Trimmed Density Ratio Estimation

Authored by:

Kenji Fukumizu
Taiji Suzuki
Song Liu
Akiko Takeda

Abstract

Density ratio estimation is a vital tool in both machine learning and statistical community. However, due to the unbounded nature of density ratio, the estimation proceudre can be vulnerable to corrupted data points, which often pushes the estimated ratio toward infinity. In this paper, we present a robust estimator which automatically identifies and trims outliers. The proposed estimator has a convex formulation, and the global optimum can be obtained via subgradient descent. We analyze the parameter estimation error of this estimator under high-dimensional settings. Experiments are conducted to verify the effectiveness of the estimator.

1 Paper Body

Density ratio estimation (DRE) [18, 11, 27] is an important tool in various branches of machine learning and statistics. Due to its ability of directly modelling the differences between two probability density functions, DRE finds its applications in change detection [13, 6], two-sample test [32] and outlier detection [1, 26]. In recent years, a sampling framework called Generative Adversarial Network (GAN) (see e.g., [9, 19]) uses the density ratio function to compare artificial samples from a generative distribution and real samples from an unknown distribution. DRE has also been widely discussed in statistical literatures for adjusting non-parametric density estimation [5], stabilizing the estimation of heavy tailed distribution [7] and fitting multiple distributions at once [8]. However, as a density ratio function can grow unbounded, DRE can suffer from robustness and stability issues: a few corrupted points may completely mislead the estimator (see Figure 2 in Section 6 for example). Considering a density ratio $p(x)/q(x)$, a point x that is extremely far away from the high density region of q may have an almost infinite ratio value and DRE results can be dominated by such points. This makes DRE performance very sensitive to rare pathological data or small modifications of the dataset. Here we give two examples:

Cyber-attack In change detection applications, a density ratio p(x)/q(x) is used to determine how the data generating model differs between p and q. Consider a ?hacker? who can spy on our data may just inject a few data points in p which are extremely far away from the high-density region of q. This would result excessively large p(x)/q(x) tricking us to believe there is a significant change from q(x) to p(x), even if there is no change at all. If the generated outliers are also far away from the ?

This work was done when Song Liu was at The Institute of Statistical Mathematics, Japan

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

high density region of p(x), we end up with a very different density ratio function and the original parametric pattern in the ratio is ruined. We give such an example in Section 6. Volatile Samples The change of external environment may be responded in unpredictable ways. It is possible that a small portion of samples react more ?aggressively? to the change than the others. These samples may be skewed and show very high density ratios, even if the change of distribution is relatively mild when these volatile samples are excluded. For example, when testing a new fertilizer, a small number of plants may fail to adapt, even if the vast majority of crops are healthy. Overly large density ratio values can cause further troubles when the ratio is used to weight samples. For example, in the domain adaptation setting, we may reweight samples from one task and reuse them in another task. Density ratio is a natural choice of such ?importance weighting? scheme [28, 25]. However, if one or a few samples have extremely high ratio, after renormalizing, other samples will have almost zero weights and have little impact to the learning task. Several methods have been proposed to solve this problem. The relative density ratio estimation [33] estimates a ?biased? version of density ratio controlled by a mixture parameter ?. The relative density ratio is always upper-bounded by ?1 , which can give a more robust estimator. However, it is not clear how to de-bias such an estimator to recover the true density ratio function. [26] took a more direct approach. It estimates a thresholded density ratio by setting up a tolerance t to the density ratio value. All likelihood ratio values bigger than t will be clipped to t. The estimator was derived from Fenchel duality for f -divergence [18]. However, the optimization for the estimator is not convex if one uses log-linear models. The formulation also relies on the non-parametric approximation of the density ratio function (or the log ratio function) making the learned model hard to interpret. Moreover, there is no intuitive way to directly control the proportion of ratios that are thresholded. Nonetheless, the concept studied in our paper is inspired by this pioneering work. In this paper, we propose a novel method based on a ?trimmed Maximum Likelihood Estimator? [17, 10]. This idea relies on a specific type of density ratio estimator (called log-linear KLIEP) [30] which can be written as a maximum likelihood formulation. We simply ?ignore? samples that make the empirical likelihood take exceedingly large values. The trimmed density ratio estimator can be formulated as a convex optimization and translated into a weighted M-estimator. This helps us develop a simple

subgradient-based algorithm that is guaranteed to reach the global optimum. Moreover, we shall prove that in addition to recovering the correct density ratio under the outlier setting, the estimator can also obtain a ?corrected? density ratio function under a truncation setting. It ignores ?pathological? samples and recovers density ratio only using ?healthy? samples. Although trimming will usually result a more robust estimate of the density ratio function, we also point out that it should not be abused. For example, in the tasks of two-sample test, a diverging density ratio might indicate interesting structural differences between two distributions. In Section 2, we explain some preliminaries on trimmed maximum likelihood estimator. In Section 3, we introduce a trimmed DRE. We solve it using a convex formulation whose optimization procedure is explained in Section 4. In Section 5, we prove the estimation error upper-bound with respect to a sparsity inducing regularizer. Finally, experimental results are shown in Section 6 and we conclude our work in Section 7.

2

Preliminary: Trimmed Maximum Likelihood Estimation

Although our main purpose is to estimate the density ratio, we first introduce the basic concept of trimmed estimator using density functions as examples. Given n samples drawn from a distribution

n i.i.d. P , i.e., X := x(i) i=1 ? P, x ? Rd , we want to estimate the density function p(x). Suppose the true density function is a member of exponential family [20], Z p(x; ?) = exp [h?, f (x)i ? log Z(?)] , Z(?) = q(x) exph?, f (x)idx (1) where f (x) is the sufficient statistics, Z(?) is the normalization function and q(x) is the base measure. Maximum Likelihood Estimator (MLE) maximizes the empirical likelihood over the entire dataset. In contrast, a trimmed MLE only maximizes the likelihood over a subset of samples according to 2

their likelihood values (see e.g., [10, 31]). This paradigm can be used to derive a popular outlier detection method, one-class Support Vector Machine (one-SVM) [24]. The derivation is crucial to the development of our trimmed density ratio estimator in later sections. Without loss of generality, we can set the log likelihood function as log p(x(i) ; ?) ? ?0 , where ?0 is a constant. As samples corresponding to high likelihood values are likely to be inliers, we can trim all samples whose likelihood is bigger than ?0 using a clipping function [?]? , i.e., ? = arg max? Pn [log p(x(i) ; ?) ? ?0]? , where [‘]? returns ‘ if ‘ ? 0 and 0 otherwise. This ? i=1 optimization has a convex formulation:

min h, 1i, s.t. ?i, log p x(i) ; ? ? ?0 ? i , (2) ?,?0

where is the slack variable measuring the difference between log p x(i) ; ? and ?0 . However, formulation (2) is not practical since computing the normalization term Z(?) in (1) is intractable for a general f and it is unclear how to set the trimming level ?0 . Therefore we ignore the normalization term and introduce other control terms: min

?,?0,? ?0

1 1 k?k2 ? ?? + h, 1i s.t. ?i, h?, f (x(i))i ? ? ? i . 2 n

(3)

The ‘2 regularization term is introduced to avoid ? reaching unbounded values. A new hyper parameter ? ? (0, 1] replaces ?0 to control the number of

trimmed samples. It can be proven using KKT conditions that at most 1 ? ? fraction of samples are discarded (see e.g., [24], Proposition 1 for details). Now we have reached the standard formulation of one-SVM. This trimmed estimator ignores the large likelihood values and creates a focus only on the low density region. Such a trimming strategy allows us to discover ?novel? points or outliers which are usually far away from the high density area.

3

Trimmed Density Ratio Estimation

In this paper, our main focus is to derive a robust density ratio estimator following a similar trimming strategy. First, we briefly review the a density ratio estimator [27] from the perspective of KullbackLeibler divergence minimization. 3.1

Density Ratio Estimation (DRE) (1)

i.i.d.

(n)

(1)

(n)

i.i.d.

For two sets of data Xp := {xp , . . . , xp p } ? P, Xq := {xq , . . . , xq q } ? Q, asp(x;?) sume both the densities p(x) and q(x) are in exponential family (1). We know q(x;?pq) ? exp [h? p ? ? q , f (x)i] . Observing that the data x only interacts with the parameter ? p ? ? q through f , we can keep using f (x) as our sufficient statistic for the density ratio model, and merge two parameters ? p and ? q into one single parameter ? = ? p ? ? q . Now we can model our density ratio as Z r(x; ?) := exp [h?, f (x)i ? log N (?)] , N (?) := q(x) exph?, f (x)idx, (4) R where N (?) is the normalization term that guarantees q(x)r(x; ?)dx = 1 so that q(x)r(x; ?) is a valid density function and is normalized over its domain. Interestingly, despite the parameterization (changing from ? to ?), (4) is exactly the same as (1) where q(x) appeared as a base measure. The difference is, here, q(x) is a density function from which Xq are drawn so that N (?) can be approximated accurately from samples of Q. Let us define nq h i h i X b (?) , N b (?) := 1 exp h?, f (xq(j))i . r?(x; ?) := exp h?, f (x)i ? log N nq j=1

(5)

Note this model can be computed for any f even if the integral in N (?) does not have a closed form . 3

In order to estimate ?, we minimize the Kullback-Leibler divergence between p and q ? r? : Z Z p(x) min KL [p—q ? r?] = min p(x) log dx = c ? max p(x) log r(x; ?)dx ? ? ? q(x)r(x; ?) np X 1 log r?(x(i) ? c ? max p ; ?) ? np i=1

(6)

where c is a constant irrelevant to ?. It can be seen that the minimization of KL divergence boils down to maximizing log likelihood ratio over dataset Xp . Now we have reached the log-linear Kullback-Leibler Importance Estimation Procedure (log-linear KLIEP) estimator [30, 14]. 3.2

Trimmed Maximum Likelihood Ratio

As stated in Section 1, to rule out the influences of large density ratio, we trim samples with large likelihood ratio values from (6). Similarly to one-SVM in (2), we can consider a trimmed MLE Pnp (i) ?? = arg max? i=1 [log r?(xp ; ?) ? t0]? where t0 is a threshold above which the likelihood ratios are ignored. It has a convex formulation: min h, 1i, s.t. ?x(i) ?(x(i) p ? Xp , log r p ; ?) ? t0 ? i .

?,?0

(7)

(7) is similar to (2) since we have only replaced p(x; ?) with r?(x; ?). However, the ratio model ? while the normalization term Z in p(x; ?) r?(x; ?) in (7) comes with a tractable normalization term N is in general intractable. Similar to (3), we can directly control the trimming quantile via a hyper-parameter ?: 1 h, 1i ? ? ? t + ?R(?), s.t. ?xp(i) ?Xp , log r?(x(i) p ; ?) ? t ? i ?,?0,t?0 np min

(8)

where R(?) is a convex regularizer. (8) is also convex, but it has np number of non-linear constraints and the search for the global optimal solution can be time-consuming. To avoid such a problem, one could derive and solve the dual problem of (8). In some applications, we rely on the primal parameter structure (such as sparsity) for model interpretation, and feature engineering. In Section 4, we translate (8) into an equivalent form so that its solution is obtained via a subgradient ascent method which is guaranteed to converge to the global optimum. One common way to construct a convex robust estimator is using a Huber loss [12]. Although the proposed trimming technique rises from a different setting, it shares the same guiding principle with Huber loss: avoid assigning dominating values to outlier likelihoods in the objective function. In Section 8.1 in the supplementary material, we show the relationship between trimmed DRE and binary Support Vector Machines [23, 4].

4

Optimization

The key to solving (8) efficiently is reformulating it into an equivalent max min problem. Proposition 1. Assuming ? is chosen such that t? ¿ 0 for all optimal solutions in (8), then ?? is an optimal solution of (8) if and only if it is also the optimal solution of the following max min problem: max ?

min h inp w? 0, n1p ,h1,wi=?

L(?, w) ? ?R(?), L(?, w) :=

np X

wi ? log r?(x(i) p ; ?).

(9)

i=1

? w) ? as a saddle point of (9): The proof is in Section 8.2 in the supplementary material. We define (?, ? w) ? = 0, w ? ? ?? ?R(?) ? ? arg ?? L(?,

min

w?[0, n1p]np ,hw,1i=?

where the second ?? means the subgradient if R is sub-differentiable. 4

? w), L(?,

(10)

Algorithm 1 Gradient Ascent and Trimming max Input: Xp , Xq , ? and step sizes {?it }it it=1 ; Initialize ? 0 , w 0 , Iteration counter: it = 0, Maximum number of iterations: itmax , Best objective, parameter pair (Obest = ??, ? best , wbest) . while not converged and n ito? itmax do

(i)

Obtain a sorted set xp 1 np , ?i

np

i=1

(1)

(2)

(np)

so that log r?(xp ; ? it) ? log r?(xp ; ? it) ? ? ? ? log r?(xp ; ? it).

wit+1,i = ? ?np . wit+1,i = 0, otherwise. Gradient ascent with respect to ?: ? it+1 = ? it + ?it ? ?? [L(? it , wit+1) ? ?R(? it)], Obest = max(Obest , L(? it+1 , wit+1)) and update (? best , wbest) accordingly. it = it + 1. end while Output: (? best , wbest)

Now the ?trimming? process of our estimator can be clearly seen from (9): The max procedure estimates a density ratio given the currently assigned weights w, and the min procedure trims the large log likelihood ratio values by assigning corresponding wi to 0 (or values smaller than n1p). For simplicity, we only consider the cases where ? is a multiple of n1p . Intuitively, 1 ? ? is the proportion of likelihood ratios that are trimmed thus ? should not be greater than 1. Note if we set ? = 1, (9) is equivalent to the standard density ratio estimator (6). Downweighting outliers while estimating the model parameter ? is commonly used by robust estimators (See e.g., [3, 29]). ? w) ? is straightforward. It is easy to solve with respect to w or ? while the other The search for (?, is fixed: given a parameter ?, the optimization with respect to w is a linear programming and one of the extreme optimal solutions is attained by assigning weight n1p to the elements that correspond to the ?np -smallest log-likelihood ratio log r?(x(i) , ?). This observation leads to a simple ?gradient ascent and trimming? algorithm (see Algorithm 1). In Algorithm 1, ?? L(?, w) =

np nq X 1 X e(j) (i) Pnq (k) f (x(j) := exp(h?, f (x(i) wi ? f (x(i)) ? ? ? q), e q)i). p np i=1 e k=1 j=1

In fact, Algorithm 1 is a subgradient method [2, 16], since the optimal value function of the inner problem of (9) is not differentiable at some ? where the inner problem has multiple optimal solutions. The subdifferential of the optimal value of the inner problem with respect to ? can be a set but Algorithm 1 only computes a subgradient obtained using the extreme point solution wit+1 of the inner linear programming. Under mild conditions, this subgradient ascent approach will converge to optimal results with diminishing step size rule and it ? ?. See [2] for details. Algorithm 1 is a simple gradient ascent procedure and can be implemented by deep learning softwares such as Tensorflow2 which

benefits from the GPU acceleration. In contrast, the original problem (8), due to its heavily constrained nature, cannot be easily programmed using such a framework.

5

Estimation Consistency in High-dimensional Settings

In this section, we show how the estimated parameter ?? in (10) converges to the ?optimal parameters? ? ? as both sample size and dimensionality goes to infinity under the ?outlier? and ?truncation? setting respectively. In the outlier setting (Figure 1a), we assume Xp is contaminated by outliers and all ?inlier? samples in Xp are i.i.d.. The outliers are injected into our dataset Xp after looking at our inliers. For example, hackers can spy on our data and inject fake samples so that our estimator exaggerates the degree of change. In the truncation setting, there are no outliers. Xp and Xq are i.i.d. samples from P and Q respectively. However, we have a subset of ?volatile? samples in Xp (the rightmost mode on histogram in Figure 1b) that are pathological and exhibit large density ratio values. 2

https://www.tensorflow.org/

5

(a) Outlier Setting. Blue and red points are i.i.d.

(b) Truncation Setting. There are no outliers.

Figure 1: Two settings of theoretical analysis. In the theoretical results in this section, we focus on analyzing the performance of our estimator for high-dimensional data assuming the number of non-zero elements in the optimal ? ? is k and ? The proofs rely on a recent use the ‘1 regularizer, i.e., R(?) = k?k1 which induces sparsity on ?. development [35, 34] where a ?weighted? high-dimensional estimator was studied. We also assume the optimization of ? in (9) was conducted within an ‘1 ball of width ?, i.e., Ball(?), and ? is wisely chosen so that the optimal parameter ? ? ? Ball(?). The same technique was used in previous works [15, 35]. Notations: We denote w? ? Rnp as the ?optimal? weights depending on ? ? and our data. To lighten the notation, we shorten the log density ratio model as z? (x) := log r(x; ?), z?? (x) := log r?(x; ?) The proof of Theorem 1, 2 and 3 can be found in Section 8.4, 8.5 and 8.6 in supplementary materials. 5.1

A Base Theorem

Now we provide a base theorem giving an upperbound of k?? ? ? ? k. We state this theorem only with respect to an arbitrary pair (? ? , w?) and the pair is set properly later in Section 5.2 and 5.3. We make a few regularity conditions on samples from Q. Samples of Xq should be well behaved in terms of log-likelihood ratio values. Assumption 1. ?0 ¡ c1 ¡ 1, 1 ¡ c2 ¡ ? ?xq ? Xq , u ? Ball(?), c1 ? exph? ? + u, xq i ? c2 and collectively c2 /c1 = Cr . We also assume the Restricted Strong Convexity (RSC) condition on the covariance of X q , i.e., cov(X q) = n1q (X q ? n1q X q 1)(X q ? n1q X q 1)¿ . Note this property has been verified for various different design matrices X q , such as Gaussian or sub-Gaussian (See, e.g., [21, 22]). Assumption 2. RSC condition of cov(X q) holds for all u, i.e., there exists ?01 ¿ 0 and c ¿ 0 such that u¿ cov(X q)u ? ?01 kuk2 ? ?cnq kuk21 with high probability. Theorem 1. In addition to

Assumption 1 and 2, there exists coherence between parameter w and ? ? w): ? at a saddle point (?, ? w) ? w?), u ? ? ?? L(?, ? i ? ??2 k? h?? L(?, uk2 ? ?2 (n, d)k? uk1 , (11) ? := ?? ? ? ? , ?2 ¿ 0 is a constant and ?2 (d, n) ¿ 0. It can be shown that if where u h i ?n ? 2 max k?? L(? ? , w?)k? , 2C??c 2 ?n , ?2 (n, d) q r

??01 ?

and k?? ? ?

¿ 2Cr2 ?2 , where c ¿ 0 is a constant determined by RSC condition, ? Cr2 3 k?n k ? (??0 ?2C with probability converging to one. ? 2 2 r ?2) 1

we are guaranteed that

? for w? , the change of the gradient ?? L is limited. The condition (11) states that if we swap w Intuitively, it shows that our estimator (9) is not ?picky? on w: even if we cannot have the optimal ? to compute the gradient which is weight assignment w? , we can still use ?the next best thing?, w close enough. We later show how (11) is satisfied. Note if k?? L(? ? , w?)k? , ?2 (n, d) converge to ? In Section zero as np , nq , d ? ?, by taking ?n as such, Theorem 1 guarantees the consistency of ?. ? ? 5.2 and 5.3, we explore two different settings of (? , w?) that make ——?? ? ? k converges to zero. 6

5.2

Consistency under Outlier Setting

Setting:

Suppose dataset Xp is the union of two disjoint sets G (Good points) and B (Bad points) i.i.d.

(j)

(i)

i.i.d.

such that G ? p(x) and minj?B z?? (xp) ¿ maxi?G z?? (xp) (see Figure 1a). Dataset Xq ? ? q(x) does not contain any outlier. We set ? = —G— np . The optimal parameter ? is set such that p(x) = q(x)r(x; ? ?). We set w?i =

(i) 1 np , ?xp

?G and 0 otherwise.

Remark: Knowing the inlier proportion —G—/np is a strong assumption. However it is only imposed for theoretical analysis. As we show in Section 6, our method works well even if ? is only a rough guess (like 90%). Loosening this assumption will be an important future work. Assumption 3. ?u ? Ball(?), supx —? z ?? +u (x) ? z??? (x)— ? Clip kuk1 . This assumption says that the log density ratio model is Lipschitz continuous around its optimal parameter ? ? and hence there is a limit how much a log ratio model can deviate from the optimal model under a small perturbation u. As our estimated weights w ?i depends on the relative ranking of (i) z??? (xp), this assumption implies that the relative ranking between two points will remain unchanged under a small perturbation u if they are far apart. The following theorem shows that if we have enough clearance between ?good?and ?bad samples?, ?? converges to the optimal parameter ? ? . Theorem 2. In addition to Assumption 1, 2 and a few mild technical conditions (see Section 8.5 in the (j) (i) ? ? supplementary

material), Assumptions 3 holds. Suppose q minj?B z? (x p) ? maxi?G z? (xp) ? 3Clip ?, ? =

—G— np , nq

constants, we are guaranteed that ——?? ? ? ? k ? It can be seen that k?? ? ? ? k = O 5.3

K1 log d ??c ? —G— , 2Cr2 nq

= ?(—G—2). If ?n ? 2 ? max

p

Cr2 ??01

, where K1 ¿ 0, c ¿ 0 are ? ? 3 k?n with probability converging to 1.

log d/min(—G—, nq) if d is reasonably large.

Consistency under Truncation Setting

In this setting, we do not assume there are outliers in the observed data. Instead, we examine the ability of our estimator recovering the density ratio up to a certain quantile of our data. This ability is especially useful when the behavior of the tail quantile is more volatile and makes the standard estimator (6) output unpredictable results. Notations: Given ? ? (0, 1], we call t? (?) is the ?-th quantile of z? if P [z? ¡ t? (?))] ? ? and P [z? ? t? (?))] ? ?. In this setting, we consider ? is fixed by a user thuswe drop the subscript ? from all subsequent discussions. Let?s define a truncated domain: X(?) = x ? Rd —z? (x) ¡ t(?) , p q X (?) = Xp ? X(?) and X (?) = Xq ? X(?). See Figure 1b for a visualization of t(?) and X(?) (the dark shaded region). i.i.d.

i.i.d.

Setting: Suppose dataset Xp ? P and Xq ? Q. Truncated densities p? and q ? are the unbounded densities p and q restricted only on the truncated domain X(?). Note that the truncated densities are dependent on the parameter ? and ?. We show that under some assumptions, the parameter ?? obtained from (9) using a fixed hyperparameter ? will converge to the ? ? such that (i) q ?? (x)r(x; ? ?) = p?? (x). We also define the ?optimal? weight assignment wi? = n1p , ?i, xp ? X(? ?) and 0 otherwise. Interestingly, the constraint in (9), hw? , 1i = ? may not hold, but our ? w) ? in the feasible region so that analysis in this section suggests we can always find a pair (?, ? ? k? ? ? k converges to 0 under mild conditions. We first assume the log density ratio model and its CDF is Lipschitz continuous. Assumption 4. ?u ? Ball(?), sup —? z ?? +u (x) ? z??? (x)— ? Clip kuk. x

7

(12)

Define T (u,) := x ? Rd — —z?? (x) ? t(? ?)— ? 2Clip kuk + where 0 ¡ ? 1. We assume ?u ? Ball(?), 0 ¡ ? 1 P [xp ? T (u,)] ? CCDF ? kuk + . In this assumption, we define a ?zone? T (u,) near the ?-th quantile t(? ?) and assume the CDF of our ratio model is upper-bounded over this region. Different from Assumption 3, the RHS of (12) is with respect to ‘2 norm of u. In the following assumption, we assume regularity on P and Q. Assumption 5. ?xq ? Rd , kf (xq)k? ? Cq and ?u ? Ball(?), ?xp ? T (u, 1), kf (xp)k? ? Cp . (see Section 8.6 in the Theorem 3. In addition Assumption 1 and 2 and other mild assumptions ? 8CCDF kCp Cr2 supplementary material), Assumption 4

and 5 hold. If 1 ? ? ? , nq = ?(—X p (? ?)—2), ?01 i hq 0 q K1 log d 2Cr2 Cq —Xq X (? ?)— 2L?Cp ??c ? , , , ?n ? 2 max + 2 ?n ? p n n 2C q p q —X (?)— r

? 4C 2 where K10 ¿ 0, c ¿ 0 are constants, we are guaranteed that ——?? ? ? ? k ? ??0r ? 3 k?n with high 1 probability. q

p It can be seen that k?? ? ? ? k = O log d/min(—X (? ?)—, nq) if d is reasonably large and q

—Xq X (? ?)—/nq decays fast.

6

Experiments

6.1

Detecting Sparse Structural Changes between Two Markov Networks (MNs) [14]

In the first experiment3 , we learn changes between two Gaussian MNs under the outlier setting. The P ratio between two Gaussian MNs can be parametrized as p(x)/q(x) ? exp(? i,j?d ?i,j xi xj), where ?i,j := ?pi,j ? ?qi,j is the difference between precision matrices. We generate 500 samples as Xp and Xq using two randomly structured Gaussian MNs. One point [10, . . . , 10] is added as an Pd outlier to Xp . To induce sparsity, we set R(?) = i,j=1,i?j —?i,j — and fix ? = 0.0938. Then run DRE and TRimmed-DRE to learn the sparse differential precision matrix ? and results are plotted on Figure 2a and 2b4 where the ground truth (the position i, j, ??i,j 6= 0) is marked by red boxes. It can be seen that the outlier completely misleads DRE while TR-DRE performs reasonably well. We also run experiments with two different settings (d = 25, d = 36) and plot True Negative Rate (TNR) - True Positive Rate (TPR) curves. We fix ? in TR-DRE to 90% and compare the performance of DRE and TR-DRE using DRE without any outliers as gold standard (see Figure 2c). It can be seen that the added outlier makes the DRE fail completely while TR-DRE can almost reach the gold standard. It also shows the price we pay: TR-DRE does lose some power for discarding samples. However, the loss of performance is still acceptable. 6.2

Relative Novelty Detection from Images

In the second experiment, we collect four images (see Figure 3a) containing three objects with a textured background: a pencil, an earphone and an earphone case. We create data points from these four images using sliding windows of 48 ? 48 pixels (the green box on the lower right picture on Figure 3a). We extract 899 features using MATLAB HOG method on each window and construct an 899-dimensional sample. Although our theorems in Section 5 are proved for linear models, here f (x) is an RBF kernel using all samples in Xp as kernel basis. We pick the top left image as Xp and using all three other images as Xq , then run TR-DRE, THresholded-DRE [26], and one-SVM. In this task, we select high density ratio super pixels on image Xp . It can be expected that the super pixels containing the pencil will exhibit high density ratio values as they did not appear in the reference dataset Xq while super pixels containing the earphone case, the earphones and the background, repeats similar patches in Xq will have lower density ratio values. This is different from 3 4

Code can be found at http://allmodelsarewrong.org/software.html Figures are best viewed in color.

8

1

0.8

TNR

0.6

0.4

0.2

0 0

0.2

0.4

0.6

0.8

1

TPR

? obtained by DRE, d = 20, with(b) ? ? obtained by TR-DRE, ? = (a) ? one outlier. 90%, with one outlier.

(c) TNR-TPR plot, ? = 90%

Figure 2: Using DRE to learn changes between two MNs. We set R(?) = k ? k1 and f (xi , xj) = xi xj .

(a) Dataset

(b) ? = 97%

(c) ? = 90%

(d) ? = 85%

(e) TH-DRE

(f) one-SVM

2

Figure 3: Relative object detection using super pixels. We set R(?) = k ? k , f (x) is an RBF kernel.

a conventional novelty detection, as a density ratio function help us capture only the relative novelty. For TR-DRE, we use the trimming threshold t? as the threshold for selecting high density ratio points. It can be seen on Figure 3b, 3c and 3d, as we tune ? to allow more and more high density ratio windows to be selected, more relative novelties are detected: First the pen, then the case, and finally the earphones, as the lack of appearance in the reference dataset Xq elevates the density ratio value by different degrees. In comparison, we run TH-DRE with top 3% highest density ratio values thresholded, which corresponds to ? = 97% in our method. The pattern of the thresholded windows (shaded in red) in Figure 3e is similar to Figure 3b though some parts of the case are mistakenly shaded. Finally, one-SVM with 3% support vectors (see Figure 3f) does not utilize the knowledge of a reference dataset Xq and labels all salient objects in Xp as they corresponds to the ?outliers? in Xp .

7

Conclusion

We presents a robust density ratio estimator based on the idea of trimmed MLE. It has a convex formulation and the optimization can be easily conducted using a subgradient ascent method. We also investigate its theoretical property through an equivalent weighted M-estimator whose ‘2 estimation error bound was provable under two high-dimensional, robust settings. Experiments confirm the effectiveness and robustness of the our trimmed estimator.

Acknowledgments We thank three anonymous reviewers for their detailed and helpful comments. Akiko Takeda thanks Grant-in-Aid for Scientific Research (C), 15K00031. Taiji Suzuki was partially supported by MEXT KAKENHI (25730013, 25120012, 26280009 and 15H05707), JST-PRESTO and JST-CREST. Song Liu and Kenji Fukumizu have been supported in part by MEXT Grant-in-Aid for Scientific Research on Innovative Areas (25120012). 9

2 References

[1] F. Azmandian, J. G. Dy, J. A. Aslam, and D. R. Kaeli. Local kernel density ratio-based feature selection for outlier detection. In Proceedings of 8th Asian Conference on Machine Learning (ACML2012), JMLR Workshop and Conference Proceedings, pages 49?64, 2012. [2] S. Boyd. Subgradient methods. Technical report, Stanford University, 2014. Notes for EE364b, Stanford University, Spring 2013?14. [3] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74(368):829?836, 1979. [4] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000. [5] B. Efron and R. Tibshirani. Using specially designed exponential families for density estimation. The Annals of Statistics, 24(6):2431?2461, 1996. [6] F. Fazayeli and A. Banerjee. Generalized direct change estimation in ising model structure. In Proceedings of The 33rd International Conference on Machine Learning (ICML2016), page 2281?2290, 2016. [7] W. Fithian and S. Wager. Semiparametric exponential families for heavy-tailed data. Biometrika, 102(2):486?493, 2015. [8] K. Fokianos. Merging information for semiparametric density estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(4):941?958, 2004. [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672?2680, 2014. [10] A. S. Hadi and A. Luceno. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. Computational Statistics & Data Analysis, 25(3):251 ? 272, 1997. [11] J. Huang, A. Gretton, K. M Borgwardt, B. Sch?lkopf, and A. J Smola. Correcting sample selection bias by unlabeled data. In Advances in neural information processing systems, pages 601?608, 2007. [12] P. J. Huber. Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1):73?101, 03 1964. [13] Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. Statistical Analysis and Data Mining, 5(2):114?127, 2012. [14] S. Liu, T. Suzuki, R. Relator, J. Sese,

M. Sugiyama, and K. Fukumizu. Support consistency of direct sparse-change learning in Markov networks. Annals of Statistics, 45(3):959?990, 06 2017. [15] P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Journal of Machine Learning Research, 16:559?616, 2015. [16] A. Nedi?c and A. Ozdaglar. Subgradient methods for saddle-point problems. Journal of Optimization Theory and Applications, 142(1):205?228, 2009. [17] N. Neykov and P. N. Neytchev. Robust alternative of the maximum likelihood estimators. COMPSTAT?90, Short Communications, pages 99?100, 1990. [18] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 56(11):5847?5861, 2010. [19] S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In Advances in Neural Information Processing Systems, pages 271?279, 2016. 10

[20] E. J. G. Pitman. Sufficient statistics and intrinsic accuracy. Mathematical Proceedings of the Cambridge Philosophical Society, 32(4):567?579, 1936. [21] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. Journal of Machine Learning Research, 11:2241?2259, 2010. [22] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. IEEE Transactions on Information Theory, 59(6):3434?3447, 2013. [23] B. Scholkopf and A. J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001. [24] B. Sch?lkopf, R. C. Williamson, Smola A. J., Shawe-Taylor J., and Platt J.C. Support vector method for novelty detection. In Advances in Neural Information Processing Systems 12, pages 582?588. MIT Press, 2000. [25] A. Shimodaira. Improving predictive inference under covariate shift by weighting the loglikelihood function. Journal of Statistical Planning and Inference, 90(2):227 ? 244, 2000. [26] A. Smola, L. Song, and C. H. Teo. Relative novelty detection. In Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics (AISTATS), volume 5, pages 536?543, 2009. [27] M. Sugiyama, T. Suzuki, and T. Kanamori. Density Ratio Estimation in Machine Learning. Cambridge University Press, 2012. [28] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von B?nau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. Annals of the Institute of Statistical Mathematics, 60(4):699?746, 2008. [29] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle. Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing, 48(1):85?105, 2002. [30] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. Journal of Information Processing, 17:138?155, 2009. [31] D. L. Vandev and N. M. Neykov. About regression estimators with high breakdown point. Statistics: A Journal of Theoretical and Applied Statistics, 32(2):111?129, 1998. [32] M. Wornowizki and R. Fried. Two-sample homogeneity tests based on divergence measures. Computational Statistics, 31(1):291?313, 2016. [33] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for

robust distribution comparison. Neural Computation, 25(5):1324?1370, 2013. [34] E. Yang, A. Lozano, and A. Aravkin. High-dimensional trimmed estimators: A general framework for robust structured estimation. arXiv preprint arXiv:1605.08299, 2016. [35] E. Yang and A. C. Lozano. Robust gaussian graphical modeling with the trimmed graphical lasso. In Advances in Neural Information Processing Systems, pages 2602?2610, 2015.

11