SYNTHETIC REGRESSION DISCONTINUITY
ESTIMATING TREATMENT EFFECTS USING MACHINE LEARNING

Jörn Boehnke [1] & Pietro Bonaldi [2]

August 28, 2019

Abstract

In the standard regression discontinuity setting, treatment assignment is based on whether a unit's observable score (running variable) crosses a known threshold. We propose a two-stage method to estimate the treatment effect when the score is unobservable to the econometrician while the treatment status is known for all units. In the first stage, we use a statistical model to predict a unit's treatment status based on a continuous synthetic score. In the second stage, we apply a regression discontinuity design using the predicted synthetic score as the running variable to estimate the treatment effect on an outcome of interest. We establish conditions under which the method identifies the local treatment effect for a unit at the threshold of the unobservable score, the same parameter that a standard regression discontinuity design with known score would identify. We also examine the properties of the estimator using simulations, and propose the use machine learning algorithms to achieve high prediction accuracy. Finally, we apply the method to measure the effect of an investment grade rating on corporate bond prices by any of the three largest credit ratings agencies. We find an average 1% increase in the prices of corporate bonds that received an investment grade as opposed to a non-investment grade rating.

Keywords: Causal Identification, Machine Learning, Regression Discontinuity Design

JEL Classification: C14, C21, C45, C63

[1] University of California, Davis, Graduate School of Management

[2] Carnegie Mellon University, Tepper School of Business

We thank Ashwin Aravindakshan, Mark Egan, Jeremy Fox, Richard Freeman, Sonia Jaffe, Jacob LaRiviere, Kyle Luh, Alex MacKay, Sarah Moshary, Brian Quistorff, Karl Schurter, Philippe Sosoe, Sergiy Verstyuk, Shing Tung Yau, and all participants in research seminars / conferences at Harvard CMSA, Microsoft's Office of the Chief Economist, PennState Smeal, and Rice Jones for valuable comments and fruitful discussions.

1 Introduction

Regulators, companies, and institutions regularly make binary decisions affecting several units such as companies, individuals, and cities. These binary decisions are equivalent to assigning a treatment status to a chosen set of units. Very often, the treatment decision can be described as based on an underlying score assigned to all units; those with a score above a threshold receive treatment while those with a score below it remain untreated. We focus on cases where the score is only known to the decision maker, while treatment status is known to everyone. Examples include credit rating agencies evaluating bonds as investment grade or non-investment grade, banks granting loans after performing credit risk assessments, and prospective students or employees being chosen based on proprietary selection methods. In this paper, we propose a method to estimate the treatment effect for a unit at the threshold of the unobservable score.

We propose a two-stage method to estimate the treatment effect when the score is unobservable to the econometrician while the treatment status is known for all units. We assume that a potentially large set of observable determinants of the score is available. In the first stage, we use a statistical model to predict units' treatment status based on a continuous estimated synthetic score (classification). In the second stage, we apply a regression discontinuity (RD) design using the synthetic score as the running variable to estimate the local treatment effect. We call this method *synthetic regression discontinuity* (SynRD) design.

SynRD does not require the decision maker to explicitly compute a numerical score as long as treatment assignment can be described as if it were implicitly based on such a score. This is analogous to a utility maximization model where it is assumed that individuals rank alternatives according to the utility derived, even if they do not explicitly compute a utility function. We furthermore argue that the proposed method may also be applicable when units self-select into treatment based on their own characteristics.

We show that under continuity and smoothness of both the unobserved and the synthetic score, and

perfect prediction in the first stage, SynRD identifies the treatment effect for a unit at the threshold of the unobservable score. This parameter is the same that a standard RD design with known score would identify. We examine the properties of the estimator when the first stage prediction accuracy is imperfect using simulations. Not surprisingly, these simulations suggest that the quality of the overall estimation depends positively on the accuracy of the classification. Therefore we implement flexible machine learning algorithms to achieve a high first stage prediction accuracy.

We apply SynRD to analyze the effect of an investment grade credit rating on corporate bond prices. At different points in time, credit rating agencies analyze several characteristics of bonds and their issuers and determine their rating. We model the agencies' decision as if it were based on an explicit or implicit score assigned to each bond. In this application, we focus on a binary classification: bonds are classified either as investment grade (I) or non-investment grade (N). It is hard to measure the direct effect of a bond's grade on its price because the rating reflects information about the quality of the bond that is likely already available to investors at the time of the rating. SynRD overcomes this endogeneity by first predicting the score assigned to each bond and then estimating the local treatment effect based on this synthetic score. We find a 0.8% - 1.1% increase in price as a consequence of receiving an investment grade rating for bonds at the threshold of the synthetic score.

The remainder of this paper is structured as follows. After a discussion of the related literature in section 2, section 3 formally introduces the SynRD design. In section 4 we set up the mathematical framework underlying SynRD and prove identification for the case with perfect prediction in the first stage. Imperfect prediction is discussed in section 5 where we report simulations evaluating the performance of SynRD. Section 6 discusses the application of SynRD to credit ratings, and section 7 concludes the paper.

2 Related Literature

RD designs have broad applicability in social sciences because they allow for causal identification with observational data. Thistlethwaite and Campbell (1960) introduced the RD design to study the effect

of scholarships on career plans where award decisions were based on a test score crossing some threshold. Since the 1990s there have been several contributions to the RD literature. Imbens and Lemieux (2008) and Lee and Lemieux (2010) provide surveys of the literature.

Hahn et al. (2001) discuss the sources of identification in the RD design and show that the treatment effect can be non-parametrically identified. Imbens and Kalyanaraman (2012) derive an optimal bandwidth for local linear regression. Calonico et al. (2014) and Calonico et al. (2018) build on their work and provide standard errors and confidence intervals for RD designs that are robust to the choice of bandwidth. Calonico et al. (forthcoming) study how to include covariates in the estimation of RD designs. We follow Calonico et al. (2014) and Calonico et al. (forthcoming) to implement the second stage RD for the simulation and application, using the first stage predicted score as the running variable.

None of these previous studies address the case in which the score (also referred to as running or forcing variable) is not observable to the econometrician. We extend the argument in Hahn et al. (2001) and prove identification of the treatment effect when the first stage procedure perfectly predicts treatment status. A related strand of the literature studies RD design with a noisy score, that is, when the running variable is measured with error. Battistin et al. (2009) assume there is a fraction of units for which the score is known with zero error. Davezies and Barbanchon (2017) show the treatment effect is identified when both the noisy and the true score are observed for a subset of all treated units. These approaches are obviously not feasible in the case we are concerned with here. When the treatment status is observed (an assumption we maintain throughout), Yu (2012) and Pei and Shen (2017) find conditions under which the treatment effect is identified, despite the fact that the econometrician only observes a noisy measure of the true score used for assignment. In work under progress (not yet included in this version), we analyze whether these or related conditions apply to the second stage RD if the treatment status prediction is not perfect in the first stage.

Porter and Yu (2015) study RD when the discontinuity point (the threshold for treatment assignment) is unknown. They propose a two stage estimation where, in the first step, the discontinuity is estimated with a difference kernel estimator. They show that the first stage does not affect the asymp-

totic efficiency of the treatment effect estimator. The case we analyze in this paper is more general, since it requires a first stage estimation of not just the discontinuity (only one parameter) but also the score for every unit in the test sample.

The second area of research this work builds on is machine learning. While not yet known as "machine learning," McCulloch and Pitts (1943) was one of the first publications that laid out an estimation approach reflecting the operation of the human brain, the neural network. Among many others, Hopfield (1982) and White (1992) further established the foundations that replicate logic using a large number of simple equivalent components – neurons – for estimation purposes. Sarle (1994) provides an intuitive translation between previous statistical work and the machine learning literature.

While the method proposed in this work is not limited to any one specific statistical method, a high prediction accuracy in the first stage is important to obtain unbiased estimates of the treatment effect. It has been found that artificial neural networks achieve a very high out-of-sample prediction accuracy. The specific machine learning algorithm we employ in our application is a multilayer perceptron (MLP), an artificial neural network with multiple hidden layers. MLPs are universal approximators (White 1992). I.e., MLPs are flexible, general-purpose, non-linear models that, given enough data and enough hidden neurons, can approximate any function to any desired degree of accuracy. Regarding the practical implementation of the MLP, our work is building on Kour and Saabne (2014a), Kour and Saabne (2014b), and Hadash et al. (2018). We also utilize the literature of ordinal classification. Cheng et al. (2008) and Niu et al. (2016) derive a representation of labels to measure the cross entropy error; a number of binary classifiers predicting whether a data point is larger than a threshold. Both publications adapt a traditional neural network to learn ordinal categories.

Finally, the credit rating application presented in this paper is building on a body of literature in accounting, finance, and information systems. Closely related to the first stage estimation of this application are publications by Hájek (2011) and Huang et al. (2004). Both use machine learning tools to predict credit ratings. Hájek (2011) uses neural networks and Huang et al. (2004) compare the performance of support vector machines and neural networks in predicting credit ratings. Their main

focus is prediction rather than causal analysis.

Another strand of the literature in accounting and finance concentrates on the effect of credit ratings. Sufi (2007) analyzes the effects of the introduction of syndicated loan ratings on firm outcomes. Tang (2009) uses a credit rating refinement by Moody's as an exogenous source of variation, to analyze the effect of ratings on firms' financing and investing decisions. Almeida et al. (2017) document a sovereign ceiling policy by the rating agencies, and exploit sovereign downgrades as shocks to corporate ratings to study their effect on firms' decisions and bond yields.

Hand et al. (1992) find a -1,27% excess return (net of accrued interest) for all downgrades in their sample, and a (non-significant) -0.37% excess return for downgrades that were not contaminated by concurrent news about the issuer. More recently, using data on corporate bonds from TRACE (2002 - 2009), May (2010) report significant abnormal returns of -0.64% for all downgrades in the uncontaminated sample, when focusing only on prices of bonds that traded both on the day before and the day after the change in rating. The effect seems smaller in magnitude for bonds issued by firms rated as investment grade before the downgrade (-0.45%) than for non-investment grade firms (-0.83%). Both studies report much smaller and non-significant effects for upgrades.

Although similar in magnitude, these previous estimates differ from ours in several respects. To begin with, we are not just measuring the price changes associated with upgrades or downgrades, but instead we are estimating the causal effect on prices of downgrading a bond from investment to non-investment grade keeping other price-relevant variables constant. By implementing an RD design, our method provides causal identification at the threshold. Moreover, SynRD does not rely on exogenous shocks or natural experiments that might constrain the sample and raise concerns about external validity. An advantage of the approach we propose is that we can consider longer time windows around the rating event without having to worry about concurrent news about the issuer or other events that might affect prices directly.

3 Method

Suppose a decision maker assigns treatment to a set of units based on observable characteristics. We focus on cases where the treatment assignment is either explicitly or implicitly based on a numerical score computed for each unit as a function of these characteristics. This is analogous to a utility maximization model where it is assumed that individuals choose the option that provides the highest utility, even if they do not explicitly compute a utility function. A unit receives treatment if its score is greater than or equal to a predetermined threshold. When the score is observable, the traditional RD design identifies the local average treatment effect for those units with scores equal to the threshold. We propose a two-stage method to estimate this parameter when the score is unobservable and treatment status is known for all units. We call it synthetic regression discontinuity (SynRD), because it is based on a predicted synthetic score recovered from a statistical model designed for predicting treatment assignment.

As an illustration, consider a simple case where the decision maker assigns treatment to all units based on two characteristics, both measured as real numbers. Each unit can be described as a pair of characteristics (x_1, x_2), and the decision maker can be thought of as having preferences over the set of all units regarding whether they should receive treatment. Such preferences are represented using a continuous score $s(x_1, x_2) \in \mathbb{R}$. Moreover, there is a threshold τ_s such that the decision maker is indifferent between assigning treatment or not to any unit with $s(x_1, x_2) = \tau_s$. Figure 1 illustrates this simple model using level sets of score s, analogous to a standard indifference curve map. All units with a score greater than or equal to $\tau_s = 0.5$ receive treatment, otherwise they are assigned to the control group. We arbitrarily assume that the units right at the threshold where the decision maker is indifferent are treated. This assumption is innocuous since, the probability of a unit having a score equal to the threshold is zero under reasonable assumptions.

More generally, let $U \in \mathcal{U}$ be a vector of unit characteristics available to the decision maker.[1] For now we do not impose any restriction on the set $\mathcal{U}$. In principle, characteristics may be represented

[1] We follow the convention of denoting random variables with uppercase letters, and the values they take with lowercase letters.

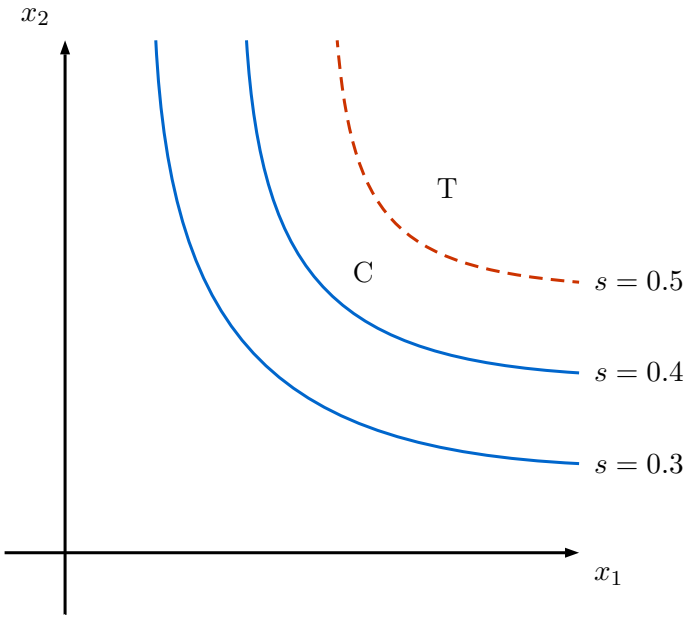

Figure 1: Decision maker preferences regarding treatment assignment over the set of unit characteristics.

by real numbers or discrete variables. Also, let $s : \mathcal{U} \to \mathbb{R}$ be the score used by the decision maker for treatment assignment. That is, a unit with characteristics u is treated if and only if $s(u) \geq \tau_s$.

We are interested in identifying the causal effect of treatment on an observed outcome Y. We follow the Rubin causal model and define two potential outcomes for each unit. Let $Y^{(0)}$ denote the outcome in the absence of treatment, and $Y^{(1)}$ the outcome if treated. Both are assumed to be well defined for each unit, regardless of whether the unit is treated or not, and their values do not depend on treatment assignment.

If the score, as well as the treatment status, are viewed as a random variables (for the econometrician), the observed outcome can be expressed as $Y = Y^{(1)}T + Y^{(0)}(1-T)$, where T is an indicator of treatment status. Similarly, the treatment effect on Y for units at the threshold can be defined by

$$\gamma = E[Y^{(1)}|S = \tau_s] - E[Y^{(0)}|S = \tau_s]$$

The question we address in this paper is how to estimate γ when the score is unobservable. The

solution we propose is a two stage estimator. We assume that there is a vector of unit characteristics $W \in \mathcal{W}$ available to the econometrician. In principle, W does not necessarily coincide with U, since the econometrician might not know or be able to observe all determinants of the true score. Still, the treatment status of all units is known. In the first stage, we propose to estimate (train) a statistical model (an ML algorithm) to predict treatment assignment. This estimation must be performed on a subset of all available units, the training sample. The outcome is a continuous function $p : \mathcal{W} \to \mathcal{I} \subseteq \mathbb{R}$ that assigns a predicted score $p(w)$ to any unit with characteristics w in an interval $\mathcal{I}$ on the real line. The predicted score must be defined in a way such that the statistical model predicts a unit with characteristics w belongs to the treatment group if and only if $p(w) \geq \tau_p$ for a given threshold $\tau_p \in \mathcal{I}$. For instance, if the statistical model is a logistic regression, then the model predicts $T(w) = 1$ if and only if $\frac{\exp(w'\hat{\beta})}{1+\exp(w'\hat{\beta})} \geq 0.5$, where $\hat{\beta}$ is a vector of parameters that minimizes a loss function on the training sample. The final outcome of the first stage is a predicted score p for all units in the testing sample (that does not overlaps with the training sample), and its corresponding predicted classification.

In the second stage, we propose to estimate γ on the testing sample under suitable assumptions as follows:

$$\gamma^{SynRD} = \lim_{q \downarrow \tau_p} E[Y|P = q, T = 1] - \lim_{q \uparrow \tau_p} E[Y|P = q, T = 0] \tag{1}$$

This is obviously an RD estimate (see, for example, Hahn et al., 2001). The crucial difference is that we use the predicted score P as the running variable instead of the unobservable S. Notice that we also condition on the observed treatment status T. Since the statistical model in the first stage might incorrectly predict the treatment status for some units, the misclassified units must be excluded from the testing sample in the second stage.

We now state an assumption that we will use further in section 4 to derive properties of the estimator in equation 1. Let $W = (X, D)$, where X takes values in $\mathcal{X} \subseteq \mathbb{R}^n$ and D in $\mathcal{D}$ for some finite set $\mathcal{D}$. X is a vector of n observable characteristics represented as real numbers, and D is a discrete or categorical characteristic taking only finitely many values. Since there is only a finite number of variables that

the econometrician could include in the model, we assume that there is only one discrete observable characteristic without loss of generality.

Assumption 3.1 *Existence of conditional joint density functions. For each $d \in \mathcal{D}$, the random variables $(Y^{(0)}, X)$ and $(Y^{(1)}, X)$ have each a continuous probability density function, conditional on $D = d$. Let f_d^0 and f_d^1 denote these two functions, respectively, for each each $d \in \mathcal{D}$.*

We introduce further assumptions in section 4, as well as a precise formal definition of the SynRD estimator and a derivation of some of its properties. In particular, we show that under perfect prediction γ^{SynRD} identifies the average treatment effect for the units right at the threshold of the true score used by the decision maker.

4 Perfect Prediction

In this section we show that, under suitable conditions, when the procedure in the first stage perfectly predicts treatment assignment, SynRD identifies the local treatment effect for a unit at the threshold of the unobservable score, i.e., the same parameter that a standard RD would identify when the score is known.

Treatment status is based on a continuous score function s that assigns a real number to every unit given a vector of characteristics $x \in \mathbb{R}^n$. We focus first on the case where all unit characteristics are real numbers. At the end of the section we discuss how to extend the main result to include the case where some characteristics are discrete (take only finitely many values).

A unit is treated if and only if its corresponding score $s(x)$ is greater or equal than a predefined threshold τ_s. Similarly, the statistical model in the first stage assigns a predicted score $p(x)$ to a unit with observable characteristics x, and predicts the unit belongs to the treatment group if and only if $p(x) \geq \tau_p$. Since this section studies identification with perfect prediction, we only consider here the case where the statistical model includes all the characteristics of the units used for treatment assignment.

We identify a unit with its vector of characteristics $x \in \mathcal{X}$, and state several conditions and results

in terms of subsets of $\mathcal{X}$ constructed as preimages of the functions p and s. Thus, for a given $A \subseteq \mathbb{R}$, let $s^{-1}(A) = \{x \in \mathcal{X} : s(x) \in A\}$ and $p^{-1}(A) = \{x \in \mathcal{X} : p(x) \in A\}$. We also use the convention $s^{-1}(a) = s^{-1}(\{a\})$ and $p^{-1}(a) = p^{-1}(\{a\})$ for the preimages of points.

In standard RD, the parameter of interest is defined in terms of expectations of the potential outcomes $Y^{(0)}$ and $Y^{(1)}$, conditional on the score S as it approaches the threshold. An underlying assumption is that the conditional expectations $E[Y^{(i)}|S = r]$ are continuous functions of r at the threshold for $i = 0, 1$. It is implicitly assumed that the score S is a random variable, and that the conditional expectations exist. Under this assumptions, the local treatment effect γ at the threshold τ_s is

$$\gamma = \lim_{r \downarrow \tau_s} E[Y^{(1)}|S = r] - \lim_{r \uparrow \tau_s} E[Y^{(0)}|S = r] = E[Y^{(1)}|S = \tau_s] - E[Y^{(0)}|S = \tau_s]$$

Our goal in this section is to prove that γ can also be computed as

$$\lim_{q \downarrow \tau_p} E[Y^{(1)}|P = q] - \lim_{q \uparrow \tau_p} E[Y^{(0)}|P = q]$$

However, in contrast to the traditional RD, we cannot simply assume the existence and continuity of $E[Y^{(i)}|P = q]$ for $i \in \{0, 1\}$. We must instead derive it from properties of the function p which might vary depending on the choice of the statistical method used to predict treatment assignment.

Notice that, in our context, the primitive random variable is the vector of characteristics describing a unit, $X \in \mathbb{R}^n$. Both the true score S and the predicted score P are defined as functions of X. Hence, when conditioning on values of these scores, we are actually conditioning on values of the random variable X. In particular, $E[Y^{(i)}|S = r] = E[Y^{(i)}|X \in s^{-1}(r)]$ and $E[Y^{(i)}|P = q] = E[Y^{(i)}|X \in p^{-1}(q)]$.

A difficulty arises because the sets $p^{-1}(q)$ may be of measure zero. Under assumption 4.3, $p^{-1}(q)$ contains no critical points for all $q \in (\underline{q}, \bar{q})$. Therefore $p^{-1}(q)$ is of Lebesgue measure zero. However, given the existence of a join density f^i for X and $Y^{(i)}$, we can define the conditional expectation in

terms of integrals of the join density over the manifold[2] $p^{-1}(q)$. That is, for all $q \in \left(\underline{q}, \bar{q}\right)$, let

$$E[Y^{(i)}|P=q] = E\left[Y^{(i)}|X \in p^{-1}(q)\right] = \frac{\int y(\int_{p^{-1}(q)} f^i(x,y)\,dx)dy}{\int(\int_{p^{-1}(q)} f^i(x,y)\,dx)dy} \tag{2}$$

We will use this definition to show that under the assumptions introduced in this chapter, the expectations $E[Y^{(i)}|P=q]$ are continuous at the threshold τ_p.

We now state a set of assumptions that are sufficient for identification of the local treatment effect.

Assumption 4.1 *Perfect prediction. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the open and non-empty space of characteristics used to assign and predict treatment status. For all $x \in \mathcal{X}$, $s(x) \geq \tau_s$ if and only if $p(x) \geq \tau_p$*

Assumption 4.2 *Continuity of s and p. The functions s and p are continuous in $\mathcal{X}$.*

Assumption 4.3 *Local smoothness of s and p. There are real numbers $\underline{r}$, $\bar{r}$, $\underline{q}$ and $\bar{q}$ such that:*

1. *$\underline{r} < \tau_s < \bar{r}$ and $\underline{q} < \tau_p < \bar{q}$.*
2. *The preimages $s^{-1}([\underline{r}, \bar{r}])$ and $p^{-1}([\underline{q}, \bar{q}])$ are compact.*
3. *When restricted to $s^{-1}((\underline{r}, \bar{r}))$, the score s is smooth and has no critical points, and the same holds for the predicted score p when restricted to $p^{-1}((\underline{q}, \bar{q}))$.*

Under these assumptions, the set of units right at the threshold of the true score coincides with those at the threshold of the predicted score. We state this result formally in lemma 4.4 below. This is a key step to prove identification, since the parameter that we want to recover is the local average treatment effect for units right at the threshold of the true score. It is worth noticing that perfect prediction as defined in assumption 4.1 is not enough. For instance, if the predicted score is constant and equal to the threshold in an open subset of $\mathcal{X}$ where s is always larger than τ_s, and the scores coincide elsewhere, then perfect prediction still holds but the level sets at τ_s and τ_p are obviously different.

[2]A manifold is a topological space that locally resembles the Euclidean space near each point. More precisely in our case, $M \subseteq \mathbb{R}$ is called a (smooth) manifold if it is locally diffeomorphic to $\mathbb{R}^n$ for some fixed $n \in \mathbb{N}$.

Lemma 4.4 *Under Assumptions 4.1, 4.2, and 4.3,* $s^{-1}(\tau_s) = p^{-1}(\tau_p)$.

Proof: Let $x \in s^{-1}(\tau_s)$, then $s(x) = \tau_s$, then $p(x) \geq \tau_p$. Suppose $p(x) > \tau_p$, then, since p is continuous, there is an $\epsilon > 0$ such that for all $x' \in B_\epsilon(x)$, $p(x') > \tau_p$. By assumption (no critical points), x is not a critical point of s. It follows that for some $x' \in B_\epsilon(x)$, $s(x') < \tau_s$, which contradicts perfect prediction. Then, $p(x) = \tau_p$. Therefore, $s^{-1}(\tau_s) \subseteq p^{-1}(\tau_p)$. A similar argument shows that $p^{-1}(\tau_p) \subseteq s^{-1}(\tau_s)$. ■

If lemma 4.4 did not hold, $s(x)$ might not converge to τ_s as $p(x)$ approaches τ_p. Hence, we need the level sets to be equal for identification. Moreover, given how we define conditional expectations, we also need:

$$\int_{p^{-1}(q)} f^i(x,y)dx \to \int_{p^{-1}(\tau_p)} f^i(x,y)dx$$

as q approaches τ_p. As illustrated in figure 2, for a well-behaved function p, the level sets $p^{-1}(\tau_p - t)$ should get arbitrarily close to $p^{-1}(\tau_p)$ as t goes to zero. Consequently, the integrals should converge to the desired limit. To prove this result formally, we apply methods from differential topology, specifically from Morse theory. Intuitively, we use a change of variables formula to rewrite the integral over the manifold $p^{-1}(\tau_p - t)$ as an integral over $p^{-1}(\tau_p)$, for all t close enough to zero. This allows us to transform sequences of the form $\{\int_{p^{-1}(q_t)} f^i(x,y)dx\}_t$ into sequences of the form $\{\int_{p^{-1}(q)} f^i_t(x,y)dx\}_t$. Finally, we show that the sequence $\{f^i_t\}$ converges to $f^{(i)}$ point-wise, and that the dominated convergence theorem holds once we express each integral over the compact manifold $p^{-1}(\tau_p)$ as the sum of integrals over open sets in $\mathbb{R}^{n-1}$.

Lemma 4.5 *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be continuous, positive and bounded from above. Under assumptions 4.2, and 4.3, the function* $H(t) = \int_{p^{-1}(\tau_p - t)} f(x)\,dx$ *is continuous at* $t = 0$.

Proof: We will prove first that H is right-continuous at $t = 0$.

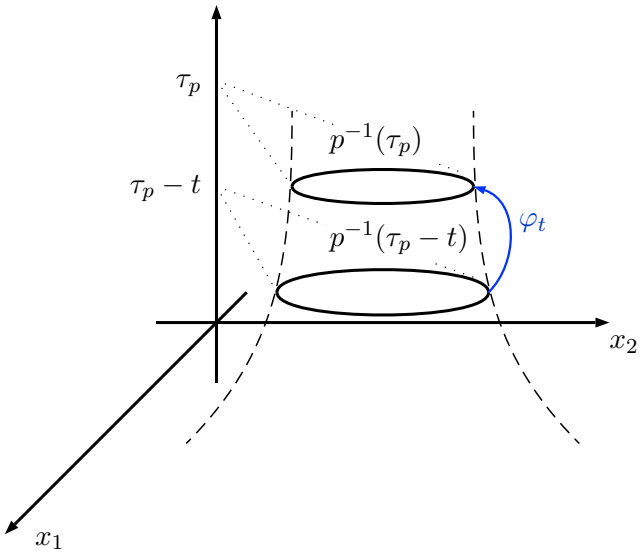

Figure 2: Schematic layout of the idea for the proof under perfect prediction, assumption 4.1.

Let M be the manifold defined by $M = p^{-1}\left(\left(\underline{q}, \bar{q}\right)\right)$, and $V : M \to \mathbb{R}^n$ be a smooth vector field with

$$V(x) = \frac{\nabla p(x)}{|\nabla p(x)|^2}$$

for all x in $p^{-1}\left(\left[\underline{q} + \epsilon, \tau_p\right]\right)$, for some $\epsilon > 0$ with $\underline{q} + \epsilon < \tau_p$, and zero outside a compact neighborhood of this set. V generates a unique 1-parameter group of diffeomorphisms of M onto itself (see lemma 2.4 in Milnor's Morse Theory). Let $\varphi : \mathbb{R} \times M \to M$ denote the group, and $\varphi_t : M \to M$ any of its elements. By construction, φ is C^∞ thus it is continuous as a function of t for all $x \in M$. Moreover, for any $x \in M$, if $\varphi_t(x) \in p^{-1}\left(\left[\underline{q} + \epsilon, \tau_p\right]\right)$, then $\frac{d\varphi_t(x)}{dt} = V(x)$, hence

$$\frac{dp(\varphi_t(x))}{dt} = \left\langle \frac{d\varphi_t(x)}{dt}, \nabla p(x) \right\rangle = \langle V(x), \nabla p(x) \rangle = 1$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbb{R}^n$. Then $p(\varphi_t(x))$ is linear with derivative 1, as a function of t, for any fixed $x \in M$, as long as $\varphi_t(x) \in p^{-1}\left(\left[\underline{q} + \epsilon, \tau_p\right]\right)$.

Let $\underline{q} + \epsilon \leq q < \tau_p$ and lets fix an x such that $p(x) = q$. Then $p(\varphi_t(x)) = t + k$, for some constant k, if $\varphi_t(x) \in p^{-1}\left(\left[\underline{q} + \epsilon, \tau_p\right]\right)$. Since φ_0 is the identity function on M, $k = p(\varphi_0(x)) = p(x) = q$. It

follows that, $p(\varphi_t(x)) = t + q$, for all $0 \leq t \leq \tau_p - q$. Therefore, for all $x \in p^{-1}(q)$, $p\left(\varphi_{\tau_p - q}(x)\right) = \tau_p$, thus $\varphi_{\tau_p - q}$ carries $p^{-1}(q)$ diffeomorphically onto $p^{-1}(\tau_p)$.

We can now apply the change of variables formula to obtain

$$\int_{p^{-1}(\tau_p - t)} f(x)\, dx = \int_{p^{-1}(\tau_p)} f(\varphi_t(x)) \left|\det\left(D\varphi_t(x)\right)\right| dx$$

where $D\varphi_t$ is the Jacobian of φ_t.

We need to show now that $\int_{p^{-1}(\tau_p)} f(\varphi_t(x)) \left|\det\left(D\varphi_t(x)\right)\right| dx$ is right-continuous at $t = 0$. Let $f_t(x) = f(\varphi_t(x)) \left|\det\left(D\varphi_t(x)\right)\right|$. f is continuous and for any given $x \in p^{-1}(\tau_p)$, $\varphi_t(x)$ is smooth as a function of t. Hence, for each $x \in p^{-1}(\tau_p)$, as t approaches 0 from above, $\varphi_t(x) \to \varphi_0(x) = x$ and $f_t(x) \to f(x)$. That is, $\lim_{t \downarrow 0} f_t(x) = f(x)$. Let $f_0(x) = f(x)$. We will show now that the integral and the limit can be switched.

Since $p^{-1}(\tau_p)$ is a compact manifold of dimension $n-1$, there is a finite partition of unity $\phi_1, ..., \phi_l$ subordinated to a covering of $p^{-1}(\tau_p)$ by coordinate patches $\{(U_\alpha, \varrho_\alpha)\}$ such that:

$$\int_{p^{-1}(\tau_p)} f_t(x)dx = \sum_{i=1}^{l} \left(\int_{U_{\alpha_i}} \phi_i(\varrho_{\alpha_i}(x)) f_t(\varrho_{\alpha_i}(x)) \left|\det(D\varrho_{\alpha_i}(x))\right| dx \right)$$

for $t \geq 0$, where $U_{\alpha_i} \subset \mathbb{R}^{n-1}$, for all $i = 1, .., l$.(see Munkres, 1991, chap. 5)

For all $t \geq 0$ close enough to 0, the integrand on the right hand side of the equation above is continuous as a function of x, positive, and uniformly bounded from above. The integral is taken with respect to the Lebesgue measure over a non-empty open subset of $\mathbb{R}^{n-1}$, thus the dominated convergence theorem implies that:

$$\lim_{t \downarrow 0} \int_{p^{-1}(\tau_p)} f_t(x)dx = \int_{p^{-1}(\tau_p)} \lim_{t \downarrow 0} f_t(x)dx = \int_{p^{-1}(\tau_p)} f(x)dx$$

It follows that $H(t) = \int_{p^{-1}(\tau_p - t)} f(x)\, dx$ is right-continuous at $t = 0$. In order to prove that H is also left-continuous at $t = 0$, the proof above needs to be modified slightly in order to build a 1-parameter

group of diffeomorphisms whose elements carry the level sets of the form $p^{-1}(\tau_p - t)$ diffeomorphically onto $p^{-1}(\tau_p)$, for t close enough to zero from below. The rest of the proof follows straightforwardly. ■

Theorem 4.6 *Under assumptions 4.1, 4.2, and 4.3,*

$$E[Y^{(1)}|S = \tau_s] - E[Y^{(0)}|S = \tau_s] = \lim_{q \downarrow \tau_p} E[Y^{(1)}|P = q] - \lim_{q \uparrow \tau_p} E[Y^{(0)}|P = q]$$

Proof: Lemma 4.5, and the dominated convergence theorem imply that $\lim_{q\downarrow\tau_p} E[Y^{(1)}|P = q] = E[Y^{(1)}|X \in p^{-1}(\tau_p)]$ and $\lim_{q\uparrow\tau_p} E[Y^{(0)}|P = q] = E[Y^{(0)}|X \in p^{-1}(\tau_p)]$. Lemma 4.4 implies that $E[Y^{(i)}|X \in p^{-1}(\tau_p)] = E[Y^{(i)}|X \in s^{-1}(\tau_s)] = E[Y^{(i)}|S = \tau_s]$, for $i = 1, 2$. ■

So far we have assumed that the space of characteristics is a subset of $\mathbb{R}^n$. However, some characteristics cannot be defined meaningfully without including finite sets in many potential applications of SynRD. Since the true score assigned to a unit is determined by finitely many characteristics, we can assume without loss of generality that there is only one discrete characteristic besides the real-valued ones. Correspondingly, let $\mathcal{D}$ be a non-empty finite set and D a discrete random variable taking values in $\mathcal{D}$. We extend the space of all characteristics to $\mathcal{X} \times \mathcal{D}$. We present conditions under which the result of theorem 4.6 holds when both the true and predicted scores also depend on D. Briefly, as long as the projection over $\mathcal{D}$ of the level sets of the functions $s : \mathcal{X} \times \mathcal{D} \to \mathbb{R}$ and $p : \mathcal{X} \times \mathcal{D} \to \mathbb{R}$ remain constant in open sets around τ_s and τ_p respectively, and lemma 4.5 holds for every fixed d in the projection, the result in theorem 4.6 also holds after performing one additional finite sum over the values of d in the projection. We state this idea formally below.

We first modify assumptions 4.1, 4.2 and 4.3 to allow for discrete variables. Correspondingly, we introduce some additional notation. For any $(x, d) \in \mathcal{X} \times \mathcal{D}$ and $A \subseteq \mathcal{X} \times \mathcal{D}$, let $\mathrm{pj_d}(x, d) = d$ and $\mathrm{pj_d}(A) = \{d \in \mathcal{D} : (x, d) \in A \text{ for some } x \in \mathcal{X}\}$. Also let the functions $s_d : \mathcal{X} \to \mathbb{R}$ and $p_d : \mathcal{X} \to \mathbb{R}$ be defined by $s_d(x) = s(x, d)$ and $p_d(x) = p(x, d)$ for each $d \in \mathcal{D}$.

Assumption 4.1′ *Perfect prediction. For all* $(x, d) \in \mathcal{X} \times \mathcal{D}$, $s(x, d) \geq \tau_s$ *if and only if* $p(x, d) \geq \tau_p$.

Assumption 4.2′ *Continuity of* s *and* p. *The functions* s_d *and* p_d *are continuous in* $\mathcal{X}$, *for each fixed* $d \in \mathcal{D}$.

Assumption 4.3′ *Local smoothness of* s_d *and* p_d. *There are real numbers* $\underline{r}$, $\bar{r}$, $\underline{q}$, *and* $\bar{q}$ *such that:*

1. $\underline{r} < \tau_s < \bar{r}$ *and* $\underline{q} < \tau_p < \bar{q}$.
2. *The sets* $s_d^{-1}([\underline{r}, \bar{r}])$ *and* $p_d^{-1}([\underline{q}, \bar{q}])$ *are compact in* $\mathbb{R}^n$.
3. *For each fixed* $d \in \mathrm{pj_d}(s^{-1}((\underline{r}, \bar{r})))$ *the score* s_d *is smooth and has no critical points when restricted to* $s_d^{-1}((\underline{r}, \bar{r}))$. *Similarly, for each fixed* $d \in \mathrm{pj_d}(p^{-1}((\underline{q}, \bar{q})))$ *the score* p_d *is smooth and has no critical points when restricted to* $p_d^{-1}((\underline{q}, \bar{q}))$.
4. *For any two* r *and* r' *in* $(\underline{r}, \bar{r})$, $\mathrm{pj_d}(s^{-1}(r)) = \mathrm{pj_d}(s^{-1}(r'))$. *Similarly, for any two* q *and* q' *in* $(\underline{r}, \bar{r})$ $\mathrm{pj_d}(p^{-1}(q)) = \mathrm{pj_d}(s^{-1}(q'))$.

For all $q \in (\underline{q}, \bar{q})$, we redefine the conditional expectations in equation 2 to allow for discrete characteristics as follows:

$$E\left[Y^{(i)} | (X, D) \in p^{-1}(q)\right] = \frac{\sum_{d \in \mathrm{pj_d}(p^{-1}(\tau_p))} \Pr[D = d] \Pr[X \in p_d^{-1}(q) | D = d] \left(\frac{\int y (\int_{p_d^{-1}(q)} f_d^i(x, y) dx) dy}{\int (\int_{p_d^{-1}(q)} f_d^i(x, y) dx) dy} \right)}{\sum_{d \in \mathrm{pj_d}(p^{-1}(\tau_p))} \Pr[D = d] \Pr[X \in p_d^{-1}(q) | D = d]}$$

where f_d^i is the joint density of $Y^{(i)}$ and X conditional on $D = d$. Notice that assumption 4.3′ allows us to sum over $\mathrm{pj_d}(p^{-1}(\tau_p))$ regardless of the value of q.

Under assumptions 4.1′, 4.2′, and 4.3′, lemma 4.5 implies that $\int_{p_d^{-1}(q)} f_d^i(x, y) dx$ converges to $\int_{p_d^{-1}(\tau_p)} f_d^i(x, y) dx$ as $q \to \tau_p$ for all $d \in \mathrm{pj_d}(p^{-1}(\tau_p))$. It can be shown similarly that $\Pr[X \in p_d^{-1}(q) | D = d]$ converges to $\Pr[X \in p_d^{-1}(\tau_p) | D = d]$. Therefore, the result of theorem 4.6 naturally extends to combinations of continuous and discrete variables.

5 Imperfect Prediction / Simulations

In this section, we extend the rigorous mathematical analysis to the case of imperfect prediction, i.e., without requiring assumption 4.1 (assumption 4.1′). For the moment we explore the case of imperfect prediction by means of simulations. Our overall findings indicate that the performance of our treatment effect estimator relies on the accuracy of the first stage estimation.

Monte Carlo Simulation

We generate data that fits our model and then apply SynRD to estimate the local average treatment effect. More precisely, we generate 20,000 observations (units), and use 10,000 of them to train a ML algorithm for treatment assignment prediction. We then predict treatment assignment and compute a synthetic score for the other 10,000 observations (testing sample). Finally, we estimate the treatment effect on the testing sample using RD with the synthetic score as the running variable. We repeat this process T times and report summary statistics of the corresponding estimates.

Each generated dataset contains 250 variables. 125 of them follow a multivariate standard normal distribution. We denote them $X = (X_1, ..., X_K)$, with $K = 125$. The other 125 are indicator variables, correlated with each other and the continuous variables, denoted by $D = (D_1, ..., D_K)$.

We assign a score $s(x, d)$ to each observation (x, d) following

$$\begin{aligned} w(x,d) &= \sum_{k=1}^{K} \alpha_k \log(x_k) + d_i'\delta + \sum_{k=1}^{K} \gamma_k d_k \log(x_k) \\ S(x,d) &= \frac{\exp(w(x,d) + \sigma\epsilon)}{1 + \exp(w(x,d) + \sigma\epsilon)} \end{aligned} \tag{3}$$

where α, δ, γ, and σ are constants, and ϵ is normally distributed and correlated with X and D. We use a capital letter for the score $S(x, d)$ to make explicit that it is a random variable since it also depends on ϵ. By definition, $0 < S(x, d) < 1$. Crucially, we do not include ϵ at any stage in the estimation. Hence, it allows us to assess the properties of SynRD when some of the determinants of the

true score are unobservable to the econometrician. In such case, the statistical model in the first stage does not achieve perfect prediction. The error scaling factor σ allows to control the relative weight of unobservables as determinants of the score.

Based on the score, we create κ categories and assign one of them to each observation. When $\kappa = 4$,

$$C(x,d) = \begin{cases} 1 & 0 < S(x,d) < 0.25 \\ 2 & 0.25 \leq S(x,d) < 0.5 \\ 3 & 0.5 \leq S(x,d) < 0.75 \\ 4 & 0.75 \leq S(x,d) < 1 \end{cases}$$

and similarly for other values of κ by partitioning the unit interval accordingly.

We also use the score to assign treatment, $T(x,d) = \mathbf{1}\left(S(x,d) \geq 0.5\right)$. When κ is even, our treatment group is comprised of the upper half categories.

The outcome of interest is

$$Y(x,d) = g(S(x,d)) + \eta(x,d)T(x,d) + \beta z + \xi \tag{4}$$

where $\eta(x,d)$ is the treatment effect for observation (x,d), z is unobservable and correlated with x and d, ξ is independent noise, and $g(\cdot)$ is a non-linear function.

We consider first a simple case with $g(s) = s^{\theta}$, and $\eta(x,d) = 2$ for all observations (homogeneous treatment effects). A randomly chosen simulated outcome with $\theta = 2$ is shown in figure 3.

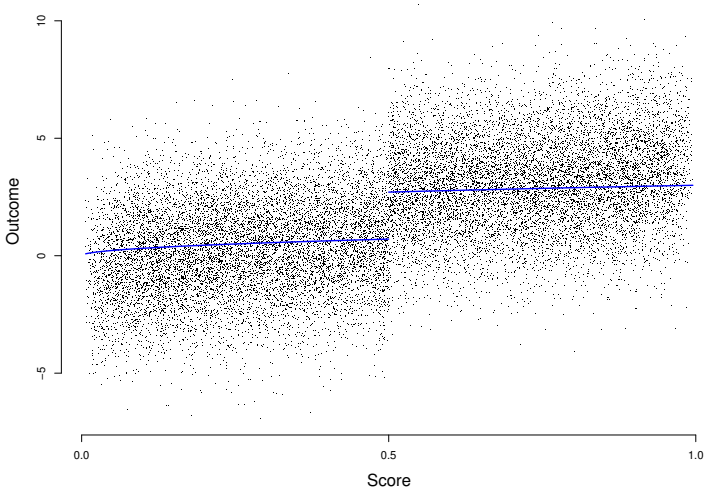

Figure 3: The simulated outcome as a function of the true score. The outcome is computed as $Y(x,d) = g(S(x,d)) + \eta(x,d)T(x,d) + \beta z + \xi$. ξ is independent mean zero noise. The line represents $E[Y(x,d)|S]$.

Machine Learning Algorithm

In the first stage we use a feed-forward neural network (multilayer perceptron) to predict the K-way classification, using as input the variables x and d, but crucially excluding ϵ. We train the algorithm on 10,000 observations, and test it on the remaining 10,000.

The neural network is comprised of four layers arranged sequentially. The input layer has a rectified linear unit (ReLU)[3] activation function and 64 units. The two hidden layers have the same activation and number of units as the input layer. The output layer uses a logistic activation function and only one unit.

The choice of this specific network architecture is somewhat arbitrary, except for the output layer. We use a logistic function with only one unit (output) because we want the algorithm to produce a single predicted score for each observation. Moreover, we encode the category of each observation as the midpoint of the corresponding interval in the score-axis. That is, if $\kappa = 4$ we encode all observations in category 1 as 0.125, all observations in category 2 as 0.375, and so forth. Hence, the encoding preserves

[3]ReLU is a standard activation function defined as $f(x) = \max(x, 0)$.

the order of the categories which is, by construction, a non-decreasing function of the simulated score. For training, we use mean square error as the loss function. We implement the neural network in Keras (Chollet et al., 2015).[4]

The neural network produces a synthetic score $p(x, d)$ between 0 and 1 for every observation (x, d) in the testing sample. We then obtain the predicted categories using the following criteria, for the case with $\kappa = 4$:

$$C^{pred}(x,d) = \begin{cases} 1 & 0 < p(x,d) < 0.25 \\ 2 & 0.25 \leq p(x,d) < 0.5 \\ 3 & 0.5 \leq p(x,d) < 0.75 \\ 4 & 0.75 \leq p(x,d) < 1 \end{cases}$$

and similarly for any other number of categories. It is worth noticing that the thresholds used for predicting classification depend on how categories are encoded and on the choice of a loss function, but not on the actual thresholds used to simulate categories. If the numbers coincide here is only because we want to put the true and synthetic scores on the same scale for comparison, but our method does not requires knowledge of the true thresholds.

Simulation Results

We explore the sensitivity of the estimation to changes in features of the data by conducting separate Monte Carlo simulations for different combinations of parameters in the data generating process (DGP). We run a simulation for each combination of parameters $\kappa \in \{2, 8, 20\}$, $\sigma \in \{0.5, 5, 50\}$, $\beta \in \{0.5, 2\}$, and $\theta \in \{0.5, 2\}$. All other parameters are kept constant. Crucially, the treatment effect η is 2 for all units, across all simulations.

For each combination of parameters, we randomly generate 1,000 datasets. Then, for each dataset, the neural network predicts treatment status and computes the synthetic score for all 10,000 observations

[4]We have also used other ML algorithms to predict classification, such as random forests (Breiman, 2001), but so far we have achieved the highest prediction accuracy with neural networks.

in the testing sample. In the second stage, we drop all misclassified observations, and estimate the treatment effect using RD with the synthetic score as the running variable.

We measure prediction accuracy as the fraction of units in the testing sample for which the algorithm in the first stage correctly predicts their treatment status. In the second stage, we obtain an estimate of the treatment effect, $\hat{\eta}$, and we define estimation bias as $E\left[|\hat{\eta} - \eta| \,|\Theta\right]$, where the mean is taken over all simulations with fixed values for some parameters of the DGP. Figure 4 depicts the relationship between prediction accuracy and estimation bias. The correlation between the two is -0.94. The shape of the marker indicate the number of classes. The color of the marker indicates the error scaling coefficient σ. The number of classes and the scaling factor are interdependent: a large error scaling factor is associated with lower prediction accuracy and larger estimation bias. The number of classes plays a more dominant role for smaller error scaling factors. Larger number of classes help overcome some of the decline in prediction accuracy. For low error scaling factors, the effect of the number of classes dominates the effect of the error scale σ on prediction accuracy and estimation bias.

Table 3 presents the results for 36 different combinations of parameters. As a benchmark, we include two columns with the estimated coefficient ($E(\eta_{\text{true}})$) and the corresponding standard error (SD(η_{true})) from a standard RD regression using the true (generated) score as the running variable. Besides the results on estimation bias already reported, 3 shows a loss of efficiency when using SynRD (SD(η_{sim})), due to additional variance from the first stage estimation of the synthetic score.

Figure 5 depicts the relationship between the true and predicted score. The scatter plots depict all (true-score, predicted-score)-pairs for the 10 simulations with prediction accuracy closest to the mean prediction accuracy. The top-left plot represents 2 classes with an error scale of $\sigma = 50$. The top-right plot represents 8 classes with an error scale of $\sigma = 50$. The bottom-left plot represents 8 classes with an error scale of $\sigma = 5$. The bottom-right plot represents 20 classes with an error scale of $\sigma = 0.5$. In the case of perfect prediction, we expect the point clouds of the upper-right quadrant and lower-left quadrant to connect precisely through the intersection of the dashed horizontal and vertical red lines. Given the noise levels in our simulations, this level of accuracy cannot be achieved. However, the fewer

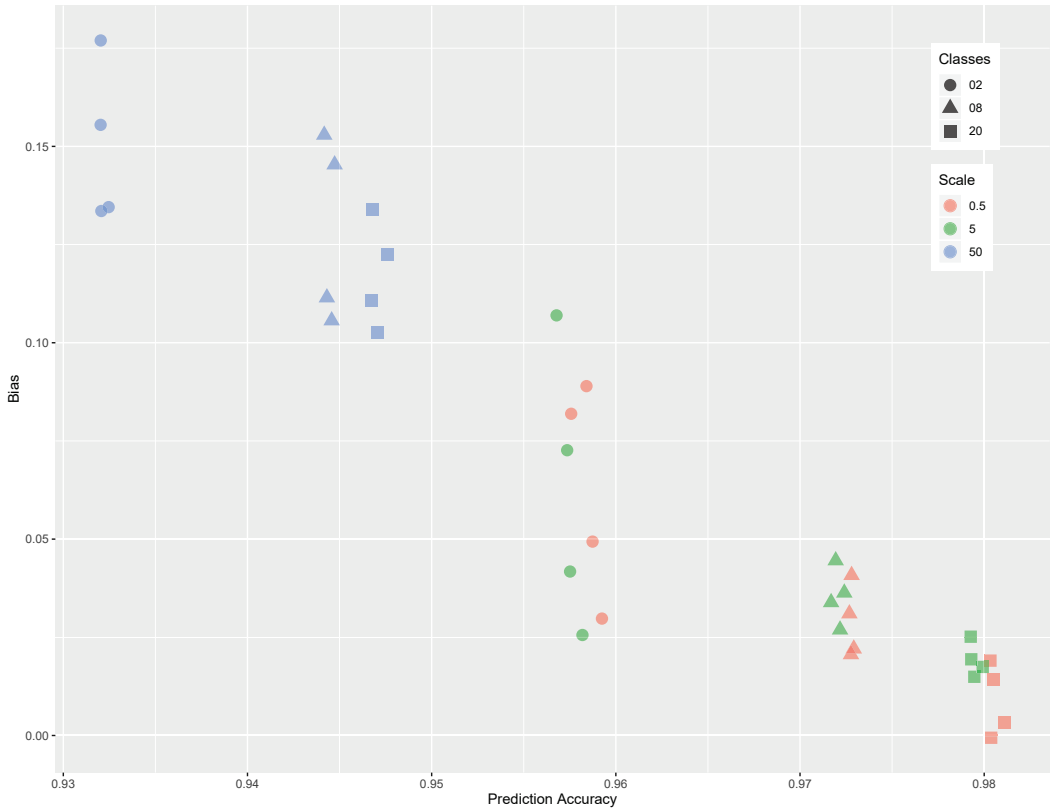

Figure 4: Simulation Results – Relationship between the prediction accuracy and the estimation bias. The correlation between the two is -0.94. The shape of the marker indicate the number of classes. The color of the marker indicates the error scaling coefficient σ. The number of classes and the scaling factor are interdependent: a large error scaling factor is associated with lower prediction accuracy and larger estimation bias. The number of classes plays a more dominant role for smaller error scaling factors. Larger number of classes help overcome some of the decline in prediction accuracy. For low error scaling factors, the effect of the number of classes dominate the effect of the error scale σ on prediction accuracy and estimation bias.

points lie in the top-left and bottom-right quadrants, the better the second stage RD results will be. Figure 5 confirms our intuition that a larger number of classes and lower noise level result in a more defined transition of the point cloud through the intersection of the red lines. Moreover, table 3 shows the mean effect sizes and prediction accuracies based on various parameter configurations.

6 Application: Price Effects of Investment Grade Credit Ratings

As an application of SynRD, we consider the effect of credit ratings on corporate bond prices. More precisely, we are interested in measuring the effect on its price of a bond being rated investment grade as

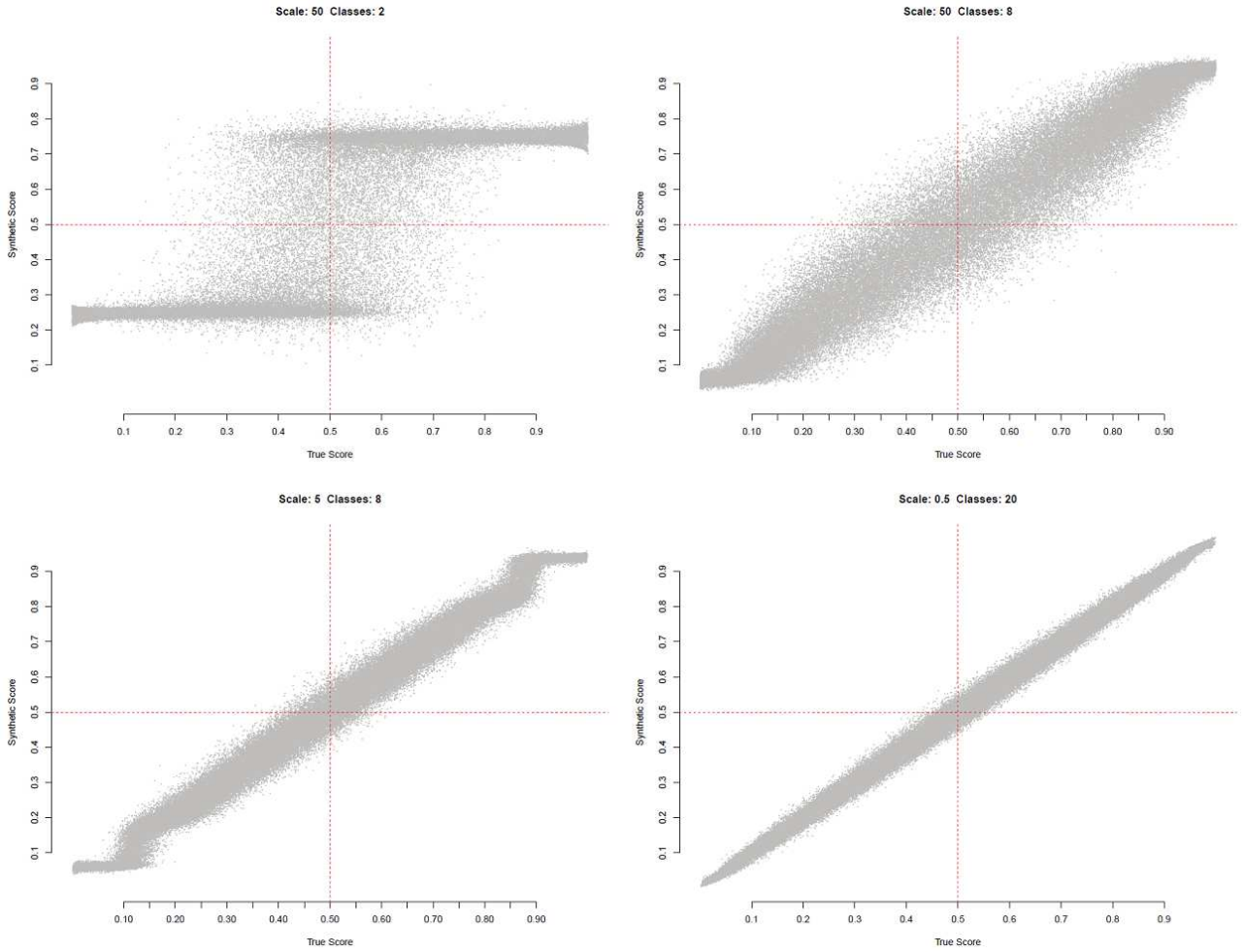

Figure 5: Score Comparison – Relationship between the true and predicted score. The scatter plots depict all (true-score,predicted-score)-pairs for the 10 simulations with prediction accuracy closest to the mean prediction accuracy. The top-left plot represents 2 classes with an error scale of $\sigma = 50$. The top-right plot represents 8 classes with an error scale of $\sigma = 50$. The bottom-left plot represents 8 classes with an error scale of $\sigma = 5$. The bottom-right plot represents 20 classes with an error scale of $\sigma = 0.5$. We can see that a larger number of classes and lower noise level result in a more defined transition of the point cloud through the intersection of the red lines.

opposed to non-investment grade by any of the three major credit rating agencies (Fitch, Moody's, and Standard & Poor's). In an ideal experiment, we would like to compare two bonds that are identical in all price-relevant characteristics, and that receive simultaneous but different credit ratings, for arbitrary reasons, with only one qualifying as investment grade. This would address the endogeneity arising from the fact that credit rating agencies use information that is very likely to directly affect prices and to be at least partially known by bond market participants. Obviously, such an experiment is not feasible. Instead, we propose to use SynRD to estimate the price effect of an investment grade rating for bonds

Num Classes	Scale (σ)	β	θ	$\mathbb{E}(\eta_{\text{true}})$	$\text{SD}(\eta_{\text{true}})$	$\mathbb{E}(\eta_{\text{sim}})$	$\text{SD}(\eta_{\text{sim}})$	Test Acc	Num Sims
2	0.5	0.5	0.5	2	0.05	2.06	0.44	0.96	914
2	0.5	0.5	2	2	0.05	2.08	0.48	0.96	914
2	0.5	2	0.5	2	0.14	2.04	1.27	0.96	913
2	0.5	2	2	2	0.14	2.08	1.2	0.96	913
2	5	0.5	0.5	2	0.05	2.04	0.65	0.96	913
2	5	0.5	2	2	0.05	2.07	0.5	0.96	913
2	5	2	0.5	1.99	0.14	2.11	1.31	0.96	913
2	5	2	2	2.01	0.14	2.03	1.29	0.96	913
2	50	0.5	0.5	2	0.05	2.14	0.3	0.93	913
2	50	0.5	2	2	0.06	2.18	0.34	0.93	913
2	50	2	0.5	2	0.17	2.15	0.92	0.93	913
2	50	2	2	1.99	0.16	2.16	0.89	0.93	913
8	0.5	0.5	0.5	2	0.05	2.02	0.11	0.97	913
8	0.5	0.5	2	2	0.05	2.04	0.11	0.97	913
8	0.5	2	0.5	2	0.15	2.02	0.33	0.97	913
8	0.5	2	2	2	0.14	2.03	0.33	0.97	913
8	5	0.5	0.5	2	0.05	2.03	0.11	0.97	913
8	5	0.5	2	2	0.05	2.04	0.11	0.97	913
8	5	2	0.5	2	0.15	2.04	0.33	0.97	913
8	5	2	2	2	0.15	2.03	0.32	0.97	913
8	50	0.5	0.5	2	0.06	2.11	0.11	0.94	913
8	50	0.5	2	2	0.06	2.14	0.11	0.94	913
8	50	2	0.5	2	0.16	2.11	0.32	0.94	913
8	50	2	2	2	0.16	2.15	0.33	0.94	913
20	0.5	0.5	0.5	2	0.05	2.02	0.11	0.98	913
20	0.5	0.5	2	2	0.05	2.01	0.1	0.98	913
20	0.5	2	0.5	2	0.15	2	0.32	0.98	913
20	0.5	2	2	2	0.14	2.01	0.29	0.98	913
20	5	0.5	0.5	2	0.05	2.02	0.1	0.98	913
20	5	0.5	2	2	0.05	2.02	0.1	0.98	913
20	5	2	0.5	2	0.14	2.02	0.3	0.98	913
20	5	2	2	2	0.14	2.03	0.28	0.98	913
20	50	0.5	0.5	2	0.06	2.1	0.11	0.95	913
20	50	0.5	2	2	0.06	2.14	0.11	0.95	913
20	50	2	0.5	2	0.17	2.11	0.32	0.95	913
20	50	2	2	2	0.15	2.13	0.33	0.95	913

Table 1: Simulation Results – Mean effect sizes and prediction accuracies based on various configurations.

at the threshold between investment and non-investment grade.

We assume the rating agencies decision making process can be described as follows. First, they analyze information on the bond and its issuer and synthesize it into a single continuous credit score. Then they compare this score to a set of predetermined ordered thresholds to decide the rating they assign to the bond. It is worth emphasizing that this assumption does not require rating agencies to explicitly compute such a score, only that they behave as if they do. We think of it as analogous to a utility maximization model where it is assumed that individuals choose from a set of alternatives according to the utility they derive from them, even if they do not explicitly compute a utility function.

Estimation of the treatment effect follows the method described in section 3. In the first step, we

use abundant information that was available to the rating agencies when they rated the bonds to train a machine learning algorithm to predict credit ratings. The training sample is comprised of tens of thousands of rating events. We treat every instance in which a bond receives a new rating by any of the three major credit rating agencies as a different rating event. The algorithm is tailored to produce a score for each rating event, such that all bonds with a score below a given threshold are classified by the algorithm as investment grade (and only those). Then, we compute this score for all rating events in the validation sample and keep only those that are correctly classified by the algorithm.

In the second stage, we estimate the local treatment effect on bond prices using a sharp RD. The outcome of interest is the change in the price of a bond right after a rating event. Crucially, the running variable is the score predicted in the first stage. We drop the misclassified observations. Hence, the sample contains only those rating events in the validation sample that are correctly classified as investment grade or non-investment grade using the predicted score.[5]

Intuitively, the higher the prediction accuracy on the validation sample in the first stage, the more likely is the algorithm to assign a score close to the threshold precisely to those rating events where the respective rating agency was indifferent or close to indifferent between an investment and a non-investment grade rating. As we prove in section 4, with perfect prediction SynRD identifies the treatment effect for those units right at the threshold of the true underlying score.

Data Description

We compile an extensive dataset from several sources covering rating events from January 2001 to June 2018. All information about rating events comes from Mergent Fixed Income Securities Database (FISD). This includes the issue (a bond identified with a 9 digit CUSIP), the issuer (6 digit CUSIP), the rating date when the event is made public, the bond rating, and the corresponding rating agency. FISD also provides bond characteristics such as maturity, coupon rate, seniority and covenants.

[5]We follow a key common practice in machine learning of not using the test sample until the predictive model is fully tuned. Hence, in this preliminary draft we only report the results of the second stage estimation on the validation sample. We have not used the test sample so far as we want to retain the possibility to adjust our first stage algorithm without knowledge of the final outcome.

We match a subset of issuers (only publicly traded corporations listed in the US and Canada) to Standard & Poor's Compustat by CUSIP, and collect data from their last four quarterly financial statements filed prior to the rating event.

RavenPack provides news analytics covering publicly traded companies in the US and Canada. It classifies news according to sentiment, relevance, topic, novelty, and market impact. RavenPack uses news from many sources for its analysis including the financial press, newswires, announcements by governments and regulators, and press releases. It computes an *Event Sentiment Score* (ESS) for each news item. We use the ESS to compute an average sentiment score for both old and novel news about the issuer. The average scores are computed for the last week, month, quarter, and two quarters prior to the rating event.

Average bond prices and yields are computed using daily data from FINRA's TRACE. At the predicting stage, we include average prices and yields for the two quarters prior to the rating event. We also obtain issuers' stock prices, shares outstanding, and earnings per share from CRSP, and we use monthly averages for the last 6 months prior to the rating event for prediction.

The dataset contains 145,850 observations (rating events) and 2,524 variables. Various indicators for categorical variables and for missing values contribute to the large number of variables. The data was split randomly (by sampling at the rating event level) into a training sample containing 46,837 observations, a validation sample containing 46,740 observations, and a test sample containing the remaining 52,273 observations. To reduce bond heterogeneity, we only include ratings for debentures (137,478) and medium term notes (8,372) issued by corporations trading in the US and Canada and denominated in the issuer's domestic currency. The distribution of credit ratings in our sample can be seen in figure 6. We use the same letter designations as S&P's and Fitch (AAA, AA+, AA, AA-, etc.), and apply a standard conversion table to Moody's designations. Every bond rated BBB- (Baa3) or above is considered investment grade.

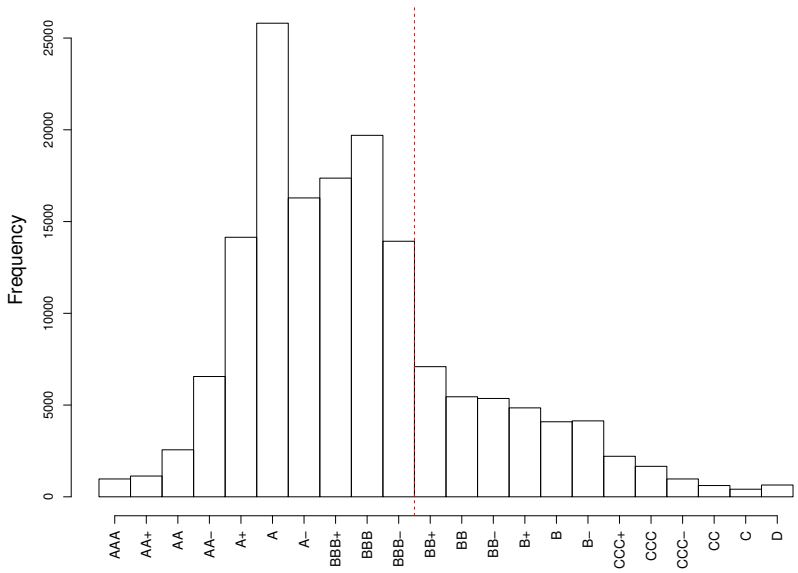

Figure 6: Distribution of credit ratings. We use the letter designations of S&P's and Fitch (AAA, AA+, AA, AA-, etc.). We apply a standard conversion table to Moody's designations. Every bond rated BBB- (Baa3) or above is considered investment grade.

Stage 1: Prediction

In the first stage, we train a machine learning algorithm on the training sample. The objective is to correctly predict ratings based on a set of bond and issuer characteristics, akin to the estimation of a statistical model for a discrete outcome. The model is subsequently validated by its prediction accuracy, defined as the fraction of correctly classified ratings. For each rating event in the validation sample we predict the rating and evaluate the out-of-sample performance of the model.

For the first stage predictions, we use an artificial neural network similar to that used in the simulations (see section 5). The choice of a specific network architecture is somewhat arbitrary and could be modified to increase prediction accuracy in the validation sample. Crucially, though, the output layer has a single unit and a logistic (sigmoid) activation function. That way, the algorithm assigns a single number from zero to one to each observation –its predicted score.

Table 2 presents an overview of the first stage prediction results for the validation sample, comparing

the true and predicted ratings. The overall prediction accuracy is 81.7%. However, for the estimation of the treatment effect, we are only interested in the investment vs non-investment grade classification, where the algorithm achieves a 97.9% prediction accuracy. Notice that when a bond is misclassified, the predicted rating tends to be close to the true one. For instance, out of a total 852 BBB- ratings that are incorrectly classified, 680 (79.8%) are predicted as BBB or BB+.

True Rating \ Predicted Rating	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-	B+	B	B-	CCC+	CCC	CCC-	CC	C	D
AAA	**307**	8	4	0	3	2	4	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0
AA+	4	**339**	17	5	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA	3	13	**764**	59	8	4	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA-	0	1	33	**1881**	87	25	8	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0
A+	3	3	7	93	**4211**	177	46	9	7	3	0	0	0	0	0	0	0	0	0	0	0	0
A	0	2	0	10	110	**7793**	334	71	16	2	1	0	0	0	0	0	0	0	0	0	0	0
A-	0	0	0	1	10	170	**4682**	288	58	13	1	0	0	0	0	0	1	0	0	0	0	0
BBB+	0	0	0	0	4	28	264	**4875**	416	44	5	0	2	2	1	0	0	0	0	0	0	0
BBB	0	0	0	0	0	11	67	397	**5307**	391	48	7	4	1	1	0	0	0	0	0	0	0
BBB-	0	0	0	0	1	3	12	83	489	**3655**	191	52	13	4	2	1	1	0	0	0	0	0
BB+	0	0	0	0	0	1	0	8	56	274	**1596**	179	64	22	3	3	0	0	0	0	0	0
BB	0	0	0	0	0	0	2	4	10	41	186	**1127**	212	80	22	7	0	0	0	0	0	0
BB-	0	0	0	0	0	0	2	1	6	25	56	231	**1047**	250	69	17	11	0	0	0	0	0
B+	0	0	0	0	0	0	1	2	9	6	12	64	247	**887**	231	60	7	1	0	0	0	4
B	0	0	0	0	0	0	0	1	2	8	7	29	100	245	**713**	160	23	4	3	0	1	0
B-	0	0	0	0	0	1	0	0	1	1	9	6	40	119	316	**714**	111	29	5	1	0	0
CCC+	0	0	0	0	0	0	0	0	0	2	1	5	7	23	81	166	**351**	62	21	4	1	3
CCC	0	0	0	1	0	0	0	0	0	0	2	1	6	15	24	54	115	**215**	53	24	4	2
CCC-	0	0	0	0	0	0	0	0	0	0	0	0	1	4	6	10	31	73	**168**	13	6	3
CC	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5	10	14	24	33	**84**	12	4
C	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	5	11	19	34	**59**	12
D	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	5	7	10	22	26	51	**93**

Table 2: First stage prediction accuracy. Investment grade bonds that have been incorrectly classified as non-investment grade, and vice versa, are shown in red.

Stage Two: Treatment Effect Estimation

The ML algorithm in the first stage produces an estimate of the continuous score underlying a bond's rating. By design, any bond with a predicted score below 0.5 is classified as invest grade (I), otherwise its predicted rating is non-investment grade (N). Following the method proposed in section 3, we use the predicted score as the running variable in an RD design with 0.5 as the I-N threshold (the same threshold used for prediction). Since we observe the rating of all bonds in the validation sample, we condition on treatment status as in equation 1. Hence, we drop the misclassified rating events which guarantees there is a discontinuity in the probability of treatment assignment at the threshold. The

fraction of dropped observations goes to zero as prediction accuracy increases.

Table 3 shows the local treatment effect estimates, using the bias-corrected RD estimator of Calonico et al. (2014) and Calonico et al. (forthcoming). The former propose a method to perform inference that is robust to the choice of bandwidth for the estimation of the local polynomials near the threshold. The latter allows for the inclusion of additional controls. The estimated coefficients represent the percentage change in quantity-weighted average bond prices one and four weeks after the rating event, compared to a period of the same duration before it. According to our estimates, an investment grade rating causes close to a 0.8% - 1.1% increase in price, on average, across a wide variety of corporate bonds in the U.S. and Canada. To address further concerns about bond heterogeneity, we also present the results for a restricted sample that excludes medium term notes, floating rate notes, convertibles, asset backed securities, exchangeable bonds, payment-in-kind bonds, originally issued discount issues, secured lease obligation issues, bonds with refund protection, and private placements. The estimated effects are slightly smaller on the restricted sample, when we focus on price changes within one week of the rating event, and almost identical across both samples when we extend the time window to four weeks. Adding controls for maturity, coupon rate and previous rating has virtually no effect on the size of the coefficients and their standard errors. Besides conventional standard errors for RD, we report robust standard errors following Calonico et al. (2014). It is worth noticing that in this preliminary draft reported standard errors do not account yet for variation in the first stage prediction outcomes.

Previous studies have found significant effects of rating changes in bond returns. As described in section 2, Hand et al. (1992) find excess returns for downgrades raging from a non-significant -0.37% to a significant -1,27%, depending on the sample considered. Similarly, using data on corporate bonds from TRACE (2002 - 2009), May (2010) report significant abnormal returns of -0.64% for downgrades not contaminated by concurrent news, when focusing only on prices of bonds that traded both on the day before and the day after the change in rating. An advantage of the approach we propose is that we can consider longer time windows around the rating event without having to worry about concurrent news about the issuer or other events that might affect prices directly. The application of SynRD allows for causal identification of the effect on prices of downgrading a bond from investment to non-investment

Change in price (%)	One week				Four weeks			
	Sample 1[a]		Sample 2[b]		Sample 1		Sample 2	
Treatment effect	0.011***	0.011***	0.008***	0.009***	0.008*	0.008*	0.008**	0.009**
Std. error[c]	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Std. error (robust)[d]	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)
Maturity		Yes		Yes		Yes		Yes
Coupon		Yes		Yes		Yes		Yes
Previous Rating		Yes		Yes		Yes		Yes
Observations	25368	24991	23088	22774	28562	28175	25711	25395

Notes: Treatment effects represent the change in average prices one and four weeks after the rating event compared to a period of the same duration before it. A coefficient of 0.01 reflects a 1% increase in price. The standard errors reported in this preliminary version do not account yet for variation in the first stage prediction outcome.

[a] Sample 1 is comprised of rating events for debentures and medium term notes issued by corporations trading in the US and Canada and denominated in the issuer's domestic currency.

[b] Sample 2 excludes medium term notes, floating rate notes, convertibles, asset backed securities, exchangeable bonds, payment-in-kind bonds, originally issued discount issues, secured lease obligation issues, bonds with refund protection, and private placements.

[c] Conventional RD standard errors.

[d] Standard errors that are robust to the bandwidth choice (see Calonico et al., 2014).

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 3: Effect of an investment grade rating vs a non-investment grade credit rating.

grade, keeping other price-relevant variables constant.

7 Conclusion

This paper introduces synthetic regression discontinuity design, a two-stage method for estimating the local treatment effect using machine learning when the score used by the decision maker is unobservable. We establish conditions under which the method identifies the local treatment effect for a unit at the threshold of the unobservable score, the same parameter that a standard RD design with known score would identify. We examine the properties of the estimator when the first stage prediction accuracy is lower than 100% using simulations. The method requires a high first stage prediction accuracy when predicting the observable treatment assignment. Finally, we apply SynRD to measure the effect of an investment grade rating on corporate bond prices by any of the three largest credit ratings agencies. We find a 0.8% - 1.1% increase in the price of bonds rated as investment grade as opposed to non-investment

grade, across a wide variety of corporate bonds.

References

Heitor Almeida, Igor Cunha, Miguel A Ferreira, and Felipe Restrepo. The real effects of credit ratings: The sovereign ceiling channel. *The Journal of Finance*, 72(1):249–290, 2017.

Erich Battistin, Agar Brugiavini, Enrico Rettore, and Guglielmo Weber. The retirement consumption puzzle: Evidence from a regression discontinuity approach. *American Economic Review*, 99(5):2209–26, December 2009.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

Sebastian Calonico, Matias D Cattaneo, and Max H Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522):767–779, 2018.

Sebastian Calonico, Matias D Cattaneo, Max H Farrell, and Rocio Titiunik. Regression discontinuity designs using covariates. *Review of Economics and Statistics*, forthcoming.

Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. A neural network approach to ordinal regression. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1279–1284. IEEE, 2008.

François Chollet et al. Keras. `https://keras.io`, 2015.

Laurent Davezies and Thomas Le Barbanchon. Regression discontinuity design with continuous measurement error in the running variable. *Journal of Econometrics*, 200(2):260 – 281, 2017. ISSN 0304-4076. Measurement Error Models.

Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.

Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

Petr Hájek. Municipal credit rating modelling by neural networks. *Decision Support Systems*, 51(1): 108–118, 2011.

John R. M. Hand, Robert W. Holthausen, and Richard W. Leftwich. The effect of bond rating agency announcements on bond and stock prices. *The Journal of Finance*, 47(2):733–752, 1992.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Zan Huang, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, 37(4):543–558, 2004.

Guido Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, 79(3):933–959, 2012.

Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635, 2008.

George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pages 312–318. IEEE, 2014a.

George Kour and Raid Saabne. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 417–422. IEEE, 2014b.

David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

Anthony D. May. The impact of bond rating changes on corporate bond prices: New evidence from the over-the-counter market. *Journal of Banking & Finance*, 34(11):2822 – 2836, 2010.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

James R Munkres. *Analysis on manifolds*. Advanced Book Classics. Addison-Wesley Publishing Company, Advanced Book Program, 1991. ISBN 9780201510355.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016.

Zhuan Pei and Yi Shen. The devil is in the tails: Regression discontinuity design with measurement error in the assignment variable. In *Regression Discontinuity Designs: Theory and Applications*, pages 455–502. Emerald Publishing Limited, 2017.

Jack Porter and Ping Yu. Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics*, 189(1):132 – 147, 2015. ISSN 0304-4076.

Warren S Sarle. Neural networks and statistical models, 1994.

Amir Sufi. The real effects of debt certification: Evidence from the introduction of bank loan ratings. *The Review of Financial Studies*, 22(4):1659–1691, 2007.

Tony T Tang. Information asymmetry and firms' credit market access: Evidence from moody's credit rating format refinement. *Journal of Financial Economics*, 93(2):325–351, 2009.

Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317, 1960.

Halbert White. *Artificial neural networks: approximation and learning theory*. Blackwell Publishers, Inc., 1992.

Ping Yu. Identification in regression discontinuity designs with measurement error. Manuscript, 2012.