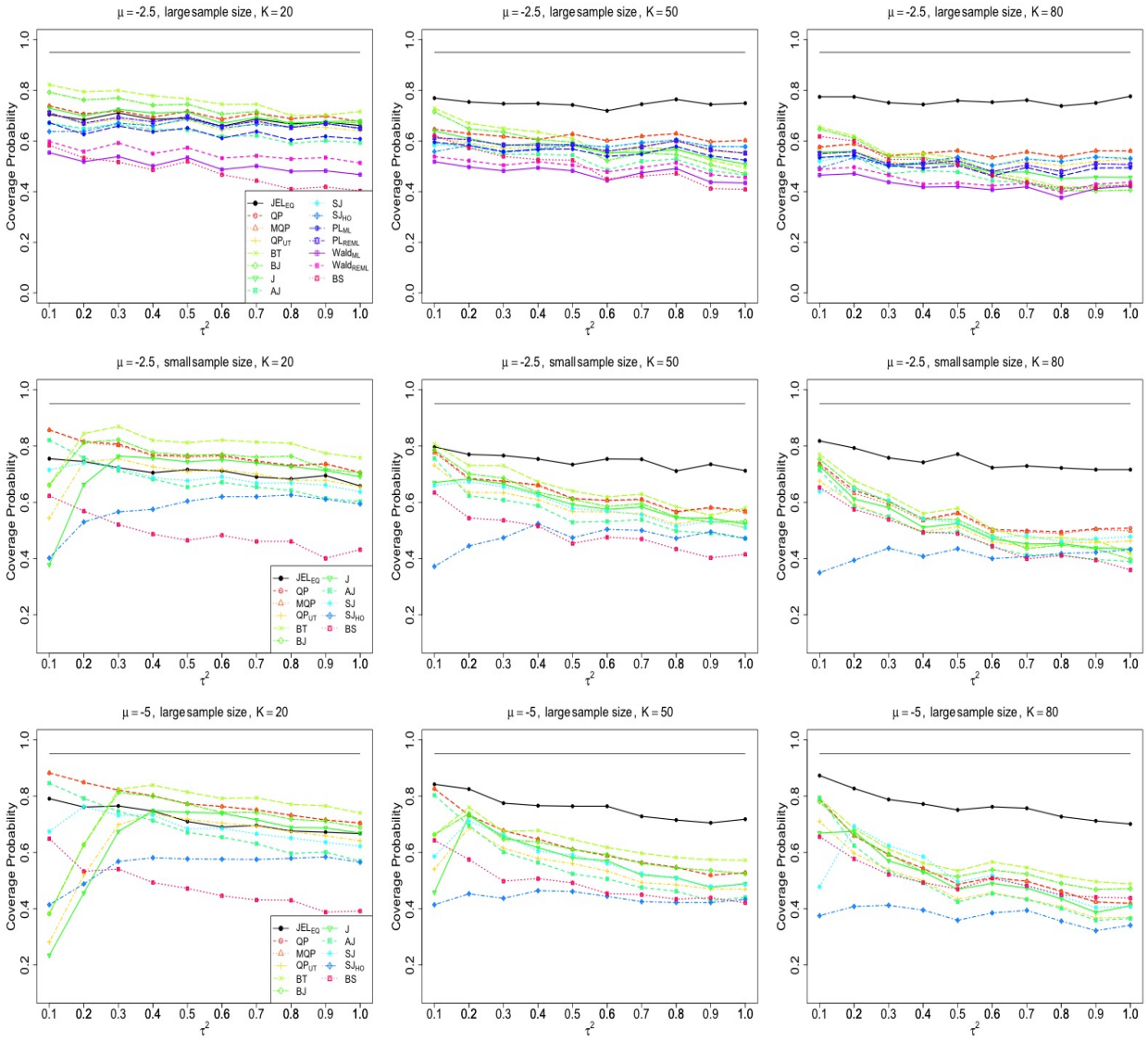


Figure A.2. Empirical coverage probabilities of 95% CIs of the between-study heterogeneity τ^2 constructed using different methods for settings with $\omega = 0$ (i.e., larger variability in the treatment than in the control) and effect sizes θ_i's from T_3 distributions.

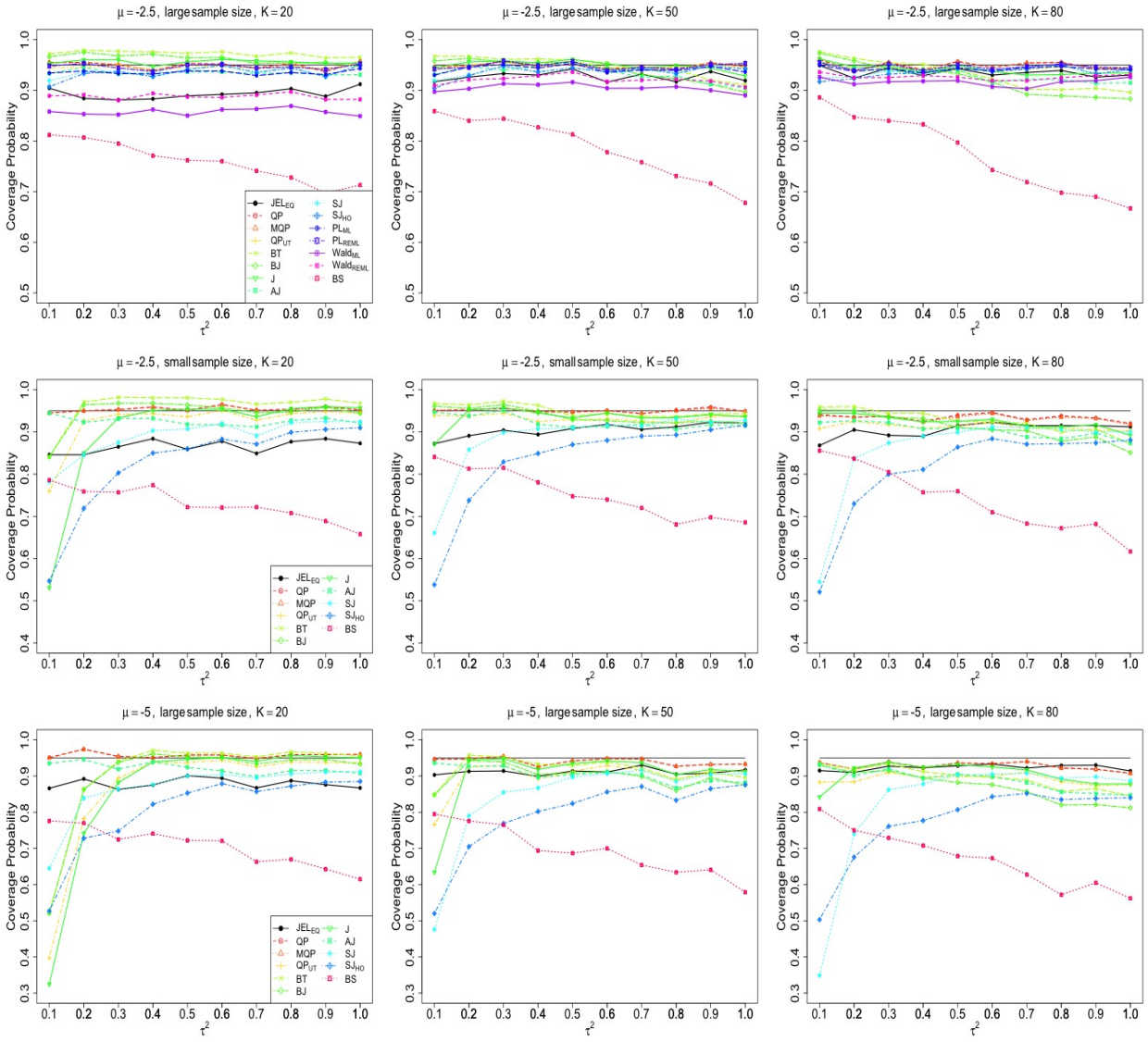


Figure A.3. Empirical coverage probabilities of 95% CIs of the between-study heterogeneity τ^2 constructed using different methods for settings with $\omega = 1$ (i.e., smaller variability in the treatment than in the control) and effect sizes θ_i's from normal distributions.

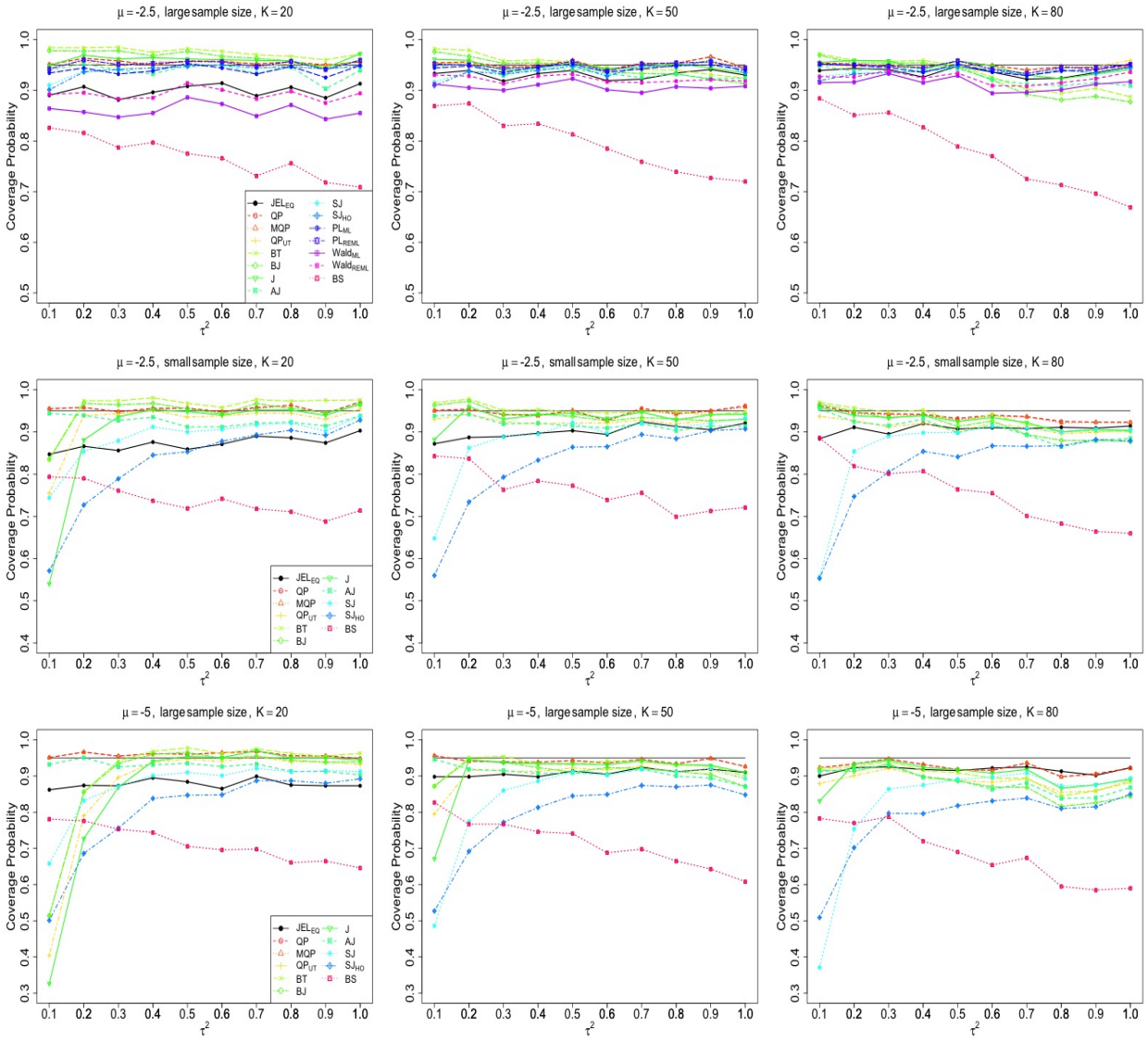


Figure A.4. Empirical coverage probabilities of 95% CIs of the between-study heterogeneity τ^2 constructed using different methods for settings with $\omega = 0$ (i.e., larger variability in the treatment than in the control) and effect sizes θ_i's from normal distributions.

A.3. Datasets

	Left-eye		Right-eye			Left-eye		Right-eye	
Study	# events	# subjects	# events	# subjects	Study	# events	# subjects	# events	# subjects
1	93	223	17	777	28	11	67	11	133
2	140	445	91	788	29	18	136	20	175
3	16	59	14	128	30	38	77	14	48
4	102	699	97	1995	31	19	35	6	39
5	17	55	14	94	32	19	191	12	364
6	10	62	2	172	33	411	1486	198	3661
7	7	22	2	171	34	9	60	8	131
8	4	19	2	40	35	26	68	34	124
9	2	26	3	42	36	4	47	2	67
10	3	21	4	43	37	18	49	19	94
11	4	33	3	92	38	31	302	8	551
12	2	22	1	50	39	11	191	21	374
13	3	22	2	47	40	37	227	27	287
14	27	128	0	261	41	467	2968	300	3764
15	9	68	11	109	42	16	84	15	348
16	5	28	2	40	43	10	42	2	86
17	20	157	13	340	44	563	3266	320	7247
18	6	20	8	42	45	5	19	1	38
19	8	39	5	61	46	19	48	6	138
20	241	1828	211	3651	47	46	94	32	203
21	19	46	10	43	48	89	232	70	454
22	2	35	2	86	49	25	41	53	121
23	8	37	2	58	50	141	515	54	1573
24	13	30	15	57	51	23	112	17	522
25	10	107	20	206	52	32	160	13	453
26	429	2957	311	4729	53	20	183	12	388
27	10	22	10	58	54	30	159	8	455

Table A.1. Data used for meta analysis of the relationship between handedness and eye-dominance [9]. Here, left-handedness is defined as an event.

	with lung cancer		without lung cancer			with lung cancer		without lung cancer	
Study	# events	# subjects	# events	# subjects	Study	# events	# subjects	# events	# subjects
1	22	138	27	297	23	6	29	4	29
2	26	178	18	199	24	110	1095	84	626
3	17	150	22	172	25	220	1921	141	1343
4	9	169	14	241	26	116	249	115	260
5	0	47	5	122	27	55	429	94	766
6	17	358	8	257	28	21	211	10	211
7	17	164	20	200	29	16	317	12	353
8	38	388	35	353	30	11	213	7	213
9	6	93	13	151	31	25	200	30	264
10	15	85	14	163	32	5	151	6	151
11	30	251	20	264	33	9	319	2	381
12	29	282	54	541	34	3	93	15	253
13	1	112	1	119	35	69	617	136	1257
14	35	362	44	419	36	5	89	9	108
15	71	229	65	197	37	19	462	6	379
16	62	446	70	622	38	7	100	12	125
17	31	235	39	233	39	13	118	22	290
18	0	12	0	23	40	97	788	92	788
19	22	228	38	288	41	23	142	26	190
20	13	89	19	119	42	33	198	27	233
21	4	227	3	227	43	5	270	5	270
22	15	112	18	151	44	9	150	4	152

Table A.2. Data used for meta analysis of the relationship between the GSTP1 gene and lung cancer [18]. Here, the GG genotype of GSTP1 is defined as an event.

BIBLIOGRAPHY

[1] Yueheng An and Yichuan Zhao. Jackknife empirical likelihood for the difference of two volumes under roc surfaces. *Annals of the Institute of Statistical Mathematics*, 70 (4):789–806, 2018.

[2] Dulal K Bhaumik, Anup Amatya, Sharon-Lise T Normand, Joel Greenhouse, Eloise Kaizar, Brian Neelon, and Robert D Gibbons. Meta-analysis of rare binary adverse event data. *Journal of the American Statistical Association*, 107(498):555–567, 2012.

[3] BJ Biggerstaff and RL Tweedie. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in medicine*, 16(7):753–768, 1997.

[4] Brad J Biggerstaff and Dan Jackson. The exact distribution of cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in medicine*, 27(29): 6093–6110, 2008.

[5] Linda Bolier, Merel Haverman, Gerben J Westerhof, Heleen Riper, Filip Smit, and Ernst Bohlmeijer. Positive psychology interventions: a meta-analysis of randomized controlled studies. *BMC public health*, 13(1):119, 2013.

[6] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111, 2010.

[7] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2011.

[8] Adrien Bougouin, Florian Boudin, and Béatrice Daille. *"TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction"*. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I13-1062.

[9] D Bourassa, Ian Mcmanus, and Philip Bryden. Handedness and eye-dominance: A meta-analysis of their relationship. *Laterality*, 1:5–34, 03 1996. doi: 10.1080/713754206.

[10] Sergey Brin and Lawrence Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine. Comput. Netw. ISDN Syst.*, 30(1–7):107–117, April 1998. ISSN 0169-7552. doi: 10.1016/S0169-7552(98)00110-X. URL https://doi.org/10.1016/S0169-7552(98)00110-X.

[11] Cornelia Caragea, Florin Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. *Citation-enhanced keyphrase extraction from research papers: A supervised approach*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, 2014.

[12] William G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10(1):101–129, 1954. doi: 10.2307/3001666.

[13] Jonathan J Deeks. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in medicine*, 21(11):1575–1600, 2002.

[14] Rebecca Dersimonian and Raghu Kacker. Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, 28(2):105–114, 2007. doi: 10.1016/j.cct.2006.04.004.

[15] Rebecca Dersimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986. doi: 10.1016/0197-2456(86)90046-2.

[16] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

[17] Charles Elkan and Keith Noto. *Learning classifiers from only positive and unlabeled data*. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[18] Xu Feng, Bao-Shi Zheng, Jun-Jie Shi, Jun Qian, Wei He, and Hua-Fu Zhou. Association of glutathione s-transferase p1 gene polymorphism with the susceptibility of lung cancer. *Molecular biology reports*, 39(12):10313–10323, 2012.

[19] Corina Florescu and Cornelia Caragea. *A Position-Biased PageRank Algorithm for Keyphrase Extraction*. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 4923–4924. AAAI Press, 2017.

[20] Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. *Domain-Specific Keyphrase Extraction*. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pages 668–673, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[21] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955. URL https://books.google.com/books?id=ZXL6AQAAQBAJ.

[22] Anshit Goyal, Mohamed Elminawy, Panagiotis Kerezoudis, M. Victor, Yagiz Yolcu, Mohammed Ali Alvi, and Mohamad Bydon. Impact of obesity on outcomes following lumbar spine surgery: A systematic review and meta-analysis. *Clinical Neurology and Neurosurgery*, 177, 12 2018. doi: 10.1016/j.clineuro.2018.12.012.

[23] Heikki Haario, Eero Saksman, and Johanna Tamminen. *Componentwise adaptation for high dimensional MCMC. Computational Statistics*, 20(2):265–273, 2005.

[24] Rebecca Hardy and Simon Giles Thompson. A likelihood approach to meta-analysis with random effects. *Statistics in medicine*, 15 6:619–629, 1996.

[25] Rebecca J Hardy and Simon G Thompson. Detecting and describing heterogeneity in meta-analysis. *Statistics in medicine*, 17(8):841–856, 1998.

[26] Curtis Hardyck and Lewis F Petrinovich. Left-handedness. *Psychological bulletin*, 84 (3):385, 1977.

[27] Larry V Hedges and Ingram Olkin. Statistical methods for meta-analysis. 1985.

[28] David C Hoaglin. Misunderstandings about q and 'cochran's q test' in meta-analysis. *Statistics in medicine*, 35(4):485–495, 2016.

[29] David C. Hoaglin. Shortcomings of an approximate confidence interval for moment-based estimators of the between-study variance in random-effects meta-analysis. *Research Synthesis Methods*, 7(4):459–461, 2016. doi: 10.1002/jrsm.1205.

[30] Anette Hulth. *Improved Automatic Keyword Extraction Given More Linguistic Knowledge*. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119355.1119383. URL https://doi.org/10.3115/1119355.1119383.

[31] Dan Jackson. Confidence intervals for the between-study variance in random effects meta-analysis using generalised cochran heterogeneity statistics. *Research Synthesis Methods*, 4(3):220–229, 2013.

[32] Dan Jackson and Jack Bowden. Confidence intervals for the between-study variance in random-effects meta-analysis using generalised heterogeneity statistics: should we use unequal tails? *BMC medical research methodology*, 16(1):118, 2016.

[33] Dan Jackson and Ian R. White. When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, 60(6):1040–1058, 2018. doi: 10.1002/bimj.201800071.

[34] Dan Jackson, Jack Bowden, and Rose Baker. Approximate confidence intervals for moment-based estimators of the between-study variance in random effects meta-analysis. *Research synthesis methods*, 6(4):372–382, 2015.

[35] Bing-Yi Jing, Junqing Yuan, and Wang Zhou. Jackknife empirical likelihood. *Journal of the American Statistical Association*, 104(487):1224–1232, 2009. doi: 10.1198/jasa.2009.tm08260. URL https://doi.org/10.1198/jasa.2009.tm08260.

[36] E. L. Kaplan and Paul Meier. *Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association*, 53(282):457–481, 1958.

[37] Guido Knapp, Brad Biggerstaff, and Joachim Hartung. Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical journal. Biometrische Zeitschrift*, 48:271–285, 04 2006. doi: 10.1002/bimj.200510175.

[38] Decong Li, Sujian Li, Wenjie Li, Wei Wang, and Weiguang Qu. *A Semi-Supervised Key Phrase Extraction Approach: Learning from Title Phrases through a Document Semantic Network*. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort'10, pages 296–300, USA, 2010. Association for Computational Linguistics.

[39] Lie Li and Xinlei Wang. Meta-analysis of rare binary events in treatment groups with unequal variability. *Statistical Methods in Medical Research*, 28:096228021772124, 07 2017. doi: 10.1177/0962280217721246.

[40] Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. *Clustering to Find Exemplar Terms for Keyphrase Extraction*. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP'09, pages 257–266, USA, 2009. Association for Computational Linguistics. ISBN 9781932432596.

[41] Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4): 719–748, 1959.

[42] Julian John McAuley and Jure Leskovec. *From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews*. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW'13, pages 897–908, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320351. doi: 10.1145/2488388.2488466. URL https://doi.org/10.1145/2488388.2488466.

[43] Rada Mihalcea and Paul Tarau. *Textrank: Bringing order into text*. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, 2004.

[44] Michael A. Newton, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. *Detecting differential gene expression with a semiparametric hierarchical mixture method*. *Biostatistics*, 5(2):155–176, 04 2004. ISSN 1465-4644. doi: 10.1093/biostatistics/5.2.155. URL https://doi.org/10.1093/biostatistics/5.2.155.

[45] Marc Orlitzky, Frank L Schmidt, and Sara L Rynes. Corporate social and financial performance: A meta-analysis. *Organization studies*, 24(3):403–441, 2003.

[46] Art B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988. doi: 10.1093/biomet/75.2.237.

[47] Art B. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18 (1):90–120, 03 1990. doi: 10.1214/aos/1176347494. URL https://doi.org/10.1214/aos/1176347494.

[48] Michael Reiss and Gilfe Reiss. Ocular dominance: some family data. *Laterality: Asymmetries of Body, Brain and Cognition*, 2(1):7–16, 1997.

[49] Faysal Satter and Yichuan Zhao. Jackknife empirical likelihood for the mean difference of two zero-inflated skewed populations. *Journal of Statistical Planning and Inference*, 2020.

[50] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[51] Kurex Sidik and Jeffrey N. Jonkman. Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54 (2):367–384, 2005. doi: 10.1111/j.1467-9876.2005.00489.x.

[52] Kurex Sidik and Jeffrey N. Jonkman. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in medicine*, 26 9:1964–81, 2007.

[53] Teresa C Smith, David J Spiegelhalter, and Andrew Thomas. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in medicine*, 14(24): 2685–2699, 1995.

[54] Theodore Stankowich and Daniel T Blumstein. Fear in animals: a meta-analysis and review of risk assessment. *Proceedings of the Royal Society B: Biological Sciences*, 272(1581):2627–2634, 2005.

[55] Peter D Turney. *Learning algorithms for keyphrase extraction. Information retrieval*, 2 (4):303–336, 2000.

[56] Wolfgang Viechtbauer. Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26(1):37–52, 2007. doi: 10.1002/sim.2514.

[57] S. D. Walter and R. J. Cook. A comparison of several point estimators of the odds ratio in a single 2 x 2 contingency table. *Biometrics*, 47(3):795–811, 1991. doi: 10.2307/2532640.

[58] Xiaojun Wan and Jianguo Xiao. *Single Document Keyphrase Extraction Using Neighborhood Knowledge*. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 855–860. AAAI Press, 2008. ISBN 9781577353683.

[59] Rui Wang, Wei Liu, and Chris McDonald. *Corpus-independent generic keyphrase extraction using word embedding vectors*. In *Software Engineering Research Conference*, volume 39, pages 1–8, 2014.

[60] Penny F Whiting, Robert F Wolff, Sohan Deshpande, Marcello Di Nisio, Steven Duffy, Adrian V Hernandez, J Christiaan Keurentjes, Shona Lang, Kate Misso, Steve Ryder, et al. Cannabinoids for medical use: a systematic review and meta-analysis. *Jama*, 313(24):2456–2473, 2015.

[61] Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. *Finding Advertising Keywords on Web Pages*. In *Proceedings of the 15th International Conference on World Wide Web*, WWW'06, pages 213–222, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933239. doi: 10.1145/1135777.1135813. URL https://doi.org/10.1145/1135777.1135813.

[62] Xue Yu and Yichuan Zhao. Empirical likelihood inference for semi-parametric transformation models with length-biased sampling. *Computational Statistics & Data Analysis*, 132:115–125, 2019.

[63] Chiyu Zhang, Min Chen, and Xinlei Wang. Statistical methods for quantifying between-study heterogeneity in meta-analysis with focus on rare binary events. *Statistics and Its Interface*, 13(4):449–464, 2020.

[64] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. *Learning with Local and Global Consistency*. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 321–328, Cambridge, MA, USA, 2003. MIT Press. URL http://dl.acm.org/citation.cfm?id=2981345.2981386.