Statistical Methods for Aggregation of Indirect Information

Professors Jun S. Liu **Simeng Han**

Statistical Methods for Aggregation of Indirect Information

Abstract

How to properly aggregate indirect information is more and more important. In this dissertation, we will present two aspects of the issue: indirect comparison of treatment effects and aggregation of ordered-based rank data.

In Chapter 1, we study causal inference via indirect comparison. Comparing the efficacy of different drugs is a very important problem in drug development and health care industries. Indirect comparison is an effective approach to avoid high costs of direct comparison via head-to-head trials. A major challenge in indirect comparison, however, is that the unit populations in different trials are often different. When this happens, we need to match the patient population across the two trials of interest. However, in practice, it's very often that only summary statistics, instead of original individual-level data, are available for some trials. For this challenging scenario, most classic matching methods fail. Methods based on weighting adjustment can still be applied, but have to be modified to fit the new challenges. In this dissertation, we will systematically study statistical issues related to casual inference via indirect comparison: assumptions under which the causal effect of interest is estimable, potential methods to estimate the causal effect, and relative efficiency of these methods.

In Chapter 2, we studied the problem of ranking aggregation, i.e., combining several base rankers to get an aggregated ranking function. Most methods in the literature assume that the base rankers of interest are equally reliable, however, it is

desirable to distinguish the high quality base rankers from the low quality ones and treat them differently in the analysis. Some methods achieve this end by assigning pre-given weights to base rankers. But there are no systematic and principled strategies for designing a proper weighting scheme for a practical problem. We proposed a Bayesian approach, called BARD, to overcome this limitation. BARD measures the reliability of the base rankers in a quantitative way, and makes use of this information to improve the aggregated ranker. Both simulation studies and real data applications show that BARD significantly outperforms existing methods when equality of base rankers varies greatly.

Contents

Acknowledgments

First, I would like to thank Professor Jun S. Liu, who has been always encouraging me and supportive during my years at Harvard. He is not only a great advisor who can provide valuable guidance on my academic research training, but also a wonderful mentor for my life. And I want to show my appreciation to Professor David Harrington and Professor Lee-Jen Wei for serving on my dissertation committee and provided valuable comments and suggestions. I also want to thank Dr. Ke Deng and Professor Kate J. Li for the great collaboration and discussion on the BARD project described in Chapter 2.

Staying at Harvard will be a precious piece of my memory because of all the members of the department. I would like to take this opportunity to thank all the faculty members Professor Donald Rubin, Carl Morris, Xiao-Li Meng, Samuel Kou, Joe Blitzstein, Tirthankar Dasgupta, Edoardo Airoldi, and Natesh Pillai. It is my great honor to learn from them and work with them, not only statistical knowledge but also the way of thinking and teaching. Support from staff in the department also played important role for my life at Harvard. Thank Betsey Cogswell, Dale Rinkel, Maureen Stanton, James Matejek, Steve Finch, and Alice Moses for their kind help. Another important piece of this memory is friendship. I would like to thank my friends, lab mates and fellow students in the department: Yang Chen, Daniel Fernandez, Valeria Espinosa, Jonathan Hennessy, Ming Hu, Bo Jiang, Joseph Kelly, Yang Li, Nathan Stein,Thomas Tong, Xufei Wang, Samuel Wong, Di Wu, Jiexing Wu, Xianchao Xie, and Xiaojin Xu.

Last but not least, I would like to thank my family. I want to thank my parents Jie Han and Linna Zhang for their great support. I also want to thank my husband

Ke Deng for the unreserved love.

Thank you everyone for being here with me on this journey.

To my parents.

Chapter 1

Statistical Methods for Indirect Comparison of Treatment Effects

1.1 Introduction

Comparing the efficacy of different drugs is a very important problem in drug development and health care industries. For example, before a new drug D_N comes into the market, evidences must be shown that its performance is better or at least comparable to well accepted drugs already in the market. A straightforward solution to this mission is to run a series of head-to-head trials, in each of which the new drug D_N is directly compared to one well accepted drug D_O by a randomized experiment. An obvious limitation of this strategy, however, is that the time and economic costs of running multiple head-to-head trials are often too expensive.

Indirect comparison is an effective approach to avoid high costs of direct comparison via head-to-head trials. For example, if all drugs in the market have been

compared to a common baseline drug (e.g., a standard placebo) for a common patient population, by running just one extra head-to-head trial over the common patient population to measure the performance of the new drug with respect to the common baseline drug, we can compare the new drug with all conventional drugs indirectly. It can be showed that under certain conditions, even if the baseline drugs used in different head-to-head trials are different, indirect comparison is still a proper strategy to estimate the relative efficacy among drugs.

A major challenge in indirect comparison, however, is that the unit populations in different trials are often different. When this happens, we need to match the patient population across the two trials of interest. The formulation of this across-trial population matching problem is very similar to that of the classic across-arm subpopulation matching problem widely encountered in non-randomized experiments. And, many popular methods designed for across-arm matching can be naturally extended to across-trial matching. However, since the data generating mechanism of the two-trial scenario is different from that of the one-trial case, at the conceptual level, there are subtle differences between the across-trial matching and the across-arm matching.

The formulation of indirect comparison can be further complicated when only summary statistics, instead of complete individual-level data, are available for trials in which the conventional drugs were compared to baseline drugs. This is a very common scenario in practical health care studies. Due to space constraints and/or commercial concerns, detailed individual-level data are often not provided in publications and technical reports about clinical trials. Although we usually have full control of the clinical trial for the new drug, and thus, can get access to complete individual-

level data, the incomplete observations of the old trials will pose great challenges in casual inference of indirect comparison. Due to the lack of information, most classic matching methods fail in this scenario; methods based on weighting adjustment can still be applied, but have to be modified to fit the new challenges.

In this Chapter, we will systematically study statistical issues related to casual inference via indirect comparison: assumptions under which the causal effect of interest is identifiable, potential methods to estimate the causal effect, and relative efficiency of these methods. The following of this chapter is organized as follows. First, we go over Rubin casual model and the basic elements of casual inference via a head-to-head trial in Section 1.2. In Section 1.3, we briefly review classic approaches for matching unit populations across the two arms of a head-to-head trial. The problem of indirect comparison is formally introduced in Section 1.4, its links to and differences from the direct comparison are discussed. Section 1.5 deals with a more complicated scenario of indirect comparison where individual level data are available for just one trial. Approaches to match unit population across the two trials in this challenging case are discussed, frequentist approaches to estimate casual effect based on the weighted samples are proposed. Considering that it may not be wise to match all available covariates in a practical problem, principles and methods to select covariates to be matched are given in Section 1.6. Simulation studies are presented in Section 1.7 to provide numerical evidences for comparing the performances of different approaches. Analysis of a real-like data example is illustrated in Section 1.8. Finally, we discuss and summarize our study in Section 1.9.

1.2 Elements of Causal Inference

1.2.1 Potential outcomes and causal effects

Potential outcomes. Following the framework of Rubin causal model (Rubin 1974), we use (Y_i^c, Y_i^t) to denote the potential outcomes of an *experiment unit* (or, simply a *unit*) under the *treatment* condition t and *control* condition c. In the drug comparison problem, for example, t is the new drug D_N, c is an old drug D_O, and each patient is an experiment unit.

Individual Casual Effect. The *Individual Casual Effect (ICE)* of treatment t with respect to control c for experiment unit i is defined as

$$ICE_i(t, c) \triangleq Y_i^t - Y_i^c. \tag{1.1}$$

However, only one of the two potential outcomes can be observed: if the i-th unit receives treatment t, we observe Y_i^t; if the i-th unit receives control c, we observe Y_i^c; and, (Y_i^c, Y_i^t) can never be obtained simultaneously. Thus, the $ICE_i(t, c)$ is not estimatable in practice.

Unit population. In practice, we are often more interested in the the casual effect of the treatments on a population of units instead of a specific unit. Use $\mathcal{P}$ to denote the unit population of interest. The population $\mathcal{P}$ can be either a finite population or an infinite population. Except for the outcomes (Y_i^c, Y_i^t), a unit $i \in \mathcal{P}$ is often associated with some covariates, which provide background information about the unit. Use X_i to denote these covariates. Thus, every unit $i \in \mathcal{P}$ associates with

a vector

$$(X_i, Y_i^c, Y_i^t).$$

Note that (X_i, Y_i^c, Y_i^t) are the properties of unit i itself and cannot be specified or controlled by us. The unit population $\mathcal{P}$ naturally induces a probability space of (X, Y^c, Y^t). In this dissertation, we use $\pi(x, y^c, y^t)$ to denote the joint distribution of (X, Y^c, Y^t) over the unit population $\mathcal{P}$.

Average Causal Effect. Given the unit population $\mathcal{P}$, the *Average Causal Effect (ACE)* of treatment t with respect to control c on $\mathcal{P}$ is defined as the average ICE over unit population $\mathcal{P}$, i.e.,

$$ACE_{\mathcal{P}}(t, c) \triangleq E_{\mathcal{P}}(ICE_i) = E_{\mathcal{P}}(Y_i^t - Y_i^c), \tag{1.2}$$

where $E_{\mathcal{P}}$ denotes the expectation with respect to the unit population $\mathcal{P}$.

It's easy to check that the value of ACE defined in Eq. (1.2) only depends on the characteristic distribution π of population $\mathcal{P}$, i.e.,

$$ACE_{\mathcal{P}}(t, c) = E_{\pi}(Y^t - Y^c) \triangleq \int y^t d\pi(y^c) - \int y^c d\pi(y^c),$$

where $\pi(y^c)$ and $\pi(y^t)$ are marginal distributions of Y^c and Y^t, respectively.

1.2.2 Causal inference via a randomized experiment

Randomized experiment (RE) is an effective technique to achieve proper estimation of ACE. In a random experiment, we **randomly assign** treatment t or control c

to units **randomly sampled** from the unit population $\mathcal{P}$. Use T_i to denote the treatment assignment for unit i ($T_i = t$ or c). The n units involved in the experiment can be divided into two arms: **the control arm** $\mathcal{A}_c = \{i : T_i = c\}$ and the **treatment arm** $\mathcal{A}_t = \{i : T_i = t\}$. Let $n_c = \#\mathcal{A}_c,\ n_t = \#\mathcal{A}_t$. Let

$$Y_i^{obs} = Y_i^t \cdot I(i \in \mathcal{A}_t) + Y_i^c \cdot I(i \in \mathcal{A}_c)$$

be the *observed response* for unit i. Let $\mathbf{Y}^{obs} = \{Y_1^{obs}, \cdots, Y_n^{obs}\}$ be the observed responses for the n units involved in the experiment.

To claim that an experiment is a randomized experiment, we need two conditions:

(1) Random selection of units: the units involved in the experiment are random samples from the unit population $\mathcal{P}$, i.e.,

$$(X_i, Y_i^c, Y_i^t) \sim \pi(x, y^c, y^t);$$

(2) Random assignment of treatment: $\{T_i\}_i$ are independent of each other, and

$$(Y_i^c, Y_i^t) \perp T_i, \quad \forall\ i \in \mathcal{A}_c \cup \mathcal{A}_t.$$

Note that although the *random assignment* condition is greatly emphasized in the literature, the *random selection* condition is often ignored by many researchers.

In practice, however, unless the concrete model of potential outcomes is known, there is no general way to construct a random experiment other than carrying out a *completely randomized experiment* (CRE), in which the treatment assignment is

purely random and independent of any other factors, and thus, independent of the potential outcomes (Y_i^c, Y_i^t).

Given the mechanism of a random experiment, the joint distribution of (X, Y^c, Y^t) can be extended to a higher dimensional distribution of (X, Y^c, Y^t, T) where the treatment assignment T is also covered. In the following, we will use π to denote the either the joint distribution for (X, Y^c, Y^t, T) or a marginal distribution of it. The specific meaning of π can be determined based on the context.

The most attractive property of a random experiment is that asymptotically the unit subpopulations in the treatment and control arms are both identical to the target population $\mathcal{P}$. Thus, we have

$$ACE_{\mathcal{P}}(t, c) = E_\pi(Y^t - Y^c) = E_\pi(Y^{obs} \mid T = t) - E_\pi(Y^{obs} \mid T = c),$$

which can be asymptotically unbiasedly estimated by

$$\widehat{ACE}_{RE} = \frac{1}{n_t} \sum_{i \in \mathcal{A}_t} Y_i^{obs} - \frac{1}{n_c} \sum_{i \in \mathcal{A}_c} Y_i^{obs}. \tag{1.3}$$

1.2.3 Non-randomized experiments

In practice, random experiments are often infeasible or imperfectly carried out. A more realistic scenario is that the treatment assignment T is marginally dependent of potential outcomes (Y^c, Y^t), but conditionally independent of (Y^c, Y^t) given observed covariates X in the extend joint distribution of (X, Y^c, Y^t, T). This condition is

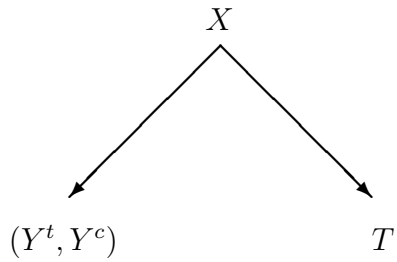


Figure 1.1: A graphical illustration of the relationship among covariates X, potential outcomes (Y^t, Y^c) and treatment assignment T in typical nonrandomized experiments.

formally proposed by Rubin (1974) as the assumption of *strong ignorability*:

$$(Y^t, Y^c) \perp T \mid X, \text{ and } 0 < P(T = t \mid X) < 1 \text{ for all } X, \tag{1.4}$$

which assumes that the randomization is properly carried out within each subpopulation defined by the covariates X.

Let $\mathcal{P}_x \triangleq \{i : X_i = x\}$ be the stratification of unit population $\mathcal{P}$ based on the value of covariates X. The stratification-level ACE is defined as

$$ACE_{\mathcal{P}_x}(t, c) \triangleq E_\pi(Y^t - Y^c \mid X = x). \tag{1.5}$$

As long as the treatment assignment is strongly ignorable given covariates X, we have

$$ACE_{\mathcal{P}_x}(t, c) = E_\pi(Y^{obs} \mid X = x, T = t) - E_\pi(Y^{obs} \mid X = x, T = c),$$

i.e., $ACE_{\mathcal{P}_x}(t, c)$ is identifiable on each unit stratification $\mathcal{P}_x$.

In fact, $\{ACE_{\mathcal{P}_x}\}_x$ are basic building bricks of population-level casual effects. For

example, the average casual effect of the treatment on the sampling unit population can be organized as

$$ACE_{\mathcal{P}}(t,c) \triangleq E_{\pi}(Y^t - Y^c) = \int ACE_{\mathcal{P}_x}(t,c)d\pi(x), \tag{1.6}$$

where $\pi(x)$ is the marginal distribution of covariates X in population $\mathcal{P}$. Sometimes, we are interested in the *average effect of the treatment on the treated* (ATT) that is defined as

$$ATT_{\mathcal{P}}(t,c) \triangleq ACE_{\mathcal{P}_t}(t,c) = \int ACE_{\mathcal{P}_x}(t,c)d\pi_t(x), \tag{1.7}$$

where $\pi_t(x)$ is the marginal distribution of covariates X in $\mathcal{P}_t$, the unit population of the treatment arm $\mathcal{A}_t$. Similarly, we can define the *average effect of the treatment on the controlled* (ATC) as

$$ATC_{\mathcal{P}}(t,c) \triangleq ACE_{\mathcal{P}_c}(t,c) = \int ACE_{\mathcal{P}_x}(t,c)d\pi_c(x), \tag{1.8}$$

where $\pi_c(x)$ is the marginal distribution of covariates X in $\mathcal{P}_c$, the unit population of the treatment arm $\mathcal{A}_c$.

The above equations indicate that a population-level casual effect always can be estimated by two steps:

(1) estimate $ACE_{\mathcal{P}_x}(t,c)$ in each stratification $\mathcal{P}_x$ based on Eq. (1.5),

(2) average out all stratifications according to the unit population of interest to get the population-level casual effect.

These equations also indicate that the marginal distribution of covariates is sufficient to represent a unit population. A major limitation of these equations, however, is that there are often too many stratifications. For example, if some covariates are continuous, the number of stratifications is infinite. This fact makes the estimate based on these equations inefficient or infeasible in many cases.

1.2.4 Propensity score of treatment assignment

A more efficient way to estimate $ACE_{\mathcal{P}}(t, c)$ is to create stratifications of population π based on *propensity score* instead of covariates. Formally, propensity score of treatment assignment T is defined as

$$e(X) = P(T = t \mid X). \tag{1.9}$$

Rosenbaum and Rubin (1983) pointed out that propensity score $e(X)$ has the following properties:

(a) $e(X)$ is also a *balancing score*, i.e., $X \perp T \mid e(X)$;

(b) if the treatment assignment is strongly ignorable given X, then it is strongly ignorable given $e(X)$, i.e.,

$$(Y^t, Y^c) \perp T \mid e(X), \text{ and } 0 < P(T = 1 \mid e(X)) < 1 \text{ for all } e(X),$$

indicating that the ACE of t with respect to c can be effectively estimated on

the stratification of $e(X)$, i.e.,

$$\begin{aligned} ACE_{\mathcal{P}_e}(t,c) &\triangleq E_\pi\{Y^t - Y^c \mid e(X) = e\} \\ &= E_\pi\{Y^t \mid e(X) = e, T = t\} - E_\pi\{Y^c \mid e(X) = e, T = c\} \end{aligned} \tag{1.10}$$

(c) The stratification based on $e(X)$ is the finest stratification of experiment units that satisfies both (a) and (b).

A joint distribution of $(e(X), Y^c, Y^t, T)$ can be induced from the joint distribution of (X, Y^c, Y^t, T). Let $\pi(e)$ is the marginal distribution of propensity score $e(X)$. Because

$$ACE_{\mathcal{P}}(t,c) \triangleq E_\pi(Y^t - Y^c) = \int ACE_{\mathcal{P}_e}(t,c)d\pi(e), \tag{1.11}$$

a modified two-step algorithm to estimate $ACE_\pi(t,c)$ can be obtained as follows:

(1) estimate $ACE_{\mathcal{P}_e}(t,c)$ in each stratification $\mathcal{P}_e$ based on Eq. (1.10),

(2) get $ACE_{\mathcal{P}}(t,c)$ by averaging out all stratifications based on Eq. (1.11).

Compared to the two-step algorithm in the previous subsection, the modified algorithm based on propensity score enjoys a better statistical efficiency as less stratifications are created.

In practice, propensity score of treatment assignment $e(X)$ is usually unknown and needs to be estimated from the observed data. When all covariates involved are discrete, $e(X)$ can be estimated empirically if the sample size is large enough. When some of the covariates are continuous, however, $e(X)$ cannot be estimated empirically

anymore. A popular solution to this problem is to specify a parametric form for propensity score $e(X)$, e.g., the logistic regression model with unknown parameter β:

$$e(X) = logisitc(X'\beta). \tag{1.12}$$

1.2.5 Bayesian causal inference

Suppose $\{X_i, T_i, Y_i^{obs}\}_{i \in \mathcal{A}_N \cup \mathcal{A}_O}$ are data observed in an experiment. Once the data generating mechanism is explicitly known, causal inference can be carried out in a Bayesian fashion (Rubin, 1978). The data generating mechanism usually contains three components: (1) model for covariates, (2) model for treatment assignment, and (3) model for potential outcomes. Here, we specify the three components as follows:

$$\begin{aligned} X_i &\sim \pi_\theta(x), \\ T_i \mid X_i = x &\sim Bernoulli(e(x)), \\ Y_i^c \mid X_i = x &\sim Bernoulli(R_c(x)), \\ Y_i^t \mid X_i = x &\sim Bernoulli(R_t(x)), \end{aligned}$$

where π_θ is a parametric distribution with unknown parameter θ, propensity score $e(x)$ and the two response surfaces $R_c(x)$ and $R_t(x)$ are specified as logistic and probit

models respectively:

$$\begin{aligned} e(x) &= logistic(x'\beta), \\ R_c(x) &= \Phi(x'\alpha_c), \\ R_t(x) &= \Phi(x'\alpha_t). \end{aligned}$$

The above model lead to the following likelihood for the observed data:

$$\begin{aligned} f(X,T,Y^{obs}) &= \prod_{i=1}^{n} f(X_i,T_i,Y_i^{obs}) = \prod_{i=1}^{n} f(X_i)\cdot f(T_i \mid X_i)\cdot f(Y_i^{obs} \mid X_i) \\ &= \prod_{i=1}^{n} \pi_\theta(X_i)\cdot \prod_{i\in\mathcal{A}_t} e_\beta(X_i) R_t(X_i)^{Y_i^{obs}}[1-R_t(X_i)]^{1-Y_i^{obs}} \\ &\quad \cdot \prod_{i\in\mathcal{A}_c}[1-e_\beta(X_i)] R_c(X_i)^{Y_i^{obs}}[1-R_c(X_i)]^{1-Y_i^{obs}}. \end{aligned}$$

Giving the model parameters $(\theta,\beta,\alpha_t,\alpha_c)$ a proper prior distribution, their posterior distribution can be derived by standard Bayesian inference techniques. Note that if $(\theta,\beta,\alpha_t,\alpha_c)$ are independent of each other in the prior distribution, they will be still independent in the posterior distribution since they are separated in the above likelihood function.

Because the estimand

$$ACE_{\pi_\theta}(t,c) = \int \big[R_t(x) - R_c(x)\big] d\pi_\theta(x) = h(\alpha_t,\alpha_c,\theta)$$

is a function of parameters $(\alpha_t,\alpha_c,\theta)$, the posterior distribution of $ACE_{\pi_\theta}(t,c)$ can be naturally derived from the posterior distribution of $(\alpha_t,\alpha_c,\theta)$.

The above approach needs to be modified when the unit population of interest (refer to as π^*) is different from π_θ, the population from which the units are selected. For example, if π^* is a pre-given distribution of X, the estimand becomes

$$ACE_{\pi^*}(t,c) = \int \big[R_t(x) - R_c(x)\big] d\pi^*(x) = h(\alpha_t, \alpha_c),$$

whose posterior distribution only depends on the posterior distribution of (α_t, α_c). If the *average treatment effect on the treated* (i.e., ATT) is of interest, we have

$$\pi^*(x) = \pi_t(x) = P(X = x \mid T = t) = \frac{e(x) \cdot \pi_\theta(x)}{\int e(x) d\pi_\theta(x)}$$

is a function of (β, θ). Thus, the posterior distribution of the estimand

$$ACE_{\pi^*}(t,c) = \int \big[R_t(x) - R_c(x)\big] d\pi_t(x) = h(\alpha_t, \alpha_c, \beta, \theta)$$

can be derived from the posterior distribution of $(\alpha_t, \alpha_c, \beta, \theta)$.

This approach can be applied to data collected in a randomized experiment as well. In a randomized experiment, the treatment assignment T is marginally independent of the potential outcomes, i.e., $(Y^c, Y^t) \perp T$. Thus, the model for treatment assignment is simplified as:

$$T_i \sim Bernoulli(\rho).$$

Assume that the covariates X are also observed in the randomized experiment, and the models of potential outcomes (i.e., $Y^c \mid X$ and $Y^t \mid X$) have the same form, we

come up with the following simplified likelihood:

$$\begin{aligned} f(X,T,Y^{obs}) &= \prod_{i=1}^{n} f(X_i,T_i,Y_i^{obs}) = \prod_{i=1}^{n} f(X_i)f(T_i)f(Y_i^{obs} \mid X_i) \\ &= \rho^{n_t}(1-\rho)^{n_c} \cdot \prod_{i=1}^{n} \pi_\theta(X_i) \cdot \prod_{i\in\mathcal{A}_t} R_t(X_i)^{Y_i^{obs}}[1-R_t(X_i)]^{1-Y_i^{obs}} \\ &\quad \cdot \prod_{i\in\mathcal{A}_c} R_c(X_i)^{Y_i^{obs}}[1-R_c(X_i)]^{1-Y_i^{obs}}. \end{aligned}$$

Based on the likelihood, Bayesian inference for $ACE_{\pi_\theta}(t,c)$ or $ACE_{\pi^*}(t,c)$ can be achieved in a similar way.

Note that Bayesian inference of ACE is robust to the design of the experiment:

Proposition 1 *With the following models for covariates and potential outcomes:*

$$X_i \sim \pi_\theta(x),\ Y_i^c \mid X_i = x \sim Bernoulli(R_c(x)),\ Y_i^t \mid X_i = x \sim Bernoulli(R_t(x)),$$

given the observed data (X,T,Y^{obs}) from an experiment, if the model parameters are independent of each other in the prior distribution, the result of Bayesian inference for $ACE_{\pi_\theta}(t,c)$ keeps unchanged no matter we assume the experiment is randomized or conditionally randomized given the covariates X.

1.3 A Review of Matching Methods for Direction Comparison

Let $\pi(x)$ be the covariate distribution over the unit population of interest. Define the arm-level covariate distributions as:

$$\begin{aligned} \pi_t(x) &= P(X = x \mid T = t), \\ \pi_c(x) &= P(X = x \mid T = c). \end{aligned}$$

If the units involved in the experiment are randomly sampled from the target population π, and the treatment is randomly assigned to each unit, theoretically we will expect

$$\pi_t(x) = \pi_c(x) = \pi(x) \text{ for } \forall\ x.$$

In practice, however, due to randomness of unit sampling and treatment assignment, the empirical arm-level distributions $\tilde{\pi}_t$ and $\tilde{\pi}_c$ can be very different from each other even in a randomized experiment. The same thing happens, if the experiment of interet is not randomized at the first place. In either case, directly applying the naive estimate

$$\widehat{ACE}_{RE} = \frac{1}{n_t} \sum_{i \in \mathcal{A}_t} Y_i^{obs} - \frac{1}{n_c} \sum_{i \in \mathcal{A}_c} Y_i^{obs}$$

would lead to biased estimation of the causal effect of interest.

Bayesian inference is a good choice for this tricky situation when data generating mechanism of the experiment is known. When the mechanism is not exactly known, however, full Bayesian inference becomes infeasible unless untestable assumptions

are made. Another line in the literature to effectively reduce bias due to unbalanced covariate distribution is *matching*. Formally, *matching* refers to any method that aims to equate or "balance" the distribution of covariates in the treated and control groups. Balancing can be achieved by either selecting a subgroup of units that fit the target distribution, or assigning weights to units to generate properly weighted samples of the target distribution. Once the covariate distribution is successfully balanced in the treated and control groups, we essentially convert the original data from a poorly randomized experiment to equivalent data from a perfectly randomized experiment, based on which downstream analysis can be easily carried out without worrying the potential risks caused by unbalanced covariate distribution. From practice point of view, a major advantage of matching methods is that we can avoid to specify concrete models for covariates and potential outcomes.

In a well randomized experiment, the covariate distribution is balanced across the treated and control arms. Thus, we have

$$\begin{aligned} E_\pi(Y^t) &= E_\pi(Y^t \mid T=t) = E_\pi(Y^{obs} \mid T=t), \\ E_\pi(Y^c) &= E_\pi(Y^c \mid T=c) = E_\pi(Y^{obs} \mid T=c), \end{aligned}$$

which guarantees that $\widehat{ACE}_{RE}$ is a proper estimate of $ACE_{\mathcal{P}}(t,c)$. In a conditionaly

randomized experiment, however, we only have

$$\begin{aligned} E_\pi(Y^t) &= \int E(Y^t \mid X) d\pi(x), \\ E_\pi(Y^{obs} \mid T = t) &= E_\pi(Y^t \mid T = t) \\ &= \int E(Y^t \mid X = x, T = t) d\pi(x \mid T = t) \\ &= \int E(Y^t \mid X) d\pi(x \mid T = t). \end{aligned}$$

Therefore, $E_\pi(Y^t) \neq E_\pi(Y^{obs} \mid T = t)$ unless $\pi(x) = \pi(x \mid T = t)$ for all x. Similarly, $E_\pi(Y^c) \neq E_\pi(Y^{obs} \mid T = c)$ unless $\pi(x) = \pi(x \mid T = c)$ for all x. Therefore, for a conditionally randomized experiment, a sufficient condition to make $\widehat{ACE}_{RE}$ proper is

$$\pi_t(x) \triangleq \pi(x \mid T = t) = \pi(x \mid T = c) \triangleq \pi_c(x) \text{ for } \forall\ x, \text{ or, } X \perp T.$$

Considering that $(Y^c, Y^t) \perp T \mid X$ in a conditionally randomized experiment, and

$$\left\{(Y^c, Y^t) \perp T \mid X;\ X \perp T\right\} \Longrightarrow (Y^c, Y^t) \perp T,$$

the effort of matching π_t and π_c in fact converts data from a conditional randomized experiment to equivalent data from a perfectly randomized experiment. In this section, we will give a brief review of major matching approaches in the literature. A more comprehensive review can be found in Stuart (2010).

1.3.1 Nearest neighbor matching

One of the most common and easiest to implement methods is the *nearest neighbor matching* (NNM) proposed by Rubin (1973a). In NNM, we select for each treated unit i its nearest neighbors in the controlled units according to a pre-given distance D_{ij} to measure the similarity of unit i and unit j. Let $\mathcal{N}_i$ be the neighbors of treated unit i selected from the controlled units, we estimate the individual casual effect on unit i as

$$\widehat{ICE}_i(t,c) \triangleq Y_i^{obs} - \frac{1}{\#\mathcal{N}_i}\sum_{j\in\mathcal{N}_i} Y_j^{obs}, \tag{1.13}$$

based on which the average casual effect of treatment on the treated (i.e., ATT) can be estimated by

$$\widehat{ATT}_{NNM} = \frac{1}{n_t}\sum_{i\in\mathcal{A}_t} \widehat{ICE}_i(t,c). \tag{1.14}$$

NNM is generally the most effective method for settings where the goal is to select units for follow-up, since it simply discards all controlled units that are not selected. This strategy is useful particularly in cases where the controlled arm is much larger than the treatment arm so that it's economically infeasible to measure the response and covariates for all units in the controlled arm. In practice, the estimation in Eq. (1.13) can be further improved by a weighted version where the selected units in $\mathcal{N}_i$ are weighted based on their distances to treated unit i.

1:1 nearest neighbor matching. In the simplest form of NNM, we only select one controlled unit for each treated unit i, i.e. $\#\mathcal{N}_i = 1$. Algorithmically, the selection procedure goes as follows:

- Step 1. Randomly arrange the treated units with a certain order.

- Step 2. Scan through all treated units according to the given order.

- Step 3. For each treated unit i, find from the controlled units that have not been selected yet the one with the smallest distance to it, and let $\mathcal{N}_i$ be the selected unit.

A common complaint regarding 1 : 1 matching is that it can discard a large number of observations and thus would apparently lead to reduced power. However, the reduction in power is often minimal, for two main reasons. First, in a two-sample comparison of means, the precision is largely driven by the smaller group size (Cohen, 1988). So if the treatment group stays the same size, and only the control group decreases in size, the overall power may not actually be reduced very much (Ho et al., 2007). Second, the power increases when the groups are more similar because of the reduced extrapolation and higher precision that is obtained when comparing groups that are similar versus groups that are quite different (Snedecor and Cochran, 1980).

Another complication of the above method is that the order in which the treated units are matched may change the quality of the matches. *Optimal matching* (Rosenbaum, 2002) avoids this issue by taking into account the overall set of matches when choosing individual matches, minimizing a global distance measure. Generally, greedy matching performs poorly when there is intense competition for controls, and performs well when there is little competition (Gu and Rosenbaum, 1993). Gu and Rosenbaum (1993) find that optimal matching does not in general perform any better than greedy matching in terms of creating groups with good balance, but does do better at reducing the distance within pairs. Thus, if the goal is simply to find

well-matched groups, greedy matching may be sufficient. However, if the goal is to create well-matched pairs, then optimal matching may be preferable.

Ratio matching. When there are large numbers of control individuals, it is sometimes possible to get multiple good matches for each treated individual, called *ratio matching* (Smith, 1997; Rubin and Thomas, 2000). Selecting the number of matches involves a bias-variance trade-off. Selecting multiple controls for each treated individual will generally increase bias since the 2nd, 3rd and 4th closest matches are, by definition, further away from the treated individual than is the 1st closest match. On the other hand, utilizing multiple matches can decrease variance due to the larger matched sample size. Approximations in Rubin and Thomas (1996) can help determine the best ratio. In settings where the outcome data has yet to be collected and there are cost constraints, researchers must also balance cost considerations. More methodological work needs to be done to more formally quantify the trade-offs involved. In addition, $k : 1$ matching is not optimal since it does not account for the fact that some treated individuals may have many close matches while others have very few. A more advanced form of ratio matching, variable ratio matching, allows the ratio to vary, with different treated individuals receiving differing numbers of matches (Ming and Rosenbaum, 2001). Variable ratio matching is related to full matching, described below.

An additional concern is that, without any restrictions, $k : 1$ matching can lead to some poor matches, if, for example, there are no control individuals with propensity scores similar to a given treated individual. One strategy to avoid poor matches is to impose a caliper and only select a match if it is within the caliper. This can

lead to difficulties in interpreting effects if many treated individuals do not receive a match, but can help avoid poor matches. Rosenbaum and Rubin (1985a) discuss those trade-offs.

Distance measurements. Different distance measurements can be used to select the nearest neighbors. In practice, the following four distances are widely used. Sometimes, these distance measures can also be combined.

1. Exact:

$$D_{ij} = \begin{cases} 0, & \text{if } X_i = X_j, \\ \infty, & \text{if } X_i \neq X_j. \end{cases}$$

2. Mahalanobis:

$$D_{ij} = (X_i - X_j)' \Sigma^{-1} (X_i - X_j),$$

where Σ is the variance covariance matrix of X in the full control group. If X contains categorical variables, they should be converted to a series of binary indicators, although the distance works best with continuous variables.

3. Propensity score:

$$D_{ij} = |e_i - e_j|,$$

where e_k is the propensity score for unit k.

4. Linear propensity score:

$$D_{ij} = |logit(e_i) - logit(e_j)|.$$

Rosenbaum and Rubin (1985b), Rubin and Thomas (1996) and Rubin (2001) have found that matching on the linear propensity score can be particularly effective in

terms of reducing bias.

With or without replacement. Another key issue is whether controls can be used as matches for more than one treated individual: whether the matching should be done with replacement or without replacement. Matching with replacement can often decrease bias because controls that look similar to many treated individuals can be used multiple times. This is particularly helpful in settings where there are few control individuals comparable to the treated individuals (e.g., Dehejia and Wahba, 1999). Additionally, when matching with replacement, the order in which the treated individuals are matched does not matter. However, inference becomes more complex when matching with replacement, because the matched controls are no longer independentsome are in the matched sample more than once and this needs to be accounted for in the outcome analysis, for example, by using frequency weights. When matching with replacement, it is also possible that the treatment effect estimate will be based on just a small number of controls; the number of times each control is matched should be monitored.

1.3.2 Subclassification

Subclassification forms groups of individuals who are similar, for example, as defined by quintiles of the propensity score distribution. It can estimate either the ACE or the ATT. One of the first uses of subclassification was Cochran (1968), which provides analytic expressions for the bias reduction possible using subclassification on a univariate continuous covariate, showing that using just five subclasses removes at least 90% of the initial bias due to that covariate. Rosenbaum and Rubin (1985b)

extended that to show that creating five propensity score subclasses removes at least 90% of the bias in the estimated treatment effect due to all of the covariates that went into the propensity score. Based on those results, the current convention is to use 510 subclasses. However, with larger sample sizes more subclasses (e.g.,1020) may be feasible and appropriate (Lunceford and Davidian, 2004).

A more sophisticated form of subclassification named as *full matching* selects the number of subclasses automatically (Rosenbaum, 1991; Hansen, 2004; Stuart and Green, 2008). Full matching creates a series of matched sets, where each matched set contains at least one treated individual and at least one control individual (and each matched set may have many from either group). Full matching is optimal in terms of minimizing the average of the distances between each treated individual and each control individual within each matched set. Full matching may have appeal for researchers who are reluctant to discard some of the control individuals but who want to obtain optimal balance on the propensity score. To achieve efficiency gains, Hansen (2004) introduces restricted ratios of the number of treated individuals to the number of control individuals in each matched set.

1.3.3 Weighting adjustment

In the literature, weighting adjustment matching has been widely discussed in different context (Czajka et al., 1992; Robins, Hernan and Brumback, 2000; Lunceford and Davidian, 2004; Hirano, Imbens and Ridder, 2003). Weights that were estimated by using the inverse of propensity score $e_i = P(T = t \mid X_i)$, were known as inverse

probability of treatment weighting (IPTW).

$$\hat{w}_i = \frac{1}{\hat{e}_i} \cdot I(T_i = t) + \frac{1}{1 - \hat{e}_i} \cdot I(T_i = c).$$

The estimated average causal effect becomes:

$$\widehat{ACE}_{WA-PS} = \frac{1}{\sum_{i \in \mathcal{A}_t} \hat{w}_i} \sum_{i \in \mathcal{A}_t} Y_i^{obs} \cdot \frac{1}{\hat{e}_i} - \frac{1}{\sum_{i \in \mathcal{A}_c} \hat{w}_i} \sum_{i \in \mathcal{A}_c} Y_i^{obs} \cdot \frac{1}{1 - \hat{e}_i}, \quad (1.15)$$

In the case that ATT was of interest, weights were constructed based on the odds ratio of treatment assignment,

$$\hat{w}_i = 1 \cdot I(T_i = t) + \frac{\hat{e}_i}{1 - \hat{e}_i} \cdot I(T_i = c).$$

And the estimated average effect of the treatment on the treated becomes

$$\widehat{ATT}_{WA-PS} = \frac{1}{n_t} \sum_{i \in \mathcal{A}_t} Y_i^{obs} - \frac{1}{\sum_{i \in \mathcal{A}_c} \hat{w}_i} \sum_{i \in \mathcal{A}_c} Y_i^{obs} \cdot \frac{\hat{e}_i}{1 - \hat{e}_i}. \quad (1.16)$$

As we will show in details in Section 1.5, the major advantage of weighting adjustment matching over NNM and subclassification is that it can be applied to the scenario when individual-level data are not available in one arm. The limitation of weighting adjustment matching, however, is that it may lead to a large variation of the estimate sine the weights for some units can be extremely large or small from time to time. Moreover, accuracy of the propensity score estimates (or the correctness of the model for propensity score) could be a critical issue when propensity score is used for weighting adjustment.

In practice, the above weighting adjustment approach can be modified to fit more complicated scenarios or improve the efficiency. For example, we can truncate the weights above up to a maximum value to avoid the estimate to be dominated by few data points with very large weights. In the literature, this strategy is called *weight trimming* (Potter, 1993; Scharfstein, Rotnitzky and Robins, 1999). Other methods include the *kernel weighting* (Imbens, 2000; Imbens, 2004) and *doubly-robust methods* (Bang and Robins, 2005).

1.3.4 Another perspective to understand propensity score weighting

Let π, π_t and π_c be the covariate distributions of the target population $\mathcal{P}$, treatment population $\mathcal{P}_t$, and control population $\mathcal{P}_c$, respectively. Suppose that π_t and π_c share a common support.

It's easy to check that:

$$\begin{aligned}
E_{\mathcal{P}_t}[Y^{obs} \cdot \frac{\pi(X)}{\pi_t(X)} \mid T = t] &= \int E[Y^t \cdot \frac{\pi(x)}{\pi_t(x)} \mid X = x, T = t] d\pi_t(x) \\
&= \int E(Y^t \mid X = x, T = t) d\pi(x) \\
&= \int E(Y^t \mid X = x) d\pi(x) = E_{\mathcal{P}} Y^t, \\
E_{\mathcal{P}_c}[Y^{obs} \cdot \frac{\pi(X)}{\pi_c(X)} \mid T = c] &= \int E[Y^c \cdot \frac{\pi(x)}{\pi_c(x)} \mid X = x, T = c] d\pi_c(x) \\
&= \int E(Y^c \mid X = x, T = c) d\pi(x) \\
&= \int E(Y^c \mid X = x) d\pi(x) = E_{\mathcal{P}} Y^c.
\end{aligned}$$

Thus, we have

$$\begin{aligned} AEC_{\mathcal{P}}(t,c) &= \int E(Y^t - Y^c \mid X = x) d\pi(x) \\ &= E_{\mathcal{P}_t}[Y^{obs} \cdot \frac{\pi(X)}{\pi_t(X)} \mid T = t] - E_{\mathcal{P}_c}[Y^{obs} \cdot \frac{\pi(X)}{\pi_c(X)} \mid T = c]. \quad (1.17) \end{aligned}$$

Defining weight of unit i as

$$w_i = \frac{\pi(X_i)}{\pi_t(X_i) \cdot I(T_i = t) + \pi_c(X_i) \cdot I(T_i = c)},$$

Eq. (1.17) leads to the following estimate of $AEC_\pi(t,c)$

$$\widehat{ACE}^*_{WA} = \frac{1}{n_t} \sum_{i \in \mathcal{A}_t} Y_i^{obs} \cdot w_i - \frac{1}{n_c} \sum_{i \in \mathcal{A}_c} Y_i^{obs} \cdot w_i.$$

The above estimate can be expressed alternatively in terms of propensity score. If ATT is of interest, we have $\pi = \pi_t$, and thus $w_i = 1$ for all $i \in \mathcal{A}_t$; for $i \in \mathcal{A}_c$, however, given its propensity score $e_i = P(T = t \mid X_i)$, we have

$$w_i = \frac{\pi_t(X_i)}{\pi_c(X_i)} = \frac{P(X_i \mid T = t)}{P(X_i \mid T = c)} = \frac{P(X_i)P(T = t \mid X_i)/P(T_i = t)}{P(X_i)P(T = c \mid X_i)/P(T_i = c)} = \frac{e_i}{1 - e_i} \cdot \frac{1 - \alpha}{\alpha},$$

where $\alpha = P(T = t)$ is the relative proportion of the treatment population over the pooled population. If ACE is of interest, we have $\mathcal{P} = \mathcal{P}_t \cup \mathcal{P}_c$, and

$$\pi(x) = \alpha \cdot \pi_t(x) + (1 - \alpha) \cdot \pi_c(x).$$

Thus, for unit i we have

$$\begin{aligned} w_i &= \frac{\alpha \cdot \pi_t(X_i) + (1-\alpha) \cdot \pi_c(X_i)}{\pi_t(X_i) \cdot I(T_i = t) + \pi_c(X_i) \cdot I(T_i = c)} \\ &= \big[\alpha + (1-\alpha)\frac{\pi_c(X_i)}{\pi_t(X_i)}\big] \cdot I(T_i = t) + \big[\alpha \cdot \frac{\pi_t(X_i)}{\pi_c(X_i)} + (1-\alpha)\big] \cdot I(T_i = c) \\ &= \alpha\big[1 + \frac{1-e_i}{e_i}\big] \cdot I(T_i = t) + (1-\alpha)\big[\frac{e_i}{1-e_i} + 1\big] \cdot I(T_i = c) \\ &= \frac{\alpha}{e_i} \cdot I(T_i = t) + \frac{1-\alpha}{1-e_i} \cdot I(T_i = c). \end{aligned}$$

Although start from different perspective, these provided consistent weights as shown in Eq. 1.15 and Eq. 1.16 other than a normalizing constant.

In practice, covariate distributions π_t, π_c and propensity scores $\{e_i\}_i$ are usually unknown, and need to be estimated from the data. Let $\hat{\pi}_t$ and $\hat{\pi}_c$ be proper estimates of π_t and π_c, respectively. Weight w_i can be estimated by:

$$\hat{w}_i = \frac{\pi(X_i)}{\hat{\pi}_t(X_i) \cdot I(T_i = t) + \hat{\pi}_c(X_i) \cdot I(T_i = c)}. \tag{1.18}$$

1.4 Casual Inference via Indirect Comparison

1.4.1 The problem setting

A head-to-head trial is the most straightforward way to compare the efficacy of a new drug D_N to a well accepted drug D_O. With a well designed randomized experiment on a patient population $\mathcal{P}$, we can efficiently estimate the relative efficacy of D_N with respect to D_O, which is define as $ACE_{\mathcal{P}}(D_N, D_O)$.

In practice, however, there are often many, say m, well accepted drugs in the

market. To establish reputation for the new drug, the drug developer must provide evidences to show that the new drug is better than the m comparator drugs already in the market. If we choose to achieve this by direct comparison via head-to-head trials, m trials will be needed, each for one comparator drug. Clearly, this strategy is economically and timely infeasible when m is large.

In this case, *indirect comparison* could be a better strategy to achieve the same purpose. The simplest version of indirect comparison goes as follows: if both the new drug D_N and a traditional drug D_O have been compared to a common baseline drug (e.g., a standard placebo) denoted as B for the target patient population $\mathcal{P}$ in two randomized experiments/trials denoted as $\mathcal{T}_N$ and $\mathcal{T}_O$, we can compare them indirectly by

$$ACE_{\mathcal{P}}(D_N, D_O) = ACE_{\mathcal{P}}(D_N, B) - ACE_{\mathcal{P}}(D_O, B), \tag{1.19}$$

since the effect of the common baseline drug B in two trials cancels out.

A more realistic scenarios is: different baseline drugs, say B_N and B_O, are used in the two trials $\mathcal{T}_N$ and $\mathcal{T}_O$. In this setting, we have two trials $\mathcal{T}_N$ and $\mathcal{T}_O$ for four treatments D_N, D_O, B_N and B_O, where $\mathcal{T}_N$ is a randomized experiment for D_N and B_N, and $\mathcal{T}_O$ is a randomized experiment for D_O and B_O. The observed data for a patient i in an indirect comparison include $(X_i, I_i, T_i, Y_i^{obs})$, where X_i stands for the covariates, $I_i \in \{\mathcal{T}_N, \mathcal{T}_O\}$ is the trial assignment, $T_i \in \{D_N, D_O, B_N, B_O\}$ is the treatment assignment, and Y_i^{obs}, which has the same meaning as in the previous sections, takes values in $\{Y_i^N, Y_i^O, Y_i^{B_N}, Y_i^{B_O}\}$. There are totally four arms in the two trials of an indirect comparison, two for each trial. We use the following notations to

denote the four arms:

$$\mathcal{A}_N^+ \triangleq \{i : I_i = \mathcal{T}_N, T_i = D_N\},\ \mathcal{A}_B^+ \triangleq \{i : I_i = \mathcal{T}_N, T_i = B_N\};$$

$$\mathcal{A}_O^- \triangleq \{i : I_i = \mathcal{T}_O, T_i = D_O\},\ \mathcal{A}_B^- \triangleq \{i : I_i = \mathcal{T}_O, T_i = B_O\}.$$

And, the sizes of the four arms are denoted as

$$n_N^+ = \#\mathcal{A}_N^+,\ n_B^+ = \#\mathcal{A}_B^+,\ n_O^- = \#\mathcal{A}_O^-, \text{ and } n_B^- = \#\mathcal{A}_B^-.$$

In this case, Eq. (1.19) becomes a biased estimator of $ACE_{\mathcal{P}}(D_N, D_O)$, as the effect of B_N and B_O cannot naturally cancel out. However, as long as $\mathcal{T}_N$ and $\mathcal{T}_O$ both cover the target population $\mathcal{P}$, we can always put $\mathcal{A}_N^+$ and $\mathcal{A}_O^-$ together to generate an artificial trial for comparing D_N and D_O. It's easy to check that the artificial trial with $\mathcal{A}_N^+$ as the treatment arm and $\mathcal{A}_O^-$ as the controlled arm is still a randomized experiment of D_N and D_O over the target population $\mathcal{P}$, i.e.,

$$T \perp (Y^{D_N}, Y^{D_O}).$$

Based on this fact, $ACE_{\mathcal{P}}(D_N, D_O)$ can be properly identified from the artificial trial by

$$\widehat{ACE}_{IC} = \frac{1}{n_N^+} \sum_{i \in \mathcal{A}_N^+} Y_i^{obs} - \frac{1}{n_O^-} \sum_{i \in \mathcal{A}_O^-} Y_i^{obs}. \tag{1.20}$$

A major challenge in indirect comparison, however, is that the unit population of trail $\mathcal{T}_N$ is often different from that of trial $\mathcal{T}_O$. Let $\mathcal{P}_O$ and $\mathcal{P}_N$ be the unit

population of trial $\mathcal{T}_O$ and $\mathcal{T}_N$ respectively. When $\mathcal{P}_N \neq \mathcal{P}_O$, the estimator defined in (1.20) becomes a biased estimator. To remove the bias, we need to match the patient population of the two trials. We call this type of matching problem as *across-trial matching* to distinguish it from the classic *across-arm matching* problem that matches covariate distributions of two arms from one trial.

1.4.2 Ignobility of trial assignment

Since both $\mathcal{T}_O$ and $\mathcal{T}_N$ are randomized experiments, they satisfy the following conditions:

(1) Random unit selection within trials: The units in trial $\mathcal{T}_O$ are i.i.d. samples from a unit population $\mathcal{P}_O$, the units in trial $\mathcal{T}_N$ are i.i.d. samples from another unit population $\mathcal{P}_N$. Let $\pi_O(x)$ and $\pi_N(x)$ be the covariate distributions of population $\mathcal{P}_O$ and $\mathcal{P}_N$, respectively. This condition can be formally expressed as:

$$X_i \sim \pi_O(x) \cdot I(i \in \mathcal{T}_O) + \pi_N(x) \cdot I(i \in \mathcal{T}_N);$$

(2) Random treatment assignment within trials: $\{T_i\}_i$ are independent of each other, and the treatment assignment is ignorable in both trials, i.e.,

$$(Y^{D_N}, Y^{B_N}) \perp T \mid I = \mathcal{T}_N,$$
$$(Y^{D_O}, Y^{B_O}) \perp T \mid I = \mathcal{T}_O.$$

To achieve across-trial matching, however, we need an extra condition below on trial assignment:

(3) Ignorable trial assignment: the trial assignment is strongly ignorable given the covariants X, i.e.,

$$(Y^{D_N}, Y^{D_O}) \perp I \mid X, \text{ and}$$

$$\pi_N \text{ and } \pi_O \text{ have the same support as the target population } \mathcal{P}.$$

Intuitively, this condition says that the covariates X are sufficient to represent difference between the two patient populations.

To better illustrate the meaning of the condition of "ignorable trial assignment", we provide the following example. An old drug D_O has been proved to be effective for asians with respect to a standard placebo B by a randomized clinical trial $\mathcal{T}_O$. Now, we want to check whether a new drug D_N is better than the old drug D_O for asians via indirect comparison. However, we cannot find enough asian patients to carry out a randomized clinical trial to compare the new drug D_N to the placebo B. Instead, we carry out the clinical trial (i.e., $\mathcal{T}_N$) for a group of caucasians with a similar covariate distribution. Assume that four covariates are considered: age (X_1), gender (X_2), blood pressure (X_3), and income in US dollars (X_4). Here, $\mathcal{P}_O$ is the population of asians, and $\mathcal{P}_N$ is the population of caucasians.

Because caucasians are very different from asians racially and genetically, the relative efficacy of the new drug D_N for caucasians obtained in $\mathcal{T}_N$ may have nothing to do with its relative efficacy for asians at all. However, if we can somehow argue that the genetic effect is very marginal here, and the efficacy of drugs is mainly determined by the basic body conditions of a patient, which can be well represented by age (X_1), gender (X_2) and blood pressure (X_3), and his/her living condition, which can be

surrogated by the income level, we can comfortably extend the results obtained from the caucasian population to an asian population with a similar covariate distribution.

More precisely, let $\{\mathcal{P}_x^N\}_x$ be the the unit stratifications defined by the value of covariates in population $\mathcal{P}_N$, and $\{\mathcal{P}_x^O\}_x$ be the the unit stratifications defined by the value of covariates in population $\mathcal{P}_O$. The assumption of ignorable trial assignment guarantees that the statistical properties of potential outcomes Y^{D_N} and Y^{D_O} are exactly same in $\mathcal{P}_x^N$ and $\mathcal{P}_x^O$ for any x. In other words, with respect to Y^{D_N} and Y^{D_O} only, we have

$$\mathcal{P}_x^N = \mathcal{P}_x^O = \mathcal{P}_x \text{ for } \forall\ x.$$

Thus, the covariates X serve as a bridge between the two populations $\mathcal{P}_N$ and $\mathcal{P}_O$, with which the two populations are comparable at the unit stratification defined by the value of covariates $\{\mathcal{P}_x\}_x$.

In practice, the condition of "random treatment assignment within trials" can be relaxed to a looser condition below:

(2*) Ignorable treatment assignment within trials: $\{T_i\}_i$ are independent of each other, and the treatment assignment is ignorable in both trials, i.e.,

$$(Y^{D_N}, Y^{B_N}) \perp T \mid (I = \mathcal{T}_N, X),$$
$$(Y^{D_O}, Y^{B_O}) \perp T \mid (I = \mathcal{T}_O, X);\ \text{and}$$

$$0 < P(T = D_O \mid X, I = \mathcal{T}_O) < 1 \text{ and } 0 < P(T = D_N \mid X, I = \mathcal{T}_N) < 1 \text{ for all } X.$$

For simpleness, however, in this dissertation we always assume that both trials are perfectly randomized, i.e., within each trial the patient populations in two arms of the trial are identical.

1.4.3 Across-trial matching vs across-arm matching

The across-trial matching problem looks almost same as the across-arm matching problem discussed in the previous section: both aim to balance distribution of covariates across two unit subpopulations (arms or trials). However, the two problems do have some subtle but critical differences.

In the across-arm matching problem, the data are generated in three steps:

- Covariates X: draw X randomly from distribution π derived from unit population $\mathcal{P}$;
- Potential outcomes $(Y^c, Y^t) \mid X$: potential outcomes are summarized/modeled at the level of unit stratifications $\{\mathcal{P}_x\}_x$, i.e., the following response surfaces must be specified

$$E_{\mathcal{P}_x^t}(Y^t),\ E_{\mathcal{P}_x^t}(Y^c),\ E_{\mathcal{P}_x^c}(Y^t),\ E_{\mathcal{P}_x^c}(Y^c);$$

- Treatment assignment $T \mid X$: treatment T is randomly assigned to units within each unit stratification $\mathcal{P}_x$, thus, satisfies $(Y^c, Y^t) \perp T \mid X$, which guarantees the statistical property of potential outcomes in each unit stratification is

identical across two arms, i.e.,

$$E_{\mathcal{P}_x^t}(Y^t) = E_{\mathcal{P}_x^c}(Y^t) \text{ and } E_{\mathcal{P}_x^t}(Y^c) = E_{\mathcal{P}_x^c}(Y^c) \text{ for } \forall\ x.$$

The variable that defines the two arms, T, is randomly assigned to units within each unit stratification $\mathcal{P}_x$, after involved units are randomly selected from the unit population $\mathcal{P}$.

In the across-trial matching problem, however, the data are generated in four steps:

- Trial assignment I: specify number of units in both trials n_O and n_N.
- Covariates $X \mid I$: for unit in trail $\mathcal{T}_N$, draw X randomly from distribution π_N derived from unit population $\mathcal{P}_N$; for unit in trail $\mathcal{T}_O$, draw X randomly from distribution π_O derived from unit population $\mathcal{P}_O$.
- Potential outcomes in two trials $(Y^{D_N}, Y^{B_N}, Y^{D_O}, Y^{B_O}) \mid (X, I)$: potential outcomes are summarized/modeled at the level of unit stratifications in both trials $\{\mathcal{P}_x^N\}_x \cup \{\mathcal{P}_x^O\}_x$, i.e., the following response surfaces must be specified

$$E_{\mathcal{P}_x^N}(Y^{D_N}),\ E_{\mathcal{P}_x^N}(Y^{B_N}),\ E_{\mathcal{P}_x^O}(Y^{D_O}),\ E_{\mathcal{P}_x^O}(Y^{B_O});$$

and, since $(Y^{D_N}, Y^{D_O}) \perp I \mid X$, we have

$$E_{\mathcal{P}_x^O}(Y^{D_N}) = E_{\mathcal{P}_x^N}(Y^{D_N}) \text{ and } E_{\mathcal{P}_x^N}(Y^{D_O}) = E_{\mathcal{P}_x^O}(Y^{D_O}) \text{ for } \forall\ x.$$

- Treatment assignment $T \mid I$: treatment T is completely randomly assigned to units within each trial, thus, satisfies $(Y^{D_N}, Y^{B_N}, Y^{D_O}, Y^{B_O}) \perp T \mid I = \mathcal{T}_N$.

The variable that defines the two trials, I, is not randomly assigned to units. Instead, it's specified before involved units are selected, and actually determines how units are selected.

Covariates X play a similar role in both problems: create unit stratifications based on which the two subpopulations of interest (two arms or two trials) can be compared and matched. However, they function via different mechanisms in the two problems. In across-trial matching, because the trial assignment I is specified before covariates X are sampled, to guarantee that $(Y^c, Y^t) \perp I \mid X$, X should cover as many predicting factors of (Y^c, Y^t) as possible. In cross-arm matching, however, since the treatment assignment T can be better controlled, X only needs to contain variables given which the treatment T is conditionally randomly assigned.

1.4.4 The Bayesian approach

To avoid heavy notations, in this section, we pose the following two extra assumptions for indirect comparison: (1) the baseline drugs in the two trials are same (i.e., $B_N = B_O = B$), and (2) the orignal ignorable trial assignment assumption is enhanced to a slightly stronger version:

$$(Y^{D_N}, Y^{D_O}, Y^B) \perp I \mid X.$$

With these extra assumptions, the generic model for indirect comparison can be specified as follows:

$$\begin{aligned}
X \mid I &\sim \pi_{\theta_N}(x) \cdot I(I = \mathcal{T}_N) + \pi_{\theta_O}(x) \cdot I(I = \mathcal{T}_O), \\
T \mid I &\sim Bernoulli(\rho_N) \cdot I(I = \mathcal{T}_N) + Bernoulli(\rho_O) \cdot I(I = \mathcal{T}_O), \\
Y^{D_N} \mid X = x &\sim Bernoulli(R_N(x)), \\
Y^{D_O} \mid X = x &\sim Bernoulli(R_O(x)), \\
Y^{B} \mid X = x &\sim Bernoulli(R_B(x)),
\end{aligned}$$

where π_θ is a parametric distribution with unknown parameter θ, ρ_N and $\rho_O \in (0, 1)$, and the three response surfaces $R_N(x)$, $R_O(x)$ and $R_B(x)$ are specified as probit models below:

$$\begin{aligned}
R_N(x) &= \Phi(x'\alpha_N), \\
R_O(x) &= \Phi(x'\alpha_O), \\
R_B(x) &= \Phi(x'\alpha_B).
\end{aligned}$$

The above model lead to the following likelihood for the observed data:

$$
\begin{aligned}
f(X,T,Y^{obs} \mid I) &= \prod_{i\in\mathcal{T}_N} f(X_i,T_i,Y_i^{obs} \mid I=\mathcal{T}_N)\cdot\prod_{i\in\mathcal{T}_O} f(X_i,T_i,Y_i^{obs} \mid I=\mathcal{T}_O) \\
&= \prod_{i\in\mathcal{T}_N}\pi_{\theta_N}(X_i)\cdot\prod_{i\in\mathcal{T}_O}\pi_{\theta_O}(X_i)\cdot\rho_N^{n_N^+}(1-\rho_N)^{n_B^+}\cdot\rho_O^{n_O^-}(1-\rho_O)^{n_B^-} \\
&\quad\cdot\prod_{i\in\mathcal{A}_N^+}R_N(X_i)^{Y_i^{obs}}[1-R_N(X_i)]^{1-Y_i^{obs}} \\
&\quad\cdot\prod_{i\in\mathcal{A}_O^-}R_O(X_i)^{Y_i^{obs}}[1-R_O(X_i)]^{1-Y_i^{obs}} \\
&\quad\cdot\prod_{i\in\mathcal{A}_B^+\cup\mathcal{A}_B^-}R_B(X_i)^{Y_i^{obs}}[1-R_B(X_i)]^{1-Y_i^{obs}}.
\end{aligned}
$$

Giving the model parameters $(\theta_N,\theta_O,\rho_N,\rho_O,\alpha_N,\alpha_O,\alpha_B)$ a proper prior distribution, their posterior distribution can be derived by standard Bayesian inference techniques. Note that if $(\theta_N,\theta_O,\rho_N,\rho_O,\alpha_N,\alpha_O,\alpha_B)$ are independent of each other in the prior distribution, they will be still independent in the posterior distribution since they are separated in the above likelihood function.

Because the estimand

$$
ACE_{\pi_{\theta_O}}(D_N,D_O)=\int\big[R_N(x)-R_O(x)\big]d\pi_{\theta_O}(x)=h(\alpha_N,\theta_O,\alpha_B)
$$

is a function of parameters $(\alpha_N,\alpha_P,\theta_O)$, the posterior distribution of $ACE_{\pi_{\theta_O}}(D_N,D_O)$ can be naturally derived from the posterior distribution of $(\alpha_N,\theta_O,\alpha_B)$.

1.4.5 Match unit populations by weighting adjustment

Matching methods developed for balancing the treated and controlled arms of a non-randomized experiment can be naturally applied to this more complicated scenarios with two trials and four arms.

Totally, there are six unit populations in the system: two trial-level populations π_N and π_O for $\mathcal{T}_N$ and $\mathcal{T}_O$, and four arm-level populations π_N^t, π_N^c, π_O^t, π_O^c for $\mathcal{A}_N^+$, $\mathcal{A}_P^+$, $\mathcal{A}_O^-$ and $\mathcal{A}_P^-$, respectively. Since we have assumed that both trials $\mathcal{T}_N$ and $\mathcal{T}_O$ are perfectly randomized, we have

$$\pi_N^t = \pi_N^c = \pi_N,\ \pi_O^t = \pi_O^c = \pi_O.$$

In the across-trial matching problem, we aim to estimate $ACE_{\mathcal{P}_O}(D_N, D_O)$ by matching the two trial-level populations π_N and π_O.

By definition, we have

$$ACE_{\mathcal{P}_O}(D_N, D_O) \ = \ \int E_{\pi_O}(Y^{D_N} \mid X = x) d\pi_O(x) - E_{\pi_O}(Y^{D_O} \mid T = D_O).$$

Based on the assumptions we make for the across-trial matching problem, we have

$$\begin{aligned}
& \int E_{\pi_O}(Y^{D_N} \mid X = x) d\pi_O(x) \\
= & \int E_{\pi_N}(Y^{D_N} \mid X = x) \cdot \frac{\pi_O(x)}{\pi_N(x)} d\pi_N(x) \\
= & \int E_{\pi_N}(Y^{D_N} \mid X = x, T = t) \cdot \frac{\pi_O(x)}{\pi_N(x)} d\pi_N(x) \\
= & \int E_{\pi_N}(Y^{D_N} \cdot \frac{\pi_O(x)}{\pi_N(x)} \mid X = x, T = t) d\pi_N^t(x) \\
= & E_{\pi_N^t}(Y^{D_N} \cdot \frac{\pi_O(X)}{\pi_N(X)} \mid T = t).
\end{aligned}$$

Thus, the following estimate

$$\frac{1}{n_N^+} \sum_{i \in \mathcal{A}_N^+} Y_i^{obs} \cdot \omega_i - \frac{1}{n_O^-} \sum_{i \in \mathcal{A}_O^-} Y_i^{obs} \tag{1.21}$$

is an unbiased estimate of $ACE_{\mathcal{P}_O}(D_N, D_O)$, where

$$\omega_i = \frac{\pi_O(X_i)}{\pi_N(X_i)} = \frac{P(X = X_i \mid I = \mathcal{T}_O)}{P(X = X_i \mid I = \mathcal{T}_N)}.$$

Following the spirit of Section 1.3.4, it is easy to check that

$$\begin{aligned}
\omega_i &= \frac{e_i}{1 - e_i} \cdot \frac{1 - \alpha}{\alpha}, \\
e_i &= \frac{P(I = \mathcal{T}_N \mid X = X_i)}{P(I = \mathcal{T}_N \mid X = X_i) + P(I = \mathcal{T}_O \mid X = X_i)}, \\
\alpha &= \frac{P(I = \mathcal{T}_N)}{P(I = \mathcal{T}_N) + P(I = \mathcal{T}_O)}.
\end{aligned}$$

Here, e_i stands for the propensity score of trial assignment.

1.5 Indirect Comparison without Individual-Level Data

1.5.1 Problem setting

In practice, the situation can be further complicated as detailed data at patient level is often available in just one trial. For example, it's very often that the developer the new drug D_N have a full control for trial $\mathcal{T}_N$, but only a limited access to summary statistics (usually, the first one or two moments) of the old trial $\mathcal{T}_O$, which was carried out by another drug developer who keeps many details confidential due to various concerns. In this new setting, the observed data contain two components: (1) detailed data of the new trial $\{X_i, I_i, T_i, Y_i^{obs}\}_{i\in\mathcal{T}_N}$, (2) summary statistics of the old trial: $\{\bar{X}_{\mathcal{T}_O}, S_{\mathcal{T}_O}, n_O^-, n_B^-, m_O^-, m_B^-\}$, where $\bar{X}_{\mathcal{T}_O}$ and $S_{\mathcal{T}_O}$ are the sample mean and sample covariance matrix of the unobserved covariates X in trial $\mathcal{T}_O$, $n_O^- = \#(\mathcal{A}_O^-)$ and $n_B^- = \#(\mathcal{A}_B^-)$ are sizes (numbers of individuals) of $\mathcal{A}_O^-$ and $\mathcal{A}_B^-$, and $m_O^- = \sum_{i\in\mathcal{A}_O^-} Y_i^{obs}$ and $m_B^- = \sum_{i\in\mathcal{A}_P^-} Y_i^{obs}$ numbers of patients with positive response in $\mathcal{A}_O^-$ and $\mathcal{A}_B^-$.

For this more challenging scenario, most classic matching methods, e.g., the nearest neighbor matching and substratification, fail due to the lack of detailed data at patient level.

1.5.2 The Bayesian and semi-Bayesian approaches

A full Bayesian inference of this problem can be achieved by following the framework of Section 1.4.4 and treating the unobserved $\{T_i, X_i, Y_i^{obs}\}_{i\in\mathcal{T}_O}$ as missing data. Samples from posterior distribution of unknown parameters $(\alpha_O, \alpha_N, \pi_O, \pi_N)$ can be

obtained using the data argumentation strategies (Tanner and Wong, 1987). However, this sampling strategy suffers from slow convergence because little information is available for parameter α_O of response surface

$$R_O(x) = \Phi(x'\alpha_O).$$

To overcome the limitation of Bayesian method, we propose a semi-Bayesian approach below. In the semi-Bayesian approach, we choose to ignore part of the observed data to simplify the Bayesian inference. First, we ignore the detailed generating mechanism of potential outcome Y^{D_O} where the covariates X are involved in, i.e.,

$$Y^{D_O} \mid X = x \sim Bernoulli(R_O(x)).$$

Instead, we directly model the overall statistical property of Y^{D_O} in unit population $\mathcal{P}_O$. To be concrete, we assume that for $\forall\ i \in \mathcal{P}_O$,

$$Y_i^{D_O} \sim Bernoulli(r_O),$$

where $r_O \in [0, 1]$ is the overall probability of $Y^{D_O} = 1$ in population $\mathcal{P}_O$. Mathematically, it's easy to check that

$$r_O = \int R_O(x) d\pi_O(x).$$

Second, we choose to work on $(\{X_i\}_{i\in\mathcal{T}_N\cup\mathcal{T}_O}, \{Y_i^{obs}\}_{i\in\mathcal{A}_N^+\cup\mathcal{A}_O^-})$ only, responses of baseline drugs $\{Y_i^{obs}\}_{i\in\mathcal{A}_B^+\cup\mathcal{A}_B^-}$ are ignored as they do not provide useful information

to the casual effect of interest. Under the simplified model, the generating mechanism of the selected data becomes:

$$\begin{aligned} X \mid I &\sim \pi_{\theta_N}(x) \cdot I(I = \mathcal{T}_N) + \pi_{\theta_O}(x) \cdot I(I = \mathcal{T}_O), \\ U &\sim N(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p, 1), \\ Y^{D_N} &= I(U > 0), \\ Y^{D_O} &\sim Bernoulli(r_O). \end{aligned}$$

Given prior $\pi(\theta_N, \theta_O, r_O, \alpha_0, \cdots, \alpha_p)$ to the model parameters, our goal is to draw samples (via Gibbs sampling) from the posterior below

$$\begin{aligned} & f(\theta_N, \theta_O, r_O, \alpha_0, \cdots, \alpha_p, U \mid \{X_i\}_{i \in \mathcal{T}_N \cup \mathcal{T}_O}, \{Y_i^{obs}\}_{i \in \mathcal{A}_N^+ \cup \mathcal{A}_O^-}) \\ \propto\ & f(\{X_i\}_{i \in \mathcal{T}_N \cup \mathcal{T}_O}, \{Y_i^{obs}\}_{i \in \mathcal{A}_N^+ \cup \mathcal{A}_O^-}, U \mid \theta_N, \theta_O, r_O, \alpha_0, \cdots, \alpha_p) \cdot \pi(\theta_N, \theta_O, r_O, \alpha_0, \cdots, \alpha_p) \\ =\ & \prod_{i \in \mathcal{T}_N} \pi_{\theta_N}(X_i) \cdot \prod_{i \in \mathcal{T}_O} \pi_{\theta_O}(X_i) \cdot \prod_{i \in \mathcal{A}_N^+} f(Y_i^{obs} \mid U_i) f(U_i \mid \alpha, X_i) \cdot r_O^{m_B^-} (1 - r_O)^{n_B^- m_B^-} \\ & \cdot \pi(\theta_N, \theta_O, r_O, \alpha_0, \cdots, \alpha_p). \end{aligned} \tag{1.22}$$

The details of Gibbs sampling when X follows multivariate Gaussian is given in Appendix A.1.

1.5.3 The Signorovitch's method of weighting adjustment

Recently, Signorovitch et al. (2010) proposed *Matching Adjusted Indirect Comparison* (MAIC), a method based on weighting adjustment, to tangle this challenging

problem. Recall the result of section 1.4.5:

$$\frac{1}{n_N^+}\sum_{i\in\mathcal{A}_N^+} Y_i^{obs}\cdot\omega_i - \frac{1}{n_O^-}\sum_{i\in\mathcal{A}_O^-} Y_i^{obs}$$

is an unbiased estimate of $ACE_{\pi_O}(D_N, D_O)$, where

$$\begin{aligned}
\omega_i &= \frac{e_i}{1-e_i}\cdot\frac{1-\alpha}{\alpha}, \\
e_i &= \frac{P(I=\mathcal{T}_N \mid X=X_i)}{P(I=\mathcal{T}_N \mid X=X_i)+P(I=\mathcal{T}_O \mid X=X_i)}, \\
\alpha &= \frac{P(I=\mathcal{T}_N)}{P(I=\mathcal{T}_N)+P(I=\mathcal{T}_O)}.
\end{aligned}$$

Without loss of generality, we assume that trail assignment is balanced, i.e.,

$$P(I=\mathcal{T}_N) = P(I=\mathcal{T}_O), \text{ or more specifically, } n_N = n_O.$$

Thus, we have $\alpha = 0.5$, and ω_i only depends on e_i. Unfortunately, since only the first two moments of covariates X are available for arms $\mathcal{A}_O^-$ and $\mathcal{A}_P^-$, conditional probability $P(I=\mathcal{T}_N \mid X=x)$ cannot be directly estimated. Thus, the propensity score for trial assignment e_i is not identifiable until extra assumptions are made.

MAIC proposed by Signorovitch et al. (2010) avoids this dilemma by assuming a logistic model for the generalized propensity score e_i, which leads to the the following parametric form of the weight ω:

$$\omega_i(\beta) = \frac{e_i}{1-e_i} = \exp(A+\beta' x_i), \tag{1.23}$$

where $\beta = (\beta_1, \cdots, \beta_p)'$ are parameters to be specified, and A is a normalizing constant satisfying

$$\sum_{i \in \mathcal{A}_N^+} \omega_i(\beta) = \sum_{i \in \mathcal{A}_N^+} \exp(A + \beta' x_i) = n_N^+.$$

Once β is specified, we can calculate the weights $\{\omega_i\}_i$ accordingly, and adjust π_N, the patient distribution of trial $\mathcal{T}_N$, to its weighted version

$$\pi_N^*(\beta) = \big\{\big(x, \omega_x(\beta)\big) : x \sim \pi_N\big\}.$$

The authors argue that a proper β should lead to a same mean vector for π_O and $\pi_N^*(\beta)$, i.e.,

$$E_{\pi_O} X = E_{\pi_N}\big(X \cdot \omega_X(\beta)\big).$$

With finite samples, the above constrain leads to the following estimation equation for parameter $\beta = (\beta_1, \cdots, \beta_p)$:

$$\frac{1}{n_O} \sum_{i \in \mathcal{T}_O} X_i = \frac{1}{n_N} \sum_{i \in \mathcal{T}_N} \big(X_i \cdot \omega_i(\beta)\big), \tag{1.24}$$

which can be solved by iterative methods, such as the Newton-Raphson algorithm. As β has been specified, $\{\omega_i\}_{i \in \mathcal{A}_N^+}$ are fixed numbers. Signorovitch et al. (2010) claim that the standard error of estimator

$$\hat{r}_N = \frac{1}{n_N^+} \sum_{i \in \mathcal{A}_N^+} Y_i^{obs} \cdot \omega_i$$

can be "derived by" (estimated by) a *sandwich estimator* (Liang and Zeger, 1986)

$$\frac{1}{(\sum_{i\in\mathcal{A}_{N+}}\omega_i)^2}\sum_{i\in\mathcal{A}_N^+}\omega_i^2\cdot(Y_i-\hat{r}_N)^2.$$

This can be understood as *effective sample size* of weighted samples $\{x_i,\omega_i(\beta)\}_{i\in\mathcal{A}_N^+}$ with respect to the estimate given in (1.21) is $\frac{(\sum_{i\in\mathcal{A}_{N+}}\omega_i)^2}{\sum_{i\in\mathcal{A}_{N+}}\omega_i^2}$, based on which the standard error of the estimate can be easily calculated.

The correctness of the MAIC method can be shown by the following theorem:

Theorem 1 *If the logistic model for propensity score of trial assignment, i.e., equation (1.23), is correct, MAIC gives a consistent and asymptotically unbiased estimate of* $ACE_{\pi_O}(t,c)$ *with probability one when both* n_O *and* n_T *go to infinity.*

MAIC provides a point estimation of the estimand $ACE_{\pi_O}(D_N,D_O)$ with a parametric assumption on the propensity score of trial assignment. Although the method is straightforward, it has the following limitations:

- The logistic model for propensity score of trial assignment, i.e., equation (1.23), cannot be verified in a practical problem.
- The method matches the first moment of covariates across the two trials, but ignores the potential impact of the correlation structures of these covariates.
- The method, which matches all available covariates, tends to overmatch the data and lead to loss of statistical efficiency. For example, if a covariate is irrelevant to the response or is already balanced across the two trials except for random noise, matching it will just introduce noise into the system.

- Solving the estimation equation (1.24) involves numerical iterations, which are time costuming and may result in no numerical solution in some cases.

- The uncertainty (i.e., variance) of the proposed estimator is not properly evaluated: the effective sample size claimed by the authors is not accurate and may lead to a wrong variance estimation of the proposed estimator.

1.5.4 A novel approach

To overcome the limitations of Signorovitch's method, we propose the following novel approach. In the new approach, we directly estimate the distribution of covariates X in the two trials $\pi_O(x)$ and $\pi_N(x)$, and assign weights to data points from trial $\mathcal{T}_N$ based on the estimated covariates distributions. We will show that the new approach results in a consistent estimator as well, and overcomes many limitations of the Signorovitch's method.

To be concrete, we assume that distribution of covariates X in the two trials, $\pi_O(x)$ and $\pi_N(x)$, share the same parametric form $\pi_\theta(x)$, i.e.,

$$\pi_O(x) = \pi_{\theta_O}(x) \quad \text{and} \quad \pi_N(x) = \pi_{\theta_N}(x),$$

where θ_O and θ_N are unknown parameters whose values can be inferred from the data. Once θ_O and θ_N are specified, it's straightforward to see that the following weight assignment to unit $i \in \mathcal{T}_N$ leads to a proper weight adjustment mechanism

$$\omega_i = \frac{\pi_{\theta_O}(X_i)}{\pi_{\theta_N}(X_i)}. \tag{1.25}$$

In practice, θ_N and θ_O are unknown, and need to be estimated from the observed covariates $\{X_i\}_{i\in\mathcal{T}_N}$ and summary statistics $\{\bar{X}_{\mathcal{T}_O}, S_{\mathcal{T}_O}, n_O^-, n_P^-, m_O^-, m_P^-\}$. Let $\hat{\theta}_O$ and $\hat{\theta}_N$ be proper estimates of θ_O and θ_N. Replacing the unknown θ_O and θ_N with their estimates $\hat{\theta}_O$ and $\hat{\theta}_N$, we get the following estimated weights

$$\hat{\omega}_i = \frac{\pi_{\hat{\theta}_O}(X_i)}{\pi_{\hat{\theta}_N}(X_i)}, \tag{1.26}$$

which leads to the estimate below

$$\frac{1}{n_N^+} \sum_{i\in\mathcal{A}_N^+} Y_i \cdot \hat{\omega}_i - \bar{Y}_{\mathcal{A}_O^-}. \tag{1.27}$$

It can be showed that (1.27) is a proper estimator under certain conditions.

Theorem 2 *If $\hat{\theta}_O$ and $\hat{\theta}_N$ are consistent estimates of θ_O and θ_N, (1.27) is a consistent and asymptotically unbiased estimate of $ACE_{\pi_O}(t, c)$.*

Theorem 2 requires that consistent estimates of θ_O and θ_N can be obtained from the observed data for covariates: $\{X_i\}_{i\in\mathcal{T}_N}$ and $\{\bar{X}_{\mathcal{T}_O}, S_{\mathcal{T}_O}, n_O^-, n_P^-, m_O^-, m_P^-\}$. For θ_N, this requirement is easy to satisfy as long as the assumed model π_θ is identifiable, since iid samples from π_{θ_N}, i.e., $\{X_i\}_{i\in\mathcal{T}_N}$, are available. For θ_O, however, this requirement is equivalent to assuming that the underlying distribution π_θ is completely determined by its first two moments. Therefore, we have the following corollary immediately:

Corollary 1 *If the assumed distribution of covariates π_θ cannot be fully determined by its first two moments, there does not exist a consistent estimate of θ_O, and thus, no consistent and asymptotically unbiased estimate of $ACE_{\pi_O}(t, c)$ can be obtained*

until more assumptions are made.

To distinguish this novel approach from the original MAIC approach, in the following of this dissertation, we refer to the novel approach as MAIC_N.

1.5.5 Estimating population distributions π_{θ_O} and π_{θ_N}

Theorem 2 and Corollary 1 put quite strong constraint on π_θ, the model of covariates. Fortunately, many widely used distributions (such as normal, lognormal, exponential, beta, Poisson and so on) do satisfy this constraint. For example, assume that the covariates X come from a multi-normal distribution, i.e.,

$$X_i \sim N(\mu_N, \Sigma_N) \cdot I(i \in \mathcal{T}_N) + N(\mu_O, \Sigma_O) \cdot I(i \in \mathcal{T}_O).$$

Here, $\theta_N = (\mu_N, \Sigma_N)$ and $\theta_O = (\mu_O, \Sigma_O)$. Given the observed data for covariates: $\{X_i\}_{i \in \mathcal{T}_N}$ and $\{\bar{X}_{\mathcal{T}_O}, S_{\mathcal{T}_O}, n_O^-, n_P^-, m_O^-, m_P^-\}$, it's easy to see that the MLEs of θ_N and θ_O are:

$$\hat{\theta}_N = (\hat{\mu}_N, \hat{\Sigma}_N) \quad \text{and} \quad \hat{\theta}_O = (\hat{\mu}_O, \hat{\Sigma}_O),$$

where

$$\begin{aligned}
\hat{\mu}_N &= \bar{X}_{\mathcal{T}_N} = \frac{1}{n_N}\sum_{i\in\mathcal{T}_N} X_i, \\
\hat{\Sigma}_N &= S_{\mathcal{T}_N} = \frac{1}{n_N}\sum_{i\in\mathcal{T}_N} (X_i - \bar{X}_{\mathcal{T}_N})(X_i - \bar{X}_{\mathcal{T}_N})^T; \\
\hat{\mu}_O &= \bar{X}_{\mathcal{T}_O} = \frac{1}{n_O}\sum_{i\in\mathcal{T}_O} X_i, \\
\hat{\Sigma}_O &= S_{\mathcal{T}_O} = \frac{1}{n_O}\sum_{i\in\mathcal{T}_O} (X_i - \bar{X}_{\mathcal{T}_O})(X_i - \bar{X}_{\mathcal{T}_O})^T.
\end{aligned}$$

In some cases, for trial $\mathcal{T}_O$, the sample covariance matrices $S_{\mathcal{T}_O}$ is not available. Instead, only the sample variances of covariates, i.e., the diagonal elements of $S_{\mathcal{T}_O}$, are given. In this scenario, we will assume that the covariates X share the same correlation structure in the two trials $\mathcal{T}_O$ and $\mathcal{T}_N$ to avoid the identifiability problem. Since the variance estimates are orthogonal to the estimate of correlation structure, we can estimate the common correlation structure based on $\mathcal{T}_N$, and plug it into the trial $\mathcal{T}_O$.

In practice, the problem can be further simplified if the p covariates are assumed to be independent of each other. For example, assume that the p covariates come from independent Gaussian distributions, i.e., for $\forall\ j \in \{1, \cdots, p\}$,

$$X_{i,j} \sim N(\mu_j, \sigma_j^2) \cdot I(i \in \mathcal{T}_N) + N(\nu_j, \kappa_j^2) \cdot I(i \in \mathcal{T}_O).$$

Here, $\theta_N = \{\mu_j, \sigma_j^2\}_{j=1}^p$ and $\theta_O = \{\nu_j, \kappa_j^2\}_{j=1}^p$. Given the observed data for covariates: $\{X_i\}_{i\in\mathcal{T}_N}$ and $\{\bar{X}_{\mathcal{T}_O}, S_{\mathcal{T}_O}, n_O^-, n_P^-, m_O^-, m_P^-\}$, it's easy to see that the MLEs of θ_N and

θ_O are:

$$\hat{\theta}_N = \{\hat{\mu}_j, \hat{\sigma}_j^2\}_{j=1}^p \quad \text{and} \quad \hat{\theta}_O = \{\hat{\nu}_j, \hat{\kappa}_j^2\}_{j=1}^p,$$

where

$$\begin{aligned}
\hat{\mu}_j &= \bar{X}_{\mathcal{T}_N}(j) = \frac{1}{n_N} \sum_{i \in \mathcal{T}_N} X_{i,j}, \\
\hat{\sigma}_j^2 &= S_{\mathcal{T}_N}(j,j) = \frac{1}{n_N} \sum_{i \in \mathcal{T}_N} (X_{i,j} - \hat{\mu}_j)^2; \\
\hat{\nu}_j &= \bar{X}_{\mathcal{T}_O}(j) = \frac{1}{n_O} \sum_{i \in \mathcal{T}_O} X_{i,j}, \\
\hat{\kappa}_j^2 &= S_{\mathcal{T}_O}(j,j) = \frac{1}{n_O} \sum_{i \in \mathcal{T}_O} (X_{i,j} - \hat{\nu}_j)^2.
\end{aligned}$$

If covariates X comes from other distributions instead of normal distribution, θ_N and θ_O can be estimated in a similar way. For cases where MLEs of θ_N and θ_O are difficult to get, we can use moment estimation instead. Compared to the MAIC method, MAIC$_N$ enjoys the following advantages:

- The model assumption can be (at least partially) verified in a practical problem;
- Both the marginal distributions and the correlation structures of covariates are considered;
- The computation involved is straightforward.

1.5.6 Variance Estimation via bootstrap

To derive the (asymptotic) variance of the estimate obtained by MAIC or MAIC$_N$ theoretically, a joint model for (X, Y^{D_N}) must be specified for units in $\mathcal{A}_N^+$ to describe

the dependence structure of X and Y^{D_N}. (In general, the weights $\{\omega_i\}_{i\in\mathcal{A}_N^+}$, which are functions of $\{X_i\}_{i\in\mathcal{A}_N^+}$, are depend of the potent outcomes $\{Y_i\}_{i\in\mathcal{A}_N^+}$.) Considering that in practice, the joint model for (X, Y^{D_N}) is often unknown or difficult to specify, we propose to achieve estimate variance using the bootstrap distribution generated by the the following bootstrap procedure:

Step 1. Resample individual-level data points in trial $\mathcal{T}_N$ by non-parametric bootstrap;

Step 2. Regenerate samples of covariates in $\mathcal{T}_O$ by parametric bootstrap, i.e., draw n_O i.i.d. samples from $\pi_{\hat{\theta}_O}$, calculate the summary statics (e.g., mean and covariance matrix) of the resampled data;

Step 3. Run MAIC or MAIC$_N$ for the resampled data set (individual-level data in trial $\mathcal{T}_N$ and summary statics of covariates in trial $\mathcal{T}_O$) to get an estimation of the relative efficacy of D_N with respect to D_O;

Step 4. Repeat above steps for K times to get the bootstrap distribution of the corresponding estimate.

1.5.7 Hybrid Bayesian inference with bootstrap distributions

Another strategy to achieve semi-Bayesian inference is to hybrid Bayesian inference with bootstrap distributions. The algorithm goes s follows:

- Step 1. Draw M realizations of θ_N and θ_O from the following "posterior distri-

butions"

$$\begin{aligned} f(\theta_N \mid \{X_i\}_{i\in\mathcal{T}_N}) &\propto \pi(\theta_N)\cdot f(\{X_i\}_{i\in\mathcal{T}_N} \mid \theta_N), \\ f(\theta_O \mid \bar{X}_{\mathcal{T}_O}, S_{\mathcal{T}_O}) &\propto \pi(\theta_o)\cdot f(\bar{X}_{\mathcal{T}_O}, S_{\mathcal{T}_O} \mid \theta_O), \end{aligned}$$

where $\pi(\theta_N)$ and $\pi(\theta_O)$ are prior distributions.

- Step 2. For the m-th realization of (θ_N, θ_O), generate K bootstrap samples for units in $\mathcal{A}_N^+$, denoted as $\{X_i^{(1)}, Y_i^{(1)}\}_{i=1}^{n_N^+}, \cdots, \{X_i^{(K)}, Y_i^{(K)}\}_{i=1}^{n_N^+}$, by non-parametric bootstrap. For the k-th bootstrap sample $\{X_i^{(k)}, Y_i^{(k)}\}_{i=1}^{n_N^+}$, calculate the point estimate

$$\widehat{ACE}_k = \frac{1}{n_N^+}\sum_{i=1}^{n_N^+} Y_i^{(k)} \cdot \frac{\pi_{\theta_N}(X_i^{(k)})}{\pi_{\theta_O}(X_i^{(k)})}. \tag{1.28}$$

  $\hat{F}_m \triangleq \{\widehat{ACE}_1, \cdots, \widehat{ACE}_K\}$ forms the bootstrap distribution of the estimate for $E_{\mathcal{P}_O} Y^{D_N}$ given m-th realization of (θ_N, θ_O).

- Step 3. Pool the bootstrap distributions from different realizations of (θ_N, θ_O) together, i.e., define

$$\hat{F} = \bigcup_{m=1}^{M} \hat{F}_m. \tag{1.29}$$

  Treat $\hat{F}$ as the "posterior distribution" of $E_{\mathcal{P}_O} Y^{D_N}$.

For many practical problems, Step 1 is not trivial. Appendix A.2 provides the details of Step 1 when covariates X follows multivariate Gaussian distribution.

1.6 Selecting Covariates to be Matched

In practice, it's not wise to match all available covariates across the two trials. For example, if one of the following two scenarios happens for a covariate X_j, we should not match it:

Scenario 1. The distribution difference of X_j across the two trials is due to random noise;

Scenario 2. X_j has no impact to potential outcomes Y^t and Y^c given the other covariates.

To exclude scenario 1, we propose the following screening procedure. For each covariate X_j, test the hypotheses below:

$$H_0 : X_j \text{ follows the same distribution in } \mathcal{T}_N \text{ and } \mathcal{T}_O,$$
$$H_1 : X_j \text{ follows different distributions in } \mathcal{T}_N \text{ and } \mathcal{T}_O.$$

If H_0 is rejected for X_j for a pre-given significant level (e.g., 0.05), put X_j into the candidate set for further investigation; otherwise, remove X_j from consideration. When unit level data are available in both trials, The above hypothesis test can be achieved parametrically (when parametric model of X_j is known) or non-parametrically (when parametric model of X_j is unknown). When unit level data are available only in trial $\mathcal{T}_N$, however, non-parametric test becomes difficult. In this case, we may want to specify a parametric model for X_j and do parametric test. For example, if X_j follows Gaussian distribution in both trials, the above test can be achieved by testing (1)

whether X_j has the same variance across $\mathcal{T}_N$ and $\mathcal{T}_O$, and (2) whether X_j has the same mean across $\mathcal{T}_N$ and $\mathcal{T}_O$.

Covariates in the candidate set generated in the previous step will be further investigated to exclude scenario 2. Statistically, this is equivalent to selecting the smallest subset of covariates from the candidate set given which all other covariates are independent of Y^{D_N} and Y^{D_O}. Logistic regression, which models the conditional distribution of Y^{D_N} or Y^{D_O} given covariates X, is a popular tool for this purpose. Based on the data from arm $\mathcal{A}_N^+$, a subset of covariates denoted as $\mathcal{X}_N$ that have the best prediction power for Y^{D_N} can be selected using *Best Subset Regression* or *Stepwise Regression* strategy to minimized AIC (Akaike Information Criteria) or BIC (Bayesian Information Criteria) score. Similarly, we can select a subset $\mathcal{X}_{B_N}$ for Y^{B_N} based on arm $\mathcal{A}_B^+$, a subset $\mathcal{X}_{B_O}$ for Y^{B_O} based on arm $\mathcal{A}_B^-$, and a subset $\mathcal{X}_O$ for Y^{D_O} based on arm $\mathcal{A}_O^-$. To properly compare the efficacy of D_N and D_O, we only need to match covariates in $\mathcal{X}_N \cup \mathcal{X}_O$.

When individual-level data in $\mathcal{A}_O^-$ and $\mathcal{A}_B^-$ are unavailable, however, $\mathcal{X}_O$ and $\mathcal{X}_{B_O}$ cannot be identified. In this case, if it's acceptable to assume that $\mathcal{X}_O = \mathcal{X}_N$ or $\mathcal{X}_O \subseteq \mathcal{X}_N \cup \mathcal{X}_{B_N}$, we can match covariates in $\mathcal{X}_N$ or $\mathcal{X}_N \cup \mathcal{X}_{B_N}$ instead.

1.6.1 Variable selection via a joint Bayesian analysis

In this subsection, we propose to check the two criteria simultaneously by a joint Bayesian analysis. To simplify the problem, we assume that $\mathcal{X}_O \subseteq \mathcal{X}_N$. Thus, to judge whether covariate X_j should be matched, we only need to answer the following two questions:

Question 1. Whether the distribution difference of X_j across $\mathcal{T}_N$ and $\mathcal{T}_O$ is due to random noise?

Question 2. whether X_j has no impact to Y^{D_N} given the other covariates?

Based on answers to the two questions, the p covariates can be divided into four groups:

- G_1: covariates that say No to both question 1 and question 2;
- G_2: covariates that say Yes to question 1 but No to question 2;
- G_3: covariates that say No to question 1 but Yes to question 2;
- G_4: covariates that say Yes to both question 1 and question 2.

Clearly, only covariates in G_1 need to be matched.

Let J_j be the group indicator of covariate X_j, where $J_j = 1$ if $X_j \in G_1$, $J_j = 2$ if $X_j \in G_2$, $J_j = 3$ if $X_j \in G_3$, and $J_j = 4$ if $X_j \in G_4$. Assuming that the p covariates are independent of each, we propose to work on the following likelihood function:

$$
\begin{aligned}
& f(\{X_i\}_{i\in\mathcal{T}_N\cup\mathcal{T}_O}, \{Y_i^{obs}\}_{i\in\mathcal{A}_N^+} \mid J, \Theta) \\
= & \prod_{j\in G_1\cup G_3} \Big[\prod_{i\in\mathcal{T}_N} f(X_{i,j} \mid \theta_{N,j}) \prod_{i\in\mathcal{T}_O} f(X_{i,j} \mid \theta_{O,j}) \Big] \cdot \prod_{j\in G_2\cup G_4} \prod_{i\in\mathcal{T}_N\cup\mathcal{T}_O} f(X_{i,j} \mid \theta_j) \\
& \cdot \prod_{i\in\mathcal{A}_N^+} f(Y_i^{obs} \mid \beta' X_{i,G_1\cup G_2}),
\end{aligned}
$$

where the model parameters $\Theta = (\theta_N, \theta_O, \theta, \beta)$ are defined as:

$$\begin{aligned}
\theta_N &= (\theta_{N,1}, \cdots, \theta_{N,p}) \text{ are the } \mathcal{T}_N\text{-specific parameters,} \\
\theta_O &= (\theta_{O,1}, \cdots, \theta_{O,p}) \text{ are the } \mathcal{T}_O\text{-specific parameters,} \\
\theta &= (\theta_1, \cdots, \theta_p) \text{ are the common parameters across two trials,} \\
\beta &= \{\beta_0\} \cup \{\beta_j\}_{j \in G_1 \cup G_2} \text{ are coefficients in probit regression of } Y_i.
\end{aligned}$$

Giving (J, Θ) a proper prior distributions $\pi(J, \Theta)$, samples from posterior distribution below can be obtained by Gibbs sampling

$$f(J, \Theta \mid \{X_i\}_{i \in \mathcal{T}_N \cup \mathcal{T}_O}, \{Y_i^{obs}\}_{i \in \mathcal{A}_N^+}) \propto \pi(J, \Theta) \cdot f(\{X_i\}_{i \in \mathcal{T}_N \cup \mathcal{T}_O}, \{Y_i^{obs}\}_{i \in \mathcal{A}_N^+} \mid J, \Theta).$$

Note that since the dimension of β, which equals to $\#(G_1 \cup G_2)$, changes with J, we need to use advanced MCMC techniques (e.g., *reversible jump MCMC*) to guarantee the convergence of the sampler. A comprehensive review of these techniques can be found in Liu (2001).

With the same assumption that the p covariates are independent of each, we can also achieve variable selection by extending the naive Bayes method. To be concrete,

we propose to the following inverse model of $\{X_i\}_{i\in\mathcal{T}_N\cup\mathcal{T}_O}$ when $\{Y_i^{obs}\}_{i\in\mathcal{A}_N^+}$ are given:

$$
\begin{aligned}
& f(\{X_i\}_{i\in\mathcal{T}_N\cup\mathcal{T}_O} \mid \{Y_i^{obs}\}_{i\in\mathcal{A}_N^+}; J, \Theta) \\
= \; & \cdot \prod_{j\in G_1} \Big[\prod_{i\in\mathcal{T}_O} f(X_{i,j} \mid \theta_{O,j}) \prod_{i\in\mathcal{A}_N^+} f(X_{i,j} \mid \theta_{N,j}) \prod_{i\in\mathcal{A}_N^+} f(X_{i,j} \mid \theta_j^+)^{Y_i^{obs}} f(X_{i,j} \mid \theta_j^-)^{1-Y_i^{obs}} \Big] \\
& \cdot \prod_{j\in G_2} \Big[\prod_{i\in\mathcal{T}_O\cup\mathcal{A}_N^+} f(X_{i,j} \mid \theta_j) \prod_{i\in\mathcal{A}_N^+} f(X_{i,j} \mid \theta_j^+)^{Y_i^{obs}} f(X_{i,j} \mid \theta_j^-)^{1-Y_i^{obs}} \Big] \\
& \cdot \prod_{j\in G_3} \Big[\prod_{i\in\mathcal{T}_O} f(X_{i,j} \mid \theta_{O,j}) \prod_{i\in\mathcal{T}_N} f(X_{i,j} \mid \theta_{N,j}) \Big] \cdot \prod_{j\in G_4} \prod_{i\in\mathcal{T}_O\cup\mathcal{T}_N} f(X_{i,j} \mid \theta_j),
\end{aligned}
$$

where the model parameters $\Theta = \{\theta_{N,j}, \theta_{O,j}, \theta_j^+, \theta_j^-, \theta_j\}_j$ are defined as follows:

$\theta_{N,j}$ is the $\mathcal{T}_N$-specific parameters for $\{X_{i,j}\}_{i\in\mathcal{T}_N}$,

$\theta_{O,j}$ is the $\mathcal{T}_O$-specific parameters for $\{X_{i,j}\}_{i\in\mathcal{T}_O}$,

θ_j^+ is the $(Y_i^{obs}=1)$-specific parameters for $\{X_{i,j}\}_{i\in\mathcal{A}_N^+, Y_i^{obs}=1}$,

θ_j^- is the $(Y_i^{obs}=0)$-specific parameters for $\{X_{i,j}\}_{i\in\mathcal{A}_N^+, Y_i^{obs}=0}$,

θ_j is the common parameters for $\{X_{i,j}\}_{i\in\mathcal{T}_N\cup\mathcal{T}_O}$.

Let $\Theta_j = (\theta_{N,j}, \theta_{O,j}, \theta_j^+, \theta_j^-, \theta_j)$, and define

$$L_j(J_j, \Theta_j) = f(\{X_{i,j}\}_{i\in\mathcal{T}_N\cup\mathcal{T}_O} \mid \{Y_i^{obs}\}_{i\in\mathcal{A}_N^+}; J_j, \Theta_j),$$

the above model indicates that

$$f(\{X_i\}_{i\in\mathcal{T}_N\cup\mathcal{T}_O} \mid \{Y_i^{obs}\}_{i\in\mathcal{A}_N^+}; J, \Theta) = \prod_{j=1}^{p} L_j(J_j, \Theta_j),$$

and $L_j(J_j, \Theta_j)$ can take four different values below depending on the status of J_j:

$$
\begin{aligned}
J_j = 1: &\prod_{i \in \mathcal{T}_O} f(X_{i,j} \mid \theta_{O,j}) \prod_{i \in \mathcal{A}_B^+} f(X_{i,j} \mid \theta_{N,j}) \prod_{i \in \mathcal{A}_N^+} f(X_{i,j} \mid \theta_j^+)^{Y_i^{obs}} f(X_{i,j} \mid \theta_j^-)^{1-Y_i^{obs}}, \\
J_j = 2: &\prod_{i \in \mathcal{T}_O \cup \mathcal{A}_B^+} f(X_{i,j} \mid \theta_j) \prod_{i \in \mathcal{A}_N^+} f(X_{i,j} \mid \theta_j^+)^{Y_i^{obs}} f(X_{i,j} \mid \theta_j^-)^{1-Y_i^{obs}}, \\
J_j = 3: &\prod_{i \in \mathcal{T}_O} f(X_{i,j} \mid \theta_{O,j}) \prod_{i \in \mathcal{T}_N} f(X_{i,j} \mid \theta_{N,j}), \\
J_j = 4: &\prod_{i \in \mathcal{T}_O \cup \mathcal{T}_N} f(X_{i,j} \mid \theta_j).
\end{aligned}
$$

When J_j jumps from one status to another status, $L_j(J_j, \Theta_j)$ changes accordingly. Giving J a non-informative uniform prior on the sampling space $\{1, 2, 3, 4\}^p$ and Θ_j a prior $\pi(\Theta_j)$, the posterior distribution of J_j is completely determined by vector

$$L_j \triangleq \big(L_j(1), L_j(2), L_j(3), L_j(4)\big).$$

where

$$L_j(J_j) = \int L_j(J_j, \Theta_j) d\pi(\Theta_j).$$

When X_{ij} follows normal, and $\pi(\Theta_j)$ is the conjugate prior, the value of L_j can be obtained analytically. In a practical problem, variable selection can be achieved by specifying J as its the posterior mode.

The Bayesian variable selection approach can also be integrated into the hybrid Bayesian approach proposed in Section 1.5.7 as follows:

- Step 1. Draw T samples $\mathcal{J} = \{J_1, \cdots, J_T\}$ from the posterior distribution of J.
- Step 2. For each $J \in \mathcal{J}$, denote the selected covariates in G_1 as $\mathcal{X}_J$. Run hybrid

Bayesian for the selected covariates in $\mathcal{X}_J$ to get hybrid-bootstrap samples of the estimand $ACE_{\mathcal{P}}(t, c)$ given J, which is denoted as $\hat{F}_J$.

- Step 3. Pool different versions of hybrid-bootstrap samples together, we get the following hybrid-bootstrap samples below

$$\hat{F} = \bigcup_{J \in \mathcal{J}} \hat{F}_J.$$

The $\hat{F}$ obtained in this way contains the uncertainty in variable selection, trial population estimation as well as data point sampling.

1.6.2 Variable screening via SIRI

When there are too many possible variables, the procedure could lead to heavy computation. In this subsection, we propose a step-wise variable screening strategy to solve this problem. Logistic regression, which models the conditional distribution of response Y given predictors X, is not the only solution to our problem. Recently, Jiang and Liu (2013) proposed a novel method named SIRI (*Sliced Inverse Regression with Interactions*) to solve the same problem, but, via inverse modeling (i.e., modeling the conditional distribution of predictors X given response Y). Both simulation studies and theoretical analysis show that SIRI can achieve variable selection effectively even when interactions among covariates are also considered and greatly avoids overfitting, which bothers logistic regression a lot.

To be concrete, SIRI partitions the p covariates into two groups G_0 and G_1, where the covariates in G_0 are independent of Y, while the covariates in G_1 are dependent

of Y. And, for simpleness, SIRI models the conditional distribution of covariates X given response Y with multi-normal distributions below:

$$\begin{aligned} X_{G_0} \quad &| \quad Y \sim N(\mu, \Sigma), \text{ and} \\ X_{G_1} \quad &| \quad Y \sim N(\mu_0, \Sigma_0) \cdot I(Y = 0) + N(\mu_1, \Sigma_1) \cdot I(Y = 1), \end{aligned}$$

where (μ, Σ), (μ_0, Σ_0) and (μ_1, Σ_1) are unknown parameters. In practice, the group partition G_0 and G_1 are also unknown. Let J_j be the group indictor of covariate X_j, where $J_j = 0$ if $X_j \in G_0$, and $J_j = 1$ if $X_j \in G_1$. SIRI aims to get the posterior distribution of $J = (J_1, \cdots, J_p)$ given the observed data $\{X_i, Y_i\}_i$.

In practice, when Bayesian inference is too time-consuming, we can simplify SIRI with a series of screening tests. For example, the selection of $\mathcal{X}_N$ via SIRI can be achieved in two steps:

Step1. For each covariate X_j in the candidate set, calculate the p-value P_j of the hypothesis test below

$$\begin{aligned} H_0 &: X_j \text{ is independent of } Y^{D_N} \text{ in } \mathcal{A}_N^+, \\ H_1 &: X_j \text{ is not independent of } Y^{D_N} \text{ in } \mathcal{A}_N^+. \end{aligned}$$

Rank the p covariates by their p-values decreasingly. Denote the ranked covariates as $X_{(1)}, X_{(2)}, \cdots, X_{(p)}$, and the ranked p-values as $P_{(1)}, \cdots, P_{(p)}$. Initialize $\mathcal{X}_N = \{X_{(1)}\}$ if $P_{(1)} < 0.05$, and let $\mathcal{X}_N = \emptyset$ otherwise.

Step 2. Scan through $X_{(1)}, X_{(2)}, \cdots, X_{(p)}$ until $X_{(p)}$ is processed. If $X_{(k)} \in \mathcal{X}_N$, move

to $X_{(k+1)}$; otherwise, test the hypotheses below by likelihood ratio test

$$H_0 : X_{(k)} \mid \mathcal{X}_N, Y^{D_N} \sim N(\mathcal{X}_N\beta, \sigma^2);$$

$$H_1 : X_{(k)} \mid \mathcal{X}_N, Y^{D_N} \sim N(\mathcal{X}_N\beta_0, \sigma^2) \cdot I(Y^{D_N} = 0) + N(\mathcal{X}_N\beta_1, \sigma^2) \cdot I(Y^{D_N} = 1),$$

where $\beta_0 \neq \beta_1$.

Let $L = \frac{P_{M_1}(X_{(k)}|\mathcal{X}_N, Y^{D_N})}{P_{M_0}(X_{(k)}|\mathcal{X}_N, Y^{D_N})}$ be the likelihood ratio statistics, it can be showed that $2L \sim \chi^2(q)$, where $q = \#\mathcal{X}_N$ is the number of covariates already selected. Update $\mathcal{X}_N$ with $\mathcal{X}_N \cup \{X_{(p)}\}$ if H_0 is rejected for $X_{(p)}$.

1.7 Simulation Studies

In this section, we will evaluate the performance of the following methods via simulation under different settings:

- BASE: Direct comparison without matching
- MAIC: Signovitch's matching adjusted indirect comparison method
- MAIC$_B$: MAIC with bootstrap
- MAIC$_N$: Novel matching adjusted indirect comparison method
- BM: Semi-Bayesian model with independent assumption
- BMIC: Hybrid semi-Bayesian method with bootstrap

For direct comparison without matching, we applied Bayesian model for each trial separately, assuming the two trials of interest share the same unit population even

though the actual unit populations are different.

1.7.1 Simulation 1

In this simulation study, we simulated 10 continuous covariates $X = (X_1, \ldots, X_{10})$ as population characteristics under two scenarios where population structure are same and different, respectively.

- Scenario 1A:
  - New trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} N(0.5, 1)$
  - Old trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} N(0.5, 1)$
- Scenario 1B:
  - New trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} N(0.5, 1)$
  - Old trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} N(-0.5, 1)$

Once obtained X, outcome Y was simulated from same probit model for both new drug trial and old drug:

$$P(Y^{D_O} = 1|X, T) = P(Y^{D_N} = 1|X, T) = \Phi(-2 + X\beta + 0.5 \cdot I_{\{T=t\}}),$$

where $\beta = (1, -0.5, 0.5, 0, 0, 0, 0, 0, 0, 0)^T$

Under each scenario above, different settings of sample size (n) were applied (100, 200, 300, 500). Within each trial, half of individuals were assigned to treatment arm and control arm, respectively. 1000 data sets were simulated for each setting of (X, Y, n). All methods were applied to the same 1000 datasets for each setting.

Table 1.1: Scenario 1A, data generating models for X are the same for both trials, sample size for each trial varied from 100 to 500.

	BASE	MAIC	MAIC_B	MAIC_N	BM	BMIC	Sample Size
% coverage [length]	95.5% [0.335]	94.9% [0.388]	98.7% [0.501]	97.6% [0.434]	92.8% [0.303]	98.1% [0.471]	
mean (SE)	-0.002 (0.087)	-0.003 (0.096)	-0.002 (0.094)	-0.003 (0.093)	0.010 (0.083)	-0.005 (0.091)	100
% coverage [length]	94.7% [0.239]	96.9% [0.257]	97.1% [0.269]	97.5% [0.272]	96.4% [0.227]	99.0% [0.304]	
mean (SE)	-0.002 (0.062)	-0.001 (0.060)	-0.001 (0.060)	-0.001 (0.060)	0.005 (0.055)	-0.002 (0.060)	200
% coverage [length]	95.9% [0.197]	97.0% [0.207]	97.9% [0.212]	97.6% [0.214]	97.2% [0.189]	98.6% [0.238]	
mean (SE)	0.002 (0.048)	0.002 (0.045)	0.002 (0.045)	0.002 (0.045)	0.004 (0.043)	0.001 (0.046)	300
% coverage [length]	94.2% [0.153]	96.2% [0.157]	96.5% [0.159]	96.4% [0.160]	96.9% [0.148]	97.8% [0.176]	
mean (SE)	0.000 (0.040)	0.000 (0.037)	0.000 (0.037)	0.000 (0.037)	0.001 (0.034)	-0.001 (0.037)	500

Based on the simulations, we calculate the practical coverage of confidence/credible intervals (CIs), the average length of CIs, the mean and standard error of the mean efficacy estimator. Note that the true efficacy difference is 0 as the efficacy simulation mechanisms are same over the two trials.

The results are summarized in Table 1.1 and Table 1.2. From Table 1.1, we observed that when the two populations are identical (thus, no matching is needed at the first place), all methods have comparable performance in terms of coverage and length. Additionally, as sample size goes larger, all methods provide more accurate coverage (closer to 95%) with shorter length. When sample size goes to 500, with and without Bootstrap procedure, MAIC provided almost same coverage and length. When the two population distributions are different, MAIC provides lower coverage than other methods, as shown in Table 1.2,

In both scenarios, BM method consistently provides more accurate coverage with shorter length. However, this benefit may rely on the fact that the model assumed by BM (i.e., the probit link function) is identical to the true model we used for simulation. To avoid this problem, in the following simulation studies, logistic link function was applied instead.

Table 1.2: Scenario 1B, data generating models for X are different for both trials, sample size for each trial varied from 100-500 .

	BASE	MAIC	MAIC_B	MAIC_N	BM	BMIC	Sample Size
% coverage [length]	43.2% [0.290]	89.4% [0.509]	88.0% [0.457]	93.4% [0.611]	78.4% [0.330]	88.8% [0.539]	
mean (SE)	0.158 (0.074)	0.105 (0.145)	0.122 (0.098)	0.052 (0.157)	0.057 (0.137)	0.045 (0.161)	100
% coverage [length]	14.9% [0.207]	86.2% [0.469]	92.3% [0.426]	90.6% [0.479]	90.9% [0.274]	88.1% [0.447]	
mean (SE)	0.159 (0.053)	0.083 (0.139)	0.101 (0.091)	0.045 (0.143)	0.025 (0.079)	0.043 (0.144)	200
% coverage [length]	4.8% [0.170]	82.6% [0.446]	94.0% [0.410]	90.7% [0.405]	93.6% [0.227]	90.7% [0.391]	
mean (SE)	0.159 (0.045)	0.073 (0.139)	0.089 (0.090)	0.031 (0.129)	0.017 (0.061)	0.030 (0.129)	300
% coverage [length]	0.3% [0.131]	83.0% [0.443]	96.7% [0.404]	89.0% [0.357]	93.3% [0.175]	89.0% [0.354]	
mean (SE)	0.159 (0.034)	0.069 (0.132)	0.080 (0.085)	0.032 (0.120)	0.010 (0.046)	0.034 (0.124)	500

1.7.2 Simulation 2

In this study, we designed a set of simulation to evaluate the impact of correlation among covariates. The data generating mechanism is as follows:

- New trial: $(X_1, \ldots, X_{10}) \sim MVN(0.5 \cdot \vec{1}, \Sigma)$,
- Old trial: $(X_1, \ldots, X_{10}) \sim MVN(-0.5 \cdot \vec{1}, \Sigma)$,

where $\vec{1} = (1,1,1,1,1,1,1,1,1,1)^T$, $\Sigma = (\sigma_{i,j})$, and $\sigma_{ij} = 0.5 \cdot I_{i \neq j} + I_{i=j}$.

Once X was obtained, outcome Y was simulated from same probit model for both trials:

$$P(Y^{D_O} = 1|X,T) = P(Y^{D_N} = 1|X,T) = \Phi(-2 + X\beta + 0.5 \cdot I_{\{T=t\}}),$$

where $\beta = (1, -0.5, 0.5, 0, 0, 0, 0, 0, 0, 0)^T$.

Similar to the previous study, 1000 independent data sets were simulated for sample size $n = 100, 200, 300$, and 500, respectively. Again, all methods were applied to these simulated data sets. In MAIC_N, the covariance matrix is estimated by assuming that the two populations share the same correlation structure. And, BM and

BMIC methods with covariance estimation are referred to as BM-COV and BMIC-COV, respectively.

We found that MAIC has a lower coverage when sample size is small. After integrating covariance estimation, BM and BMIC outperformed other methods. MAIC_N also provided relatively accurate coverage with shorter interval length.

Table 1.3: Data generating models for X are different for two trials, covariates are highly correlated. Sample size in each trail varied from 100 to 500.

	BASE	MAIC	MAIC_B	MAIC_N	BM-COV	BMIC-COV	Sample Size
% coverage [length]	59.3% [0.320]	83.2% [0.527]	98.1% [0.545]	95.1% [0.470]	91.5% [0.323]	92.6% [0.418]	
mean (SE)	0.138 (0.084)	0.035 (0.164)	0.068 (0.105)	0.015 (0.104)	0.017 (0.094)	0.007 (0.157)	100
% coverage [length]	13.9% [0.188]	88.4% [0.271]	98.1% [0.369]	95.7% [0.242]	94.3% [0.193]	91.1% [0.227]	
mean (SE)	0.145 (0.046)	-0.001 (0.079)	0.005 (0.073)	0.002 (0.061)	0.004 (0.050)	-0.002 (0.085)	200
% coverage [length]	2.9% [0.145]	92.3% [0.207]	95.3% [0.225]	94.4% [0.192]	93.8% [0.149]	89.7% [0.181]	
mean (SE)	0.145 (0.038)	0.004 (0.055)	0.004 (0.054)	0.006 (0.050)	0.003 (0.038)	0.006 (0.073)	300
% coverage [length]	0.0% [0.103]	95.1% [0.140]	93.8% [0.140]	94.9% [0.133]	96.0% [0.106]	89.6% [0.125]	
mean (SE)	0.145 (0.025)	0.001 (0.037)	0.001 (0.036)	0.002 (0.034)	0.001 (0.026)	0.001 (0.049)	500

1.7.3 Simulation 3

In this study, we evaluate the influence of the data generating model for Y on different methods. From previous simulation study, we found that different methods performance significantly different only when sample size is relative small. Therefore, in this section, we only focus on the cases where sample size $n = 100$.

Here, we still simulated 10 continuous covariates $(X_1, \ldots, X_{10})$ from the following distribution:

- New Trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} N(0.5, 1)$,
- Old Trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} N(-0.5, 1)$.

The link function between X and Y, however, is specified in different ways as showed in the table below:

Table 1.4: Model Specification for Y

Effective Covariates	$P(Y^{D_O}=1\|X,T)=P(Y^{D_N}=1\|X,T)$	β
X_1	$logistic(-2+0.5\cdot I_{\{T=t\}}+X\beta)$	(1,0,0,0,0,0,0,0,0,0)
X_1, X_2, X_3	$logistic(-2+0.5\cdot I_{\{T=t\}}+X\beta)$	(1,-0.5,0.5,0,0,0,0,0,0,0)
$X_1, X_2, X_3, X_4\cdot X_5$	$logistic(-2+0.5\cdot I_{\{T=t\}}+X\beta+0.5X_3\cdot X_4)$	(1,-0.5,0.5,0,0,0,0,0,0,0)
$X_1, \ldots, X_{10}$	$logistic(-2+0.5\cdot I_{\{T=t\}}+X\beta)$	(-.1,-.15,-.25,.2,.1,0.2,0.25,0.35,0.45,-.05)

After changing the generating model of Y from probit link to logistic link, BM method performances worse as showed in Table 1.5. Furthermore, BM method is more sensitive to the interaction term in the logistic model for Y than the other methods, as no specific model of Y is assumed by the other methods.

We can also see that MAIC had a comparatively low performance in terms of coverage and average length of CI. In most cases, incorporating bootstrap procedure brings benefit of more accurate estimation by bring back the uncertain of the weight estimation and sample variance. However, sometimes, when MAIC result was not stable, (say the truth is on the boundary) due to solving high dimensional non-linear equations, perturbation of MAIC may bring the prediction farther from the truth.

Table 1.5 also shows that incorporating variable selection can indeed reduce the estimation variance, especially when few covariates were associated with Y. For example, in the case of only 1 or 3 covariates were associated with Y, the BMIC method with variable selection (called BMIC-V) provides a shorter CI with a higher coverage.

Table 1.5: Data generating models for X are different for both trials, number of influential variables varied from 1 to 10.

	BASE	MAIC	MAIC$_B$	MAIC$_N$	BM	BMIC	BMIC-V	Effective Covariates
% coverage [length]	52.0% [0.316]	85.3% [0.557]	91.3% [0.510]	92.9% [0.697]	84.7% [0.468]	88.8% [0.617]	96.6% [0.664]	
mean (SE)	0.153 (0.081)	0.100 (0.171)	0.115 (0.113)	0.048 (0.182)	0.063 (0.164)	0.041 (0.184)	0.050 (0.137)	X_1
% coverage [length]	58.0% [0.323]	88.8% [0.588]	93.6% [0.518]	93.4% [0.728]	85.3% [0.474]	90.3% [0.647]	96.6% [0.684]	
mean (SE)	0.144 (0.081)	0.098 (0.159)	0.109 (0.107)	0.049 (0.184)	0.058 (0.159)	0.043 (0.187)	0.046 (0.145)	X_1, X_2, X_3
% coverage [length]	66.4% [0.343]	87.9% [0.616]	95.1% [0.544]	90.5% [0.758]	78.8% [0.445]	87.4% [0.678]	95.7% [0.714]	
mean (SE)	0.139 (0.086)	0.067 (0.174)	0.083 (0.122)	0.007 (0.198)	-0.042 (0.150)	0.002 (0.201)	0.015 (0.159)	$X_1, X_2, X_3, X_4 \cdot X_5$
% coverage [length]	16.2% [0.303]	88.1% [0.532]	80.3% [0.487]	94.0% [0.642]	84.6% [0.367]	90.0% [0.560]	97.3% [0.636]	
mean (SE)	0.227 (0.074)	0.135 (0.155)	0.161 (0.101)	0.072 (0.163)	0.044 (0.124)	0.060 (0.165)	0.089 (0.143)	$X_1, \ldots, X_{10}$

1.7.4 Simulation 4

In order to test the robustness of the normal assumption, we generated covariates X from distributions other than normal in this study. To be concrete, we simulated 10 continuous covariates $(X_1, \ldots, X_{10})$ from relocated t-distributions as follows:

- Scenario 1A:
  - New Trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} t_4 - 0.5$
  - Old Trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} t_4 - 0.5$
- Scenario 1B:
  - New Trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} t_4 - 0.5$
  - Old Trial: $X_1, \ldots, X_{10} \overset{i.i.d.}{\sim} t_4 + 0.5$

The logistic regression model below was used to simulate outcome Y given X:

$$P(Y^{D_O} = 1|X, T) = P(Y^{D_N} = 1|X, T) = logistic(X\beta + 0.5 \cdot I_{\{T=t\}}),$$

where $\beta = (2, -1, 1, 0, 0, 0, 0, 0, 0, 0)^T$.

Table 1.6: Scenario 4A, data generating models for X are the same for both trials, sample size for each trial varied from 100 to 500.

	BASE	MAIC	MAIC$_B$	MAIC$_N$	BM	BMIC	Sample Size
% coverage [length]	95.9% [0.376]	94.9% [0.496]	99.7% [0.626]	98.5% [0.714]	95.7% [0.330]	98.1% [0.703]	
mean (SE)	0.003 (0.095)	0.004 (0.120)	0.005 (0.108)	0.006 (0.136)	0.009 (0.080)	0.005 (0.146)	100
% coverage [length]	95.1% [0.270]	96.1% [0.304]	97.5% [0.327]	98.3% [0.596]	97.3% [0.246]	98.2% [0.575]	
mean (SE)	0.002 (0.070)	0.003 (0.073)	0.004 (0.071)	-0.001 (0.112)	0.007 (0.055)	-0.002 (0.127)	200
% coverage [length]	94.4% [0.222]	96.4% [0.240]	97.0% [0.247]	98.8% [0.548]	96.8% [0.205]	98.1% [0.518]	
mean (SE)	0.004 (0.058)	0.004 (0.061)	0.004 (0.058)	0.009 (0.111)	0.010 (0.046)	0.009 (0.126)	300
% coverage [length]	94.0% [0.172]	97.8% [0.181]	97.4% [0.183]	98.8% [0.509]	97.3% [0.162]	98.2% [0.466]	
mean (SE)	-0.001 (0.045)	0.000 (0.041)	0.000 (0.041)	0.003 (0.101)	0.006 (0.035)	0.001 (0.117)	500

Table 1.7: Scenario 4B, data generating models for X are different for both trials, sample size for each trial varied from 100-500 .

	BASE	MAIC	MAIC$_B$	MAIC$_N$	BM	BMIC	Sample Size
% coverage [length]	34.7% [0.367]	84.0% [0.731]	87.6% [0.619]	97.3% [0.903]	82.7% [0.387]	95.3% [0.834]	
mean (SE)	-0.226 (0.092)	-0.136 (0.196)	-0.163 (0.125)	-0.115 (0.195)	-0.036 (0.139)	-0.107 (0.206)	100
% coverage [length]	10.9% [0.263]	83.8% [0.701]	94.0% [0.605]	96.4% [0.780]	93.9% [0.321]	94.2% [0.724]	
mean (SE)	-0.220 (0.068)	-0.083 (0.201)	-0.109 (0.128)	-0.089 (0.183)	-0.018 (0.086)	-0.087 (0.197)	200
% coverage [length]	2.6% [0.216]	84.0% [0.700]	95.0% [0.611]	96.0% [0.749]	91.4% [0.273]	94.5% [0.698]	
mean (SE)	-0.221 (0.057)	-0.077 (0.200)	-0.097 (0.125)	-0.097 (0.171)	-0.026 (0.077)	-0.094 (0.186)	300
% coverage [length]	0.0% [0.168]	82.7% [0.676]	97.2% [0.609]	97.1% [0.715]	95.2% [0.216]	94.6% [0.656]	
mean (SE)	-0.220 (0.042)	-0.049 (0.203)	-0.063 (0.123)	-0.080 (0.165)	-0.016 (0.052)	-0.080 (0.180)	500

We followed exactly the same simulation and data analysis strategies. The results are summarized in Table 1.6 and Table 1.7, from which we can see that even if the underlining distribution is not perfect normal, the newly proposed methods still consistently outperforms MAIC.

1.8 Real-Like Data Example

In this section, we will show the application of the above methods in a pseudo real data. The pseudo real data were generated by perturbing a real data set without changing its main structure. The real dataset contains data from two trials, one for treatment A, one for treatment B. For treatment A, we have the individual level data

(95 patients for treatment, 108 patients for control/placebo); for treatment B, we only have the aggregated data from the literature (96 patients for treatment, 112 patients for control/placebo). There are 4 continuous variables (X_1, X_2, X_3, X_4) including age, BMI, etc., and 7 binary variables $(X_5, \ldots, X_{11})$ including gender, race indicator and so on in both trials.

Figure 1.2 shows histograms of the four continuous variables, from which it is easy to tell the normal assumption is reasonable for the third and forth variable, but not for the first two variables. We did a log transformation for X_1 and X_2. The histograms of the transformed data are showed in Figure 1.3, which indicates that the normal assumption becomes reasonable after the transformation. Therefore, we will assume that X_1 and X_2 follow log-normal distribution, X_3 and X_4 follow normal distribution, and all other covariates follow bernoulli distribution.

In order to incorporate the correlation structure among covariates into analysis, we assume that under the log scale, the two population distribution share same correlation structure, even though the variances of covariates may be different.

We applied MAIC, MAIC_B and MAIC_N to estimate the efficacy difference for the perturbed real data. And, the bootstrap procedure was conducted on the log scale with consideration of correlation structure. As a sensitivity analysis of the covariance structure here, the estimation under independent assumption is also reported in the first column of Table 1.8. From Table 1.8, we can see that all methods provide consistent results, i.e., treatment A is significantly better than treatment B after matching. And, because the correlation among covariates are small in this case, we get similar results with or without considering the covariance structure.

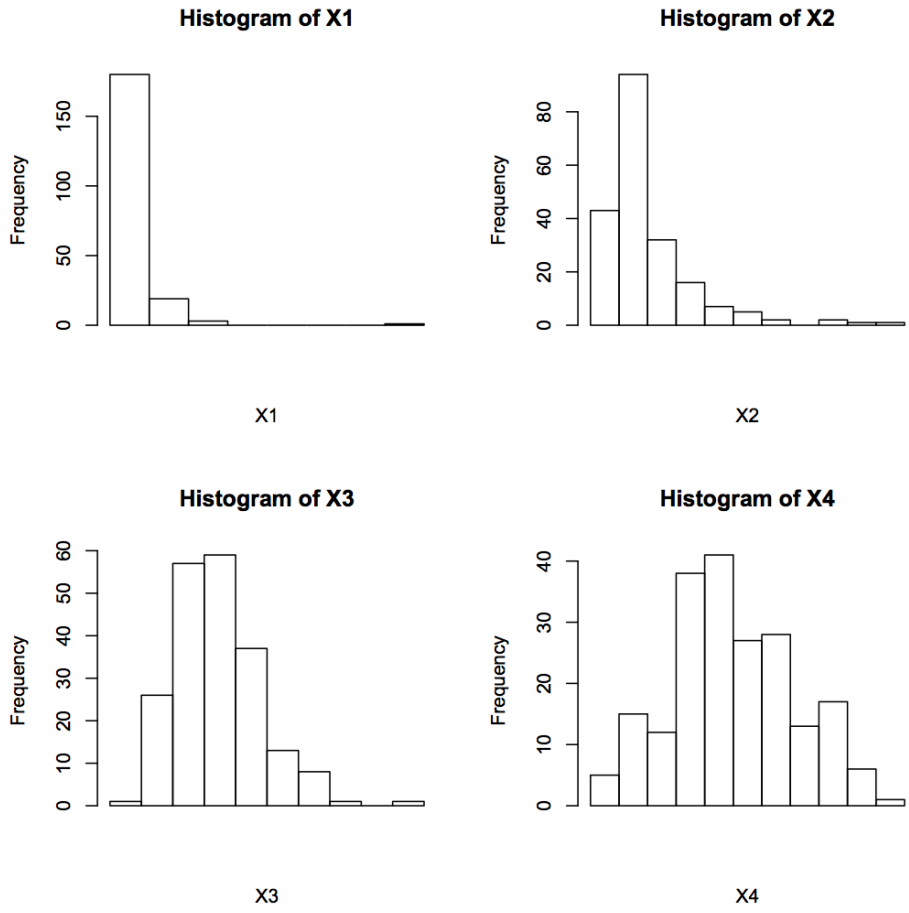


Figure 1.2: Histogram of continuous variables

1.9 Discussion

In this Chapter we systematically studied the problem of indirect comparison. The new proposed methods showed advantage over MAIC in terms of higher coverage rate and shorter length, especially in the case of sample size was smaller and two trial distributions were different which indicated matching adjusted meant was necessary. Additionally, the new proposed methods can incorporate covariance structure and variable selection scheme which could increase the precision of the estimation.

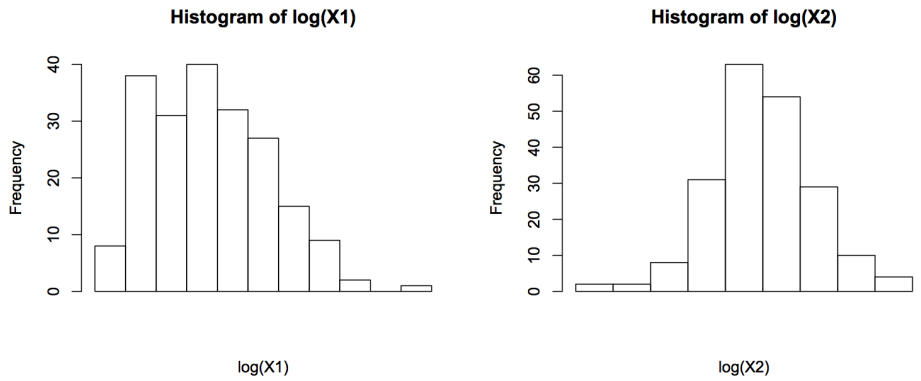


Figure 1.3: Histogram of continuous variables after transformation

Table 1.8: Efficacy difference of treatment A versus treatment B.

	Independent		Correlated	
	Mean	95% CI	Mean	95% CI
MAIC	0.170	(0.023, 0.317)	-	-
MAIC_B	0.177	(0.041, 0.309)	0.180	(0.038, 0.312)
MAIC_N	0.183	(0.016, 0.323)	0.178	(0.024, 0.319)

The simulation study also showed advantage for symmetric X even if the normal assumption was not perfect satisfied. In the real life, if the distribution is not symmetric, such as log normal, MAIC_N can still handle the problem based on the moment estimator of the parameters. Or data transformation could be applied before conduct matching analysis. After transformation, the covariance structure can still be applied by assuming same correlation structure across trials.

Chapter 2

Bayesian Aggregation of Ordered-Based Rank Data

2.1 Introduction

Rank aggregation, whose goal is to generate a "better" aggregated ranking function (referred to as *aggregated ranker*) from multiple ranking functions (referred to as *base rankers*) of a group of entities, is encountered in many disciplines. The earliest efforts on rank aggregation can be traced back to studies on social choice theory and political elections in the eighteenth century (Borda, 1781). Since the middle 1990s, rank aggregation has drawn a lot of attentions with the rise of internet and web search engines. Score-based rank aggregation methods for meta-search (Shaw and Fox, 1994; Manmatha et al, 2001; Montague and Aslam, 2001; Manmatha and Sever, 2002), document analysis (Hull et al., 1996; Vogt and Cottrel, 1999), and similarity search in database (Fagin et al., 2001), which take score information from individual

base rankers as input to generate an aggregated ranker, form the first wave of modern rank aggregation studies.

However, considering that usually only order information is available in meta-search, order-based methods, which need only the order information from base rankers, became popular quickly. The first generation of order-based methods construct the aggregated ranking function based on simple statistics of the ranked lists from base rankers. For example, Van Erp and Schomaker (2000), and Aslam and Montague (2001) proposed to use a democratic voting procedure called *Borda count* (i.e., the average rank across all base rankers) to generate the aggregated rank; while Fagin, Kumar and Sivakumar (2003b) suggested the use of median rank. To strive for better performance, more complicated methods were proposed later, including Markov-chain-based methods (Dwork et al, 2001), fuzzy-logic-based method (Ahmad and Beg, 2002), genetic algorithm (Beg, 2004), and graph-based method (Lam and Leung, 2004). In addition, as an important special case, the problem of combining the top-d lists has been given extra attention in Dwork et al. (2001) and Fagin et al. (2003).

Randa and Straccia (2003) compared the performance of score-based methods and rank-based methods in the context of meta-search, and found that Markov-chain-based methods do achieve results comparable to score-based methods, and significantly outperform methods based on Borda count. The success of Markov-chain-based methods quickly made Dwork et al. (2001) a classic. These methods were later applied to bioinformatics problems, and a number of their variants and extensions were proposed to fit more complicated situations (Sese and Morishita, 2001;

DeConde et al., 2006; Lin and Ding, 2009). Meanwhile, Freund et al. (2003) proposed to achieve preference aggregation via boosting with the guidance of "feedback" that provides information about which pair of entities should be ranked above or below one another.

In practice, the problem of rank aggregation can become even more challenging because of the diverse quality of the base rankers. For example, in a meta-search study, some searching engines are more powerful than others; or, in a meta-analytic bioinformatic study, some labs collect and/or analyze data in a more efficient way than other labs do. In some extreme cases, some base rankers may be noninformative or even misleading. For example, "paid placement" and "paid inclusion" are very popular among search engines. These low quality base rankers, referred to as *spam rankers*, may disturb the rank aggregation procedure significantly if they are not treated properly. However, little attention has been given to this point. Giving different base rankers different weights, as mentioned in Aslam and Montague (2001) and Lin and Ding (2009), is the only method available that takes the diverse quality of base rankers into consideration. But, a clear disadvantage of this method is that there are no systematic and principled strategies for designing a proper weighting scheme for a practical problem. Recently, supervised rank aggregation (Liu et al., 2007) has been proposed to specify weights to base rankers, but it is achieved at the price of a set of training data, which may often be unavailable.

In this Chapter, we focus on the order-based rank aggregation problem and propose a novel Bayesian method to tackle it. By reformulating the original rank aggregation problem into a Bayesian model selection problem and attaching a quality

parameter to each base ranker, we can estimate the quality of base rankers jointly with rank aggregation. Compared to existing methods in the literature, our method is superior in the sense of having an explicit model, being adaptive to the heterogeneity of base rankers, and achieving a better statistical efficiency. The remainder of the article is organized as follows. In Section 2.2, we formally define the rank aggregation problem, and briefly review the existing methods. In Section 2.3, we propose our Bayesian approach for rank aggregation. Section 2.4 provides tools for model diagnosis. The performance of our method is evaluated via simulations in Section 2.5. Real data applications are given in Section 2.6 to demonstrate the application potentials of the new method in practice. We conclude our study with discussions in Section 2.7.

2.2 An Overview of Existing Methods

Let $U = \{1, 2, \cdots, n\}$ be the "universe" (set) of n entities of interest. An ordered list (or simply, a list) τ with respect to U is a ranking of entities in a subset $S \subseteq U$, i.e., $\tau = [x_1 \preceq x_2 \preceq \cdots \preceq x_d]$, where $x_t \in S$, "$i \preceq j$" means that i is ranked higher than j. Let $\tau(i)$ be the position or rank of entity $i \in \tau$ (a highly ranked element has a low-numbered position in the list).

We call τ a *full list* if it contains all the elements in U; and a *partial list* otherwise. An important special case of partial lists is the *top-d lists*. For a list τ and a subset T of the universe U, the projection of τ with respect to T (denoted as $\tau_{|T}$) will be a new list that contains only entities from T. Note that if τ happens to contain all elements in T, then $\tau_{|T}$ is a full list with respect to T. In rank aggregation study, our

goal is to generate an aggregated list α from a group of full/partial lists $\{\tau_1, \cdots, \tau_m\}$.

2.2.1 Methods based on summary statistics

Many rank aggregation methods are based on simple summary statistics of the m given base rankers. Let $\{\tau_k(i)\}_{1\leq k\leq m}$ be the ranks of entity i in the base rankers, to determine the rank of entity i in the aggregated list, the arithmetic mean, geometric mean, or median of $\{\tau_k(i)\}_{1\leq k\leq m}$ are often used. In this paper we refer to the above three methods as AriM, GeoM and MedR, respectively. These naive methods are straightforward, but easy to be disturbed by spam rankers.

2.2.2 Optimization-based methods and Markov-chain-based methods

Dwork et al. (2001) propose to report the list that minimizes an objective function as the aggregated rank list, i.e., let

$$\alpha = \arg\min_{\sigma\in\mathscr{A}_U} d(\sigma; \tau_1, \cdots, \tau_m),$$

where $\mathscr{A}_U$ is the space of all allowable rankings of entities in U, and the objective function d can be either the *Spearman's footrule distance* (Diaconis and Graham, 1977)

$$d_F(\sigma; \tau_1, \cdots, \tau_m) \triangleq \frac{1}{m}\sum_{k=1}^{m} F(\sigma_{|\tau_k}, \tau_k),$$

where $F(\sigma_{|\tau_k}, \tau_k) = \sum_{i \in \tau_k} |\sigma_{|\tau_k}(i) - \tau_k(i)|$, or the *Kendall tau distance* (Diaconis, 1988)

$$d_K(\sigma; \tau_1, \cdots, \tau_m) \triangleq \frac{1}{m} \sum_{k=1}^{m} K(\sigma_{|\tau_k}, \tau_k),$$

where $K(\sigma_{|\tau_k}, \tau_k)$ is the bubble sort distance between $\sigma_{|\tau_k}$ and τ_k. The aggregation obtained by optimizing the Kendall distance is called *Kemeny optimal aggregation*, and the one obtained by optimizing the Spearman's footrule distance is called *footrule optimal aggregation*. In fact, the idea of generating the aggregated ranking by optimizing the Kendall distance can go back to the Mallows model in 1950s (Mallows 1957), which is generalized by Fligner and Verducci (1986) and later Meila et al. (2007).

Considering that it is computationally expensive to solve the above optimization problems (the Kemeny optimal aggregation is NP-Hard, and the footrule optimal aggregation needs an expensive polynomial algorithm), Dwork et al. (2001) also propose a few Markov-chain-based methods as fast alternatives to provide suboptimal solutions. The basic idea behind these methods is to construct a transition probability matrix $P = \{p_{ij}\}_{i,j \in U}$ based on $\{\tau_1, \cdots, \tau_m\}$, where p_{ij} is the transition probability from entity i to entity j, and use the stationary distribution of P to generate the aggregated ranked list, i.e., let

$$\alpha = sort(i \in U \text{ by } \pi_i \downarrow), \text{ where } \pi = (\pi_1, \cdots, \pi_n) \text{ satisfies } \pi P = \pi,$$

and the symbol $\downarrow$ means that the entities are sorted in descending order. In practice, a few Markov-chain-based methods have been developed by constructing P in different

ways as proposed by Dwork et al. (2001) and Deconde et al. (2006):

MC_1: If the current state is entity i, then the next state is chosen uniformly from the set of all entities that are ranked higher than (or equal to) i by some base rankers that rank i.

MC_2: If the current state is entity i, then the next state is chosen by first picking a base ranker τ uniformly from all base rankers containing entity i, then picking a entity j uniformly from the set $\{j \in \tau : \tau(j) \leq \tau(i)\}$.

MC_3: If the current state is entity i, then the next state is chosen as follows. First pick base ranker τ uniformly from all base rankers containing entity i, then uniformly pick an entity j that is ranked by τ. If $\tau(j) \leq \tau(i)$, then go to j; otherwise, stay in i.

MC_4: If the current state is entity i, then the next state is chosen as follows. First, pick an entity j uniformly from the union of all entities ranked by the base rankers. If $\tau(j) \leq \tau(i)$ for a majority of the base rankers that rank both i and j, then go to j; otherwise, stay in i.

MC_T: MC_T is identical to MC_4, except that the move from i to j at the last step is not a deterministic procedure based on the majority vote, but a stochastic procedure in which the probability to accept j is proportional to the percentage of base rankers that rank j higher than i among all base rankers that rank both i and j.

2.2.3 Rank aggregation of weighted lists

Considering that the base rankers of interest may not be equally knowledgeable or reliable in practice, methods based on weighted lists are also proposed. In these methods, each base ranker τ_k is assigned a weight w_k ($0 \leq w_k \leq 1$ and $\sum_{1 \leq k \leq m} w_k = 1$), and the base rankers with larger weights play more important roles in generating the aggregated list. Aslam and Montague (2001) propose to generate the aggregated list based on the weighted average of the m lists (known as Borda Fuse), i.e., let $\alpha = sort\big(i \in U \text{ by } \sum_{k=1}^{m} w_k \tau_k(i) \downarrow \big)$. Lin and Ding (2009) extend the objective function of Dework et al. (2001) to a weighted fashion, and generated the aggregated list as follows:

$$\alpha = \arg\min_{\sigma \in \mathscr{A}_U} d(\sigma; \tau_1, \cdots, \tau_m; w) = \arg\min_{\alpha \in \mathscr{A}_U} \sum_{k=1}^{m} w_k d(\alpha_{|\tau_k}, \tau_k),$$

where $d(\alpha_{|\tau_k}, \tau_k) = F(\alpha_{|\tau_k}, \tau_k)$ or $K(\alpha_{|\tau_k}, \tau_k)$. The authors also propose to use Cross Entropy Monte Carlo (CEMC) to solve the above optimization problem. (The optimization based on Spearman's footrule distance is denoted as CEMC_F, and the optimization based on Kendall distance is denoted as CEMC_K.) CEMC is an iterative algorithm to solve difficult optimization problems, details of the algorithm can be found in Rubinstein and Kroese (2004).

Although assigning weights to base rankers is a sensible way of handling the quality difference among them, it can be quite difficult to design a proper weight specification scheme in practice, especially when little or no prior knowledge on base rankers is available. The *supervised rank aggregation* (SRA) of Liu et al. (2007) solves this

problem at the price of extra training data. In SRA, the true relative ranks of some entities are provided as training data, and the weights $\{w_k\}_{1\leq k\leq m}$, which are treated as parameters instead of pre-fixed constants in these models, are optimized with the help of the training data as well as the aggregated list σ. A problem of SRA is that no training data are available in many applications.

2.2.4 Rank aggregation via boosting

Another line of using training data to achieve rank aggregation in the literature is the *RankBoost* method of Freund et al. (2003). Similar to SRA, RankBoost assumes that besides the rank lists $\{\tau_1, \cdots, \tau_m\}$, we also have a *feedback function* of the form $\Phi : U \times U \to \mathbf{R}$, where $\Phi(i, j) > 0$ means that entity i should be ranked above entity j, $\Phi(i, j) < 0$ means the opposite, and a value of zero indicates no preference between i and j. Different from SRA, RankBoost does not tend to assign weights to different rankings themselves. Instead, RankBoost follows the boosting idea to generate a series of *weak rankings* from $\{\tau_1, \cdots, \tau_m\}$, and construct the final ranking by a weighted average of these weak rankings.

2.3 A Bayesian Model for Rank Aggregation

2.3.1 Assumptions and the model

Here we propose a Bayesian approach, called *Bayesian Aggregation of Rank Data* (BARD), to tackle the challenging problem of aggregating rankings at different quality levels. This section focuses on the rank aggregation of full lists; a more complicated

scenario where partial lists and training data are also involved will be discussed in the next section. Our BARD method reformulates the ranking problem as follows. We assume that the set U is composed of two non-overlapping subsets: set U_R representing relevant entities (with true signals) and set U_B representing noisy background entities. Task of each base ranker is to distinguish the relevant entities from the background ones. By integrating the rankings from the base rankers, we try to infer the set of relevant entities.

Let I_i be the group indicator of entity $i \in U$, where $I_i = 1$ if $i \in U_R$, and $I_i = 0$ if $i \in U_B$. We make the following assumptions for base rankers $\tau_1, \cdots, \tau_m$:

- Given the group indicators of entities $I = \{I_i\}_{i \in U}$, the rankers $\tau_1, \cdots, \tau_m$ are conditionally independent;
- In each base ranker τ_k, the relative ranks of background entities $\tau_k^0 \triangleq \tau_{k|U_B}$ is purely random (i.e., uniformly distributed) in the space of all permutations;
- The relative rank of an entity $i \in U_R$ among the background entities $\tau_k^{1|0}(i) \triangleq \tau_{k|\{i\} \cup U_B}(i)$ follows a power law distribution, i.e. $P(\tau_k^{1|0}(i) = t) \propto t^{-r_k}$, where a larger γ_k ($\gamma_k > 0$) means that ranker τ_k can better distinguish relevant entities from the background ones[1];
- Given $\tau_k^{1|0} \triangleq \{\tau_k^{1|0}(i)\}_{i \in U_R}$, the relative ranks of background entities $\tau_k^1 \triangleq \tau_{k|U_R}$ is purely random in the space of all allowable values.

Because, the triplet $(\tau_k^0, \tau_k^{1|0}, \tau_k^1)$ gives an equivalent representation of the information in a full list τ_k when I is given (the equivalency is illustrated in Figure 2.1

[1]Note that by requiring that $\gamma_k > 0$, we assume that each base ranker τ_k is capable of distinguishing the relevant entities from the background ones more or less.

U	I
E_1	1
E_2	1
E_3	1
E_4	0
E_5	0
E_6	0
E_7	0
E_8	0
E_9	0
E_{10}	0

$\Longleftrightarrow$

τ_k	τ_k^0	$\tau_k^{1\|0}$	τ_k^1
2	-	2	1
3	-	2	2
5	-	3	3
1	1	-	-
4	2	-	-
6	3	-	-
7	4	-	-
8	5	-	-
9	6	-	-
10	7	-	-

Figure 2.1: An equivalent representation of a full rank list τ_k via the triplet $(\tau_k^0, \tau_k^{1|0}, \tau_k^1)$.

with a toy example), the above assumptions lead to the following likelihood function immediately:

$$\begin{aligned} P(\tau_1, \cdots, \tau_m \mid I, \gamma) &= \prod_{k=1}^{m} P(\tau_k \mid I, \gamma_k) \\ &= \prod_{k=1}^{m} P(\tau_k^0, \tau_k^{1|0}, \tau_k^1 \mid I, \gamma_k) \\ &= \prod_{k=1}^{m} P(\tau_k^0 \mid I) \times P(\tau_k^{1|0} \mid I, \gamma_k) \times P(\tau_k^1 \mid \tau_k^{1|0}; I), \quad (2.1) \end{aligned}$$

where $P(\tau_k^0 \mid I)$ and $P(\tau_k^1 \mid \tau_k^{1|0}; I)$ are uniform distributions on the corresponding spaces of allowable configurations, and

$$P(\tau_k^{1|0} \mid I, \gamma_k) = \prod_{i \in U_R} P(\tau_k^{1|0}(i) \mid I, \gamma_k) \quad \text{where} \quad P(\tau_k^{1|0}(i) = t \mid I, \gamma_k) \propto t^{-\gamma_k}. \quad (2.2)$$

In practice, however, both I and γ are unknown, and it is our main goal to estimate them from the observed data $\{\tau_1, \cdots, \tau_m\}$. Letting $\pi(I, \gamma)$ be the prior distribution,

the Bayes' rule leads to the following posterior distribution of (I, γ):

$$P(I, \gamma \mid \tau_1, \cdots, \tau_m) \propto P(\tau_1, \cdots, \tau_m \mid I, \gamma)\pi(I, \gamma).$$

Since the marginal probability

$$\rho_i \triangleq P(I_i = 1 \mid \tau_1, \cdots, \tau_m) \tag{2.3}$$

is a good measurement of the importance of entity i, we generate the aggregated list as

$$\alpha = sort(i \in U \text{ by } \rho_i \downarrow). \tag{2.4}$$

On the other hand, the posterior mean

$$\overline{\gamma}_k \triangleq \int \gamma_k P(\gamma_k \mid \tau_1, \cdots, \tau_m) d\gamma_k \tag{2.5}$$

gives the estimation of the quality of base ranker τ_k.

The identifiability of the BARD model comes from the following intuition. For a fixed group of n entities, if we have a larger number of independent rankings generated from the posited model, we will expect a clear gap between the average rank of a relevant entity (across all rankings) and the average rank of a background entity. Thus, it will be straightforward to distinguish the relevant entities from the background entities, which will in turn help us to determine the quality of each ranking, even though we do not really need this quality information to discover relevant entities in the first place. In a practical problem where m is small, having the quality

information of different rankings, however, becomes more useful for discovering relevant entities. Therefore, estimating the group indicators of entities I and estimating the quality of rankings γ can in fact help each other.

2.3.2 Motivations and intuitions behind the model

Compared with the existing methods, BARD is unique in the following features: (1) partitioning the entities into two groups U_R and U_B, (2) modeling the relative rank of a relevant entity among background entities $\tau_k^{1|0}(i)$ (i.e., the between-group rankings) with a power-law distribution, and (3) modeling the within-group rankings τ_k^1 and τ_k^0 with the uniform distribution. In this subsection, we explain in details why we introduce these features and how these features help us better resolve the rank aggregation problem.

First, the partition of U into U_R and U_B is directly motivated by the observation that behind a ranking problem there is often a partitioning problem. For example, in the page-ranking problem, conceptually there is a binary answer for each web page whether it is truly relevant to a given search task (e.g., a group of key words) or not. By ranking the web pages, what we really want to achieve is to better distinguish the truly relevant web pages from the other web pages. Every year, each grant committee of NSF or NIH ranks hundreds of grant proposals; but, at the end of the decision procedure, some top proposals are funded and the others are dismissed due to the limited resources. Again, ranking is just an intermedia step of the whole decision procedure whose final goal is to partition the grant proposals into funded and unfunded groups.

Second, the power-law model $\tau_k^{1|0}(i)$ is a convenient approximation reasonably reflective of reality. In a real problem, the distribution of $\tau_k^{1|0}(i)$ depends on many factors and can take different forms in different problems. But we need to find a computationally affordable model for $\tau_k^{1|0}(i)$. We reason that it needs to satisfy the following simple requirement: it should give higher probability to a better rank for a relevant item and be no worse than assigning it a random (uniform) rank. That is, the probability function should be a monotone decreasing function. Two obvious choices are exponential or polynomial, of which polynomial is more robust and therefore chosen here. A large range of numerical investigations also support the adoption of the power law distribution. For example, we generate each ranker τ_k as the order of $\{X_{k,1}, \cdots, X_{k,n}\}$, i.e.,

$$\tau_k = sort(i \in U \text{ by } X_{k,i} \downarrow),$$

where $X_{k,i}$ is generated from two different distributions $F_{k,0}$ and $F_{k,1}$ via the following mechanism:

$$X_{k,i} \sim F_{k,0} \cdot I(i \in U_B) + F_{k,1} \cdot I(i \in U_R), \ \forall \ i \in U.$$

Figure 2.2 shows that the linear trend in the log-log plot of t versus $h(t) = P(\tau_k^{1|0}(i) = t \mid \tau_k^0; I, \gamma_k)$ is quite stable across different specifications of $F_{k,0}$ and $F_{k,1}$.

Third, by modeling τ_k^1 and τ_k^0 with the uniform distribution, BARD ignores the detailed information on the internal rankings within subset U_R and U_B, and only takes the relative rankings between the two subsets into consideration. In other words, we choose to ignore all information in the data that is irrelevant to distinguish the relevant entities from the background ones. This strategy greatly reduces the

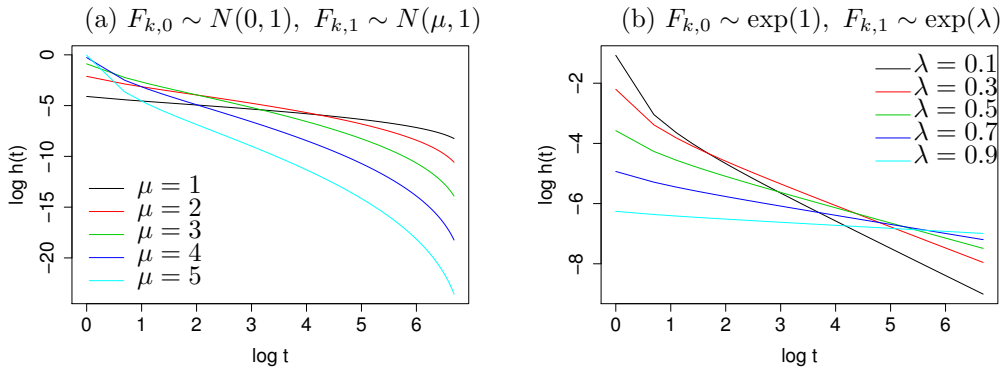


Figure 2.2: Log-log plots of relative rank $\tau_k^{1|0}(i) = t$ versus the corresponding probability $h(t) = P(\tau_k^{1|0}(i) = t \mid I, \gamma_k)$ under different scenarios. In each plot, we set $|U_B| = 1000$, thus the range of t is $\{1, \cdots, 1001\}$. The values of $h(t)$ are calculated via numerical integration.

model complexity and computation burden while loss only marginal information in the data. In some scenarios, we can even argue that the internal rankings within the background group are just noise, and thus, should be ignored to stabilize the analysis.

2.3.3 Details of the Bayesian computation

Let $n_I = \sum_{i=1}^n I_i$ be the number of relevant entities defined by I. Note that the relative rank order of all background entities, τ_k^0, follows the uniform distribution, i.e.,

$$P(\tau_k^0 = \tau) = \frac{1}{(n - n_I)!}. \tag{2.6}$$

Recall that, for entity i with $I_i = 1$ (i.e., relevant entity), $\tau_k^{1|0}(i)$ denotes the relative rank of entity i among all the background entities and takes value in $\in \{1, 2, \cdots, n - n_I + 1\}$. According to model (2), we have

$$P(\tau_k^{1|0}(i) = t_i \mid I, \gamma_k) = \frac{t_i^{-\gamma_k}}{C(\gamma_k, n_I)},$$

and

$$P(\tau_k^{1|0} \mid I, \gamma_k) = \prod_{i \in U_R} P(\tau_k^{1|0}(i) \mid I, \gamma_k) \tag{2.7}$$

where the normalizing constant $C(\gamma_k, n_I) = \sum_{t=1}^{n-n_I+1} t^{-\gamma_k}$.

Let $\mathscr{A}_{U_R}$ be the space of all allowable rankings of entities in U_R. Let $\mathscr{A}_{U_R}(\tau_k^{1|0})$ be the configurations of τ_k^1 that are compatible with a given $\tau_k^{1|0}$. $\mathscr{A}_{U_R}(\tau_k^{1|0})$ is a subset of $\mathscr{A}_{U_R}$ due to the constraints introduced by $\tau_k^{1|0}$. For example, given $\tau_k^{1|0} = (2, 2, 3)$ as shown in Figure 2.1, τ_k^1 will have only two possible configurations: $(1, 2, 3)$ or $(2, 1, 3)$, since only the relative position of the first two entities E_1 and E_2 is not fixed given $\tau_k^{1|0}$. In general, we have the following assignment based on the "purely random assumption":

$$P(\tau_k^1 = \tau \mid \tau_k^{1|0}; I) = \frac{1}{\prod_{t=1}^{n-n_I+1} n_{\tau_k,t}^{1|0}!} \cdot I\big(\tau \in \mathscr{A}_{U_R}(\tau_k^{1|0})\big), \tag{2.8}$$

where $n_{\tau_k,t}^{1|0} = \sum_{i \in U_R} I(\tau_k^{1|0}(i) = t)$.

Putting (2.6), (2.7) and (2.8) together, we have

$$\begin{aligned} P(\tau_k \mid I, \gamma_k) &= P(\tau_k^0 \mid I) \times P(\tau_k^{1|0} \mid I, \gamma_k) \times P(\tau_k^1 \mid \tau_k^{1|0}; I) \\ &= \Big\{(n - n_I)! \times A_{\tau_k,I} \times \Big(C(\gamma_k, n_I)\Big)^{n_I} \times \Big(B_{\tau_k,I}\Big)^{\gamma_k}\Big\}^{-1}, \end{aligned}$$

where

$$A_{\tau_k,I} \triangleq \prod_{t=1}^{n-n_I+1} (n_{\tau_k,t}^{1|0}!) \;\text{ and }\; B_{\tau_k,I} \triangleq \prod_{i \in U_R} \tau_k^{1|0}(i).$$

Thus, the joint likelihood of $\{\tau_1, \cdots, \tau_m\}$ is

$$P(\tau_1, \cdots, \tau_m \mid I, \gamma) = \Big[(n - n_I)!\Big]^{-m} \times \prod_{k=1}^{m} \Big\{ A_{\tau_k, I} \times \Big(C(\gamma_k, n_I)\Big)^{n_I} \times \Big(B_{\tau_k, I}\Big)^{\gamma_k} \Big\}^{-1}. \tag{2.9}$$

We give I an informative prior

$$\pi(I) \propto \exp^{-\frac{(n_I - n \cdot p)^2}{2\sigma^2}},$$

where p is the hyperparameter representing the expected percentage of relevant entities in U, and give σ^2 is a pre-given hyperparameter (whose default value is $\sigma^2 = \frac{1}{\sqrt{m}}$). We let $\{\gamma_k\}_{1 \le k \le m}$ have an independent exponential prior, i.e., $\pi(\gamma) = \prod_{1 \le k \le m} f(\gamma_k)$, where $f(\gamma_k) = \lambda e^{-\lambda \gamma_k}$, λ is the mean of the exponential distribution. In BARD, we use $\lambda = 1$ as the default setting, and allow users to specify the value of λ based on their own judgement for a practical problem. We also tested using a uniform prior in hyper-cube $[0, 10]^m$ for γ, which resulted in a very similar performance to the exponential prior.

Given the above prior distributions, we get the joint posterior distribution of (I, γ):

$$\begin{aligned} P(I, \gamma \mid \tau_1, \cdots, \tau_m) &\propto \pi(I)\pi(\gamma)P(\tau_1, \cdots, \tau_m \mid I, \gamma) \\ &= \frac{\pi(I)}{\big[(n - n_I)!\big]^m} \cdot \prod_{k=1}^{m} \frac{f(\gamma_k)}{A_{\tau_k, I} \times \Big(C(\gamma_k, n_I)\Big)^{n_I} \times \Big(B_{\tau_k, I}\Big)^{\gamma_k}}, \end{aligned} \tag{2.10}$$

which induces the following conditional distributions:

$$\begin{aligned} P(\gamma_k \mid \tau_1, \cdots, \tau_m; I, \gamma_{[-k]}) &= P(\gamma_k \mid \tau_k; I) \\ &\propto e^{-\lambda\gamma_k} \times \Big(C(\gamma_k, n_I)\Big)^{-n_I} \times \Big(B_{\tau_k, I}\Big)^{-\gamma_k}, \quad (2.11) \\ P(I_i \mid \tau_1, \cdots, \tau_m; I_{[-i]}, \gamma) &\sim \text{Bernoulli}\Big(\frac{q_i(\gamma)}{q_i(\gamma)+1}\Big), \quad (2.12) \end{aligned}$$

where

$$q_i(\gamma) = \frac{\pi(I_{[I_i=1]})}{\pi(I_{[I_i=0]})} \cdot \prod_{k=1}^{m} \frac{P(\tau_k \mid I_{[I_i=1]}, \gamma_k)}{P(\tau_k \mid I_{[I_i=0]}, \gamma_k)}.$$

These distributions enable us to draw samples from $P(I, \gamma \mid \tau_1, \cdots, \tau_m)$ via Gibbs sampling. The posterior probabilities, $P(I_i \mid \tau_1, \cdots, \tau_m)$ and $P(\gamma_k \mid \tau_1, \cdots, \tau_m)$, can be obtained from the Monte Carlo samples and used to generate the aggregated rank list and reliability measures of base rankers. Since the conditional distribution shown in (2.11) is not a standard distribution, we use the random-walk Metropolis algorithm to draw samples from it (see Liu (2001) for a comprehensive review).

2.3.4 Extensions to partial lists and supervised rank aggregation

Since a partial list can be viewed as an incomplete version of a full list, the aggregation of partial lists can be treated as a missing data problem and solved via data augmentation strategies (Tanner and Wong, 1987). To be precise, we let $\{\tau_1^P, \cdots, \tau_m^P\}$ be the m partial lists of interest, and let $\{\tau_1^*, \cdots, \tau_m^*\}$ be their unobserved underlying

full lists. We are interested in drawing samples from the following target distribution:

$$P(I, \gamma \mid \tau_1^P, \cdots, \tau_m^P) \propto \pi(I)\pi(\gamma)P(\tau_1^P, \cdots, \tau_m^P \mid I, \gamma),$$

which can be achieved via Gibbs sampling based on the following conditional distributions:

$$\begin{aligned} P(\tau_1^*, \cdots, \tau_m^* \mid \tau_1^P, \cdots, \tau_m^P; I, \gamma) &= \prod_{k=1}^{m} P(\tau_k^* \mid \tau_k^P; I, \gamma_k), \\ P(I, \gamma \mid \tau_1^*, \cdots, \tau_m^*) &\propto \pi(I)\pi(\gamma)\prod_{k=1}^{m} P(\tau_k^* \mid I, \gamma_k). \end{aligned}$$

Given that the distribution $P(I, \gamma \mid \tau_1^*, \cdots, \tau_m^*)$ has been analyzed in the previous section, we only need to focus on $P(\tau_1^*, \cdots, \tau_m^* \mid \tau_1^P, \cdots, \tau_m^P; I, \gamma)$, or more concretely, $P(\tau_k^* \mid \tau_k^P; I, \gamma_k)$ here. Let Ω_k be the set of full lists that are compatible with τ_k^P, we have

$$P(\tau_k^* \mid \tau_k^P; I, \gamma_k) \propto P(\tau_k^* \mid I, \gamma_k) \cdot I(\tau_k^* \in \Omega_k).$$

Again, we can use random walk Metropolis algorithm to draw samples from this distribution.

BARD can also be applied to the scenario where training data are available. Let $\{\tau_1, \cdots, \tau_m\}$ be the m lists (full or partial) of interest, and $i_1 \preccurlyeq i_2 \preccurlyeq \cdots \preccurlyeq i_s$ be the training information, which gives the true relative rank of s entities $\{i_1, i_2, \cdots, i_s\}$ in U. In BARD, a natural way to make use of the training information is to put constraints on I with respect to $i_1 \preccurlyeq i_2 \preccurlyeq \cdots \preccurlyeq i_s$, i.e., if $I_{i_t} = 1$, then $I_{i_{t'}} = 1$ for all $t' \leq t$. Incorporating the training data into the analysis may help BARD better

estimate the quality parameters $\{\gamma_k\}_{1\leq k\leq m}$ of the m base rankers, and thus, improve the final results.

2.4 Model Diagnostics and Remedies

2.4.1 Detecting violation of the independence assumption

Although we will show in Section 2.5 that BARD is reasonably robust to the violation of the "independent rankers" assumption, it is desirable to detect a severe violation of the assumption and further improve BARD based on this information. Standard correlation measures such as the Spearman and the Kendall correlations do not work here because any pair of informative rankings are inherently correlated since they are supposed to capture the same signal. This type of correlation is not what we are interested in. Instead, we want to detect pairs of rankings that are "over-correlated" relative to their quality levels.

Consider all the ranks $\{\tau_1(i), \cdots, \tau_m(i)\}$ entity i received from all the rankers. It forms a natural distribution on the rank space $\{1, \cdots, n\}$, denoted as Q_i. If entity i has a strong positive/negative signal, a significant proportion of the rankers would give it a high/low rank, so that Q_i skews towards the left/right tail; if entity i belongs to the background, Q_i should be close to be uniform. To capture these key features of Q_i, we fit Q_i with a rescaled Beta distribution:

$$Q_i(t) \propto dBeta\left(\frac{t}{n+1}; \alpha_i, \beta_i\right) \cdot I(t \in \{1, 2, \cdots, n\}),$$

where $dBeta(x; \alpha, \beta)$ is the density of the Beta distribution with parameters (α, β).

Assuming that $\{\frac{\tau_1(i)}{n+1}, \cdots, \frac{\tau_m(i)}{n+1}\}$ are i.i.d draws from distribution $Beta(\alpha_i, \beta_i)$, we denote the estimated parameters as $(\hat{\alpha}_i, \hat{\beta}_i)$, and the fitted distribution as $Q(\hat{\alpha}_i, \hat{\beta}_i)$ ($\hat{Q}_i$ for short).

For any pair of base rankers τ_{j_1} and τ_{j_2}, without loss of generality, we assume that $\tau_{j_1}(i) \leq \tau_{j_2}(i)$. Given the fitted Beta distribution $\hat{Q}_i$, we use the quantity below to measure excessive correlatedness of them at entity i:

$$V_{j_1 j_2}^{(i)} \triangleq \sum_{\tau_{j_1}(i) \leq t \leq \tau_{j_2}(i)} Q(t; \hat{\alpha}_i, \hat{\beta}_i).$$

Intuitively, $V_{j_1 j_2}^{(i)}$ corresponds to the probability that a random sample from $\hat{Q}_i$ falls into the interval $[\tau_{j1}(i), \tau_{j2}(i)]$. A smaller $V_{j_1 j_2}^{(i)}$ means a smaller probability that the two independent rankers agree with each other by chance at entity i, hence a stronger evidence of non-independence. Note that $V_{j_1 j_2}^{(i)}$ accounts for not only the distance between $\tau_{j_1}(i)$ and $\tau_{j_2}(i)$, but also their relative probabilities based on $\hat{Q}_i$. We can estimate the p-value $P_{j_1 j_2}^{(i)} \triangleq P(V_{xy} < V_{j_1 j_2}^{(i)})$ using Monte Carlo simulation, and summarize the overall evidence for the pair of rankers by the *coordination coefficient*:

$$\zeta_{j_1 j_2} \triangleq -\frac{1}{n} \sum_{i=1}^{n} \log P_{j_1 j_2}^{(i)}.$$

A larger $\zeta_{j_1 j_2}$ means that rankers τ_{j_1} and τ_{j_2} are "over-correlated." Alternatively, we can use the method of posterior predictive checking (Rubin 1984) to generate the Bayesian *coordination coefficient*, which will be computationally more demanding.

Under the null hypothesis that the two rankers are independent, we have by the Central Limit Theorem that $\zeta_{j_1 j_2}$ follows $N(1, \frac{1}{n})$ approximately, which can be used to

set a threshold for $\zeta_{j_1 j_2}$ to claim that τ_{j_1} and τ_{j_2} are not independent. The procedure for discovering correlated rankings can be summarized as follows:

- For each entity $i \in U$, fit a rescaled Beta distribution $\hat{Q}_i$ for $\{\tau_1(i), \cdots, \tau_m(i)\}$;
- For each ranker pair τ_{j_1} and τ_{j_2}, calculate the coordination coefficient $\zeta_{j_1 j_2}$ based on $\{\hat{Q}_i\}_{i \in U}$;
- If $\zeta_{j_1 j_2}$ is larger than a threshold (e.g., significance level 0.05 with Bonfferoni correction), we say that τ_{j_1} and τ_{j_2} belong to a "block" of correlated rankers.

2.4.2 A hierarchical model for the correlated base rankers

Once the underlying correlation structure among the rankers are detected, we can modify BARD to avoid the negative impact of the correlation. Assume that the correlated base rankers fall into M blocks $\{G_1, \cdots, G_M\}$, where the rankers within a block are highly correlated while the rankers from different blocks are conditionally independent given the entity membership I. Let G_0 be all the other conditionally independent rankers. To simplify the problem, we assume that every base ranker provides a complete ranking list in this paper. The more general scenario with partial lists involved can be solved based on a similar principle.

Let κ_j be the representative ranker of group G_j, and let $\gamma_j > 0$ denote the quality measure of the ranker block G_j. We modify the BARD model into the following

hierarchical form:

$$\begin{aligned} P(\kappa_j \mid I, \gamma_j) &= P(\kappa_j^0 \mid I) P(\kappa_j^{1|0} \mid I, \gamma_j) P(\kappa_j^1 \mid I, \kappa_j^{1|0}), \\ P(\tau_k \mid \kappa_j, \beta_j) &\propto \exp\left\{ -\frac{\beta_j}{|G_j|} \cdot d(\tau_k, \kappa_j) \right\}. \end{aligned}$$

where $\beta_j > 0$ measures the average magnitude of correlation between κ_j and base rankers in group G_j, $|G_j|$ is the number of rankers in group G_j, and $d(\tau_k, \kappa_j)$ is the Spearman's footrule distance or Kendall tau distance between τ_k and κ_j. The joint likelihood can be weritten as

$$P(\kappa_1, \cdots, \kappa_M; \tau_1, \cdots, \tau_m \mid I, \gamma_j) = \prod_{k \in G_0} P(\tau_k \mid I, \gamma_k) \cdot \prod_{j=1}^{M} \Big[P(\kappa_j \mid I, \gamma_j) \prod_{k \in G_j} P(\tau_k \mid \kappa_j) \Big].$$

In words, the model assumes that the base rankers within each block G_j are conditionally independent of each other given the common ranker κ_j.

Given the prior distribution

$$\pi(I, \gamma, \beta) = \pi(I) \prod_{j=1}^{M} \pi(\gamma_j) \pi(\beta_j),$$

the joint posterior distribution is

$$P(I, \gamma, \beta \mid \tau_1, \cdots, \tau_m) \propto \pi(I, \gamma, \beta) P(\tau_1, \cdots, \tau_M \mid I, \gamma, \beta).$$

An MCMC sampler for simulating from this distribution can be implemented based

on the following conditional distributions:

$$\begin{aligned}
P(\kappa_j \mid I, \gamma_j, \beta_j, \{\tau_k\}_{k\in G_j}) &\propto P(\kappa_j \mid I, \gamma_j) \prod_{k\in G_j} P(\tau_k \mid \kappa_j) \\
&= P(\kappa_j^0 \mid I) P(\kappa_j^{1|0} \mid I, \gamma_j) P(\kappa_j^1 \mid I, \kappa_j^{1|0}) \exp\{-\frac{\beta_j}{m_j} \sum_{k\in G_j} d(\tau_k, \kappa_j)\}, \\
P(\beta_j \mid \kappa_j, \{\tau_k\}_{k\in G_j}) &\propto \pi(\beta_j) \exp\{-\frac{\beta_j}{m_j} \sum_{k\in G_j} d(\tau_k, \kappa_j)\}; \text{ and,} \\
P(I, \gamma \mid \kappa_1, \cdots, \kappa_M) &\propto \pi(I)\pi(\gamma) \prod_{j=1}^{M} P(\kappa_j \mid I, \gamma).
\end{aligned}$$

A random walk Metropolis algorithm can be used to sample from $P(\kappa_j \mid I, \gamma_j, \beta_j, \{\tau_k\}_{k\in G_j})$. With a non-informative prior for β_j, $P(\beta_j \mid \kappa_j, \{\tau_k\}_{k\in G_j})$ becomes an exponential distribution. Sampling from distribution $P(I, \gamma \mid \kappa_1, \cdots, \kappa_M)$ can be achieved by the technique developed in Section 2.3. In this paper, we use BARD_{HM} to denote this modification of BARD with hierarchical model.

2.5 Simulation Studies

2.5.1 Simulation under the BARD model

Let $U = \{1, \cdots, n\}$, of which the first 10% are the relevant entities (i.e., $U_R = \{1, \cdots, [n/10]\}$). We generate the base rankers $\{\tau_k\}_{1\le k\le m}$ via the following scheme:

$$\tau_k = sort(i \in U \text{ by } X_{k,i} \downarrow) \text{ where } X_{k,i} \sim N(0,1) \cdot I(i \in U_B) + N(\mu_k, 1) \cdot I(i \in U_R).$$

We examine two scenarios: (A) $\mu_k = \mu$ for all k, and (B) $\mu_k = \mu \cdot I(k \leq \frac{m}{2})$. In scenario A, the base rankers are equally reliable; in scenario B, however, only the first 50% base rankers are informative. The parameter μ controls the signal strength of the data set (a larger μ means that we have more information to distinguish relevant entities from irrelevant ones). We generate both full lists and top-d lists ($d = 0.2 \cdot n$) for each scenario and test four cases: full lists from scenario A (denoted as A_F), top-d lists from scenario A (denoted as A_P), full lists from scenario B (denoted as B_F), and top-20 lists from scenario B (denoted as B_P).

We first evaluate the impact of signal strength δ on the performance of BARD. Fixing $n = 100$ and $m = 10$, we tried four different values of μ ($\mu = 0.5,\ 1.0,\ 1.5$ and 2.0) for each of the above four cases. Under each configuration, $1,000$ independent datasets were simulated. To each data set, we applied three naive methods (AriM, GeoM, MedR), four Markov-chain-based methods (MC_1, MC_2, MC_3, MC_4), two optimization-based methods (CEMC_F and CEMC_K), and BARD with $\lambda = 1$ under three different choices of the hyperparameter p ($p_1 = 0.05$, $p_2 = 0.10$, and $p_3 = 0.15$), respectively. Additionally, we include BARD with the constraint of equal quality, i.e., $\gamma_1 = \cdots = \gamma_m$ (denoted by BARD_C), in the comparison. For each method, its average coverage rate across the 100 parallel experiments under different configurations is calculated to evaluate the performance. (The coverage rate of an aggregated list is defined as the percentage of true relevant entities covered by the top-10 entities.)

The results are summarized into Table 2.1, from which we can see that: (1) when the quality of base rankers is same (i.e., scenario A), BARD_C slightly outperforms

Table 2.1: Average coverage rates of different rank aggregation methods.

Configuration				Naive methods			MC-based methods				CEMC		BARD_C			BARD		
Case	m	n	δ	AriM	GeoM	MedR	MC_1	MC_2	MC_3	MC_4	$d=F$	$d=K$	p_1	p_2	p_3	p_1	p_2	p_3
A_F	10	100	0.5	0.48	0.47	0.43	0.11	0.45	0.48	0.28	0.46	0.47	0.46	0.44	0.43	0.41	0.40	0.40
A_F	10	100	1.0	0.83	0.82	0.77	0.15	0.79	0.84	0.27	0.78	0.80	0.81	0.80	0.78	0.75	0.74	0.73
A_F	10	100	1.5	0.98	0.98	0.94	0.32	0.95	0.98	0.25	0.89	0.90	0.95	0.96	0.96	0.94	0.95	0.94
A_F	10	100	2.0	1.00	1.00	1.00	0.71	0.99	1.00	0.25	0.93	0.93	0.98	1.00	0.99	0.98	1.00	0.99
A_P	10	100	0.5	0.10	0.13	0.14	0.16	0.15	0.08	0.05	0.38	0.41	0.40	0.41	0.38	0.36	0.37	0.37
A_P	10	100	1.0	0.14	0.22	0.22	0.29	0.27	0.18	0.08	0.72	0.75	0.74	0.73	0.72	0.67	0.68	0.69
A_P	10	100	1.5	0.25	0.37	0.36	0.55	0.54	0.44	0.20	0.91	0.91	0.92	0.92	0.92	0.89	0.89	0.90
A_P	10	100	2.0	0.44	0.58	0.54	0.80	0.80	0.73	0.38	0.96	0.96	0.97	0.99	0.99	0.96	0.98	0.98
B_F	10	100	0.5	0.26	0.26	0.24	0.10	0.26	0.26	0.19	0.25	0.25	0.25	0.25	0.24	0.24	0.24	0.24
B_F	10	100	1.0	0.45	0.49	0.42	0.11	0.48	0.46	0.26	0.45	0.45	0.47	0.46	0.44	0.51	0.51	0.50
B_F	10	100	1.5	0.63	0.70	0.61	0.12	0.70	0.63	0.29	0.63	0.62	0.67	0.65	0.63	0.79	0.79	0.78
B_F	10	100	2.0	0.74	0.84	0.74	0.12	0.84	0.75	0.29	0.74	0.73	0.81	0.78	0.74	0.93	0.94	0.93
B_P	10	100	0.5	0.11	0.13	0.13	0.12	0.13	0.08	0.06	0.23	0.24	0.24	0.24	0.24	0.22	0.23	0.23
B_P	10	100	1.0	0.13	0.17	0.17	0.17	0.17	0.09	0.06	0.43	0.45	0.44	0.43	0.41	0.45	0.45	0.45
B_P	10	100	1.5	0.16	0.21	0.23	0.24	0.23	0.15	0.08	0.63	0.65	0.65	0.63	0.61	0.71	0.72	0.71
B_P	10	100	2.0	0.19	0.26	0.28	0.29	0.28	0.20	0.10	0.80	0.80	0.80	0.78	0.75	0.88	0.88	0.88

Remark: (1) in CEMC, $d = F$ stands for CEMC_F, and $d = K$ stands for CEMC_K; (2) BARD_C stands for BARD with constraint that $\gamma_1 = \cdots = \gamma_m$; (3) for both BARD_C and BARD, we tried 3 values for hyper-parameter p, i.e., $p_1 = 0.05$, $p_2 = 0.10$ and $p_3 = 0.15$ with hyper-parameter $\lambda = 1$.

BARD and achieves a similar performance as CEMC, which is claimed to be "optimal" in this case; (2) when the quality of base rankers varies greatly (i.e., scenario B), BARD uniformly outperforms all the other methods, and the benefit becomes larger with the increase of the signal strength μ; (3) both BARD_C and BARD are robust to the choice of the hyperparameter p. Figure 2.3 displays the box-plots of $\{\bar{\gamma}_k\}_k$ obtained by BARD from the 100 parallel runs under different configurations, suggesting that BARD is capable of efficiently estimating the quality of base rankers when the signal strength is reasonably large (e.g., $\delta \geq 1.0$). We also applied BARD and BARD_C with $\lambda = 2$ to each of the simulated data set and obtained very consistent results, indicting that BARD is robust to hyper-parameter λ.

Second, we check the impact of data size (i.e., n and m) to the performance of

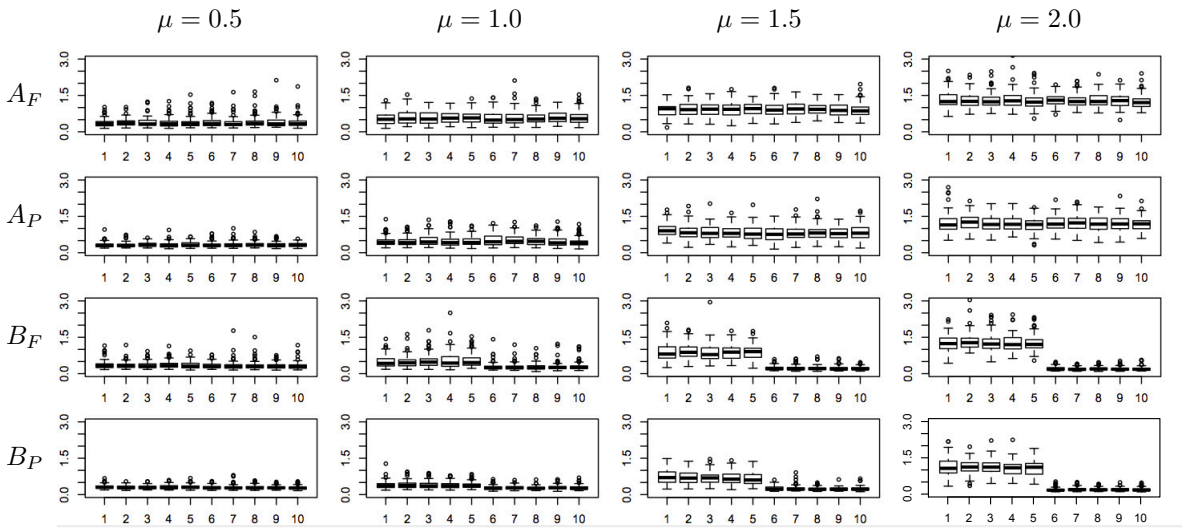


Figure 2.3: The box-plots of $\{\bar{\gamma}_k\}_k$ estimated by BARD from the 100 parallel runs under different configurations when $m = 10$ and $n = 100$ with hyper-parameters $p = 0.1$ and $\lambda = 1$.

Table 2.2: Impact of data size to the performances of different methods.

Configuration				Naive methods			MC-based methods				CEMC		BARD_C			BARD		
Case	m	n	δ	AriM	GeoM	MedR	MC_1	MC_2	MC_3	MC_4	$d = F$	$d = K$	p_1	p_2	p_3	p_1	p_2	p_3
A_F	10	200	1.0	0.84	0.83	0.77	0.18	0.79	0.84	0.33	0.83	0.83	0.81	0.80	0.78	0.77	0.76	0.75
A_F	10	100	1.0	0.84	0.82	0.78	0.15	0.79	0.84	0.26	0.83	0.84	0.82	0.79	0.78	0.74	0.73	0.73
A_F	20	100	1.0	0.96	0.96	0.91	0.10	0.95	0.96	0.32	0.96	0.96	0.94	0.94	0.93	0.89	0.91	0.89
A_P	10	200	1.0	0.13	0.22	0.21	0.35	0.32	0.22	0.09	0.74	0.75	0.73	0.73	0.72	0.68	0.69	0.69
A_P	10	100	1.0	0.13	0.20	0.21	0.23	0.23	0.14	0.06	0.74	0.74	0.73	0.73	0.72	0.68	0.69	0.69
A_P	20	100	1.0	0.17	0.24	0.22	0.64	0.58	0.52	0.20	0.90	0.92	0.90	0.89	0.87	0.84	0.84	0.83
B_F	10	200	1.0	0.46	0.51	0.42	0.11	0.50	0.47	0.27	0.44	0.46	0.49	0.48	0.46	0.55	0.55	0.54
B_F	10	100	1.0	0.46	0.51	0.42	0.11	0.50	0.47	0.29	0.46	0.46	0.49	0.46	0.45	0.51	0.50	0.49
B_F	20	100	1.0	0.63	0.67	0.57	0.10	0.67	0.64	0.33	0.63	0.64	0.63	0.60	0.54	0.69	0.68	0.66
B_P	10	200	1.0	0.13	0.18	0.18	0.17	0.15	0.08	0.05	0.43	0.44	0.45	0.44	0.41	0.47	0.48	0.46
B_P	10	100	1.0	0.13	0.17	0.17	0.15	0.13	0.08	0.05	0.43	0.44	0.43	0.43	0.41	0.46	0.47	0.45
B_P	20	100	1.0	0.15	0.17	0.18	0.43	0.39	0.34	0.15	0.61	0.61	0.59	0.56	0.42	0.62	0.61	0.57

BARD. We fixed the signal strength $\mu = 1.0$, and tried two alternative combinations: $(m, n) = (10, 200)$ and $(m, n) = (20, 100)$. The results are summarized into Table 2.2, from which we can see that most of methods tested are not sensitive to the increase of n, although an increase of m does lead to a better performance for most methods. More importantly, BARD performs quite robust to different choices of n and m compared with the other methods.

2.5.2 Robustness of BARD

An important assumption in our model development is the mutual independence among the rankers in consideration, which can often be violated in real problems. To test how well our method tolerates the violation of this assumption, we simulated 20 rankings $\{\tau_1, \cdots, \tau_{20}\}$ falling into three groups

$$G_1 = \{\tau_1, \tau_2, \tau_3, \tau_4\},\ G_2 = \{\tau_5, \tau_6, \tau_7, \tau_8\},\ G_0 = \{\tau_9, \cdots, \tau_{20}\},$$

where the rankings in G_0 are independently generated, the rankings in G_1 and G_2, however, have very strong within group correlation. More precisely, we let $U = \{1, \cdots, 100\}$; let the relevant entities be $U_R = \{1, \cdots, 10\}$, and let the background entities be composed of two subsets: the "neutral" set $U_{B_1} = \{11, \cdots, 90\}$ and the "negative" set $U_{B_2} = \{91, \cdots, 100\}$. We define $\delta_i = I(i \in U_R) - I(i \in U_I)$, implying that $\delta_i = 1$ for $i \in U_R$, 0 for $i \in U_{B_1}$, and -1 for $i \in U_{B_2}$.

- A ranking τ_k in G_0 is simulated by generating $X_{k,i} \sim N(\delta_i \cdot \mu_k, 1)$, and setting $\tau_k = sort(i \in U \text{ by } X_{k,i} \downarrow)$, where $\mu_k \geq 0$ represents the quality of τ_k since a larger μ_k means that τ_k can better distinguish the relevant entities from background ones;

- The rankings in G_1 and G_2 were generated via two steps: first, we generated a common ranking

$$\kappa = sort(i \in U \text{ by } X_{j,i} \downarrow) \text{ where } X_{j,i} \sim N(\delta_i \cdot \mu, 1);$$

  and then, manipulated κ with random transpositions to generate a group of correlated rankings. Let $\mathcal{M}(\cdot)$ denote a random transposition operation. The aforementioned manipulation can be written as $\tau_k = \mathcal{M}^s(\tau)$ where s is number of such operations used. Note that a small s indicates a stronger correlation among the rankings.

In the simulation, we set $\mu = \mu_9 = \cdots = \mu_{12} = 0.5$, $\mu_{13} = \cdots = \mu_{16} = 1.0$, $\mu_{17} = \cdots = \mu_{20} = 1.5$, and tried three different values (20, 60 and 100) for s. We simulated 1,000 data sets for each configuration. Table 2.3 shows a typical data set simulated

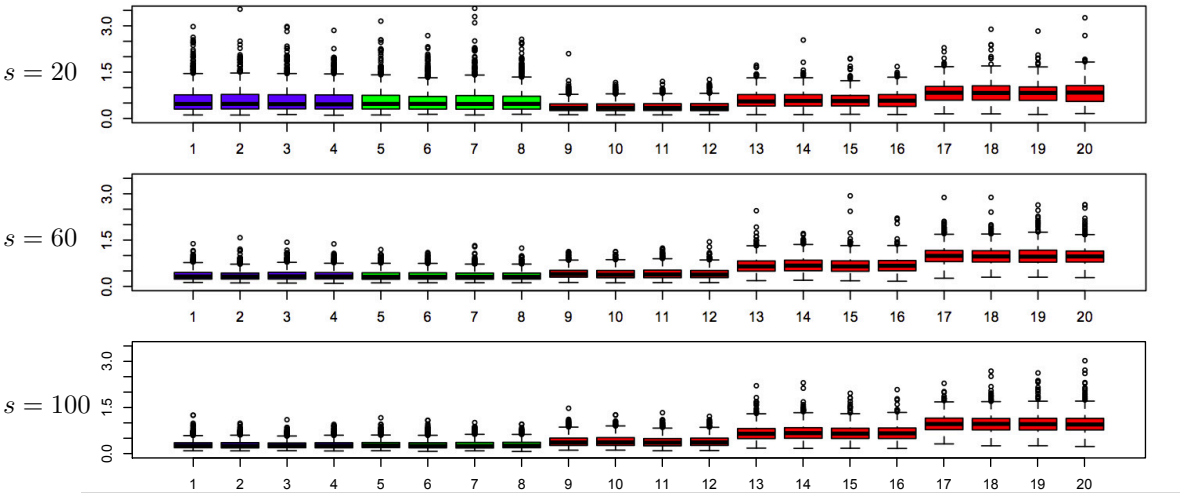


Figure 2.4: Box-plots of $\{\bar{\gamma}_k\}_k$ estimated by BARD from 1,000 parallel runs when some base rankers are dependent of each other. The data sets are simulated from the mechanism described in Section 2.5.2, where the 20 rankers belongs to three blocks $G_1 = \{\tau_1, \cdots, \tau_4\}$, $G_2 = \{\tau_5, \cdots, \tau_8\}$ and $G_0 = \{\tau_9, \cdots, \tau_{20}\}$.

when $s = 60$, from which we can see that the rankings from G_1 or G_2 are quite similar to each other for many entities. We applied BARD, BARD$_{HM}$ as well as other methods to each of these simulated data sets. The results are summarized in Table 2.4 and Figure 2.4. From Table 2.4, we can see that: (1) BARD$_{HM}$ uniformly outperforms all other methods; (2) BARD performs reasonably well even when correlations among the rankers within G_1 and G_2 are very strong (i.e., $s = 20$), and approaches the performance of BARD$_{HM}$ when the correlation is weaker (i.e., $s = 60$ or 100). These results are consistent with the information provided by Figure 2.4, from which we can see that BARD tends to overestimate the quality of the rankers in G_1 and G_2 when the correlation within G_1 and G_2 is very strong (i.e., $s = 20$). All together, these results indicate that BARD$_{HM}$ is efficient to deal with correlated rankers, and BARD is robust to the model assumptions in terms of the average coverage rate.

Table 2.3: A typical simulated data set for testing the robustness of BARD.

	G_1				G_2				G_0											
Entity	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}	τ_{11}	τ_{12}	τ_{13}	τ_{14}	τ_{15}	τ_{16}	τ_{17}	τ_{18}	τ_{19}	τ_{20}
1	57	57	45	27	56	63	56	31	82	4	5	53	22	4	1	69	44	10	6	29
2	14	100	14	15	31	56	67	56	70	42	89	4	15	11	29	2	1	26	28	7
3	27	55	55	87	94	1	5	1	34	89	36	80	12	9	58	35	22	16	78	12
4	4	17	28	14	90	90	86	90	20	5	2	36	6	2	43	11	21	51	59	16
5	55	85	4	55	86	24	90	50	15	63	32	2	21	36	48	23	20	13	21	31
6	5	5	5	75	49	49	32	99	48	22	53	78	13	6	45	17	58	49	1	60
7	73	15	99	25	17	53	13	73	21	67	19	22	1	46	4	19	3	3	16	10
8	31	52	53	57	76	26	17	17	57	83	23	68	3	1	21	76	8	2	30	18
9	22	92	87	10	73	17	26	72	27	30	3	74	16	77	2	1	10	7	4	2
10	62	10	77	77	7	76	76	29	8	18	63	66	32	20	5	91	41	21	5	34
11	70	24	86	24	6	6	16	26	64	38	66	33	47	56	92	36	39	56	45	62
12	41	88	24	86	34	42	6	22	38	58	49	97	36	40	14	55	54	53	81	33
13	8	76	82	34	30	66	3	89	59	72	38	40	25	43	76	26	86	61	15	54
14	25	68	31	42	11	29	11	30	73	68	100	25	94	92	40	46	59	92	32	43
15	79	82	76	82	81	83	48	94	32	12	46	52	68	96	50	59	81	69	10	55
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
86	78	19	50	78	66	41	63	42	87	41	27	92	100	18	20	33	11	88	50	39
87	83	71	19	51	63	7	66	63	9	8	21	15	29	21	88	27	43	23	22	8
88	19	95	10	8	26	30	39	7	30	76	45	26	14	66	62	5	36	65	63	52
89	13	8	78	83	42	44	45	44	96	57	70	95	48	75	10	24	6	45	39	35
90	42	42	34	4	69	69	69	69	100	95	96	29	27	41	61	87	17	93	46	44
91	9	87	1	76	89	14	85	11	75	100	59	58	81	99	57	49	69	82	47	83
92	94	94	72	94	96	87	89	87	83	29	58	82	97	47	98	86	73	100	86	49
93	72	72	12	72	80	85	50	3	36	36	6	69	54	42	65	74	25	85	95	92
94	23	22	18	23	92	92	7	92	90	98	88	20	66	55	56	64	100	40	100	95
95	77	59	21	84	88	88	87	32	52	59	42	72	72	100	44	100	84	79	12	93
96	84	80	46	95	91	2	10	51	97	50	76	24	95	97	87	88	98	84	82	89
97	49	99	84	22	10	91	91	96	84	21	75	65	99	31	83	99	89	55	84	99
98	95	7	3	49	64	55	79	58	33	77	97	89	90	86	84	62	92	99	80	80
99	74	74	95	74	37	58	55	33	85	79	98	55	24	58	51	95	96	98	74	100
100	39	39	39	39	58	37	58	76	53	16	74	96	46	98	78	66	67	97	97	96

Remark: The 100 entities belongs to three subsets $U_R = \{1, 2, \cdots, 10\}$, $U_{B_1} = \{11, \cdots, 90\}$ and $U_{B_2} = \{91, \cdots, 100\}$. The entities in U_R have strong positive signal, the entities in U_{B_2} have strong negative signal, the entities in U_{B_1} do not have strong signal. The 20 rankings fall into three blocks $G_1 = \{\tau_1, \cdots, \tau_4\}$, $G_2 = \{\tau_5, \cdots, \tau_8\}$ and $G_0 = \{\tau_9, \cdots, \tau_{20}\}$. Rankings from different blocks are generated independently, the rankings in block G_0 are generated independently, while the rankings within G_1 or G_2 come from a common ranking with random manipulations. The quality of rankings in G_1 and G_2 is relatively low, while G_0 contains rankings at different quality levels.

Table 2.4: BARD is robust to the assumption of "independent rankers."

	Naive methods			MC-based methods				CEMC		BARD			BARD_{HM}		
s	AriM	GeoM	MedR	MC_1	MC_2	MC_3	MC_4	$d = F$	$d = K$	p_1	p_2	p_3	p_1	p_2	p_3
20	0.74	0.75	0.66	0.11	0.73	0.74	0.32	0.70	0.71	0.72	0.70	0.65	0.87	0.85	0.84
60	0.78	0.81	0.72	0.10	0.80	0.79	0.32	0.75	0.75	0.83	0.84	0.82	0.87	0.86	0.85
100	0.78	0.81	0.72	0.10	0.81	0.78	0.32	0.75	0.75	0.85	0.85	0.83	0.85	0.85	0.85

2.5.3 Discover highly correlated rankers

Next, we test the performance of the proposed coordination coefficient. Figure 2.5 shows the empirical distribution as well as the fitted Beta distribution of Q_i for three typical entities from Table 2.3 (entity 1, 11 and 91), suggesting that the Beta-distribution approximation does effectively capture the key feature of different types of entities. We calculated the Spearman correlation matrix, Kendall correlation matrix and coordination coefficient matrix for the data set shown in Table 2.3, and summarized the results in Figure 2.6 (b). Similar results for other two data sets simulated under different correlation levels ($s = 20$ and 100) are also shown in Figure 2.6. We observe that the proposed method based on the coordination coefficient worked well in all cases, whereas the correlation coefficients were effective only when the dependence is extremely strong.

2.6 Real Data Applications

2.6.1 Aggregating rankings of cancer-related genes

In the first application, we use BARD to aggregate lists of cancer-related genes found in five prostate cancer studies. The first six columns of Table 2.5 present the rankings of the top-25 ranked genes that were found in DeConde et al. (2006) to be

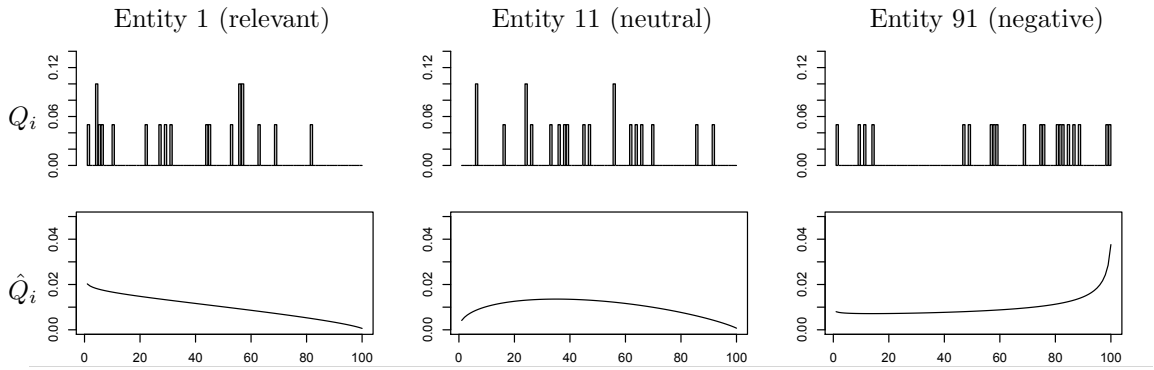
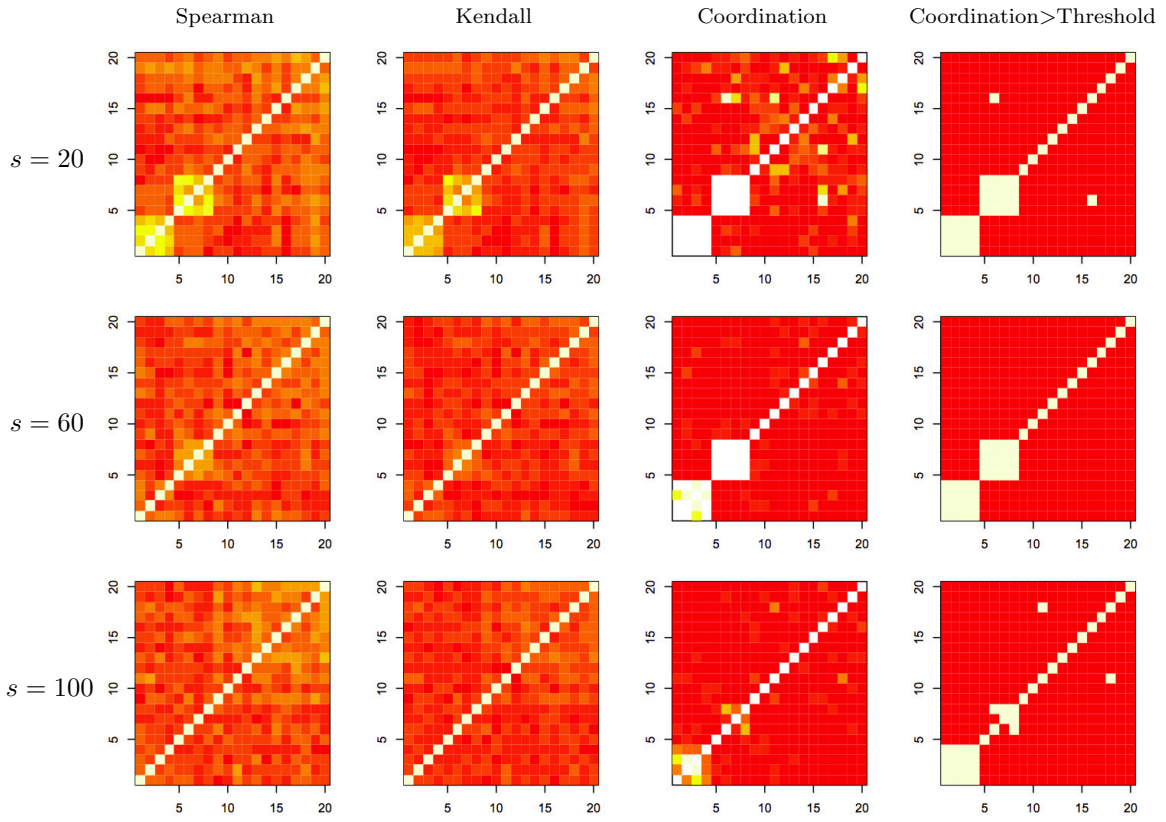


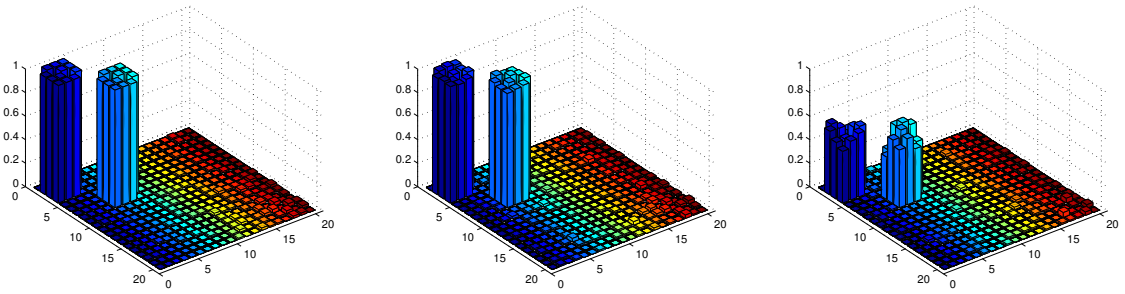
Figure 2.5: The natural distribution of $\{\tau_k(i)\}_{k=1}^m$ and the fitted Beta distribution $\hat{Q}_i$ for three typical entities ($i = 1, 11$ and 91) in Table 2.3.

upregulated in prostate tumors compared to normal prostate tissues from five studies (Dhanasekaran et al., 2001; Luo et al., 2001; Welsh et al., 2001; Singh et al., 2002; True et al., 2006). These five studies rely on different technologies, and their results show that they are quite different in the genes selected to be included in the top-25 list. Lin and Ding (2009) analyzed this dataset, found that the gene list in Luo et al. (2001) is the least common compared to the other four studies, and downgraded its weight in their analysis .

Letting U be the 89 genes appeared in the five top-25 lists, and applying BARD with $\lambda = 1$ to this dataset, we obtain consistent results under different choices of the hyperparameter p ($p = \frac{10}{89}$, $\frac{15}{89}$, and $\frac{20}{89}$). As shown in Table 2.5, the top genes selected by BARD under different configurations do reflect the consensus of the base rankers, and are robust to the choices of p. As illustrated by Figure 2.7, the gene lists from Welsh et al. (2001) and Dhanasekaran et al. (2001) are relatively reliable, while that from Luo et al. (2001) does suffer from low quality. However, the Markov-chain-based methods (MC_1, MC_2, MC_3, MC_4 and MC_T) give very poor results when applied to



(a) Performance of different measurements at different dependence levels



(b) Discovery rate of the coordination-coefficient based method from 100 simulated data sets

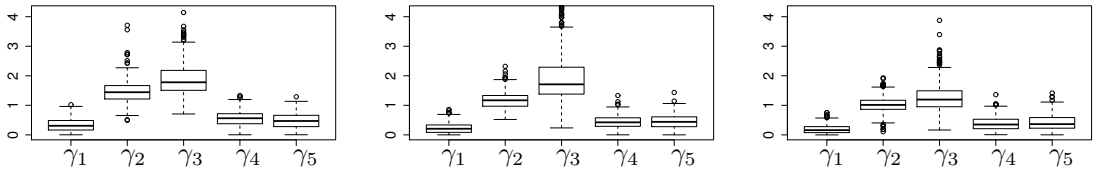
Figure 2.6: Performance of the coordination-coefficient based method for simulated data generated from the mechanism described in section 2.5.2, where the 20 rankings fall into three groups $G_1 = \{\tau_1, \tau_2, \tau_3, \tau_4\}$, $G_2 = \{\tau_5, \tau_6, \tau_7, \tau_8\}$ and $G_0 = \{\tau_9, \cdots, \tau_{20}\}$. The rankings in G_0 are independently generated; the rankings in G_1 and G_2 have strong within group correlation, since each ranking group are generated from a common ranking with s random transposition operations. A smaller s means a stronger within group correlation. We simulated 100 data sets for $s = 20$, 60 and 100 respectively, and applied the proposed method based on the coordination coefficient to each of the 300 simulated data sets. The pair-level discover rates are summarized into figure (b); detailed comparisons with Spearman and Kendall correlation measurements for three typical data sets are illustrated in figure (a). From the figure, we can see that the proposed method works reasonably well for all cases, while the the Spearman or Kendall correlation coefficients are effective only when the dependence is extremely strong.

this dataset: in all these methods, the stationary distribution π of the transition matrix P degenerated to a point mass at gene OGT, i.e., $\pi_i = 1$ if i =OGT and $\pi_i = 0$ for all the other genes, indicating that except OGT, all other genes cannot be effectively distinguished.

Table 2.5: Bayesian rank aggregation of top-25 genes from five prostate cancer studies.

Individual top-25 genes from five prostate cancer studies						*Top genes reported by BARD*				*Original ranks*				
Rank	Luo(L)	Welsh(W)	Dhana(D)	True(T)	Singh(S)	Entity	$\rho_i^{(10)}$	$\rho_i^{(15)}$	$\rho_i^{(20)}$	L	W	D	T	S
1	HPN	HPN	OGT	AMACR	HPN	HPN	1.00	1.00	1.00	1	1	4	2	1
2	AMACR	AMACR	AMACR	HPN	SLC25A6	AMACR	1.00	1.00	1.00	2	2	2	1	-
3	CYP1B1	OACT2	FASN	NME2	EEF2	FASN	1.00	0.98	1.00	-	5	3	-	9
4	ATF5	GDF15	HPN	CBX3	SAT	OACT2	0.99	0.97	0.97	-	3	7	-	-
5	BRCA1	FASN	UAP1	GDF15	NME2	GDF15	0.95	0.97	0.95	-	4	13	5	17
6	LGALS3	ANK3	GUCY1A3	MTHFD2	LDHA	UAP1	0.97	0.97	0.94	-	8	5	-	25
7	MYC	KRT18	OACT2	MRPL3	CANX	OGT	0.96	0.97	0.94	-	-	1	-	-
8	PCDHGC3	UAP1	SLC19A1	SLC25A6	NACA	NME1	0.64	0.93	0.93	14	12	14	9	-
9	WT1	GRP58	KRT18	NME1	FASN	KRT18	0.97	0.97	0.92	-	7	9	-	11
10	TFF3	PPIB	EEF2	COX6C	SND1	STRA13	0.51	0.80	0.82	-	13	11	-	-
11	MARCKS	KRT7	STRA13	JTV1	KRT18	EEF2	0.47	0.89	0.81	-	-	10	14	3
12	OS-9	NME1	ALCAM	CCNG2	RPL15	PPIB	0.17	0.54	0.72	-	10	23	-	-
13	CCND2	STRA13	GDF15	AP3S1	TNFSF10	SLC19A1	0.28	0.65	0.59	-	25	8	-	-
14	NME1	DAPK1	NME1	EEF2	SERP1	CANX	0.03	0.20	0.47	-	16	-	-	7
15	DRRK1A	TMEM4	CALR	RAN	GRP58	GUCY1A3	0.32	0.66	0.46	-	-	6	-	-
16	TRAP1	CANX	SND1	PRKACA	ALCAM	GRP58	0.07	0.24	0.42	-	9	-	-	15
17	FMO5	TRA1	STAT6	RAD23B	GDF15	STAT6	0.01	0.12	0.34	-	-	17	-	-
18	ZHX2	PRSS8	TCEB3	PSAP	TMEM4	NME2	0.04	0.12	0.32	-	-	-	3	5
19	RPL36AL	EMTPD6	EIF4A1	CCT2	CCT2	TCEB3	0.00	0.08	0.31	-	-	18	-	-
20	ITPR3	PPP1CA	LMAN1	G3BP	SLC39A6	TMEM4	0.01	0.07	0.28	-	15	-	-	18
21	GCSH	ACADSB	MAOA	EPRS	RPL5	CALR	0.01	0.15	0.27	-	-	15	-	-
22	DDB2	PTPLB	ATP6V0B	CKAP1	RPS13	SND1	0.01	0.17	0.25	-	-	16	-	10
23	TFCP2	TMEM23	PPIB	LIG3	MTHFD2	EIF4A1	0.00	0.06	0.25	-	-	19	-	-
24	TRAM1	MRPL3	FMO5	SNX4	G3BP2	ANK3	0.06	0.10	0.24	-	6	-	-	-
25	YTHDF3	SLC19A1	SLC7A5	NSMAF	UAP1	MAOA	0.00	0.07	0.22	-	-	21	-	-

Remark: Totally, 89 distinct genes appear in the top-25 lists of the five studies, which are referred to as Luo, Welsh, Dhana, True, and Singh, respectively. And, $\rho_i^{(k)}$ stands for vector ρ obtained from BARD with hyperparameter $p = \frac{k}{89}$.



(a) $p = \frac{10}{89}$ (b) $p = \frac{15}{89}$ (c) $p = \frac{20}{89}$

Figure 2.7: Posterior distributions of $\{\gamma_1, \cdots, \gamma_5\}$ obtained from BARD under different hyper-parameter p for the dataset of cancer-related genes.

2.6.2 Aggregating rankings of NBA teams

Ranking sports teams has attracted tremendous attention from both sports analysts and academics. Numerous ranking methods have been proposed for different sports, including NBA, NFL, MLB, NCAA football, etc. (see Langville and Meyer 2012 for a comprehensive review). Our BARD method produces an aggregated ranking considering the results of any number of ranking methods, which can be used to provide better predictions of game outcomes and to evaluate the effectiveness of different sports statistics in generating rankings.

Here, we use BARD to aggregate power rankings of NBA teams. We collected 6 rankings from professional sports web sites and 28 rankings generated by a small group of Harvard students for the 30 NBA teams in the 2011-2012 season. The 6 professional rankings are generated in December 2011 after the preseason games of the 2011-2012 season. The 28 amateur rankings were generated as follows in July 2012 after the 2011-2012 season: we sent out a questionnaire to all graduate students in the Department of Statistics and a small group of summer school students who were taking the summer course STAT 100 at Harvard, asking them to select the best 8 NBA teams of the 2011-2012 season and rank them top-down based on his/her own knowledge or opinion without checking online information or consulting others.

We also asked each student to classify himself/herself into one of the following four groups in the survey: (1) "Avid fans" who never missed NBA games, (2) "Fans" who watched NBA games frequently, (3) "Infrequent watchers" who watched NBA games occasionally, and (4) the "Not-interested" who never watched NBA games in the past season. We received 28 responses, amounting to a 47% response rate. The data are displayed in Table 2.6. We expect BARD to give higher γ_ks to rankings from professional web sites and students who paid more attention to NBA games. Moreover, using the ranking of these teams in 2011-2012 playoffs as a surrogate of the unknown "true" power ranking of these teams, we can evaluate the performance of BARD in a quantitative way.

We applied BARD to the dataset with hyperparameter $p = \frac{16}{30}$ and $\lambda = 1$ to "predict" which teams can make their appearance in the playoffs. (Each season, 16 teams enter the playoffs based on their performances in the regular season.) The results are summarized into Figure 2.8. From sub-figure (a), we can see that BARD figures out the quality difference among different rankers successfully: the boxplots of γ_is show a clear decreasing trend with the decrease in knowledge level of the rankers. We also observed an interesting phenomena from these boxplots: ranker 11 (i.e., S_5) is an outlier in the group of *Avid fans*, which precisely reflects the fact that S_5 gave high ranks to Warriors and Wizards, two teams that failed to enter the playoffs.

Moreover, the aggregated ranking does outperform individual rankings in terms of being closer to the "truth," even though the amateur rankings from the students contain considerable amount of noise. The aggregated ranking makes only one mistake: putting Rockets instead of Jazz into the playoffs list. Among the six professional

Table 2.6: Power rankings of NBA teams for the 2011-2012 season collected from 6 professional sport-ranking web sites and a survey of 28 Harvard students.

		Professional						Avid fans						Fans						Infrequent watchers							Not-interested individuals								
N.o.	Team	P_1	P_2	P_3	P_4	P_5	P_6	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}	S_{17}	S_{18}	S_{19}	S_{20}	S_{21}	S_{22}	S_{23}	S_{24}	S_{25}	S_{26}	S_{27}	S_{28}
1	Heat	1	2	1	1	1	1	1	1	2	3	1	3	-	1	2	1	3	1	3	1	4	1	4	2	1	2	-	4	-	-	1	1	1	2
2	Thunder	3	3	2	3	2	3	2	2	3	2	-	2	-	-	4	2	7	4	-	-	-	2	-	-	2	-	-	-	-	2	-	-	-	-
3	Spurs	7	10	11	5	8	7	6	5	5	5	-	6	-	6	5	4	-	-	5	-	8	6	3	6	-	-	-	-	-	-	-	-	-	-
4	Celtics	5	11	10	9	9	5	4	8	1	4	2	5	2	3	1	3	4	3	-	4	2	3	2	-	4	4	-	2	-	-	-	2	-	-
5	Clippers	8	5	6	10	5	6	-	6	-	8	-	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	4	-
6	Lakers	6	7	7	6	6	8	3	7	6	1	3	1	1	2	7	7	1	2	1	2	1	4	5	1	3	1	5	1	-	-	8	4	2	-
7	Pacers	14	13	14	14	13	12	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	-	-	-	5	8	-	-
8	76ers	15	16	13	15	15	15	-	-	-	-	-	-	-	4	-	-	8	6	-	-	-	-	-	5	6	-	7	6	6	4	3	3	-	-
9	Mavericks	2	1	3	2	4	4	-	-	7	7	-	8	-	-	8	6	6	5	-	8	3	-	1	-	-	-	3	-	-	-	4	-	5	-
10	Bulls	4	4	4	4	3	2	5	4	8	6	4	4	3	-	3	-	-	8	4	-	-	-	-	-	-	3	-	3	3	5	-	5	-	1
11	Knicks	9	6	9	8	7	13	-	3	4	-	5	-	-	-	-	-	2	-	-	-	-	-	-	4	7	8	-	-	2	8	-	7	7	6
12	Grizzlies	10	8	8	7	11	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	-	-	-	-	-
13	Nuggets	19	9	5	13	10	9	-	-	-	-	8	-	-	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-	-	-	8	-
14	Magic	11	12	17	11	14	11	-	-	-	-	-	-	6	-	-	-	-	-	-	5	-	-	-	-	5	-	4	-	-	-	6	6	-	-
15	Hawks	12	18	12	18	12	18	7	-	-	-	-	-	-	-	-	5	-	-	-	-	-	-	-	-	8	7	-	-	5	-	-	-	-	8
16	Jazz	18	23	26	27	28	19	-	-	-	-	-	-	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
17	TrailBlazers	13	14	15	12	16	14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18	Rockets	21	15	16	16	17	17	-	-	-	-	-	-	5	7	6	-	-	-	2	6	-	-	-	3	-	5	-	5	-	3	-	-	-	7
19	Bucks	16	17	20	17	20	16	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	-	-	-	-	-	-	-	-	6	-	-	-	-
20	Suns	20	22	19	21	19	21	-	-	-	-	-	-	-	8	-	-	-	-	7	7	6	8	8	-	-	-	-	-	-	-	7	-	-	-
21	Nets	17	19	24	20	24	23	-	-	-	-	-	-	-	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	7	-	-	-	3	5
22	Warriors	22	21	23	19	22	20	-	-	-	-	6	-	-	-	-	-	-	-	6	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-
23	Timberwolves	23	20	22	22	23	24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7	7	-	-	-	-	-	-	4	7	-	-	-	4
24	Hornets	27	28	18	23	18	25	-	-	-	-	-	-	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
25	Pistons	25	25	25	24	25	22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	-	-	-	8	-	-	2	-	-	-
26	Kings	29	24	21	26	21	26	-	-	-	-	-	-	4	5	-	-	-	-	8	-	5	5	-	-	-	-	-	-	1	-	-	-	-	-
27	Wizards	28	27	28	25	27	27	-	-	-	-	7	-	-	-	-	-	-	-	-	-	-	-	-	7	-	-	-	7	-	-	-	-	6	-
28	Raptors	24	26	29	28	30	28	-	-	-	-	-	-	-	-	-	8	-	-	-	-	-	-	-	-	-	-	8	-	-	1	-	-	-	-
29	Cavaliers	26	29	27	29	26	29	-	-	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-	6	-	-	-	-	-	-	-	-
30	Bobcats	30	30	30	30	29	30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	-	-	-	-	-	-	-	-	-	-	-

Remark: The 30 NBA teams are arranged in the table based on their performance in the playoffs of the season, i.e., the top 16 teams reached the playoffs, the top 8 teams survived the first round of the playoffs, and so on. The 6 professional power rankings $(P_1, \cdots, P_6)$ are downloaded from FOXSports.com, ESPN.com, SI.com, NBA.com, midwestsportsfans.com, and jsonline.com, respectively. They are based on the preseason games before the the regular 2011-2012 season. (More details about these professional power rankings are listed in the Appendix A.3.) The 28 rankings by Harvard students $(S_1, \cdots, S_{28})$ are collected by a survey after the 2011-2012 season was finished, in which each student was asked to select the best 8 NBA teams in the 2011-2012 season and rank them top-down based on his/her own knowledge without checking online information or consulting others. To collect information about how much the students followed NBA games in the 2011-2012 season, we also asked every student to classify himself/herself into one of the following four groups in the survey: (1) "Avid fans" who never missed NBA games, (2) "Fans" who watched NBA games frequently, (3) "Infrequent watchers" who watched NBA games occasionally, and (4) the "Not-interested" who never watched NBA games in the past season. In addition, we encouraged the students to do random guess if they really have no ideas about these teams.

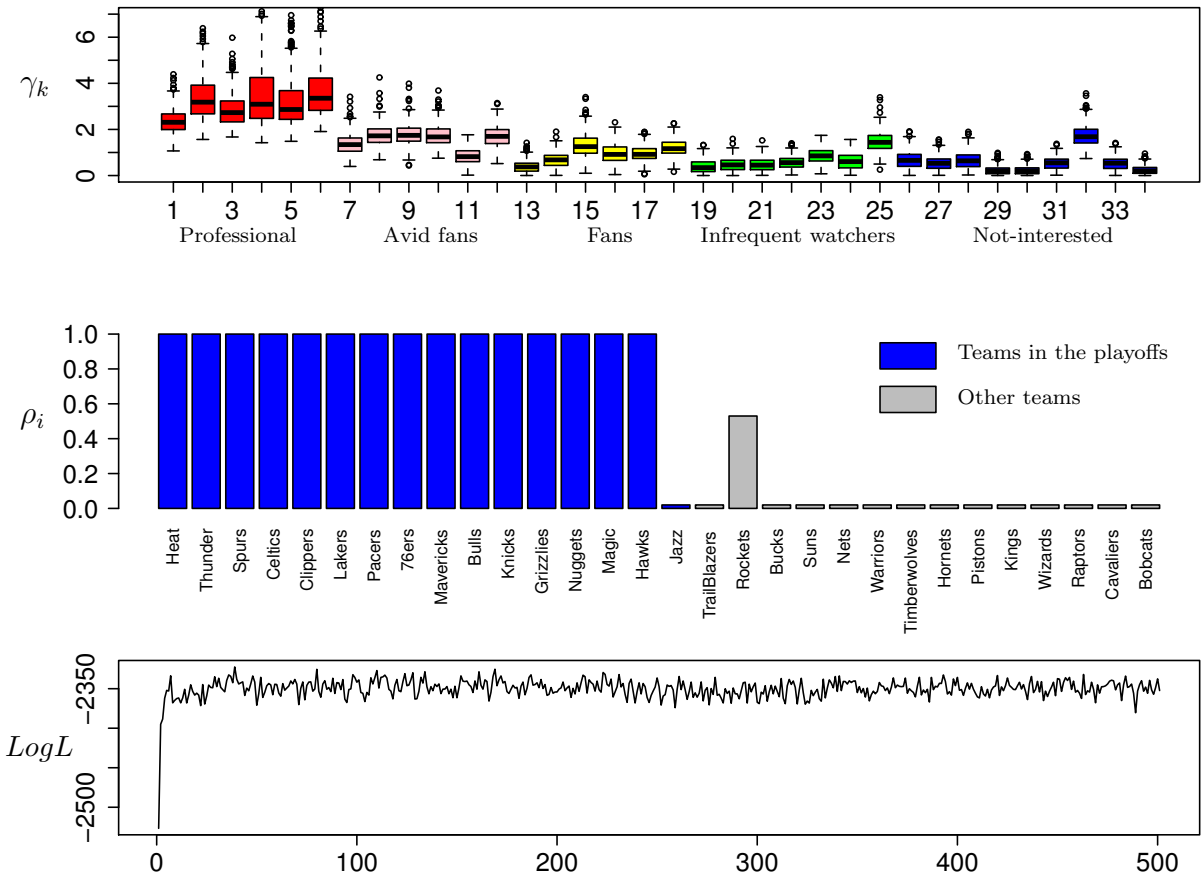


Figure 2.8: Results from BARD for aggregating 34 rankings of 30 NBA teams in season 2011-2012.

rankings, however, only P_5 achieves the same result as the aggregated ranking; the other five rankings make two mistakes: P_1 misses Nuggets and Jazz, P_3 misses Magic and Jazz, P_2, P_4 and P_6 miss Hawkes and Jazz.

2.7 Discussion

In this chapter, we propose the Bayesian rank aggregation (BARD) method for the rank aggregation problem. By giving each base ranker a specific quality parameter and estimating these parameters using the data, BARD measures the reliability of

the base rankers in a quantitative way and makes use of this information to improve the aggregated rank list. Compared to the methods in the literature, BARD works significantly better when the equality of base rankers varies greatly. Both simulation studies and real data applications demonstrate the usefulness and superiority of BARD.

BARD assumes that (1) the entities involved can potentially be divided into two subsets, the relevant entities U_R and the background entities U_B; (2) given the group indicators of entities $I = \{I_i\}_{i \in U}$, the rankers $\tau_1, \cdots, \tau_m$ are conditionally independent; (3) in each base ranker τ_k, the internal ranks of the relevant/background entities are assigned randomly, the rank of a relevant entity among the background entities, however, follows a power law distribution. To apply BARD to a practical problem, we need to check whether the above assumptions (especially, the first two) hold approximately. BARD is reasonably robust if the dependence among base rankers is not strong. However, if the dependence is very strong, BARD may report biased result. The methods developed in Section 2.4 provide diagnosis tools for this critical issue. Facing a practical problem, we should try to understand the mechanism behind the base rankers as much as we can to make a good judgement whether the base rankers work independently or not.

BARD also requires that all base rankers involved have a common objective, that is to distinguish the relevant entities from the background entities by giving the former ones higher ranks than the later ones. If the data collected in practice do not satisfy this requirement (e.g., the rankings from different base rankers have different goals, or are purely based on the opinion of the rankers themselves), BARD may not be an

appropriate tool to use.

In general, BARD is robust to different choices of p, the expected percentage of relevant entities in U, when p comes from a proper range (e.g., $[0.01, 0.2]$). In some practical problems, the choice of p is obvious. If not, we need to try different choices of p from a proper range and check how robust the results are before a conclusion can be made.

The framework of BARD supports us to deal with full rankings, partial rankings and rankings with ties as well. It is possible to further extend this framework to problems with more complicated structures. For example, if some covariates of the entities of interest are also observed, which can potentially influence the rankings of some base rankers, it will be desirable to link these covariates to the quality parameters of corresponding base rankers to achieve a better performance.

Remark: this Chapter is based on a published paper:
Deng K., Han S., Li J.K., and Liu J.S. Bayesian Aggregation of Order-Based Rank Data. (2014) *JASA*. Published online: Jan 14, 2014. DOI: 10.1080/01621459.2013.878660.

Appendix A

Appendix

A.1 Technical Details of the Semi-Bayesian Approach for Indirect Comparison without Individual-Level Data

When both π_{θ_N} and π_{θ_O} are multivariate Gaussian distributions, the following conditional distributions can be derived from the posterior distribution (1.22) Gibbs sampler:

$$
\begin{aligned}
U_i^{\mathcal{T}_N} \mid others &\sim N(\alpha_0 + \alpha_1 X_1^{\mathcal{T}_N} + \cdots + \alpha_p X_p^{\mathcal{T}_N}, 1) \cdot \left[I(Y_i^{\mathcal{T}_N}, U_i > 0) + I(Y_i^{\mathcal{T}_N} = 0, U_i \leq 0) \right] \\
\alpha_j \mid others &\sim N\Big(\frac{\sum X_{ij}^{\mathcal{T}_N} (U_i - X_{i[-j]}^{\mathcal{T}_N} \alpha_{[-j]})}{\sum (X_{ij}^{\mathcal{T}_N})^2}, \frac{1}{\sum (X_{ij}^{\mathcal{T}_N})^2} \Big) \\
\theta_O \mid others &\propto \prod_{i \in \mathcal{A}_O^-} \pi_{\theta_O}(X_i) \cdot \pi(\theta_O) \text{ Population estimation can be found in Appendix A.2}
\end{aligned}
$$

$$
\begin{aligned}
E_{\pi_O} E[Y^{\mathcal{T}_N} \mid X^{\mathcal{T}_N}, T = t] &= E_{\pi_O}[g_1(\alpha_0 + \alpha_1 X_1^{\mathcal{T}_N} + \cdots + \alpha_p X_p^{\mathcal{T}_N})] \\
&= E_{\pi_O}[\Phi(\alpha_0 + \alpha_1 X_1^{\mathcal{T}_N} + \cdots + \alpha_p X_p^{\mathcal{T}_N})] \\
&= \int_{\pi_O} \Phi(\alpha_0 + \alpha_1 X_1^{\mathcal{T}_N} + \cdots + \alpha_p X_p^{\mathcal{T}_N}) dx
\end{aligned}
$$

Procedures:

- Get one sample for all the parameters $\alpha_0, \alpha_1, \cdots, \alpha_p, \Theta_1, \Theta_2$
- For each sample, calculate $\int_{\pi_2} \Phi(\alpha_1 X_1^{\mathcal{T}_N} + \cdots + \alpha_p X_p^{\mathcal{T}_N} + \alpha) dx$ by Monte Carlo integration, i.e.,
  1. Sample 1,000 X from $\pi_O(X \mid \Theta_2)$
  2. Calculate $\Phi(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p)$ for each sample
  3. Take the average of above quantity
- Repeat above steps 1,000 times to get 1,000 samples from the posterior distribution of estimand $E_{\pi_O} E[Y^{\mathcal{T}_N} \mid X, T = t]$

A.2 Estimation of Population Distribution

A.2.1 Independence Assumption

At the first stage, we assume covariates are independent of each other. We will infer the underlying distribution (posterior distribution of parameters) of each covariate in both trials based on the summary statistics. And posterior of propensity weights can be calculated based on the distribution of covariates in both trials. Therefore the

posterior distribution of adjusted efficacy can be inferenced. Assume all the covariates are independent of each other, and follow the following model:

$$\begin{aligned}
X_{ij}^{\mathcal{T}_N} &\sim N(\mu_j^{\mathcal{T}_N}, [\sigma_j^{\mathcal{T}_N}]^2), \quad i = 1, \ldots, n_N \quad j = 1, \ldots, p \\
X_{ij}^{\mathcal{T}_O} &\sim N(\mu_j^{\mathcal{T}_O}, [\sigma_j^{\mathcal{T}_O}]^2), \quad i = 1, \ldots, n_O \quad j = 1, \ldots, p \\
X_{ij}^{\mathcal{T}_N} &\sim Bern(c_j^{\mathcal{T}_N}), \qquad i = 1, \ldots, n_N \quad j = p+1, \ldots, p+p_d \\
X_{ij}^{\mathcal{T}_O} &\sim Bern(c_j^{\mathcal{T}_O}), \qquad i = 1, \ldots, n_O \quad j = p+1, \ldots, p+p_d
\end{aligned} \tag{A.1}$$

Given non-informative prior $f(\mu, \sigma^2) \propto \sigma^{-2}, f(c) \sim Beta(\alpha, \beta)$, because sample mean and sample variance are sufficient statistics for normal distribution It is easy to get the posterior distribution: $(\mathcal{T} = \mathcal{T}_N, \mathcal{T}_O)$

$$\begin{aligned}
[\sigma_j^{(\mathcal{T})}]^2 | X_{j=1,\ldots,p}^{(\mathcal{T})} &\sim Inv - \chi^2(n_N - 1, \textstyle\sum_{j=1}^{p}((x_{ij}^{(\mathcal{T})} - \bar{x}_{\cdot j}^{(\mathcal{T})})^2)/(n_N - 1) \quad j = 1, \ldots, p \\
\mu_j^{(\mathcal{T})} | [\sigma_j^{(\mathcal{T})}]^2, X_{j=1,\ldots,p}^{(\mathcal{T})} &\sim N(\bar{x}_{\cdot j}^{(\mathcal{T})}, [\sigma_j^{(\mathcal{T})}]^2/n_N) \quad j = 1, \ldots, p \\
c_j^{(\mathcal{T})} | X_{p+1,\ldots,p+p_d}^{(\mathcal{T})} &\sim Beta(\alpha + \textstyle\sum_i x_{ij}^{(\mathcal{T})}, \beta + \sum_i (1 - x_{ij}^{(\mathcal{T})})) \quad j = p+1, \ldots, p+p_d
\end{aligned} \tag{A.2}$$

A.2.2 Correlation Matrix is same in both population for continuous covariates

In the previous method, all variables are assumed independently. However, in most of the cases, the covariance matrix may not be the identity matrix. Since full observation is only available in one of the two population, how to model the covariance matrix is a question to us.

As we all know that in the problem of matching, covariates in two distributions

are same, the correlation between variables shouldn't be changed a lot. Therefore, a reasonable assumption will be made here: Correlation matrix for both distributions are identical.

$$\begin{aligned}
X_i^{\mathcal{T}_N}|\mu,\Sigma &\sim N(\mu,\Sigma) \quad j=1,\ldots,p \\
X_i^{\mathcal{T}_O}|D,\eta,\Sigma &\sim N(\eta, D^{\frac{1}{2}}\Sigma D^{\frac{1}{2}}) \quad j=1,\ldots,p \\
\Sigma &\sim Inv-Wishart(\Lambda_0) \\
\mu|\Sigma &\sim N(\mu_0,\Sigma/\kappa_0) \\
\eta|\Sigma,D &\sim N(\eta_0, D^{\frac{1}{2}}\Sigma D^{\frac{1}{2}}/\kappa_0)
\end{aligned} \tag{A.3}$$

Where $D = diag(r_1^2, r_2^2, \ldots, r_p^2)$.

Given flat prior

$$\begin{aligned}
f(\mu,\Sigma,\eta,D) \propto\ & |\Sigma|^{-\frac{v_0+p+1}{2}} \exp[-\frac{1}{2}tr(\Lambda_0\Sigma^{-1})] \\
& |\Sigma/\kappa_0|^{-1/2} \exp[-\frac{1}{2}(\mu-\mu_0)^T(\Sigma/\kappa_0)^{-1}(\mu-\mu_0)] \\
& |D^{\frac{1}{2}}\Sigma D^{\frac{1}{2}}/\kappa_0|^{-1/2} \exp[-\frac{1}{2}(\eta-\eta_0)^T(D^{\frac{1}{2}}\Sigma D^{\frac{1}{2}}/\kappa_0)^{-1}(\eta-\eta_0)] \\
& |D|^{-1}
\end{aligned}$$

Therefore, for $j=1,\ldots,p$

$$\begin{aligned}
\Sigma|D,X_j^{\mathcal{T}_N},X_j^{\mathcal{T}_O} &\sim Inv-Wishart(n_N+n_O+v_0,\Lambda_n^{-1}) \\
f(r_k|r_{[-k]},\Sigma,X_j^{\mathcal{T}_N},X_j^{\mathcal{T}_O}) &\propto (r_k^2)^{-\frac{n_O+1}{2}} \exp[-\tfrac{1}{2}(\tfrac{A_{kk}B_{kk}}{r_l^2} + 2\textstyle\sum_{l\neq k}\frac{A_{kl}B_{kl}}{\sqrt{r_l^2}}\frac{1}{\sqrt{r_k^2}})] \\
\mu|\Sigma,X_j^{\mathcal{T}_N},X_j^{\mathcal{T}_O} &\propto N(\tfrac{n_N\bar{X}_{\mathcal{T}_N}+\kappa_0\mu_0}{n_N+\kappa_0}, \Sigma/(n_N+\kappa_0)) \\
\eta|D,\Sigma,X_j^{\mathcal{T}_N},X_j^{\mathcal{T}_O} &\propto N(\tfrac{n_O\bar{X}_{\mathcal{T}_O}+\kappa_0\eta_0}{n_O+\kappa_0}, D^{\frac{1}{2}}\Sigma D^{\frac{1}{2}}/(n_N+\kappa_0))
\end{aligned} \tag{A.4}$$

Where,

$$
\begin{aligned}
\Lambda_n &= S_{\mathcal{T}_N} + \Lambda_0 + \frac{n_N \kappa_0}{n_N + \kappa_0}(\bar{X}_{\mathcal{T}_N} - \mu_0)(\bar{X}_{\mathcal{T}_N} - \mu_0)^T \\
&\quad + D^{-\frac{1}{2}}(S_{\mathcal{T}_O} + \frac{n_O \kappa_0}{n_O + \kappa_0}(\bar{X}_{\mathcal{T}_O} - \eta_0)(\bar{X}_{\mathcal{T}_O} - \eta_))^T)D^{-\frac{1}{2}}) \\
S_{\mathcal{T}_N} &= \sum_{i \in \mathcal{T}_N}(X_i^{\mathcal{T}_N} - \bar{X}_{\mathcal{T}_N})((X_i^{\mathcal{T}_N} - \bar{X}_{\mathcal{T}_N}))^T \\
S_{\mathcal{T}_O} &= \sum_{i \in \mathcal{T}_O}(X_i^{\mathcal{T}_O} - \bar{X}_{\mathcal{T}_O})((X_i^{\mathcal{T}_O} - \bar{X}_{\mathcal{T}_O}))^T \\
A &= \Sigma^{-1} \\
B &= S_{\mathcal{T}_O} + \frac{n_O \kappa_0}{n_O + \kappa_0}(\bar{X}_{\mathcal{T}_O} - \eta_0)(\bar{X}_{\mathcal{T}_O} - \eta_0))^T
\end{aligned}
$$

How to derive $S_{\mathcal{T}_O}$

In the above equations, everything can be easily derived except $S_{\mathcal{T}_O}$. As we know that

However, $S_{\mathcal{T}_O}$ is not observed. But because of BASU theorem, sample correlation is independent of sample mean and sample variance, $X^{\mathcal{T}_O}$ is easily imputed from $N(\eta, D^{\frac{1}{2}}\Sigma D^{\frac{1}{2}})$ and then resealed to match the sample mean and sample variance. $S_{\mathcal{T}_O}$ can be easily computed via the imputed data.

A.3 Detailed information about the professional rankings of NBA teams used in Section 2.6.2.

Ranking	Provider	Link
P_1	FOXSports.com	http://msn.foxsports.com/nba/powerRankings/2011-2012/PRE
P_2	ESPN.com	http://espn.go.com/nba/powerrankings/_/week/0
P_3	SI.com	http://sportsillustrated.cnn.com/2011/writers/britt_robson/12/20/preseason.power.rankings/index.html
P_4	NBA.com	http://www.nba.com/2011/news/powerrankings/12/21/preseason/index.html
P_5	midwestsportsfans.com	http://www.midwestsportsfans.com/2011/12/nba-power-rankings-preseason-edition/
P_6	jsonline.com	http://www.jsonline.com/sports/136175388.html

Bibliography

Ahmad N., and Beg M. M. S. (2002), Fuzzy Logic Based Rank Aggregation Methods for the World Wide Web, In *Proceedings of the International Conference on Artificial Intelligence in Engineering and Technology*, Malaysia, 2002, 363-368.

Aslam, J. A., and Montague, M. (2001), Models for Metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 276-284.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61, 962-972.

Beg, M. M. S. (2004), Parallel Rank Aggregation for the World Wide Web. *World Wide Web*. Kluwer Academic Publishers, vol 6, issue 1, 5-22. March 2004.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. The Annals of Statistics, 9, 1196-1217.

Borda, J. C. (1781), Mémoire sur les élections au scrutin. *Histoire del' Académie Royale des Sciences*, 1781

Bucher, H. C., Guyatt, G. H., Griffith, L. E., et al. (1997). The results of direct and

indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol, 50, 683-91

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics, 24, 295-313.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Earlbaum, Hillsdale, NJ.

Czajka, J. C., Hirabayashi, S., Little, R. and Rubin, D. B. (1992). Projecting from advance data using propensity modeling. J. Bus. Econom. Statist.,10, 117-131.

Davison, A. C. (1997).Bootstrap methods and their application. Cambridge University Press.

DeConde, R. P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., and Etzioni, R. (2006), Combining results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology* 5, article 15.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. J. Amer. Statist. Assoc., 94, 1053-1062.

Deng K., Han S., Li J.K., and Liu J.S. (2014) Bayesian Aggregation of Order-Based Rank Data. J. Amer. Statist. Assoc., Published online: Jan 14, 2014

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. S., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* 412, 822-826.

Diaconis, P. (1988), Group Representation in Probability and Statistics. *IMS Lecture Series* 11, IMS, 1988.

Diaconis, P., and Graham, R. (1977), Spearman's footrule as a measure of disarrary. *Journal of the Royal Statistical Society*, Series B, 39(2): 261-268.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001), Rank Aggregation Methods for the Web. In *Proceedings of the 10th international conference on World Wide Web*, 613-622.

Eckert, L. and Falissard, B. (2006) Using meta-regression in performing indirect-comparisons: comparing escitalopram with venlafaxine XR. Curr Med Res Opin, 22 (11): 2313-21.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7, 1-26.

Efron, B. and Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman and Hall. New York, NY.

Fagin, R., Kumar, R., and Sivakumar, D. (2003), Comparing top k lists. *SIAM Journal of Discrete Mathematics* 17, 134-160.

Fagin, R., Kumar, R., and Sivakumar, D. (2003b), Efficient Similarity Search and Classification via Rank Aggregation. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 301-312.

Fagin, R., Lotem, A., and Naor, M. (2001), Optimal Aggregation Algorithm for Mid-

dleware. In *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 102-113.

Fligner, M. A., and Verducci, J. S. (1986), Distance based ranking models. *Journal of the Royal Statistical Society*, Series B, 48, 359-369.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003), An Efcient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, 4, 933-969.

Glenny, A. M., Altman, D. G., Song, F., et al. (2005) Indirect comparisons of competing interventions. Health Technol Assess, 9 (26): 1-134.

Gu, X. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. J. Comput. Graph. Statist. 2, 405-420.

Hall, P. (1992). The Bootstrap and Edgeworth Expansion. Springer-Verlag. New York, NY.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. J. Amer. Statist. Assoc., 99, 609-618.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71, 1161-1189.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15 199-236.

Hull, D. A., Pedersen, J. O., and Schütze, H. (1996), Method Combination for Document Filtering. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 279-287.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. Biometrika, 87, 706-710.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and Statistics, 86, 4-29.

Jiang, B. and Liu, J. S. (2013). Sliced Inverse Regression with Variable Selection and Interaction Detection. arXiv preprint arXiv:1304.4056.

Lam, K. W., and Leung, C. H. (2004), Rank Aggregation for Metasearch Engines. In *Proceedings of the 13th International Conference on World Wide Web*, 384-385.

Langville, A.N., and Meyer, C.D. (2012), *Who's #1?: The Science of Rating and Ranking*. Princeton University Press, Princeton, NJ.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13-22.

Lin, S. L., and Ding J. (2009), Integration of Ranked Lists via Cross Entropy Monte Carlo with Applications to mRNA and microRNA Studies. *Biometrics* 65, 9-18.

Liu, J. S. (2001), Monte Carlo Strategies in Scientific Computing. *Springer Series in Statistics*, Springer-Verlag, Newyork.

Liu, Y., Liu, T., Qin, T., Ma, Z., and Li, H. (2007), Supervised rank aggregation. In *Proceedings of the 16th international conference on World Wide Web*, 481-490.

Lumley T. (2002) Network meta-analysis for indirect treatment comparisons. Stat Med, 21, 2313-24.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Stat. Med., 23, 2937-2960.

Luo, J., Duggan, D. J., Chen, Y., Sauvageot, J., Ewing, M., Bittner, M. L., Trent, J. M., and Isaacs, W. B. (2001). Human prostate cancer and benign prostatic hyperplasia: Molecular dissection by gene expression profiling. *Cancer Research* 61, 4683-4688.

Mallows, C. L. (1957), Non-null ranking models. *Biometrika*, 44, 114-130.

Manmatha, R., and Sever, H. (2002), A Formal Approach to Score Normalization for Meta-search. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, 98-103.

Manmatha, R.., Rath, T., and Feng, F. (2001), Modeling Score Distributions for Combining the Outputs of Search Engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 267-275.

Meila, M., Phadnis, K., Patterson, A., and Bilmes, J. (2007), Consensus ranking under the exponential model. In *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence* (UAI), 2007

Ming, K. and Rosenbaum, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. J. Comput. Graph. Statist., 10, 455-463.

Montague, M., and Aslam, J. A. (2001), Relevance Score Normalization for Metasearch. In *Proceedings of the 10th Conference on Information and Knowledge Management*, 427-433.

Moore, R. A., McQuay, H. J. (1997). Single-patient data metaanalysis of 3453 postoperative patients: oral tramadol versus placebo, codeine and combination analgesics. Pain, 69: 287-94.

Nixon, R. M., Bansback, N., Brennan, A. (2007). Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. Stat Med, 26, 1237-54.

Pignon, J. P., Arriagada, R., Ihde, D. C., et al. (1992). A meta-analysis of thoracic radiotherapy for small-cell lung cancer. N Engl J Med, 327: 1618-24.

Potter, F. J. (1993). The effect of weight trimming on nonlinear survey estimates. In Proceedings of the Section on Survey Research Methods of American Statistical Association. Amer. Statist. Assoc., San Francisco, CA.

Randa, M. E., and Straccia, U. (2003), Web metasearch: Rank vs. Score based Rank Aggregation Methods. In *Proceedings of the 2003 ACM Symposium on Applied Computing*, 841-846.

Robins, J. M., Hernan, M. A. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 11, 550-560.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. J. Roy. Statist. Soc. Ser. B, 53, 597-610.

Rosenbaum, P. R. (2002). Observational Studies, 2nd ed. Springer, New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41-55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Amer. Statist., 39, 33-38.

Rubin, D. B. (1973). Matching to remove bias in observational studies. Biometrics, 29, 159-184.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66 688-701.

Rubin, D. B. (1978) Bayesian inference for casual effect: the role of randomization. The Annals of Statistics, 6(1), 34-58.

Rubin, D. B. (1984), Bayesianly justiable and relevant frequency calculations for the applied statistician. *Ann. Statist*, 12, 1151-1172.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. Health Services & Outcomes Research Methodology, 2, 169-188.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores, relating theory to practice. Biometrics, 52, 249-264.

Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. J. Amer. Statist. Assoc., 95, 573-585.

Rubinstein, R. Y., and Kroese, D. P. (2004), *The Cross-Entropy Method. A Unied Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York: Springer.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. J. Amer. Statist. Assoc., 94, 1096-1120.

Sese, J., and Morishita, S. (2001), Rank Aggregation Method for Biological Databases. *Genome Informatics* 12, 506-507.

Shaw, J. A., and Fox, E. A. (1994), Combination of Multiple Searches. In *Proceedings of the 2nd Text Retrieval Conference*, 243-252.

Signorovitch, J. E., Wu, E. Q., Yu, A. P., et al. (2010). Comparative Effectiveness Without Head-to-Head Trials. Pharmacoeconomics, 28 (10): 935-945.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D′Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 203-209.

Singh, K. (1981). On the asymptotic accuracy of Efrons bootstrap. The Annals of Statistics, 9, 1187-1195.

Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. Sociological Methodology, 27, 325-353.

Snedecor, G. W. and Cochran, W. G. (1980). Statistical Methods, 7th ed. Iowa State Univ. Press, Ames, IA.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Science, 25(1), 1-21.

Stuart, E. A. and Green, K. M. (2008). Using full matching to estimate causal effects in non-experimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. Developmental Psychology, 44, 395-406.

Sutton, A., Ades, A. E., Cooper, N., et al. (2008). Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics, 26: 753-67.

Tanner, M. A. and Wong, W. H. (1987), The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* 82(398), 528-540.

Tanner, M. A. and Wong, W. H. (1987), The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association 82(398), 528-540.

True, L., Coleman, I., Hawley, S., Huang, A., Gifford, D., Coleman, R., Beer, T., Gelman, E., Datta, M., Mostaghel, E., Knudsen, B., Lange, P., Vessella, R., Lin, D., Hood, L., and Nelson, P. (2006). A molecular correlate to the gleason grad-

ing system for prostate adenocarcinoma. *Proceedings of the National Academy of Sciences of the USA*, 103, 10991-10996.

Turner, R. M., Omar, R. Z., Yang, M., et al. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med, 19: 3417-32.

Van Erp M., and Schomaker, L. (2000), Variants of the Borda Count Method for Combining Ranked Classifier Hypotheses. In *Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition*, 443-452.

Vogt, C., and Cottrel, G. W. (1999), Fusion via a Linear Combination of Scores. *Information Retrieval* 3, 151-173.