Domain Adaptation Approaches for Exposure Models

Ron Sarafian

Ben-Gurion University of the Negev

ronsar@post.bgu.ac.il

Acknowledgements

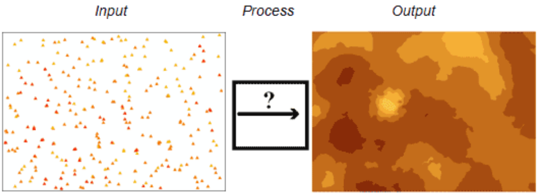
- Dr. Jonathan Rosenblatt
- Prof. Itai Kloog

Spatial Prediction

Spatial prediction

Examples:

- **Exposure Science**: Predicting air-pollution with satellite products
- **Precision Agriculture**: Predicting field's stress with drones imagery
- **Remote Sensing**: Classifying Land-Use with spectral data



Definition

the indirect measurement of some geo-located **output** (a.k.a labels), in places where this output is unknown, but some **input** data (a.k.a features) are available.

Spatial prediction as a supervised learning problem

- **Training set**: pairs of observed **inputs and outputs**, measured at known locations. (e.g., AOD and PM from ground monitoring stations)
- **Predictor**: obtained by applying some statistical **learning** algorithm, on the training set (e.g., Kriging; LMMs; Random Forest; Deep Networks)
- **Test set** (a.k.a prediction set): observed **inputs** in some area (e.g., $1km^2$ grid of AOD in Israel)
- **Predictions**: obtained by applying the **predictor** on the test set data

Formal supervised learning setup

Denote:

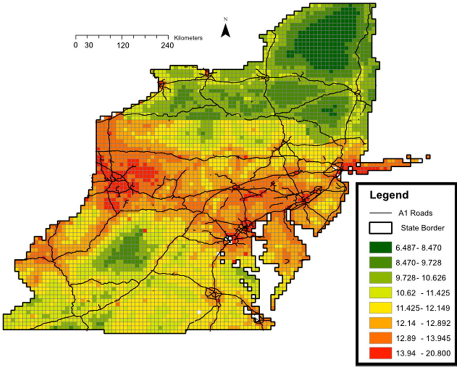
- $s \in S$ a **location**
- $y(s) \in Y$ **output** at location s
- $x(s) \in X$ **input** at location s
- $D = \{s_i, x_i, y_i\}$, $i = 1, ..., N$ a **training set**
- $D^* = \{s_j^*, x_j^*, y_j^*\}$, $j = 1, ..., M$ a **test set**

(ERM:) find $h : X \to Y$ that **minimize the Empirical Risk** w.r.t to some loss L (e.g., squared):

$$\hat{h} := \arg\min_{h \in H} \sum_{i=1}^{N} L\Big[h\big(x(s_i)\big), y(s_i)\Big]$$

Prediction

$$\hat{h}(D^*)$$

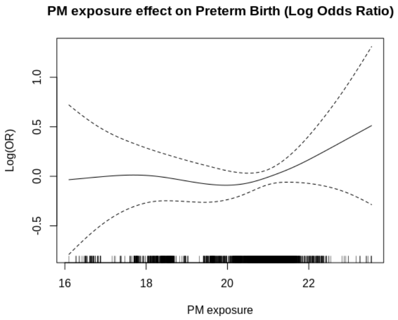


Predictions as Covariates

Spatial predictions themselves may serve as covariates in a subsequent research (*two-stage studies*).

Example (Epidemiology):

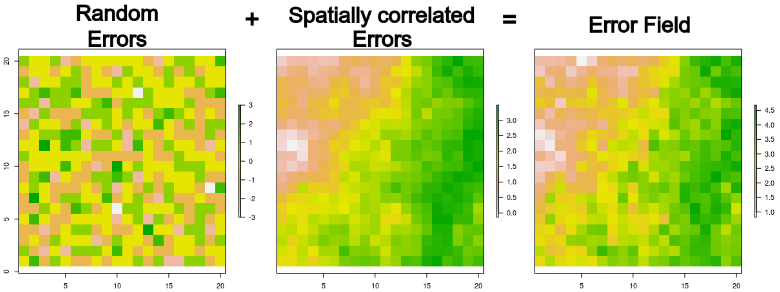
1. **First stage**: Predicting PM
2. **Second stage**: Estimating PM effect on health



Prediction errors are Epidemiological Errors-in-variables

Bad predictions in the exposure stage, are *errors-in-variables* (a.k.a *measurement errors*) in the epidemiological stage.

Errors in variables might lead to **erroneous second stage conclusions**



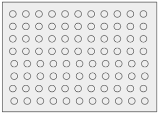
$\Rightarrow$ Prediction accuracy is important, from **epidemiological perspective!**

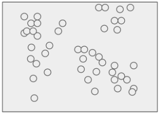
What is Prediction Accuracy?

What is considered an **accurate** predictor?

A one that does a small error on **test data**. However:

- What is the test set? (where should we predict)





- Does one accuracy measure fits all test sets?
- Test's outputs are not available. Can we use holdout methods?
- Does accurate predictions mean accurate second stage estimation?

Our talk

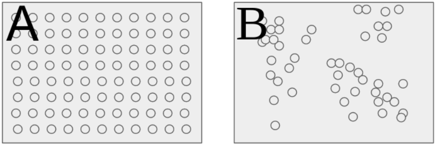
1. **Performance estimation**: Defining what is "good"
    - Spatial prediction as a Domain Adaptation problem
    - Empirical estimator and Cross-validation
2. **Spatial adaptation**: Using our framework to improve prediction
    - Real data results
3. **Optimal Design adaptation**: Using our framework to improve second-stage estimation
    - Algorithm
    - Real data Preliminary results

Performance Estimation

Prediction task and the test set

The **prediction task** determines the test set, and vice-versa.

E.g., different spatial prediction tasks:



Predicting a grid

Predicting specific locations (e.g., residence of patients)

- Do we expect a predictor to have the same accuracy in both tasks?
- If one predictor is good at predicting **A**, is it also good at **B**?

Intuition: Prediction task should determine the evaluation metric.

When Training and Test data are are not similar

Test set's outputs are unknown... **How can we estimate predictor's performance?**

Standard machine-learning answer: Use **holdouts** approaches (e.g., cross-validation)

But what if the training and test data are **not so similar**?

Formally:

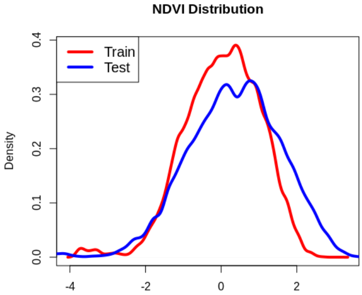
If **training data distribution** $\neq$ **test data distribution**, then **naive** cross-validation is **biased**

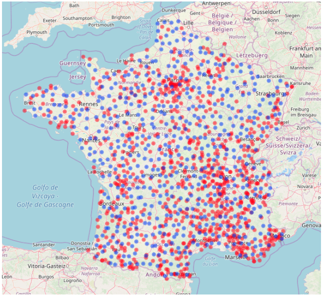
When training data distribution $\neq$ test data distribution

Example:

monitoring stations (training set) are mostly in **urban** areas, but region of interest (test set) consist of mostly **rural** areas.

Inputs (e.g., NDVI) **distribution differ** between urban and rural regions





Task-tailored performance estimation

Can we design a **criterion** that is tailored to the **prediction task**?

Potential usages:

- Quality assessment
- Model training
- Model selection

Our observation:

Any prediction task implies certain spatial **importance weights** that can be incorporated into the **decision-theoretic learning** framework.

Our criterion borrows ideas from the literature on **Domain Adaptation**.

Domain Adaptation

Domain Adaptation

Designed to learn a predictor from one population that follows a source distribution, and apply it in some other population that follows a target distribution. (Pan et. al., 2009)

Example: Train classifiers with amazon photos; predict on mobile camera photos (Gonq et al., CVPR 2012)

Domain Adaptation

Domain adaptation deals with learning

how to learn a predictor on one domain, and adapt it to another?

We use it for model validation

how to estimate the performance of a given predictor with samples from the "wrong" distribution?

- We will view the training set as samples from the **source domain**, and the test set as samples from the **target domain**.
- We will then **reweight** each data point from the training/source so they resemble a sample from the test/target. (importance weighting)

Data Generating Distribution

Data generating distribution differ between the **source** and the **target**.

- Training samples: $P_S = P_S(s, x, y)$
- Test samples: $P_T = P_T(s, x, y)$

The researcher may specify a prediction task by specifying P_T.

The Target Risk

How good is the predictor on the average target point?

$$\epsilon_T(h) := E_{P_T} L(h(x), y) \quad (1)$$

Two major problems:

1. Specifying P_T is extremely hard
2. We only have samples from P_S, how can we estimate $\epsilon_T(h)$?

Specifying $P_T(s,x,y)$

$$P_T(s,x,y) = \underbrace{P_T(s)}_{\text{spatial}}\,\underbrace{P_T(x,y|s)}_{\text{local}}$$

Specifying $P_T(s,x,y)$ is greatly simplified if one observes that:

$P_T(x,y|s)$ does not depend on the researcher's task.

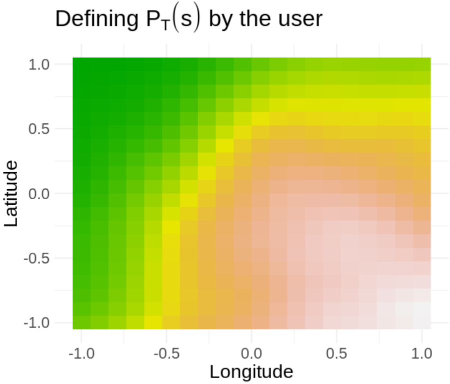
- Thus $P_T(x,y|s) = P_S(x,y|s)$ (Covariate-shift)
- $P_S(x,y|s)$ can be estimated from the training data.

$P_T(s)$ is merely a 2D spatial weights function

- Can be thought of as the **target importance** allocated to location s

$P_T(s)$: Target Importance

Specifying the spatial target distribution $P_T(s)$ by our self:



Importance-Weighted Source Risk (IWSR)

The **IWSR** empirical estimator of the **Target Risk**:

$$\epsilon_S^{IW}(h) := N^{-1} \sum_{i=1}^{N} \omega(s_i) L(h(x_i), y_i), \tag{2}$$

We show (Sarafian et. al., 2020):

If $\omega(s_i) = \frac{P_T(s_i)}{P_S(s_i)}$, then IWSR is an **unbiased** estimate of the **target risk**:

$$E[\epsilon_S^{IW}(h)] = \epsilon_T(h) \tag{3}$$

- $P_S(s)$ is estimated from the data
- $P_T(s)$ is specified by the user

IWSR unbiasedness

In expectation over training sets:

$$E\left[\epsilon_S^{IW}(h)\right] = \int_S \int_X \int_Y P_S(s,x,y)\left[\frac{1}{N}\sum_{i=1}^{N}\omega(s)L(h(x),y)\right]dy\ dx\ ds$$

$$= \int_S P_S(s)\ \omega(s) \int_X \int_Y P_S(x,y|s)\left[\frac{1}{N}\sum_{i=1}^{N}L(h(x),y)\right]dy\ dx\ ds$$

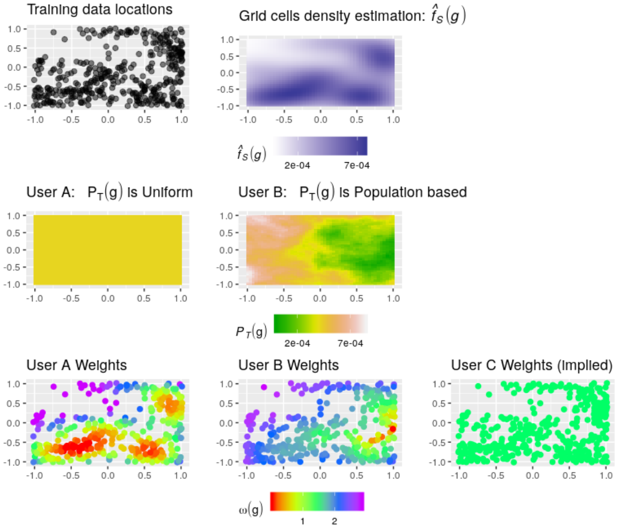
$$= \int_S P_S(s)\frac{P_T(s)}{P_S(s)}E_{P_T(x,y|s)}\left[L(h(x),y)\right]\Big)\ ds$$

$$= \int_S P_T(s)E_{P_T(x,y|s)}\left[L(h(x),y)\right]\ ds$$

$$= E_{P_T(s,x,y)}\left[L(h(x),y)\right] = \epsilon_T(h)$$

* shared support condition: $\nexists s$ s.t. $P_T(s) \neq 0$, and $P_S(s) = 0$

Toy Example


Training data locations
Grid cells density estimation: $\hat{f}_S(g)$
$\hat{f}_S(g)$
2e-04
7e-04
User A: $P_T(g)$ is Uniform
User B: $P_T(g)$ is Population based
$P_T(g)$
2e-04
7e-04
User A Weights
User B Weights
User C Weights (implied)
$\omega(g)$
1
2

Cross-validation

The IWSR is not immune to the usual overfitting problem.
Remedy: Compute it on held-out samples.

Importance-weighted k-fold cross-validation (IWKF)

$$\epsilon_S^{IWKF} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|D_k|}\sum_{i\in D_k}\omega(s_i)L(h_{D_{-k}}(x_i), y_i) \tag{4}$$

We show (Sarafian et. al., 2020):

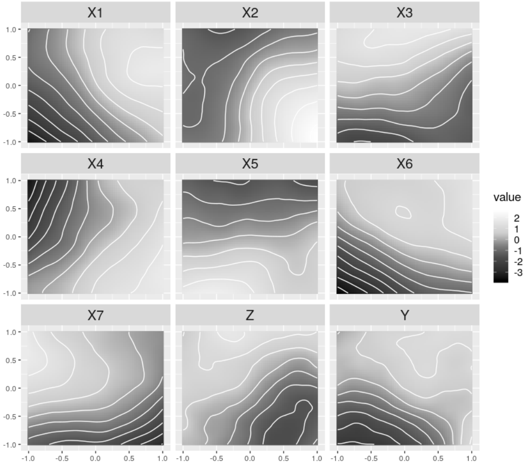
When some conditions hold, for any training set D of size N

$$E_D[\epsilon_S^{IWKF}(h)] = G_{N-\frac{N}{K}} \tag{5}$$

$G_{N-\frac{N}{K}}$ is the expected target risk over all training sets of size $N - \frac{N}{K}$

Simulated data analysis (1)

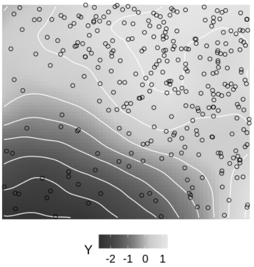
input data: $x = (x_1, ..., x_7)$
hidden field: z
output data: $y = g(x, \theta) + \delta z$

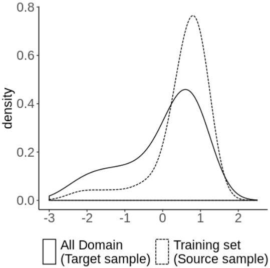


Simulated data analysis (2)

Setup:

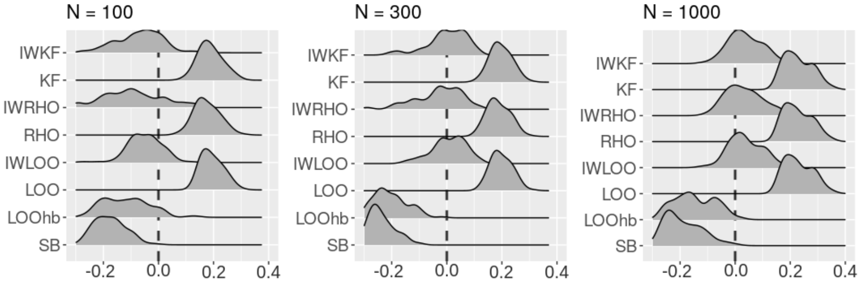
1. Simulate new data
2. Simulate new locations
3. Fit $\hat{h}(x)$
4. Estimate $\epsilon_T(\hat{h})$
5. Compare $\epsilon_T(\hat{h}) - \hat{\epsilon}(\hat{h})$
6. Repeat





Simulated data analysis (3)

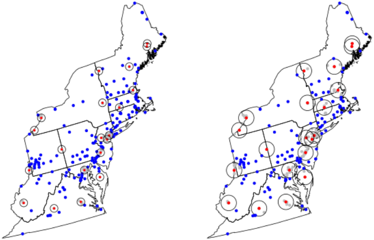
Density of $\epsilon_T(\hat{h}) - \hat{\epsilon}(\hat{h})$ for IW estimators vs. naive / blocking approaches:



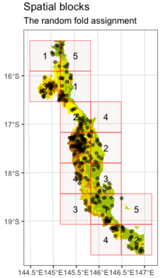
- **IWSR** cross-validation estimates are **unbiased**
- **Naive** cross-validation estimates **underestimate** the target risk
- **Blocking** cross-validation estimates **overestimate** the target risk

Comparing to other validation approaches

- Some authors suggest spatial variants of h-block cross-validation (Burman, 1994) or other spatial folding mechanisms, to enforce **spatially uncorrelated training** and **validation** sets.



Sarafian et al. (2019)



Valavi et al. (2018)

- We observe that removing **training**-**validation** correlation is **not** (always) **desired** – it impose unnecessary extrapolation.

Comparing to other validation approaches

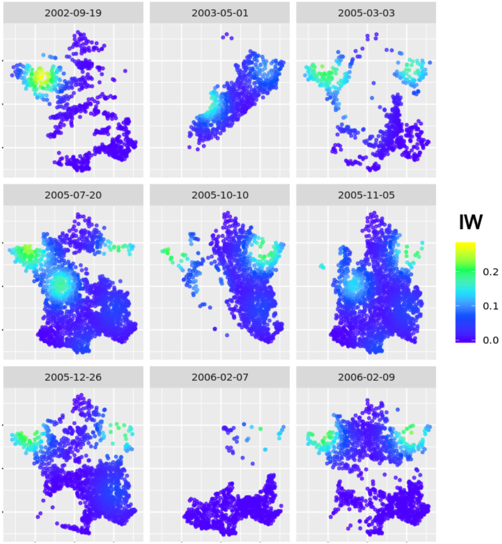
Our observation:

There is no harm in spatially correlated **training**-**validation** sets, as long as the **source**-**target** data experience the same correlation structure.

IWSR maintains the desired structure between sets through weighting; ensuring that **validation** sets would have the **same probabilistic properties** as the **target** data.

Spatial adaptation

Improving prediction by spatial adaptation



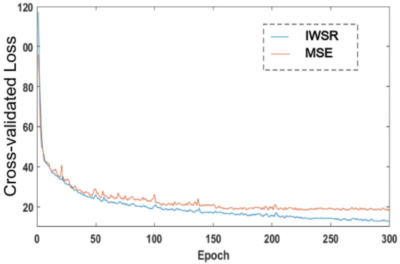
Steps:

- Estimate $P_S(s)$
- Estimate $P_T(s)$
- Calculate $\omega(s)$

IWSR minimization

- Find $h : X \to Y$ that **minimize IWSR** w.r.t to the loss L:

$$\hat{h} := \arg\min_{h \in H} \sum_{i=1}^{N} \omega(s_i) L\Big[h\big(x(s_i)\big), y(s_i)\Big]$$



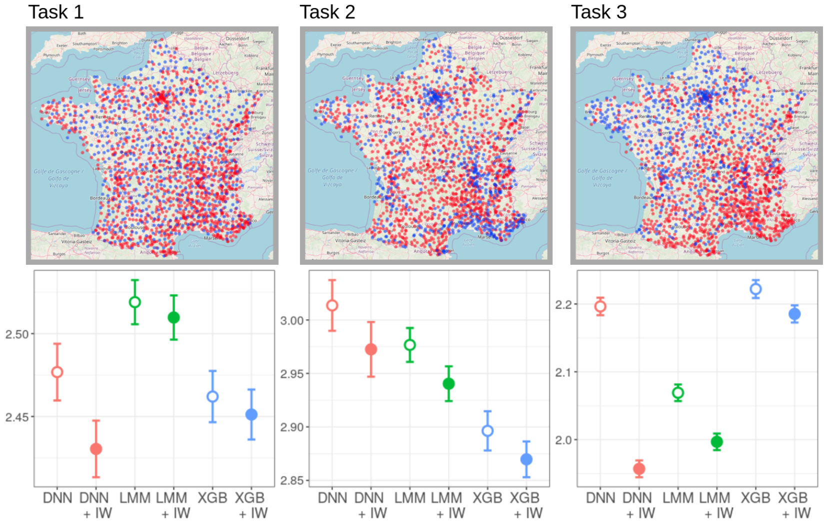
Learning through minimizing the IWSR improves spatial prediction

Real Data Analysis (1)

Air-temperature prediction from remote sensing data and ground stations in France.

- 3 different spatial target distributions where defined, from which test locations where sampled
- Training sets were sampled from the remaining locations
- 3 type of predictors: LMM; XGB; DNN, each trained by minimizing the MSE or the IWSR
- Their performance are compared the average test set loss (the "target risk").

Real Data Analysis (2)



Optimal Design adaptation

The Second Stage

PM-health example:

- Epidemiologist assumes a parametric model $M(y, z; \beta)$
- y is PM
- z is Low-birth-weight
- β are model's parameters, including PM effect.
- $\hat{\beta}$ parameters' estimates

For example, M may be a *Logistic regression*:

$$z|y \sim Binom(p = \frac{e^{\beta_0+\beta_1 y}}{1+e^{\beta_0+\beta_1 y}}).$$

So the second-stage goal is **accurate estimation**, e.g.:

$$min\{E||\beta - \hat{\beta}||^2\}.$$

The Second Stage

Note: For the epidemiologist, y **is unknown**, and is replaced by $\hat{y} = h(x)$.

Potentially, the **one-stage** problem is to find $h^* : X \to Y$, s.t.:

$$h^* := \arg\min_{h \in H} \left\{ E \left|\left| \hat{\beta}(h(x); z) - \beta \right|\right|^2 \right\}.$$

But we restrict our self to the **two-stage** framework:

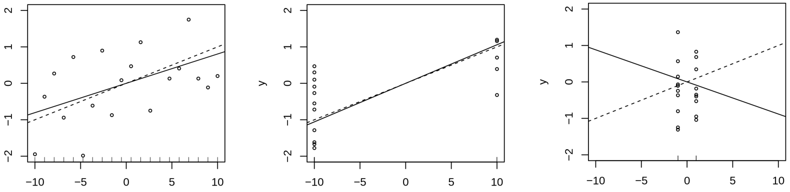
- Can we make first-stage that **improve the second-stage**? Yes.
- Does accurate **first-stage** prediction necessarily mean accurate **second-stage** estimation? No.

Adapting to the second stage

The idea: Instead of adapting to the locations of subjects, we now **adapt to the second stage** itself.
Do all samples have the same contribution to $E[(\beta_1 - \hat{\beta}_1)^2]$? No.

Example: Linear second-stage model: $z = \beta_0 + \beta_1 y + \varepsilon$

It turns out that $E[(\beta_1 - \hat{\beta}_1)^2]$ is minimized when $Var[y]$ is maximized



Optimal Design

Optimal design deals with the identification of **favorable sampling points**. I.e., points that **yield accurate estimates**

Usually we look for points $\tilde{\xi}$, s.t.:

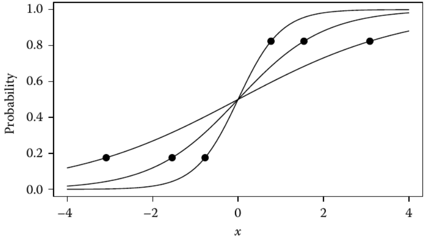
$$\tilde{\xi} := \arg\min_{\xi}\{F(Var[\hat{\beta}_{\xi}])\}$$

- A-optimality: $F() := trace()$
- D-optimality: $F() := det()$
- E-optimality: $F() := \lambda_1()$

Optimal Design

- **Linear Models**
  - Finite sample variance is known
  - Optimum is independent of the unknown parameters β
- **Non-linear Models**
  - Finite sample variance is unknown $\rightarrow$ Use limiting variance
  - Optimum depends on unknown β $\rightarrow$ sequential design, prior, etc.

Example: D-optimal points in Logistic Regression:



Algorithm - Informal

1. Esther **provides** exposures.
2. Ephraim uses them to **estimate** exposure effects.
3. Ephraim uses **optimal design** theory to mark data points of importance.
4. Esther uses **weighted ERM** to improve predictions at those locations.

Algorithm

function EE ESTIMATOR$(\mathcal{A}, \mathcal{M}, D^*, \mathcal{K}, w^0)$
    $\hat{x}^1 \leftarrow \mathcal{A}_{D^*, w^0}$ ▷ Initialize exposures
    for $l \in \{1, ..., L\}$ **do**
        $\hat{\beta}^l \leftarrow \mathcal{M}_{\hat{x}^l}$ ▷ Estimate EE with current exposures
        $\tilde{\xi}^l \leftarrow \arg\max_\xi \left\{ \det(I(\beta; \xi, \hat{\beta}^l)) \right\}$ ▷ Find D-optimal design
        $w_i^l \leftarrow \max_{x \in \tilde{\xi}} \{\mathcal{K}(x, x_i)\}, \forall i \in \{1, \ldots, n^*\}$ ▷ Weight x_i using distance from $\tilde{\xi}$
        $\hat{x}^{l+1} \leftarrow \mathcal{A}_{D^*, w^l}$ ▷ Update exposures using current weights
    end for
    return $\hat{\beta}^L$
end function

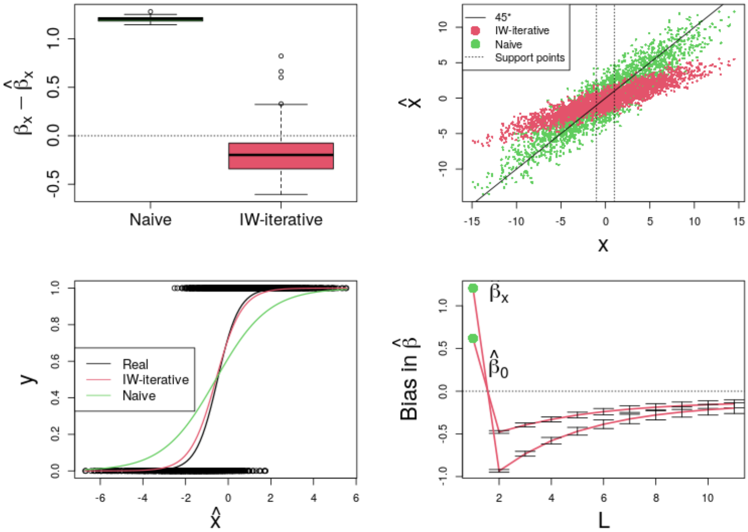
Simulation Results

- **First stage**
  - Data: Gaussian Process: $y \sim N\big(x'\theta, \text{Matérn}(s;\alpha)\big)$
  - Predictors: lm / gam / RF
- **Second stage**
  - Data: Binomial: $z|y \sim Binom\big(p = \frac{e^{\beta_0+\beta_1\hat{y}}}{1+e^{\theta_0+\theta_1 y}}\big).$
  - Model: Logistic Regression
  - Univariate exposure
  - No other epidemiological covariates

Simulation

function EE ESTIMATOR($\mathcal{A}, \mathcal{M}, D^*, \mathcal{K}, w^0$)
  $\hat{x}^1 \leftarrow \mathcal{A}_{D^*,w^0}$ ▷ Initialize exposures
  for $l \in \{1, ..., L\}$ **do**
    $\hat{\beta}^l \leftarrow \mathcal{M}_{\hat{x}^l}$ ▷ Estimate EE with current exposures
    $\tilde{\xi}^l \leftarrow \arg\max_{\xi} \left\{ \det(I(\beta; \xi, \hat{\beta}^l)) \right\}$ ▷ Find D-optimal design
    $w_i^l \leftarrow \max_{x \in \tilde{\xi}} \{\mathcal{K}(x, x_i)\}, \forall i \in \{1, \ldots, n^*\}$ ▷ Weight x_i using distance from $\tilde{\xi}$
    $\hat{x}^{l+1} \leftarrow \mathcal{A}_{D^*,w^l}$ ▷ Update exposures using current weights
  end for
  return $\hat{\beta}^L$
end function

Simulation Results



Simulation Results

Things to note

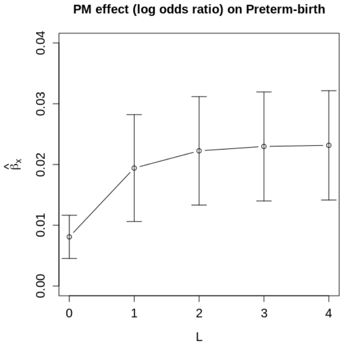
- Estimation error of β is smaller using our iterative estimator.
- Predicted exposure is worse on average. A paradox reported by Szpiro, Paciorek, and Sheppard (2011).
- More impactful in non-linear models.
- Multiple random initializations to deal with overfitting.

Empirical Results

Applying the IW-iterative algorithm on real data:

- Southern Israel, 2004-2014
- Soroka birth data, all births.
- **Response:** Preterm birth (binary).
- **Exposure:**Average PM2.5 exposure in the last 30 days of pregnancy (pmlast30).
- **Controls:** time trends, regional characteristics, mother characteristics (socioeconomic status, age, ethnicity), infant sex.

Empirical Results



Disclaimers

- Model misspecification
- Covariates
- Confounders.

Extantions

- Other epidemiological models (LM, GLM, Survival,. . .) only requires existence of optimal design theory.
- From spatio to spatio-temporal
- Multivariate exposure
- Tailor predictor to covariates.
- Private exchanges of covariates
- Other domains

Summary

Take Home (1)

- What is a good spatial predictor? It is not well defined.
  - Depending on the **task / test set**
  - The **target distribution** define the task
  - **Domain Adaptation** can be used to adapt to this distribution
  - IWSR is an **unbiased** estimate of the **target risk**
- Minimizing the IWSR in the ERM framework is recommended
  - Improve prediction by **adapting the learning** to the task (e.g., entire grid / specific locations)

Take Home (2)

- Can we use the Domain Adaptation to improve the second stage?
  - Yes, using the **iterative-IW algorithm**:
  - **D-optimal** Design points to improve prediction
  - **ERM with IWSR** to adapt prediction

References

Sarafian, Ron, et al. "Gaussian Markov Random Fields versus Linear Mixed Models for satellite-based PM2.5 assessment: Evidence from the Northeastern USA." Atmospheric Environment (2019).

Sarafian, Ron, et al. "A Domain Adaptation Approach for Performance Estimation of Spatial Predictions." IEEE Transactions on Geoscience and Remote Sensing (2020).

Hough, I., Sarafian, R., Shtein, A., Zhou, B., Lepeule, J., Kloog, I. (2021). Gaussian Markov random fields improve ensemble predictions of daily 1 km PM2. 5 and PM10 across France. Atmospheric Environment, 264, 118693.

Sarafian, Ron, Itai Kloog, and Jonathan D. Rosenblatt. "Optimal-design domain-adaptation for exposure prediction in two-stage epidemiological studies." Journal of Exposure Science Environmental Epidemiology (2022): 1-8.

Burman, Prabir, Edmond Chow, and Deborah Nolan. "A cross-validatory method for dependent data." Biometrika 81.2 (1994): 351-358.

Valavi, Roozbeh, et al. "blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models." Biorxiv (2018): 357798.

Thank you!