Stratified Learning: A general-purpose method for learning under covariate shift with applications to observational cosmology

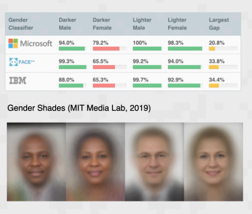
Speaker: Maximilian Autenrieth[1],
Supervisor: David A. van Dyk[1]
Co-Supervisor: Roberto Trotta[1,2], David Stenning[3]

Imperial College London[1], SISSA (Trieste)[2], Simon Fraser University[3]

June 15, 2021

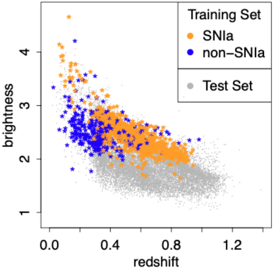
Cases of covariate shift:

- **Natural language processing:** Annotated data (e.g. Wall street journal) is specialized.
- **Computer vision/Facial recognition:** Web-scraped images non-representative
- **Clinical studies/Medical imaging:** Configurations vary between centers.
- **Astronomy:** Follow-up of astronomical sources not at random.



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Gender Shades (MIT Media Lab, 2019)



Definitions and Notation:

- Feature space $\mathcal{X} \subset \mathbb{R}^F$, $F > 0$, and label space $\mathcal{Y}$ the with $K > 1$ classes (or subset of $\mathbb{R}^K$ in multivariate regression case).
- Different domains defined as different joint probability distributions $p(x, y)$ over same feature-label space $\mathcal{X} \times \mathcal{Y}$ (Kouw and Loog 2019).

Transductive, unsupervised domain adaptation:

- Source data: $D_S = \{(x_S^{(i)}, y_S^{(i)})\}_{i=1}^{n_s}$ with n_s labelled samples, from joint distribution $p_S(x, y)$ (Domain $\mathcal{D}_S$),
- Target data: $D_T = \{x_T^{(i)}\}_{i=1}^{n_t}$ with n_t unlabelled samples, from joint distribution $p_T(x, y)$ (Domain $\mathcal{D}_T$).

Definition 1.1 (Moreno-Torres et al. (2012))

Covariate shift is defined as $p_S(y|x) = p_T(y|x)$ but $p_S(x) \neq p_T(x)$.

Notation: $p_S(x, y) := p(x, y|s = 1)$, binary variable S indicating source selection.

Univariate regression example (Shimodaira 2000):

Definition 1.2 (Moreno-Torres et al. (2012))

Covariate shift is defined as $p_S(y|x) = p_T(y|x)$ but $p_S(x) \neq p_T(x)$.

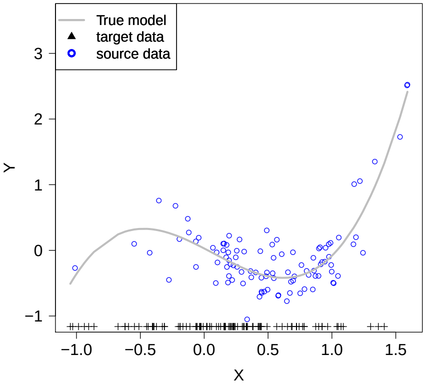
Simulated data:

- Source: $X_S \sim N(0.5, 0.5^2)$
- Target: $X_T \sim N(0.2, 0.5^2)$

Outcome generation:

- $y = -x + x^3 + \epsilon$, with $\epsilon \sim N(0, 0.3^2)$.

- 100 i.i.d. samples from X_S and X_T, along with y_S available.



Univariate regression example (Shimodaira 2000):

Definition 1.3 (Moreno-Torres et al. (2012))

Covariate shift is defined as $p_S(y|x) = p_T(y|x)$ but $p_S(x) \neq p_T(x)$.

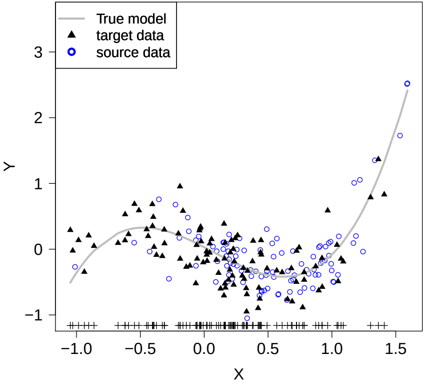
Simulated data:

- Source: $X_S \sim N(0.5, 0.5^2)$
- Target: $X_T \sim N(0.2, 0.5^2)$

Outcome generation:

- $y = -x + x^3 + \epsilon$,
  with $\epsilon \sim N(0, 0.3^2)$.

- 100 i.i.d. samples from X_S and X_T, along with y_S available.



Problem setting and objective:

Let $f : \mathcal{X} \rightarrow \mathbb{R}^K$ be the training function, f an element of the hypothesis space $\mathcal{H}$. Then,

- $\ell : \mathbb{R}^K \times \mathcal{Y} \rightarrow [0, \infty)$ is the loss function
- $\mathcal{R}(f) := \mathbb{E}[\ell(f(x), y)]$ is the risk function

Objective: Accurately predicting target labels y_T, by minimizing target risk

$$\mathcal{R}_T(f) := \mathbb{E}_{(x,y)\sim p_T(x,y)}[\ell(f(x), y)], \tag{1}$$

via labelled source data D_S and unlabelled target data D_T.

Univariate regression example (Shimodaira 2000):

Simulated data:

- Source: $X_S \sim N(0.5, 0.5^2)$
- Target: $X_T \sim N(0.2, 0.5^2)$

Outcome generation:

- $y = -x + x^3 + \epsilon$,
  with $\epsilon \sim N(0, 0.3^2)$.

Objective:

- Ordinary least square regression to predict y_T.
- 100 i.i.d. samples from X_S and X_T, along with y_S available.

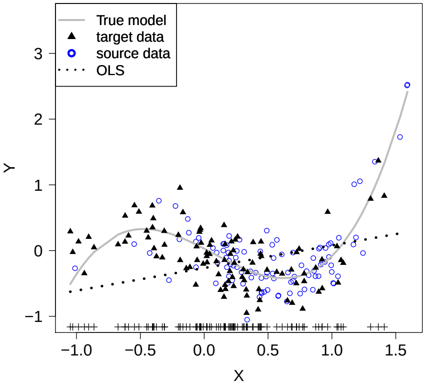


Figure: Illustrative univariate example.

Previous methods – Importance weighting:

Under **covariate shift** conditions:

Proposition 1 (Shimodaira (2000), Bickel et al. (2009))

If the support of $p_T(x)$ is contained in $p_S(x)$, then

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_T}\left[\ell(f(x),y)\right] = \mathbb{E}_{(x,y)\sim\mathcal{D}_S}\left[\frac{p_T(x)}{p_S(x)}\ell(f(x),y)\right]. \tag{2}$$

Proposition 2 (Bias Correction (Zadrozny 2004))

Let (x, y, s) be examples drawn from a distribution $\mathcal{D}$, with feature-label-selection space $\mathcal{X}\times\mathcal{Y}\times\mathcal{S}$. Then,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(f(x),y)\right] = \mathbb{E}_{(x,y)\sim\hat{\mathcal{D}}}\left[\ell(f(x),y)|s=1\right], \tag{3}$$

$$\text{with } \hat{\mathcal{D}}(x,y,s) := \frac{P(s=1)}{P(s=1|x)}\mathcal{D}(x,y,s). \tag{4}$$

Importance weight estimation and limitations:

- KLIEP – Kullback-Leibler Importance estimation procedure, minimizing the KL-divergence (Sugiyama et al. 2008).
- uLSIF – unconstrained least-squares importance fitting, minimizing the L2-norm as domain discrepancy (Kanamori et al. 2009).
- NN – Nearest-Neighbor (NN) importance weight estimator (Kremer et al. 2015; Lima et al. 2008; Loog 2012).
- IPS – Importance weights estimated through probabilistic classification of source set assignment (Kanamori et al. 2009).

Issue of importance weighting: Large (noisy) weights cause high variance and unreliable target predictions.

Illustration – Importance weighting:

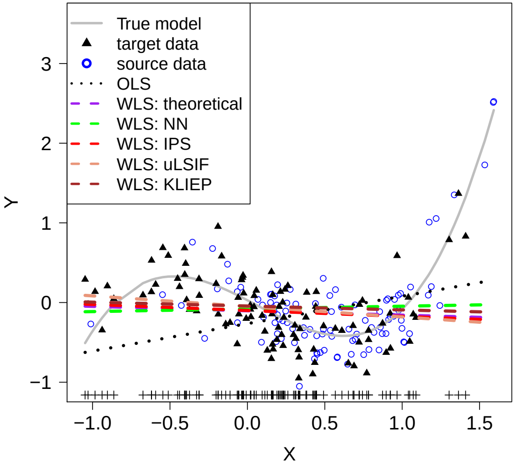


Figure: Illustrative weighted model fit.

Preliminaries – Propensity scores in causal inference:

- Rosenbaum and Rubin (1983) introduce propensity score (PS):

$$e(X) = P(Z = 1|X).$$

- Treatment assignment Z is strongly ignorable, if

$$\text{(i)}\ \ (Y_1, Y_0) \perp\!\!\!\perp Z|X \qquad \text{and} \qquad \text{(ii)}\ \ 0 < e(X) < 1. \tag{5}$$

- Rosenbaum and Rubin (1983) demonstrate:
  - **[Theorem 1]** PS is a balancing score, that is $x \perp\!\!\!\perp z|e(x)$
  - **[Theorem 4]** If (5) holds, conditional on PS, treatment effects unbiased
- PS methods for unbiased treatment effects:
  (i) Inverse probability of treatment weighting (IPTW),
  (ii) PS covariate adjustment, (iii) matching and (iv) stratification on PS

Methodology – Stratified Learning (StratLearn):

In our context we define the propensity score as:

$$e(x_i) := P(s_i = 1 | x_S, x_T), \text{ with } 0 < e(x_i) < 1. \tag{6}$$

Proposition 3 (Learning conditional on the propensity score)

Under covariate shift conditions, conditional on the propensity score:

$$p_T(x, y | e(x)) = p_S(x, y | e(x)). \tag{7}$$

That is, given $e(x)$ the joint source and target distributions are the same. It directly follows, for any loss function $\ell = \ell(f(x), y)$, that

$$E_{(x,y)\sim p_T(x,y|e(x))}[\ell(f(x), y)] = E_{(x,y)\sim p_S(x,y|e(x))}[\ell(f(x), y)]. \tag{8}$$

Methodology – StratLearn:

Verification of Proposition 3: Propensity score is a balancing score (Rosenbaum and Rubin 1983) [Theorem 1], in our case:

$$x \perp\!\!\!\perp s | e(x), \tag{9}$$

Under covariate shift conditions, it follows:

$$\begin{aligned} p_S(x, y|e(x)) :=& p(x, y|e(x), s = 1) \\ =& p(y|x, e(x), s = 1)p(x|e(x), s = 1) \qquad (10) \\ =& p(y|x, e(x), s = 0)p(x|e(x), s = 0) \qquad (11) \\ =& p(x, y|e(x), s = 0) \\ =:& p_T(x, y|e(x)). \end{aligned}$$

Thus, conditional on the propensity score, the source and target data have the same joint distribution. Equation (8) follows directly.

Methodology – StratLearn:

Subdivide ("stratify") target and source data in k subgroups according to quantiles of the propensity scores. Then supervised learning in each stratum ("stratified learner").

Stratification: For $j \in 1, \ldots, k$, we divide D_S and D_T into

$$D_{Sj}^{(k)} = \{(x, y) \in D_S : q_{k-j} < e(x) \leq q_{k-j+1}\} \tag{12}$$

$$D_{Tj}^{(k)} = \{x \in D_T : q_{k-j} < e(x) \leq q_{k-j+1}\}, \tag{13}$$

where q_j is the j'th k-quantile of $\{e(x_i) : x_i \in (x_S \cup x_T)\}$ and $q_0 = 0, q_k = 1$. As a consequence of Proposition 3, we have

$$p_{T_j}(y, x) \approx p_{S_j}(y, x), \text{ for } j \in 1, \ldots, k, \tag{14}$$

where subscript S_j means that we condition on assignment to the j'th source stratum (analogously for target T_j). Then,

$$\mathbb{E}_{(x,y)\sim p_{T_j}(x,y)}[\ell(f(x), y)] \approx \mathbb{E}_{(x,y)\sim p_{S_j}(x,y)}[\ell(f(x), y)], \text{ for } j \in 1, \ldots, k. \tag{15}$$

StratLearn – Technical details:

- Logistic regression to estimate PS (alternative: ML methods), including all suspected confounders as main effects.
- $k = 5$ strata – empirical evidence by Cochran (1968): five strata enough to remove 90 percent of bias.
- Given strata, model f_j fitted to source data D_{S_j}, to predict respective target samples in D_{T_j}, for $j \in 1, \ldots, k$.
- Model hyperparameters for f_j through empirical risk minimization on source D_{S_j} (e.g. cross-validation).
- When higher strata have insufficient source data for model training, source data from one or more adjacent stratum/strata added.

Illustration – Importance weighting:

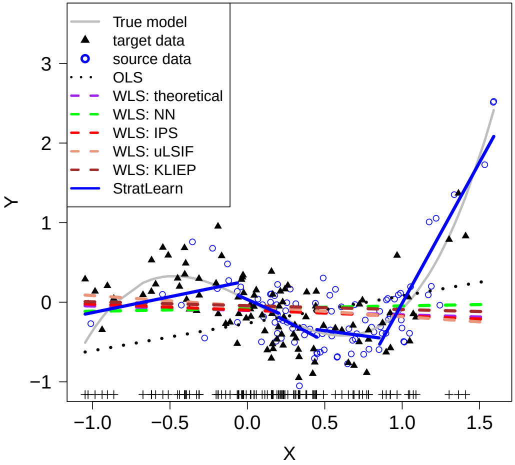


Figure: Illustrative *StratLearn* fit.

Balance diagnostics:

Covariate balance (following causal inference literature):

- Balance measures to verify propensity score model and/or suitability of choice of covariates (Austin 2011; Rosenbaum and Rubin 1984)
- e.g. standardized mean differences, Kolmogorov-Smirnov test statistics, comparison of higher order moments and interaction terms

Remark 1 (Outcome balance:)

In covariate shift framework

- *Potential outcomes are identical ($Y_0 \equiv Y_1$), no "treatment effect"*
- *Only source data is observed ($Y_1 \equiv Y$)*
- *Given $e(x)$, with $0 < e(x) < 1$, and covariate shift conditions, source data assignment is 'strongly ignorable'*
- *Then, conditional on PS, source and target outcome are the same in expectation [invoking Rosenbaum and Rubin (1983), Theorem 4].*

Univariate regression example (Shimodaira 2000):

Simulated data:

- Source: $X_S \sim N(0.5, 0.5^2)$
- Target: $X_T \sim N(0.2, 0.5^2)$

Outcome generation:

- $y = -x + x^3 + \epsilon$,
  with $\epsilon \sim N(0, 0.3^2)$.

Objective:

- Ordinary (weighted) least square regression to predict y_T.
- 100 i.i.d. samples from X_S and X_T, along with y_S available.

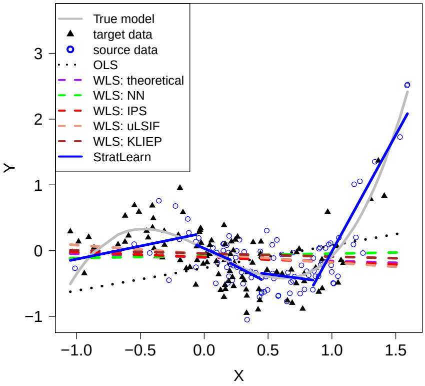


Figure: Representative model fit.

Univariate regression example:

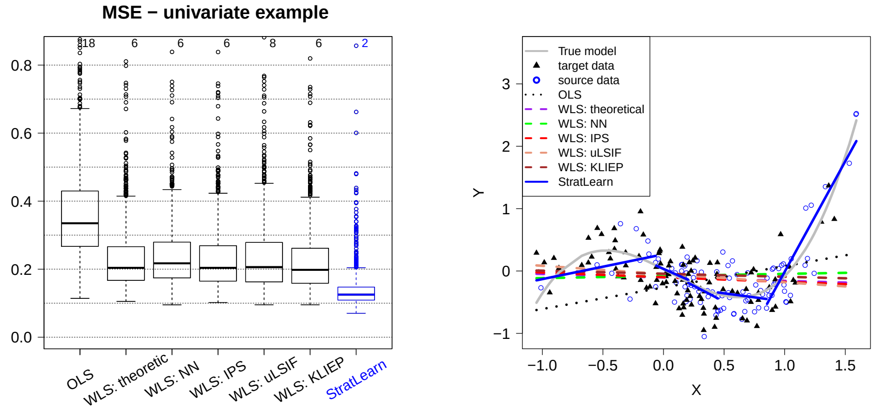


Figure: Left: Boxplot of the target MSE, obtained by $m = 1000$ Monte Carlo simulations. Right: Representative model fit.

Supernova classification – updated SPCC:

Objective: Reliable identification of Supernovae Type Ia (SNIa) based on photometric light curve (LC) data, given non-representative spectroscopically confirmed source data.

Data: Updated "Supernova photometric classification challenge" (SPCC) (Kessler et al. 2010)

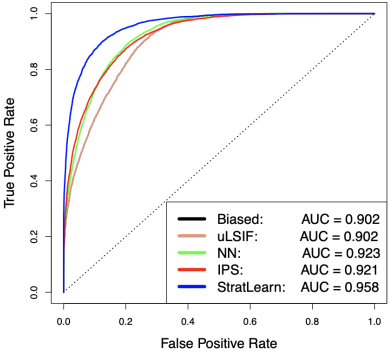
- LC data of 21,319 simulated supernovae of type Ia, Ib, Ic and II.
- Source data: 1102 spectroscopically confirmed SNe with known types
- Target data: 20,216 SNe with photometric information alone

Preprocessing:

- Gaussian process fit of LCs (four color bands C= (g,r,i,z)) combined with diffusion map to extract 100 covariates, plus redshift and a measure of brightness (Revsbech et al. 2018; Richards et al. 2012)

Supernova classification – StratLearn results:

Random forest classification, cross validation to select hyperparameter



Stratum	Set	Number of SNe	Number of SNIa	Prop. of SNIa
1	Source	958	518	0.54
	Target	3306	1790	0.54
2	Source	120	28	0.23
	Target	4144	927	0.22
3	Source	13	4	0.31
	Target	4250	540	0.13
4	Source	7	4	0.57
	Target	4257	610	0.14
5	Source	4	4	1
	Target	4259	662	0.16

Figure: Left: Comparison of target ROC curves on updated SPCC data. Right: Composition of the five strata on updated SPCC data (Kessler et al. 2010).

State-of-the-art:

Lochner et al. (2016): AUC=0.855; Pasquet et al. (2019): AUC=0.939;

Revsbech et al. (2018) (“STACCATO”): AUC=0.94;

Original SPCC data:

Stratum	Set	Number of SNe	Number of SNIa	Prop. of SNIa
1	Source	996	794	0.80
	Target	2470	1759	0.71
2	Source	210	56	0.27
	Target	3256	1010	0.31
3	Source	9	0	0
	Target	3457	385	0.11
4	Source	2	1	0.50
	Target	3464	258	0.07
5	Source	0	0	NA
	Target	3466	180	0.05

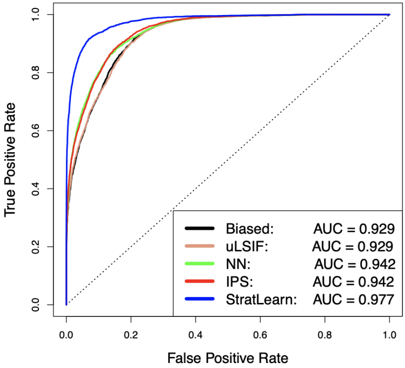


Figure: Left: Composition of the five strata on *original* SPCC data (Kessler et al. 2010). Right: Comparison of target ROC curves on *original* SPCC data.

State-of-the-art:

Revsbech et al. (2018) (“STACCATO”): AUC=0.961;

Photo-z conditional density estimation

Objective:
Conditional density estimation $f(z|x)$ of redshift, z, given photometric magnitudes x, in the presence of covariate shift.

Data (following Izbicki et al. (2017)):

- 467,710 galaxies (Sheldon et al. 2012), spectroscopic redshift z, five photometric covariates x (source D_S).
- Target D_T by rejection sampling from D_S, with $p(s=0|x) = f_{B(13,4)}(x_{(r)})/\max_{x_{(r)}} f_{B(13,4)}(x_{(r)})$.
- Additional $k \in \{10, 50\}$ i.i.d. standard normal covariates as potential confounders.
- Source: $|D_S^{\text{train}}| = 2800$, $|D_S^{\text{val}}| = 1200$; Target: $|D_T^{\text{test}}| = 6000$

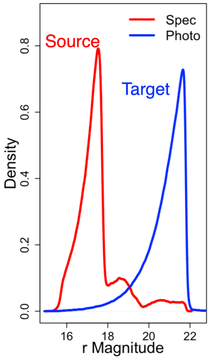


Photo-z conditional density estimation

***Generalized* risk** optimization (Izbicki et al. 2017) w.r.t:

$$\hat{R}_S(\hat{f}) = \frac{1}{n_T}\sum_{k=1}^{n_T}\int \hat{f}^2(z|x_T^{(k)})dz - 2\frac{1}{n_S}\sum_{k=1}^{n_S}\hat{f}(z_S^{(k)}|x_S^{(k)})\hat{w}(x_S^{(k)}), \tag{16}$$

Conditional density estimation models:

- hist-NN, ker-NN, Series
- Comb (combination model):

$$\hat{f}^{\alpha}(z|x) = \sum_{k=1}^{p}\alpha_k\hat{f}_k(z|x),\ \text{with constraints (i): } \alpha_i \geq 0,\ \text{and (ii): } \sum_{k=1}^{p}\alpha_k = 1, \tag{17}$$

StratLearn:

- Minimize (16) in each source stratum separately (with $w(x) \equiv 1$).
- *StratLearn* version of Comb, optimizing (17) on each source stratum (with $w(x) \equiv 1$), including *StratLearn* versions of ker-NN and Series.

Photo-z – Target results:

The target risk $\hat{R}_T(\hat{f})$ is computed as

$$\hat{R}_T(\hat{f}) = \frac{1}{n_T}\sum_{k=1}^{n_T}\int \hat{f}^2(z|x_T^{(k)})dz - 2\frac{1}{n_T}\sum_{k=1}^{n_T}\hat{f}(z_T^{(k)}|x_T^{(k)}), \qquad (18)$$

where z_T is the true target redshift, used for evaluation purposes only.

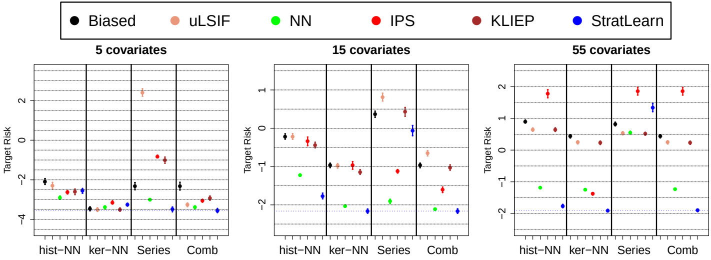


Figure: Target risk ($\hat{R}_T$) of photometric redshift estimation.

Photo-z – Target results:

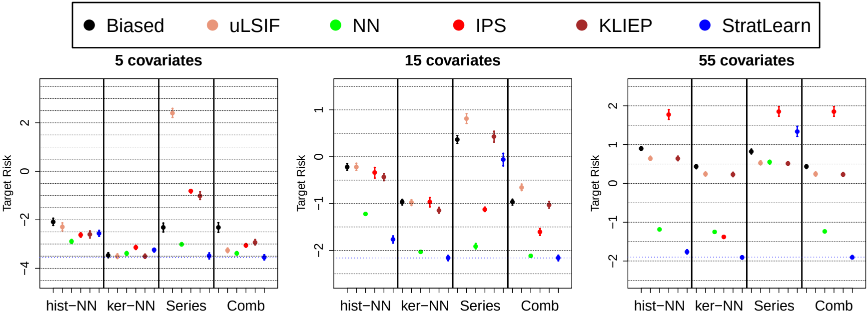


Figure: Target risk ($\hat{R}_T$) of photometric redshift estimation models, using different sets of predictors. Bars give the mean ± 2 bootstrap standard errors (from 400 bootstrap samples).

Photo-z – Strata composition:

Table: Composition of *StratLearn* strata for medium covariate shift on SDSS data, using estimated propensity scores with different sets of predictors.

		5 covariates	15 covariates	55 covariates
Stratum	Set	#galaxies (Mean z)	#galaxies (Mean z)	#galaxies (Mean z)
1	Source	1631 (0.06)	1583 (0.06)	1620 (0.06)
	Target	7 (0.05)	9 (0.05)	7 (0.05)
2	Source	1500 (0.09)	1515 (0.09)	1546 (0.09)
	Target	112 (0.08)	113 (0.09)	98 (0.08)
3	Source	618 (0.20)	641 (0.20)	594 (0.21)
	Target	1481 (0.23)	1499 (0.23)	1480 (0.23)
4	Source	116 (0.30)	114 (0.28)	108 (0.28)
	Target	2196 (0.27)	2215 (0.27)	2258 (0.27)
5	Source	135 (0.33)	147 (0.32)	132 (0.33)
	Target	2204 (0.33)	2164 (0.34)	2157 (0.34)
All	Source	4000 (0.11)	4000 (0.11)	4000 (0.11)
	Target	6000 (0.28)	6000 (0.28)	6000 (0.28)

Summary:

- *StratLearn* provides statistically principled framework for supervised learning under covariate shift (alternative to importance weighting)
- Especially advantageous in presence of high dimensional covariate space
- Examples demonstrate advantage of using small subset of source data chosen for its similarity to individuals in target data – markedly different to widespread practice of including all possible available data when fitting ML models.

Future work:

- Balance diagnostics via *Remark 1*, based on predicted outcome
- Matching on the propensity score
- Application of Deep Learning in *StratLearn* framework
- Model selection and strata combination
- Employ SNIa probabilities in secondary analysis in (pragmatic and fully) hierarchical Bayesian framework to estimate cosmological parameters

Future work – Balance diagnostics via predicted outcome

Remark 1 (Outcome balance:)

In covariate shift framework

- *Potential outcomes are identical ($Y_0 \equiv Y_1$), no "treatment effect"*
- *Only source data is observed ($Y_1 \equiv Y$)*
- *Given $e(x)$, with $0 < e(x) < 1$, and covariate shift conditions, source data assignment is 'strongly ignorable'*
- *Then, conditional on PS, source and target outcome are the same in expectation [invoking Rosenbaum and Rubin (1983), Theorem 4].*

- Target labels y_T in practice not observed, but source labels y_S and target and source label predictions ($\hat{y}_T$ and $\hat{y}_S$) are given

Future project: Model diagnostics using predicted outcomes ($\hat{y}_T$ and $\hat{y}_S$)

Future work – Balance diagnostics via predicted outcome

Table: Strata composition on updated SPCC data.

Stratum	Set	Number of SNe	Number of SNIa	Prop. of SNIa
1	Source	996	794	0.80
	Target	2470	1759	0.71
2	Source	210	56	0.27
	Target	3256	1010	0.31
3	Source	9	0	0
	Target	3457	385	0.11
4	Source	2	1	0.50
	Target	3464	258	0.07
5	Source	0	0	NA
	Target	3466	180	0.05

Table: Strata composition on SDSS photometric redshift data

Stratum	Set	5 covariates #galaxies (Mean z)
1	Source	1631 (0.06)
	Target	7 (0.05)
2	Source	1500 (0.09)
	Target	112 (0.08)
3	Source	618 (0.20)
	Target	1481 (0.23)
4	Source	116 (0.30)
	Target	2196 (0.27)
5	Source	135 (0.33)
	Target	2204 (0.33)
All	Source	4000 (0.11)
	Target	6000 (0.28)

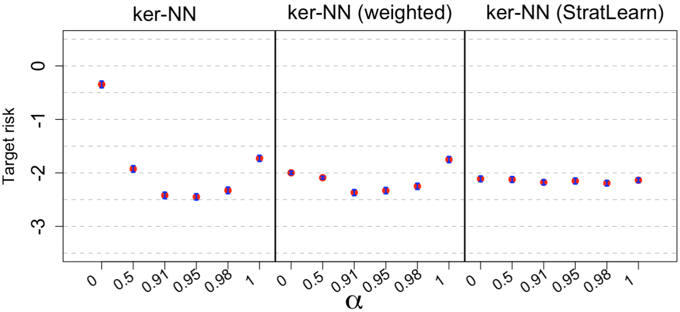
Future work – Matching on the propensity score

Ker-NN estimator (Izbicki et al. 2017):

$$\hat{f}(z|x) \propto \sum_{k \in \mathcal{N}_N(x)} \hat{w}(x_S^{(k)}) K_\epsilon (z - z_S^{(k)}),[1] \tag{19}$$

Nearest neighbor $\mathcal{N}_N(x_T^{(i)})$ by distance:

$$(1-\alpha) d(x_T^{(i)}, x_s^{(j)}) + \alpha d(e(x_T^{(i)}), e(x_s^{(j)})) \tag{20}$$

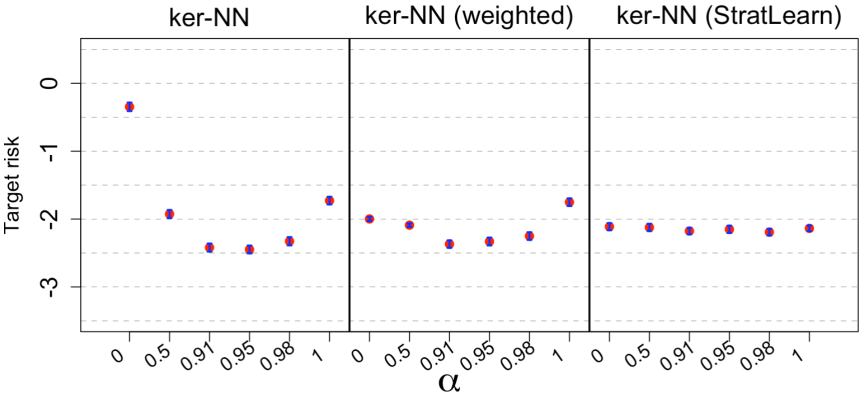


[1] with kernel smoother $K_\epsilon(z - z^{(k)}) = \exp(-(z - z^{(k)})^2)/4\epsilon$

Future work – Matching on the propensity score

Nearest neighbor $\mathcal{N}_N(x_T^{(i)})$ by distance:

$$(1-\alpha)d(x_T^{(i)}, x_s^{(j)}) + \alpha d(e(x_T^{(i)}), e(x_s^{(j)}))$$



References I

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Bickel, S., Brückner, M., and Scheffer, T. (2009). Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313.

Izbicki, R., Lee, A. B., Freeman, P. E., et al. (2017). Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. *The Annals of Applied Statistics*, 11(2):698–724.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445.

References II

Kessler, R., Bassett, B., Belov, P., Bhatnagar, V., Campbell, H., Conley, A., Frieman, J. A., Glazov, A., González-Gaitán, S., Hlozek, R., et al. (2010). Results from the supernova photometric classification challenge. *Publications of the Astronomical Society of the Pacific*, 122(898):1415.

Kouw, W. M. and Loog, M. (2019). A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*.

Kremer, J., Gieseke, F., Pedersen, K. S., and Igel, C. (2015). Nearest neighbor density ratio estimation for large-scale applications in astronomy. *Astronomy and Computing*, 12:67–72.

Lima, M., Cunha, C. E., Oyaizu, H., Frieman, J., Lin, H., and Sheldon, E. S. (2008). Estimating the redshift distribution of photometric galaxy samples. *Monthly Notices of the Royal Astronomical Society*, 390(1):118–130.

References III

Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., and Winter, M. K. (2016). Photometric supernova classification with machine learning. *The Astrophysical Journal Supplement Series*, 225(2):31.

Loog, M. (2012). Nearest neighbor-based importance weighting. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE.

Moreno-Torres, J. G., Raeder, T., Alaiz-RodríGuez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

Pasquet, J., Pasquet, J., Chaumont, M., and Fouchez, D. (2019). Pelican: deep architecture for the light curve analysis. *Astronomy & Astrophysics*, 627:A21.

References IV

Revsbech, E. A., Trotta, R., and van Dyk, D. A. (2018). Staccato: a novel solution to supernova photometric classification with biased training sets. *Monthly Notices of the Royal Astronomical Society*, 473(3):3969–3986.

Richards, J. W., Homrighausen, D., Freeman, P. E., Schafer, C. M., and Poznanski, D. (2012). Semi-supervised learning for photometric supernova classification. *Monthly Notices of the Royal Astronomical Society*, 419(2):1121–1135.

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

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.

References V

Sheldon, E. S., Cunha, C. E., Mandelbaum, R., Brinkmann, J., and Weaver, B. A. (2012). Photometric redshift probability distributions for galaxies in the sdss dr8. *The Astrophysical Journal Supplement Series*, 201(2):32.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM.

Thank you very much for your time!

Additional univariate regression simulations (Shimodaira 2000):

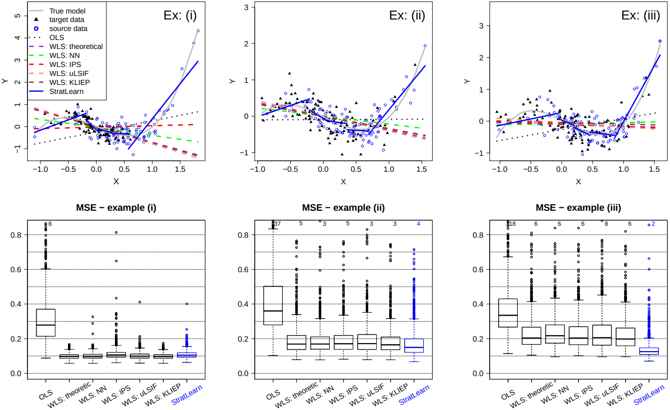


Figure: Top: Representative fit for each of the target parameter setting (i)-(iii). Bottom: Boxplot of the target MSE ($m = 1000$ Monte Carlo simulations)

PhotoZ: Varying strengths of covariate shift

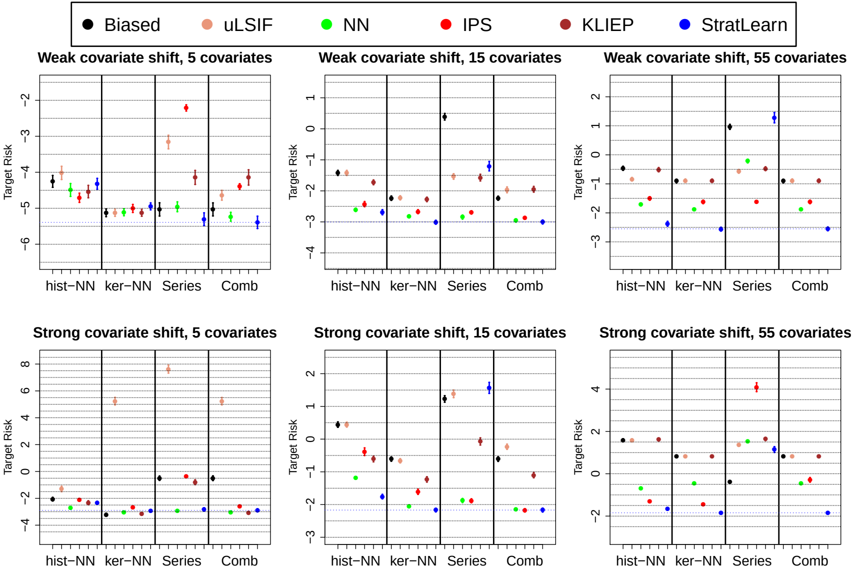
- Weak covariate shift:

$$p(s = 0|x) = f_{B(9,4)}(x_{(r)}) / \max_{x_{(r)}} f_{B(9,4)}(x_{(r)}) \tag{21}$$

- Strong covariate shift:

$$p(s = 0|x) = f_{B(18,4)}(x_{(r)}) / \max_{x_{(r)}} f_{B(18,4)}(x_{(r)}), \tag{22}$$

PhotoZ – varying strengths of covariate shift:



StratLearn under violation of the covariate shift assumption (UCI data examples):

Table: Composition of the five *StratLearn* strata for the UCI wine and UCI parkinson data. The number of samples/subjects in source and target stratum, as well as the mean outcome ("quality score" and "UPDRS score") are presented.

		UCI Wine data	UCI Parkinson data
Stratum	Set	# samples (Mean "quality")	# subjects (Mean "UPDRS")
1	Source	1299 (5.98)	627 (22.00)
	Target	0 (0.00)	174 (29.15)
2	Source	1300 (5.92)	486 (26.36)
	Target	0 (0.00)	315 (25.21)
3	Source	1300 (5.93)	314 (25.53)
	Target	0 (0.00)	487 (24.86)
4	Source	999 (5.63)	269 (27.38)
	Target	301 (5.49)	532 (28.15)
5	Source	0 (0.00)	181 (27.06)
	Target	1298 (5.67)	619 (30.00)
All	Source	4898 (5.88)	1877 (24.98)
	Target	1599 (5.64)	2127 (27.58)

StratLearn under violation of the covariate shift assumption (UCI data examples):

Table: MSE of target predictions on UCI Wine and Parkinson data, based on ordinary least squares regression (OLS), various importance weighted least squares regression methods (WLS), and our proposed StratLearn method.

Method \ Data	UCI wine data	UCI Parkinson data
OLS (Biased)	1.024	130.88
WLS:uLSIF	2.363	120.81
WLS:KLIEP	3.968	116.72
WLS:NN	2.377	117.47
WLS:IPS	0.660	112.80
StratLearn	0.715	114.97