The s-value: evaluating stability with respect to distributional shifts

Suyash Gupta,
Stanford University

December, 2023

Joint work with

Dominik Rothenhaeusler

Motivation

What is the problem?

Replicability crisis due to-

failure to account for multiple testing, publication bias, problematic incentives, distributional shifts

What is the problem?

Replicability crisis due to-

failure to account for multiple testing, publication bias, problematic incentives, **distributional shifts (this talk)**

Example

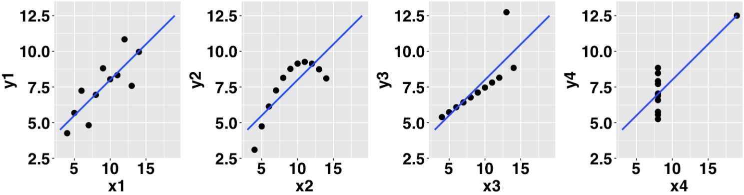


Figure: Anscombe's quartet (F.J.Anscombe, 1973)

$Y = \beta_0 + X\beta_1$	OLS estimate (β_1)	p-values
Set 1	0.5	0.00217
Set 2	0.5	0.00217
Set 3	0.5	0.00217
Set 4	0.5	0.00217

Table: Identical estimates despite varying distributions

Example

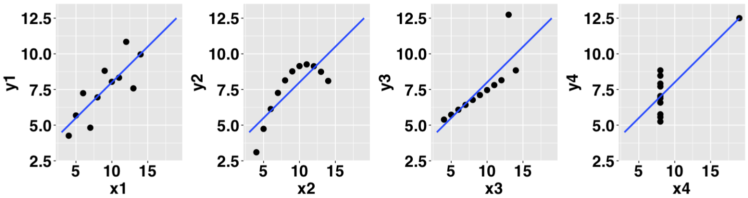


Figure: Anscombe's quartet (F.J.Anscombe, 1973)

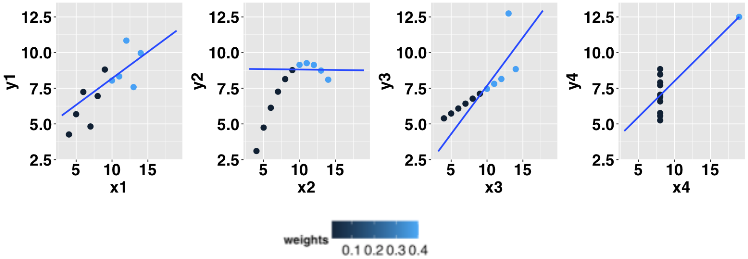


Figure. Covariate shift changes the OLS estimates differently (set 2 is the most unstable)

Example

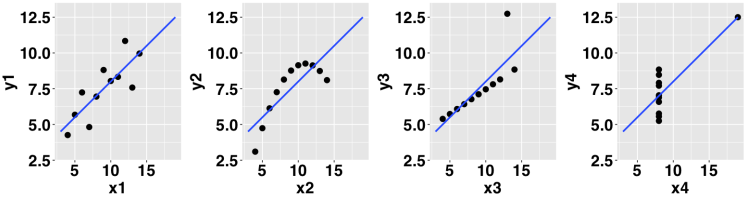


Figure: Anscombe's quartet (F.J.Anscombe, 1973)

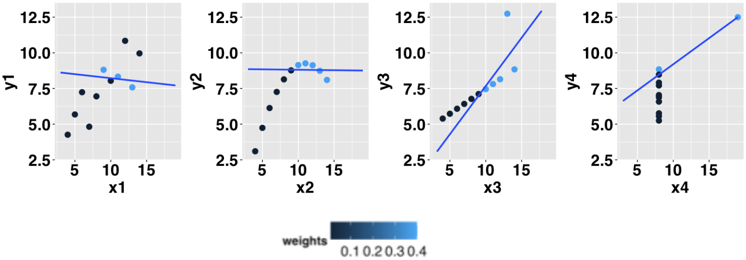


Figure: More general shifts make sets 1 and 4 unstable.

Take home messages?

- Classical measures convey little about distributional stability of estimators.
- Considering overall distributional shifts maybe a bit conservative.
- Often not all aspects of distribution shifts.
- We should also consider less conservative shifts like shift in marginal distribution of observed covariates.

Sources of distributional instability

- model misspecification
- presence of confounding variables
- selection bias that changes across settings
- heterogeneity etc.

Our contributions

- We propose a measure of stability that
  quantifies the distributional instability of a statistical estimand.
- We develop measures for both overall and conditional shifts in distributions.
- We use the above measures to guide transfer learning procedure for better estimation under shifted distribution

Notation

Consider probability distributions with finite support (size K) (for ease)

$$P \leftarrow \text{Probability distribution}$$

$$w \leftarrow \text{corresponding weights (K dimensional)}$$

We will use them interchangeably (again for ease)!

(Marginal) s-value

Consider training distribution P^0, parameter θ such that $\theta(P^0) > 0$.

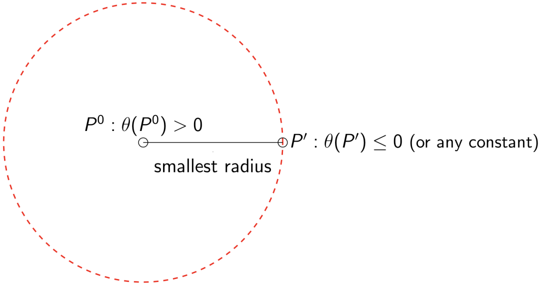


Figure: KL Divergence ball

(marginal) stability value, $s(\theta, P^0)$= exp(-smallest radius) $\in [0, 1]$.

(Marginal) s-value of mean of a random variable

Distn. P^0, weights w^0, finite support $Z_1, \ldots, Z_K$, s-value for mean $\mu(w^0)$:

$$s(\mu, P^0) = \exp\left\{ -\min_w \sum w_k \log\left(\frac{w_k}{w_k^0}\right) \right\} \text{ s.t. } \mu(w) = \sum w_k Z_k = 0$$

(Convex in probability weights w) easy!

(Marginal) s-value of mean of a random variable

Distn. P^0, weights w^0, finite support $Z_1, \ldots, Z_K$, s-value for mean $\mu(w^0)$:

$$s(\mu, P^0) = \exp\left\{ -\min_w \sum w_k \log\left(\frac{w_k}{w_k^0}\right) \right\} \text{ s.t. } \mu(w) = \sum w_k Z_k = 0$$

(Convex in probability weights w) easy!

Theorem (*Donsker and Varadhan, 1976; Owen, 2001*)

It turns out that

$$s(\mu, P^0) = \inf_\lambda \mathbb{E}_{P^0}[e^{\lambda Z}]. \tag{1}$$

Optimal weights are given by

$$w_k^* \propto e^{\lambda^* Z_k} w_k^0$$

where λ^* is the minimizer in (1).

$$\frac{\partial \mu}{\partial w_k} = Z_k$$

(Marginal) s-value of more general parameter?

For training distribution P^0 (weights w^0), parameter θ, s-value is

$$s(\theta, P^0) = \exp\left\{ -\min_w \sum w_k \log\left(\frac{w_k}{w_k^0}\right) \right\} \text{ s.t. } \theta(w) = 0.$$

(Maybe non-convex in w) easy? (may be not!)

Example: Linear regression coefficients (non-linear in weights).

Related work (Empirical Likelihood (Owen, 2001))

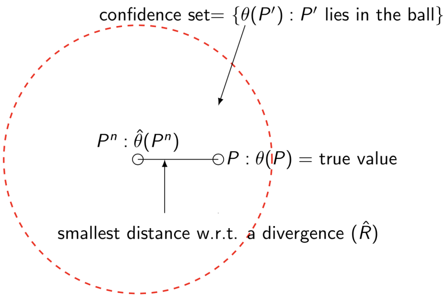


Figure. Divergence ball. (Use asymptotic distribution of $\hat{R}$ to construct confidence set)

Related work (Empirical Likelihood (Owen, 2001))

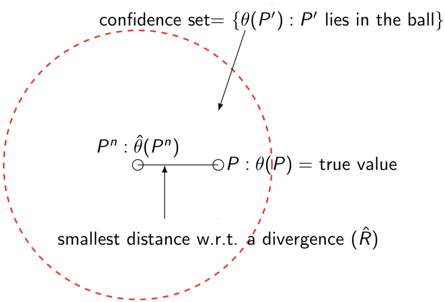


Figure. Divergence ball. (Use asymptotic distribution of $\hat{R}$ to construct confidence set)

- Considers asymptotically negligible shift in (marginal) distributions $O(\frac{1}{n})$.

Related work (Empirical Likelihood (Owen, 2001))

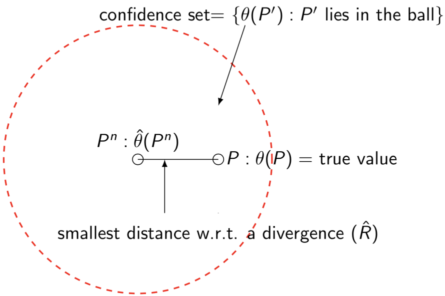


Figure. Divergence ball. (Use asymptotic distribution of $\hat{R}$ to construct confidence set)

- Considers asymptotically negligible shift in (marginal) distributions $O(\frac{1}{n})$.
- Hence, can approximate non-linear parameters linearly $\theta(w) \approx \theta(w^n) + \langle \nabla\theta(w^n), w \rangle$.
- We are interested in large shifts (both marginal and conditional)!

(Conditional) s-value

Consider training distribution P^0, parameter θ such that $\theta(P^0) > 0$ and covariate E

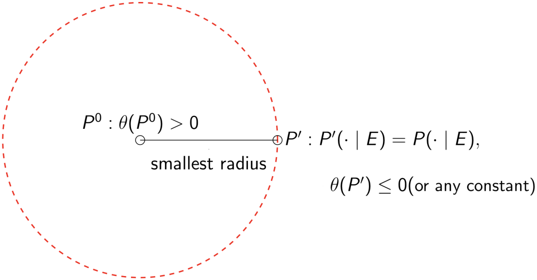


Figure: KL Divergence ball

(conditional) stability value, $s_E(\theta, P^0)$= exp(-smallest radius) $\in [0, 1]$.

(Conditional) *s*-value of mean of a random variable

For $(Z, E) \sim P^0$, s-value for the mean $(\mu(w^0) = \mathbb{E}_{P^0}[Z])$ conditional on E

$$s_E(\mu, P^0) = \exp\left\{ -\min_w \sum w_k \log\left(\frac{w_k}{w_k^0}\right) \right\} \text{ s.t. } \mu(w) = \sum w_k Z_k = 0 \ ,$$
$$P(\cdot \mid E) = P^0(\cdot \mid E)$$

(Conditional) s-value of mean of a random variable

For $(Z, E) \sim P^0$, s-value for the mean ($\mu(w^0) = \mathbb{E}_{P^0}[Z]$) conditional on E

$$s_E(\mu, P^0) = \exp\left\{ -\min_w \sum w_k \log\left(\frac{w_k}{w_k^0}\right) \right\} \text{ s.t. } \mu(w) = \sum w_k Z_k = 0 \ ,$$

$$P(\cdot \mid E) = P^0(\cdot \mid E)$$

Theorem

It turns out that

$$s(\mu, P^0) = \inf_\lambda \mathbb{E}_{P^0}[e^{\lambda \mathbb{E}_{P^0}[Z|E]}]. \qquad (2)$$

Optimal distribution is given by

$$w_k^* \propto e^{\lambda^* \mathbb{E}_{P^0}[Z|E=E_k]} w_k^0$$

where λ^* is the minimizer in (2).

R.V. $Z = \frac{\partial \mu}{\partial w}$ $\quad \mathbb{E}_{P^0}[Z \mid E = E_k] = \mathbb{E}_{P^0}[(\frac{\partial \mu}{\partial w}) \mid E = E_k]$

Linear parameters

Parameters that are linear in weights/ can be written as mean of a random variable-

$$\theta(P) = \mathbb{E}_P[\phi(Z)],\ s\text{-value} = \inf_{\lambda} \mathbb{E}_{P^0}[e^{\lambda\phi(Z)}]$$

Linear parameters

Parameters that are linear in weights/ can be written as mean of a random variable-

$$\theta(P) = \mathbb{E}_P[\phi(Z)], \; s\text{-value} = \inf_{\lambda} \mathbb{E}_{P^0}[e^{\lambda\phi(Z)}]$$

- AIPW estimator for average treatment effect (under covariate shift) (Jeong and Namkoong, 2020)
- Predictive coverage (Cauchois et al., 2020)
- Predictive risk on a validation set (conditional shifts) (Subbaswamy et al., 2021)

They all consider worst achievable value within a given amount of shift. We can use binary search to find s-values.

Non-linear parameters

$$s(\theta, P^0) = \exp\left\{ -\min_w \sum w_k \log\left(\frac{w_k}{w_k^0}\right) \right\} \text{ s.t. } \theta(w) = 0 \text{ (or } c\text{)}.$$

- Problem is non-convex if θ is non-linear.
- Can we at least obtain a local optima?

Non-linear parameters

$$s(\theta, P^0) = \exp\left\{ -\min_w \sum w_k \log\left(\frac{w_k}{w_k^0}\right) \right\} \text{ s.t. } \theta(w) = 0 \text{ (or } c).$$

- Problem is non-convex if θ is non-linear.
- Can we at least obtain a local optima?

Theorem (*Marginal case*)

If w^* is a locally optimal solution to the above problem then w^* satisfies

$$w_k^* \propto e^{\lambda \frac{\partial \theta(w^*)}{\partial w_k}}$$

for some constant λ. Further, if w^* is of above form and $sign(\lambda) \cdot sign(\theta(P^0)) = -1$, then w^* is a local optima.

Non-linear parameters

$$s(\theta, P^0) = \exp\left\{ -\min_w \sum w_k \log\left(\frac{w_k}{w_k^0}\right) \right\} \text{ s.t. } \theta(w) = 0 \text{ (or } c\text{)},$$
$$P(\cdot \mid E) = P^0(\cdot \mid E)$$

- Problem is non-convex if θ is non-linear.
- Can we at least obtain a local optima?

Theorem (*Conditional case*)

If w^* is a locally optimal solution to the above problem then w^* satisfies

$$w_k^* \propto e^{\lambda \mathbb{E}_{P^0}\left[\frac{\partial \theta(w^*)}{\partial w} | E = E_k\right]}$$

for some constant λ. Further, if w^* is of above form and $sign(\lambda) \cdot sign(\theta(P^0)) = -1$, then w^* is a local optima.

Algorithm for non-linear setting

Minimize the lagrangian ($\delta > 0$ if $\theta(w^0) > 0$)

$$\delta\theta(w) + \sum w_k \log\left(\frac{w_k}{w_k^0}\right)$$

Use MM algorithm with a (carefully chosen) convex majorizer.

Algorithm for non-linear setting

Minimize the lagrangian ($\delta > 0$ if $\theta(w^0) > 0$)

$$\delta\theta(w) + \sum w_k \log\left(\frac{w_k}{w_k^0}\right)$$

Use MM algorithm with a (carefully chosen) convex majorizer.

Assumption

θ is continuously differentiable and M smooth, that is, for weights w, w^0,

$$|\theta(w) - \theta(w^0) - \langle \nabla\theta(w^0), w - w^0\rangle| \leq \frac{M}{2}\left\|w - w^0\right\|_2^2 .$$

Algorithm for non-linear setting

Minimize the lagrangian ($\delta > 0$ if $\theta(w^0) > 0$)

$$\delta\theta(w) + \sum w_k \log\left(\frac{w_k}{w_k^0}\right)$$

Use MM algorithm with a (carefully chosen) convex majorizer.

Assumption

θ is continuously differentiable and M smooth, that is, for weights w, w^0,

$$|\theta(w) - \theta(w^0) - \langle\nabla\theta(w^0), w - w^0\rangle| \leq \frac{M}{2}\left\|w - w^0\right\|_2^2.$$

Using equivalence of ℓ_1 and ℓ_2 norms and by Pinsker's inequality, we have

$$|\theta(w) - \theta(w^0) - \langle\nabla\theta(w), w - w^0\rangle| \leq L\sum w_k \log\frac{w_k}{w_k^0}.$$

Proposition

Let w^1 be the minimizer of the above majorizer, w^1 is given by

$$w_k^1 \propto e^{-\frac{\delta}{1+L\delta}\frac{\partial\theta(w^0)}{\partial w_k}}(w_k^0)^{\frac{L\delta}{1+L\delta}}.$$

Proposition

Let w^1 be the minimizer of the above majorizer, w^1 is given by

$$w_k^1 \propto e^{-\frac{\delta}{1+L\delta}\frac{\partial\theta(w^0)}{\partial w_k}}(w_k^0)^{\frac{L\delta}{1+L\delta}}.$$

Recall

$$s(\theta, P^0) = \exp\left\{ -\min_w \sum w_k \log\left(\frac{w_k}{w_k^0}\right) \right\} \text{ s.t. } \theta(w) = 0 \text{ (or } c).$$

Proposition

If all stationary points of lagrangian are isolated, then the iterates of MM algorithm converge to some w^*, and w^* is a local optima to the above problem, where the $\theta(w)$ is constrained to equal $\theta(w^*)$.

Obtain δ via binary search for which convergent solution satisfies $\theta(w^*) = 0$ (or c).

Example

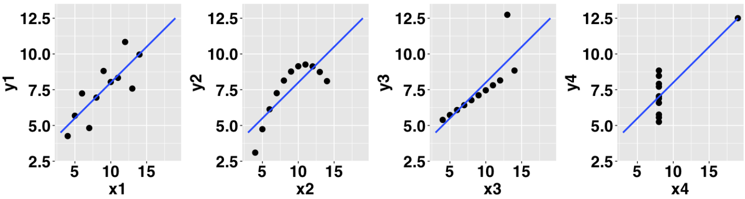


Figure: Anscombe's quartet (F.J.Anscombe, 1973)

$Y = \beta_0 + X\beta_1$	OLS estimate (β_1)	p-values	s	s_X
Set 1	0.5	0.00217	0.465	0
Set 2	0.5	0.00217	0.63	0.63
Set 3	0.5	0.00217	0	0
Set 4	0.5	0.00217	0	0

Table. OLS estimate, p-values, marginal and conditional s-values of the regression coefficient for each set.

Example

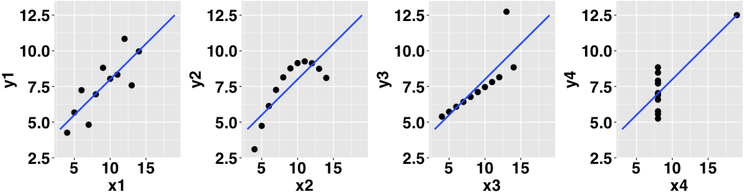
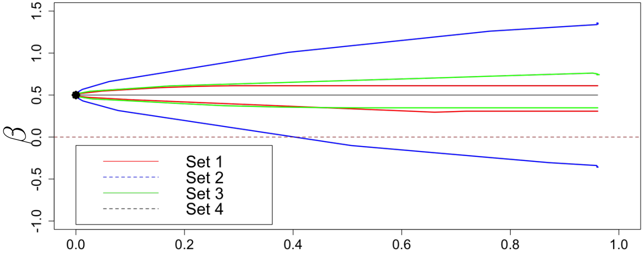


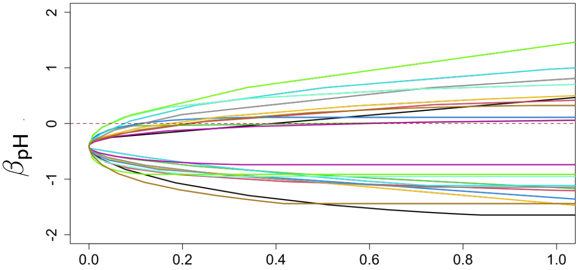
Figure: Anscombe's quartet (F.J.Anscombe, 1973)



Conditional shift with respect to KL divergence

Wine quality data (Linear regression example)

- Two subgroups– red wine and white wine.
- Response- wine quality (1 to 10), covariates- some continuous features of wine



Shift in marginal distribution of a given covariate with

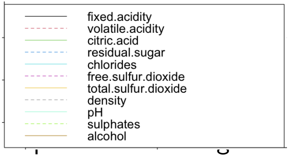
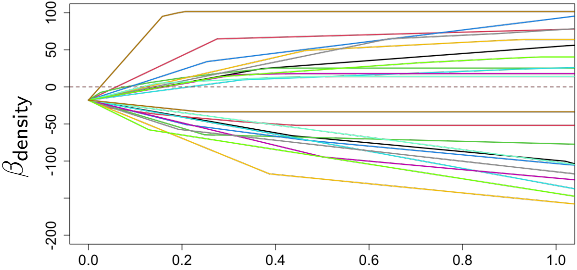


Figure: Minimum and maximum achievable value within a given shift.



Shift in marginal distribution of a given covariate with

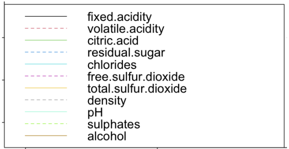
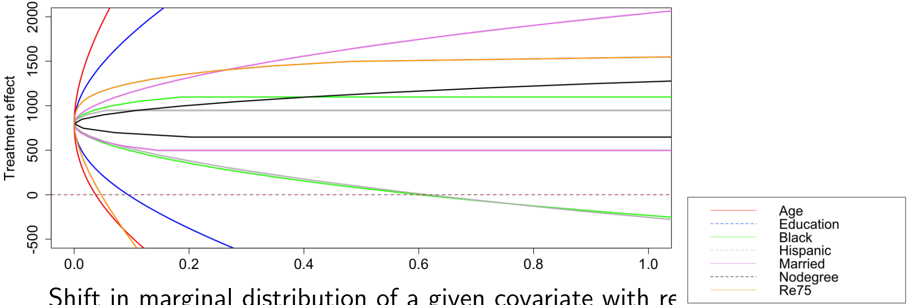


Figure: Minimum and maximum achievable value within a given shift.

National supported work demonstration data (NSW) (Lalonde, 1986)

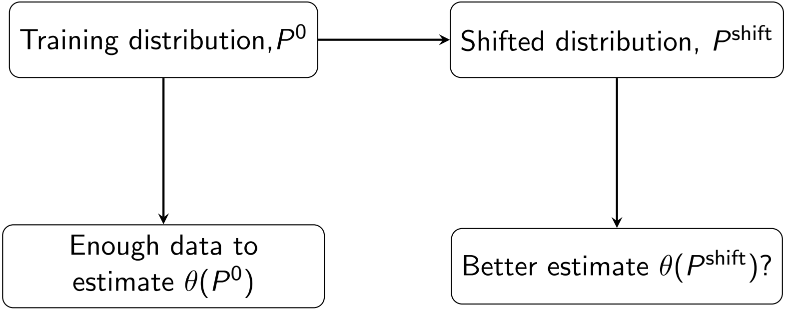
Average treatment effect of an employment program on trainee earnings.



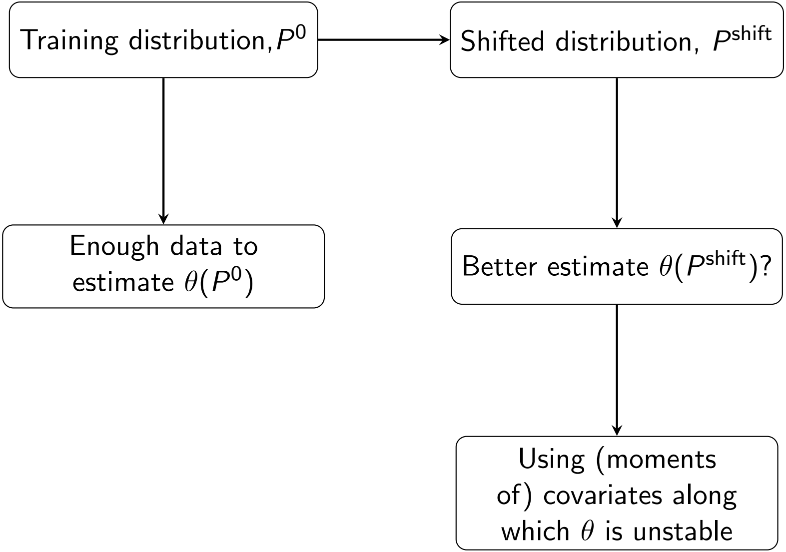
Shift in marginal distribution of a given covariate with re ce

Figure: Minimum and maximum achievable value within a given shift.

Parameter transfer



Parameter transfer



Parameter transfer under covariate shift

1. (Intuitively) match moments of covariates along which parameter is unstable (X_S).
2. Regularize by still being as close to training distribution as possible.

$$P_{\text{proj}} = \arg\min_{P'} D_{KL}(P' \| P^0) \text{ such that } \mathbb{E}_{P'}[g(X_S)] = \mathbb{E}_{P^0}[g(X_S)].$$

3. Compute $\theta(P_{\text{proj}})$.

Likelihood ratio based reweighting with full covariate information-
ATE-Dahabreh et al. (2019)
Predictive coverage-Barber et al. (2019)

Parameter transfer

Proposition (*Transfer of parameters*)

Assume that $t \mapsto \theta(tP_0 + (1-t)P)$ *is continuously differentiable with derivative* $\mathbb{E}_{P_0}[\phi_t(Z)] - \mathbb{E}_P[\phi_t(Z)]$ *for* ϕ_t *the influence function at* $tP_0 + (1-t)P$*. Let* $\epsilon_t = \inf_b \|\phi_t - b^\intercal g(X_S)\|_\infty$*. Then, any distribution* P' *that satisfies* $\mathbb{E}_{P'}[g(X_S)] = \mathbb{E}_P[g(X_S)]$*,*

$$|\theta(P') - \theta(P)| \leq \|\epsilon\|_\infty = 2 \sup_{t \in [0,1]} \mid \epsilon_t \mid .$$

Parameter transfer

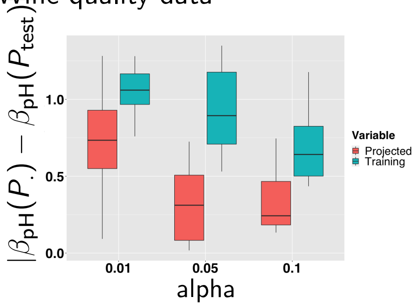
Proposition (*Transfer of parameters under conditional shifts*)

Let X_S be a variable such that $P[\bullet|X_S] = P_0[\bullet|X_S]$ and let $K = g(X_S)$. Assume that $t \mapsto \theta(tP_0 + (1-t)P)$ is continuously differentiable with derivative $\mathbb{E}_{P_0}[\phi_t(Z)] - \mathbb{E}_P[\phi_t(Z)]$ for ϕ_t the influence function at $tP_0 + (1-t)P$. Let $\epsilon_t = \inf_b \|\mathbb{E}[\phi_t|S] - b^\intercal g(X_S)\|_\infty$. Then,

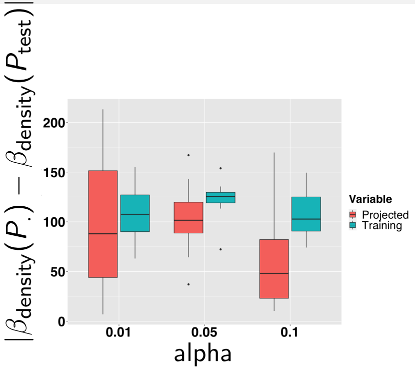
$$|\theta(P_{proj}) - \theta(P)| \leq 2\|\epsilon\|_\infty. \tag{3}$$

Parameter transfer

Wine quality data



(a)

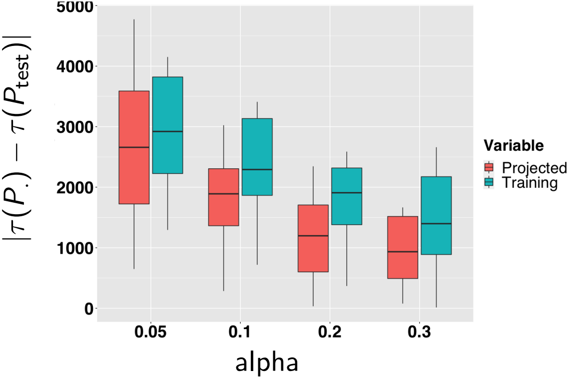


(b)

Figure. Wine quality data- transfer of regression coefficient of "pH" and "density". We add randomly chosen alpha proportion of samples from white to red wine to construct the training set.

Parameter transfer

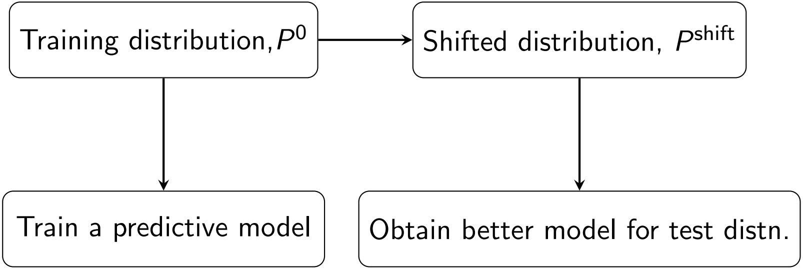
NSW data



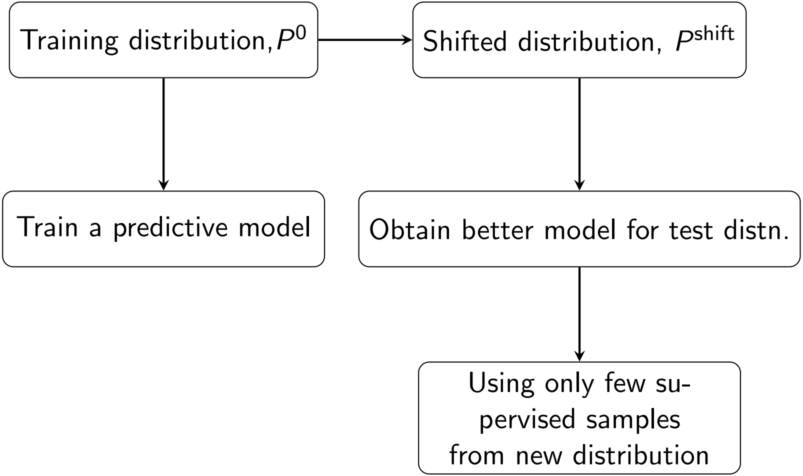
(a)

Figure. Transfer of ATE τ from training to test distribution. We use splits by (Dehejia and Wahba, 1999) and add randomly chosen alpha proportion of samples from one split to the other to construct the training set.

Model transfer



Model transfer



Model transfer

- Obtain a small test set with only few supervised samples $\{(X_i^s, Y_i^s)\}_{i=1}^m$.
- Let $R(P) = \sum_{i=1}^m \ell(f(\theta(P), X_i^s), Y_i^s)$ denote risk on test set for model $f(\theta, \cdot)$ obtained under distribution P.
- Try to transfer $R(P^0)$ to $R(P^{\text{shift}})$.

Model transfer

- Obtain a small test set with only few supervised samples $\{(X_i^s, Y_i^s)\}_{i=1}^m$.
- Let $R(P) = \sum_{i=1}^m \ell(f(\theta(P), X_i^s), Y_i^s)$ denote risk on test set for model $f(\theta, \cdot)$ obtained under distribution P.
- Try to transfer $R(P^0)$ to $R(P^{\text{shift}})$.
- Obtain

$$P^{\text{proj}} = \arg\min_{P \in \mathcal{P}} D_{KL}(P \| P_{0,n}) \text{ such that } \frac{1}{m} \sum_{i=1}^m \ell(f(\theta(P), X_i^s), Y_i^s) \leq \gamma, \tag{4}$$

choose $\gamma \in \mathbb{R}$ via cross-validation.

Model transfer

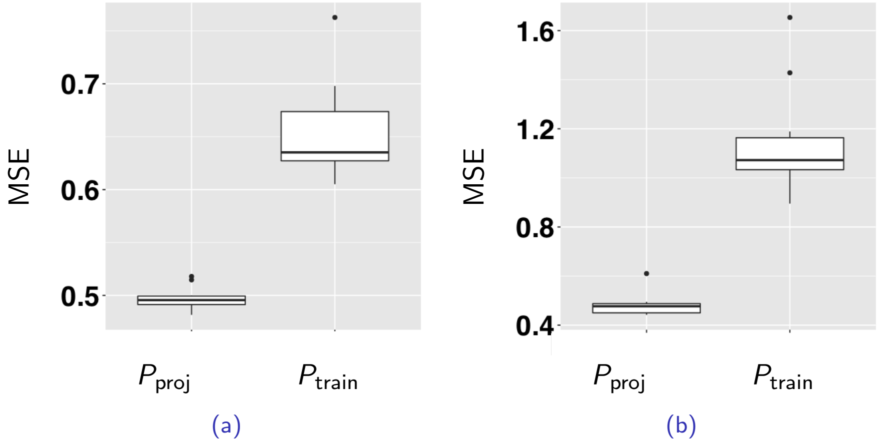


Figure. Wine quality data. MSE on new test set when predictive model is trained under a projected distribution vs training distribution. We mix α proportion of samples from one group to the other in each case.

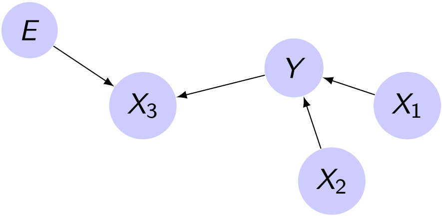


Figure: Causal graphical structure

Conclusion

- Shift in data generating distribution is inevitable due to which statistical knowledge may fail to generalize.
- We developed measures to understand distributional instability and further suggested steps to deal with it.

Based on the paper-

The s-value: evaluating stability with respect to distributional shifts. - Suyash Gupta and Dominik Rothenhaeusler. 2021.