Mahalanobis Matching and Equal Percent Bias Reduction

A thesis presented by:

Seth Viren Neel

Advisor: Natesh S. Pillai
sethneel@college.harvard.edu
(401) 632-7495

to
The Mathematics Department

in partial fulfillment of the requirements
for the degree with honors
of Bachelor of Arts

Harvard College
March 2015

MAHALANOBIS MATCHING AND EQUAL PERCENT BIAS REDUCTION

SETH NEEL

Abstract. Section 1 comprises a literature review, which introduces the reader to Mahalanobis Matching and the Rubin causal framework, summarizes key papers on affinely invariant matching methods, and introduces results which will be used later in the thesis. Section 2.1 computes a new approximation for the expected bias after exact matching with one binary covariate. Section 3.1 reviews work on conditionally affinely invariant matching methods from [RT92a]. In Section 3.2, under the assumptions for conditionally affinely invariant matching, we write the bias reduction in a novel form, and give an interpretation for how quickly the matching approaches equal percent bias reduction (EPBR). In Section 4 we look at the probability that additional special covariates do not affect the matching. We compute an asymptotic lower bound for this probability, in the case when these additional covariates are bounded. Section 5 contains simulation results illustrating the approximation in Section 2, the asymptotic bound in Section 4, and the performance of Mahalanobis matching on ellipsoidal and discrete covariates.

Date: March 22, 2015.
Key words and phrases. Mahalanobis distance, matching methods, propensity scores.

Contents

1. Background 3
1.1. Estimating Treatment Effect in Observational Studies 3
1.2. Propensity Score Matching 6
1.3. Mahalanobis Matching and Ellipsoidal Distributions 9
1.4. Equal Percent Bias Reduction 12
1.5. Results 15
2. Baby Steps with Binaries 16
3. Mahalanobis Matching with Proportional Ellipsoidal Distributions and Additional Covariates 19
3.1. Review 19
3.2. Convergence to EPBR 22
4. Principality in Mahalanobis Matching 24
4.1. A Lower Bound for p_I 25
5. Simulations 28
5.1. Mahalanobis Matching with Ellipsoidal Covariates 29
5.2. Mahalanobis Matching with Ellipsoidal and Discrete Covariates 30
5.3. Matching on Discrete Covariates 32
5.4. Convergence to Principality: Discrete Case 33
5.5. Matching on a Single Binary Covariate 34
6. Further Questions 35
7. Appendix 36
7.1. Integrals from Section 2.1 36
7.2. R Code 36
Acknowledgments 37
References 37

1. Background

1.1. **Estimating Treatment Effect in Observational Studies.** In practice, due to political, temporal, or moral reasons, many experimental studies are not randomized. When the treatment assignment mechanism is not random, naturally covariates cannot be assumed to have the same distribution in the treatment and control groups. Because the treatment effect typically depends on the observed covariates, naively taking the difference in sample means between treatment and control outcomes will produce biased estimates. The development of methods to adjust for said bias and to accurately estimate causal effects is paramount to the application of statistics in fields like medicine and social science, and is a major theme in econometrics [Zha04]. In studies with a randomized treatment assignment mechanism both observed and latent covariates are equally distributed in treatment and control groups; thus taking the difference in average treatment effect provides an unbiased estimate of treatment effect. By "treatment effect" we mean $r_1 - r_0$ in the Rubin causal framework, where we assume that every unit has a single set of potential treatment and untreated outcomes (r_0, r_1). This is called the *stable unit-treatment value assumption* [Rub80b]; in particular it dictates that there is no interaction in the treatment effects between units. While this is sometimes not the case, as in systems where units are competing, it can often be addressed in observational studies through design [Stu10]. Because in any experiment a unit is either treated or untreated, (r_0, r_1) is never fully observed, and hence estimating the treatment effect $r_1 - r_0$ is actually a missing data problem [Rub76a]. The technique of matching attempts to solve this missing data problem by estimating unobserved outcomes using the outcomes for units that are as similar as possible. If we let Z

denote the indicator of treatment, then in this setting:

$$\mathbb{E}(r_1^{obs}|z=1) - \mathbb{E}(r_0^{obs}|z=0) \neq \mathbb{E}(r_1) - \mathbb{E}(r_0)$$

The righthand quantity is what we typically want to estimate, the average treatment effect on the treated (ATT), and is the quantity of concern in this thesis. If we assume that $(r_0, r_1) \perp\!\!\!\perp Z|X$, then, since in randomized studies $Z \perp\!\!\!\perp X$, we have:

$$\mathbb{E}(r_1|z=1) - \mathbb{E}(r_0|z=0) = \mathbb{E}_x(\mathbb{E}(r_1|z=1,x)|z=1) - \mathbb{E}_x(\mathbb{E}(r_0|z=0,x)|z=0) =$$

$$\mathbb{E}_x\mathbb{E}(r_1|x,z=1) - \mathbb{E}_x\mathbb{E}(r_0|x,z=0) = \mathbb{E}_x\mathbb{E}(r_1|x) - \mathbb{E}_x\mathbb{E}(r_0|x) = \mathbb{E}(r_1) - \mathbb{E}(r_0),$$

and thus the ATT is an unbiased estimator for $\mathbb{E}(r_1) - \mathbb{E}(r_0)$ in randomized studies, assuming $(r_0, r_1) \perp\!\!\!\perp Z|X$. The last assumption, along with the assumption that for each value of the covariates there is a chance the unit will receive treatment, is called *ignorability*, and it is discussed in the context of propensity scores in Section 1.2. Matching methods subsample or match to improve covariate balance in the treatment and control groups and mimic the effect of randomization, and consequently reduce bias in the estimation of treatment effect.

Matching methods have been applied since the 1940's, but began receiving a rigorous treatment only in the early 1970's [Stu10]. It is important to note that matching methods do not take into account outcome variables, and thus they necessarily preclude selection bias on the part of the researcher trying to obtain certain results. Modern matching methods include exact matching on the covariates, coarsened exact matching, matching on estimated propensity scores via logistic regression, nearest-neighbor Mahalanobis metric matching, and Mahalanobis matching within propensity score calipers [Stu10]. Clearly, finding an exact covariate match for a given treated unit is preferable if possible, but this is often not the case. Exact

matching suffers from the aptly termed "curse of dimensionality," where as the number of covariates grows the matching must match on all variables simultaneously, and thus performs increasingly poorly. A large step forward came with the introduction of the propensity score in [RR83], which addresses this problem by producing a scalar summary of each covariate that, upon matching, gives an unbiased estimate of treatment effect assuming ignorability. We highlight the propensity score in Section 1.2.

Both propensity score and Mahalanobis matching share the property of being *affinely invariant*, which means that the matching is preserved under any full rank transformation of the covariates in the treatment and control groups. This is a virtue; we want our matching to be robust whether temperature is reported in Fahrenheit or Celsius, height in centimeters or inches, etc. Pivotal work on affinely invariant matching methods is given in [RT92a] and [RS06]. In [RT92a], affinely invariant matching methods for proportionally ellipsoidal distributions are shown to be equal percent bias reducing (EPBR), and formulas for the variance and second moments of a linear combination of the covariates after the matching are given. Multivariate normal and multivariate t-distributions are examples of ellipsoidally symmetric distributions that are often encountered in hypothesis testing. In [RS06], most of the results in [RT92a] are extended to discriminant mix of proportionally ellipsoidally symmetric distributions (DMPES), a generalization of ellipsoidal distributions. Mahalanobis distance and ellipsoidal distributions are introduced in Section 1.3, and equal percent bias reduction is introduced in Section 1.4. The review of material in [RT92a] most relevant to the thesis is postponed until Section 3.1, where it complements new work in Section 3.2. Section 1.5 summarizes the main contributions of this thesis.

1.2. **Propensity Score Matching.** We follow along with [RR83] in introducing the propensity score, and explaining its pivotal properties in analysis of observational studies. We then summarize work in [RT92b] on linear propensity score matching with normal covariates, and discuss an extension of the technique presented in [RT00]. Remember that in this thesis we are not primarily concerned with propensity scores. Nevertheless we feel that it is too important a subject in matching methods to not give it at least a cursory treatment.

Definition 1. *The propensity score is the probability a given unit received the treatment assignment:* $e(x) = \mathbb{P}(z = 1|x)$.

Note that in non-randomized studies we do not know the treatment assignment mechanism, and hence in practice $e(x)$ must be estimated, but we disregard this concern for now. We show that if we assume strongly ignorable treatment assignment, at a given level of the propensity score taking the difference in treatment and control outcomes is an unbiased estimate of the treatment effect. This property is in a sense universal to the propensity score, as it is the *coarsest* score for which it holds. Setting up this statement formally requires us to define a related concept, that of the balancing score $b(x)$.

Definition 2. *A balancing score is a function of* x *such that* $x \perp\!\!\!\perp z|b(x)$.

This says that, conditional on a balancing score $b(x)$, the distribution of x in the treatment and control distributions is the same. Note that trivially any one-to-one function of x is a balancing score since from it we can recover the exact value of x. The propensity score $e(x)$ is a balancing score, and in particular Rubin and Rosenbaum show that for any balancing score $b(x)$ there exists a function h such that $h(b(x)) = e(x)$; hence $e(x)$ is the coarsest balancing score. We follow Section 2 in [RR83] and prove:

- $e(x)$ is the coarsest balancing score
- Conditioning on a level of the balancing score, the average difference between treatment and control means is an unbiased estimator for the treatment effect.

Theorem 1.1. [RR83] *Let $b(x)$ be a function of x. Then $b(x)$ is a balancing score if and only if there exists f such that $f(b(x)) = e(x)$.*

Proof. Suppose that $b(x)$ is a balancing score, but no such function f exists. Then necessarily there are m, n such that $b(m) = b(n)$, but $e(m) \neq e(n)$, which implies that $\mathbb{P}(z = 1|m) \neq \mathbb{P}(z = 1|n)$. It follows that $b(x)$ is not a balancing score because, conditional on the value of $b(m) = b(n)$, Z is not independent of the vectors m, n, which is a contradiction. Conversely, suppose that there exists f such that $f(b(x)) = e(x)$. We want to show that $\mathbb{P}(z = 1|b(x)) = P(z = 1|x) = e(x)$. But $\mathbb{P}(z = 1|b(x)) = \mathbb{E}(Z|b(x))$ and $e(x) = \mathbb{E}(Z|x)$. Then, by the law of iterated expectation, $\mathbb{P}(z = 1|b(x)) = \mathbb{E}(Z|b(x)) = \mathbb{E}(\mathbb{E}(Z|x)|b(x)) = \mathbb{E}(e(x)|b(x)) = e(x)$, where the last equality follows since $e(x)$ is a function of $b(x)$. Hence any finer score than $e(x)$, and in particular $e(x)$ itself, is a balancing score. □

This theorem is significant because it tells us that, if we match on $e(x)$, we can include other functions of x into the score and still maintain the property of being a balancing score [RR83]. We now show that if the treatment effect is strongly ignorable, given x, then it is strongly ignorable given any balancing score.

Theorem 1.2. [RR83] *Suppose that $(r_1, r_0) \perp\!\!\!\perp z|x$, and $\forall x,\ 0 < \mathbb{P}(z = 1|x) < 1$. Then $(r_1, r_0) \perp\!\!\!\perp z|b(x)$, and $\forall x,\ 0 < \mathbb{P}(z = 1|b(x)) < 1$.*

Proof. We again follow [RR83]. We want to show that, given $\mathbb{P}(z = 1|(r_1, r_0), x) = p(z = 1|x)$, that $\mathbb{P}(z = 1|(r_1, r_0), b(x)) = \mathbb{P}(z = 1|b(x))$. Note that $e(x) = f(b(x)) \implies$

$\mathbb{P}(z = 1|b(x)) = e(x)$. Hence we want to show that $\mathbb{P}(z = 1|(r_1, r_0), b(x)) = e(x)$. Again, by the law of iterated expectation, $\mathbb{P}(z = 1|(r_1, r_0), b(x)) = \mathbb{E}(Z|r_1, r_0, b(x)) = \mathbb{E}_x(\mathbb{E}(Z|r_1, r_0, x)|\{r_1, r_0, b(x)\})$, which by ignorability is equal to $\mathbb{E}_x(\mathbb{E}(Z|x)|\{r_1, r_0, b(x)\}) = \mathbb{E}(e(x)|(r_0, r_1), b(x)) = e(x)$, as desired, since knowing $b(x)$ means we know $e(x)$ by Theorem 1.1. □

The theorem tells us if the treatment effect is strongly ignorable given x then matching on balancing score gives us an unbiased estimate of treatment effect since $(r_0, r_1)|b(x) \perp\!\!\!\perp Z \implies \mathbb{E}(r_1|b(x), z = 1) - \mathbb{E}(r_0|b(x), z) = \mathbb{E}(r_1|b(x)) - \mathbb{E}(r_0|b(x))$, as desired. This in turn allows us to use the law of total probability to compute the ATT:

$$\mathbb{E}(r_1 - r_0) = \mathbb{E}_{b(x)}(\mathbb{E}(r_1 - r_0|b(x))) = \mathbb{E}_{b(x)}(\mathbb{E}(r_1|z = 1, b(x))) - \mathbb{E}_{b(x)}(\mathbb{E}(r_0|z = 0, b(x))),$$

i.e., if we can match on $b(x)$ and compute the sample difference in matched treatment and control means, we obtain an unbiased estimate of the ATT. We have now established the key property of matching on propensity scores, so we conclude our review with some remarks on matching on propensity scores in practice, and by briefly summarizing work in the area. Recall that since we do not know the treatment assignment mechanism, when dealing with actual data $e(x)$ must be modeled, typically by logistic regression. Then matching is done on these estimated propensity scores $\hat{e}(x)$, or on a transformation $\text{logit}(\hat{e}(x))$, the estimated linear propensity score. As the sample sizes N_t, N_c increase, $\hat{e}(x)$ will become a better estimate of the true propensity score $e(x)$. In [RT92b], Rubin and Thomas compute an analytic approximation for the expected bias reduction when matching on estimated linear propensity scores for normal distributions, and go on in [RT96] to show that these approximations hold well even when the data deviates strongly from normality. In [RT00] Rubin and Thomas compute approximations for the matched distributions

when combining propensity score matching with exact matching on special "prognostic" covariates, and show that this hybrid approach is more effective at reducing bias than matching on prognostic covariates or propensity scores alone.

1.3. **Mahalanobis Matching and Ellipsoidal Distributions.** In 1936 P.C. Mahalanobis, founder of the Indian Statistical Institute, introduced the Mahalanobis distance. Mahalanobis distance is a measure of how far a point x is from a distribution $\mathcal{F}$, and it is a multivariate generalization of how many standard deviations a univariate point is from the mean μ of $\mathcal{F}$. In one dimension this is $\frac{x-\mu}{\sigma}$, and in n dimensions the Mahalanobis distance M is defined as $M^2 = (x-\mu)'\Sigma^{-1}(x-\mu)$ where Σ is the covariance matrix of $\mathcal{F}$. One immediately notices that the quantity M^2 appears in the probability density function of the multivariate normal distribution, and in fact if $X \sim \mathcal{N}(\mu, \Sigma)$, then $f_x = c \cdot e^{-\frac{1}{2}M^2}$. This relationship to the normal is why Mahalanobis distance is ubiquitous in statistics; for example the reader may recognize it from the decision boundary for linear discriminant analysis (LDA) in classification. It also is the basis of the intuition behind defining Mahalanobis distance: the distance treats points equally that lie on the same level set of the ellipse $(x-\mu)'\Sigma_c^{-1}(x-\mu)$, or in the case of an ellipsoidal distribution like the multivariate normal, have the same density. An ellipsoidal distribution with mean μ and covariance matrix Σ has density function (if it exists) equal to $k \cdot g((x-\mu)'\Sigma^{-1}(x-\mu))$ where g is any function mapping $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that defines a valid probability density function. Alternatively an ellipsoidal distribution is one such that there exists a linear transformation of the coordinates such that the transformed distribution is spherically symmetric [Dem69]. Spherically symmetric distributions are invariant under orthogonal transformations of their coordinate systems, for example, the standard multivariate normal distribution. In [RT92a], Rubin and Thomas define

proportionally ellipsoidal distributions, which are simply pairs of ellipsoidal distributions whose covariance matrices are proportional. As we will see in Section 1.4, these proportional ellipsoidal distributions have a special role to play in matching, and so here we prove a canonical form that was utilized in [RT92a], and had been cited in the literature previously.

Theorem 1.3. *[Canonical Form] Let X_t, X_c be proportionally ellipsoidal distributions with means μ_t, μ_c and covariance matrices $\Sigma_t \propto \Sigma_c$. Then there exists an affine transformation A such that, in the transformed distributions AX_t, AX_c, we have:*

- $\mu_t' \propto (1, \dots, 1)$
- $\mu_c' = 0$
- $\Sigma_t' = \sigma^2 I$
- $\Sigma_c'' = I$

Proof. Affine transformations are simply compositions of scalar translations and linear transformations. First subtract μ_c from both distributions. Let U be a matrix such that $U\Sigma_c U' = I$, where we obtain U by taking the Cholesky decomposition of Σ_c^{-1}. Then multiplying by U our new distributions are $U(X_t - \mu_c), U(X_c - \mu_c)$. Finally let T be any orthogonal matrix projecting the vector $U(\mu_t - \mu_c)$ onto the subspace spanned by $(1, \dots, 1)$. Our final transformed distributions are $TU(X_t - \mu_c), TU(X_t - \mu_c)$. Then in these distributions $\mu_c' = 0, \mu_t' = TU(\mu_t - \mu_c) \propto (1, \dots, 1), \Sigma_c' = TU\Sigma_c U'T' = TIT' = TT' = I$, and similarly $\Sigma_t' \propto \Sigma_c' = I$. □

In the matching setting, typically a distance measure between treatment and control units is defined, and then when iterating randomly through the treated units, control units are subsampled with or without replacement to minimize this distance measure from the given treated unit. Along with distances based on propensity

scores, Mahalanobis distances are one of the most common choices of distance measure. Letting x_t, x_c denote treatment and control units respectively, we define the Mahalanobis distance between a treatment and control unit as $\sqrt{(x_t - x_c)' S^{-1} (x_t - x_c)}$, where S is either the control covariance matrix Σ_c, or the pooled covariance matrix between the treatment and control groups. When the goal is on estimating the ATT and we do not subsample the treated units, we take $S = \Sigma_c$, and so we define the Mahalanobis distance with respect to Σ_c for the rest of this thesis [Stu10].

Previous work on Mahalanobis matching has primarily involved computing Monte Carlo values for bias reduction with normal distributions and comparing them to values from other matching methods, or in the context of Mahalanobis matching as an affinely invariant matching method (defined in Section 1.4). In [Rub79] Monte Carlo simulations show that when the response variable Y is a nonlinear function of the covariates (bivariate X), Mahalanobis matching has superior performance to matching on the linear discriminant, when combined with regression adjustment. Regression adjustment for matching just means that instead of estimating the treatment effect by taking the difference in sample means from the matched samples, we also adjust for the remaining bias in the samples by subtracting $\beta \cdot (\bar{X}_{mt} - \bar{X}_{mc})$, where β is the regression coefficient of X in the regression of outcome on covariates. Motivated by the success of Mahalanobis matching in [Rub79], in [Rub80a] Rubin computes Monte Carlo bias reductions for bivariate normal covariates and univariate normal covariates with Mahalanobis matching, and compares the values to the theoretical maximum bias reduction for affinely invariant matching methods derived in [Rub76c]. [GR93] studies Mahalanobis matching as one of several matching methods, and reiterates the finding that Mahalanobis matching performs well relative to discriminant matching and propensity methods when the covariates are

low-dimensional ([Rub79], [Zha04]), but found that the bias reduction is not optimal when the covariates deviate from normality or the number of dimensions grows high.

The past work on Mahalanobis matching has centered on how effective Mahalanobis matching is at reducing bias in various distributional settings relative to other matching methods. Beyond proving that Mahalanobis matching is equal percent bias reducing (EPBR) in specific distributional settings, a study of how this property holds up as covariates deviate from normality and grow in dimension is lacking. In the next section, we introduce the key concepts of equal percent bias reduction and affinely invariant matching methods.

1.4. **Equal Percent Bias Reduction.** Let X_t, X_c denote the treatment and control distributions, and let μ_t, μ_c respectively be their finite means. Now let μ_{mt}, μ_{mc} denote the expected sample means of X_t, X_c in the matched treatment and matched control populations.

Definition 3. *We say that a matching method is equal percent bias reducing (EPBR) if there exists a constant λ such that $\mu_t - \mu_c = \lambda(\mu_{mt} - \mu_{mc})$.*

One might ask: Why do we care if a matching method is EPBR? For example, it seems intuitive that a matching method which reduced bias in each coordinate by exactly 5 percent would always be inferior to a matching method that reduced bias by more than 80 percent in every coordinate, although not by exactly the same percent in every coordinate. Now suppose that we know our outcome variable is a linear function of our covariates, but we are not confident which linear function in particular it is. Then, as is remarked upon (but not proven) in [Rub76b], if a matching method reduces bias in every coordinate but is not EPBR, then there

exists a linear outcome direction $\gamma' X$ in which the matching increases bias by an arbitrarily large percent. We give a short proof below.

Proof. Suppose that $\mu_t - \mu_c \neq \lambda(\mu_{mt} - \mu_{mc})$. Then $\exists \gamma'$ such that $\gamma' \cdot (\mu_t - \mu_c) = 0$, but $\gamma' \cdot (\mu_{mt} - \mu_{mc}) \neq 0$. This is clear since $(\mu_t - mu_c)^\perp \neq (\mu_{mt} - \mu_{mc})^\perp$ and they are the same dimension. But then the matching has created bias when there was initially no bias, which corresponds to increasing bias by an infinite amount. □

Because we are guaranteed to avoid situations like this in matching settings that are EBPR, a substantial number of papers have dealt with methods and distributional assumptions in which the EPBR property can be proven. In [Rub76b] Rubin considers eleven distributional assumptions, and in each case describes corresponding matching methods which are EPBR. In a related paper of the same year, Rubin computes the maximum attainable percent bias reduction given fixed sample sizes and multivariate distributions under EPBR matching methods [Rub76c]. In [Rub76b], Rubin introduces the notion of an *exchangeable* matching method, which is one whose rules are invariant under permutations of the indices of the matching variables. For these matching methods, multiplying the treatment and control data matrices by the same permutation matrix does not change which control indices were selected by the matching. In [RT92a], Rubin and Thomas extend this definition to affinely invariant matching methods. Let $(\mathcal{X}_t, \mathcal{X}_c)$ denote the treatment and control data matrices with each column representing a unit, and (T, C) denote the treatment and control indices selected by the matching method. Note that if we do not subsample the treated units then T is just the set $\{1, \ldots, N_t\}$.

Definition 4. [RT92a] *Let A be any affine transformation. An affinely invariant matching method m is one such that if $(\mathcal{X}_t, \mathcal{X}_c) \overset{m}{\Longrightarrow} (T, C)$, then $(A\mathcal{X}_t, A\mathcal{X}_c) \overset{m}{\Longrightarrow} (T, C)$.*

This is an instructive definition, since as it turns out that many common matching methods are affinely invariant. The most important of these are Mahalanobis matching, discriminant matching, and matching on estimated propensity scores using logistic regression.

Lemma 1.4. *Mahalanobis matching is affinely invariant.*

Proof. Suppose that we are not subsampling on the treated units, and so we use Σ_c^{-1} in our inner product. The same proof goes forward when using the pooled covariance matrix. Then, if we multiply our covariates by the linear transformation A, our new covariance matrix in the control units is $\Sigma_c' = A\Sigma_c A^T$. Let x_t, x_c be treatment and control units respectively. If we multiply our data matrices by A, then x_t, x_c become Ax_t, Ax_c. The Mahalanobis distance between these two units with respect to Σ_c' is $\sqrt{(Ax_t - Ax_c)'(A\Sigma_c A^T)^{-1}(Ax_t - Ax_c)} = \sqrt{(x_t - x_c)'A^T((A^T)^{-1}\Sigma_c^{-1}A^{-1})A(x_t - x_c)} = \sqrt{(x_t - x_c)'\Sigma_c^{-1}(x_t - x_c)}$. Then, since the pairwise distance between units is invariant under any linear transformation A, and it is clearly invariant under any scalar translation since these do not change the covariance matrix, the distance is invariant under any affine transformation. We are matching on exactly this pairwise distance, and so the selected indices will be invariant as well. □

In [Rub76b] the theorem labeled Case 9 states that when a matching method is exchangeable, if X_t, X_c are exchangeably distributed, then the matching method is EPBR. This is obvious since symmetry of the distribution and matching method force $\mu_t, \mu_c, \mu_{mt}, \mu_{mc} \propto (1, \dots, 1)$, and hence $\mu_t - \mu_c \propto (\mu_{mt} - \mu_{mc})$. Then it is clear that, given an affinely invariant matching method, if X_t, X_c are proportionally ellipsoidal we can assume the exchangeable canonical form defined in Section 1.3 without changing the matching. Hence by this earlier work we know that affinely invariant matching methods on proportional ellipsoidal distributions must be EPBR.

In [RT92a], Rubin and Thomas derive a sharpened version of this result, giving an expression for the bias reduction. Their technique also enables them to obtain results decomposing the variance of matched linear outcome variables. In the second part of their paper they generalize their results to the case when X_t, X_c are composed of s special covariates and r covariates whose conditional distribution on the special covariates is ellipsoidal. Discussion of their results in the general case is postponed until Section 3.1 of the thesis. In [RS06] the theorems in [RT92a] are extended to DMPES distributions. DMPES distributions are mixtures of ellipsoidally symmetric distributions, where each component covariance matrix is proportional, and the best linear discriminant between any two components is proportional. The results in [RS06]are in the vein of [RT92a], in that a canonical exchangeable form of DMPES is assumed, and linear outcome variables are decomposed along and orthogonal to the best linear discriminant. The idea for this thesis was a suggestion from Donald Rubin to Natesh Pillai, wondering if the result on equal percent bias reduction obtained for the distributional settings in [RT92a] and [RS06] could be extended to the case when some of the covariates are discrete.

1.5. **Results.** Simulation results in Section 5 indicate that Mahalanobis matching on discrete covariates, and on combinations of ellipsoidal and discrete covariates does not have the EPBR property. However, simulated and theoretical results also show that there are situations in which Mahalanobis matching is very close to equal percent bias reducing. In light of these findings, the original contributions of this thesis branch out along three different paths:

- A precise asymptotic calculation for the expected matched bias under exact matching on one binary covariate (Section 2, Section 5.5)

- Characterizations of how quickly matching with ellipsoidal and discrete covariates approaches EPBR (Section 3 Corollaries 3.4, 3.5, Lemma 3.6, Sections 5.1, 5.2, 5.3)
- Introduction of a new measure which we call *principality*, which measures the probability that in a combination of discrete and ellipsoidal covariates the discrete covariates affect which control units are subsampled. An asymptotic lower bound for this probability is obtained, in the case when additional covariates are bounded. This result has applications to the EBPR context, and also to the situation in which latent covariates are present, but which we do not consider here (Section 4, Section 5.4).

Although perhaps the above results would not typically be organized into the same paper (particularly Section 2), in this thesis it is the author's intention to present all of the original work completed during the $2014-2015$ academic year, all motivated by the same initial conjecture concerning Mahalanobis matching. We also note that the *Matching* package in R can handle Mahalanobis and propensity score matching, but given the need to experiment with many different matching settings the author produced original code for all of the simulations. This code is attached after the References section.

2. Baby Steps with Binaries

We derive an approximation for the matched bias after exact matching without replacement on one binary covariate, which holds for large samples. In particular, we compute $\mathbb{E}(\bar{B}_{mt} - \bar{B}_{mc})$, where B_c, B_t are Bernoulli random variables, and m denotes exact matching. This amounts to computing $\mathbb{E}(\bar{B}_{mc})$ since we do not subsample the treated units. Let $B_t \sim Bern(p_1), B_c \sim Bern(p_2)$. The Mahalanobis distance between two Bernoulli random variables is $\frac{1}{p_2(1-p_2)}(b_t - b_c)^2 = \frac{1}{p_2(1-p_2)} I_{b_t = b_c}$.

Hence in the matching either the treated unit will find an exact match, or it will select a control unit at random as a match. So with one binary covariate, Mahalanobis matching is just exact matching. This is now a combinatorics problem, and we calculate $\mathbb{E}(\bar{X}_{mc}) = \frac{1}{N_t}\mathbb{E}(\sum x_{mc})$, or at least a very close approximation. Let X_{mc} denote $\sum x_{mc}$, X_t denote $\sum x_t$, and X_c be defined similarly. Then $\mathbb{E}(X_{mc}) = \mathbb{E}(\mathbb{E}(X_{mc}|X_t))$. First we calculate $\mathbb{E}(X_{mc}|X_t)$. If $X_t \leq X_c \leq N_c - N_t + X_t$ then $X_{mc} = X_t$, if $X_t > X_c$ then $X_{mc} = X_c$, and if $X_c > N_c - N_t + X_t$ then $X_{mc} = N_t - N_c + X_c$. So, by the law of conditional expectation,

$$\begin{aligned}\mathbb{E}(X_{mc}|X_t) = \mathbb{P}(X_t \leq X_c \leq N_c - N_t + X_t)X_t + \mathbb{P}(X_t > X_c)\mathbb{E}(X_c|X_c < X_t) \\ + \mathbb{P}(X_c > N_c - N_t + X_t)\mathbb{E}(N_c - N_t + X_c|X_c > N_c - N_t + X_t).\end{aligned}$$

Since N_c and N_t are large, and the treatment and control units are i.i.d, we can use the Central Limit Theorem to approximate X_t, X_c; let $X_t \sim N(\mu_t, \sigma_t^2), X_c \sim N(\mu_c, \sigma_c^2)$, where $\mu_t = p_t \cdot N_t, \mu_c = p_c \cdot N_c, \sigma_t^2 = p_t(1-p_t) \cdot N_t, \sigma_c^2 = p_c(1-p_c) \cdot N_c$. Let ϕ be the probability density function of the standard normal, and Φ be the cumulative distribution function of the standard normal. Then we can approximate this conditional expectation as

$$\begin{aligned}\mathbb{E}(X_{mc}|X_t) = \Phi(\frac{X_t - \mu_c}{\sigma_c})\mathbb{E}(X_c|X_t > X_c) + (\Phi(\frac{X_t + N_c - N_t - \mu_c}{\sigma_c}) - \Phi(\frac{X_t - \mu_c}{\sigma_c}))X_t \\ + (1 - \Phi(\frac{X_t + N_c - N_t - \mu_c}{\sigma_c}))(N_c - N_t + \mathbb{E}(X_c|X_c > N_c - N_t + X_t)).\end{aligned}$$

The expectation of a normal distribution $X \sim N(\mu, \sigma)$ in the upper tail is

$$\mathbb{E}(X|X > a) = \mu + \sigma \cdot \frac{\phi((a-\mu)/\sigma)}{1 - \Phi((a-\mu)/\sigma)},$$

and the expectation of a normal distribution truncated in the lower tail is

$$\mathbb{E}(X|X < a) = \mu - \sigma \cdot \frac{\phi((a-\mu)/\sigma)}{\Phi((a-\mu)/\sigma)}.$$

Then our conditional expectation becomes

$$\begin{aligned}(1)\qquad \mathbb{E}(X_{mc}|X_t) &= \mu_c\Phi(\frac{X_t-\mu_c}{\sigma_c}) + (\Phi(\frac{X_t+N_c-N_t-\mu_c}{\sigma_c}) - \Phi(\frac{X_t-\mu_c}{\sigma_c}))X_t \\ &+ \mu_c(1-\Phi(\frac{X_t+N_c-N_t-\mu_c}{\sigma_c})).\end{aligned}$$

Let $a = \frac{\sigma_t}{\sigma_c}, b = \frac{\mu_c-\mu_t}{\sigma_c}, b' = \frac{\mu_c-\mu_t+N_t-N_c}{\sigma_c}$. Rewriting (1) we obtain:

$$\begin{aligned}(2)\qquad \mathbb{E}(X_{mc}) &= \mu_c\mathbb{E}(\Phi(aZ-b)) + \sigma_t\mathbb{E}(\Phi(aZ-b')Z) + \mu_t\mathbb{E}(\Phi(aZ-b')) \\ &- \sigma_t\mathbb{E}(\Phi(aZ-b)Z) - \mu_t\mathbb{E}(\Phi(aZ-b)) - \mu_c\cdot(\mathbb{E}(\Phi(aZ-b'))) + \mu_c.\end{aligned}$$

To take the expectation with respect to X_t of the right-hand side in (2), we need to compute integrals of the form

$$\int_{-\infty}^{\infty} z\Phi(az-b)\phi(z)dz, \text{ and } \int_{-\infty}^{\infty}\Phi(az-b)\phi(z)dz.$$

Both integrals have known closed forms which enable us to compute the exact bias after matching. We leave these manipulations to the first section of the Appendix. We substitute in the expressions:

$$\int z\Phi(az-b)\phi(z)dz = \phi(\frac{-b}{\sqrt{1+a^2}})\frac{a}{\sqrt{a^2+1}},$$

and

$$\int_{-\infty}^{\infty}\Phi(az-b)\phi(z)dz = \Phi(\frac{-b}{\sqrt{1+a^2}}),$$

obtaining Theorem 2.1.

Theorem 2.1. *Suppose that $X_t \sim$ Bern$(p_t), X_c \sim$ Bern(p_c), with N_t, N_c treatment and control units respectively. Let $\mu_t = N_t\cdot p_t, \mu_c = N_c\cdot p_c, \sigma_t = N_t\cdot p_t(1-p_t), \sigma_c =$*

$N_c \cdot p_c(1-p_c), a = \frac{\sigma_t}{\sigma_c}, b = \frac{\mu_c - \mu_t}{\sigma_c}, b' = \frac{\mu_c - \mu_t + N_t - N_c}{\sigma_c}$. *Then by the above calculations,*

$$\mathbb{E}(\bar{X}_{mt} - \bar{X}_{mc}) \approx p_t - \frac{1}{N_t}((\mu_c - \mu_t) \cdot [\Phi(\frac{-b}{\sqrt{1+a^2}}) - \Phi(\frac{-b'}{\sqrt{1+a^2}})] + \frac{a\sigma_t}{\sqrt{1+a^2}} \cdot [\phi(\frac{-b'}{\sqrt{1+a^2}}) - \phi(\frac{-b}{\sqrt{1+a^2}})] + \mu_c).$$

Simulations illustrating this approximation appear in Section 5.5. Note that when $N_t = N_c$ the matched control units are just randomly sampled control units, and hence the expected matched bias should be $p_t - p_c$; indeed letting $N_t = N_c$ in our approximation returns $p_t - p_c$. To get from the expected matched bias to the expected bias reduction simply divide by $p_t - p_c$. Finally, note that our expression for matched bias is a function of only $\mu_t - \mu_c, N_t, \sigma_t, \sigma_c$ and N_c, as opposed to p_t, p_c (this is clear since taking the μ_c out of the parentheses we get $p_t - \frac{\mu_c}{N_t} = \frac{\mu_t - \mu_c}{N_t}$).

3. Mahalanobis Matching with Proportional Ellipsoidal Distributions and Additional Covariates

3.1. **Review.** We summarize the work in [RT92a] on conditionally affinely invariant matching methods, and then describe how it applies to Mahalanobis matching. We follow [RT92a] in our exposition. Suppose our covariates are partitioned into $(X^{(s)}, X^{(r)})$. A *conditionally affinely invariant matching method* with respect to $(X^{(s)}, X^{(r)})$ is defined by the property that if $((\mathcal{X}_t^{(s)}, \mathcal{X}_t^{(r)}), (\mathcal{X}_c^{(s)}, \mathcal{X}_c^{(r)})) \stackrel{m}{\Longrightarrow} (T, C)$, then for any affine transformation A, $((\mathcal{X}_t^{(s)}, \mathcal{A}\mathcal{X}_t^{(r)}), (\mathcal{X}_c^{(s)}, \mathcal{A}\mathcal{X}_c^{(r)})) \stackrel{m}{\Longrightarrow} (T, C)$, where (T, C) are the treatment and control indices selected by the matching applied to the data matrices $\mathcal{X}$. Correspondingly, we define proportionally conditionally ellipsoidal distributions $((\mathcal{X}_t^{(s)}, \mathcal{X}_t^{(r)}), (\mathcal{X}_c^{(s)}, \mathcal{X}_c^{(r)}))$ such that:

- $X^{(r)}|X^{(s)}$ is ellipsoidally distributed with conditional mean a linear function of $X^{(s)}$ and constant conditional covariance matrix

- $\Sigma_t^{(r|s)} \propto \Sigma_c^{(r|s)}$
- For each i, the linear regression of $X_i^{(r)}$ on $X^{(s)}$ is the same in the treatment and control groups.

Given these assumptions, in conjunction with a conditionally affinely invariant matching method, we can assume a canonical form for the distributions where

- $\Sigma_t = \begin{bmatrix} \Sigma_t^{(s)} & 0 \\ 0 & \sigma^2 I \end{bmatrix}$
- $\Sigma_c = \begin{bmatrix} \Sigma_c^{(s)} & 0 \\ 0 & I \end{bmatrix}$
- $\mu_c^{(r)} \propto 1, \mu_t^{(r)} = 0.$

This canonical form follows from the one for proportionally ellipsoidal distributions with a small trick. First apply the affine transformation $X^{(r)} = X^{(r)} - BX^{(s)}$, where B is the regression of each component of $X^{(r)}$ on $X^{(s)}$, so that now $X^{(r)}$ is uncorrelated with $X^{(s)}$ in both the treatment and control populations. Then transform the uncorrelated components. From this point forward, assume this canonical form. Represent an arbitrary linear combination Y as $Y = \rho\mathcal{Z} + (\sqrt{1-\rho^2})\mathcal{W}$ where $\mathcal{Z}$ is the standardized projection of Y onto the subspace $\{X^{(s)}, Z\}$, and Z is the standardized discriminant uncorrelated with $X^{(s)}$; in our canonical form it is simply $1'X^{(r)}/\sqrt{p-s}$.

Theorem 3.1. [RT92a]

$$\frac{\mathbb{E}(\bar{Y}_{mt} - \bar{Y}_{mc})}{\mathbb{E}(\bar{Y}_{rt} - \bar{Y}_{rc})} = \frac{\mathbb{E}(\bar{\mathcal{Z}}_{mt} - \bar{\mathcal{Z}}_{mc})}{\mathbb{E}(\bar{\mathcal{Z}}_{rt} - \bar{\mathcal{Z}}_{rc})}.$$

Proof. Using the representation $Y = \rho\mathcal{Z}+(\sqrt{1-\rho^2})\mathcal{W}$, we have that $\mathbb{E}(\bar{Y}_{mt}-\bar{Y}_{mc}) = \rho\mathbb{E}(\bar{\mathcal{Z}}_{mt} - \bar{\mathcal{Z}}_{mc}) + (\sqrt{1-\rho^2})\mathbb{E}(\bar{\mathcal{W}}_{mt} - \bar{\mathcal{W}}_{mc})$. We claim that $\mathbb{E}(\bar{\mathcal{W}}_{mt} - \bar{\mathcal{W}}_{mc}) = 0$. We are not subsampling the treated units, hence $\mathbb{E}(\bar{\mathcal{W}}_{mt}) = \mathbb{E}(\bar{\mathcal{W}}_{rt}) = 0$. Now let $W_c = \gamma' X_c = (\gamma^{(s)}, \gamma^{(r)})(X_c^{(s)}, X_c^{(r)})^T$, then by construction $\mathcal{W}$ is orthogonal to $\{Z, X_c^{(s)}\}$

and in particular $X_c^{(s)}$, and hence $\gamma^{(s)} = 0$. Then $\mathbb{E}(\bar{\mathcal{W}}_{mt} - \bar{\mathcal{W}}_{mc}) = \gamma^{(r)}(\bar{X}_{mt}^{(r)} - \bar{X}_{mc}^{(r)}) \propto \gamma^{(r)}1'$, since the matching on $X^{(r)}$ is exchangeable, and the joint distribution of $(X^{(s)}, X^{(r)})$ is exchangeable in $X^{(r)}$ in the treatment and control distributions in the canonical form. But $W \perp\!\!\!\perp Z \implies \gamma^{(r)}1' = 0 \implies \mathbb{E}(\bar{\mathcal{W}}_{mt} - \bar{\mathcal{W}}_{mc}) = 0$ as desired. The result follows from the fact that $\mathbb{E}(\bar{\mathcal{W}}_{rt} - \bar{\mathcal{W}}_{rc}) = 0$ by construction. □

Corollary 3.2. [RT92a] *Specialize to the case when there are no special covariates. Then if* $Z = (1, \ldots, 1)'X'$,

$$\frac{\mathbb{E}(\bar{Y}_{mt} - \bar{Y}_{mc})}{\mathbb{E}(\bar{Y}_{rt} - \bar{Y}_{rc})} = \frac{\mathbb{E}(\bar{Z}_{mt} - \bar{Z}_{mc})}{\mathbb{E}(\bar{Z}_{rt} - \bar{Z}_{rc})}.$$

In particular, Z is the same for each outcome direction Y, and so this shows that the matching is EPBR, with bias given by the reduction in bias along the best linear discriminant Z. Starting from the decomposition $Y = \rho\mathcal{Z} + (\sqrt{1-\rho^2})\mathcal{W}$, taking the variance of both sides, and using the same exchangeability tricks used to prove Theorem 3.1, allow [RT92a] to obtain:

Corollary 3.3. [RT92a]

$$\frac{var(\bar{Y}_{mt} - \bar{Y}_{mc})}{var(\bar{Y}_{rt} - \bar{Y}_{rc})} = \rho^2 \frac{var(\bar{\mathcal{Z}}_{mt} - \bar{\mathcal{Z}}_{mc})}{var(\bar{\mathcal{Z}}_{rt} - \bar{\mathcal{Z}}_{rc})} + (1 - \rho^2) \frac{var(\bar{\mathcal{W}}_{mt} - \bar{\mathcal{W}}_{mc})}{var(\bar{\mathcal{W}}_{rt} - \bar{\mathcal{W}}_{rc})}.$$

The variable W is the component of Y uncorrelated with Z, and has the same distribution for all Y; hence the above corollary shows that the reduction in variance after matching varies only with the correlation ρ between Y and Z. Based on the work in [RT92a], Corollary 3.4 naturally follows:

Corollary 3.4. *Let* $\mathcal{Y} = \{\gamma' X |\ \rho_{YX^{(s)}} = 0_s\}$, *then matching restricted to* $Y \in \mathcal{Y}$ *is EPBR.*

Proof. Recall that $\mathcal{Z}$ is the projection of Y onto $\{Z, X^{(s)}\}$, hence if $Y \in X^{(s)\perp}$, then $\text{Proj}_{(X^{(s)},Z)}Y = Proj_Z(Y)$, which in turn gives:

$$\frac{\mathbb{E}(\bar{Y}_{mt} - \bar{Y}_{mc})}{\mathbb{E}(\bar{Y}_{rt} - \bar{Y}_{rc})} = \frac{\mathbb{E}(\bar{Z}_{mt} - \bar{Z}_{mc})}{\mathbb{E}(\bar{Z}_{rt} - \bar{Z}_{rc})}, \text{ which is constant for all } Y \in \mathcal{Y}.$$

□

Corollary 3.4 also gives insight into the performance of the matching method when $X^{(s)}$ is weakly correlated to Y; as $\rho_{X^{(s)}Y} \to 0_s$, the matching restricted over these outcome directions will reduce bias approximately equally in each direction.

3.2. **Convergence to EPBR.** Let (X_t, X_c) be proportionally conditionally ellipsoidal, as in Section 3.1. Then we assume the canonical form, that is, $\Sigma_t^{(r)} \propto I, \Sigma_c^{(r)} = I, \mu_t^{(r)} \propto 1, \mu_c^{(r)} = 0, X^{(r)} \perp\!\!\!\perp X^{(s)}$. By symmetry it stands that, since the first p coordinates are jointly exchangeably distributed in the treatment and control distributions, we have that $\bar{X}_{mt}^{(p)}, \bar{X}_{mc}^{(p)} \propto 1$. Then let $Y = \gamma' X$ be a random linear outcome variable, $||\gamma||= 1$, and write

$$\frac{Y^{(p)}}{||\gamma^{(p)}||} = Z^p \cdot \rho_Y^p + \sqrt{1 - \rho_Y^p} \cdot W_p$$

where Z^p is the standardized best linear discriminant, W is in the orthogonal complement of Z^p, and ρ_Y^p is the correlation of $Y^{(p)}$ with Z^p. But then by the same symmetry arguments in [RT92a] we obtain

$$\mathbb{E}(\bar{Y}_{mt}^{(p)} - \bar{Y}_{mc}^{(p)}) = \rho_Y^p \cdot \mathbb{E}(\bar{Z}_{mt}^p - \bar{Z}_{mc}^p) \cdot ||\gamma^{(p)}||, \tag{3}$$

which implies by the linearity of expectation:

Corollary 3.5.

$$\mathbb{E}(\bar{Y}_{mt} - \bar{Y}_{mc}) = \rho_Y^p \cdot \mathbb{E}(Z_{mt}^p - Z_{mc}^p) \cdot ||\gamma^{(p)}|| + \gamma^{(s)} \cdot \mathbb{E}(\bar{X}_{mt}^{(s)} - \bar{X}_{mc}^{(s)}).$$

Although Corollary 3.5 doesn't represent a large conceptual leap from Theorem 3.1, this new form is instructive. For a matching to be EPBR means that, for any linear outcome variable $Y = \gamma' X$, the ratio of $\mathbb{E}(\bar{Y}_{mt} - \bar{Y}_{mc})$ to $\mathbb{E}(\bar{Y}_{rt} - \bar{Y}_{rc}) = \gamma' \cdot (\mu_t - \mu_c)$ is constant for all Y. Then as (3) approaches $\rho_Y \cdot \mathbb{E}(\bar{Z}_{mt} - \bar{Z}_{mc})$ and as $\mathbb{E}(\bar{Y}_{rt} - \bar{Y}_{rc})$ approaches $\rho_Y \cdot \mathbb{E}(Z_{rt} - Z_{rc})$ the matching becomes equal percent bias reducing. Specifically, $\mathbb{E}(\gamma^{(s)}) = 0_s$ and $\text{var}(\gamma_i^{(s)}) \to 0$ at the rate $O(\frac{1}{p})$ as $p \to \infty$. So from Corollary 3.5 we gain the intuition that for fixed s the matching should approach equal percent bias reduction at rate $\frac{1}{p}$. It is also clear that, if the bias in the special covariates is small relative to the bias in the ellipsoidal covariates, the matching is approximately EPBR:

$$\mathbb{E}(\bar{Y}_{rt} - \bar{Y}_{rc}) = \rho_Y^p \cdot \mathbb{E}(Z_{rt}^p - Z_{rc}^p) \cdot ||\gamma^{(p)}|| + \gamma^{(s)} \cdot \mathbb{E}(\bar{X}_{rt}^{(s)} - \bar{X}_{rc}^{(s)}) \approx \rho_Y^p \cdot \mathbb{E}(Z_{rt}^p - Z_{rc}^p) \cdot ||\gamma^{(p)}||,$$

and assuming

$$||\mathbb{E}(\bar{X}_{mt}^{(s)} - \bar{X}_{mc}^{(s)})|| < ||\mathbb{E}(\bar{X}_{rt}^{(s)} - \bar{X}_{rc}^{(s)})||, \tag{4}$$

then it follows

$$\mathbb{E}(\bar{Y}_{mt} - \bar{Y}_{mc}) \approx \rho_Y^p \cdot \mathbb{E}(Z_{mt}^p - Z_{mc}^p) \cdot ||\gamma^{(p)}|| \implies \frac{\mathbb{E}(\bar{Y}_{mt} - \bar{Y}_{mc})}{\mathbb{E}(\bar{Y}_{rt} - \bar{Y}_{rc})} \approx \frac{\mathbb{E}(\bar{Z}^p{}_{mt} - \bar{Z}_{mc})}{\mathbb{E}(\bar{Z}^p{}_{rt} - \bar{Z}_{rc})},$$

as desired. Note that assumption (4) is quite plausible; this is the intuition that is the basis for matching on Mahalanobis distance in the first place. Below we prove a related but weaker statement than (4) in the case when matching with proportional covariance matrices and subsampling with replacement, which holds for all distributions in the treated and control populations.

Lemma 3.6. *Suppose that $\Sigma_c \propto I$, that we do not subsample treated units, and we subsample the control units with replacement.Then for any distribution of X_t, X_c we*

have

$$||\mathbb{E}(\bar{X}_{mt} - \bar{X}_{mc})|| \leq \mathbb{E}(\|\bar{X}_{rt} - \bar{X}_{rc}\|),$$

with equality if and only if $N_c = 1$.

Proof. For simplicity assume that our matching method subsamples the control units with replacement, so that each matched control unit x^j_{mc} has the same distribution. Then

$$\|\mathbb{E}(x_{mt} - x_{mc})\| = \|\mathbb{E}(x_t - x_c \mid x_t - x_c = \text{argmin}_{j \in 1 \dots N_c} \|x_t - x^j_c\|)\|$$

since $\Sigma_c \propto I$ implies that matching on Mahalanobis distance is the same as matching on the Euclidean norm. Let $x_t - x_c$ be denoted by y. Then $||y||\colon \mathbb{R}^p \to \mathbb{R}$ is a convex function, so by Jensen's inequality:

$$\|\mathbb{E}(y | y = \text{argmin}_{j \in 1 \dots N_c} \|y_j\|)\| \leq \mathbb{E}(\|y\| | y = \text{argmin}_{j \in 1 \dots N_c} \|y_j\|) =$$

$$\mathbb{E}(\min_{j \in 1 \dots N_c} ||y_j||) \leq \mathbb{E}(||y||),$$

where the last inequality is strict unless $N_c = 1$. But $\mathbb{E}(\|y\|) = \mathbb{E}(\|\bar{X}_{rt} - \bar{X}_{rc}\|)$, as desired. □

Lemma 3.6 is not the precise statement that we need (4), but it does lend credence to the notion that if the initial bias is small in norm, the norm of the matched bias will also be small. Finally, note that all of the results in this section hold for any affinely invariant matching method, including but not limited to Mahalanobis matching.

4. Principality in Mahalanobis Matching

Suppose that X_t, X_c are proportionally conditionally ellipsoidally distributed, and assume that they have the canonical form given in Section 3.1. Let $\mathcal{X}_t$ and $\mathcal{X}_c$ again

denote the data matrices consisting of N_t sampled treated units and N_c control units respectively.

Definition 5. *We say that a matching is principal in* $X^{(s)}$ *if* $(\mathcal{X}_t, \mathcal{X}_c) \overset{m}{\Longrightarrow} (T, C)$ *and* $(\mathcal{X}_t^{(r)}, \mathcal{X}_c^{(r)}) \overset{m}{\Longrightarrow} (T, C)$.

That is, applying the matching method to the first r covariates selects the same matched control units as applying the matching method to the full data matrix. Let p_I denote the probability that a given matching is principal in its special covariates, under Mahalanobis matching and subsampling the control units with replacement. Then in Section 4.1 we compute an asymptotic lower bound for p_I in the case where the $X^{(s)}$ are bounded. We are interested in p_I because it can give insight into scenarios when additional covariates affect the matching mechanism in a substantive way. Specifically, if Y denotes a linear outcome variable, then as $p_I \to 1$,

$$\mathbb{E}(\bar{Y}_{mt} - \bar{Y}_{mc}) \to \rho_Y^p \cdot \mathbb{E}(Z_{mt}^p - Z_{mc}^p) \cdot \|\gamma^{(p)}\| + \gamma^{(s)} \cdot \mathbb{E}(\bar{X}_{rt}^{(s)} - \bar{X}_{rc}^{(s)}),$$

since we can assume that $X^{(s)}, X^{(r)}$ are uncorrelated in canonical form, and so matching based on $X^{(r)}$ amounts to subsampling the $X^{(s)}$ randomly. Above, m denotes Mahalanobis matching on just the first p covariates. Another application of principality could be to matching in the case when there are latent covariates that are linearly related to the outcome variable, but are not accounted for in the matching. Bounding p_I provides a measure of the extent to which excluding these variables actually affects the matching [Dylan Small, The Wharton School, *personal communication*, 2/25/14].

4.1. **A Lower Bound for p_I.** We now obtain an asymptotic bound for p_I in the case when the $X^{(s)}$ are bounded, with covariance matrix $\Sigma^{(s)}$, and inverse covariance matrix $\Sigma^{(-s)}$. Let $I^{(j)}$ be the indicator variable of the event that the matching on

the j^{th} treated unit is principal, and let $\mathbb{E}(I_{(j)}) = p$. Then it is clear that, given the matching on the k^{th} unit is principal, it is more likely that the matching on the j^{th} unit is principal, since it indicates that the special covariates contributed less to the Mahalanobis distance. By this reasoning we have the bound $p_I \geq p^{N_t}$. So it remains to obtain an asymptotic lower bound for p, the probability that the matching on a single unit x_t is principal. For $j \in 1, 2, \ldots, N_c$, let $\epsilon_{(j)} = ||x_t^{(p)} - x_{cj}^{(p)}||^2$, which is just the Mahalanobis distance between the first p coordinates of x_t and unit j in the control population, since in the canonical form $\Sigma_c^{(p)} = I$. Let $\epsilon_{(1)}$ denote the minimal $\epsilon_{(j)}$. Then the matching is principal if for all other treated units j, the sum of $\epsilon_{(j)}$ and the distance coming from the binary covariates is greater than the sum of $\epsilon_{(1)}$ and the distance coming from the binary covariates. Recall that $X_c^{(s)}, X_c^{(p)}$ are uncorrelated, thus the probability the matching is principal is precisely

$$\mathbb{P}(\ \forall\ j,\ \ \epsilon_{(j)} - \epsilon_{(1)} > (x_{(1)}^{(s)} - x_t^{(s)})'\Sigma_c^{(-s)}(x_{(1)}^{(s)} - x_t^{(s)}) - (x_{(j)}^{(s)} - x_t^{(s)})'\Sigma_c^{(-s)}(x_{(j)}^{(s)} - x_t^{(s)})$$

Now suppose $(x_{(1)}^{(s)} - x_t^{(s)})'\Sigma_c^{(-s)}(x_{(1)}^{(s)} - x_t^{(s)}) - (x_{(j)}^{(s)} - x_t^{(s)})'\Sigma_c^{(-s)}(x_{(j)}^{(s)} - x_t^{(s)}) \leq c_s$, then $p \geq \mathbb{P}(\ \forall\ j,\ \ \epsilon_{(j)} - \epsilon_{(1)} > c_s)$. This probability is intractable as written, since the difference in two order statistics of a normal distribution is not well-understood. However, suppose we approximate $\epsilon_{(1)}$, the minimum of the $\epsilon_{(j)}$'s. It is of course a random variable, but as N_c and p get large the variance of $\epsilon_{(1)}$ approaches 0 at rate $O(1/\log(N_c))$, and the random variable converges to its expected value, which we denote c_1 [Pet00]. Then our bound becomes

$$\begin{aligned} p \geq \mathbb{P}(\ \forall\ j,\ \ \epsilon_{(j)} - \epsilon_{(1)} > c_s) &\approx \mathbb{P}(\ \forall\ j,\ \ \epsilon_{(j)} > c_s + c_1 | \epsilon_{(j)} > c_1) \\ &\geq \mathbb{P}(\epsilon_{(j)} > c_s + c_1 | \epsilon_{(j)} > c_1)^{N_c - 1}, \end{aligned}$$

where we have to add the last inequality since the $\epsilon_{(j)}$ are not independent, they are conditionally independent given x_t. Finally note that $\epsilon_{(j)} = \sum_{i=1}^{p} (x_t^i - x_c^i)^2$, where

$x_t^i - x_c^i$ are i.i.d $\sim N(\mu_t, 1+\sigma^2)$. For large p we use the central limit theorem to obtain $\epsilon_j \approx N((1+\sigma^2+\mu_t^2)p, (4\mu_t^2(\sigma^2+1)+2(\sigma^2+1)^2)p) = N(\bar{\mu}, \bar{\sigma}^2)$. Thus

$$\mathbb{P}(\epsilon_{(j)} > c_s + c_1 | \epsilon_{(j)} > c_1) \approx \frac{1-\Phi(\frac{c_s+c_1-\bar{\mu}}{\bar{\sigma}})}{1-\Phi(\frac{c_1-\bar{\mu}}{\bar{\sigma}})}.$$

It now remains to approximate c_1, and to compute an upper bound c_s. In [Roy82] an approximation for the first order statistic of N_c normals with mean $\bar{\mu}$ and variance $\bar{\sigma}^2$ is given as $\bar{\mu} + \Phi^{-1}(\frac{1-\alpha}{N_c-2\alpha+1})\bar{\sigma}$, with $\alpha = 0.375$. In [Har61], values of α are given for $n \leq 400$ which give an accurate approximation for the first moment of $\epsilon_{(1)}$ to within .001. Note that these approximations assume that the $\epsilon_{(j)}$'s are independent when in our case they are positively correlated. To address this, one can obtain the first moment of correlated normal order statistics via the technique in [OS62]. However, for our purposes it is sufficient to use the approximation assuming the $\epsilon_{(j)}$ are independent, since if they are positively correlated that will result in a higher minimum value $\epsilon_{(1)}$, and so our lower bound p would still hold. Using the Central Limit Theorem approximation for $\epsilon_{(j)}$ we take $c_1 = \bar{\mu} + \Phi^{-1}(\frac{1-\alpha}{N_c-2\alpha+1})\bar{\sigma}$, with α given in [Har61]. Finally we compute an upper bound $(x_{(j)}^{(s)} - x_t^{(s)})' \Sigma_c^{(-s)} (x_{(j)}^{(s)} - x_t^{(s)}) \leq c_s$. If $X^{(s)}$ is discrete or if each component of $X^{(s)}$ is bounded in absolute value by a constant k, letting $a = (x_{(j)}^{(s)} - x_t^{(s)})$ and $\Sigma = \Sigma_c^{(-s)}$, then

$$a'\Sigma a = \sum_{l,m} a_l a_m \sigma_{lm} \leq 4k^2 |\sum_{lm} (\sigma_{lm})|.$$

For example, when $X^{(s)}$ consists of Bernoulli random variables, $k = 1$, and we take $c_s = |\sum_{lm}(\sigma_{lm})|$. Then

$$p_I \geq \frac{1-\Phi(\frac{c_s+c_1-\bar{\mu}}{\bar{\sigma}})}{1-\Phi(\frac{c_1-\bar{\mu}}{\bar{\sigma}})}^{N_t(N_c-1)},$$

with $\bar{\mu}, \bar{\sigma}, c_s, c_1$ all defined above. Simplifying we obtain the below theorem.

Theorem 4.1. *Let X_t, X_c be conditionally elliposidally proportional, where $X^{(s)}$ are binary covariates. Define $\bar{\sigma}$ as above. Let p_I denote the probability that $X^{(s)}$ is principal. Then*

$$p_I \geq \left(\frac{1-\Phi(c_s/\bar{\sigma}+\Phi^{-1}(\frac{1-\alpha}{N_c-2\alpha+1}))}{\frac{N_c-\alpha}{N_c-2\alpha+1}}\right)^{N_t(N_c-1)},$$

where $c_s = |\sum_{lm}(\sigma_{lm})|$.

Studying the approximation in Theorem 4.1, we obtain a corollary:

Corollary 4.2. *For large N_c, as $p \to \infty, p_I \to 1$, where $p_I \to 1$ at rate $O(1/p)$.*

Proof. As $p \to \infty, c_s/\bar{\sigma} \to 0$, since $\bar{\sigma} = pc$. Then $\left(\frac{1-\Phi(c_s/\bar{\sigma}+\Phi^{-1}(\frac{1-\alpha}{N_c-2\alpha+1}))}{\frac{N_c-\alpha}{N_c-2\alpha+1}}\right)^{N_t(N_c-1)} \to 1^{N_t(N_c-1)} = 1$, at the rate that $c_s/\bar{\sigma} \to 0$, which is $O(\frac{1}{p})$. □

Note that the corollary holds not only when the $X^{(s)}$ are binary, but whenever they are bounded as well.

5. Simulations

In Sections 5.1 − 5.3 nearest-neighbor Mahalanobis matching is used. Nearest-neighbor matching randomly iterates through the treated units, selecting for each match the closest control unit, without replacement. In Section 5.4 control units are subsampled with replacement. We do not subsample treated units, and so throughout in the Mahalanobis distance we take $S = \Sigma_c$. In Section 5.1 we present simulation results that verify previous work in [RT92a] and [RS06], specifically that Mahalanobis matching on proportional ellipsoidal distributions is EPBR. In Section 5.2 we present simulation results that show that for conditionally ellipsoidal distributions the EPBR property fails, largely as a function of the number of additional discrete covariates introduced and the initial bias in the discrete covariates. When the initial bias is dominated by the normal covariates, the simulation results appear

approximately EPBR. In Section 5.3 we investigate Mahalanobis Matching on purely discrete covariates, and show that bias reduction both decreases in magnitude over most directions relative to matching with normal distributions, and fails to retain the property of EPBR. In Section 5.4 we compute the probability that additional discrete covariates affect the matching when the covariates are conditionally proportionally ellipsoidal, and show that in practice the asymptotic lower bound computed in Section 4.1.1 generally holds. In Section 5.5 simulations validate the asymptotic formula for bias reduction derived in Section 2. In each section we describe simulation conditions, and briefly discuss the simulation results. Code is attached at the end of the Appendix.

5.1. **Mahalanobis Matching with Ellipsoidal Covariates.** In [RT92a], Rubin and Thomas show that applying an affinely invariant matching method to any proportionally ellipsoidal distribution is EPBR. We verify these results via a simulation where random outcome directions are generated (uniform distribution on the unit sphere), and then control and treated data are drawn from normal distributions with proportional covariance matrices. The positive definite covariance matrix we use (Σ_t) was generated by the *clusterGeneration* package in R, and remains fixed throughout the simulations. For the matching to be EPBR the average bias reduction along each of the different outcome directions should be the same. Our initial conditions for the simulations were: $N_\rho = 20, \mu_t = (3,2,1,4,2), \mu_c = (2,4,1,5,6), N_t = 50, N_c = 500, \Sigma_c = 3 \cdot \Sigma_t,$ nsims $= 100$ where ρ is the number of outcome directions tested, and nsims is the number of simulations that bias reductions were averaged over. The bias reduction is approximately constant across outcome directions, with a standard deviation of only 0.06. This result is unsurprising since it has been proven theoretically, but is included here for completeness and to give the reader an idea of what EPBR in outcome directions looks like.

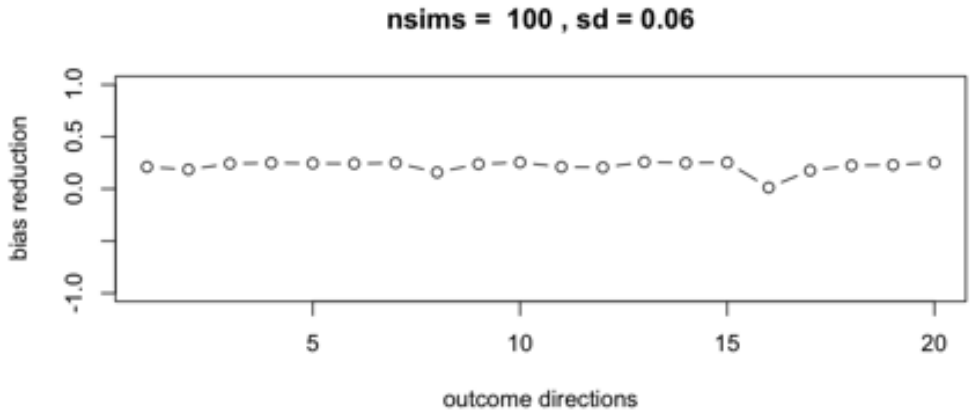

FIGURE 1. Bias reduction for five normal covariates

5.2. **Mahalanobis Matching with Ellipsoidal and Discrete Covariates.** We explore the case of matching with conditionally ellipsoidal distributions where the additional covariates are independent and binary. Figure 4 shows the bias reduction with three normal covariates, three binary covariates, and means for normal and binary covariates chosen uniformly. The normal covariates have proportional covariance matrices in the treatment and control distributions, with constant of proportionality 2 in all simulations. Figure 5 shows results under the same settings but with the initial bias in the binary covariates scaled down by a factor of 0.09. Figure 6 displays matching in the setting of one binary covariate and nine normal covariates. The same covariance matrices and outcome directions are used in the first two simulations.

In the first simulation the standard deviation of the bias reduction across each outcome direction is 0.26, significantly larger than the .06 value observed in Section 5.1. When the exact same matching settings are used but the discrete covariate means are scaled down by a factor of 0.09 the standard deviation drops to a minuscule 0.02, which is indistinguishable from the results in Section 5.1. The third

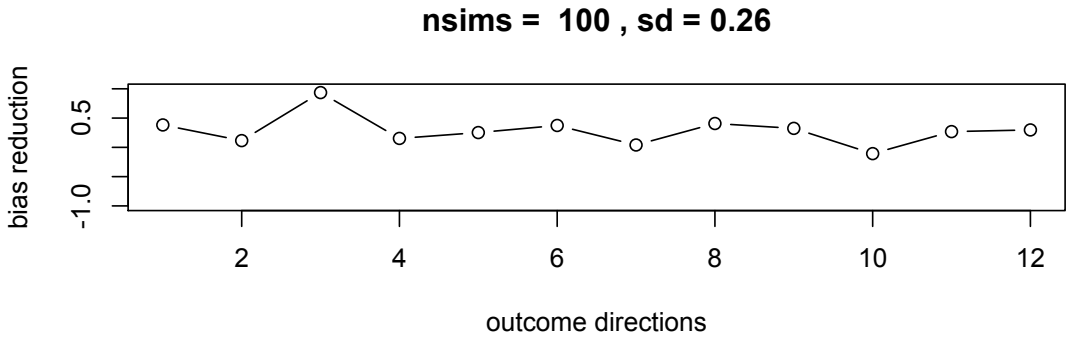

FIGURE 2. Bias reduction for three normal and three binary covariates

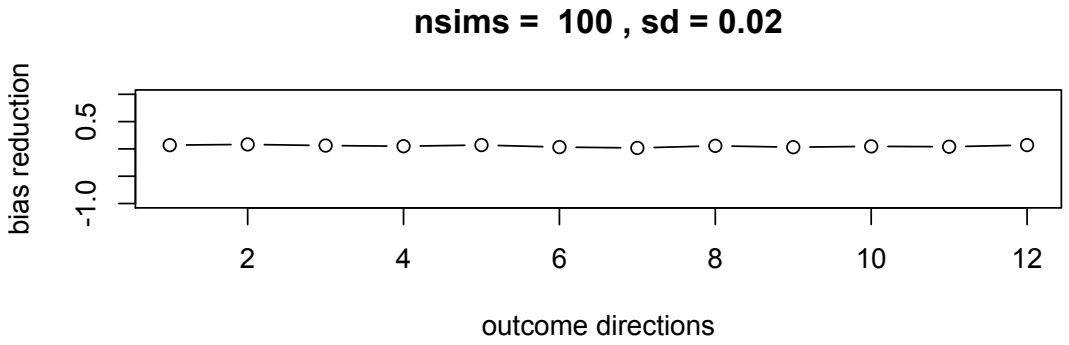

FIGURE 3. Bias reduction for three normal and three binary covariates with binary covariate means scaled towards 0

simulation examines how the bias reduction drops when the number of normal covariates grows relative to the number of binary covariates; and it does indeed drop to 0.16. Note that we do not compare the second and third simulations because in the third simulation the discrete covariate means are not scaled down.

To summarize, simulation results show that when discrete covariate means are non-negligible, EPBR breaks down. As a result, the appearance of EPBR increases

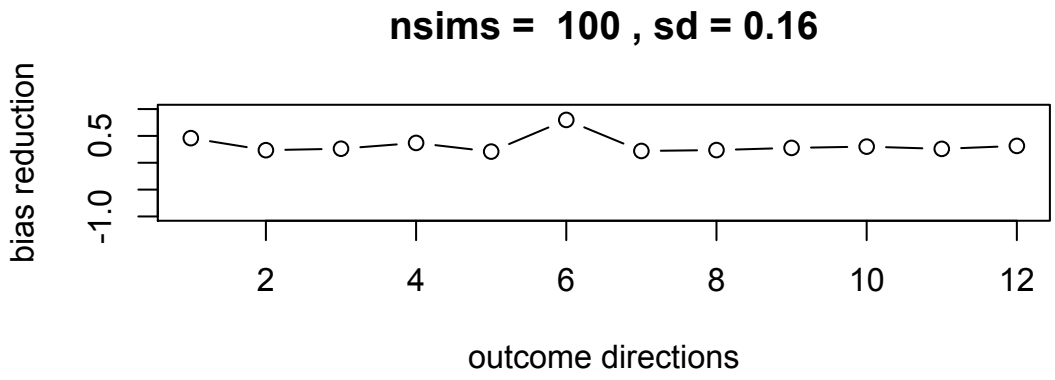

FIGURE 4. Bias reduction for nine normal and one binary covariate

with both scaling down discrete covariate means and increasing the number of normal covariates; although it appears that the former has a larger effect. These results are explained by the discussion following Corollary 3.5.

5.3. **Matching on Discrete Covariates.** We perform Mahalanobis matching in the setting where the treatment and control distributions are binary covariates, and we do not subsample the treated units. We perform 500 simulations, where each treatment and control vector consists of 10 independent Bernoulli covariates, with probabilities p_t, p_c. The parameters were set to $N_t = 50, N_c = 500, N_\rho = 10$, and the initial probability vectors were chosen uniformly. The results are displayed below. Across the 10 different outcome directions the mean reduction in bias was 0.98, with four of the matchings actually increasing bias. The standard deviation in ρ was 0.20, showing that the bias reduction was markedly non-constant. In contrast to the previous cases, it is clear that Mahalanobis matching on binary covariates is certainly not EPBR, and in most outcome directions has a negligible effect on reducing bias. This simulation fits into the picture of how proportionally conditionally ellipsoidal distributions lose the EPBR property the more discrete

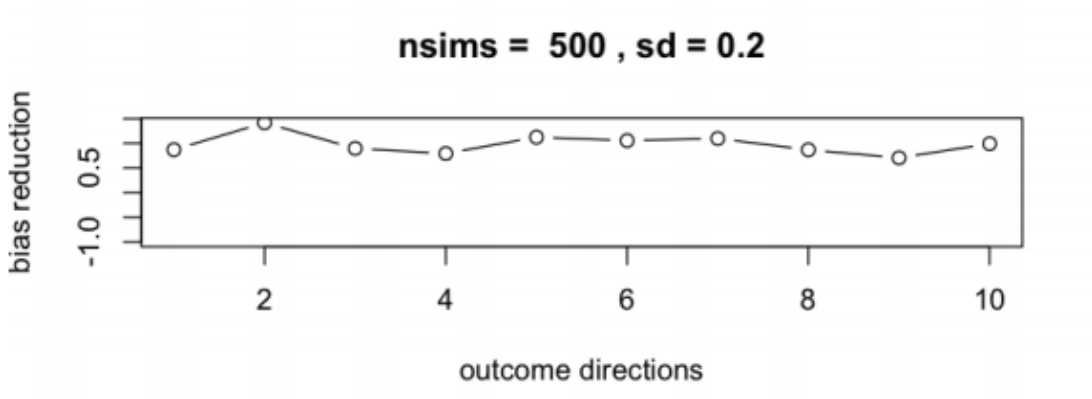

FIGURE 5. Bias reduction for ten discrete covariates

covariates are added to the matching. Gu and Rosenbaum remark in [GR93] that when Mahalanobis matching on binary covariates, the method fails due to extreme sensitivity to outliers. Specifically, if one of the binary covariates has mean p and $p \to 0$, then if the covariate realizes in the sample it will dominate the Mahalanobis distance since the variance $p(1-p)$ approaches 0. Since one coordinate is dominating the distance it is not surprising that in the setting of many binary covariates it neither reduces bias effectively nor is equal percent bias reducing.

5.4. **Convergence to Principality: Discrete Case.** The asymptotic lower bound for p_I relies on two approximations: the convergence of the sum of p chi-squared distributions to a normal distribution, and the actual value of $\epsilon_{(1)}$ being very close to its expected value. These two conditions require sufficiently large p and N_c for the bound to actually be a lower bound. For the lower bound to be close to the true probability, it is clear that the Mahalanobis distance should frequently approach the maximum distance, and that the $\epsilon_{(j)}$'s for a given x_t be roughly independent. Figure 5 displays Monte Carlo values for the probability of principality for a unit x_t, and below it the calculated lower bound. We start with 100 normal covariates, two binary covariates, and with $N_c = 500, \Sigma_c = 0.5 \cdot \Sigma_t$, and with the means randomly

1	2	3	4	5	6	7	8	9	10
0.45	0.65	0.95	0.75	0.75	0.85	0.75	0.9	0.85	0.8
0.002	0.695	0.75	0.57	0.741	0.758	0.625	0.583	0.716	0.643

FIGURE 6. Approximate upper bound vs. simulated probability of Principality

drawn. In most of the outcome directions the lower bound is relatively tight, and in all of the directions it is in fact a lower bound, except for column 2, where the values are still very close. In column 5 the bound is within 0.09, and in all of the columns but column 1 and 8 the bound is within 0.2 of the simulated probability. The reason for column 2 not being a lower bound can likely be attributed to the number of simulations being set at only 20 due to time constraints. The distance of the bound in column 1 from the simulated bound could be explained by several factors including a large value of x_t which leads to large correlation in the $\epsilon'_{(j)}s$, or to the maximal Mahalanobis bound being very high relative to the observed Mahalanobis distances.

5.5. **Matching on a Single Binary Covariate.** We illustrate Theorem 2.1 in the case where $N_t = 100, N_c = 500$. The approximation relies on the sum of N_t, N_c binary covariates approaching the normal distribution, and hence the approximation holds only when N_t, N_c are reasonably large. We simulate the final bias, averaging over 100 simulations, and compare this sample average to our approximation. We vary p_t from 0.75 to 0.95 and N_c from 0.10 to 0.15. We keep a difference in p_t and p_c of 0.65 because otherwise our simulated bias will nearly always be 0 since $N_c >> N_t$. The results in Figure 6 show the approximation and then the simulated bias in each column, with increasing values of p_t, p_c from left to right. The mean difference between the approximation and the simulated bias is only 0.0048, which convinces us of the precision of our approximation.

1	2	3	4	5	6	7	8	9	10
0.244	0.239	0.235	0.23	0.224	0.219	0.214	0.209	0.204	0.199
0.255	0.244	0.235	0.232	0.231	0.219	0.221	0.21	0.21	0.208

FIGURE 7. Approximated bias after matching vs. simulated values

6. Further Questions

This thesis on equal percent bias reduction and Mahalanobis matching presents a foundation from which to explore several promising directions. We now summarize some possibilities for future investigation.

Problem 1. *In Theorem* 2.1 *the asymptotic formula for matched bias is a function of* $\sigma_c, \sigma_t, N_t, N_c$, *and* $\mu_t - \mu_c$, *where* $\mu_i = p_i \cdot N_i$. *We conjecture that for* $n > 1$ *binary covariates the matched bias is a function of* $\mu_t - \mu_c, N_t, N_c, \Sigma_c, \Sigma_t$.

Problem 2. *The lower bound for* p_I *can be tightened by using the approximate moment for* $\epsilon_{(1)}$ *that accounts for correlation between the* $\epsilon_{(j)}$*'s, via the technique in* [OS62]. *Obtaining an upper bound for* p_I *would also be of theoretical interest.*

Problem 3. *Can the asymptotic bound for* p_I *be extended to the case when additional covariates are not bounded? Specifically, if the additional covariates are normally distributed (but not proportionally), can we make a probabilistic statement bounding the sample Mahalanobis distance, perhaps using techniques similar to the limit laws for singular values of random matrices?*

Problem 4. *Applications of the concept of principality to latent bounded covariates with a linear relationship to the outcome variable are yet to be explored.*

Problem 5. *Why precisely does EPBR break down for binary covariates, past the reasoning from* [GR93] *given in Section* 5.3*? Perhaps a combinatorial approach in the style of Section* 2 *could be useful.*

7. Appendix

7.1. Integrals from Section 2.1.

- $\int_{-\infty}^{\infty} z\Phi(az-b)\phi(z)dz = \int_{-\infty}^{\infty}\int_{\infty}^{az-b} z\phi(z)\phi(y)dydz = \int_{-\infty}^{\infty}\int_{\frac{y+b}{a}}^{\infty} z\phi(z)\phi(y)dzdy =$

  $\int_{-\infty}^{\infty}\phi(y)\phi(\frac{y+b}{a})dy = \frac{a}{\sqrt{a^2+1}}\phi(\frac{-b}{\sqrt{a^2+1}})$,
  where the last equality comes from completing the square and the fact that the normal probability density function integrates to 1.
- Let Y, Z be distributed i.i.d standard normal.
  $\int_{-\infty}^{\infty}\Phi(az-b)\phi(z)dz = \int_{-\infty}^{\infty} P(Y \leq aZ-b|Z=z)\phi(z) = \int_{-\infty}^{\infty} P(Z=z|Y \leq aZ-b)P(Y \leq aZ-b) = P(Y \leq aZ-b)\int_{-\infty}^{\infty} P(Z=z|Y \leq aZ-b) = P(Y \leq aZ-b)$,

  since the conditional probability density function of Z integrates to 1, and the last expression is easily seen to be $\Phi(\frac{-b}{\sqrt{1+a^2}})$.

7.2. **R Code.** Attached is the R code written to simulate Mahalanobis matching with discrete and normal covariates, exact matching on one binary covariate, and to compute the proportion of times additional discrete covariates are principal, in the sense of Section 4.

R Code: Mahalanobis Matching with Normal and Binary Covariates

```
#test
#initialize parameters
install.packages("clusterGeneration")
require(clusterGeneration)
install.packages("Matrix")
require("Matrix")
N_rho = 10
N_t = 50
N_c = 500
nbin = 3
nnorm = 9
mult = 5
n = nnorm + nbin
GAMMA = mvrnorm(N_rho,rep(0,n),diag(n))
prob_t = runif(nbin,0,1)
prob_c = runif(nbin,0,1)
mu_t = runif(nnorm,0,1)
mu_c = mult*runif(nnorm,0,1)
alpha = 2
sigma_t = genPositiveDefMat(nnorm,"eigen")$Sigma
sigma_c = alpha*sigma_t
cov_bin = diag(prob_c*(1-prob_c), nbin, nbin)
cov = bdiag(cov_bin, sigma_c)
COR = rep(NA, N_rho)
rho = rep(NA, N_rho)
# simulations
nsims = 500
for(j in 1:N_rho)
{
 gamma = GAMMA[j,]
 gamma = gamma/sqrt(sum(gamma^2))
 post = rep(NA,nsims)
 bias = rep(NA, nsims)
 for(m in 1:nsims)
 {
      # generate random samples
      X_t = matrix(nrow = n, ncol = N_t)
      X_c = matrix(nrow = n, ncol = N_c)
      if(nbin != 0)
      {
            for(i in 1:nbin)
            {
                  X_t[i,] = rbinom(N_t,1,prob_t[i])
                  X_c[i,] = rbinom(N_c,1,prob_c[i])
            }
      }
      for(k in 1 : N_t)
      {
            X_t[(nbin+1):(n),k] = mvrnorm(1,mu_t, sigma_t)

      }
      for(u in 1:N_c)
      {
            X_c[(nbin+1):(n),u] = mvrnorm(1,mu_c, sigma_c)
      }

      # compute the initial bias in the random samples
      bias[m] = gamma %*% (rowMeans(X_c)-rowMeans(X_t))
```

```
      matched_treated = matrix(data = NA, nrow = n, ncol = N_t)
      for(l in 1:N_t)
      {
            x_1 <- X_t[,l]
            distances = mahalanobis(t(X_c),x_1,cov)
            index = which(distances == min(distances))
            matched_treated[,l] = X_c[,index[1]]
                  X_c <- X_c[,-index[1]]
      }

      post[m] = gamma %*% (rowMeans(matched_treated)-rowMeans(X_t))
}
rho[j] = sum(post)/sum(bias)
COR[j] = cov(post,bias)
}
plot(rho, main = paste("nsims = ",nsims,", sd =",round(sd(rho),2)), type = "b", xlab =
"outcome directions", ylab = "bias reduction", ylim = c(min(-
1,min(rho)),max(1,max(rho))))
```

R Code: Exact Matching on 1 Binary Covariate

```
# Exact Matching on 1 Binary Covariate
install.packages("clusterGeneration")
require(clusterGeneration)
install.packages("Matrix")
require(Matrix)
ndir = 10
table = matrix(nrow = 2, ncol = ndir)
for(u in 1:ndir)
{
p_t = .75 + .2*(u/ndir)
p_c = .1 + .05*(u/ndir)
# number of treated cov
Nt = 100
# number of control cov
Nc = 500
# number of simulations to compute average bias reduction over
nsims = 100
post = rep(NA, nsims)
bias = rep(NA, nsims)

 for(i in 1:nsims)
 {

      # sample Nt treated units

      X_t = rbinom(Nt,1,p_t)
      # sample Nc control units
      X_c = rbinom(Nc,1,p_c)
      # initial bias
     bias[i] =  mean(X_t)-mean(X_c)
     matched_treated = rep(NA, Nt)
 # select Nt closest matches from control population without replacement
 for(l in 1:Nt)
 {
      x_1 <- X_t[l]
      index = 1
```

```
      for(k in 1:length(X_c))
      {
            if(x_1 == X_c[k])
            {
                  index = k
                  break
            }
      }
      matched_treated[l] = X_c[index]
            X_c <- X_c[-index]
 }
    # compute the posterior bias
 post[i] = mean(X_t) - mean(matched_treated)

 }

# theoretical bias: E(Y_mt-Y_mc)
mu_t = p_t*Nt
mu_c = p_c*Nc
sigma_c = sqrt(Nc*p_c*(1-p_c))
sigma_t = sqrt(Nt*p_t*(1-p_t))
a_1 = sigma_t/sigma_c
a_2 = sqrt(1 + a_1^2)
b = (mu_c-mu_t)/sigma_c
b_2 = (mu_c-mu_t + Nt - Nc)/sigma_c

approx = p_t -1/(Nt)*((mu_c-mu_t)*(pnorm(-b/a_2)-pnorm(-b_2/a_2)) +
 sigma_t*(a_1/a_2)*(dnorm(-b_2/a_2)-dnorm(-b/(a_2)))+ mu_c)

table[1,u] = approx
table[2,u] = mean(post)
}

table
```

R Code: Irrelevance for Mahalanobis matching

```
#test
install.packages("clusterGeneration")
require(clusterGeneration)
install.packages("Matrix")
require("Matrix")
table = matrix(nrow = 2, ncol = 10)
for(l in 1:1)
{
N_c = 500
nbin = 2
nnorm = 100
n = nnorm + nbin
prob_t = rep(1,nbin)-.02*runif(nbin,0,1)
prob_c = runif(nbin,0,1)*.02
m_t = runif(nnorm,0,1)
mu_c = runif(nnorm,0,1)
sigma_nu = .2
sigma_c = genPositiveDefMat(nnorm,"eigen")$Sigma
sigma_t = sigma_nu*sigma_c
cov_bin = diag(prob_c*(1-prob_c), nbin, nbin)
cov = bdiag(cov_bin, sigma_c)
 nsims = 20
```

```
count = 0
for(m in 1:nsims)
{
     # generate random control samples and a  treated unit.
     X_c = matrix(nrow = n, ncol = N_c)
     for(i in 1:nbin)
     {
          X_c[i,] = rbinom(N_c,1,prob_c[i])
     }
     for(u in 1:N_c)
     {
          X_c[(nbin+1):(n),u] = mvrnorm(1,mu_c,    sigma_c)
     }
       x_t = c()
       for(y in 1:nbin)
       {
           x_t = c(x_t,rbinom(1,1,prob_t[y]))
       }
       xnorm = mvrnorm(1,m_t,sigma_t)
       x_t = c(x_t, xnorm)
     distances = mahalanobis(t(X_c),x_t,cov)
     index1 = which(distances == min(distances))
     X_cnorm = X_c[(nbin+1):n,]
     distances2 = mahalanobis(t(X_cnorm),xnorm,sigma_c)
     index2 = which(distances2 == min(distances2))
     if(index1 == index2)
     {
          count = count + 1
     }

}


prob = count/nsims
mu_t = sqrt(sum(solve(sigma_c)%*% (m_t-mu_c))^2)/sqrt(nnorm)
c_s = sum(sum(solve(cov_bin)))
alpha = .375
sigma_bar = sqrt((4*mu_t^2*(sigma_nu^2+1) + 2*(sigma_nu^2 +1)^2)*nnorm)
mu_bar = nnorm*(1+sigma_nu^2 + mu_t^2)
     c_1 = mu_bar + sigma_bar*qnorm((1-alpha)/(N_c-2*alpha+1))
ign = (1-pnorm(c_s/sigma_bar + (c_1-mu_bar)/sigma_bar))/((N_c-alpha)/(N_c-2*alpha+1))


table[1,l] = prob
table[2,l] = ign^N_c
}
```

Acknowledgments

I would like to first and foremost thank my thesis advisor, Natesh Pillai, for suggesting this thesis topic, and for meeting throughout the year to discuss the results. I would like to thank Donald Rubin, Dylan Small, and Paul Rosenbaum for commenting on the results, and especially Steven Finch for numerous and extensive copy-edits to the manuscript. I would also like to thank my roommate Alden Green for frequently useful discussions and suggestions, particularly pertaining to the integrals in Section 6.1 of the Appendix. Finally, I would like to acknowledge the many amazing friends I've made during my time at Harvard, as well as my brother and parents' unwavering support.

References

[Dem69] A.P. Dempster. *Continuous Multivariate Analysis*. Addison-Wesley, Reading, Massachusetts., 1969.

[GR93] Xing Gu and Paul Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Taylor and Francis Group*, 1993.

[Har61] H. Leon Harter. Expected values of normal order statistics. *Biometrika*, 48(1):151–165, 1961.

[OS62] D.B. Owen and G.P. Steck. Moments of order statistics from the equicorrelated multivariate normal distribution. *The Annals of Mathematical Statistics*, 33(4):1286–1291, 1962.

[Pet00] Max Petzold. A note on the first moment of extreme order statistics from the normal distribution, 2000.

[Roy82] J.P. Royston. Expected normal order statistics (exact and approximate). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(2):161–165, 1982.

[RR83] Donald Rubin and Paul Rosenbaum. The central role of propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[RS06] Donald Rubin and Elizabeth Stuart. Affinely invariant matching methods with discriminant mixtures of ellipsoidally symmetric distributions. *Annals of Statistics*, 34:1814–1826, 2006.

[RT92a] Donald Rubin and Neal Thomas. Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics*, 20(2):1079–93, 1992.

[RT92b] Donald Rubin and Neal Thomas. Characterizing the effect of matching using linear propensity score methods with normal covariates. *Biometrika*, 79:797–809, 1992.

[RT96] Donald Rubin and Neal Thomas. Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52:249–264, 1996.

[RT00] Donald Rubin and Neal Thomas. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95:573–585, 2000.

[Rub76a] Donald Rubin. Inference and missing data (with discussion). *Biometrika*, 63:581–92, 1976.

[Rub76b] Donald Rubin. Multivariate matching methods that are equal percent bias reducing, i: Some examples. *Biometrics*, 32(1):109–120, 1976.

[Rub76c] Donald Rubin. Multivariate matching methods that are equal percent bias reducing, ii: Maximums on bias reduction for fixed sample sizes. *Biometrics*, 32(1):121–132, 1976.

[Rub79] Donald Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366):318–328, 1979.

[Rub80a] Donald Rubin. Bias reduction using mahalanobis-metric matching. *Biometrics*, 36:293–298, 1980.

[Rub80b] Donald Rubin. Discussion of paper by d. basu. *Journal of the American Statistical Association*, 75:591–593, 1980.

[Stu10] Elizabeth Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Sciences*, 25(1):1–23, 2010.

[Zha04] Zhong Zhao. Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. *The Review of Economics and Statistics*, 86(1):91–107, 2004.

Department of Mathematics, Harvard University, Cambridge, MA
E-mail address: sethneel@college.harvard.edu