ECML2008 Sep. 15-19, 2008

Pool-based Agnostic Experiment Design in Linear Regression

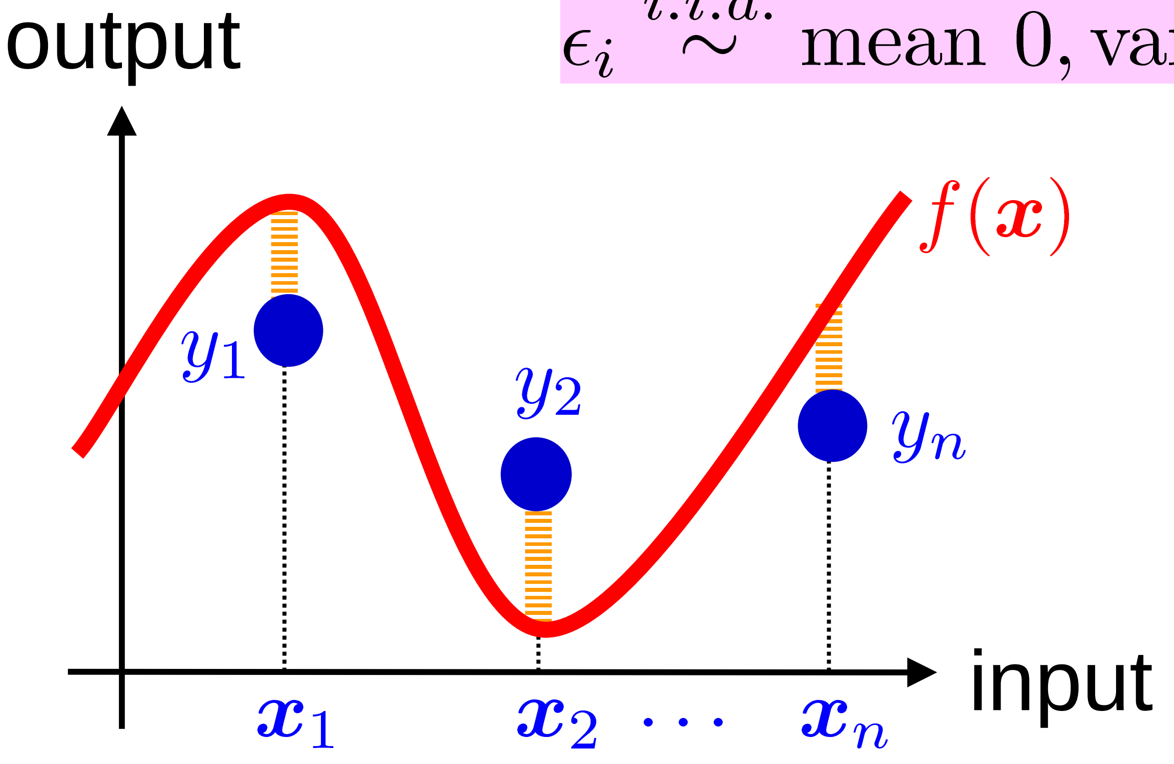
Masashi Sugiyama (Tokyo Tech.)
Shinichi Nakajima (Nikon)

Linear Regression

- Learn a real-valued function $f(\boldsymbol{x})$ from input-output training samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$.

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

$$\epsilon_i \overset{i.i.d.}{\sim} \text{mean } 0, \text{variance } \sigma^2$$



Linear Regression (cont.)

- Linear model is used for learning:

$$\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

α_i : Parameter

$\varphi_i(\boldsymbol{x})$: Basis function

- Goal: learn $\hat{\boldsymbol{\alpha}}$ so that the generalization error is minimized

$$Gen = \mathbb{E}_{\boldsymbol{\epsilon}} \int \left(f(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x}) \right)^2 p_{test}(\boldsymbol{x}) d\boldsymbol{x}$$

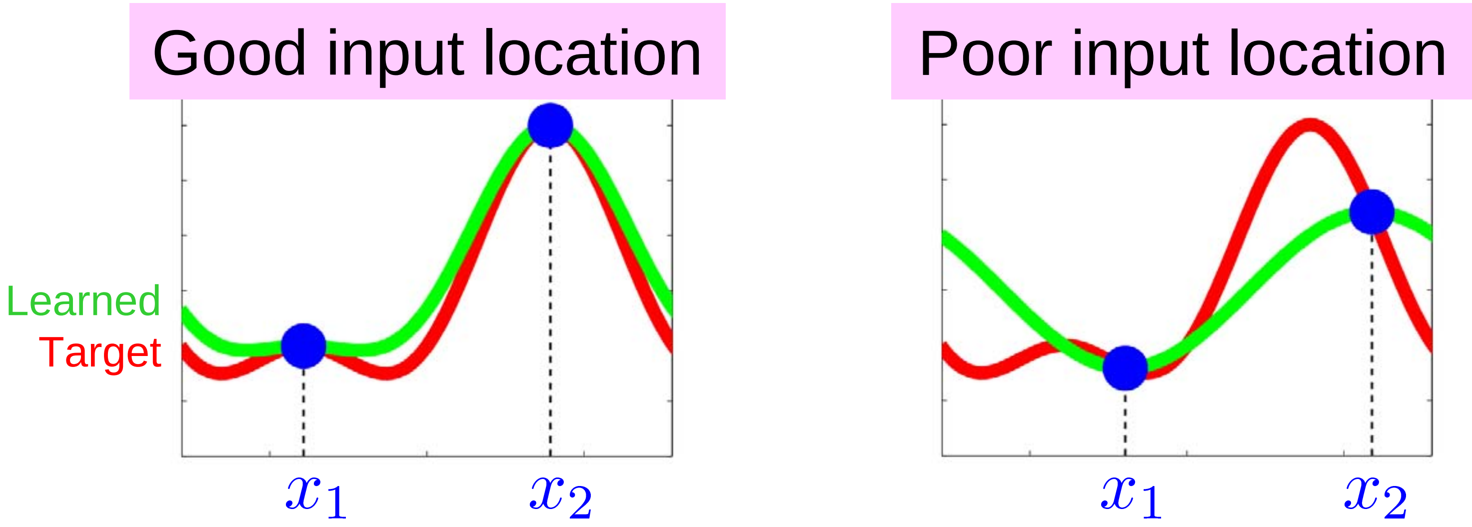
$$= \mathbb{E}_{\boldsymbol{\epsilon}} \| f - \widehat{f} \|_{p_{test}}^2$$

$\mathbb{E}_{\boldsymbol{\epsilon}}$: Expectation over noise

$p_{test}(\boldsymbol{x})$:Test input density

Experiment Design

- Quality of learned functions depends on training input location $\{\boldsymbol{x}_i\}_{i=1}^n$.



- Goal: optimize training input location

$$\min_{\{\boldsymbol{x}_i\}_{i=1}^n} Gen$$

$$Gen = \mathbb{E}_{\boldsymbol{\epsilon}} \| f - \widehat{f} \|_{p_{test}}^2$$

Challenges

$$\min_{\{\boldsymbol{x}_i\}_{i=1}^n} Gen$$

$$Gen = \mathbb{E}_{\boldsymbol{\epsilon}} \|f - \widehat{f}\|_{p_{test}}^2$$

- Gen is unknown and needs to be estimated.
- In experiment design, we do not have training output values $\{y_i\}_{i=1}^n$ yet.
- Thus we cannot use, e.g., cross-validation which requires $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$.
- Only training input positions $\{\boldsymbol{x}_i\}_{i=1}^n$ can be used in generalization error estimation!

Organization

Pool-based Agnostic Experiment Design in Linear Regression

1. Problem definition
2. Basic strategy
3. Proposed method
4. Experiments

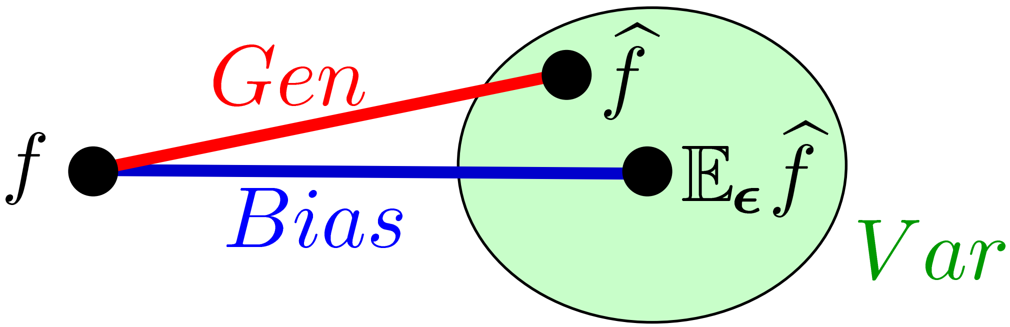
Bias and Variance

$$\underbrace{\mathbb{E}_{\boldsymbol{\epsilon}}\|f-\widehat{f}\|^2_{p_{test}}}_{Gen} = \underbrace{\|f-\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{f}\|^2_{p_{test}}}_{Bias} + \underbrace{\mathbb{E}_{\boldsymbol{\epsilon}}\|\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{f}-\widehat{f}\|^2_{p_{test}}}_{Var}$$

- $Bias$ is not estimable without $\{y_i\}_{i=1}^n$.
- For linear learning $\widehat{\boldsymbol{\alpha}} = \boldsymbol{L}\boldsymbol{y}$:

$$Var = \sigma^2 \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{L}^\top)$$

  - Noise variance σ^2 is not estimable without $\{y_i\}_{i=1}^n$.
  - $\mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{L}^\top)$ is computable from $\{\boldsymbol{x}_i\}_{i=1}^n$.



$\boldsymbol{U}_{i,j} = \langle \varphi_i, \varphi_j \rangle_{p_{test}}$

$\boldsymbol{L}$: Learning matrix

$\boldsymbol{y} = (y_1, \ldots, y_n)^\top$

Key Trick in Experiment Design

- Find a setup where $Bias = 0$ is guaranteed.
- Then

$$Gen = \underbrace{Bias}_{0} + \underbrace{Var}_{\sigma^2 \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{L}^\top)} \propto \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{L}^\top)$$

- Thus

$$\mathrm{argmin}\, Gen = \mathrm{argmin}\, \underbrace{\mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{L}^\top)}_{\text{computable before observing } \{y_i\}_{i=1}^n}$$

$\boldsymbol{U}_{i,j} = \langle \varphi_i, \varphi_j \rangle_{p_{test}}$

$\boldsymbol{L}$: Learning matrix

Traditional Method

(Fedorov 1972)

- Assume model is correct:

$$^{\exists}\boldsymbol{\alpha}^*,\ \widehat{f}(\boldsymbol{x};\boldsymbol{\alpha}^*) = f(\boldsymbol{x})$$

$$\widehat{f}(\boldsymbol{x};\boldsymbol{\alpha}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

- Use ordinary least squares (OLS) estimation:

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\widehat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

$$\widehat{\boldsymbol{\alpha}}_O = \boldsymbol{L}_O \boldsymbol{y}$$

- Experiment design criterion:

$$\min_{\{\boldsymbol{x}_i\}_{i=1}^{n}} \ \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_O\boldsymbol{L}_O^\top)$$

$$\boldsymbol{L}_O = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$$

$$\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i)$$

$$\boldsymbol{y} = (y_1, \ldots, y_n)^\top$$

$$\boldsymbol{U}_{i,j} = \langle \varphi_i, \varphi_j \rangle_{p_{test}}$$

Goal of This Work

- Pros / cons of traditional method:
  - \+ Generalization error estimation is exact.
  - \+ Easy to implement.
  - \- Correct-model assumption is not realistic.
  - \- Very poor performance when agnostic.
  - \- Test input density $p_{test}(\boldsymbol{x})$ is often unknown.
- We propose a new method that is
  - Still easy to implement,
  - Robust against agnosticity,
  - Able to work without $p_{test}(\boldsymbol{x})$.

Organization

Pool-based Agnostic Experiment Design in Linear Regression

1. Problem definition
2. Basic strategy
3. Proposed method
    1. Overcoming agnosticity
    2. Coping with pool-based setup
4. Experiments

Weak Agnostic Setup

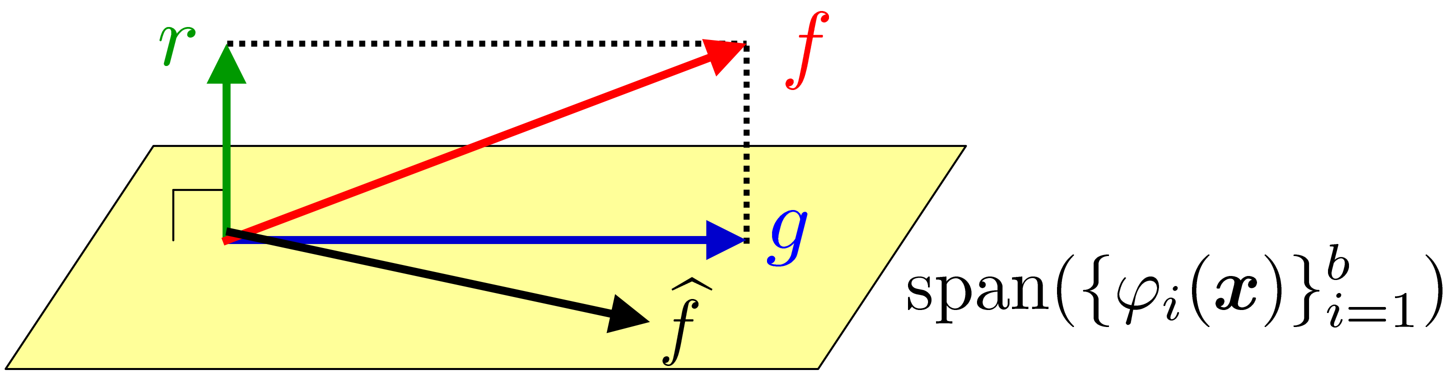
- The model is not exactly correct, but is reasonably good:

$$^{\exists}\boldsymbol{\alpha}^*, \ \|\widehat{f}(\boldsymbol{x};\boldsymbol{\alpha}^*) - f(\boldsymbol{x})\| \approx 0$$

$$\widehat{f}(\boldsymbol{x};\boldsymbol{\alpha}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

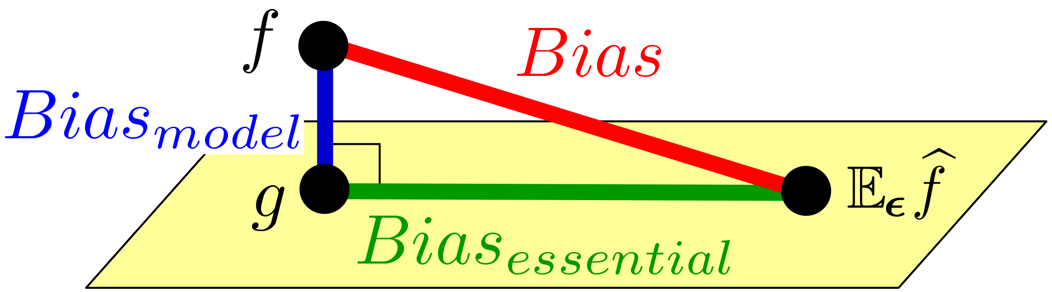
- Decomposition of target function:

$$\underbrace{f(\boldsymbol{x})}_{\text{Target function}} = \underbrace{g(\boldsymbol{x})}_{\text{Approximable part}} + \underbrace{r(\boldsymbol{x})}_{\text{Residual part}}$$



Further Decomposition of Bias

$$\underbrace{\|f - \mathbb{E}_{\epsilon}\widehat{f}\|^2_{p_{test}}}_{Bias} = \underbrace{\|f - g\|^2_{p_{test}}}_{Bias_{model}} + \underbrace{\|g - \mathbb{E}_{\epsilon}\widehat{f}\|^2_{p_{test}}}_{Bias_{essential}}$$



- $Bias_{model}$ is constant and ignorable.
- But OLS cannot make $Bias_{essential}$ zero due to "covariate shift": (Shimodaira JSPI2000)
  - Training / test inputs follow different distributions.
  - "Covariate" is another name for "input".

Importance-Weighted LS (IWLS)

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \underbrace{\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)}}_{\text{Importance}} \left(\widehat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

$$\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} p_{train}(\boldsymbol{x})$$

- Even when agnostic: $\lim_{n\to\infty} Bias_{essential} = 0$
- When weak agnostic: $Bias_{essential} \ll Var$
- Solution is given by

$$\widehat{\boldsymbol{\alpha}}_W = \boldsymbol{L}_W \boldsymbol{y}$$

$$\boldsymbol{L}_W = (\boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D}$$

$$\boldsymbol{D} = \mathrm{diag}\left(\frac{p_{test}(\boldsymbol{x}_1)}{p_{train}(\boldsymbol{x}_1)}, \ldots, \frac{p_{test}(\boldsymbol{x}_n)}{p_{train}(\boldsymbol{x}_n)} \right)$$

$$\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i) \qquad \boldsymbol{y} = (y_1, \ldots, y_n)^\top$$

Justification

(Sugiyama JMLR2006)

- For IWLS

$$Gen = \underbrace{Bias_{model}}_{constant} + \underbrace{Bias_{essential}}_{\ll Var} + \underbrace{Var}_{\sigma^2 \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_W\boldsymbol{L}_W^\top)}$$

- Thus

$$\underset{p_{train}}{\mathrm{argmin}}\, Gen \approx \underset{p_{train}}{\mathrm{argmin}}\, \underbrace{\mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_W\boldsymbol{L}_W^\top)}_{\text{computable before observing } \{y_i\}_{i=1}^n}$$

$$\boldsymbol{L}_W = (\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}$$

$$\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i) \quad \boldsymbol{U}_{i,j} = \langle \varphi_i, \varphi_j \rangle_{p_{test}}$$

$$\boldsymbol{D} = \mathrm{diag}\left(\frac{p_{test}(\boldsymbol{x}_1)}{p_{train}(\boldsymbol{x}_1)}, \ldots, \frac{p_{test}(\boldsymbol{x}_n)}{p_{train}(\boldsymbol{x}_n)}\right)$$

Organization

Pool-based Agnostic Experiment Design
in Linear Regression

1. Problem definition
2. Basic strategy
3. Proposed method
    1. Overcoming agnosticity
    2. Coping with pool-based setup
4. Experiments

Pool-based Setup

- Pool-based setup:
  - The test input density $p_{test}(\boldsymbol{x})$ is unknown.

    Importance weight is not accessible.

$$\boldsymbol{D} = \mathrm{diag}\left(\frac{p_{test}(\boldsymbol{x}_1)}{p_{train}(\boldsymbol{x}_1)}, \ldots, \frac{p_{test}(\boldsymbol{x}_n)}{p_{train}(\boldsymbol{x}_n)}\right)$$

  - But a pool of test input samples is given.

$$\{\boldsymbol{x}'_i\}_{i=1}^{N} \overset{i.i.d.}{\sim} p_{test}(\boldsymbol{x})$$

  - Training input points are chosen from the pool.

$$\{\boldsymbol{x}_i\}_{i=1}^{n} \subset \{\boldsymbol{x}'_i\}_{i=1}^{N}$$

- We assume $N \gg n$.

Computing Importance Weight

- $\{b(\boldsymbol{x}'_i)\}_{i=1}^N$: Resampling probability of $\{\boldsymbol{x}'_i\}_{i=1}^N$

$$\sum_{i=1}^N b(\boldsymbol{x}'_i) = 1, \quad b(\boldsymbol{x}'_i) \geq 0$$

- Choose $\{\boldsymbol{x}_i\}_{i=1}^n$ following $\{b(\boldsymbol{x}'_i)\}_{i=1}^N$.

$$\{\boldsymbol{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \{b(\boldsymbol{x}'_i)\}_{i=1}^N$$

- Then $\boldsymbol{D}$ can be exactly computed:

$$p_{train}(\boldsymbol{x}_i) = p_{test}(\boldsymbol{x}_i) b(\boldsymbol{x}_i)$$

$$\{\boldsymbol{x}'_i\}_{i=1}^N \overset{i.i.d.}{\sim} p_{test}(\boldsymbol{x})$$

$$\Rightarrow \frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)} = \frac{1}{b(\boldsymbol{x}_i)}$$

$$\Rightarrow \boldsymbol{D} = \mathrm{diag}\left(\frac{1}{b(\boldsymbol{x}_1)}, \ldots, \frac{1}{b(\boldsymbol{x}_n)}\right)$$

Proposed Method

- Choose resampling function based on

$$\min_{b} \mathrm{tr}(\widehat{\boldsymbol{U}} \boldsymbol{L}_W \boldsymbol{L}_W^\top)$$

$$\{\boldsymbol{x}'_i\}_{i=1}^N \overset{i.i.d.}{\sim} p_{test}(\boldsymbol{x})$$

$$\{\boldsymbol{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \{b(\boldsymbol{x}'_i)\}_{i=1}^N$$

$$\widehat{\boldsymbol{U}}_{i,j} = \frac{1}{N}\sum_{i=1}^{N} \varphi_i(\boldsymbol{x}'_i)\varphi_j(\boldsymbol{x}'_i)$$

$$\boldsymbol{L}_W = (\boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D}$$

$$\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i)$$

$$\boldsymbol{D} = \mathrm{diag}\left(\frac{1}{b(\boldsymbol{x}_1)}, \ldots, \frac{1}{b(\boldsymbol{x}_n)}\right)$$

- Advantages:
  - Robust against model misspecification.
  - Easy to implement.

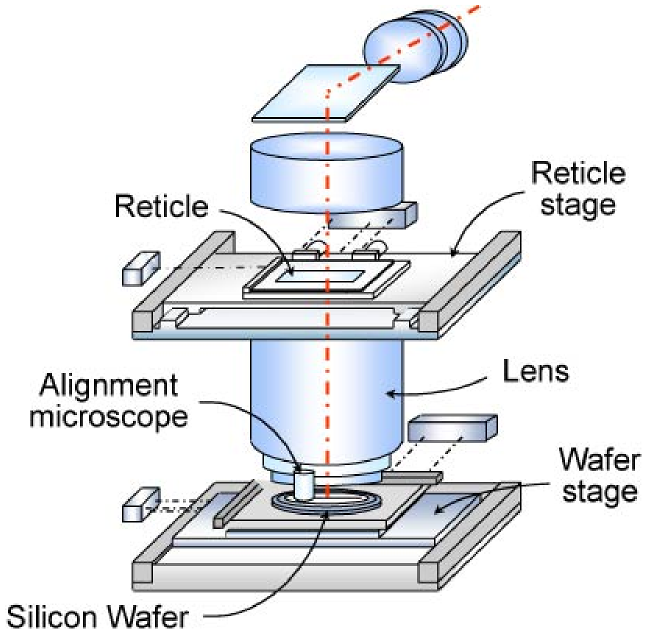
Organization

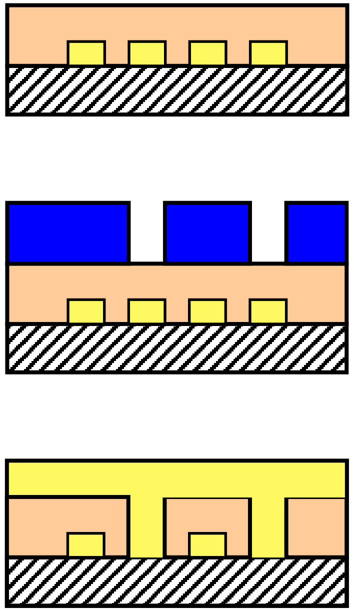
Pool-based Agnostic Experiment Design
in Linear Regression

1. Problem definition
2. Basic strategy
3. Proposed method
4. Experiments

Wafer Alignment in Semiconductor Exposure Apparatus

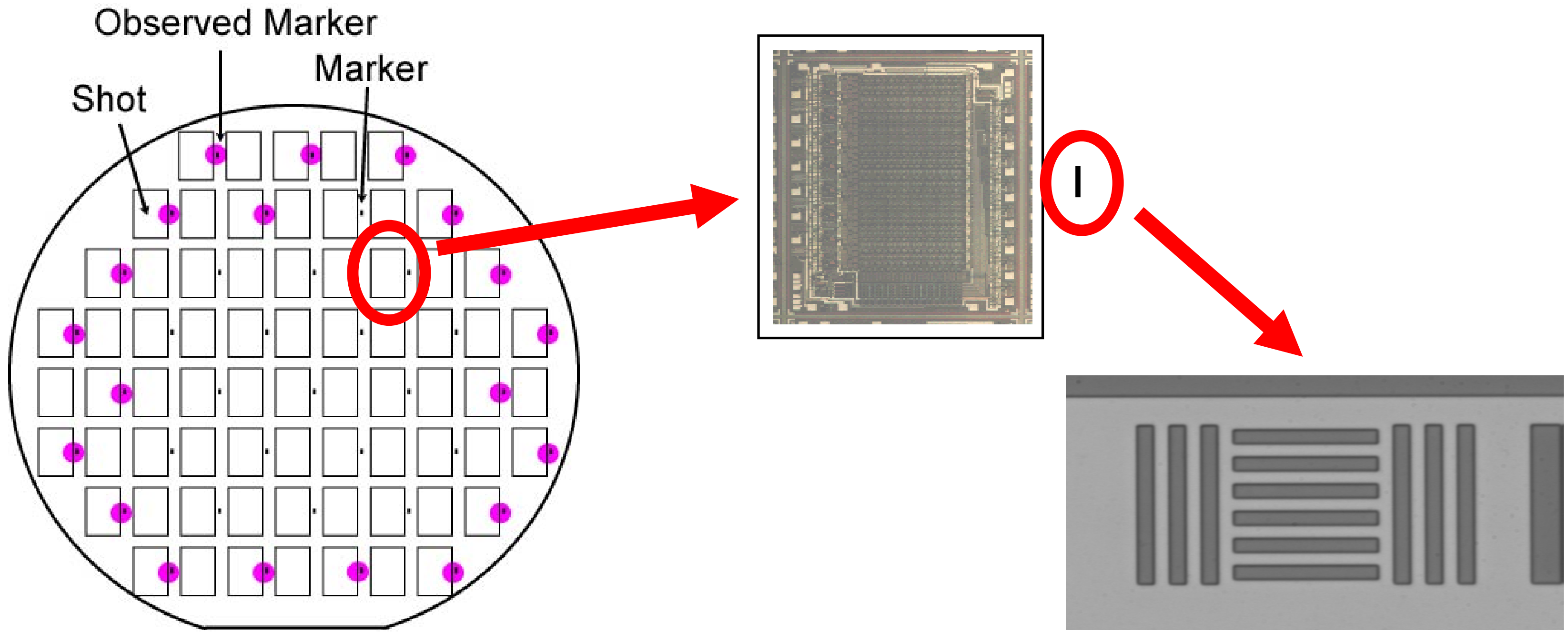
- Recent silicon wafers have layer structure.
- Circuit patterns are exposed multiple times.
- Exact alignment of wafers is very important.





Markers on Wafer

- Wafer alignment process:
  - Measure marker location printed on wafers.
  - Shift and rotate the wafer to minimize the gap.
- For speeding up, reducing the number of markers to measure is very important.

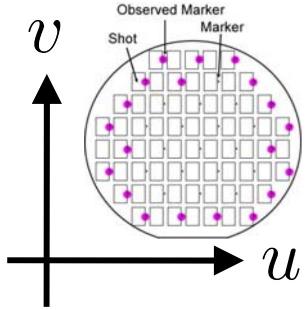


Non-linear Alignment Model

- When gap is only shift and rotation, linear model is exact:

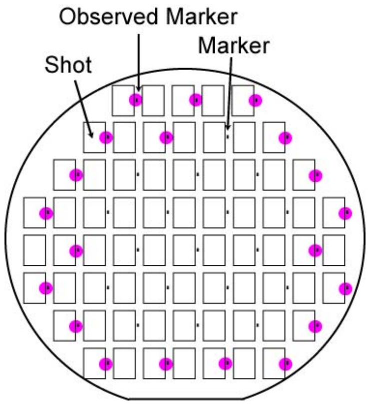
$$\Delta u \text{ or } \Delta v = \theta_0 + \theta_1 u + \theta_2 v$$

- However, non-linear factors exist, e.g.,
  - Warp
  - Biased characteristic of measurement apparatus
  - Different temperature conditions
- Exactly modeling non-linear factors is very difficult in practice!



Experimental Results

- 20 markers (out of 38) are chosen by experiment design methods.
- Gaps of all markers are predicted.
- Repeated for 220 different wafers.
- Mean (standard deviation) of the gap prediction error
- Red: Significantly better by 5% Wilcoxon test
- Blue: Worse than the baseline passive method



Model	Pool / Agnostic (Proposed)	Pool / Non-agnostic (Fedorov 1972)	"Outer" heuristic	Passive (Random)
Order 1	2.27(1.08)	2.37(1.15)	2.36(1.15)	2.32(1.11)
Order 2	1.93(0.89)	1.96(0.91)	2.13(1.08)	2.32(1.15)

Order 1: $\Delta u \text{ or } \Delta v = \theta_0 + \theta_1 u + \theta_2 v$

Order 2: $\Delta u \text{ or } \Delta v = \theta_0 + \theta_1 u + \theta_2 v + \theta_3 uv + \theta_4 u^2 + \theta_5 v^2$

- Proposed method works the best!

Conclusions

- We proposed a pool-based agnostic experiment design method for linear regression.
- Proposed method is
  - Robust against model misspecification,
  - Easy to implement.
- Proposed method is promising in
  - Extensive benchmark simulations,
  - Real-world wafer alignment task.
- Come to our poster for technical details!