

Identificación de efectos heterogéneos usando algoritmos de detección de patrones anómalos

Alejandra Caro

Agenda

1. Motivación
2. Aproximación de algoritmos de Machine Learning para identificar la heterogeneidad de efectos de tratamiento.
3. Detección de patrones: Fast Generalized subset scanning
4. Aplicación a experimentos aleatorios
5. Aplicación de otras técnicas de ML para evaluación de impacto

Implicaciones de política publica, efectos heterogéneos no identificados

Aproximación de la econometría convencional

- Requiere conocimiento *a priori* de la fuente que causa los ETH. Hay subconjuntos no obvios que tienen efectos distintos.
- Dificultad en la identificación de la **distribución** de los efectos. No se puede observar a la vez el mismo individuo en el grupo de tratamiento y de control.
- Homogeneidad de los efectos generaliza. Pero no es realista ni interesante.

No identificar heterogeneidad en los efectos puede conducir a...

- Subestimar los efectos de un tratamiento puede conducir implementar políticas en subgrupos donde hay efectos negativos (o viceversa).

Identificar fuentes de efectos heterogéneos permite el diseño y evaluación de políticas efectivas

- Diseño de programas focalizados; más efectivos?
- Implicaciones en bienestar social. Si es posible identificar sub poblaciones para las que los programas son positivos, vale la pena enfocarse en este subconjunto?

La aproximación de Machine Learning

Uso de métodos de ML para tratar el problema de heterogeneidad

Imai and Ratkovik (2013)

- Aproximan los efectos heterogeneos como un problema de selección de atributos.
- Adaptan un clasificador SVM asignando restricciones . Restricciones LASSO sobre las variables de pre-tratamiento y las que causan la heterogenidad. Esto permite observar de forma separada varios tratamientos.
- Aplicación a los datos del experimento get-out-the-vote (GOTV) para las elecciones en New Haven. The study not only include multiple feature but multiple treatments

Grimmer et al. (2013)

- Usan diferentes métodos para estimar HTE: LASSO, Elastic Nets, Bayesian Generalized linear models, Suport Vector Machines y Kernel recursive least squares.
- Prueban que las técnicas de aprendizaje conjunto tienen un performance superior que emplear cada técnica por separado.
- Note: arity of the features.

La aproximación de Machine Learning

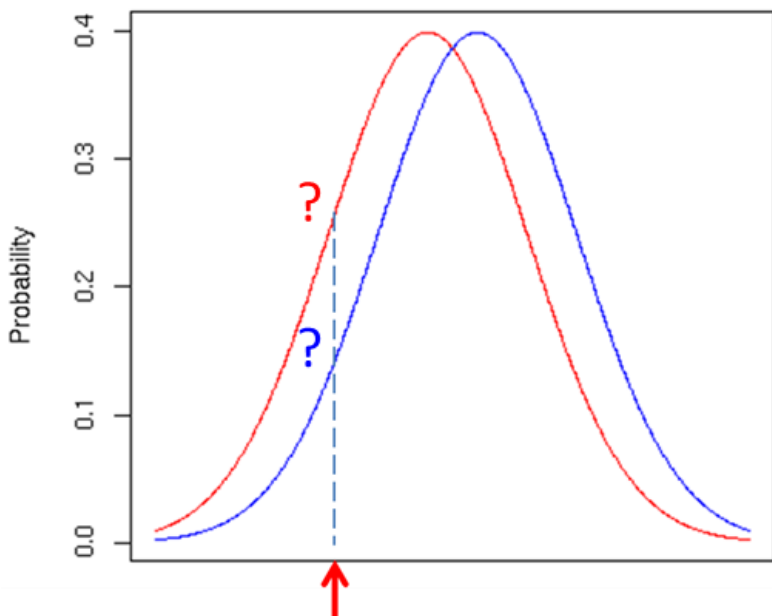
Uso de métodos de ML para tratar el problema de heterogeneidad

Esta presentación:

- ✓ Aproximar la situación como un problema de **detección de patrones** anómalos en grupos **auto-similares**.
 - ✓ **Intuición:** Grupos parecidos se deben responder de forma similar a un tratamiento!!!
- ✓ Comparar las distribuciones del grupo de control y tratamiento e identificar los sub conjuntos que observan un comportamiento disímil al modelo nulo.
- ✓ Emplear el algoritmo de escaneo generalizado en sub conjuntos para identificar los sub grupos de observaciones que vienen de otro proceso generador de datos y las variables que generan la anomalía.
- ✓ Aplicación a experimentos aleatorios. Se puede extender a otros quasi - experimentos, luego de corregir el sesgo de selección (-por ejemplo-).

Detección de patrones

Objetivo: Observar anomalías que se repiten de forma consistente en varios records.



Detectar eventos es un a tarea de
comparar hipotesis

$$H_1(D)$$

D = Subconjunto de datos, locaciones o
tiempo donde hay una anomalía

vs. H_0 : no hay eventos

- La mayoría de las técnicas de detección de anomalías busca observaciones atípicas por **separado**
- La dificultad de observar un solo punto es que no se sabe de qué distribución viene.
- Observaciones “ligeramente” atípicas son difíciles de detectar. Suelen minar la capacidad de detección del algoritmo si se analizan separadamente.
- Algoritmos de detección de patrones como un problema de búsqueda entre subconjuntos. La
- La búsqueda entre grupos mejora la capacidad de detección del sistema.
- Es una aproximación generalizada -> No aprende estructuras atípicas conocidas. Por ejemplo detección de fraude.

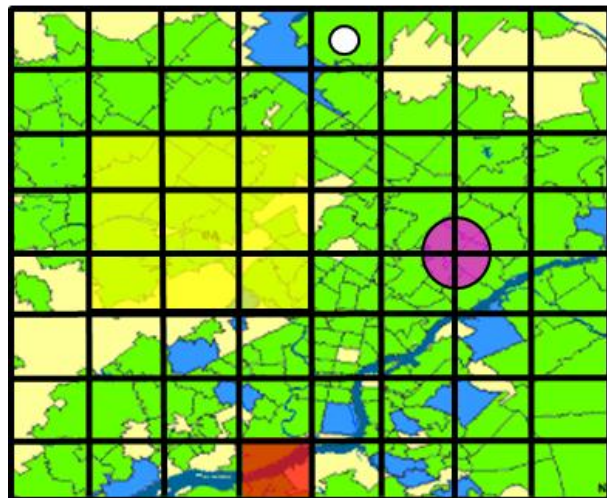
Detección de patrones

- Encontrar **subconjuntos** de datos que observan un comportamiento diferente al del resto de los datos.
- Identificar subconjuntos de records que **individualmente** son anómalos.

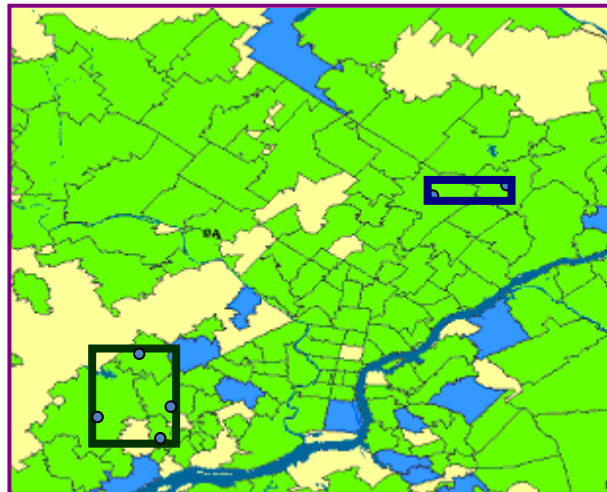
Esta aproximación...

Búsqueda de eventos en datos espaciales.

- (1) Búsqueda de Grilla. (*Parallel monitoring*) El poder del algoritmo depende de la forma de las particiones.
- (2) Escaneo en ventanas aleatorias que cambian de tamaño. *Spatial scan statistic*, Kulldorf, 1997.
- (3) *Fast generalized Scan Statistic*.



Tomado de Neill (2014)



FGSS

Escaneo en subconjuntos de datos espaciales.



Mejora poder de detección



Computacionalmente complejo.

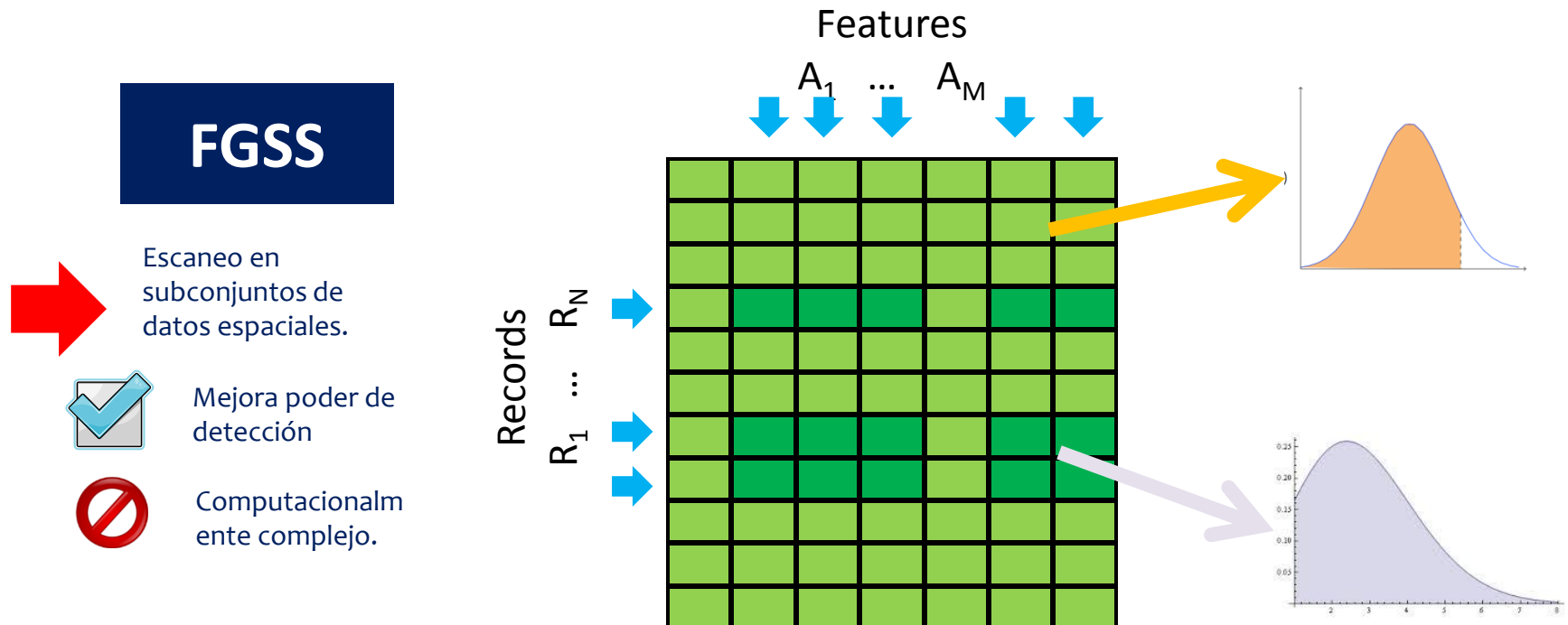
Detección de patrones

- Encontrar **subconjuntos** de datos que observan un comportamiento diferente al del resto de los datos.
- Identificar subconjuntos de records que **individualmente** son anómalos.

Esta aproximación...

Búsqueda de eventos en datos espaciales.

- (1) Búsqueda de Grilla. (*Parallel monitoring*) El poder del algoritmo depende de la forma de las particiones.
- (2) Escaneo en ventanas aleatorias que cambian de tamaño. *Spatial scan statistic, Kulldorf, 1997.*
- (3) *Fast generalized Scan Statistic.*



Detección de patrones

Fast Generalized Subset Scan (Neill, McFowland, Speakman; 2014)

- En lugar de tomar subregiones, FGSS aproxima la detección de patrones como un problema de **búsqueda entre subconjuntos**. Es posible aplicar restricciones (de proximidad, por ejemplo.)



Mejora poder de detección



Computacionalmente complejo.

En un conjunto de N observaciones y M atributos habría que buscar $2^N \times 2^M$ veces!
Computacionalmente Imposible!!!

Propiedad de Linear Time Subset Scanning (LTSS) -> No es necesario hacer búsqueda exhaustiva



Ahora si es computacionalmente posible!!!

Detección de patrones

Fast Generalized Subset Scan (Neill, McFowland, Speakman; 2014)

- En lugar de tomar subregiones, FGSS aproxima la detección de patrones como un problema de **búsqueda entre subconjuntos**. Es posible aplicar restricciones (de proximidad, por ejemplo.)



Mejora poder de detección



~~Computacionalmente complejo.~~

Propiedad de Linear Time Subset Scanning (LTSS) -> No es necesario hacer búsqueda exhaustiva

1. El tiempo de computación es lineal
2. Establecer un criterios prioridad para organizar los records.
3. Se realiza la búsqueda sobre los K más importantes records, la función de scoring garantiza que los datos escogidos son en efecto los que la maximizan.

Detección de patrones

Fast Generalized Subset Scan (Neill, McFowland, Speakman; 2014)

1. Estimar un modelo nulo M_0 que representa la el comportamiento “regular” de los datos. Este modelo se aprende del grupo de **control**.
2. Definir los subconjuntos de búsqueda. En el grupo de **tratamiento**.
3. Encontrar una función de *scoring*. Esto permite construir un ranking de observaciones anómalas, en el grupo de tratamiento respecto a lo aprendido en el modelo nulo.
4. Optimizar la función de *scoring* y encontrar los subconjuntos que la maximizan.

Detección de patrones

Objetivo: Encontrar el subconjunto de atributos y records que maximizan una función de scoring $F(S)$.

Sea $R_1 \dots R_N$ el conjunto de records y $A_1 \dots A_M$ de atributos. Para cada record R_i existe un valor v_{ij} para cada atributo A_j .

Se quiere encontrar el conjunto de datos anómalos que

$$S^* = R^* \times A^* = \operatorname{argmax}_S F(S).$$

Donde $F(S)$ es una función que define que tan anómalo es el conjunto.

Detección de patrones

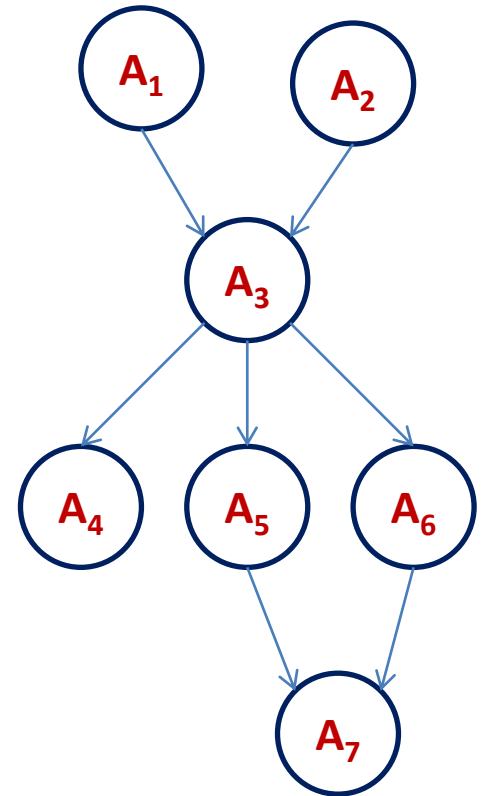
1. Aprender una distribución de los datos

Sea A_j un nodo en la red Bayesiana M_o y $A_{p(j)}$ es nodo padre de A_j . Los parámetros de la tabla de probabilidad condicional están dados por :

$$\theta_{jmk} = P_{M_o}(A_j = m | A_{p(j)} = k) \quad \forall j, m, k$$

Y computar la verosimilitud de cada observación en el grupo de tratamiento.

La mayoría de los algoritmos de detección de patrones y de detección de eventos usan redes bayesianas -> Suele ser una representación conveniente de la dependencia entre atributos.



- Redes bayesianas tienen múltiples representaciones.
- Mas atributos, más complejidad?
(Aproximación de Moore y Wong, 2003 -> "Optimal reinsertion search".).

Detección de patrones

2. Encontrar p-valores empíricos

De acuerdo a lo estimado en el modelo anterior se encuentra la función de distribución empírica del grupo de control. La distribución asintótica de los p-valores del modelo nulo es Uniforme $[0,1]$ (Probability integral transform)

$$p(l) = \frac{N_{beat} + N_{tie} + 1}{N_{train} + 1}$$

(Conveniente! Si un atributo no indica nada, entonces p-value=1)

Sea α un nivel de significancia y $n_a(p_{ij})$ un indicador binario de la significancia de p_{ij} , $N(S)$ es el número de elementos en el subconjunto S y $N_\alpha(S)$ el número de p-valores significativos.

*** p-valor está acotado!!!

N_{beat} y N_{tie} representan el conteo de *likelihoods* para cada record que tienen un valor menor o igual a l_{ij}

Detección de patrones

3. Estadístico no paramétrico

Neil et al. Definen la forma general de la función de scoring como:

$$F(S) = \operatorname{argmax} F_{\alpha}(S) = \max \phi(\alpha, N_{\alpha}(S), N(S))$$

Donde,

ϕ es convexo

ϕ es monotónicamente creciente en N_{α}

ϕ es monotónicamente decreciente en N and α

LTSS (-> computacionalmente posible), penaliza conjuntos muy grandes y niveles de significancia pequeños.

El estadístico no paramétrico de “Higher Criticism” (Donoho and Jin, 2004) que cumple estas tres condiciones se define:

$$\phi_{HC}(\alpha, N_{\alpha}, N) = \frac{N_{\alpha} - \alpha N}{\sqrt{N \alpha (1 - \alpha)}}$$

Detección de patrones

4. Búsqueda entre subconjuntos

Se maximiza recursivamente sobre los records y sobre los atributos hasta encontrar un score máximo para R dado A y para cada A dado R.

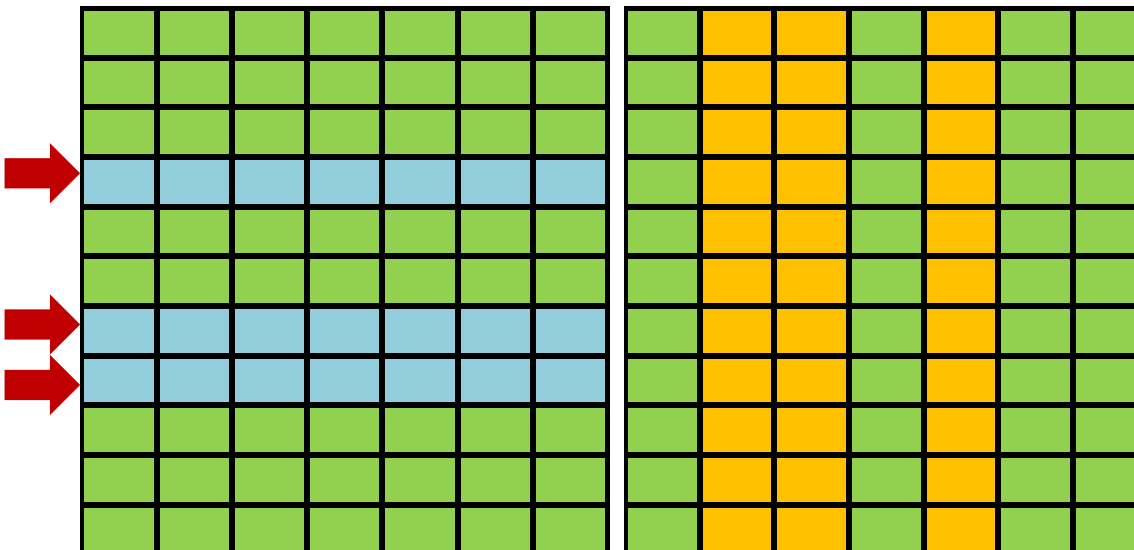
$$R = \operatorname{argmax}_{R \subseteq \{R_1, \dots, R_N\}} F(R \times A)$$

$$A = \operatorname{argmax}_{A \subseteq \{A_1, \dots, A_N\}} F(R \times A)$$

...

$$\operatorname{argmax}_{R \subseteq \{R_1, \dots, R_N\} A \subseteq \{A_1, \dots, A_N\}} F(R \times A)$$

Búsqueda no restringida
Máximo Global? ...



... 5. Restricciones de similitud

Incluir restricciones de similitud definiendo vecindarios locales para cada record y luego maximizar los subgrupos en cada vecindario.

... 6. Evaluación?

- Retos
- Test de Aleatorización es una opción

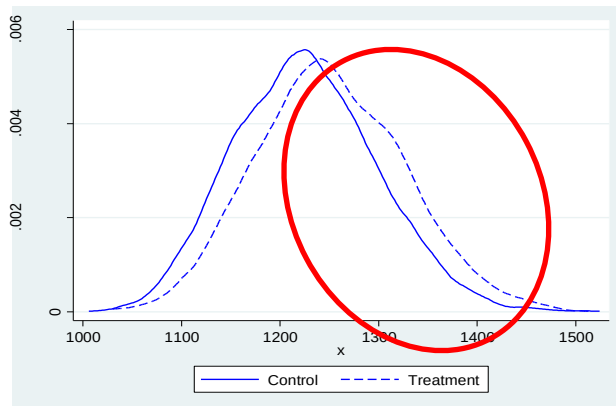
Aplicación: Detección de patrones

Aplicación a evaluación de Impacto

Uso datos del proyecto **STAR** (*Student Teacher Achievement Ratio*), un experimento a gran escala conducido en Tennessee durante 4 años en los años 80. El experimento pretendía observar el efecto de asignar menos estudiantes en cada salón de clase sobre el rendimiento académico.

- ¿Afecta la clase el rendimiento académico en lectura y matemáticas?
- ¿Existe alguna diferencia en los efectos sobre minorías?
- ¿Beneficia el programa a todos los estudiandtes de la misma manera?

Total treatment - control Kernel density

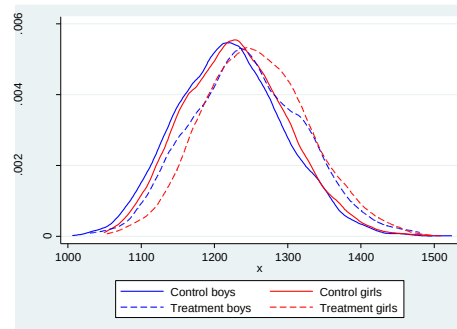


Aplicación: Detección de patrones

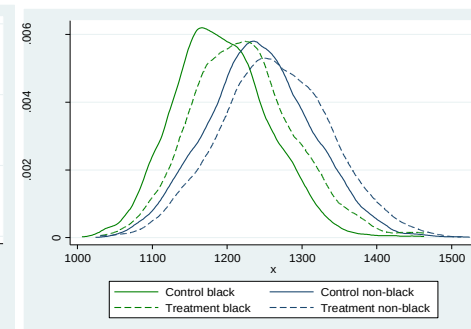
Aplicación a evaluación de Impacto

Uso datos del proyecto **STAR** (*Student Teacher Achievement Ratio*), un experimento a gran escala conducido en Tennessee durante 4 años en los años 80. El experimento pretendía observar el efecto de asignar menos estudiantes en cada salón de clase sobre el rendimiento académico.

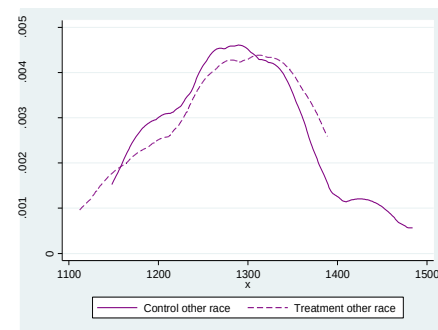
Kdristribution by gender



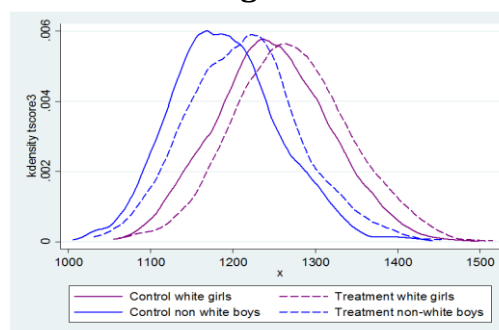
Kdristribution by race: black /non black



Effects - other race



Kdristribution: Effect on white girls



Aplicación: Detección de patrones

Al estimar varios modelos lineales controlando por raza (Blanco, negro u otra raza), experiencia del profesor y género .

Tabla 1: Modelos lineales para estimar efectos heterogéneos

VARIABLES	(1) tscore3	(2) tscore3	(3) tscore3	(4) tscore3	(5) tscore3
Treatment, T	24.59*** (1.940)	28.71*** (2.892)	29.37*** (3.424)	31.85*** (3.918)	17.50*** (3.639)
totexp3	0.156 (0.106)	-0.0844 (0.110)	0.164 (0.106)	0.154 (0.106)	-0.0109 (0.128)
boy	-10.66*** (1.811)	-8.431*** (2.306)		-8.879*** (2.199)	-10.66*** (1.811)
white	46.66*** (1.928)		48.62*** (2.324)	48.71*** (2.317)	46.70*** (1.927)
T + Boy		-5.418 (4.063)		-5.520 (3.875)	
T + white			-6.738 (4.159)	-6.599 (4.148)	
T + teaching					0.517** (0.225)
Constant	1,198*** (2.513)	1,231*** (2.296)	1,192*** (2.476)	1,196*** (2.722)	1,201*** (2.712)
Observations	5,918	5,918	5,918	5,918	5,918
R-squared	0.119	0.032	0.115	0.120	0.120

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

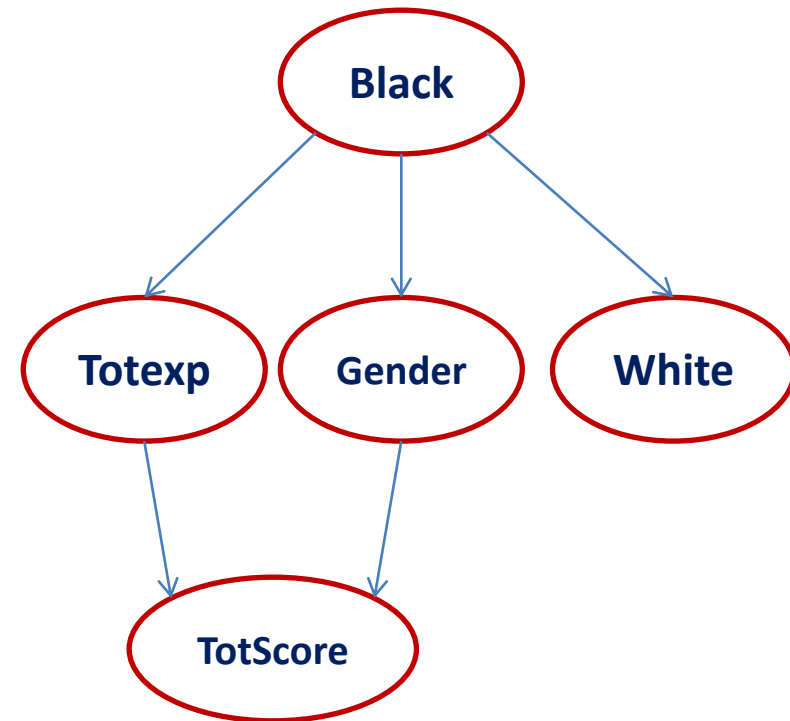
Aplicación: Detección de patrones

Resultados FGSS

Al aplicar FGSS arroja el conjunto de datos más anómalos. La tabla 2 muestra los datos en el subconjunto que maximiza la función de scoring.

Tabla 2: Subconjunto de datos “más anómalos”

TsScore	Totexp	White	Black	Boy
1304	16	1	0	0
1221	8	1	0	1
1280	14	0	1	0
1361	16	1	0	1
1303	5	1	0	1
1143	25	0	1	1
1280	17	1	0	0
1202	4	1	0	1
1356	2	1	0	1
1137	17	1	0	0
1241	18	0	1	0
1372	24	1	0	1
1084	18	0	1	0
1149	19	1	0	0
1257	5	1	0	1
1182	2	1	0	1



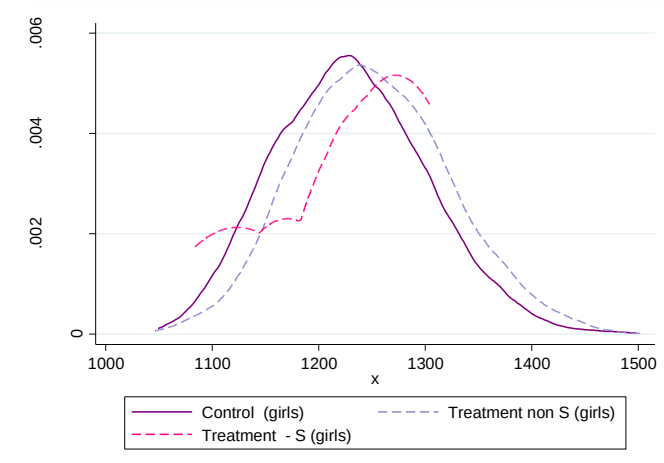
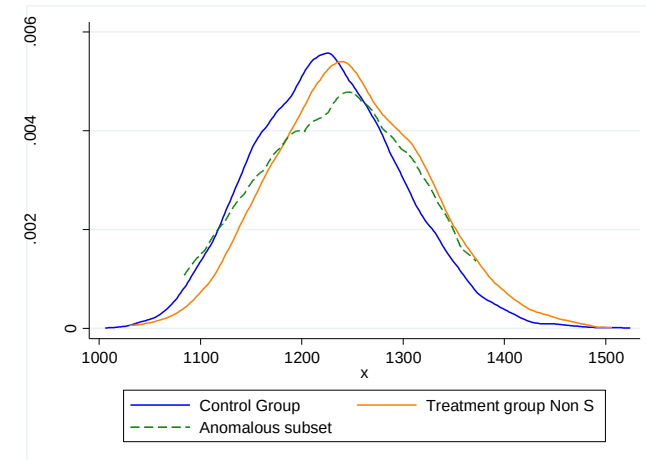
Aplicación: Detección de patrones

Resultados FGSS

Al aplicar FGSS arroja el conjunto de datos más anómalos. La tabla 2 muestra los datos en el subconjunto que maximiza la función de scoring.

Tabla 2: Subconjunto de datos “más anómalos”

TsScore	Totexp	White	Black	Boy
1304	16	1	0	0
1221	8	1	0	1
1280	14	0	1	0
1361	16	1	0	1
1303	5	1	0	1
1143	25	0	1	1
1280	17	1	0	0
1202	4	1	0	1
1356	2	1	0	1
1137	17	1	0	0
1241	18	0	1	0
1372	24	1	0	1
1084	18	0	1	0
1149	19	1	0	0
1257	5	1	0	1
1182	2	1	0	1



Aplicación: Detección de patrones

Conclusiones

1. Métodos de Machine Learning tiene un gran potencial para mejorar técnicas econométricas. En especial, para identificar los grupos que tienen efectos heterogéneos.
2. Algoritmos de detección de patrones anómalos pueden indicar la fuente de los efectos heterogeneos, sin embargo no indica nada sobre la distribución de los efectos.
3. **FGSS** es una estrategia muy útil en conjutnos con muchos atributos, donde la heterogeneidad puede estar causada por una combinacion de características.
4. **FGSS** permite identificar alteraciones **sutiles** en la distribución.

Retos: evaluación

Otras aplicaciones

1. Detección de agentes contaminantes en los ríos de PA.
2. Predicción de cambios en la salud financiera de los clientes de un banco local.



Otras proyectos en la intersección de ML y economía

1. Aplicación de métodos de ML para corrección de sesgo de selección.

PSM

