A RELEVANT S-MATCHING CLASSIFIER FOR THE COVARIATE SHIFT MACHINE LEARNING PROBLEM

Yannis G. Yatracos

Cyprus University of Technology

June 23, 2014

Running Title: Relevant S-Matching in covariate shift

Address: School of Management and Economics, Cyprus U. of Technology, P.O. Box 50329, 3603 Lemesos, Cyprus

e-mail: yannis.yatracos@cut.ac.cy

Summary

Matching methodology from causal inference is used to obtain a new, improved classifier for the Covariate Shift Machine Learning Problem. Let $\mathbf{x}$ be the covariate to be y-labeled and let $S(\mathbf{x})$ be the likelihood ratio of the $\mathbf{x}$-covariates' densities in the training and the test populations, that is equivalent to the propensity score. For loss l a classifier δ_s is obtained which minimizes, among all classifiers $d \in D$, l's risk in the training population for covariates with equal S-importance, $E[l(d(\mathbf{x}), y)|S(\mathbf{x}) = s]$. Classifier $\delta = \{\delta_s\}$ reduces the minimum of the unconditional l-risks for *both* populations. If $S(\mathbf{x}_1) = \tilde{s}$, $\delta_{\tilde{s}}$ labels $\mathbf{x}_1$ and is used for $\mathbf{x}_2$ when they S-match, i.e. $S(\mathbf{x}_1) \approx S(\mathbf{x}_2)$. When samples are available, S is used to group learning covariates *relevant* to the test data and obtain a classifier. The same holds with more than one learning populations or samples and the minimal sufficient statistic, or equivalently ratios of generalised propensity scores, allow for the fusion of learning data having covariates with equal importance.

Some key words: Covariate Shift Problem, Machine Learning, Matching, Minimal Sufficient Statistic

1 Introduction

In Machine Learning (ML), the $\mathbf{x}$-covariate and its y-label may have different joint probability distributions in the learning (called also training) and the test populations. This situation occurs often in practice and is studied, among others, in the covariate shift problem (Bickel *et al.*, 2007), sample selection bias (Zadrozny, 2004), domain adaptation (Daumé and Marcu, 2006) and distance ML (Cao *et al.*, 2009). For any loss l, the classifier minimizing, over a collection of classifiers of interest $d \in \mathcal{D}$ the (statistical) risk $El(d(\mathbf{x}), y)$ in the training population *may not be* the risk's minimizer over $\mathcal{D}$ in the test population; see, for example, Bickel *et al.* (2007, section 2). An additional problem not addressed is whether the whole learning population or its available subset (called learning "data" or sample) are relevant when obtaining a classifier for the test population or the test data.

Both problems are solved herein using tools from causal inference, the minimal sufficient statistic S or equivalently ratios of generalized propensity scores (Yatracos, 2011), to identify relevant "matching" groups of $\mathbf{x}$-covariates; $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to the same matching group when $S(\mathbf{x}_1) \approx S(\mathbf{x}_2)$. Due to sufficiency, the conditional risks on each S-matching group coincide for the learning and the test distributions of $(\mathbf{x}, y)$ and are minimized by the *same* classifier that is used to predict the y-label in the test population. When D consists, for

example, of linear classifiers in x, the classifier obtained herein via matching will consist of piecewise linear classifiers, one for each matching group. This approach solves directly the problem of obtaining the same piecewise classifier both for the training and the test populations and reduces the mean square error.

The above description is now supplemented with the comments of a reader trained in ML: "the basic idea of the paper seems sensible: learn localized models for different regions of the input space, as defined by similarity to the test distribution. Then, pick the appropriate model for a given test example, and use this to make a prediction."

Previously, $S(\mathbf{x})$ was used as weight to adjust the log-likelihood function for covariate shift and improve predictive inference (Shimodaira, 2000, p. 231); $\mathbf{x}$-covariates with the same S-value have equal "importance" (see, for example, Shimodaira, 2000 or Zadrozny, 2004). More recently, S has been used to adjust loss function l to randomized $l^* = Sl$ and obtain the same optimal classifier in the l^*-risk and l-risk minimization problems (Bickel *et al*, 2007).

In applications with learning and test data, conditional risk minimization via S allows for the use of learning data relevant to the test data and reduces potential sampling bias as well as the intensity of the optimization problem when the sizes of the training data and the covariates' dimension are large.

With k (> 1) learning $(\mathbf{x}, y)$-populations and the test population, the use of Shimodaira's S factor is not possible unless the mixture distribution of the learning populations is available. In this case, for several related "tasks", i.e. parameters in the densities of the learning populations, Bickel *et al.* (2009) provided for task t the Shimodaira-type weight $r_t(\mathbf{x}, y)$ and its estimate, in order to "train a hypothesis for task t by minimizing the expected loss over the distributions of all tasks", i.e. for the learning mixture distribution. It is seen herein that r_t is the minimal sufficient statistic for the test distribution of task t and the learning mixture distribution.

When the learning mixture distribution is unknown but the covariate shift distributional assumption holds for the $(k+1)$ populations, the k-dimensional minimal sufficient statistic S is used to obtain matching groups of $\mathbf{x}$-covariates and the corresponding classifiers. With learning samples, conditional risk minimization on S-matched groups pooled together from *all* learning populations is used to obtain the corresponding classifiers as in the case $k = 1$. This matching approach has been recently used with multiple treatments ("tasks" in ML), when the data is obtained from an observational study (Yatracos, 2011).

For the interested reader, a recent review on matching, propensity scores and causal inference is presented in Stuart (2010). In sections 2-4, results are presented for $k = 1$; a brief description of the results for $k > 1$ is in section 5.

The Figures are after the references and the proofs are in the Appendix.

2 The set-up, the assumption and the tool S

In machine learning, the training sample TR consists of the $\mathbf{x}$-covariates ($\in R^p$) with the corresponding y-labels ($\in R$), $TR = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, and the test sample TE consists only of covariates, $TE = \{\mathbf{x}_{m+1}, \ldots, \mathbf{x}_{m+n}\}$. The goal is to predict the y-label for each $\mathbf{x} \in TE$," learning" from the training sample. Without loss of generality, let $f(\mathbf{x}, y|\theta_1)$ and $f(\mathbf{x}, y|\theta_2)$ be the densities of $(\mathbf{x}, y)$, respectively, in the training and the test populations; θ_1, θ_2 are generic parameters that need not be specified, i.e. one can use $f_1(\mathbf{x}, y)$ and $f_2(\mathbf{x}, y)$ instead.

The main distributional assumption in covariate shift is that

$$f(\mathbf{x}, y|\theta_i) = p(\mathbf{x}|\theta_i)q(y|\mathbf{x}),\ i = 1, 2; \tag{1}$$

$p(\mathbf{x}|\theta_1)$ and $p(\mathbf{x}|\theta_2)$ are the densities of the $\mathbf{x}$-covariates, respectively, of the training and the test samples, and q is the conditional density of y given $\mathbf{x}$ that is independent of $\theta_i,\ i = 1, 2$.

When $p(\mathbf{x}|\theta_1)$ and $p(\mathbf{x}|\theta_2)$ have either common support or the support of the test distribution is a subset of the support of the training distribution, the

minimal sufficient statistic

$$S(\mathbf{x}) = \frac{f(\mathbf{x}, y|\theta_2)}{f(\mathbf{x}, y|\theta_1)} = \frac{p(\mathbf{x}|\theta_2)}{p(\mathbf{x}|\theta_1)}, \tag{2}$$

provides all the information for the densities of the covariates $\mathbf{x}$ and of $(\mathbf{x}, y)$. Hence, S can be used to group (i.e. match) covariates from both populations with the same or similar information (or importance) and the so-obtained risk minimizer in this group is identical for the test and training populations (Proposition 3.1).

Assume that any class of classifiers, $\mathcal{D}$, includes also randomized classifiers with form $d = \{d_s,\ s \in \mathcal{S}\}$. For example, $\mathcal{D}$ could consist of local linear classifiers depending on the S-values or not. One classifier δ_s is obtained herein that minimizes the conditional risk $E[l(d(\mathbf{x}), y)|S = s]$ over all classifiers $d \in D$ for *both* populations, due to sufficiency. The randomized classifier $\delta = \{\delta_s \in D, s \in \mathcal{S}\}$, obtained via S and its values $\mathcal{S}$, minimizes the unconditional l-risks in the treatment and test populations and the minimum risk is smaller than that obtained using the same classifier from D for all $\mathbf{x}$. If $\mathbf{x}_1$ is in the treatment population and $\mathbf{x}_2$ in the test population and $S(\mathbf{x}_1) = S(\mathbf{x}_2) = \tilde{s}$, then $\delta_{\tilde{s}}$ labels both $\mathbf{x}_1$ and $\mathbf{x}_2$. When $S(\mathbf{x}_2)$ is not equal to any of the S-values on the training population, approximate matching can be used to obtain $\mathbf{x}_2$'s class label from the training population, at least when the classifier is smooth function of S.

When the **x**-covariates' densities have no common support, see Lehmann and Casella (1998, p. 70, Theorem 9.1) for the minimal sufficient statistic.

Definition 2.1 *With the set-up already presented and m learning and test populations with densities indexed by parameters $\theta_1, \ldots, \theta_m$ which are realizations of a random variable Θ, the generalized propensity score (Imbens, 2000) is the probability*

$$P(\Theta = \theta_i|\mathbf{x}),\ i = 1, \ldots, m. \tag{3}$$

Observe that when $m = 2$,

$$S(\mathbf{x}) = \frac{p(\mathbf{x}|\theta_2)}{p(\mathbf{x}|\theta_1)} = \frac{P(\Theta = \theta_1)}{P(\Theta = \theta_2)} \frac{P(\Theta = \theta_2|\mathbf{x})}{P(\Theta = \theta_1|\mathbf{x})} \propto \frac{P(\Theta = \theta_2|\mathbf{x})}{P(\Theta = \theta_1|\mathbf{x})} \tag{4}$$

which implies due to the equality

$$P(\Theta = \theta_1|\mathbf{x}) + P(\Theta = \theta_2|\mathbf{x}) = 1$$

that the minimal sufficient statistic (2) is equivalent to the propensity score $P(\Theta = \theta_1|\mathbf{x})$ and the latter can be modeled in order to avoid the curse of dimensionality, as it is done in causal inference; see, for example, Rosenbaum and Rubin, 1983, and Stuart, 2010.

3 Large sample theory

The densities and the minimal sufficient statistic $S(\mathbf{x})$ are considered known and the results are applicable for large samples. Let $\mathcal{S}$ denote the set of values s of $S(\mathbf{x})$ for all $\mathbf{x}$-covariates. Following the literature in machine learning, $\mathbf{x}$ and y are used to denote either random variables or their realizations. No assumptions are made on the uniqueness of risk minimizers.

From sufficiency, $p(\mathbf{x}|\theta_i, S(\mathbf{x})) = p(\mathbf{x}|S(\mathbf{x}))$ is independent of $\theta_i,\ i = 1, 2$. E_{θ_i} is used below to denote expected value with respect to a density having θ_i as parameter but there is no dependence on θ_i when conditioning on $S,\ i = 1, 2$.

The propositions that follow help solving the Covariate Shift Machine Learning problem.

Proposition 3.1 *Let $l(d(\mathbf{x}), y)$ be the loss between the classifier d evaluated at $\mathbf{x}$ and y, $\mathbf{x}$'s class label. For any $s \in \mathcal{S}$, a classifier δ_s minimizes both risks*

$$E_{\theta_i}[l(d(\mathbf{x}), y)|S(\mathbf{x}) = s],\ i = 1, 2,$$

over all classifiers in $\mathcal{D}$ and the classifier $\delta = \{\delta_s,\ s \in \mathcal{S}\}$ minimizes

$$E_{\theta_i} l(d(\mathbf{x}), y),\ i = 1, 2.$$

For a given $\mathbf{x}$-covariate for which $S(\mathbf{x}) = \tilde{s}$, its predicted class label is $\delta_{\tilde{s}}(\mathbf{x})$.

Proposition 3.1 shows that with conditional l-risk minimization given $S(x) = s$, the obtained classifier $\delta = \{\delta_s,\ s \in \mathcal{S}\}$ minimizes the unconditional global l-risk in the learning and test populations.

We now relate risk minimization of the "loss" $l^* = Sl$ with that of l.

Proposition 3.2 *The risks $E_{\theta_i} l(d(\mathbf{x}), y),\ i = 1, 2$, and the scale adjusted Shimodaira's risk $E_{\theta_1} S(\mathbf{x}) l(d(\mathbf{x}), y)$ are all minimized over $\mathcal{D}$ by the same classifier $\delta = \{\delta_s,\ s \in \mathcal{S}\}$; δ_s minimizes conditional risks $E_{\theta_i}[l(d(\mathbf{x}), y)|S],\ i = 1, 2$.*

4 Some small sample theory-An Example

When $p(\mathbf{x}|\theta),\ \theta \in \tilde{\Theta}$, belong to the same family parametrized by θ, known theorems in statistics (for example, in Lehmann and Casella, 1998, or Dawid, 2011) allow to obtain the minimal sufficient statistic S. For independent $\mathbf{x}$-samples from nonparametric models, one for each $\theta \in \tilde{\Theta}$, the class of empirical distributions is minimal sufficient statistic S. In both situations S is used for S-conditional minimization of the training sample and for matching various $\mathbf{x}$-covariates and predicting their labels. When a covariate in the test sample cannot be matched exactly with those in the training sample, approximate S-matching methods can be used, for example, nearest neighbor matching, sub-classification, full matching and weighting described in Stuart (2010) .

In the example that follows, the densities of the learning sample and the test sample are assumed to be known for the S-matching.

Example 4.1 *The learning x-covariates are 100 i.i.d. observations $x_1, x_2, \ldots, x_{100}$ from the normal distribution, $p_1(x) = N(3, 4)$, with mean 3 and variance 4. Obtain 100 random i.i.d. errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{100}$ from a standard normal, $N(0, 1)$, distribution. The learning y-labels are*

$$y_i = 1.2 + 3x_i + 0.8x_i^2 + \varepsilon_i, \ i = 1, 2, \ldots, 100,$$

and will be approximated conditionally and unconditionally with linear classifiers.

The test x^*-covariates are 20 i.i.d. observations $x_1^*, x_2^*, \ldots, x_{20}^*$ from the normal, $p_2(x) = N(6, .25)$, distribution.

Compute the S-values for the learning and the test covariates; $S(x) = p_2(x)/p_1(x)$. The relevant matched learning data have x-covariates with S-values in the interval of the x^*-covariates' S-values.

In Figure 1, observe the (x, y)-learning data (the dots) and the corresponding linear classifier, i.e. the simple linear least squares regression (OLS) of y on x. Undoubtedly, the OLS-classifier (in bold), based on the relevant matched learning data, provides more accurate y-labels for the test data; the ticks on the x-axis indicate the x^*-covariates. In Figure 2, the S-values for the learning and test covariates can be seen.

5 The k-covariate shift problem

We consider first our proposed setting and then the setting in Bickel et al. (2009). Measurements $(\mathbf{x}, y)$ are available from k learning populations with joint probability densities, respectively, $f(\mathbf{x}, y|\theta_1), \ldots, f(\mathbf{x}, y|\theta_k)$. From the test population $\mathbf{x}$-covariates are only observed, having with y joint probability density $f(\mathbf{x}, y|\theta_{k+1})$. The goal is to find a y-predictor $\delta(\mathbf{x})$ minimizing over $\mathcal{D}$ the risks associated with the populations, and for each of the $\mathbf{x}$ covariates in the test population predict its label y.

Assume that the $\mathbf{x}$-covariates in the test population are included in the supports of $\mathbf{x}$-covariates in the k training populations, and that (1) holds for each of the densities of $(\mathbf{x}, y)$. The minimal sufficient statistic is

$$S(\mathbf{x}) = \left(\frac{p(\mathbf{x}|\theta_2)}{p(\mathbf{x}|\theta_1)}, \ldots, \frac{p(\mathbf{x}|\theta_{k+1})}{p(\mathbf{x}|\theta_1)}\right) \propto \left(\frac{P(\Theta = \theta_2|\mathbf{x})}{P(\Theta = \theta_1|\mathbf{x})}, \ldots, \frac{P(\Theta = \theta_{k+1}|\mathbf{x})}{P(\Theta = \theta_1|\mathbf{x})}\right), \tag{5}$$

with the proportionality obtained as in (4). In Yatracos (2011), it is shown that the matching is not changed when using in (5) as divisor $p(\mathbf{x}|\theta_j)$ instead of $p(\mathbf{x}|\theta_1),\ j \neq 1$, and directions are given for its implementation. From sufficiency,

$$p(\mathbf{x}|\theta_i, S(\mathbf{x})) = p(\mathbf{x}|S(\mathbf{x})),\ i = 1, \ldots, k+1.$$

A loss function l is available and it is desired to obtain $\delta(\mathbf{x})$ that minimizes

over $d \in \mathcal{D}$ the risk $E_{\theta_i} l(d(\mathbf{x}), y)$ for all $i = 1, ..., k+1$. The conditional risks $E_{\theta_i}[l(d(\mathbf{x}), y)|S(\mathbf{x}) = \mathbf{s}]$ coincide, are independent of $\theta_i,\ i = 1, \ldots, k+1$, are all minimized by $\delta_{\mathbf{s}}$ and the risk minimizer $\delta = \{\delta_{\mathbf{s}},\ \mathbf{s} \in \mathcal{S}\}$; $\mathcal{S}$ are the S-values.

S in (5) is k-dimensional but in reality its dimension may be reduced. Think, for example, that when estimating the mean θ of a normal distribution with possible values $\theta_1, \ldots, \theta_k$ and known variance, the minimal sufficient is one-dimensional. The proof, using $\tilde{\Theta}_0 = \{\theta_1, \theta_2\}$ and $\tilde{\Theta} = \{\theta_1, \ldots, \theta_k\}$, is based on a known result (see, e.g., Lehmann and Casella, 1998) according to which if S is minimal sufficient for $\tilde{\Theta}_0 (\subset \tilde{\Theta})$ and sufficient for $\tilde{\Theta}$ then it is also minimal sufficient for $\tilde{\Theta}$.

With small samples, S is obtained via (5) and $\mathbf{x}$-covariates from the k learning populations are pooled in groups determined by their S-values combining the available information in order to solve the corresponding relevant, conditional minimization problems.

The setting in Bickel *et al.*(2009), inspired by the targeted advertising problem, is now described using the terminology in this section. "Transfer learning in which classifiers for multiple tasks have to be learned for biased sample" is the problem. Let $u = 1$ denote "learning" and $u = -1$ denote "test" and let the learning populations have densities $f(\mathbf{x}, y|\theta_1, u = 1), \ldots, f(\mathbf{x}, y|\theta_k, u = 1)$ with $\theta_1, \ldots, \theta_k$ denoting different tasks; densities of the test populations are

similarly denoted but with $u = -1$. The conditional densities of y given $\mathbf{x}$ and θ_i are assumed to be independent of u and it holds

$$f(\mathbf{x}, y|\theta_i, u =^{+}_{-} 1) = p(\mathbf{x}|\theta_i, u =^{+}_{-} 1) \cdot q(y|\mathbf{x}, \theta_i), i = 1, \ldots, k. \tag{6}$$

By pooling the learning tasks populations the learning mixture distribution is

$$f(\mathbf{x}, y|u = 1) = \sum_{j=1}^{k} P(\Theta = \theta_j|u = 1) f(\mathbf{x}, y|\theta_j, u = 1). \tag{7}$$

To determine the test classifier for task θ_i, the densities involved are

$f(\mathbf{x}, y|\theta_i, u = -1),\ f(\mathbf{x}, y|u = 1)$ and the minimal sufficient statistic

$$S_i = \frac{f(\mathbf{x}, y|\theta_i, u = -1)}{f(\mathbf{x}, y|u = 1)} = \frac{f(\mathbf{x}, y|\theta_i, u = 1)}{f(\mathbf{x}, y|u = 1)} \frac{p(\mathbf{x}|\theta_i, u = -1)}{p(\mathbf{x}|\theta_i, u = 1)} = r_{i,1}(\mathbf{x}, y) \cdot r_{i,2}(\mathbf{x}); \tag{8}$$

the second equality in (8) is due to (6), the ratios denoted in Bickel *et al.*(2009), respectively, $r_{i,1}(\mathbf{x}, y)$ and $r_{i,2}(\mathbf{x})$ are resampling weights and it is shown that

$$r_{i,1}(\mathbf{x}, y) \propto P(\Theta = \theta_i|\mathbf{x}, y, u = 1),\ r_{i,2}(\mathbf{x}) \propto \frac{1}{P(u = 1|\mathbf{x}, \theta_i)} - 1,$$

i.e. $r_{i,1}(\mathbf{x}, y)$ and $r_{i,2}(\mathbf{x})$ can be modeled as the propensity score, $i = 1, \ldots, k$.

With the Bickel *et al.*(2009) setting and the approach in this work, the test classifier for task θ_i is obtained with conditional minimization given S_i as described in section 3, for one learning and one test population, $i = 1, \ldots, k$.

Acknowledgment

Many thanks are due to an anonymous referee who brought the covariate shift Machine Learning problem to my attention.

Appendix

Proof of Proposition 3.1. Minimization of the risks

$$E_{\theta_i} l(d(\mathbf{x}), y) = E\{E_{\theta_i}[l(d(\mathbf{x}), y)|S]\},\ i = 1, 2, \tag{9}$$

over $d \in \mathcal{D}$ is equivalent to minimization of the conditional risks

$$E_{\theta_i}[l(d(\mathbf{x}), y)|S = s],\ i = 1,\ 2, \tag{10}$$

for all values of $s \in \mathcal{S}$. Since S is sufficient for the $(\mathbf{x}, y)$-distributions, (10) is independent of θ_i and identical for $i = 1,\ 2$. In the right side of (9), the outer expectation over the S-values depends on θ_i but the inner expectation due to sufficiency is independent of $\theta_i,\ i = 1, 2$.

Therefore, there is a classifier δ_s minimizing both conditional risks in (10) and the classifier $\delta = \{\delta_s,\ s \in \mathcal{S}\}$ minimizes both (unconditional) risks in (9).

□

Proof of Proposition 3.2. Observe that

$$E_{\theta_i} l(d(\mathbf{x}), y) = E[E_{\theta_i}(l(d(\mathbf{x}), y)|S)],\ i = 1, 2. \tag{11}$$

Reproducing Shimodaira's scale factor adjustment calculations we obtain

$$E_{\theta_1} S(\mathbf{x}) l(d(\mathbf{x}), y) = \int \int \frac{p(\mathbf{x}|\theta_2)}{p(\mathbf{x}|\theta_1)} p(\mathbf{x}|\theta_1) q(y|\mathbf{x}) l(d(\mathbf{x}), y) d\mathbf{x} dy$$

$$= \int \int p(\mathbf{x}|\theta_2) q(y|\mathbf{x}) l(d(\mathbf{x}), y) d\mathbf{x} dy = E_{\theta_2} l(d(\mathbf{x}), y),$$

and it holds

$$E_{\theta_1} S(\mathbf{x}) l(d(\mathbf{x}, y)) = ES E_{\theta_1}[l(d(\mathbf{x}), y)|S]. \tag{12}$$

The risks in the left side of equalities (11) and (12) are all minimized when the identical conditional risks $E_{\theta_i}(l(d(\mathbf{x}), y)|S), i = 1, 2$, are minimized.

□

References

[1] Bickel, S., Brückner, M. and Scheffer, T. (2007) Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning.*

[2] Bickel, S., Sawade, C. and Scheffer, T. (2009) Transfer learning by distribution matching for targeted advertising. NIPS.

[3] Daumé III, H. and Marcu, D. (2006) Domain adaptation for statistical classifiers. *J. of Artificial Intelligence Research* **26**, 101-126.

[4] Cao, B., Ni, X., Sun, J.-T., Wang, G. and Yang, Q. (2010) Distance metric learning under covariate shift. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* 1204-1210.

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

[6] Lehmann, E. L. and Casella, G. (1998) *Theory of Point Estimation.* Springer, New York.

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

[8] Shimodaira, H. (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* **90**, 227-244.

[9] Stuart, E. A. (2010) Matching methods for causal inference: A review and a look forward. *Stat. Sci.* **25**, 1-21.

[10] Yatracos, Y. G. (2011) Causal inference for multiple treatments via sufficiency and ratios of generalized propensity scores. *Submitted for publication.*

[11] Zadrozny, B. (2004) Learning and evaluating classifiers under sample selection bias. *Proceedings of the 21st International Conference on Machine Learning*, ACM Press, p. 114-121.

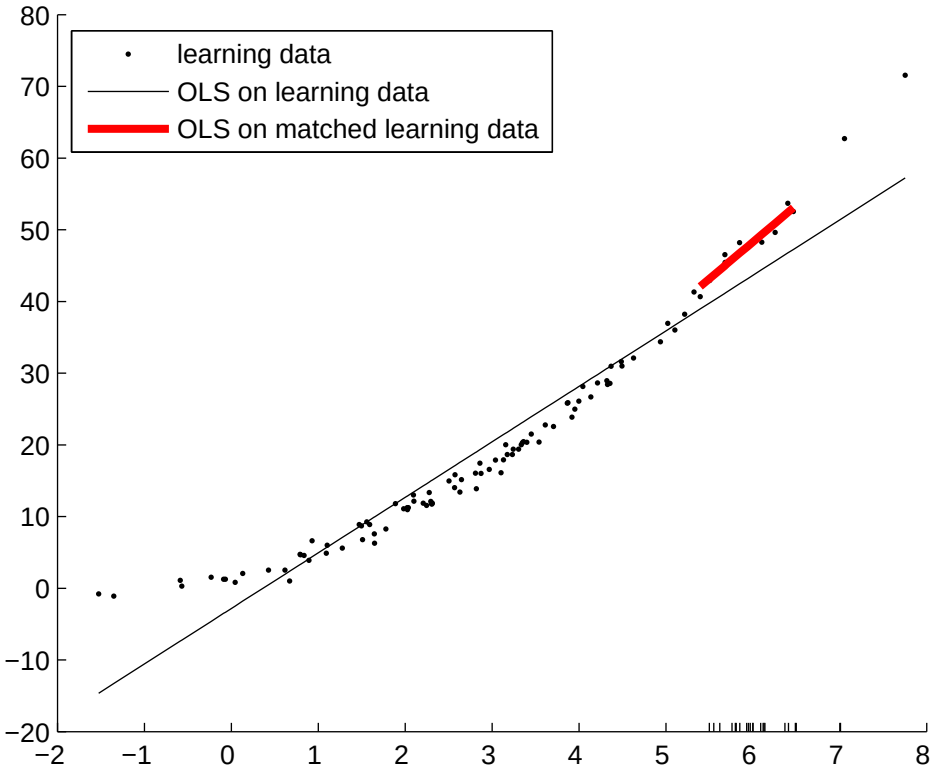

Figure 1
Test data: the ticks on the x−axis

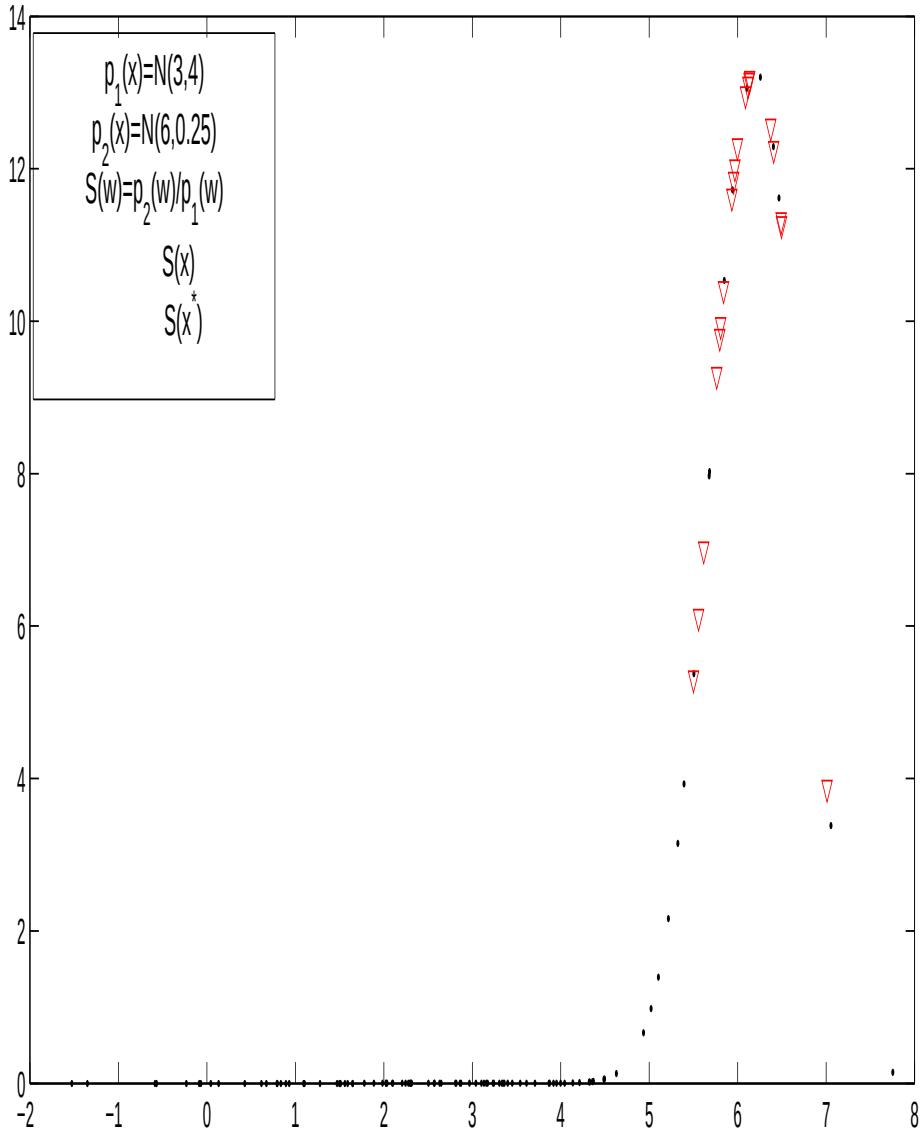

Figure 2

S−values for the x−learning and the x^*−test covariates