More robust estimation of sample average treatment effects using Kernel Optimal Matching in an observational study of spine surgical interventions

Nathan Kallus
School of Operations Research and Information Engineering and
Cornell Tech, Cornell University, New York, New York 10044

Brenton Pennicooke
NewYork-Presbyterian Hospital and
Weill Cornell Medical Center, New York, New York 10032

Michele Santacatterina*
TRIPODS Center for Data Science for Improved Decision Making
and Cornell Tech, Cornell University, New York, New York, 10044

April 2019

*Corresponding author. This material is based upon work supported by the National Science Foundation under Grants Nos. 1656996 and 1740822.

Abstract

Inverse probability of treatment weighting (IPTW), which has been used to estimate sample average treatment effects (SATE) using observational data, tenuously relies on the positivity assumption and the correct specification of the treatment assignment model, both of which are problematic assumptions in many observational studies. Various methods have been proposed to overcome these challenges, including truncation, covariate-balancing propensity scores, and stable balancing weights. Motivated by an observational study in spine surgery, in which positivity is violated and the true treatment assignment model is unknown, we present the use of optimal balancing by Kernel Optimal Matching (KOM) to estimate SATE. By uniformly controlling the conditional mean squared error of a weighted estimator over a class of models, KOM simultaneously mitigates issues of possible misspecification of the treatment assignment model and is able to handle practical violations of the positivity assumption, as shown in our simulation study. Using data from a clinical registry, we apply KOM to compare two spine surgical interventions and demonstrate how the result matches the conclusions of clinical trials that IPTW estimates spuriously refute.

Keywords: SATE, positivity assumption, model misspecification, kernel optimal matching, causal inference, non-experimental studies.

1 Introduction

Inverse probability of treatment weighting (IPTW) has been used to estimate the sample average treatment effect (SATE) of a treatment on an outcome using observational data. The key idea of IPTW is to correct for selection bias into treatment by weighting each unit in the sample by its probability of being in its treatment group conditional on covariates, *i.e.*, the propensity score (Rosenbaum and Rubin, 1983). In other words, IPTW creates a pseudo-population in which each unit has the same probability of getting treated, thus mimicking a randomized experiment. IPTW's popular use in medicine (Mansournia and Altman, 2016), epidemiology (Hernán et al., 2000), and lately also in computer science (Swaminathan and Joachims, 2015; Kallus and Zhou, 2018; Su et al., 2018) come from its theoretical appeal and interpretability. The standard way to estimate SATE via IPTW consists of predicting the propensity scores by modeling the treatment assignment mechanism, taking their inverse, and plugging the obtained set of weights into a weighted average or a weighted least squares (WLS) estimator (Horvitz and Thompson, 1952; Robins et al., 1994, 2000; Lunceford and Davidian, 2004). Wald confidence intervals are then constructed using a robust (sandwich) estimator for the standard error (Van der Vaart, 2000; Stefanski and Boos, 2002; Freedman, 2006; Tsiatis, 2007).

Positivity, which requires that for any set of covariates it is theoretically possible to observe a unit with either treatment, is key to estimating SATE (without parametric outcome models). However, IPTW's reliance on positivity can be very tenuous. In particular, if positivity is violated in even a very limited region of covariates, then the IPTW estimator for SATE has *infinite* variance (Robins et al., 2000; Cole and Hernán, 2008). Even if positivity holds theoretically, if some propensities are close to 0, then even small errors in propensity estimates can lead to outsize errors in IPTW's SATE estimate. This issue

is known as *practical violations of the positivity assumption* (Petersen et al., 2012) and it is well known that it can lead to extreme weights and large variance (Robins et al., 1995; Scharfstein et al., 1999; Robins et al., 2007; Kang and Schafer, 2007), which pose serious problems in practice.

IPTW also relies on the correct specification of the unknown treatment assignment model – a concern in almost every observational study.

One example of practical positivity violation and possible model misspecification that we study in this paper is in the evaluation of laminectomy alone compared with fusion-plus-laminectomy in patients with lumbar stenosis and lumbar spondylolisthesis. The comparison is based purely on passive observations of historical spine surgical interventions and their outcomes, as recorded in a clinical registry of spine surgeries. Lumbar stenosis is a spine pathology consisting of a compression of the lower back's nerves. Lumbar spondylolisthesis is a pathology in which one vertebra move out of position. Common spine surgical practice suggests treating patients with lumbar stenosis with laminectomy alone, while those with lumbar spondylolisthesis with fusion-plus-laminectomy. While deviations exist, this leads to very limited positivity in the data. Understanding the differing benefits of these treatments is of utmost interest because of the invasive nature of the surgeries. Registry data provide a unique opportunity to use a large number of observations to study these effects, but very limited positivity and potential misspecification remain a significant hurdle to the use of standard methodologies.

Several statistical methods have been proposed to overcome issues of practical positivity violation and potential misspecification. To control for practical positivity violation, the most popular solution is truncation, which consists of replacing outlying weights with less extreme ones (Cole and Hernán, 2008). Kang et al. (2016) and Lee et al. (2011) investigated the impact of different cutoff points in the distribution of the propensity scores

with respect to bias and efficiency. Cole and Hernán (2008) suggested truncating at high percentiles of the distribution of the estimated weights, *e.g.*, the 1st and 99th percentiles. Ju et al. (2017) proposed an adaptive truncation method based on the collaborative targeted maximum likelihood estimation methodology. Despite the fact that truncation reduces the variance of the weights and consequently that of the weighted estimator, it can also introduce substantial bias. Rather than truncating, Santacatterina and Bottai (2018) and Santacatterina et al. (2018) proposed to find the closest set of weights to the IPTW weights while controlling precision by constraining the variance of the resulting estimator or the variance of the weights.

To mitigate the effect of possible misspecification of the treatment assignment model, Imai and Ratkovic (2014) proposed to use the Covariate Balancing Propensity Score (CBPS), which, instead of plugging in logistic-regression estimates of propensities, uses IPTW with propensities predicted by the logistic model that balances covariates, found via the generalized method of moments. Lee et al. (2010) proposed to use boosted classification and regression trees to estimate the propensities. Zubizarreta (2015) proposed Stable Balancing Weights (SBW), which are the set of weights of minimal sample variance that satisfy a list of approximate moment matching conditions to a level of balance specified by the research.

In this paper, we use Kernel Optimal Matching (KOM), a subclass of the Generalized Optimal Matching (GOM) framework (Kallus, 2016), to provide weights that simultaneously mitigate the effects of possible misspecification of the treatment assignment model and control for possible practical positivity violations. We do so by minimizing the worst-case conditional mean squared error of the weighted estimator in estimating SATE over the space of weights. Specifically, KOM controls for practical positivity violations by limiting the variance of the estimate (either by penalizing or constraining it), while mitigating possible model misspecification by using flexible models to balance covariates. To use KOM,

we show how to extend the general approach of Kallus (2016), which focused only on SATT for simplicity, to the case of SATE, which requires a new, more intricate error decomposition and an approach that balances both the conditional average of the control and the treatment outcomes. Compared with the state-of-the-art methods, we find that estimating SATE with KOM has the advantages of (1) optimally balancing covariates while simultaneously controlling for precision, (2) mitigating the effect of possible misspecification of the treatment assignment model, (3) controlling for strong practical positivity violations, (4) tractably allowing for nonlinear and nonadditive covariates relationships by using kernels, (5) better handling of many covariates and higher order relationships, and (6) automatic selection of balancing levels. In particular, in the simulation study presented in Section 5, we show that both bias and mean squared error of the KOM estimates of SATE are lower than those obtained by using IPTW, truncated IPTW, Propensity Score Matching (PSM), Regression Adjustment (RA), CBPS, and SBW in most of the considered scenarios (we provide a detailed file containing the R code to compute KOM as supporting material). Motivated by this, we use KOM to address the problem of estimating the effect of spine surgical interventions using clinical registry data and find that, whereas both an unadjusted comparison and IPTW show a large significant effect, our estimates show a small insignificant effect, which actually matches the results of recent clinical trials (Ghogawala et al., 2016; Försth et al., 2016).

In the next Section, we introduce a study on the effect of two spine surgical interventions among patients with lumbar stenosis or lumbar spondylolisthesis that motivated the use of KOM. In Section 3 we introduce KOM for SATE and discuss practical guidelines (Section 4). In Section 5, we present the results of a simulation study aimed at comparing KOM with IPTW, truncated IPTW, PSM, RA, CBPS, and SBW. In Section 6 we apply KOM to estimate the effect of laminectomy alone versus fusion-plus-laminectomy on the Oswestry

Disability Index (ODI) among patients with lumbar stenosis or lumbar spondylolisthesis. We conclude with some remarks in Section 7.

2 The effect of two spine surgical interventions among patients with lumbar stenosis or lumbar spondylolisthesis

Lumbar stenosis is a pathology caused by the narrowing of the central spinal canal by overgrown and inflamed connective tissue (Resnick et al., 2014). Lumbar spondylolisthesis is a pathology caused by the slippage of one vertebra on another. These spinal pathologies can severely restrict function, walking ability, and quality of life (Waterman et al., 2012). If the symptoms due to lumbar stenosis or lumbar spondylolisthesis are no longer controlled by medications, physical therapy, or spinal injections, then surgery may be needed to improve a patient's symptoms (Waterman et al., 2012). Typically, a laminectomy alone is done to treat lumbar stenosis and a fusion-plus-laminectomy is done to treat lumbar spondylolistheses (Resnick et al., 2014; Eck et al., 2014; Raad et al., 2018). In addition, patients with leg pain are typically treated with a laminectomy alone, while patients with mechanical back pain are treated with fusion-plus-laminectomy (Resnick et al., 2014). Though there is some variation and both interventions may be used for both pathologies, the prevalence of this surgical practice leads to a practical positivity violation when evaluating the effect of laminectomy alone versus fusion-plus-laminectomy in observational data. In particular, in the case study presented in Section 6, in which we compare these two spine surgical interventions, less than 10% of patients with lumbar spondylolisthesis were treated with laminectomy alone and only 1% of those with a moderate-low leg pain were treated with

fusion-plus-laminectomy.

Due to practical and methodological challenges, randomized trials on the effect of spine surgical interventions are rare (Cook, 2009). Consequently, most assessments of spine surgical interventions must be based on observational data, in which the true treatment assignment mechanism is hardly ever known and the true causal parameter is hidden by confouding factors. The patient's principal spine pathology, *i.e.*, lumbar stenosis or lumbar spondylolisthesis, is one example of such a confounding factor in this case. Patients with lumbar stenosis, who are mainly treated with laminectomy alone, are also more likely to have a lower ODI overall (Pearson et al., 2011). Given these challenges, it is therefore of paramount importance to develop and use statistical methods that provide robust estimates of the SATE for spine surgical interventions based on observational data.

3 Kernel Optimal Matching

In this Section we propose to use KOM for estimating SATE to address the issues noted above. KOM is an approach that minimizes an estimation error objective when unknown conditional expectations are let to vary in a Reproducing Kernel Hilbert Spaces (RKHS) (Kallus, 2016). To extend this approach to SATE estimation: we analyze the conditional mean squared error (CMSE) of any weighted estimator for SATE; show that the CMSE can be decomposed in terms of the discrepancies in the conditional expectations of the two potential outcomes as well as a variance term and some additional ignorable terms; embed these conditional expectations in an RKHS to obtain an error objective that can be evaluated given observational data; and finally minimize this objective using quadratic optimization to find optimal weights. We discuss how to automatically tune the method in order to appropriately set the level of balance, the exchange between balance and variance,

and the kernel parameters.

3.1 Decomposing the CMSE for SATE

Suppose we have a simple random sample with replacement of size n from a population. Using the potential outcome framework (Imbens and Rubin, 2015), for each unit $i = 1, \ldots, n$, we let $Y_i(t)$ be the potential outcome of treatment $t \in \{0, 1\}$, X_i the observed confounders, T_i the observed treatment, and $Y_i = Y_i(T_i)$ the observed outcome. Let $X_{1:n}$ and $T_{1:n}$ denote all the observed confounders and treatment assignments. We impose the assumptions of consistency, non-interference, and ignorability (Imbens and Rubin, 2015). The assumptions of consistency and non-interference (also known as the SUTVA assumption) state that the observed outcome corresponds to the potential outcome under the treatment applied to that specific unit, *i.e.*, $Y_i = Y_i(t)$, and that the potential outcomes are well-defined. The assumption of ignorability states that the potential outcomes are independent to the treatment assignment once we condition on observed covariates. In other words, ignorability states that we have collected all potential confounders in our covariates. It suffices to impose the independence in expectation, *i.e.*, we assume only that $\mathbb{E}[Y_i(t) \mid X_i, T_i] = \mathbb{E}[Y_i(t) \mid X_i]$ for $t = 0, 1$.

We consider estimating the SATE, defined as

$$\text{SATE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i(1) - Y_i(0)), \tag{3.1}$$

by using the weighted estimator

$$\hat{\tau}_W^{\text{SATE}} = \sum_{i:T_i=1} W_i Y_i - \sum_{i:T_i=0} W_i Y_i = \sum_{i=1}^{n} W_i (2T_i - 1) Y_i, \tag{3.2}$$

which compares the reweighted average outcome among the treated and control group. Given any weights $W_{1:n}$, setting $W_i' = W_i / \sum_{j:T_j=T_i} W_i$, we have that $\hat{\tau}_{W'}^{\text{SATE}}$ is equivalent to

the WLS estimator with weights $W_{1:n}$. In particular, if $\sum_{i:T_i=1} W_i = \sum_{i:T_i=0} W_i = 1$ then $W'_{1:n} = W_{1:n}$ and $\hat{\tau}^{\text{SATE}}_W$ is already the WLS estimator.

If we were to let $W_i = T_i/e(X_i) + (1-T_i)/(1-e(X_i))$, where $e(X_i) = \mathbb{P}(T_i = 1 \mid X_i)$ is the propensity score, then $\hat{\tau}^{\text{SATE}}_{\text{W}}$ reduces to the well-known IPTW estimator (Horvitz and Thompson, 1952; Robins et al., 1994; Lunceford and Davidian, 2004). Similarly, if we normalize these weights to sum to one in each treatment group, then $\hat{\tau}^{\text{SATE}}_{\text{W}}$ reduces to the WLS estimator with IPTW weights. Instead of taking this plug-in approach, we will find the weights $W_{1:n}$ that optimize an error objective given by the CMSE.

We now decompose the error of the weighted estimator $\hat{\tau}^{\text{SATE}}_{\text{W}}$ for *any* weights $W_{1:n}$ that are a function of the covariate and treatment data, $X_{1:n}, T_{1:n}$, i.e. $W_i = W(X_{1:n}, T_{1:n})$. We start by defining $f_t(X_i) = \mathbb{E}[Y_i(t) \mid X_i]$ and $\sigma_i^2 = \text{Var}\,(Y_i \mid X_i, T_i)$. Next define the conditional average of SATE (CSATE):

$$\text{CSATE} = \mathbb{E}[\text{SATE} \mid X_{1:n}] = \frac{1}{n}\sum_{i=1}^{n}(f_1(X_i) - f_0(X_i)).$$

In our decomposition, we will separate out the error of the weighted estimator in estimating just CSATE, which is what we will actually focus on.

For any function f, we define the f-moment discrepancy between the weighted t-treated group and the whole sample as

$$B_t(W_{1:n}; f) = \sum_{i=1}^{n}\left(\mathbb{I}\left[T_i = t\right] W_i - \frac{1}{n}\right) f(X_i), \tag{3.3}$$

where $\mathbb{I}[T_i = t] \in \{0, 1\}$ is the indicator for unit i having treatment t. We now decompose the conditional bias and CMSE of $\hat{\tau}^{\text{SATE}}_{\text{W}}$.

Theorem 3.1. *Suppose $W_{1:n}$ is independent of all else given $X_{1:n}, T_{1:n}$. Then, under con-*

sistency, non-interference, and ignorability,

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}_W^{\mathrm{SATE}} - \mathrm{SATE} \mid X_{1:n}, T_{1:n}\right] &= \mathbb{E}\left[\hat{\tau}_W^{\mathrm{SATE}} - \mathrm{CSATE} \mid X_{1:n}, T_{1:n}\right] \qquad (3.4)\\
&= B_1(W_{1:n}; f_1) - B_0(W_{1:n}; f_0)\\
\mathbb{E}\left[\left(\hat{\tau}_W^{\mathrm{SATE}} - \mathrm{CSATE}\right)^2 \mid X_{1:n}, T_{1:n}\right] &= (B_1(W_{1:n}; f_1) - B_0(W_{1:n}; f_0))^2 + \sum_{i=1}^n W_i^2 \sigma_i^2 \qquad (3.5)\\
\mathbb{E}\left[\left(\hat{\tau}_W^{\mathrm{SATE}} - \mathrm{SATE}\right)^2 \mid X_{1:n}, T_{1:n}\right] &= (B_1(W_{1:n}; f_1) - B_0(W_{1:n}; f_0))^2 + \sum_{i=1}^n W_i^2 \sigma_i^2 \qquad (3.6)\\
&\quad + \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(Y_i(1) - Y_i(0) \mid X_i)\\
&\quad + \frac{1}{n}\sum_{i=1}^n W_i(2T_i - 1)\,\mathrm{Cov}(Y_i, Y_i(1) - Y_i(0) \mid X_i, T_i).
\end{aligned}$$

Theorem 3.1 shows that the bias of $\hat{\tau}_W^{\mathrm{SATE}}$ decomposes into two discrepancies: the discrepancy in the f_1 moment between the weighted treated group and the whole sample and the discrepancy in the f_0 moment between the weighted control group and the whole sample (eq. (3.4)). Next, Theorem 3.1 shows that the CMSE of $\hat{\tau}_W^{\mathrm{SATE}}$ in estimating CSATE decomposes into a conditional bias squared plus a conditional variance, where the conditional variance is simply given by the weighted squared Euclidean norm of the W vector, with components weighted appropriately by the conditional variance of the outcomes (eq. (3.5)). This allows us to understand precisely where errors due to the choice of $W_{1:n}$ arise from and help us in judicially choosing $W_{1:n}$ to minimize total error. In particular, we will next discuss an approach to minimize this total error, given some restrictions on the unknown f_0, f_1.

Theorem 3.1 also shows that the CMSE of $\hat{\tau}_W^{\mathrm{SATE}}$ in estimating SATE differs from that of estimating CSATE by two certain terms. We next argue it is safe to ignore these terms when using this CMSE objective to choose $W_{1:n}$. One additional term (the second

on the right-hand side of eq. (3.6)) corresponds to the variance of SATE in estimating CSATE (or, vice versa). In particular, this term is both small and *independent* of $W_{1:n}$, so it should not affect how we choose $W_{1:n}$ and we may ignore it. Another additional term (the third on the right-hand side of eq. (3.6)) involves both the weights W_i and the covariance of the observed outcome (Y_i) and the individual effect ($Y_i(1) - Y_i(0)$). Although this term does involve the weights, it is always small for any set of weights. In particular, if conditional variances are bounded such that $\mathrm{Var}(Y_i(t) \mid X_i) \leq \sigma^2_{\max}$ (as would be the case under homoskedasticity, for example) and if we focus our attention to weights that sum to one in each treatment group (as we do in this paper) then, applying the Cauchy-Schwarz inequality to the covariance and Hölder's inequality to the sum, we see that this term is bounded by $4\sigma^2_{\max}/n$ regardless of the choice of $W_{1:n}$. Otherwise, using the Cauchy-Schwarz inequality to bound the unknowable conditional covariance of the two potential outcomes by their respective conditional variances (see also Splawa-Neyman et al., 1990; Imai, 2008), we simply get an additional term that we could easily also take into consideration if we so choose.

3.2 Worst-case squared bias

The bias of the weighted estimator, and correspondingly its CMSE, depends on the unknown functions f_1, f_0. In this Section, we propose to minimize the worst-case CMSE and correspondingly replace the bias by its worst-case value, normalized relative to the "size" of f_1, f_0 since the bias scales linearly in these functions.

To define this "size," we embed each function in a normed space. In particular, we consider an extended seminorm $\|\cdot\|_t$, i.e., a norm on functions from the space of covariates to the space of outcomes that can also assign the values 0 and ∞ to nonzero elements. We then define the "size" of the pair f_1, f_0 as $\sqrt{\|f_1\|_1^2 + \|f_0\|_0^2}$ (i.e., we take the product of the

spaces). Given this magnitude (we discuss our specific choice below), we can define the relative worst-case squared bias as follows:

$$\mathcal{B}(W_{1:n}) = \sup_{f_0, f_1} \frac{B_1(W_{1:n}; f_1) - B_0(W_{1:n}; f_0)}{\sqrt{\|f_1\|_1^2 + \|f_0\|_0^2}} = \sqrt{\Delta_1^2(W_{1:n}) + \Delta_0^2(W_{1:n})}, \tag{3.7}$$

where

$$\Delta_t(W_{1:n}) = \sup_f \frac{B_t(W_{1:n}; f)}{\|f\|_t} = \sup_{\|f\|_t \leq 1} B_t(W_{1:n}; f)$$

is the relative worst-case discrepancy in the f moment between the weighted t-treated group and the whole sample over all f functions in the unit ball of $\|\cdot\|_t$.

In particular, we will use the norm given by an RKHS. Given a positive semidefinite (PSD) kernel $\mathcal{K}_t(x, x')$, these norms take the form

$$\|f\|_t = \inf \left\{ \sum_{i,j=1}^{\infty} \alpha_i \alpha_j \mathcal{K}_t(x_i, x_j) : f = \sum_{i=1}^{\infty} \alpha_i \mathcal{K}_t(x_i, \cdot),\ \sum_{i=1}^{\infty} \alpha_i^2 \mathcal{K}_t(x_i, x_i) < \infty \right\}.$$

Despite this complex form of the norm, the corresponding form for $\Delta_t(W)$ is rather simple. Define the matrix $K_t \in \mathbb{R}^{n \times n}$ as $K_{tij} = \mathcal{K}_t(X_i, X_j)$ (that such a matrix is PSD for any set of points is precisely the definition of a PSD kernel). Then, we have that

$$\begin{aligned}
\Delta_t^2(W_{1:n}) &= \sup_{\|f\|_t \leq 1} \left(\sum_{i=1}^{n} \left(W_i \mathbb{1}[T_i = t] - \frac{1}{n} \right) f_t(X_i) \right)^2 \\
&= \sup_{\sum_{i,j=1}^n \alpha_i \alpha_j \mathcal{K}_t(X_i, X_j) \leq 1} \left(\sum_{i=1}^{n} \left(W_i \mathbb{1}[T_i = t] - \frac{1}{n} \right) \sum_{j=1}^{n} \alpha_j \mathcal{K}_t(X_i, X_j) \right)^2 \\
&= \sup_{\alpha^T K_t \alpha \leq 1} \left(\alpha^T K_t (I_t W_{1:n} - e_n) \right)^2 \\
&= (I_t W_{1:n} - e_n)^T K_t (I_t W_{1:n} - e_n) \\
&= W_{1:n}^T I_t K_t I_t W_{1:n} - 2 e_n^T K_t I_t W_{1:n} + e_n^T K_t e_n,
\end{aligned}$$

where e_n is the length-n vector with $1/n$ in every entry and I_t is the n-by-n diagonal matrix with $\mathbb{I}[T_i = t]$ in the i^{th} diagonal entry. The second equality above follows by the

representer theorem, which states that when optimizing over an RKHS norm ball it is sufficient to restrict to span of the kernels at the points where the function is evaluated (Scholkopf and Smola, 2001). The third equality follows by rewriting using matrix notation. The fourth equality follows by basic Euclidean geometry and the fifth by expanding the matrix product.

3.3 Minimizing the worst-case CMSE

In the previous two Sections we decomposed the conditional mean squared error and defined the relative worst-case squared bias. If we also estimate (or, bound) the conditional variances σ_i^2, this immediately leads to an objective for the worst-case CMSE. We propose to estimate SATE using the weighted estimator with weights the minimize the worst-case CMSE of this estimator. We restrict to weights that sum to one in each treatment group, which is equivalent to just using the WLS estimator for any given unrestricted nonnegative weights. Formally, we let $\mathcal{W} = \{W_{1:n} \in \mathbb{R}^n : W_i \geq 0 \; \forall i, \sum_{i:T_i=1} W_i = \sum_{i:T_i=0} W_i = 1\}$ and choose the weights $W_{1:n}$ to solve the optimization problem

$$\min_{W_{1:n} \in \mathcal{W}} \sup_{\|f_1\|_1^2 + \|f_0\|_0^2 \leq 1} \mathbb{E}\left[\left(\hat{\tau}_{\mathrm{W}}^{\mathrm{SATE}} - \mathrm{CSATE}\right)^2 \mid X_{1:n}, T_{1:n}\right]$$
$$= \min_{W_{1:n} \in \mathcal{W}} \left(\Delta_1^2(W_{1:n}) + \Delta_0^2(W_{1:n}) + \sum_{i=1}^n W_i^2 \sigma_i^2\right). \tag{3.8}$$

By minimizing the worst-case CMSE, this optimization problem essentially finds weights that optimally balance the confounders (by minimizing the relative worst-case discrepancies) while simultaneously controlling precision (by regularizing the norm of $W_{1:n}$). In particular, the worst-case discrepancies $\Delta_t(W_{1:n})$ are precisely a distributional distance (specifically, an integral probability metric) between the sample distribution of covariates and the reweighted t-treated-group distribution of covariates.

If we use an RKHS norm as we have in the last section then this optimization problem reduces to a linearly-constrained convex-quadratic optimization problem:

$$\min_{\substack{W_{1:n}\geq 0,\\ W_{1:n}^T I_1 e_n = W_{1:n}^T I_0 e_n = n}} W_{1:n}^T (I_1 K_1 I_1 + I_0 K_0 I_0 + \Sigma) W_{1:n} - 2 e_n^T (K_1 I_1 + K_0 I_0) W_{1:n}, \tag{3.9}$$

where Σ is the n-by-n diagonal matrix with σ_i^2 in its i^{th} diagonal entry. This optimization problem can be easily and quickly solved by many off-the-shelf solvers (in particular, the problem can be efficiently solved by a polynomial-time algorithm). We use Gurobi (Gurobi Optimization, 2018), for example.

4 Practical guidelines for choosing kernels and conditional variances

In the previous Sections we formulated a novel KOM approach to find optimal weights for estimating SATE. This, however, depended on a choice of kernel and conditional variances. Indeed, the solutions to the optimization problem (3.9) depends on these choices.

We generally propose to use a polynomial Mahalanobis kernel:

$$\mathcal{K}_t(x, x') = \gamma_t (1 + \theta_t (x - \hat{\mu}_n)^T \hat{\Sigma}_n^{-1} (x' - \hat{\mu}_n))^d, \tag{4.1}$$

where $\hat{\mu}_n$ is the sample mean of confounders and $\hat{\Sigma}_n$ their sample covariance (in other word, we simply Studentize the data first). This kernel has a few hyperparameters: γ_t, θ_t, and d. The parameter d controls the degree of the polynomial kernel. We generally suggest to use 2 or 3 mostly based on the numerical results from simulations in the following Section. This choice for d offers the model some flexiblity to balance higher order moments of covariates, while the other hyperparameters allow us to control the relative importance of such higher

orders. In particular, KOM with polynomial kernel degree 3 outperforms IPTW, truncated IPTW, PSM, RA, CBPS and SBW with respect to both bias and MSE across all levels of practical positivity violation in our simulations in Section 5.

We suggest to choose the other two hyperparameters, γ_t and θ_t, as well as the conditional variance parameters, σ_i^2, in a data-driven way. The parameter θ_t controls the relative importance of higher-order moments: a lower value stresses more balance in lower-order moments over higher-order moments. We would like to chose this to match the level of nonlinearity of f_t. Finally, the parameter γ_t controls the overall scale of the kernel and we would like to chose it to match the scales of f_t. In particular, to achieve this, we suggest to tune γ_t and θ_t using the empirical Bayes approach of marginal likelihood (Rasmussen, 2004). Specifically, we suppose f_1, f_0 came from a Gaussian process with kernels $\mathcal{K}_1, \mathcal{K}_0$ and that each Y_i was observed from $f_{T_i}(X_i)$ with Gaussian noise of variance $\lambda_{T_i}^2$. We then choose the values for γ_t, θ_t, λ_t that maximize the likelihood of the data and we set $\sigma_i^2 = \lambda_{T_i}^2$. This has various unique benefits, such as automatically learning the structure of the data and preferring simpler models by default. This method is also implemented in the matlab package GPML. In our code, we provide a sufficient re-implementation in R.

Of course, there are many other possible choices and one of the benefits of the KOM approach is its great flexibility. For example, one may use the Gaussian or Matérn kernels (Scholkopf and Smola, 2001) instead of the polynomial, or even much more complicated kernels (Wilson and Adams, 2013). Additionally, instead of Studentizing the data, we could instead parameterize the matrix in the inner product used in the polynomial, Gaussian, or Matérn kernel (i.e., replace $\hat{\Sigma}_n^{-1}$ in eq. (4.1) by a parameter matrix Ω) and learn that matrix as part of the marginal likelihood tuning step. For example, an approach known as Automatic Relevance Detection (ARD) is to use a diagonal matrix with tunable variable-specific weights on the diagonal. This allows us to learn the importance of different variables

and appropriately stress the balance in the different variables and their interactions.

5 Simulations

In this Section, we compare the performance of KOM with IPTW, truncated IPTW (tIPTW), Propensity Score Matching (PSM), Regression Adjustment (RA), CBPS and SBW with respect to bias and MSE in estimating SATE in various linear, nonlinear, correct, and misspecified scenarios and across different levels of strength of practical violation of the positivity assumption. All bias and MSE values are computed over 500 replications.

5.1 Setup

We considered a sample size of $n = 200$. For the linear scenario we drew data from the following model: $Y_i = \alpha + \delta T_i + \sum_{k=1}^{K} X_{i,k} + N(0,1)$, where $T_i \sim \text{binom}(\pi_i)$, $\pi_i = \text{expit}(\beta(\sum_{k=1}^{K} X_{i,k}))$, $X_{k,i} \sim \text{N}(0,1)$, $k = 1, \ldots, K$, and $K = 2$. For the nonlinear scenario, we drew data from the following model: $Y_i = \alpha_1 + \delta T_i + \sum_{k=1}^{K} X_{i,k} + \sum_{k=1}^{K} X_{i,k}^2 + \sum_{k \neq m} X_{i,k} X_{i,m} + N(0,1)$, where $T_i \sim \text{binom}(\pi_i)$, $\pi_i = \text{expit}(\beta(\sum_{k=1}^{K} X_{i,k} + \sum_{k=1}^{K} X_{i,k}^2 + \sum_{k \neq m} X_{i,k} X_{i,m}))$, $X_{k,i} \sim \text{N}(0,1)$, $k = 1, \ldots, K$, and $K = 2$. The intercepts α and α_1 were chosen so that the marginal mean of Y_i was equal to 0. We set the true causal parameter $\delta = 1$. We vary β in order to vary the level of practical positivity violation.

In particular, we considered seven equally-spaced values, ranging from 0.1 to 3, for the β parameter in the treatment assignment models above. By tuning this parameter, we can easily control the strength of practical positivity violation, where higher values correspond to a strong practical positivity violation. For instance, in the linear scenario, under the weakest level considered ($\beta = 0.1$), the propensity score ranged on average between 0.5 and 0.8, while under the strongest level ($\beta = 3$) between 0.002 and 0.999 (average of min/max

over replications).

For the correct scenarios we plugged into the models the correct variables, X_1 and X_2. We refer to these scenarios as correct. To evaluate the performance under misspecification, we also generated $Z_1 = (2 + X_1)/(\exp(X_1))$ and $Z_2 = ((X_1 X_2/25) + 1)^3$ and plugged them into the models instead of the correct variables X_1 and X_2. We refer to these scenarios as misspecified.

For each scenario and in each sample, we then computed the set of KOM, IPTW, tIPTW, PSM, CBPS and SBW weights. Specifically, under the linear correct scenario, we computed the set of KOM weights by using a linear kernel (KOM-$\mathcal{K}_1$), IPTW and PSM weights by regressing the treatment on the linear terms using logistic regression, and CBPS and SBW weights by including the linear terms in the covariates fed to the methods. We refer to the last four as linear IPTW, PSM, CBPS, and SBW. Under the nonlinear correct scenario, we computed the set of KOM weights by using a polynomial degree 2 kernel (KOM-$\mathcal{K}_2$), IPTW and PSM weights by regressing the treatment on the linear, quadratic and interaction terms using logistic regression, and CBPS and SBW weights by including linear, quadratic and interaction terms in the covariates fed to the methods. We refer to the last four as quadratic IPTW, PSM, CBPS, and SBW. Under both linear and nonlinear misspecified scenarios, we computed the set of KOM weights by using a polynomial degree 3 kernel (KOM-$\mathcal{K}_3$), IPTW and PSM weights by regressing the treatment on the linear, quadratic, cubic and interaction terms (all monomials up to degree three) using logistic regression, and CBPS and SBW weights by including all monomials up to degree three in the covariates fed to the methods. We refer to the last four as cubic IPTW, PSM, CBPS and SBW. We specified the level of balance for SBW to be equal to 1/100 (Zubizarreta, 2015). If SBW failed to find a solution, we increased the level of balance to 1/10 and then to 1 if that also failed. For each scenario and each level of the strength of practical positivity

violation we also computed a set of truncated IPTW weights. Specifically, we truncated the IPTW weights at the 1st and 99th percentile of their distribution as suggested by Cole and Hernán (2008). To compute the KOM weights, we rescaled the covariates to have mean 0 and variance 1 and tuned the hyperparameters by using Gaussian process marginal likelihood, as described in our practical guidelines in Section 4.

Given a set of weights, we estimated the SATE by using a WLS estimator, regressing the outcome on the treatment, weighted by KOM, IPTW, tIPTW, PSM, CBPS, and SBW. To estimate SATE via RA, we computed, for each scenario and each level of practical positivity violation, the contrasts of means of treatment-specific predicted outcomes. We used the R interface of Gurobi to obtain the set of KOM weights, and the glm, CBPS and sbw packages to obtain the set of IPTW, tIPTW, CBPS and SBW weights respectively. We also chose Gurobi as solver to obtain the SBW weights. We used the R package Matching with the default settings (Sekhon, 2011) to perform PSM. We used lm for RA.

5.2 Results

In this section we present and discuss the simulations results obtained across levels of practical positivity violation in the correct linear and nonlinear scenarios (Section 5.2.1), in the misspecified linear scenario (Section 5.2.2) and in the misspecified nonlinear scenario (Section 5.2.3). In summary, KOM outperformed IPTW, tIPTW, PSM, RA, CBPS and SBW with respect to bias and MSE across all levels of practical positivity violation and considered scenarios. In addition, KOM outperformed the other methods especially under strong practical positivity violation.

5.2.1 Correct linear and nonlinear scenarios

Figure 1 shows squared bias (left panels) and MSE (right panels) of KOM (solid-circle), IPTW (dashed), tIPTW (long-dashed), PSM (two-dashed), RA (solid), CBPS (dot-dashed) and SBW (dotted) in the correct linear scenario (top panels) and correct nonlinear scenario (bottom panels). Under the correct linear scenario, KOM-$\mathcal{K}_1$ outperformed IPTW, tIPTW, PSM, and CBPS with respect to both bias and MSE. It is worth mentioning that, while, the bias and the MSE of IPTW, tIPTW, PSM, and CBPS increased with the levels of practical positivity violation, those of KOM-$\mathcal{K}_1$ were consistently low across all levels. Notably, in the linear scenarios, linear SBW and KOM-$\mathcal{K}_1$ performed similarly since both control a similar linear moment discrepancy of just a few (two) covariates. KOM-$\mathcal{K}_1$ also performed similarly to RA. In Section 5.2.4 we show that KOM, which optimizes these discrepancies directly, outperforms SBW with respect to both bias and MSE, and it outperforms RA with respect to MSE, in these linear scenarios when the number of confounders considered is increased. In the nonlinear correct, misspecified linear, and misspecified nonlinear scenarios, KOM also outperformed SBW and RA, as discussed below.

The lower panels of Figure 1 show the bias and the MSE across levels of practical positivity violation under the correct nonlinear scenario. KOM-$\mathcal{K}_2$ outperformed IPTW, tIPTW, PSM, CBPS and SBW with respect of both bias and MSE across all considered levels of practical positivity violation. It is worth mentioning that the bias of KOM-$\mathcal{K}_2$ was as low as that of the RA, which is theoretical zero given the fact that RA used the correct model. In addition, contrary to RA, KOM-$\mathcal{K}_2$ also resulted in a low MSE while that of RA exploded when increasing the level of practical positivity violation. Although KOM and RA can be thought as methodologically similar techniques, the results of our simulation study suggest that KOM-$\mathcal{K}_2$ can be used even in nonlinear settings without being affected by moderate or strong practical positivity violation.

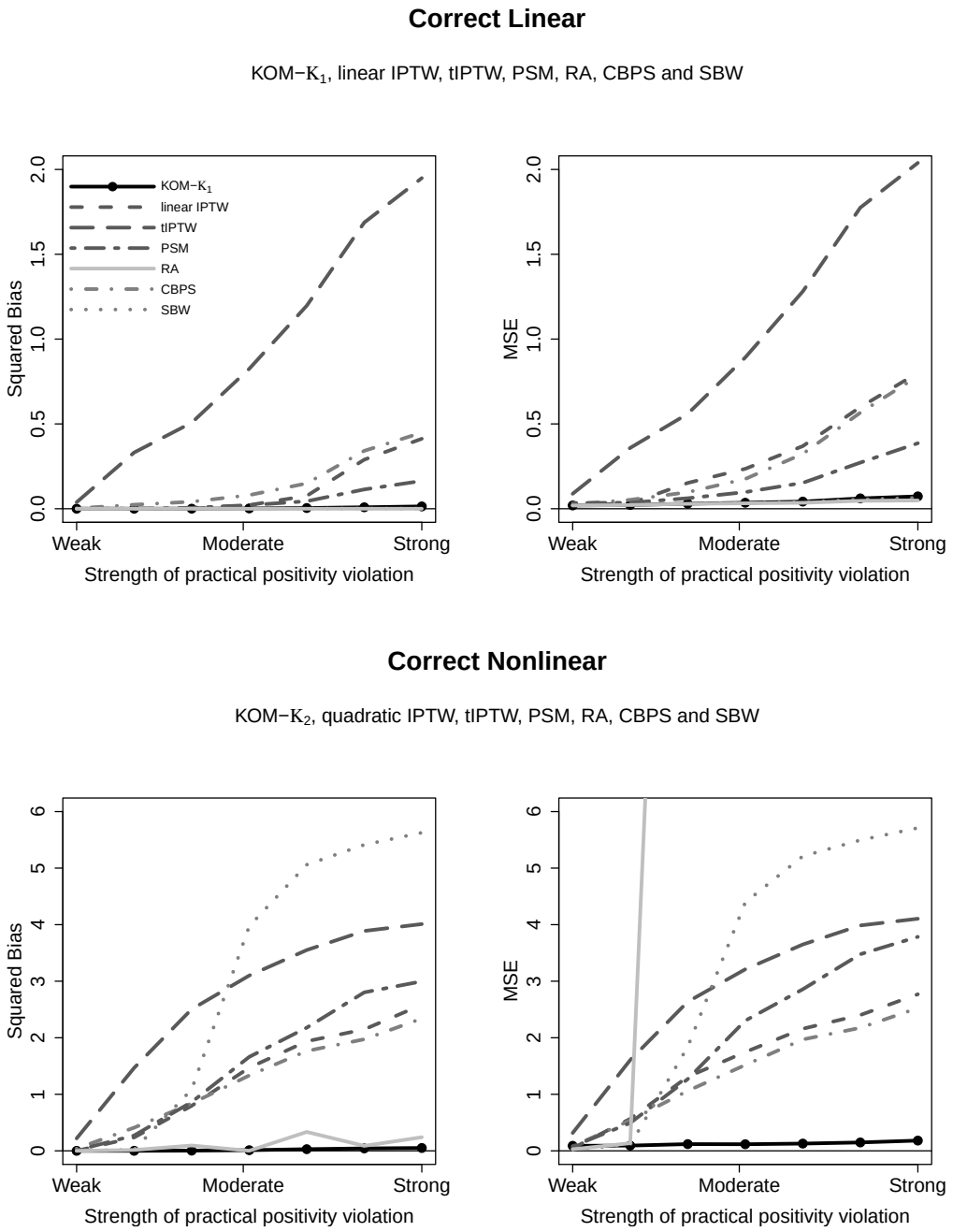

Figure 1: Squared bias (left panels) and MSE (right panels) of the estimated SATE using KOM (solid-circle), IPTW (dashed), tIPTW (long-dashed), PSM (two-dashed), RA (solid), CBPS (dashed-dotted) and SBW (dotted) when increasing the strength of practical positivity violation in the correct linear scenario (top panels) and in the correct nonlinear scenario (bottom panels), $n = 200$. Top panels shows the results when using KOM-$\mathcal{K}_1$, linear IPTW, tIPTW, PSM, RA, CBPS, and SBW. Bottom panels show the results when using KOM-$\mathcal{K}_2$, quadratic IPTW, tIPTW, PSM, RA, CBPS, and SBW.

5.2.2 Misspecified linear scenario

The top panels of Figure 2 shows squared bias (left panels) and MSE (right panels) of KOM (solid-circle), IPTW (dashed), tIPTW (long-dashed), PSM (two-dashed), RA (solid), CBPS (dashed-dotted) and SBW (dotted) in the misspecified linear scenario. In this scenario, we observed that the cubic variants of methods better handle the misspecification compared with the linear ones. We therefore focus only on the results obtained from KOM-$\mathcal{K}_3$, cubic IPTW, tIPTW, PSM, RA, CBPS, and SBW. KOM-$\mathcal{K}_3$, a polynomial degree 3 kernel for KOM, outperformed cubic IPTW, tIPTW, PSM, RA, CBPS, and SBW across all considered levels of practical positivity violation. Cubic RA resulted in very high bias and MSE across all levels (results are outside the plot region in Fig. 2).

5.2.3 Misspecified nonlinear scenario

The bottom panels of Figure 2 shows squared bias (left panel) and MSE (right panel) of KOM (solid-circle), IPTW (dashed), tIPTW (long-dashed), PSM (two-dashed), RA (solid), CBPS (dashed-dotted) and SBW (dotted) in the misspecified nonlinear scenario. KOM-$\mathcal{K}_3$ outperformed cubic IPTW, tIPTW, PSM, RA, CBPS, and SBW. Cubic RA resulted in very high bias and MSE across all levels of practical positivity violation (results are outside the plot region in Fig. 2).

In summary, KOM showed a consistently lower bias and MSE across all considered scenarios and across levels of practical positivity violation, and especially under strong practical violation. These results suggest that KOM with a polynomial degree $d \geq 2$ kernel mitigates the impact of model misspecification while being able to handle strong practical positivity violations.

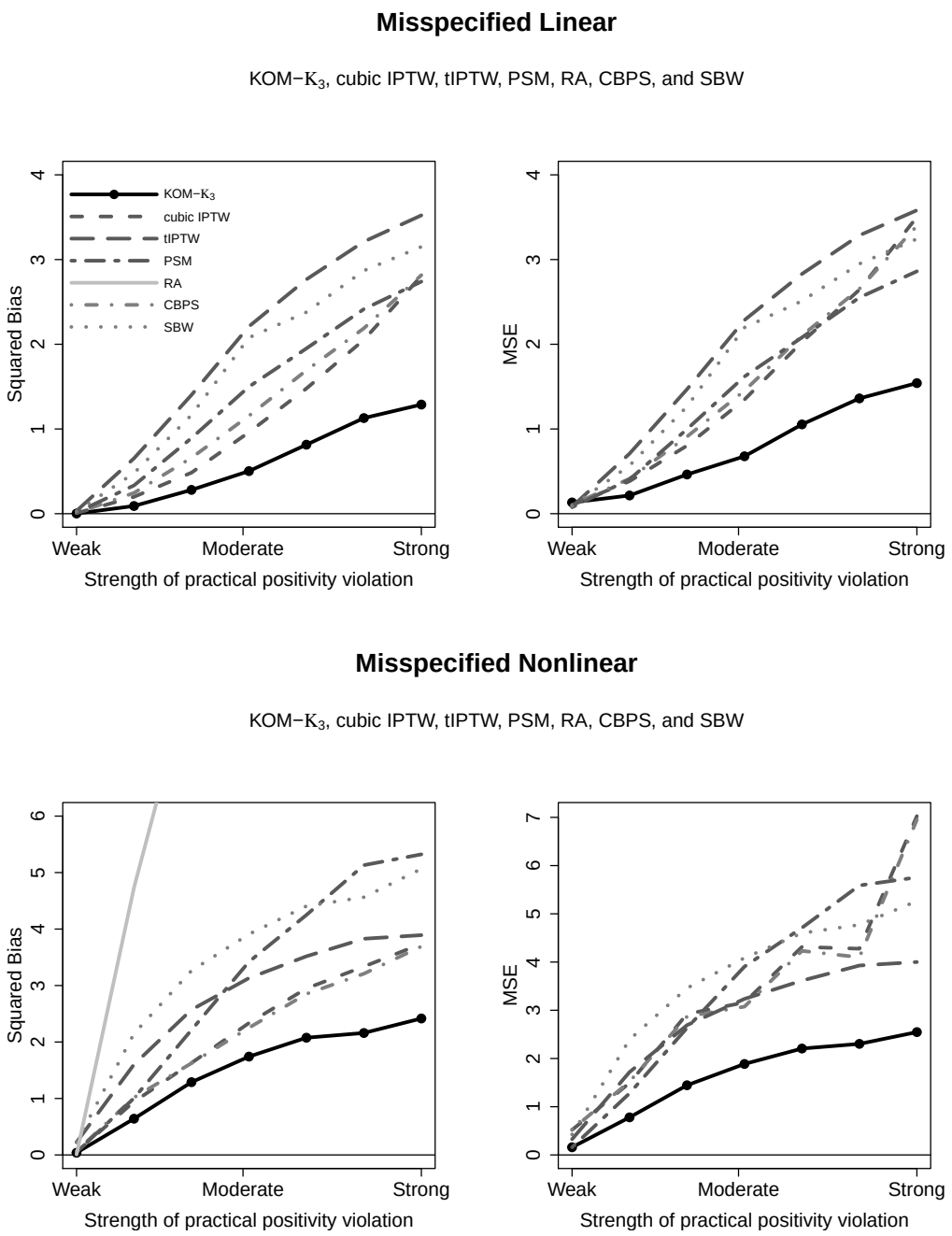

Figure 2: Squared bias (left panels) and MSE (right panels) of the estimated SATE using KOM-$\mathcal{K}_3$ (solid-circle), cubic IPTW (dashed), cubic tIPTW (long-dashed), cubic PSM (two-dashed), cubic RA (solid; outside of plot region in 3 of 4 plots), cubic CBPS (dashed-dotted), and cubic SBW (dotted) when increasing the strength of practical positivity violation in the misspecified linear scenario (top panels) and in the misspecified nonlinear scenario (bottom panels), $n = 200$.

5.2.4 Linear SBW, RA and KOM-$\mathcal{K}_1$ when increasing the number of confounders under linear scenarios

The results presented in the top panels of Figure 1 suggest that in the correct linear scenario when the number of confounders considered was equal to 2, KOM performed similarly to SBW and RA with respect to bias and MSE. Motivated by the fact that in practice (including in our own application), the number of confounders used for analysis can be much larger, in this Section we present the results of a simulation study aimed at comparing bias and MSE of KOM, SBW and RA when increasing the number of confounders. Specifically, we drew data from the following model: $Y_i = \alpha + \delta T_i + \sum_{k=1}^{K} X_{i,k} + N(0,1)$, where $T_i \sim \text{binom}(\pi_i)$, $\pi_i = \text{expit}(\beta(\sum_{k=1}^{K} X_{i,k}))$, $\delta = 1$, and $X_{k,i} \sim \text{N}(0,1), k = 1, \ldots, K$, with $K = 2, 20, 50$ and 100. We set $\beta = 2$ for a moderately strong practical positivity violation and computed bias and MSE over 500 replications in the correct linear scenario with a sample size of $n = 200$.

Figure 3 shows squared bias (left panels) and MSE (right panels) in the correct linear scenario across $K = 2, 20, 50$, and 100 number of confounders. KOM outperformed SBW with respect to bias and MSE across all considered numbers of confounders, suggesting that KOM provides lower bias and MSE compared with SBW when the number of confounders is moderate. KOM outperformed RA with respect to MSE when the number of confounders increased.

5.3 Coverage

To compute confidence intervals of a weighted estimator for SATE, Wald confidence intervals can be used together with the robust sandwich estimator (Hernán et al., 2001; Robins et al., 2000; Freedman, 2006). We next compare the empirical coverage of such

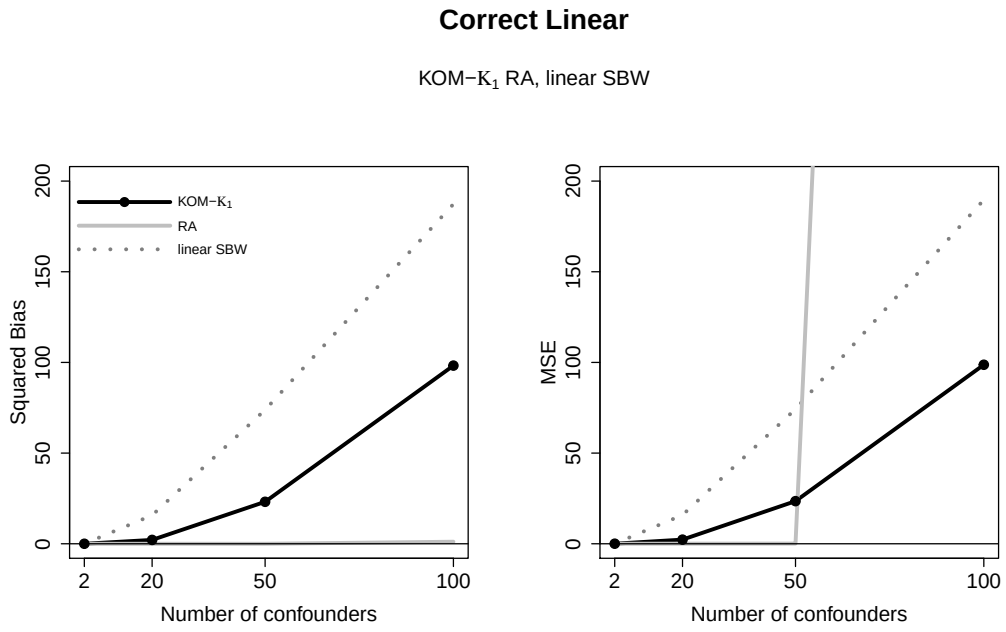

Figure 3: Squared bias (left panels) and MSE (right panels) of the estimated SATE using KOM-$\mathcal{K}_1$ (solid-circle), RA (solid) and liner SBW (dotted) when increasing the number of confounders (2, 20, 50, 100) in the linear correct scenario, $n = 200$.

Table 1: Empirical coverage of Wald 95% confidence intervals

Scenario	**Method**					
	KOM	IPTW	tIPTW	SBW	CBPS	PSM
Correct linear	0.92	0.45	0.05	0.95	0.45	0.88
Correct nonlinear	0.88	<0.01	<0.01	<0.01	<0.01	0.69
Misspecified linear	0.27	<0.01	<0.01	<0.01	<0.01	0.73
Misspecified nonlinear	0.02	0.16	<0.01	0.01	0.09	0.10

95% confidence intervals for the various methods across scenarios under the strongest practical positivity violation setting. In the case of PSM, we use the standard error estimator proposed by Abadie and Imbens (2006). Table 1 shows the results. In summary, KOM achieved desirable coverage under both linear and nonlinear correct scenarios. These results are similar to those found by Kallus (2016)[Section 4.4] in which coverage was computed keeping $X_{1:n}$ and $T_{1:n}$ fixed. Since all methods had significant bias in the misspecified scenarios, they all exhibit undercoverage, as expected. The slightly higher coverage of IPTW, PSM and CBPS with cubic logistic models simply arises from their much larger variance, leading to very wide confidence intervals. Indeed, when truncating the IPTW weights, leading to lower variance without affecting bias (see bottom panels of Fig. 2), the coverage drops to 0%.

5.4 Computational time of KOM

In this Section we report the computational time required by KOM in the simulation study described in Section 5.1. Three steps are required to compute the set of KOM weights. First, we tune the kernel's hyperparameters; second, we construct the matrices required by

problem (3.9); and third we solve problem (3.9). We computed the computational time by using the R package **rbenchmark** on a AWS EC2 C5 instance, Intel Xeon Platinum 8000 series, 3.5 GHz, 16GB RAM and a Linux Ubuntu 16.04 operating system.

In the correct linear scenario with $n = 200$, KOM required a mean computational time of 2 seconds to obtain the weights. Tuning the hyperparameters required 50% of the computational time, computing the matrices 49%, and solving the optimization problem 1%. Similar mean computational times were observed in the misspecified linear scenario and in the correct nonlinear scenario. In the misspecified nonlinear scenario, KOM required a mean computational time of 3.2 seconds to obtain the weights. Tuning the hyperparameters required 70% of the computational time, computing the matrices 29%, and solving the optimization problem 1%. The mean computational times were similar across levels of practical positivity violation.

6 Application to the study of spine surgical interventions

In this Section we apply KOM to the observational study presented in Section 2. We used data from a single-institutional subset of the Spine QOD registry (NeuroPoint Alliance, 2018). The registry was launched in 2012 with the aim of evaluating the effectiveness of spine surgery interventions on the improvement of pain, disability, and quality of life. QOD contains clinical and demographic information as well as patient-reported outcomes. We evaluated the effect of fusion-plus-laminectomy compared to laminectomy alone on the Oswestry Disability Index (ODI), an index used by surgeons to quantify disability, for the treatment of lumbar stenosis or spondylolisthesis. Previous randomized control trials have shown that fusion-plus-laminectomy and laminectomy alone have equivalent average

improvement on the ODI of patients with these conditions (Ghogawala et al., 2016; Försth et al., 2016).

6.1 Study population and models setup

We restrict our study to *primary surgery*, defined as the first spine surgery intervention for each patient. Patients were interviewed before surgical intervention, and demographic and clinical information was collected. ODI was collected at 3-month follow-up. The study subset was composed of 311 patients, 247 of which received laminectomy alone and 64 fusion-plus-laminectomy. As described in Section 2, spine surgical practice may lead to a practical violation of the positivity assumption. In our dataset, 1% of those patients with a moderate-low leg pain were treated with fusion-plus-laminectomy and less than 10% of the patients with lumbar spondylolistheses were treated with laminectomy alone.

We identified as potential confounders the following variables: lumbar stenosis (yes vs. no), lumbar spondylolistheses (yes vs. no), leg pain (score from 0 to 10), back pain (score from 0 to 10), activity outside home (yes vs. no), activity at home (yes vs. no), duration of symptom (less than 3 months vs. greater than or equal to 3 months), motor deficiency (yes vs. no), dominant symptoms (back; leg; both), and age at interview. Common statistical practice suggest using IPTW to consistently estimate the effect of laminectomy alone versus fusion-plus-laminectomy in the presence of these confounders. To apply IPTW, we used logistic regression to estimate the propensities and compute the set of IPTW weights by taking their inverse. Based on the simulation results showed in Section 5.2.2 and 5.2.3, we used a cubic logistic regression models (IPTW_3). We also compute the set of KOM weights by using a polynomial kernel degree 3 (KOM-$\mathcal{K}_3$). We tuned the kernel's hyperparameters using Gaussian process marginal likelihood and solve problem (3.9) by using quadratic optimization.

Table 2: The effect of fusion-plus-laminectomy on ODI

	Naive	IPTW_3	KOM-$\mathcal{K}_3$
$\hat{\beta}_2$ (SE)	5.1* (2.3)	9.7* (4.6)	0.5 (4.4)

* indicates statistical significance at the 0.05 level.

We considered the following model to evaluate the effect of fusion-plus-laminectomy ($T = 1$) versus laminectomy alone ($T = 0$) on ODI among patients with lumbar stenosis or spondylolisthesis,

$$\mathbb{E}[Y_i(1)] = \beta_1 + \beta_2 \mathbb{I}[T = 1], \tag{6.1}$$

where $\mathbb{I}[T = 1]$ is the indicator function for fusion-plus-laminectomy, $Y_i(T)$ is the potential outcome of observing ODI under intervention T for the i-th unit, β_1 is the effect of laminectomy alone and β_2 is the SATE. We estimated β_2 using either ordinary least squares (unweighted) or weighted least squares with weights given either by IPTW_3 or KOM-$\mathcal{K}_3$. We computed robust (sandwich) standard errors in each case. We used the R interface of Gurobi to obtain the set of KOM weights, and the glm package and the poly function to obtain the set of IPTW weights. We used the R package sandwich to estimate robust standard errors.

6.2 Results

Table 2 shows the results of our analysis. When analysing the distribution of IPTW_3 weights, a weight of more than 1,000 was assigned to $n = 28$ patients, suggesting a strong practical positivity violation. Both the naive estimator ($\hat{\beta}_2 = 5.1$; SE: 2.3) and IPTW_3 ($\hat{\beta}_2 = 9.7$; SE: 4.6) indicated a statistically significant positive effect of fusion-

plus-laminectomy compared with laminectomy alone on ODI. In contrast, and similar to the results obtained by two recent randomized controlled trials (Ghogawala et al., 2016; Försth et al., 2016), KOM-$\mathcal{K}_3$ resulted in an estimated effect that is both much smaller in magnitude and is statistically insignificant ($\beta_2 = 0.5$; SE: 4.0). Whereas an analysis based on IPTW leads to conclusions that perhaps spuriously refute experimental evidence, using KOM we conclude, in agreement with experimental evidence, that among patients with lumbar stenosis or spondylolisthesis, fusion-plus-laminectomy did not result in better ODI compared with laminectomy alone.

6.2.1 Results when changing model degree

The results of the simulation study presented in Section 5 suggested that the cubic variants of all considered methods better handled model misspecification compared with the linear ones. This led us to use cubic variants in estimating the effect of fusion-plus-laminectomy compared with laminectomy alone on ODI in the above. In this Section we study the change in these estimates if we change the degree, d, of the polynomial models considered in KOM and IPTW. Specifically, we let the degree of the polynomials range from 1, corresponding to the linear kernel for KOM and a linear logistic regression model for IPTW, to 5, corresponding to a polynomial kernel of degree 5 for KOM and a quintic logistic regression model for IPTW. The results are shown in Table 3.

IPTW (second row of Table 3) led to volatile estimates that switched back and forth in both sign and significance as we varied the degree. In contrast, KOM led to stable results that decreased in magnitude from a narrowly significant effect, similar to that of the naive estimator, to a statistically insignificant effect, similar to that of the experimental results, as we increased the degree (first row of Table 3). These results suggest first that KOM results in more stable estimates and that using KOM with a nonlinear kernel ($d \geq 2$) leads

Table 3: Effect estimates when increasing the degree of polynomials

$\hat{\beta}_2$ (SE)	Linear	Quadratic	Cubic	Quartic	Quintic
KOM	4.6* (2.3)	2.1 (2.8)	0.5 (4.4)	1.5 (4.6)	0.7 (4.8)
IPTW	2.0 (3.0)	-3.3 (4.2)	9.7* (4.6)	7.6* (3.3)	4.5 (3.7)

* indicates statistical significance at the 0.05 level.

to improved control of confounders and consequently to more coherent clinical results.

7 Conclusions

In this paper, we developed an approach using KOM to provide weights for the estimation of SATE. The method developed directly and optimally controls the total error — both bias and variance – of the estimates uniformly over a class of models given by a RKHS. This leads the method to effectively mitigate issues of possible misspecification and robustly handle moderate and strong practical positivity violations, two issues that are of central concern in many observational studies.

By using mathematical optimization, KOM optimally minimizes the conditional mean squared error of any weighted estimator with respect to the weights, resulting in a lower bias and MSE compared with IPTW, tIPTW, PSM, RA, CBPS and SBW in most of the considered scenarios of our simulation study. In addition, KOM automatically learns the structure of the data and allows the researcher to balance linear, nonlinear, additive, and non-additive covariate relationships without sacrificing performance.

Alternative formulations of the optimization problem (3.9) can be used. For instance, we may limit precision by bounding the variance of the resulting weighted estimator up to

a level specified by the researcher rather than regularizing it. Additionally, we may impose different norms on the conditional expectation functions of potential outcomes and even constrain them to be equal up to a constant shift or separately regularize their difference (effect) and their average (baseline). These may provide improvements in certain settings where such structure holds.

References

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *econometrica 74*(1), 235–267.

Cole, S. R. and M. A. Hernán (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology 168*(6), 656–664.

Cook, J. A. (2009). The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. *Trials 10*(1), 9.

Eck, J. C., A. Sharan, Z. Ghogawala, D. K. Resnick, W. C. Watters III, P. V. Mummaneni, A. T. Dailey, T. F. Choudhri, M. W. Groff, J. C. Wang, et al. (2014). Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. part 7: lumbar fusion for intractable low-back pain without stenosis or spondylolisthesis. *Journal of Neurosurgery: Spine 21*(1), 42–47.

Försth, P., G. Ólafsson, T. Carlsson, A. Frost, F. Borgström, P. Fritzell, P. Öhagen, K. Michaëlsson, and B. Sandén (2016). A randomized, controlled trial of fusion surgery for lumbar spinal stenosis. *New England Journal of Medicine 374*(15), 1413–1423.

Freedman, D. A. (2006). On the so-called huber sandwich estimator and robust standard errors. *The American Statistician 60*(4), 299–302.

Ghogawala, Z., J. Dziura, W. E. Butler, F. Dai, N. Terrin, S. N. Magge, J.-V. C. Coumans, J. F. Harrington, S. Amin-Hanjani, J. S. Schwartz, et al. (2016). Laminectomy plus fusion versus laminectomy alone for lumbar spondylolisthesis. *New England Journal of Medicine 374*(15), 1424–1434.

Gurobi Optimization, L. (2018). Gurobi optimizer reference manual.

Hernán, M. A., B. Brumback, and J. M. Robins (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology 11*(5), 561–570.

Hernán, M. A., B. Brumback, and J. M. Robins (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association 96*(454), 440–448.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association 47*(260), 663–685.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in medicine 27*(24), 4857–4873.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 243–263.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ju, C., J. Schwab, and M. J. van der Laan (2017). On adaptive propensity score truncation in causal inference. *arXiv preprint arXiv:1707.05861*.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*.

Kallus, N. and A. Zhou (2018). Policy evaluation and optimization with continuous treatments. *arXiv preprint arXiv:1802.06037*.

Kang, J., W. Chan, M.-O. Kim, and P. M. Steiner (2016). Practice of causal inference with the propensity of being zero or one: assessing the effect of arbitrary cutoffs of propensity scores. *Communications for Statistical Applications and Methods 23*(1), 1–20.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Lee, B. K., J. Lessler, and E. A. Stuart (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine 29*(3), 337–346.

Lee, B. K., J. Lessler, and E. A. Stuart (2011). Weight trimming and propensity score weighting. *PloS one 6*(3), e18174.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine 23*(19), 2937–2960.

Mansournia, M. A. and D. G. Altman (2016). Inverse probability weighting. *Bmj 352*, i189.

NeuroPoint Alliance, I. (2018). Qod spine surgery registry.

Pearson, A., E. Blood, J. Lurie, W. Abdu, D. Sengupta, J. W. Frymoyer, and J. Weinstein (2011). Predominant leg pain is associated with better surgical outcomes in degenerative spondylolisthesis and spinal stenosis: results from the spine patient outcomes research trial (sport). *Spine 36*(3), 219.

Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research 21*(1), 31–54.

Raad, M., C. J. Donaldson, M. H. El Dafrawy, D. M. Sciubba, L. H. Riley III, B. J. Neuman, K. M. Kebaish, and R. L. Skolasky (2018). Trends in isolated lumbar spinal stenosis surgery among working us adults aged 40–64 years, 2010–2014. *Journal of Neurosurgery: Spine*, 1–7.

Rasmussen, C. E. (2004). Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pp. 63–71. Springer.

Resnick, D. K., W. C. Watters, P. V. Mummaneni, A. T. Dailey, T. F. Choudhri, J. C. Eck, A. Sharan, M. W. Groff, J. C. Wang, Z. Ghogawala, et al. (2014). Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. part 10: lumbar fusion for stenosis without spondylolisthesis. *Journal of neurosurgery: Spine 21*(1), 62–66.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science 22*(4), 544–559.

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

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association 90*(429), 106–121.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Santacatterina, M. and M. Bottai (2018). Optimal probability weights for inference with constrained precision. *Journal of the American Statistical Association*, 1–9.

Santacatterina, M., C. Garcia-Pareja, R. Bellocco, A. Sonnerbörg, A. M. Ekström, and M. Bottai (2018). Optimal probability weights for estimating causal eects of time-varying treatments with marginal structural cox models. In *"Under second round revision"*.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Scholkopf, B. and A. J. Smola (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the matching package for r.

Splawa-Neyman, J., D. M. Dabrowska, and T. Speed (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 465–472.

Stefanski, L. A. and D. D. Boos (2002). The calculus of m-estimation. *The American Statistician 56*(1), 29–38.

Su, Y., L. Wang, M. Santacatterina, and T. Joachims (2018, November). CAB: Continuous Adaptive Blending Estimator for Policy Evaluation and Learning. *ArXiv e-prints*.

Swaminathan, A. and T. Joachims (2015). Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pp. 814–823.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

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

Waterman, B. R., P. J. Belmont Jr, and A. J. Schoenfeld (2012). Low back pain in the united states: incidence and risk factors for presentation in the emergency setting. *The spine journal 12*(1), 63–70.

Wilson, A. and R. Adams (2013). Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pp. 1067–1075.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–922.

Supporting material

More robust estimation of sample average treatment effects using Kernel Optimal Matching in an observational study of spine surgical interventions

Nathan Kallus, Brenton Pennicooke, Michele Santacatterina

Introduction

In this document we provide the R code to compute KOM and the proof of Theorem 3.1 of the original manuscript.

R Code

Data generation

In this example we consider the nonlinear scenario, $Y_i = \alpha_1 + \delta T_i + \sum_{k=1}^{K} X_{i,k} + \sum_{k=1}^{K} X_{i,k}^2 + \sum_{k \neq m} X_{i,k} X_{i,m} + N(0,1)$, where $T_i \sim \text{binom}(\pi_i)$, $\pi_i = \text{expit}(\beta(\sum_{k=1}^{K} X_{i,k} + \sum_{k=1}^{K} X_{i,k}^2 + \sum_{k \neq m} X_{i,k} X_{i,m}))$, and $X_{k,i} \sim \text{N}(0,1), k = 1, \ldots, K$, and $K = 2$. We set $\beta = 1$, a mild violation of the positivity assumption. In `R` we have,

```
rm(list = ls())
set.seed(1)

n <- 200

#Confounders
X1 <- rnorm(n,0,1)
X2 <- rnorm(n,0,1)

#true propensities
prt <- 1/(1+exp(- (X1 + X2 + X1^2 + X2^2 + X1*X2 ) ) )

#treatment assignment model
Tr <- rbinom(n,1,prt)

truet <- 1
interc <- -mean(prt)
Y <- interc + truet*Tr + X1 + X2 + X1^2 + X2^2 + X1*X2 + rnorm(n)
X <- cbind(X1,X2)

dta <- data.frame(Y,X,Tr)
dta <- dta[order(dta$Tr,decreasing=TRUE),]
```

Tuning the hyperparameters

We defined the polynomial kernel degree 2 as

```
#Polynomial degree 2 kernel
Kpoly <-  function(vect,x){
  xs <- scale(x)
  offset <- vect[2]
  B <- (xs%*%t(xs) + offset)^2
  return(B)
}
```

Once the kernel is specified, we need to minimize the log likelihood with respect to the hyperparameters.

```
#function to be minimized - log likelohood
minus_log_likelihood <- function(vect,X,Y,nte){
  sn2 <- vect[3]^2
  gamma2 <- vect[1]^2
  K_plus <- gamma2*Kpoly(vect,X) + sn2*diag(nte)
  K_plus_inv <- try(solve(K_plus, tol = 1e-21))
  if(class(K_plus_inv) != "try-error"){
    z <- determinant(K_plus, logarithm=TRUE)
    K_plus_log_det <- as.numeric((z$sign*z$modulus)) # log-determinant of K_plus
    out <- 0.5 * ( t(Y) %*% K_plus_inv %*% Y ) + 0.5 * K_plus_log_det + (nte/2)*log(2*pi)
  }
  return(out)
}
```

The hyperparameter `sn2` is the conditional variance, `gamma2` controls the overall scale of the kernel, and `offset` is the a parameter that trades off the influence of lower-order versus higher-order terms in the polynomial. We tune the hyperparameters among the treated and among the untreated units. We used the `L-BFGS-B` algorithm to tune the hyperparameters.

```
t1 <-as.integer(dta$Tr)
t0 <-as.integer((1-dta$Tr))


# GPML optimization --- "L-BFGS-B" algorithm

Xsample <- data.frame(X,Y,Tr)

X1 <- Xsample[which(Xsample$Tr==1),]
X0 <- Xsample[which(Xsample$Tr==0),]

y1 <- Y[which(Xsample$Tr==1)]
y0 <- Y[which(Xsample$Tr==0)]

n1 <- length(which(Xsample$Tr==1))
n0 <- length(which(Xsample$Tr==0))

tol <- 1e-08

#Untreated
res.optim2_0 <- try(optim(par=c(1, 1, 1),
                          fn=minus_log_likelihood,
                          method=c("L-BFGS-B"),
```

```
                          lower = rep(tol,3),
                          hessian=TRUE,
                          control=list(trace=0, maxit=1000),
                          X=X0[,1:(dim(X0)[2]-2)],
                          Y=y0,
                          nte=n0))

#Treated
res.optim2_1 <- try(optim(par=c(1, 1, 1),
                          fn=minus_log_likelihood,
                          method=c("L-BFGS-B"),
                          lower = rep(tol,3),
                          hessian=TRUE,
                          control=list(trace=0, maxit=1000),
                          X=X1[,1:(dim(X1)[2]-2)],
                          Y=y1,
                          nte=n1))

#Save tuned hyperparameters
res.optim2_0 <- list(par=c(res.optim2_0$par[1],res.optim2_0$par[2],res.optim2_0$par[3]))
res.optim2_1 <- list(par=c(res.optim2_1$par[1],res.optim2_1$par[2],res.optim2_1$par[3]))
```

Building the matrices

Recall that we want to solve,

$$\min_{W_{1:n}\geq 0, W_{1:n}^T I_1 e_n = W_{1:n}^T I_0 e_n = n} W_{1:n}^T (I_1 K_1 I_1 + I_0 K_0 I_0 + \Sigma) W_{1:n} - 2 e_n^T (K_1 I_1 + K_0 I_0) W_{1:n},$$

where K_t is a PSD, $\Sigma = I_1 \sigma_1^2 + I_0 \sigma_0^2$, e_n is the vector of ones, and I_t is the diagonal matrix with $\mathbb{1}[T_i = t]$ in its i^{th} diagonal entry.

Then in `R` we obtain the matrices K_1 and K_0 and the standard deviations σ_1^2 and σ_0^2 by,

```
K1 <- res.optim2_1$par[1]^2*Kpoly(res.optim2_1$par,X)
K0 <- res.optim2_0$par[1]^2*Kpoly(res.optim2_0$par,X)

sigma1 <- res.optim2_1$par[3]^2
sigma0 <- res.optim2_0$par[3]^2
```

where, `res.optim2$par` is the vector of the tuned hyperparameters. Now, we want to compute $I_1 K_1 I_1 + I_0 K_0 I_0 + \Sigma$,

```
##Quadratic terms

# I1 K1 I1 + I0 K0 I0
I1KI1 <- outer(t1, t1)*K1
I0KI0 <- outer(t0, t0)*K0

#Sigma
Sigma <- sigma1*diag(t1) + sigma0*diag(t0)

#Q
Q <- ( I1KI1 + I0KI0 + Sigma )
```

We now want to compute the linear term, $e_n^T (K_1 I_1 + K_0 I_0)$

```
#Linear part
# K1 I1 + K0 I0
KI1 <- diag(t1)%*%K1
KI0 <- diag(t0)%*%K0

# en^T K1 I1 + en^T K0 I0
onesn <- rep(1/n,n)
onesnKI1 <- t(onesn)%*%KI1
onesnKI0 <- t(onesn)%*%KI0

c <- -2*(onesnKI1 + onesnKI0)
```

Solving the optimization problem

We solve the problem by using Gurobi,

```
model <- list()
model$A          <- matrix(c(((t1)/n),((t0)/n)), nrow=2, byrow=T)
model$rhs        <- c(1,1)
model$modelsense <- "min"
model$Q          <- Q
model$obj        <- c
model$sense      <- c("=")
model$lb <- rep(tol,n)
model$vtypes <- "C"

params <- list(Presolve=2,OutputFlag=0,QCPDual=0)

res <- gurobi(model,params)
kow <- res$x
summary(kow)
summary(kow*t1)
summary(kow*t0)
```

Estimating SATE

We now estimate the SATE,

```
fit <- lm(Y ~ Tr, data = dta, weights=kow)
fit$coef
sqrt(diag(sandwich(fit)))
```

Proof of Theorem 3.1

Bias of $\hat{\tau}^{\mathbf{SATE}}_{\mathbf{W}}$ with respect to SATE and CSATE

We show that

$$\begin{aligned}\mathbb{E}\left[\hat{\tau}^{\mathrm{SATE}}_{\mathrm{W}} - \mathrm{SATE} | X_{1:n}, T_{1:n}\right] &= \mathbb{E}\left[\hat{\tau}^{\mathrm{SATE}}_{\mathrm{W}} - \mathrm{CSATE} | X_{1:n}, T_{1:n}\right] \\ &= B_1(W_{1:n}; f_1) - B_0(W_{1:n}; f_0)\end{aligned}$$

which is equation (3.4) in the original manuscript.

The first equality follows by noting that $\mathbb{E}[\mathrm{SATE} \mid X_{1:n}, T_{1:n}] = \mathbb{E}[\mathrm{SATE} \mid X_{1:n}] = \mathrm{CSATE}$, where the first equality is by ignorability and second by definition.

Next, define $\epsilon_{i,t} = Y_i(t) - f_t(X_i)$, $\epsilon_i = T_i\epsilon_{i,1} + (1 - T_i)\epsilon_{i,0}$, and

$$\Xi(W_{1:n}) = \sum_{i=1}^{n}(W_iT_i - \frac{1}{n})\epsilon_{i,1} - \sum_{i=1}^{n}(W_i(1 - T_i) - \frac{1}{n})\epsilon_{i,0}.$$

Then,

$$\begin{aligned}\hat{\tau}^{\mathrm{SATE}}_{\mathrm{W}} - \mathrm{SATE} &= \sum_{i=1}^{n} W_i(T_iY_i - (1 - T_i)Y_i) - \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) \\ &= \sum_{i=1}^{n}(W_iT_i - \frac{1}{n})(f_1(X_i) + \epsilon_{i,1}) - \sum_{i=1}^{n}(W_i(1 - T_i) - \frac{1}{n})(f_0(X_i) + \epsilon_{i,0}) \\ &= \sum_{i=1}^{n}(W_iT_i - \frac{1}{n})f_1(X_i) - \sum_{i=1}^{n}(W_i(1 - T_i) - \frac{1}{n})f_0(X_i) \\ &\quad + \sum_{i=1}^{n}(W_iT_i - \frac{1}{n})\epsilon_{i,1} - \sum_{i=1}^{n}(W_i(1 - T_i) - \frac{1}{n})\epsilon_{i,0} \\ &= B(W_{1:n}; f_1) - B(W_{1:n}; f_0) + \Xi(W_{1:n}),\end{aligned}$$

where the second equality follows by consistency. For each i and each t, by the definition of f_t and by ignorability, $\mathbb{E}[\epsilon_{i,t}|X_{1:n}, T_{1:n}] = \mathbb{E}[Y_i(t)|X_i, T_i] - f_t(X_i) = \mathbb{E}[Y_i(t)|X_i] - f_t(X_i) = 0$.

Since W is assumed to be a function of $X_{1:n}, T_{1:n}$, we have that

$\mathbb{E}[\Xi(W) \mid X_{1:n}, T_{1:n}] = 0$ and hence

$$\mathbb{E}\left[\hat{\tau}^{\mathrm{SATE}}_{\mathrm{W}} - \mathrm{SATE} | X_{1:n}, T_{1:n}\right] = B_1(W_{1:n}; f_1) - B_0(W_{1:n}; f_0).$$

CMSE of $\hat{\tau}^{\mathbf{SATE}}_{\mathbf{W}}$ with respect to CSATE

We show that

$$\mathbb{E}\left[\left(\hat{\tau}^{\mathrm{SATE}}_{W} - \mathrm{CSATE}\right)^2 | X_{1:n}, T_{1:n}\right] = (B_1(W_{1:n}; f_1) - B_0(W_{1:n}; f_0))^2 + \sum_{i=1}^{n} W_i^2\sigma_i^2.$$

We define $\sigma^2_{i,t} = \text{Var}(Y_i(t) \mid X_i)$. Under consistency, non-interference and ignorability, we have $\sigma^2_i = \text{Var}(Y_i \mid X_i, T_i) = T_i\sigma^2_{i,1} + (1-T_i)\sigma^2_{i,0}$. First, notice that by conditioning on $X_{1:n}$ and $T_{1:n}$, $\hat{\tau}^{\text{SATE}}_W - \text{CSATE}$ is a constant. Then, we can decompose the CMSE as bias squared plus variance. Recall that

$$\begin{aligned}\hat{\tau}^{\text{SATE}}_{\text{W}} - \text{CSATE} = &\sum_{i=1}^{n}(W_iT_i - \frac{1}{n})f_1(X_i) - \sum_{i=1}^{n}(W_i(1-T_i) - \frac{1}{n})f_0(X_i) \\ &+ \sum_{i=1}^{n}(W_iT_i)\epsilon_{i,1} - \sum_{i=1}^{n}(W_i(1-T_i))\epsilon_{i,0}.\end{aligned}$$

To compute the variance we then consider only the two last terms. Then, for each i, j, we have $\mathbb{E}\left[W_iW_j(-1)^{T_i+T_j}\epsilon_i\epsilon_j|T_{1:n}, X_{1:n}\right] = W_iW_j(-1)^{T_i+T_j}\mathbb{E}\left[\epsilon_i\epsilon_j|T_{1:n}, X_{1:n}\right]$. When $i \neq j$, $W_iW_j(-1)^{T_i+T_j}\mathbb{E}\left[\epsilon_i\epsilon_j|T_{1:n}, X_{1:n}\right] = W_iW_j(-1)^{T_i+T_j}\mathbb{E}\left[\epsilon_i|T_{1:n}, X_{1:n}\right]\mathbb{E}\left[\epsilon_j|T_{1:n}, X_{1:n}\right] = 0$. When $i = j$, $W_iW_j(-1)^{T_i+T_j}\mathbb{E}\left[\epsilon_i\epsilon_j|T_{1:n}, X_{1:n}\right] = W_i^2\sigma_i^2$.

CMSE of $\hat{\tau}^{\text{SATE}}_{\text{W}}$ with respect to SATE

We show that

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\tau}^{\text{SATE}}_{\text{W}} - \text{SATE}\right)^2|X_{1:n}, T_{1:n}\right] = &E^2(W_{1:n}; f_1, f_0) \\ &+ \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i(1) - Y_i(0)|X_i) \\ &+ \frac{1}{n}\sum_{i=1}^{n}W_i(2T_i - 1)\text{Cov}(Y_i, Y_i(1) - Y_i(0)|X_i, T_i)\end{aligned}$$

First, notice that

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\tau}^{\text{SATE}}_{\text{W}} - \text{SATE}\right)^2|X_{1:n}, T_{1:n}\right] = &\mathbb{E}\left[\left(\hat{\tau}^{\text{SATE}}_{\text{W}} - \text{CSATE}\right)^2|X_{1:n}, T_{1:n}\right] \\ &+ \mathbb{E}\left[\left(\text{SATE} - \text{CSATE}\right)^2|X_{1:n}, T_{1:n}\right] \\ &+ \mathbb{E}\left[\left(\hat{\tau}^{\text{SATE}}_{\text{W}} - \text{CSATE}\right)\left(\text{SATE} - \text{CSATE}\right)|X_{1:n}, T_{1:n}\right].\end{aligned}$$

We have already computed the first term in the previous Section. We now compute the second term. Notice that, $\mathbb{E}\left[\left(\hat{\tau}^{\text{SATE}}_{\text{W}} - \text{CSATE}\right)|X_{1:n}, T_{1:n}\right] = \mathbb{E}\left[n^{-1}(\epsilon_{i,1} - \epsilon_{i,0})|X_{1:n}, T_{1:n}\right] = \mathbb{E}\left[n^{-1}\delta_i|X_{1:n}, T_{1:n}\right]$, where $\delta_i = \epsilon_{i,1} - \epsilon_{i,0}$. For each i, j, we have $\mathbb{E}\left[n^{-2}\delta_i\delta_j|X_{1:n}, T_{1:n}\right]$, which is equal to 0 when $i \neq j$. When $i = j$,

$$\begin{aligned}\mathbb{E}\left[\left(\text{SATE} - \text{CSATE}\right)^2|X_{1:n}, T_{1:n}\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\delta_i\right)^2|X_{1:n}, T_{1:n}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\epsilon_{i,1} - \epsilon_{i,0}\right)^2|X_{1:n}, T_{1:n}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i(1) - Y_i(0) \mid X_i)\end{aligned}$$

We now compute the last term:

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\tau}^{\text{SATE}}_{\text{W}} - \text{CSATE}\right)\left(\text{SATE} - \text{CSATE}\right)|X_{1:n}, T_{1:n}\right] \\ &= \frac{1}{n}\mathbb{E}\left[(T_i\epsilon_{i,1} - (1-T_i)\epsilon_{i,0})(\epsilon_{i,1} - \epsilon_{i,0})|X_{1:n}, T_{1:n}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}W_i(2T_i - 1)\text{Cov}(Y_i, Y_i(1) - Y_i(0) \mid X_i, T_i).\end{aligned}$$