Bayesian Nonparametrics and Data Privacy

Alejandro Jara

MiDaS: Millenium Nucleus Center for the Discovery of Data Structures

Facultad de Matemáticas, UC

midas.mat.uc.cl

September 26th, 2018

- Wide dissemination of data facilitates advances in science and public policy, enables students to develop skills at data analysis, and helps ordinary citizens learn about their communities

- Wide dissemination of data facilitates advances in science and public policy, enables students to develop skills at data analysis, and helps ordinary citizens learn about their communities
- Often, however, agencies cannot release data as collected, because doing so could reveal data subjects' identities or values of sensitive attributes

- Wide dissemination of data facilitates advances in science and public policy, enables students to develop skills at data analysis, and helps ordinary citizens learn about their communities
- Often, however, agencies cannot release data as collected, because doing so could reveal data subjects' identities or values of sensitive attributes
- Failure to protect confidentiality can have serious consequences for agencies, since they may be violating laws or institutional rules enacted to protect confidentiality

- At first glance, sharing safe data with others seems a straightforward task: simply strip unique identifiers like names, tax identification numbers, and exact addresses before releasing data

- At first glance, sharing safe data with others seems a straightforward task: simply strip unique identifiers like names, tax identification numbers, and exact addresses before releasing data
- However, these actions alone may not suffice when quasi-identifiers, such as demographic variables, employment/education histories, or establishment sizes, remain on the file

Original Data

Name	Race	Birth Date	Sex	ZIP Code	Complaint
Sean	Black	9/20/1965	Male	02141	Short of breath
Daniel	Black	2/14/1965	Male	02141	Chest pain
Kate	Black	10/23/1965	Female	02138	Painful eye
Marion	Black	8/24/1965	Female	02138	Wheezing
Helen	Black	11/7/1964	Female	02138	Aching joints
Reese	Black	12/1/1964	Female	02138	Chest pain
Forest	White	10/23/1964	Male	02138	Short of breath
Hilary	White	3/15/1965	Female	02139	Hypertension
Philip	White	8/13/1964	Male	02139	Aching joints
Jamie	White	5/5/1964	Male	02139	Fever
Sean	White	2/13/1967	Male	02138	Vomiting
Adrien	White	3/21/1967	Male	02138	Back pain

Suppressed Data

Race	Complaint
Black	Short of breath
Black	Chest pain
Black	Painful eye
Black	Wheezing
Black	Aching joints
Black	Chest pain
White	Short of breath
White	Hypertension
White	Aching joints
White	Fever
White	Vomiting
White	Back pain

Generalized Data

Race	Birth Year	Sex	ZIP Code*	Complaint
Black	1965	Male	021*	Short of breath
Black	1965	Male	021*	Chest pain
Black	1965	Female	021*	Painful eye
Black	1965	Female	021*	Wheezing
Black	1964	Female	021*	Aching joints
Black	1964	Female	021*	Chest pain
White	1964	Male	021*	Short of breath
White	1965	Female	021*	Hypertension
White	1964	Male	021*	Aching joints
White	1964	Male	021*	Fever
White	1967	Male	021*	Vomiting
White	1967	Male	021*	Back pain

Aggregated Data

Men Short of Breath	2

- The Massachusetts Group Insurance Commission (MGIC) had a bright idea back in the mid-1990s

- The Massachusetts Group Insurance Commission (MGIC) had a bright idea back in the mid-1990s
- It decided to release "anonymized" data on state employees that showed every single hospital visit

- The Massachusetts Group Insurance Commission (MGIC) had a bright idea back in the mid-1990s
- It decided to release “anonymized” data on state employees that showed every single hospital visit
- The goal was to help researchers, and the state spent time removing all obvious identifiers such as name, address, and Social Security number (SSN)

- The Massachusetts Group Insurance Commission (MGIC) had a bright idea back in the mid-1990s
- It decided to release "anonymized" data on state employees that showed every single hospital visit
- The goal was to help researchers, and the state spent time removing all obvious identifiers such as name, address, and Social Security number (SSN)
- But a graduate student in data science (Latanya Sweeney) saw a chance to make a point about the limits of anonymization

- At the time MGIC released the data, William Weld, then Governor of Massachusetts, assured the public that MGIC had protected patient privacy by deleting identifiers

- At the time MGIC released the data, William Weld, then Governor of Massachusetts, assured the public that MGIC had protected patient privacy by deleting identifiers
- In response, Latanya Sweeney started hunting for the Governor's hospital records in the MGIC data

- At the time MGIC released the data, William Weld, then Governor of Massachusetts, assured the public that MGIC had protected patient privacy by deleting identifiers
- In response, Latanya Sweeney started hunting for the Governor's hospital records in the MGIC data
- She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes

- At the time MGIC released the data, William Weld, then Governor of Massachusetts, assured the public that MGIC had protected patient privacy by deleting identifiers
- In response, Latanya Sweeney started hunting for the Governor's hospital records in the MGIC data
- She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes
- For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter

- By combining this data with the MGIC records, Sweeney found Governor Weld with ease

- By combining this data with the MGIC records, Sweeney found Governor Weld with ease
- Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code

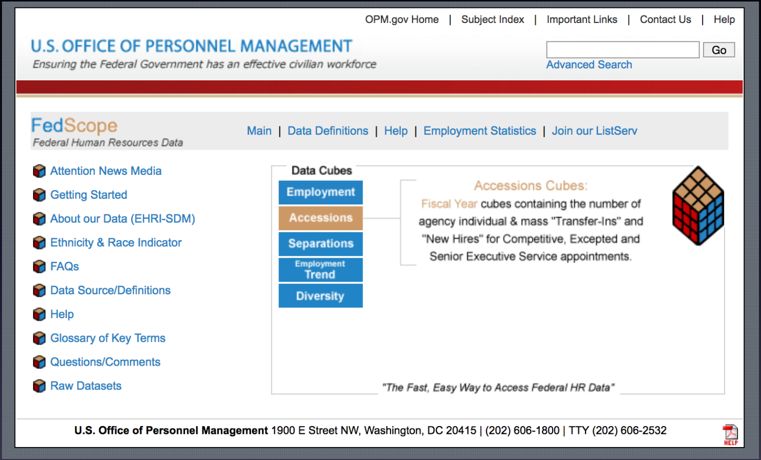
- By combining this data with the MGIC records, Sweeney found Governor Weld with ease
- Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code
- In a theatrical flourish, Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office

- By combining this data with the MGIC records, Sweeney found Governor Weld with ease
- Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code
- In a theatrical flourish, Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office
- In 2000, Sweeney showed that 87 percent of all Americans could be uniquely identified using only three bits of information: ZIP code, birthdate, and sex

- Sweeney also showed that 57 percent of American citizens are uniquely identified by their city, birth date, sex

- Sweeney also showed that 57 percent of American citizens are uniquely identified by their city, birth date, sex
- Finally, she showed that 18 percent of American citizens are uniquely identified by their county, birth date, sex

https://www.fedscope.opm.gov/, Diversity, 2018



Aggregated data: To protect the information (for example the race) they put an NA in the cells where there are less than 4 data points

Agency: IN06-INDIAN AFFAIRS

Employment as values		American Indian or Alaskan Native	Asian	Black/African American	Native Hawaiian or Pacific Islander	More Than One Race	Hispanic/Latino (H/L)	Minority
IN06-INDIAN AFFAIRS	01-ALABAMA	NA	NA	NA	NA	NA	NA	**NA**
	02-ALASKA	80	NA	NA	NA	6	NA	**90**
	04-ARIZONA	1,459	NA	7	12	4	36	**1,519**
	05-ARKANSAS	NA	NA	NA	NA	NA	NA	**NA**
	06-CALIFORNIA	160	NA	7	NA	6	26	**202**
	08-COLORADO	54	NA	NA	NA	NA	NA	**58**
	09-CONNECTICUT	NA	NA	NA	NA	NA	NA	**NA**
	10-DELAWARE	NA	NA	NA	NA	NA	NA	**NA**
	11-DISTRICT OF COLUMBIA	76	NA	NA	NA	4	NA	**85**
	12-FLORIDA	NA	NA	NA	NA	NA	NA	**NA**
	13-GEORGIA	NA	NA	NA	NA	NA	NA	**NA**
	15-HAWAII	NA	NA	NA	NA	NA	NA	**NA**
	16-IDAHO	66	NA	NA	NA	NA	NA	**67**
	17-ILLINOIS	NA	NA	NA	NA	NA	NA	**NA**
	18-INDIANA	NA	NA	NA	NA	NA	NA	**NA**
	19-IOWA	NA	NA	NA	NA	NA	NA	**NA**
	20-KANSAS	138	NA	NA	NA	NA	NA	**143**
	21-KENTUCKY	NA	NA	NA	NA	NA	NA	**NA**
	22-LOUISIANA	NA	NA	NA	NA	NA	NA	**NA**
	23-MAINE	NA	NA	NA	NA	NA	NA	**NA**
	24-MARYLAND	NA	NA	NA	NA	NA	NA	**NA**
	25-MASSACHUSETTS	NA	NA	NA	NA	NA	NA	**NA**
	26-MICHIGAN	9	NA	NA	NA	NA	NA	**9**
	27-MINNESOTA	59	NA	NA	NA	NA	NA	**63**
	28-MISSISSIPPI	4	NA	NA	NA	NA	NA	**4**
	29-MISSOURI	NA	NA	NA	NA	NA	NA	**NA**
	30-MONTANA	364	NA	NA	5	8	24	**401**
	31-NEBRASKA	19	NA	NA	NA	NA	NA	**20**
	32-NEVADA	33	NA	NA	NA	NA	NA	**35**
	33-NEW HAMPSHIRE	NA	NA	NA	NA	NA	NA	**NA**
	34-NEW JERSEY	NA	NA	NA	NA	NA	NA	**NA**
	35-NEW MEXICO	1,448	NA	4	8	6	52	**1,520**
	36-NEW YORK	NA	NA	NA	NA	NA	NA	**NA**
	37-NORTH CAROLINA	13	NA	NA	NA	NA	NA	**13**
	38-NORTH DAKOTA	386	NA	NA	NA	NA	NA	**391**

The first row is Alabama and all the cells are NA (in other words, there are less than 4 employees working for that agency in Alabama)

Employment as values		American Indian or Alaskan Native	Asian	Black/African American	Native Hawaiian or Pacific Islander	More Than One Race	Hispanic/Latino (H/L)	Minority
IN06-INDIAN AFFAIRS	01-ALABAMA	NA	NA	NA	NA	NA	NA	**NA**
	02-ALASKA	80	NA	NA	NA	6	NA	**90**
	04-ARIZONA	1,459	NA	7	12	4	36	**1,519**
	05-ARKANSAS	NA	NA	NA	NA	NA	NA	**NA**
	06-CALIFORNIA	160	NA	7	NA	6	26	**202**
	08-COLORADO	54	NA	NA	NA	NA	NA	**58**
	09-CONNECTICUT	NA	NA	NA	NA	NA	NA	**NA**
	10-DELAWARE	NA	NA	NA	NA	NA	NA	**NA**
	11-DISTRICT OF COLUMBIA	76	NA	NA	NA	4	NA	**85**
	12-FLORIDA	NA	NA	NA	NA	NA	NA	**NA**
	13-GEORGIA	NA	NA	NA	NA	NA	NA	**NA**
	15-HAWAII	NA	NA	NA	NA	NA	NA	**NA**
	16-IDAHO	66	NA	NA	NA	NA	NA	**67**
	17-ILLINOIS	NA	NA	NA	NA	NA	NA	**NA**
	18-INDIANA	NA	NA	NA	NA	NA	NA	**NA**
	19-IOWA	NA	NA	NA	NA	NA	NA	**NA**
	20-KANSAS	138	NA	NA	NA	NA	NA	**143**
	21-KENTUCKY	NA	NA	NA	NA	NA	NA	**NA**
	22-LOUISIANA	NA	NA	NA	NA	NA	NA	**NA**
	23-MAINE	NA	NA	NA	NA	NA	NA	**NA**
	24-MARYLAND	NA	NA	NA	NA	NA	NA	**NA**
	25-MASSACHUSETTS	NA	NA	NA	NA	NA	NA	**NA**
	26-MICHIGAN	9	NA	NA	NA	NA	NA	**9**
	27-MINNESOTA	59	NA	NA	NA	NA	NA	**63**
	28-MISSISSIPPI	4	NA	NA	NA	NA	NA	**4**
	29-MISSOURI	NA	NA	NA	NA	NA	NA	**NA**
	30-MONTANA	364	NA	NA	5	8	24	**401**
	31-NEBRASKA	19	NA	NA	NA	NA	NA	**20**
	32-NEVADA	33	NA	NA	NA	NA	NA	**35**
	33-NEW HAMPSHIRE	NA	NA	NA	NA	NA	NA	**NA**
	34-NEW JERSEY	NA	NA	NA	NA	NA	NA	**NA**
	35-NEW MEXICO	1,448	NA	4	8	6	52	**1,520**
	36-NEW YORK	NA	NA	NA	NA	NA	NA	**NA**
	37-NORTH CAROLINA	13	NA	NA	NA	NA	NA	**13**
	38-NORTH DAKOTA	386	NA	NA	NA	NA	NA	**391**

- From the row of Alaska, it is clear that 80 are American Indian, 6 have More than one race, and that there are 90 employees working for that agency, that is, there are 4 employees that you do not know whether to locate them in Asian, Black, Hawaiian, or Latino

- From the row of Alaska, it is clear that 80 are American Indian, 6 have More than one race, and that there are 90 employees working for that agency, that is, there are 4 employees that you do not know whether to locate them in Asian, Black, Hawaiian, or Latino

- What is silly here, is that for example if you look at Michigan you will see that everyone has the same race (American Indian)

- From the row of Alaska, it is clear that 80 are American Indian, 6 have More than one race, and that there are 90 employees working for that agency, that is, there are 4 employees that you do not know whether to locate them in Asian, Black, Hawaiian, or Latino
- What is silly here, is that for example if you look at Michigan you will see that everyone has the same race (American Indian)
- The same happens for North Carolina and Mississippi

- The example assume that a certain dataset is confidential and we want to protect privacy of the individuals

- The example assume that a certain dataset is confidential and we want to protect privacy of the individuals
- Let's assume that the ID completely identifies individuals and that we want to protect the privacy of their salaries (income)

- The example assume that a certain dataset is confidential and we want to protect privacy of the individuals
- Let's assume that the ID completely identifies individuals and that we want to protect the privacy of their salaries (income)
- Let's play with R now

- An ecological fallacy (or ecological inference fallacy) is a fallacy in the interpretation of statistical data where inferences about the nature of individuals are deduced from inference for the group to which those individuals belong

- An ecological fallacy (or ecological inference fallacy) is a fallacy in the interpretation of statistical data where inferences about the nature of individuals are deduced from inference for the group to which those individuals belong

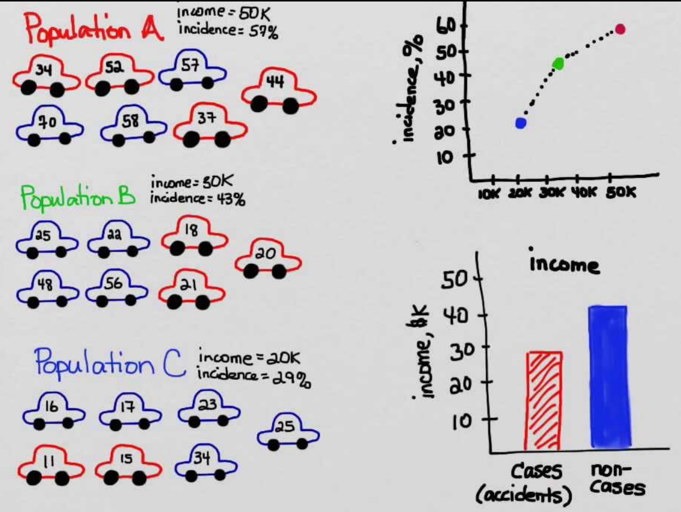
- Relationships that apply to a group level do not necessarily apply to an individual level

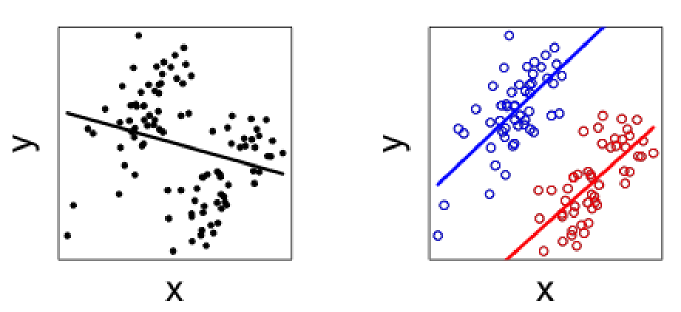
- Example of incidence of motor vehicle accident

- Example of incidence of motor vehicle accident
- Population A: Average income of 50K and incidence of 57%

- Example of incidence of motor vehicle accident
- Population A: Average income of 50K and incidence of 57%
- Population B: Average income of 30K and incidence of 43%

- Example of incidence of motor vehicle accident
- Population A: Average income of 50K and incidence of 57%
- Population B: Average income of 30K and incidence of 43%
- Population C: Average income of 20K and incidence of 29%


Population A
income = 50K
incidence = 57%
34 52 57 44
70 58 37
Population B
income = 30K
incidence = 43%
25 22 18 20
48 56 21
Population C
income = 20K
incidence = 29%
16 17 23 25
11 15 34
incidence, %
60 50 40 30 20 10
10K 20K 30K 40K 50K
income
income, $K
50 40 30 20 10
Cases (accidents)
non-cases


y
x
y
x

- The formal problem is that

$$\text{Cov}\left(\sum_{i=1}^{N} Y_i, \sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \text{Cov}(Y_i, X_i) + \sum_{i=1}^{N}\sum_{l \neq i} \text{Cov}(X_i, Y_l)$$

- William S. Robinson (1950) computed the correlation between the illiteracy rate and the proportion of the population born outside the US

- William S. Robinson (1950) computed the correlation between the illiteracy rate and the proportion of the population born outside the US
- He found a correlation of -0.53; in other words, the greater the proportion of immigrants in a state, the lower its average illiteracy

- William S. Robinson (1950) computed the correlation between the illiteracy rate and the proportion of the population born outside the US
- He found a correlation of -0.53; in other words, the greater the proportion of immigrants in a state, the lower its average illiteracy
- However, when individuals are considered, the correlation was $+0.12$ (immigrants were on average more illiterate than native citizens)

- William S. Robinson (1950) computed the correlation between the illiteracy rate and the proportion of the population born outside the US
- He found a correlation of -0.53; in other words, the greater the proportion of immigrants in a state, the lower its average illiteracy
- However, when individuals are considered, the correlation was $+0.12$ (immigrants were on average more illiterate than native citizens)
- Robinson showed that the negative correlation at the level of state populations was because immigrants tended to settle in states where the native population was more literate

- Suppose that you need to chose between two hospitals for an elderly relative's surgery

- Suppose that you need to chose between two hospitals for an elderly relative's surgery
- Out of each hospital's last 1000 patients, 900 survived in hospital A but only 800 survived in hospital B

- Suppose that you need to chose between two hospitals for an elderly relative's surgery
- Out of each hospital's last 1000 patients, 900 survived in hospital A but only 800 survived in hospital B
- It looks like hospital A is the better choice

- Suppose that you need to chose between two hospitals for an elderly relative's surgery

- Out of each hospital's last 1000 patients, 900 survived in hospital A but only 800 survived in hospital B

- It looks like hospital A is the better choice

- However, if we divide each hospital's last 1000 patients into those who arrive in good health and those who arrived in bad health, the picture starts to look very different

- Hospital A has only 100 patients who arrived in poor health, of which 30 survived

- Hospital A has only 100 patients who arrived in poor health, of which 30 survived
- Hospital B has 400 patients who arrived in poor health, of which 210 survived

- Hospital A has only 100 patients who arrived in poor health, of which 30 survived
- Hospital B has 400 patients who arrived in poor health, of which 210 survived
- So hospital B is the better choice for patients who arrived in poor health with a survival rate of 52.5%

- Formally, what we see is that for every value of z,

$$E(Y \mid Z = z, X = 1) > E(Y \mid Z = z, X = 0),$$

while

$$E(Y \mid X = 1) < E(Y \mid X = 0)$$

- What if your relative arrive in good health to the hospital?

- What if your relative arrive in good health to the hospital?
- Hospital B is still the better choice with a survival rate of 98.3% (590/600)

- What if your relative arrive in good health to the hospital?
- Hospital B is still the better choice with a survival rate of 98.3% (590/600)
- For hospital A the survival rate in this case is 96.7% (870/900)

- Let $\boldsymbol{Z} = (Z_1, Z_2, Z_3)^T \sim N_3\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (0,0,0)^T\,,$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.00 & 0.64 & 0.80 \\ 0.64 & 1.00 & 0.80 \\ 0.80 & 0.80 & 1.00 \end{pmatrix}$$

- Let $\boldsymbol{Z} = (Z_1, Z_2, Z_3)^T \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (0, 0, 0)^T,$$

  and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.00 & 0.64 & 0.80 \\ 0.64 & 1.00 & 0.80 \\ 0.80 & 0.80 & 1.00 \end{pmatrix}$$

- It is clear that $Z_1 \perp Z_2 \mid Z_3$, because

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 1.00 & 0.00 & 0.62 \\ 0.00 & 1.00 & 0.62 \\ 0.62 & 0.62 & 1.00 \end{pmatrix}$$

- Suppose that to anonymise the data we instead report

$$Y_i = \begin{cases} 1 & \text{if } V_i \geq 0, \\ 0 & \text{if } V_i < 0. \end{cases}$$

- Suppose that to anonymise the data we instead report

$$Y_i = \begin{cases} 1 & \text{if } V_i \geq 0, \\ 0 & \text{if } V_i < 0. \end{cases}$$

- Then,

$$Pr(Y_1 = 1, Y_2 = 1 \mid Y_3 = 1) = 0.6557,$$

while

$$Pr(Y_1 = 1 \mid Y_3 = 1) = Pr(Y_2 = 1 \mid Y_3 = 1) = 0.7952,$$

and

$$Pr(Y_1 = 1 \mid Y_3 = 1) \times Pr(Y_2 = 1 \mid Y_3 = 1) = 0.6323.$$

- Adding some randomly selected amount to the observed values, for example a random draw from a normal distribution with mean equal to zero

- Adding some randomly selected amount to the observed values, for example a random draw from a normal distribution with mean equal to zero

- This can reduce the possibilities of accurate matching on the perturbed data and distort the values of sensitive variables

- Adding some randomly selected amount to the observed values, for example a random draw from a normal distribution with mean equal to zero
- This can reduce the possibilities of accurate matching on the perturbed data and distort the values of sensitive variables
- The degree of confidentiality protection depends on the nature of the noise distribution; for example, using a large variance provides greater protection

- Adding some randomly selected amount to the observed values, for example a random draw from a normal distribution with mean equal to zero
- This can reduce the possibilities of accurate matching on the perturbed data and distort the values of sensitive variables
- The degree of confidentiality protection depends on the nature of the noise distribution; for example, using a large variance provides greater protection
- However, adding noise with large variance introduces measurement error that stretches marginal distributions and attenuates regression coefficients

- The basic idea of synthetic data is to replace original data values at high risk of disclosure with values simulated from probability distributions

- The basic idea of synthetic data is to replace original data values at high risk of disclosure with values simulated from probability distributions

- These distributions are specified to reproduce as many of the relationships in the original data as possible

- The basic idea of synthetic data is to replace original data values at high risk of disclosure with values simulated from probability distributions

- These distributions are specified to reproduce as many of the relationships in the original data as possible

- Synthetic data approaches come in two flavors: partial and full synthesis

- Partially synthetic data comprise the units originally surveyed with some subset of collected values replaced with simulated values

- Partially synthetic data comprise the units originally surveyed with some subset of collected values replaced with simulated values

- For example, the agency might simulate sensitive or identifying variables for units in the sample with rare combinations of demographic characteristics; or, the agency might replace all data for selected sensitive variables

- Partially synthetic data comprise the units originally surveyed with some subset of collected values replaced with simulated values
- For example, the agency might simulate sensitive or identifying variables for units in the sample with rare combinations of demographic characteristics; or, the agency might replace all data for selected sensitive variables
- Fully synthetic data comprise an entirely simulated data set; the originally sampled units are not on the file

- Partially synthetic data comprise the units originally surveyed with some subset of collected values replaced with simulated values
- For example, the agency might simulate sensitive or identifying variables for units in the sample with rare combinations of demographic characteristics; or, the agency might replace all data for selected sensitive variables
- Fully synthetic data comprise an entirely simulated data set; the originally sampled units are not on the file
- In both types, the agency generates and releases multiple versions of the data (as in multiple imputation for missing data)

- Synthetic data can provide valid inferences for analyses that are in accord with the synthesis models

- Synthetic data can provide valid inferences for analyses that are in accord with the synthesis models
- But they may not give good results for other analyses

- Synthetic data can provide valid inferences for analyses that are in accord with the synthesis models
- But they may not give good results for other analyses
- This is where Bayesian nonparametric models can play a big role

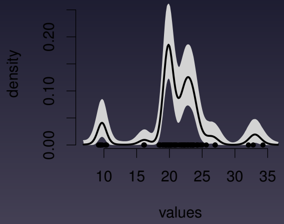
- Data sometimes is super weird:

- Data sometimes is super weird:
  - Internet transaction data distributions have a big spike at zero and spikes at other discrete values (e.g., 1 or $99)

- Data sometimes is super weird:
  - Internet transaction data distributions have a big spike at zero and spikes at other discrete values (e.g., 1 or $99)
  - Big tails that matter (e.g., $12 mil/month eBay user spend)

- Data sometimes is super weird:
  - Internet transaction data distributions have a big spike at zero and spikes at other discrete values (e.g., 1 or $99)
  - Big tails that matter (e.g., $12 mil/month eBay user spend)
  - The potential feature space is unmanageably large

- Data sometimes is super weird:
  - Internet transaction data distributions have a big spike at zero and spikes at other discrete values (e.g., 1 or $99)
  - Big tails that matter (e.g., $12 mil/month eBay user spend)
  - The potential feature space is unmanageably large
- We cannot write down simple models to explain the data



- Data are envisioned as realizations of random objects $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$

- Data are envisioned as realizations of random objects $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$
- The assumption is that $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ is drawn from a probability distribution G

- Data are envisioned as realizations of random objects $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$
- The assumption is that $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ is drawn from a probability distribution G
- Statistical models arise when G, or equivalent the density g, is known to be a member from a family

$$\mathcal{M} = \{(\mathcal{Y}, \mathcal{B}, G_{\boldsymbol{\theta}}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\},$$

labeled by a set of parameters $\boldsymbol{\theta}$ from an index set $\boldsymbol{\Theta}$

- Models that are described through a vector of a finite number of, typically, real values are referred to as finite-dimensional *parametric models*

- Models that are described through a vector of a finite number of, typically, real values are referred to as finite-dimensional *parametric models*

- Finite-dimensional parametric models can be described by the family

$$\mathcal{M} = \{(\mathcal{Y}^n, \mathcal{B}, G_{\boldsymbol{\theta}}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p\},$$

  where the dimension $p > 0$ is a finite integer

- Advantages:

- Advantages:
  - Convenience: parametric models are generally easier to work with

- Advantages:
  - Convenience: parametric models are generally easier to work with

- Advantages:
  - Convenience: parametric models are generally easier to work with
  - Efficiency: If a parametric model is correct, then parametric methods are more efficient than nonparametric methods (However, the loss in efficiency is often small)

- Advantages:
  - Convenience: parametric models are generally easier to work with
  - Efficiency: If a parametric model is correct, then parametric methods are more efficient than nonparametric methods (However, the loss in efficiency is often small)
  - Interpretation: Sometimes parametric models are easier to interpret

- Advantages:
  - Convenience: parametric models are generally easier to work with
  - Efficiency: If a parametric model is correct, then parametric methods are more efficient than nonparametric methods (However, the loss in efficiency is often small)
  - Interpretation: Sometimes parametric models are easier to interpret
- Disadvantages:

- Advantages:
  - Convenience: parametric models are generally easier to work with
  - Efficiency: If a parametric model is correct, then parametric methods are more efficient than nonparametric methods (However, the loss in efficiency is often small)
  - Interpretation: Sometimes parametric models are easier to interpret
- Disadvantages:
  - Sometimes it is hard to find a suitable parametric model

- Advantages:
  - Convenience: parametric models are generally easier to work with
  - Efficiency: If a parametric model is correct, then parametric methods are more efficient than nonparametric methods (However, the loss in efficiency is often small)
  - Interpretation: Sometimes parametric models are easier to interpret
- Disadvantages:
  - Sometimes it is hard to find a suitable parametric model
  - High risk of misspecification: assuming a wrong model

- Models that are described through infinite-dimensional parameters are referred to as *nonparametric models*

- Models that are described through infinite-dimensional parameters are referred to as *nonparametric models*

- Example (density estimation):

$$Y_1, \ldots, Y_n \mid G \overset{i.i.d.}{\sim} G,$$

where G is a probability distribution defined on $\mathbb{R}$

- Models that are described through infinite-dimensional parameters are referred to as *nonparametric models*

- Example (density estimation):

$$Y_1, \ldots, Y_n \mid G \stackrel{i.i.d.}{\sim} G,$$

where G is a probability distribution defined on $\mathbb{R}$

- In this case,

$$\boldsymbol{\theta} = G,$$

$$\boldsymbol{\Theta} \equiv \mathcal{P}(\mathbb{R}) = \{F : F \text{ is a probability distribution defined on } \mathbb{R}\}$$

- A Bayesian model is a unique probability measure on the product space "parameters $\times$ observations."

- A Bayesian model is a unique probability measure on the product space "parameters $\times$ observations."

- The sampling distribution $g_{\boldsymbol{\theta}}(\boldsymbol{y})$ is treated as a conditional distribution $g\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right)$.

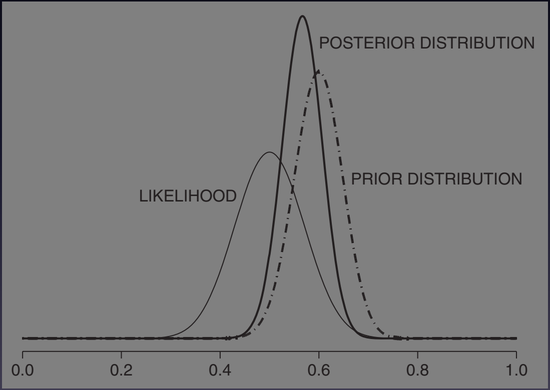
- A Bayesian model is a unique probability measure on the product space "parameters $\times$ observations."
- The sampling distribution $g_{\boldsymbol{\theta}}(\boldsymbol{y})$ is treated as a conditional distribution $g\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right)$.
- The parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is treated as random with distribution $\pi\left(\boldsymbol{\theta}\right)$ that is called the prior.

- Suppose that $Y \mid \theta \sim \text{Binomial}(n, \theta)$

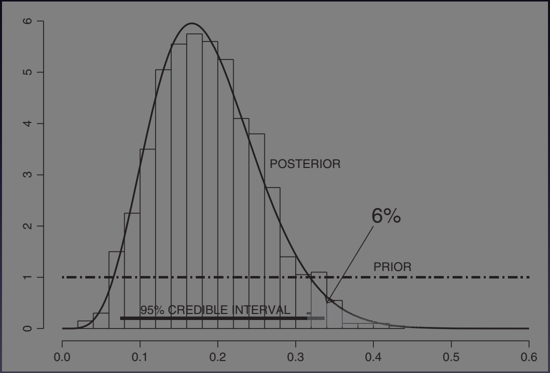
- Suppose that $Y \mid \theta \sim \text{Binomial}(n, \theta)$
- Suppose that $\theta \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$

- Suppose that $Y \mid \theta \sim$ Binomial(n, θ)
- Suppose that $\theta \mid \alpha, \beta \sim$ Beta(α, β)
- Then,
  - $\lambda_{n,\Pi}(y, \theta) = \frac{\binom{n}{y}}{B(\alpha,\beta)} \theta^{\alpha+y-1}(1-\theta)^{n-y+\beta-1}$, $\theta \in [0,1]$, $y = 0, 1, 2, \ldots, n$
  - $m(y) = \frac{\binom{n}{y} B(y+\alpha, n-y+\beta)}{B(\alpha,\beta)}$, $y = 0, 1, 2, \ldots, n$
  - $p(\theta \mid y) = \frac{1}{B(y+\alpha, n-x+\beta)} \theta^{\alpha+y-1}(1-\theta)^{n-y+\beta-1}$, $\theta \in [0,1]$
  - $m(y_0 \mid y) = \frac{\binom{n}{y} B(y+y_0+\alpha, 2n-y-y_0+\beta)}{B(y+\alpha, n-y+\beta)}$, $y_0 = 0, 1, 2, \ldots, n$

- Bayesian inference is based on the posterior distribution, which represents the updated knowledge about θ



- The interpretation of the inferences is not based on frequentist concepts



Theorem

(de Finetti, 1935) *The sequence of random objects $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots)$ is exchangeable if and only if there is a unique probability measure Π such that for all n the joint probability distribution of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ has a mixture model representation*

$$g(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) = \int \prod_{i=1}^{n} g_{\boldsymbol{\theta}}(\boldsymbol{Y}_i) d\Pi(\boldsymbol{\theta}),$$

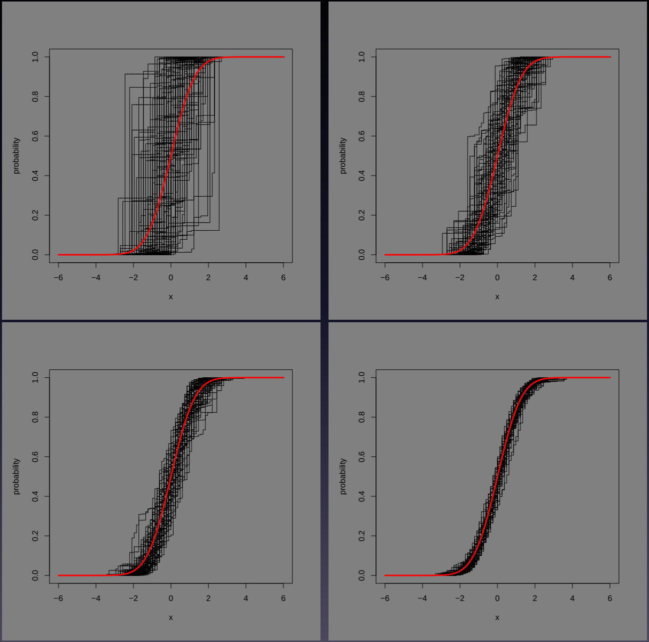
for some random variable $\boldsymbol{\theta}$

Theorem (Sethuraman, 1994)

Let $V_1, V_2, \ldots \overset{i.i.d.}{\sim}$ *Beta*$(1, \alpha)$ *and* $X_1, X_2, \ldots \overset{i.i.d.}{\sim} G_0$*. Then*

$$G(\cdot) = \sum_{i=1}^{\infty} W_i \delta_{X_i}(\cdot),$$

where, $W_1 = V_1$ *and, for* $i = 2, \ldots$*,* $W_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$*, is a Dirichlet process with parameters* (α, G_0)*.*


probability
x
1.0
0.8
0.6
0.4
0.2
0.0
−6 −4 −2 0 2 4 6

- Consider the following Polya urn model:

$$Y_1 \mid G_0 \sim G_0,$$

and, for $i = 2, 3, \ldots$,

$$Y_i \mid Y_1 \ldots, Y_{i-1}, \alpha, G_0 \sim G_n \equiv \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{Y_j} + \frac{\alpha}{\alpha + i - 1} G_0,$$

where δ_y is the Dirac measure on $(S, \mathcal{F})$ giving mass one to the point y

Theorem (Blackwell and MacQueen, 1973)

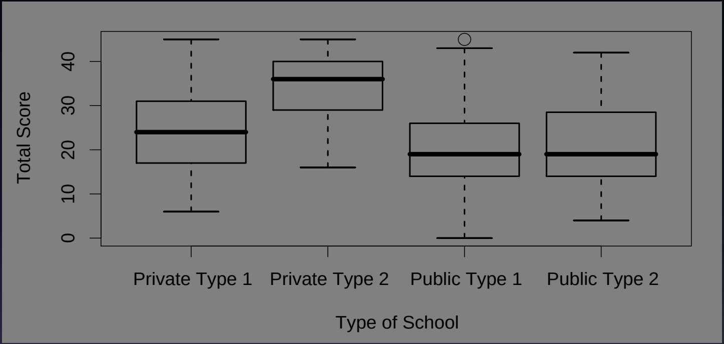
Let $Y_1 \sim G_0$ and for $i = 2, 3, \ldots, Y_i \mid Y_1 \ldots, Y_{i-1}, \alpha, G_0 \sim G_n$ as defined before. Then

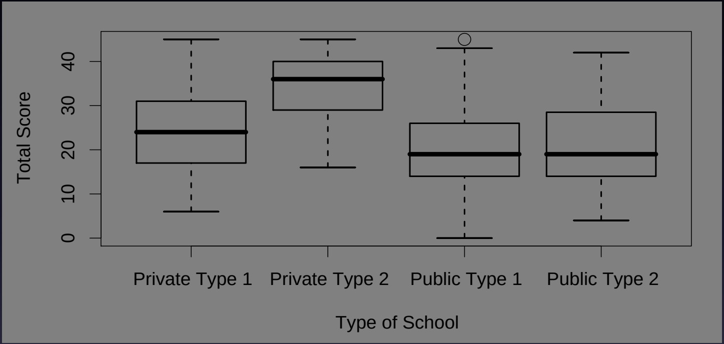
- *G_n converges almost surely to a random discrete distribution G, as $n \to \infty$*
- *G is a Dirichlet process (DP) with parameters (α, G_0)*
- *The sequence $Y_1, \ldots, Y_n$ is a sample from G*

- The SIMCE project in Chile has developed mandatory tests to assess regularly the educational progress in three levels: 4th and 8th grades in primary school, and 2th grade in secondary school

- The SIMCE project in Chile has developed mandatory tests to assess regularly the educational progress in three levels: 4th and 8th grades in primary school, and 2th grade in secondary school
- We will focus on data from the Math test, applied in 2006 to the second grade in secondary school (16 years old)

- The SIMCE project in Chile has developed mandatory tests to assess regularly the educational progress in three levels: 4th and 8th grades in primary school, and 2th grade in secondary school
- We will focus on data from the Math test, applied in 2006 to the second grade in secondary school (16 years old)
- The test consists of 45 multiple choice items with 4 alternatives, including a variety of questions ranging from problem formulation, functions,simple algebra, geometry and probability


Total Score
0
10
20
30
40
Private Type 1
Private Type 2
Public Type 1
Public Type 2
Type of School



- Models implying exchangeability of the response patterns are not suitable.

- Assume that for each of m subjects the responses to n items $\{Y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n\}$ are recorded.

- Assume that for each of m subjects the responses to n items $\{Y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n\}$ are recorded.

- Let $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in})'$ be the response pattern for subject i, where $Y_{ij} \in \{0, 1\}$.

In the Rasch model, the sampling model is given by

$$\begin{aligned} Y_{ij} \mid \lambda_{ij} &\overset{ind}{\sim} \text{Bernoulli}\,(\lambda_{ij}) \\ \lambda_{ij} &= \frac{\exp\{b_i - \beta_j\}}{1 + \exp\{b_i - \beta_j\}}, \end{aligned}$$

where b_i represents the *ability* of subject i and β_j represent the difficulty of the item j

- The ability parameters are considered as random effects whereas the difficulty parameters are interpreted as “fixed" effects

- The ability parameters are considered as random effects whereas the difficulty parameters are interpreted as “fixed" effects
- The classical specification of the model is completed by choosing a probability model for the abilities

- The ability parameters are considered as random effects whereas the difficulty parameters are interpreted as “fixed" effects

- The classical specification of the model is completed by choosing a probability model for the abilities

- The typical assumption is given by,

$$b_1, \ldots, b_m \mid G \overset{iid}{\sim} G,$$

where G is a probability distribution on $\mathbb{R}$

- We consider a dependent DP (DDP) mixture model for the distribution of the abilities b_i's,

$$g_{\boldsymbol{z}_i}(\cdot \mid \sigma^2, G_{\boldsymbol{z}_i}) = \int \frac{1}{\sigma}\phi\left(\frac{\cdot - \theta}{\sigma}\right) G_{\boldsymbol{z}_i}(d\theta),$$

where $\{G_{\boldsymbol{z}} : \boldsymbol{z} \in \mathcal{Z}\} \sim DDP$

- We consider the type of school and the gender as covariates.

- We consider the type of school and the gender as covariates.
- The type of school is a factor considering the levels:

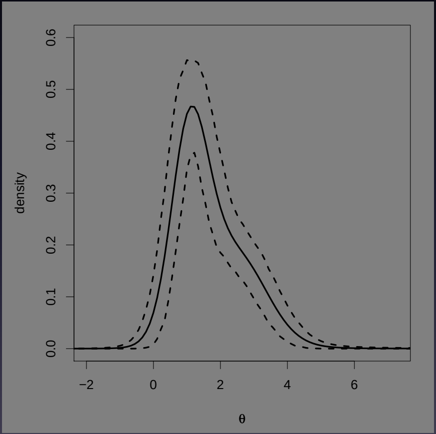
- We consider the type of school and the gender as covariates.
- The type of school is a factor considering the levels:
  - Financed by the state and administered by county governments (Public Type 1).

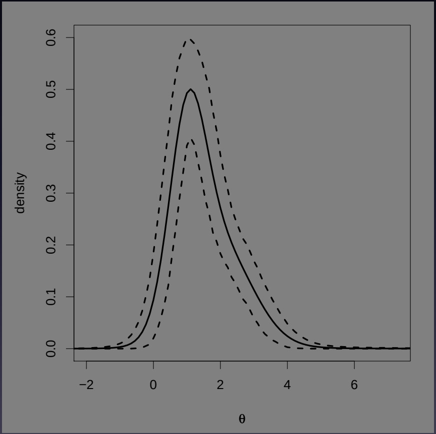
- We consider the type of school and the gender as covariates.
- The type of school is a factor considering the levels:
  - Financed by the state and administered by county governments (Public Type 1).
  - Financed by the state and administered by county corporations (Public Type 2).

- We consider the type of school and the gender as covariates.
- The type of school is a factor considering the levels:
  - Financed by the state and administered by county governments (Public Type 1).
  - Financed by the state and administered by county corporations (Public Type 2).
  - Financed by the state and administered by the private sector (Private Type 1).

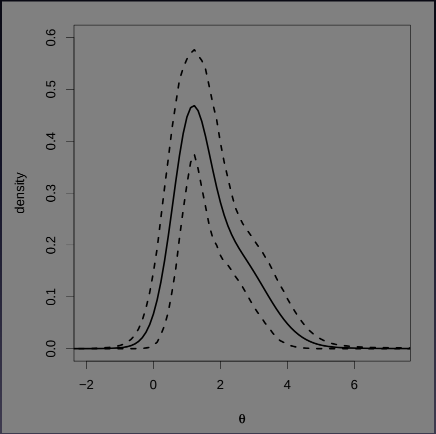
- We consider the type of school and the gender as covariates.
- The type of school is a factor considering the levels:
  - Financed by the state and administered by county governments (Public Type 1).
  - Financed by the state and administered by county corporations (Public Type 2).
  - Financed by the state and administered by the private sector (Private Type 1).
  - Fee-paying schools that operate solely on payments from parents and administered by the private sector (Private Type 2).

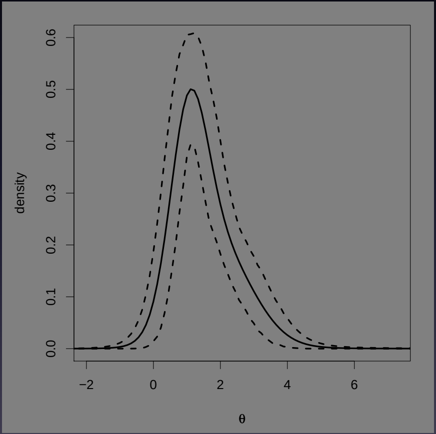
The Results - Public I



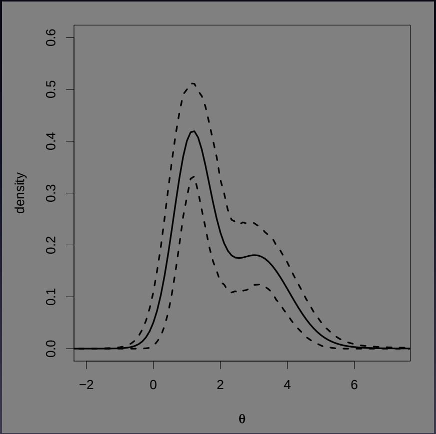


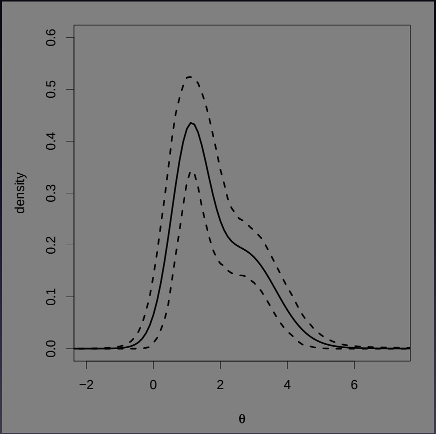
The Results - Public II



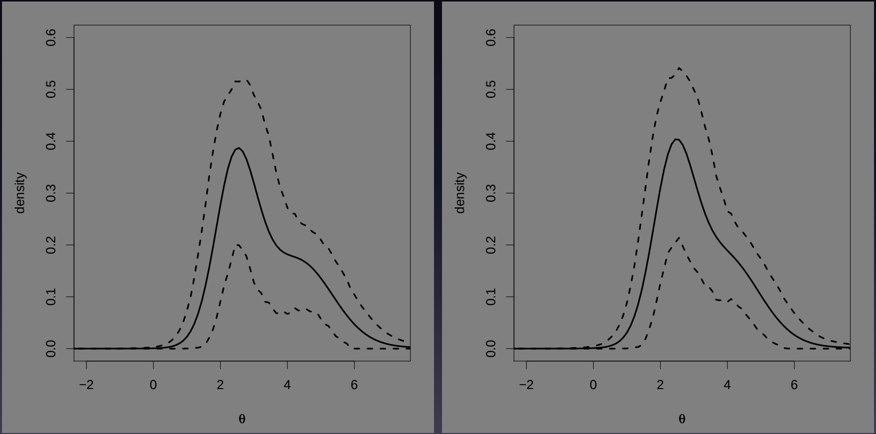


The Results - Private I





The Results - Private II

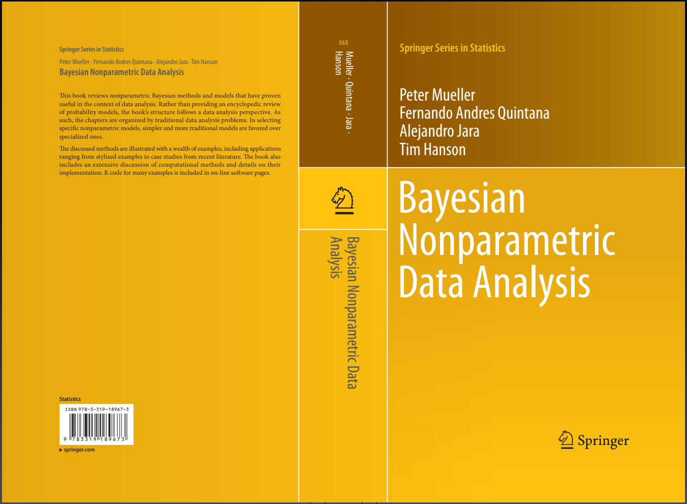


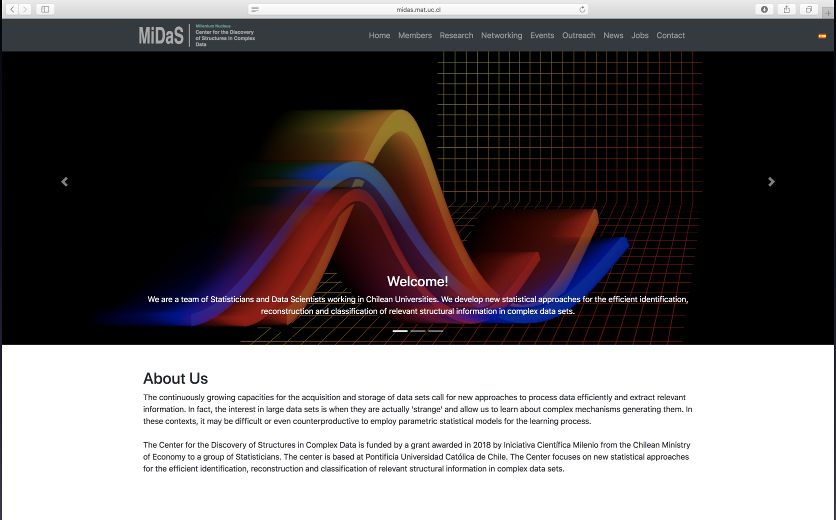
- Data can be either useful or perfectly anonymous but never both

- Data can be either useful or perfectly anonymous but never both
- One cannot eliminate the risk of disclosure, simply reduce it, unless one restricts access to the data

- Data can be either useful or perfectly anonymous but never both
- One cannot eliminate the risk of disclosure, simply reduce it, unless one restricts access to the data
- Thus techniques for disclosure limitation are inherently statistical in nature and must be evaluated using statistical tools for assessing the risk of harm to respondents

Thanks


Springer Series in Statistics
Peter Mueller · Fernando Andres Quintana · Alejandro Jara · Tim Hanson
Bayesian Nonparametric Data Analysis
This book reviews nonparametric Bayesian methods and models that have proven useful in the context of data analysis. Rather than providing an encyclopedic review of probability models, the book's structure follows a data analysis perspective. As such, the chapters are organized by traditional data analysis problems. In selecting specific nonparametric models, simpler and more traditional models are favored over specialized ones.
The discussed methods are illustrated with a wealth of examples, including applications ranging from stylized examples to case studies from recent literature. The book also includes an extensive discussion of computational methods and details on their implementation. R code for many examples is included in on-line software pages.
Statistics
ISBN 978-3-319-18967-3
springer.com
SSS
Mueller · Quintana · Jara · Hanson
Bayesian Nonparametric Data Analysis
Springer Series in Statistics
Peter Mueller
Fernando Andres Quintana
Alejandro Jara
Tim Hanson
Bayesian Nonparametric Data Analysis
Springer


midas.mat.uc.cl
MiDaS
Millenium Nucleus
Center for the Discovery of Structures in Complex Data
Home Members Research Networking Events Outreach News Jobs Contact
Welcome!
We are a team of Statisticians and Data Scientists working in Chilean Universities. We develop new statistical approaches for the efficient identification, reconstruction and classification of relevant structural information in complex data sets.
About Us
The continuously growing capacities for the acquisition and storage of data sets call for new approaches to process data efficiently and extract relevant information. In fact, the interest in large data sets is when they are actually 'strange' and allow us to learn about complex mechanisms generating them. In these contexts, it may be difficult or even counterproductive to employ parametric statistical models for the learning process.
The Center for the Discovery of Structures in Complex Data is funded by a grant awarded in 2018 by Iniciativa Científica Milenio from the Chilean Ministry of Economy to a group of Statisticians. The center is based at Pontificia Universidad Católica de Chile. The Center focuses on new statistical approaches for the efficient identification, reconstruction and classification of relevant structural information in complex data sets.

midas.mat.uc.cl

MiDaS Millenium Nucleus Center for the Discovery of Structures in Complex Data

Inicio Descripción Fechas de Interés Auspiciadores Patrocinadores Contacto

Esta es una competencia en la que equipos de alumnos de educación media de colegios chilenos resuelven problemas de análisis de datos. El objetivo de la competencia es estimular en estudiantes de educación secundaria el interés por la Estadística y la Ciencia de Datos.

Los Estadísticos y Científicos de Datos son expertos en la resolución de problemas, son capaces de analizar grandes conjuntos de datos, sacan conclusiones e informan sus hallazgos. Son buenos comunicadores orales y visuales, y pueden hacer llegar su mensaje a una amplia audiencia.

Durante la competencia, los estudiantes se 'ensuciarán las manos' realizando un análisis en profundidad de conjuntos de datos para encontrar la mejor recomendación para abordar un problema en particular.

La competencia tiene dos etapas. En la fase de preselección, los equipos deberán elaborar un informe escrito utilizando técnicas estadísticas básicas y MS Excel. Los equipos seleccionados en esta etapa, serán invitados a una semana de entrenamiento en la Facultad de Matemáticas de la Pontifica Universidad Católica de Chile. El entrenamiento incluirá técnicas modernas para la descripción y visualización de datos, y sobre el programa estadístico R. Posterior a al entrenamiento, se llevará a cabo la competencia final. Los costos de estadía y traslado de los equipos seleccionados de regiones distintas a la Metropolitana serán cubiertos por la competencia.