Stratified Learning

Improved Learning under Covariate Shift

David A. van Dyk

Statistics Section of Department of Mathematics
Imperial College London

BIRS Programme on Astrostatistics in Canada and Beyond
Banff, October 2022

Learning with Non-Representative Data

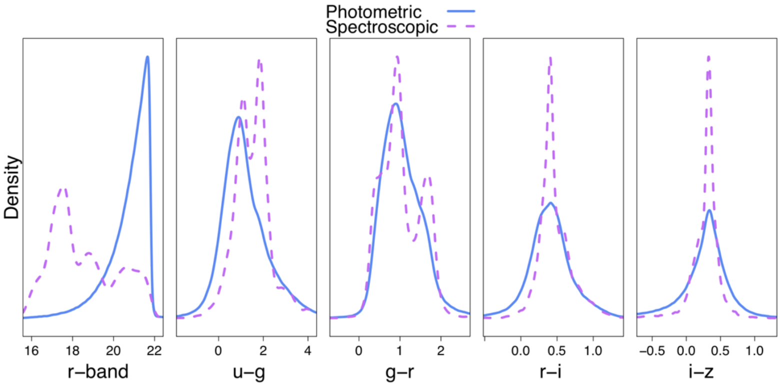
Can you learn about a population from a sample that only partially represents the population?

New general method – looking for additional applications.

Joint with: Max Autenrieth, David Stenning, and Roberto Trotta



Non-Representative Data



A General Challenge

- **Aim:** use training set (x, y) to predict target set (y from x).
- Spectroscopic data more available for bright/near objects.
- These object differ systematically from population.

Imperial College London

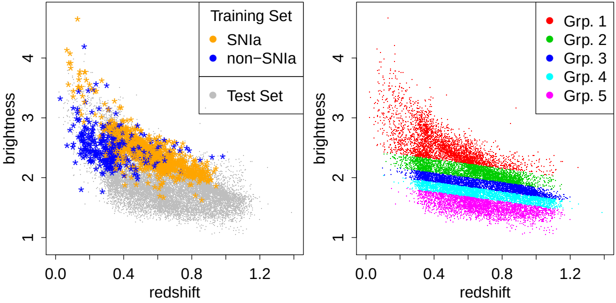
[Image Credit: Izbicki, Lee, Freeman, 2017, AoAS]

Learning with Non-Representative Data

Covariate Shift:

$$p_{\text{training}}(y \mid x) = p_{\text{target}}(y \mid x) \quad \textbf{but} \quad p_{\text{training}}(x) \neq p_{\text{target}}(x)$$

Supernovae classification:



Learning methods must be adapted to account for non-representative training data.

Imperial College London

Does a new drug improve health outcomes?

Causal Inference:

- Split subjects: treatment ($Z = 1$) and control ($Z = 0$) group
- What if treatment group differs systematically from control group, e.g., in terms of x.

$$p_{\text{treatment}}(x) \stackrel{?}{=} p_{\text{control}}(x)$$

- Randomiziation is the gold standard, not always possible.

Propensity Scores:

- Rosenbaum and Rubin (1983) define propensity scores:

$$e(x) = \Pr(Z = 1 \mid x).$$

- Demonstrate that $e(x)$ is a *balancing score*:

$$p_{\text{treatment}}(x \mid e(x)) = p_{\text{control}}(x \mid e(x)).$$

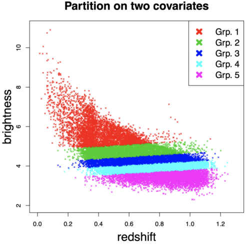
Imperial College London

... easy to diagnose in practice!

StratLearn:[1] Improved Learning under Covariate Shift

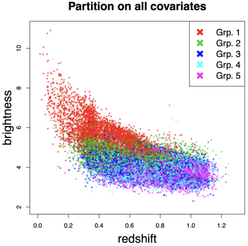
Propensity scores

- **Estimate:**
  $\hat{e}(x) = \Pr(\text{target set} \mid \text{covariates})$
- **Check:** $p_{\text{train}}(x \mid \hat{e}(x)) = p_{\text{target}}(x \mid \hat{e}(x))$
- Given $e(x)$, expected loss of predictor, $f(x)$, is same in target & training sets.



StratLearn

- Stratify target & training sets on $\hat{e}(x)$.
- Classify data separately in each strata.



Reduce covariate shift and thus expected classification/prediction error.

[1] Autenrieth, van Dyk, Trotta, and Stenning (2023). Stratified Learning: A General-Purpose Statistical Method for Improved Learning under Covariate Shift, SADM, to appear

Imperial College London

Supernova classification – updated SPCC:

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.
- Training Set: **1102 spectroscopically confirmed SNe** with known types
- Target Set: **20,216 SNe** with **photometric information** alone

Preprocessing:

- Gaussian process fit of LCs (four color bands, g, r, i, z) combined with diffusion map, plus redshift and a measure of brightness, to extract **102 covariates**

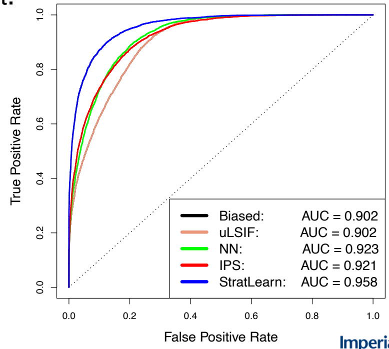
  (Revsbech et al., 2018; Richards et al.. 2012)

Results for Supernova Classification

Random forest classification, cross validation to select hyperparameter

ROC for StratLearn and several existing weighting methods.

- "Biased" ignores Covariate Shift.
- With an unbiased training set AUC = 0.965.



Weighting Methods for Covariate Shift

- Reweight training set: $p_{\text{target}}(x)/p_{\text{training}}(x)$.
- uLSIF (Kanamori et al. 2009);
- NN: Nearest-Neighbor (Kremer et al. 2015);
- IPS: probabilistic classification (Kanamori et al. 2009);

Photo-z conditional density estimation

Objective:
Conditional density estimation $f(z|x)$ of redshift given photometric magnitudes.

Significant covariate shift is magnitudes.

Data (following Izbicki et al., 2017):

- 468k galaxies (Sheldon et al. 2012), spectroscopic redshift, 5 photometric magnitudes.
- Create non-representative training set.
- Add $k \in \{10, 50\}$ i.i.d. Gaussian covariates.

What is the effect of high-dimensional irrelevant covariates?

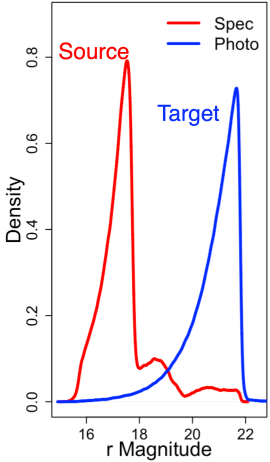
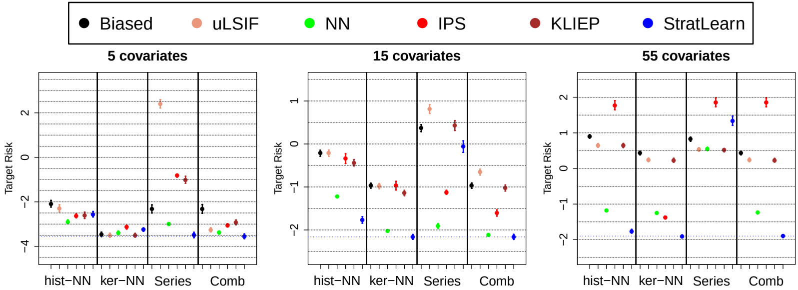


Photo-z – Stress Test:

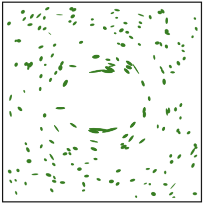


Target risk of photometric redshift estimates, using different sets of predictors.

StratLearn especially advantageous in presence of high dimensional covariate space.

Cosmic Shear Tomography

Weak Gravitational Lensing



- Large mass along line of sight creates distortion/shear in observed image.
- Shear Tomography bins galaxies on photo-z to map 3D distribution of mass.
- Resulting estimates of cosmological parameters under ΛCMD are inconsistent with those from CMD.
- A possible source of bias is binning of galaxies and the estimated redshift distribution within bins.

We use StratLearn to improve:

- Tomographic binning of galaxies
- Estimate z-distribution within bins (using hierarchical models)
- Joint work with: Benjamin Joachimi and Angus Wright.

Cosmic Shear Tomography

Confusion matrices for (a) z_B and (b) StratLearn:

(a) Target (columns) × Prediction (rows)

Prediction \ Target	1	2	3	4	5	l	r	
1	6.7% 837414	2.3% 292315	0.8% 103700	0.1% 8173	0% 4604	0.4% 52056	0.1% 10906	10.5% 1309168
2	2.4% 304578	11% 1378487	1.9% 233405	0% 4332	0.1% 7757	0% 3522	0.1% 6925	15.5% 1939006
3	2.4% 296104	6.8% 847716	10.9% 1364656	0.9% 112729	0.9% 116545	0.1% 8273	0.8% 93785	22.8% 2839808
4	0.3% 32511	0.4% 47784	5% 618554	8.4% 1051001	2.6% 329657	0% 1465	0.6% 77109	17.3% 2158081
5	0.2% 29292	0.7% 83816	2% 255205	5.7% 710042	10.3% 1288885	0% 1044	2.3% 287567	21.3% 2655851
l	0.3% 32843	0% 86	0% 1422	0% 1861	0% 527	0.2% 24629	0% 1164	0.5% 62532
r	0.5% 58912	0.5% 63239	1% 118739	1.1% 134700	4.1% 515790	0% 3614	5% 620560	12.1% 1515554
	12.8% 1591654	21.7% 2713443	21.6% 2695681	16.2% 2022838	18.1% 2263765	0.8% 94603	8.8% 1098016	12480000

(b) Target (columns) × Prediction (rows)

Prediction \ Target	1	2	3	4	5	l	r	
1	6.9% 861710	0.9% 115539	0.6% 71099	0% 6137	0% 6040	0.6% 74285	0.1% 8574	9.2% 1143384
2	3.4% 427042	17.1% 2128352	3.8% 475212	0.2% 30545	0.4% 55909	0% 5650	0.4% 56112	25.5% 3178822
3	1.2% 149750	2.4% 301283	12.2% 1519233	1.4% 173335	0.7% 81414	0% 4597	0.5% 64981	18.4% 2294593
4	0.4% 44384	0.4% 45478	3.2% 402818	11.6% 1445904	3.6% 448609	0% 3046	1% 124592	20.2% 2514831
5	0.6% 77467	0.8% 96919	1.5% 189063	2.8% 345528	12.6% 1575876	0% 4540	4.9% 607041	23.2% 2896434
l	0% 2					0% 12		0% 14
r	0.3% 31298	0.2% 25872	0.3% 38256	0.2% 21390	0.8% 95917	0% 2472	1.9% 236717	3.6% 451922
	12.8% 1591653	21.7% 2713443	21.6% 2695681	16.2% 2022839	18.1% 2263765	0.8% 94602	8.8% 1098017	12480000

Reduce bias by 40% compared with best available alternative.

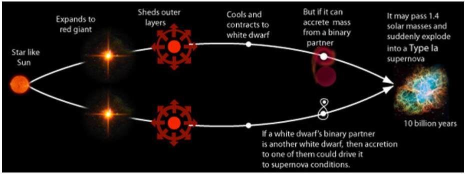
[Within bin mean of z, bias averaged across bins.]

Studying the Expansion History of Universe[2]

Type Ia Supernovae had a common "flashpoint"

Absolute magnitudes:

$$M_j^{\mathrm{Ia}} \sim \mathrm{N}(M_0^{\mathrm{Ia}}, \sigma_{\mathrm{int}}^{\mathrm{Ia}}).$$



Non-linear Regression: $m_{Bj} = g(z_j, \Omega_\Lambda, \Omega_M, H_0) + M_j^{\mathrm{Ia}}$

[function of density of dark energy and of total matter]

[part of a (second-stage) fully-Bayesian Hierarchical model ⋆]

For Non Type Ia: $M_j^{\mathrm{Ia}'} \sim \mathrm{Distribution}(M_0^{\mathrm{Ia}'}, \sigma_{\mathrm{int}}^{\mathrm{Ia}'})$ with $\sigma_{\mathrm{int}}^{\mathrm{Ia}'} \gg \sigma_{\mathrm{int}}^{\mathrm{Ia}}$

First Stage Analysis: Classify Supernova into Type Ia, non Type Ia.

[2] Shariff, Jiao, Trotta, and van Dyk (2016). BAHAMAS: New SNIa Analysis Reveals Inconsistencies with Standard Cosmology. *The Astrophysical Journal*, **827**, 1

Two-Stage Analysis

Let:

- Y_0 = data used to classify supernovae
- Y_1 = data used to fit cosmological parameters
- Z = classification of supernovae (1 for Type 1a, 0 otherwise)
- θ = cosmological parameters

Pragmatic Bayes: $\pi_0(Z, \theta) = p(Z \mid Y_0)\, p(\theta \mid Z, Y_1)$

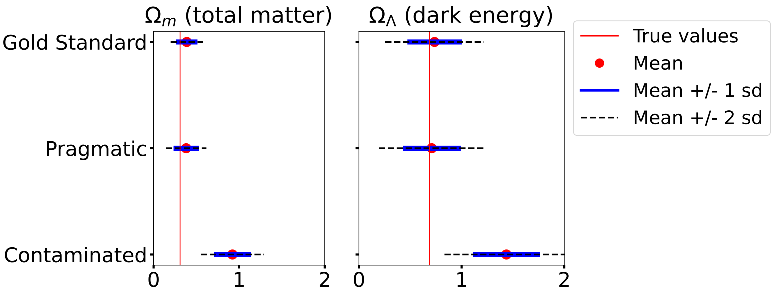
- Resample $Z^{(t)} \sim p(Z \mid Y_0)$.
- Sample $\theta^{(t)} \sim p(\theta \mid Z^{(t)} Y_1)$.

Fully Bayes: $\pi(Z, \theta) = p(Z \mid Y_0, Y_1)\, p(\theta \mid Z, Y_0, Y_1)$

- Y_1 improves classification, Z (and thus θ estimate).

Pragmatic Bayesian – Simulation Study

- Frequentist evaluation with 8 repetitions on simulated data each with 500 SNe (5% contamination).



- Pragmatic approach recovers true parameters well, with slightly increased variance compared to Gold Standard.
- Results shown consistent for other parameters.

For Further Reading I

Autenrieth, M., van Dyk, D. A., Trotta, R., and Stenning, D. C.
Stratified Learning: A General-Purpose Statistical Method for Improved Learning under Covariate Shift
Statistical Analysis and Data Mining, 2023, 1–16.

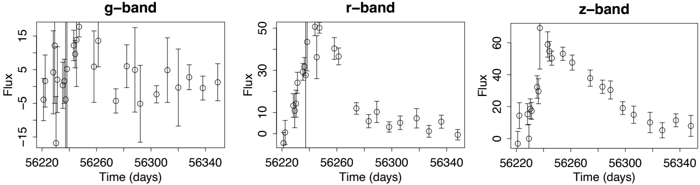
Autenrieth, M., Joachimi, B., Stenning, D. C., Trotta, R., van Dyk, D. A., and Wright, A. H.
Improved Weak Lensing Photometric Redshift Calibration via StratLearn and Hierarchical Modeling
preprint, 2023+.

Revsbech, E., Trotta, R., and van Dyk, D. A.
STACCATO: A Novel Solution to Supernova Photometric Classification...
Monthly Notices of the Royal Astronomical Society, **473**, 3969–3986, 2018.

Shariff, H., Jiao, X., Trotta, R., and van Dyk, D. A.
BAHAMAS: SNIa Analysis Reveals Inconsistencies with Standard Cosmology.
The Astrophysical Journal, **827**, 1 (25 pp), 2016.

Photometric Classification of SNe[3]

Data:

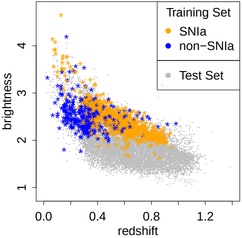


E.g., Supernova photometric classification challenges, such as Kessler (2010).

Classifier: Interpolate with Gaussian Process → Identify features w/ Diffusion Maps → Classify using Random Forest

Unfortunately Data are Subject to Covariate Shift.

$$p_{\text{training}}(x) \neq p_{\text{target}}(x).$$



[3] Revsbech, Trotta, and van Dyk (2018). STACCATO: A Novel Solution to Supernova Photometric Classification with Biased Training Samples, **473**, 3969-3986.

Imperial College London