Insurance, biases, discrimination & fairness

(with some mathematical digressions)

Arthur Charpentier

2024

UNIWERSYTET
WARSZAWSKI

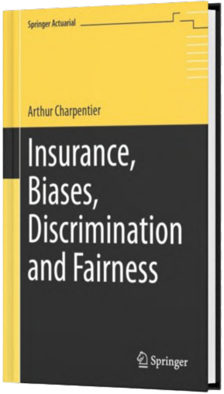
WYDZIAŁ NAUK
EKONOMICZNYCH

SCOR
FOUNDATION FOR SCIENCE

Reference book

Insurance, Biases, Discrimination and Fairness
ISBN : 978-3-031-49782-7

Pitch: Discrimination and fairness of predictive models, in insurance, in the context of data enrichment (”big data”) and opaque models (”machine learning”, not to say ”artificial intelligence”).



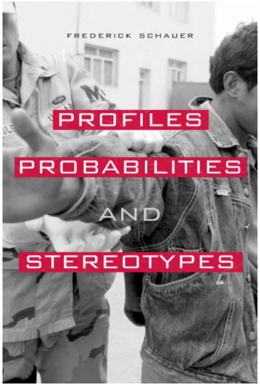
Warning: there are probably too many slides...

Preliminaries

Definition 1.1: Actuaries, Schauer (2006)

To be an actuary is to be a specialist in generalization, and actuaries engage in a form of decision making that is sometimes called actuarial. Actuaries guide insurance companies in making decisions about large categories that have the effect of attributing to the entire category certain characteristics that are probabilistically indicated by membership in the category, but that still may not be possessed by a particular member of the category.

See Barry and Charpentier (2020) on personalization of insurance prices.



Preliminaries

“– *Tu la troubles, reprit cette bête cruelle,*
Et je sais que de moi tu médis l’an passé.
– Comment l’aurais-je fait si je n’étais pas né ?
Reprit l’Agneau, je tette encor ma mère.
– Si ce n’est toi, c’est donc ton frère.
– Je n’en ai point.
– C’est donc quelqu’un des tiens.”

de La Fontaine (1668), *Le Loup et l’Agneau*.

Preliminaries

Definition 1.2: Discrimination, Merriam-Webster (2022)

Discrimination is the act, practice, or an instance of separating or distinguishing categorically rather than individually.

Discrimination is "*the act of treating different groups differently*," Frees and Huang (2021)

Definition 1.3: Prejudice, Merriam-Webster (2022)

Prejudice is (1) preconceived judgment or opinion, or an adverse opinion or leaning formed without just grounds or before sufficient knowledge; (2) an instance of such judgment or opinion; (3) an irrational attitude of hostility directed against an individual, a group, a race, or their supposed characteristics.

Preliminaries

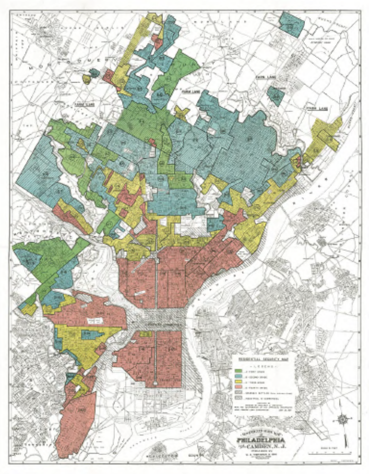
Definition 1.4: Disparate treatment, Merriam-Webster (2022)

Disparate treatment corresponds to the treatment of an individual (as an employee or prospective juror) that is less favorable than treatment of others for discriminatory reasons (as race, religion, national origin, sex, or disability).

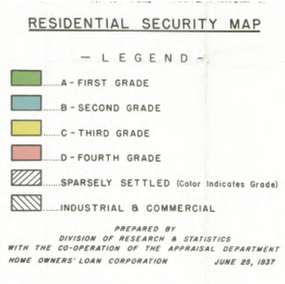
Definition 1.5: Disparate impact, Merriam-Webster (2022)

Disparate impact corresponds to an unnecessary discriminatory effect on a protected class caused by a practice or policy (as in employment or housing) that appears to be nondiscriminatory.

Motivation (1. Redlining)



1937 HOLC (Home Owners' Loan Corporation)
"residential security" map of Philadelphia



RESIDENTIAL SECURITY MAP

— LEGEND —

🟩 A - FIRST GRADE

🟦 B - SECOND GRADE

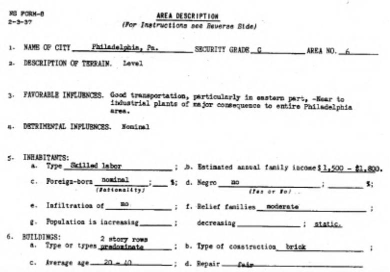
🟨 C - THIRD GRADE

🟥 D - FOURTH GRADE

▨ SPARSELY SETTLED (Color Indicates Grade)

▧ INDUSTRIAL & COMMERCIAL

PREPARED BY
DIVISION OF RESEARCH & STATISTICS
WITH THE CO-OPERATION OF THE APPRAISAL DEPARTMENT
HOME OWNERS' LOAN CORPORATION JUNE 25, 1937

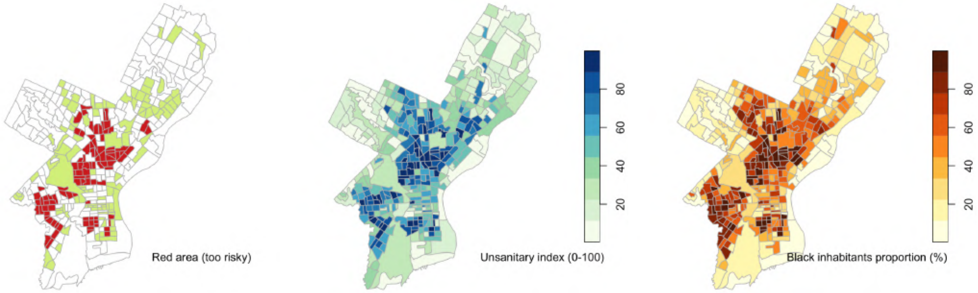


NS FORM-8
2-3-37

AREA DESCRIPTION
(For Instructions see Reverse Side)

1. NAME OF CITY Philadelphia, Pa. SECURITY GRADE C AREA NO. 6
2. DESCRIPTION OF TERRAIN. Level
3. FAVORABLE INFLUENCES. Good transportation, particularly in eastern part, -Near to industrial plants of major consequence to entire Philadelphia area.
4. DETRIMENTAL INFLUENCES. Nominal
5. INHABITANTS:
   - a. Type Skilled labor ; b. Estimated annual family income $1,500 - $1,800.
   - c. Foreign-born nominal (Nationality) ; ___ %; d. Negro no (Yes or No) ; ___ %;
   - e. Infiltration of no ; f. Relief families moderate ;
   - g. Population is increasing ___ ; decreasing ___ ; static.
6. BUILDINGS:
   - a. Type or types 2 story rows predominate ; b. Type of construction brick ;
   - c. Average age 20 - 40 ; d. Repair fair

Motivation (1. Redlining)



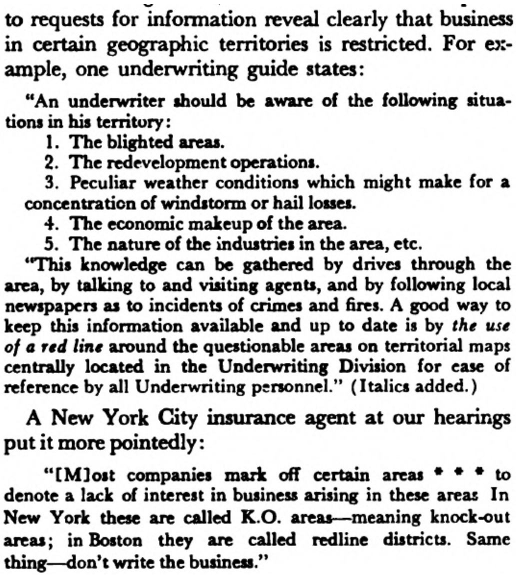
(Fictitious maps, inspired by a Home Owners' Loan Corporation map from 1937)

- Federal Home Loan Bank Board (FHLBB) "*residential security maps*" (for real-estate investments), Crossney (2016) and Rhynhart (2020)
- Unsanitary index and proportion of Black inhabitants

Motivation (1. Redlining)

Redlining was used for loans
but also insurers, Kerner (1968)

"*use of a red line around the questionable areas on territorial maps centrally located in the Underwriting Division for ease of reference by all Underwriting personnel* [...] *mark off certain areas * * * to denote a lack of interest in business arising in these areas In New York these are called K.O. areas meaning knock-out areas; in Boston they are called redline districts. Same thing – don't write the business.*"



to requests for information reveal clearly that business in certain geographic territories is restricted. For example, one underwriting guide states:

"An underwriter should be aware of the following situations in his territory:
1. The blighted areas.
2. The redevelopment operations.
3. Peculiar weather conditions which might make for a concentration of windstorm or hail losses.
4. The economic makeup of the area.
5. The nature of the industries in the area, etc.

"This knowledge can be gathered by drives through the area, by talking to and visiting agents, and by following local newspapers as to incidents of crimes and fires. A good way to keep this information available and up to date is by *the use of a red line* around the questionable areas on territorial maps centrally located in the Underwriting Division for ease of reference by all Underwriting personnel." (Italics added.)

A New York City insurance agent at our hearings put it more pointedly:

"[M]ost companies mark off certain areas * * * to denote a lack of interest in business arising in these areas In New York these are called K.O. areas—meaning knock-out areas; in Boston they are called redline districts. Same thing—don't write the business."

Motivation (1. Redlining)

Definition 1.6: Redline, Merriam-Webster (2022)

To redline is (1) to withhold home-loan funds or insurance from neighborhoods considered poor economic risks; (2) to discriminate against in housing or insurance.

See https://evolutionofraceandinsurance.org/ for some historical perspective, Squires and Velez (1988), or more recently Squires (2003)

... but still a concern see, e.g., Li (1996) about homosexuals.

Motivation (2. “Gender directive”, 2004/113/EC)

Treaty on European Union (26.10.2012, C326)

– Article 2 –

The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail.

– Article 3 –

(...) It shall combat social exclusion and discrimination, and shall promote social justice and protection, equality between women and men, solidarity between generations and protection of the rights of the child.

Motivation (2. “Gender directive”, 2004/113/EC)

Charter of Fundamental Rights of the European Union (18.12.2000 , C364)

– Article 21 (Non discrimination) –

Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

– Article 23 (Equality between men and women) –

Equality between men and women must be ensured in all areas, including employment, work and pay.
The principle of equality shall not prevent the maintenance or adoption of measures providing for specific advantages in favour of the under-represented sex.

Motivation (2. “Gender directive”, 2004/113/EC)

EU Directive (2004/113/EC), 2004 version

– Article 5 (Actuarial factors) –

1. Member States shall ensure that in all new contracts concluded after 21 December 2007 at the latest, the use of sex as a factor in the calculation of premiums and benefits for the purposes of insurance and related financial services shall not result in differences in individuals’ premiums and benefits.

2. Notwithstanding paragraph 1, Member States may decide before 21 December 2007 to permit proportionate differences in individuals’ premiums and benefits where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data. The Member States concerned shall inform the Commission and ensure that accurate data relevant to the use of sex as a determining actuarial factor are compiled, published and regularly updated.

Motivation (2. “Gender directive”, 2004/113/EC)

There was initially (2004) an opt-out clause (Article 5(2)).

Where gender is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data then proportionate differences in individual premiums or benefits are allowed.

March 2011, the European Court of Justice issued its judgement into the “Test-Achats case”. The ECJ ruled Article 5(2) was invalid.

Insurers were no longer able to use gender as a risk factor when pricing policies, "unisex pricing”.

"*Machine learning won’t give you anything like gender neutrality ‘for free’ that you didn’t explicitly ask for* ", Kearns and Roth (2019)

Motivation (2. “Gender directive”, 2004/113/EC)

“Ten Oever” judgement (*Gerardus Cornelis Ten Oever v Stichting Bedrijfspensioenfonds voor het Glazenwassers – en Schoonmaakbedrijf*, in April 1993), the Advocate General Van Gerven argued that “*the fact that women generally live longer than men has no significance at all for the life expectancy of a specific individual and it is not acceptable for an individual to be penalized on account of assumptions which are not certain to be true in his specific case*,” as mentioned in De Baere and Goessens (2011).

Schanze (2013) used the term “*injustice by generalization*,” from Britz (2008) (”Generalisierungsunrecht”)

Motivation (2. “Gender directive”, 2004/113/EC)

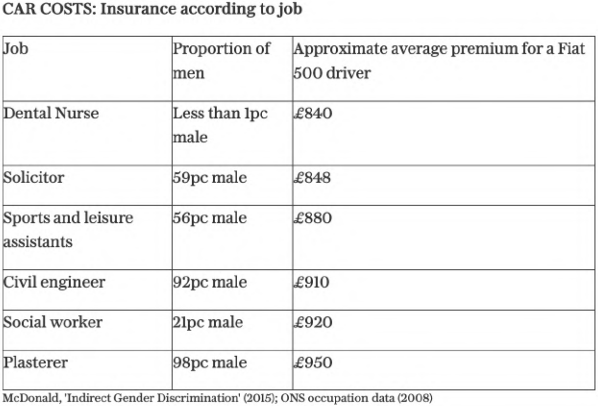
The Telegraph News Sport Money Business Opinion

Men are still charged more than women for car insurance, despite EU rule change

Car insurers are dodging European equality laws by making gender judgements based on people's jobs, an economist has found

By Kate Palmer
10 April 2015 • 12:33pm

Insurers will price by occupation, and female-dominated jobs tend to attract cheaper premiums | CREDIT: Photo: Rex Features



CAR COSTS: Insurance according to job

Job	Proportion of men	Approximate average premium for a Fiat 500 driver
Dental Nurse	Less than 1pc male	£840
Solicitor	59pc male	£848
Sports and leisure assistants	56pc male	£880
Civil engineer	92pc male	£910
Social worker	21pc male	£920
Plasterer	98pc male	£950

McDonald, 'Indirect Gender Discrimination' (2015); ONS occupation data (2008)

(data source: Mcdonald (2015))

Motivation (3. Québec)

Au Québec, Charte des droits et libertés de la personne (C-12)

– Article 20.1 –

In an insurance or pension contract, a social benefits plan, a retirement, pension or insurance plan, or a public pension or public insurance plan, a distinction, exclusion or preference based on age, sex or civil status is deemed non-discriminatory where the use thereof is warranted and the basis therefor is a risk determination factor based on actuarial data

Motivation (4. Colorado)

Andrus et al. (2021), "*What we can't measure, we can't understand*"

First Regular Session | 74th General Assembly
Colorado General Assembly

September 27, 2023, the Colorado Division of Insurance exposed a new proposed regulation entitled Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes

Motivation (4. Colorado)

– Section 4 (Definitions) –

Bayesian Improved First Name Surname Geocoding, *or "BIFSG" means, for the purposes of this regulation, the statistical methodology developed by the RAND corporation for estimating race and ethnicity.*

External Consumer Data and Information Source, *or "ECDIS" means, for the purposes of this regulation, a data source or an information source that is used by a life insurer to supplement or supplant traditional underwriting factors. This term includes credit scores, credit history, social media habits, purchasing habits, home ownership, educational attainment, licensures, civil judgments, court records, occupation that does not have a direct relationship to mortality, morbidity or longevity risk, consumer-generated Internet of Things data, biometric data, and any insurance risk scores derived by the insurer or third-party from the above listed or similar data and/or information source.*

Motivation (4. Colorado)

– Section 5 (Estimating Race and Ethnicity) –

Insurers shall estimate the race or ethnicity of all proposed insureds that have applied for coverage on or after the insurer's initial adoption of the use of ECDIS, or algorithms and predictive models that use ECDIS, including a third party acting on behalf of the insurer that used ECDIS, or algorithms and predictive models that used ECDIS, in the underwriting decision-making process, by utilizing:
1. BIFSG and the insureds' or proposed insureds' name and geolocation (information included in the applications) for life insurance shall be used to estimate the race and ethnicity of each insured or proposed insured.
2. For the purposes of BIFSG, the following racial and ethnic categories shall be used: Hispanic, Black, Asian Pacific Islander (API), and White.

Motivation (4. Colorado)

– Section 6 (Application Approval Decision Testing Requirements) –

> *Using the BIFSG estimated race and ethnicity of proposed insureds and the following methodology, insurers shall calculate whether Hispanic, Black, and API proposed insureds are disapproved at a statistically significant different rate relative to White applicants for whom the insurer, or a third party acting on behalf of the insurer, used ECDIS, or an algorithm or predictive model that used ECDIS, in the underwriting decision-making process.*
> *1. Logistic regression shall be used to model the binary underwriting outcome of either approved or denied.*
> *2. The following factors may be accounted for as control variables in the regression model: policy type, face amount, age, gender, and tobacco use.*
> *3. The estimated race or ethnicity of the proposed insureds shall be accounted for by including Hispanic, Black, and Asian Pacific Islander (API) as separate dummy variables in the regression model.*

Motivation (4. Colorado)

4. Determine if there is a statistically significant difference in approval rates for each BIFSG estimated race or ethnicity variable as indicated by a p-value of less than .05.

a. If there is not a statistically significant difference in approval rates, no further testing is required.

b. If there is a statistically significant difference in approval rates, the insurer shall determine whether the difference in approval rates is five (5) percentage points or greater as indicated by the marginal effects value of each BIFSG estimated race or ethnicity variable. (...)

Motivation (4. Colorado)

– Section 7 (Premium Rate Testing Requirements) –

Using the insureds' BIFSG estimated race and ethnicity, insurers shall determine if there is a statistically significant difference in the premium rate per $1,000 of face amount for policies issued to Hispanic, Black, and API insureds relative to White insureds for whom the insurer, or a third party acting on behalf of the insurer, used ECDIS, or an algorithm or predictive model that used ECDIS, in the underwriting decision-making process.
1. Linear regression shall be used to model the continuous numerical outcome of premium rate per $1,000 of face amount.
2. The following factors may be accounted for as control variables in the regression model: policy type, face amount, age, gender, and tobacco use.
3. The estimated race or ethnicity of the proposed insureds shall be accounted for by including Hispanic, Black, and Asian Pacific Islander (API) as separate dummy variables in the regression model.

Motivation (4. Colorado)

> *4. Determine if there is a statistically significant difference in the premium rate per \$1,000 of face amount for each BIFSG estimated race or ethnicity variable as indicated by a p-value of less than .05.*
>
> *a. If there is not a statistically significant difference in premium rate per \$1,000 of face amount, no further testing is required.*
>
> *b. If there is a statistically significant difference in premium rate per \$1,000 of face amount, determine whether the premium rate per \$1,000 of face amount is at least 5% more than the average premium rate per \$1,000 for all policies.*
>
> *i. If the difference in premium rate per \$1,000 of face amount is less than 5%, no further testing is required.*
>
> *ii. If the difference in premium rate per \$1,000 of face amount is 5% or greater, further testing is required as described in Section 8.*

Motivation (4. Colorado)

In Elliott et al. (2009), BIFSG[1], `library(eiCompare)`. , consider 12 people living near Atlanta, GA (Fulton & Gwinnett counties), and `eiCompare::wru_predict_race_wrapper`

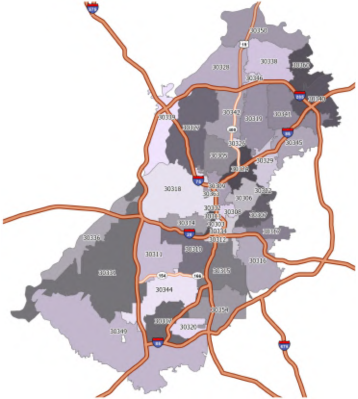
```
         last      first   county         city zipcode whi bla his asi
1     LOCKLER  GABRIELLA   Fulton      Atlanta   30318   0   0   0   0
2      RADLEY     OLIVIA   Fulton     Fairburn   30213  14  83   1   0
3      BOORSE     KEISHA   Fulton      Atlanta   30331  97   0   3   0
4         MAZ    SAVANNAH Gwinnett     Norcross   30093   5   6  76  13
5       GAULE   NATASHIA Gwinnett   Snellville   30078  67  19  14   0
6    MCMELLEN     ISMAEL Gwinnett      Lilburn   30047  73  15   6   3
7     RIDEOUT     LUQMAN Gwinnett   Snellville   30078  77  18   2   0
8  WASHINGTON       BRYN Gwinnett     Norcross   30093   0  95   3   0
9   KULENOVIC     EVELYN Gwinnett       Buford   30518 100   0   0   0
10  HERNANDEZ   SAMANTHA Gwinnett       Duluth   30096   3   1  94   1
11       LONG     BESSIE Gwinnett       Duluth   30096  53  39   1   1
12         HE       JOSE Gwinnett Lawrenceville   30045   2   3   4  89
```

---

[1] Bayesian Improved First Name Surname Geocoding

Motivation (3. Colorado)

We have 12 people,
in two counties near Atlanta
(about 10 zip-codes)



Motivation (3. Colorado)

Use `eiCompare::wru_predict_race_wrapper` on a revised dataset with the same name "Savannah Maz"

```
   last    first   county          city zipcode whi bla his asi
1   MAZ SAVANNAH   Fulton       Atlanta   30318   0   0   0 100
2   MAZ SAVANNAH   Fulton      Fairburn   30213  13  61  22   3
3   MAZ SAVANNAH   Fulton       Atlanta   30331   3  77  19   1
4   MAZ SAVANNAH Gwinnett      Norcross   30093   5   6  76  13
5   MAZ SAVANNAH Gwinnett    Snellville   30078  13  18  69   0
6   MAZ SAVANNAH Gwinnett       Lilburn   30047  28  22  34  16
7   MAZ SAVANNAH Gwinnett    Snellville   30078  53   3  40   3
8   MAZ SAVANNAH Gwinnett      Norcross   30093   5   6  76  13
9   MAZ SAVANNAH Gwinnett        Buford   30518  79   4  14   2
10  MAZ SAVANNAH Gwinnett        Duluth   30096  32   8  38  22
11  MAZ SAVANNAH Gwinnett        Duluth   30096  55  19  22   5
12  MAZ SAVANNAH Gwinnett Lawrenceville   30045  15  19  62   4
```

Motivation (3. Colorado)

Use `eiCompare::wru_predict_race_wrapper` on a revised dataset with the same name "Bryn Washington"

```
         last first   county          city zipcode whi bla his asi
1  WASHINGTON  BRYN   Fulton       Atlanta   30318   0   0   0 100
2  WASHINGTON  BRYN   Fulton      Fairburn   30213   0  99   0   0
3  WASHINGTON  BRYN   Fulton       Atlanta   30331   0  99   0   0
4  WASHINGTON  BRYN Gwinnett      Norcross   30093   0  95   3   0
5  WASHINGTON  BRYN Gwinnett    Snellville   30078   0  96   1   0
6  WASHINGTON  BRYN Gwinnett       Lilburn   30047   1  98   0   0
7  WASHINGTON  BRYN Gwinnett    Snellville   30078   6  87   2   0
8  WASHINGTON  BRYN Gwinnett      Norcross   30093   0  95   3   0
9  WASHINGTON  BRYN Gwinnett        Buford   30518   7  92   1   0
10 WASHINGTON  BRYN Gwinnett        Duluth   30096   2  96   1   0
11 WASHINGTON  BRYN Gwinnett        Duluth   30096   1  96   0   0
12 WASHINGTON  BRYN Gwinnett Lawrenceville   30045   0  98   1   0
```

Motivation (3. Colorado)

Use `eiCompare::wru_predict_race_wrapper` on a revised dataset with the same name "Samantha Hernandez"

```
        last    first   county          city zipcode whi bla his asi
1  HERNANDEZ SAMANTHA   Fulton       Atlanta   30318   0   0   0 100
2  HERNANDEZ SAMANTHA   Fulton      Fairburn   30213   2  12  85   0
3  HERNANDEZ SAMANTHA   Fulton       Atlanta   30331   0  16  81   0
4  HERNANDEZ SAMANTHA Gwinnett      Norcross   30093   0   0  99   0
5  HERNANDEZ SAMANTHA Gwinnett    Snellville   30078   1   1  97   0
6  HERNANDEZ SAMANTHA Gwinnett       Lilburn   30047   3   3  92   1
7  HERNANDEZ SAMANTHA Gwinnett    Snellville   30078   5   0  94   0
8  HERNANDEZ SAMANTHA Gwinnett      Norcross   30093   0   0  99   0
9  HERNANDEZ SAMANTHA Gwinnett        Buford   30518  17   1  81   0
10 HERNANDEZ SAMANTHA Gwinnett        Duluth   30096   3   1  94   1
11 HERNANDEZ SAMANTHA Gwinnett        Duluth   30096   8   4  86   0
12 HERNANDEZ SAMANTHA Gwinnett Lawrenceville   30045   1   2  97   0
```

Motivation (3. Colorado)

Use `eiCompare::wru_predict_race_wrapper` on a revised dataset with the same name "Jose He"

```
   last first   county           city zipcode whi bla his asi
1    HE  JOSE   Fulton        Atlanta   30318   0   0   0 100
2    HE  JOSE   Fulton       Fairburn   30213   2   9   2  84
3    HE  JOSE   Fulton        Atlanta   30331   1  27   3  55
4    HE  JOSE Gwinnett       Norcross   30093   0   0   2  98
5    HE  JOSE Gwinnett     Snellville   30078  13  18  30   0
6    HE  JOSE Gwinnett        Lilburn   30047   1   1   1  97
7    HE  JOSE Gwinnett     Snellville   30078   8   1   3  86
8    HE  JOSE Gwinnett       Norcross   30093   0   0   2  98
9    HE  JOSE Gwinnett         Buford   30518  19   1   2  78
10   HE  JOSE Gwinnett         Duluth   30096   1   0   0  98
11   HE  JOSE Gwinnett         Duluth   30096   6   2   1  85
12   HE  JOSE Gwinnett  Lawrenceville   30045   2   3   4  89
```

Motivation (5. Motor Insurance in the U.S.)

via The Zebra (2022),

California

Allowed (with applicable limitations): driving experience, marital status, address/zip code

Prohibited (or effectively prohibited): gender, age, credit history, education, occupation, employment status, residential status, insurance history

Notes & Clarifications: California's insurance commissioner banned gender as of January 2019. Occupation and education are permitted for use in group plans (i.e. for alumni associations and other membership programs).

Georgia

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): occupation, education, and employment status

Notes & Clarifications: none

Hawaii

Allowed (with applicable limitations): address/zip code, insurance history

Prohibited (or effectively prohibited): gender, age, years of driving experience, credit history, education, occupation, employment status, marital status, residential status

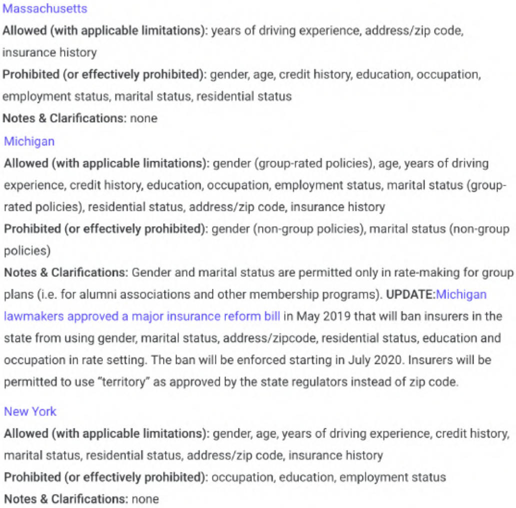
Notes & Clarifications: none

Illinois

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, education, occupation, employment status, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): none

Notes & Clarifications: none



Massachusetts

Allowed (with applicable limitations): years of driving experience, address/zip code, insurance history

Prohibited (or effectively prohibited): gender, age, credit history, education, occupation, employment status, marital status, residential status

Notes & Clarifications: none

Michigan

Allowed (with applicable limitations): gender (group-rated policies), age, years of driving experience, credit history, education, occupation, employment status, marital status (group-rated policies), residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): gender (non-group policies), marital status (non-group policies)

Notes & Clarifications: Gender and marital status are permitted only in rate-making for group plans (i.e. for alumni associations and other membership programs). **UPDATE:**Michigan lawmakers approved a major insurance reform bill in May 2019 that will ban insurers in the state from using gender, marital status, address/zipcode, residential status, education and occupation in rate setting. The ban will be enforced starting in July 2020. Insurers will be permitted to use "territory" as approved by the state regulators instead of zip code.

New York

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): occupation, education, employment status

Notes & Clarifications: none

see also ➔ Avraham et al. (2013)

Motivation (5. Motor Insurance in the U.S.)

	CA	HI	GA	NC	NY	MA	PA	FL	TX	AL	ON	NB	NL	QC
Gender	☐	☐	☑	☐	☑	☐	☐	☑	☑	☑	☑	☐	☐	☑
Age	☐	☐	☑	☐*	☑	☐	☑	☑	☑	☑*	☑	☐	☐	☑
Driving experience	☑	☐	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Credit history	☐	☐	☑	☑	☑	☐	☑*	☑	☑	☐*	☐	☑*	☐	☑
Education	☐	☐	☐	☐	☐	☐	☑	☑	☑	☑	☑	☑	☑	☑
Occupation	☐	☐	☐	☑	☐	☐	☑	☑	☑	☑	☑	☑	☑	☑
Employment status	☐	☐	☐	☑	☐	☐	☑	☑	☑	☑	☑	☑	☑	☑
Marital status	☑	☐	☑	☑	☑	☐	☑	☑	☑	☑	☑	☑	☑	☑
Housing situation	☐	☐	☑	☑	☑	☐	☑	☑	☑	☐	☐	☑	☑	☑
Address/ZIP code	☑	☑	☑	☑	☑	☑	☑	☑	☑	☐	☐	☑	☑	☑
Insurance history	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑

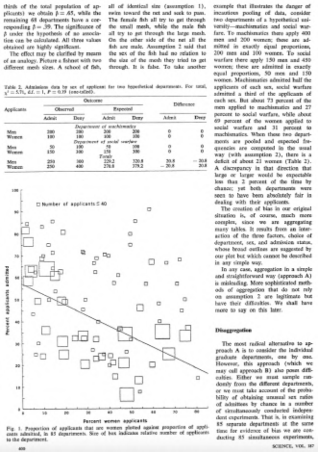
CA: California, HI: Hawaii, GA: Georgia, NC: North Carolina, NY: New York, MA: Massachusetts, PA: Pennsylvania, FL: Florida, TX: Texas, AL: Alberta, ON: Ontario, NB: New-Brunswick, NL: Newfoundland-Labrador, QC: Québec,

Motivation (6. Admission in Graduate Program, UC Berkeley)

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell



Bickel et al. (1975)

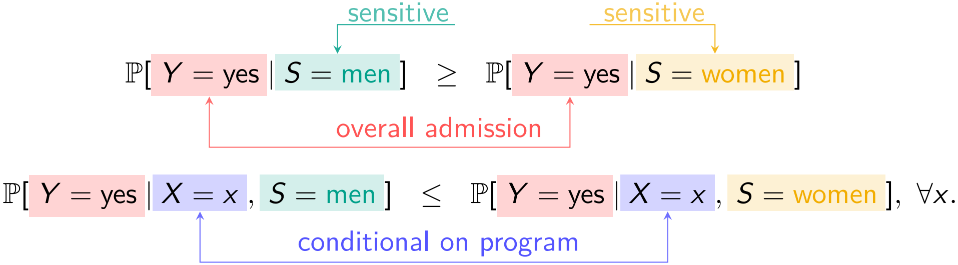
Motivation (6. Admission in Graduate Program, UC Berkeley)

	Total	Men	Women	Proportions
Total	5233/12763 ∼ 41%	3714/8442 ∼ **44%**	1512/4321 ∼ 35%	66%-34%
Top 6	1745/4526 ∼ 39%	1198/2691 ∼ **45%**	557/1835 ∼ 30%	59%-41%
A	597/933 ∼ 64%	512/825 ∼ 62%	89/108 ∼ **82%**	88%-12%
B	369/585 ∼ 63%	353/560 ∼ 63%	17/ 25 ∼ **68%**	96%- 4%
C	321/918 ∼ 35%	120/325 ∼ **37%**	202/593 ∼ 34%	35%-65%
D	269/792 ∼ 34%	138/417 ∼ 33%	131/375 ∼ **35%**	53%-47%
E	146/584 ∼ 25%	53/191 ∼ **28%**	94/393 ∼ 24%	33%-67%
F	43/714 ∼ 6%	22/373 ∼ 6%	24/341 ∼ **7%**	52%-48%

Data from Bickel et al. (1975) (discussed as an illustration of ”Simpson’s paradox”)

Formalize the later, S is the (binary) genre, Y the admission and X the program (category),

Motivation (6. Admission in Graduate Program, UC Berkeley)



sensitive

sensitive

$$\mathbb{P}[\,Y = \text{yes} \mid S = \text{men}\,] \;\geq\; \mathbb{P}[\,Y = \text{yes} \mid S = \text{women}\,]$$

overall admission

$$\mathbb{P}[\,Y = \text{yes} \mid X = x, S = \text{men}\,] \;\leq\; \mathbb{P}[\,Y = \text{yes} \mid X = x, S = \text{women}\,], \;\forall x.$$

conditional on program

“*the bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects*,” Bickel et al. (1975)

Motivation (6'. Admission in hospitals)

Consider the following mortality rates in two hospitals (fake data)

	Total	Healthy	Pre-condition	Proportions
Hospital A	$800/1000 = 80\%$	$590/600 \sim$ **98%**	$210/400 \sim$ **53%**	60%-40%
Hospital B	$900/1000 =$ **90%**	$870/900 \sim 97\%$	$30/100 \sim 30\%$	90%-10%

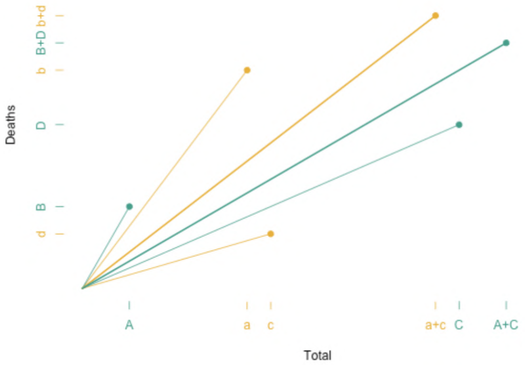
There is no mathematical "paradox", *per se*.

We could have

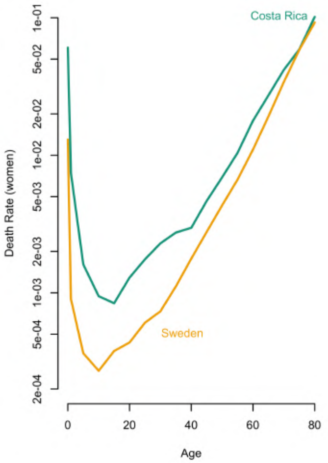
$$\frac{A}{B} \geq \frac{a}{b} \text{ and } \frac{C}{D} \geq \frac{c}{d}$$

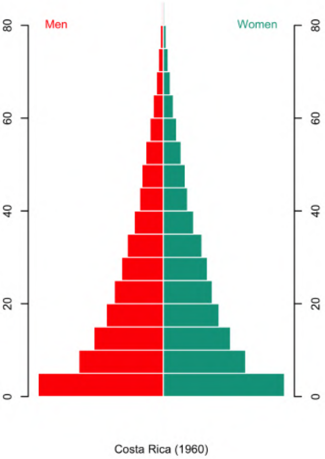
and at the same time

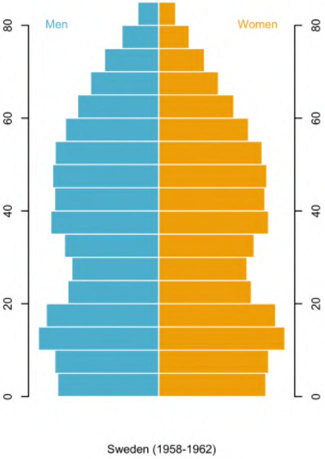
$$\frac{A+C}{B+D} \leq \frac{a+c}{b+d}$$



Motivation (6”. Mortality in Costa Rica and Sweden)





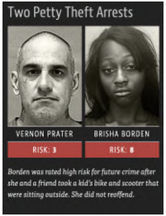


Overall mortality rate for women, 8.12‰ in Costa Rica, against 9.29‰ in Sweden.

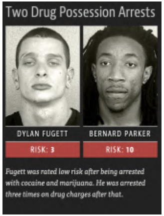
Motivation (7. Propublica, Actuarial Justice)

Concept of "actuarial justice"
as coined in Feeley and Simon (1994)

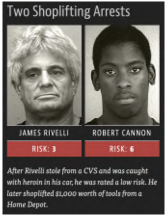
Correctional **O**ffender **M**anagement
Profiling for **A**lternative **S**anctions
(**COMPAS**), Perry (2013)



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.



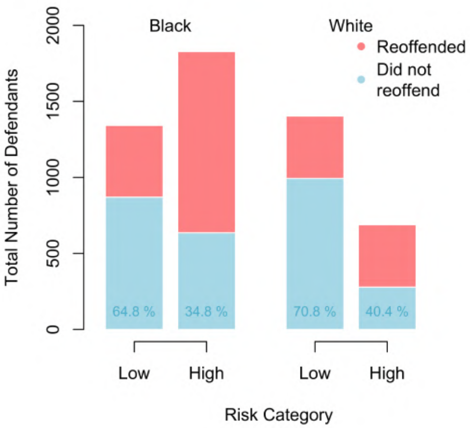
Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.



After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted $1,000 worth of tools from a Home Depot.

https://github.com/propublica/compas-analysis

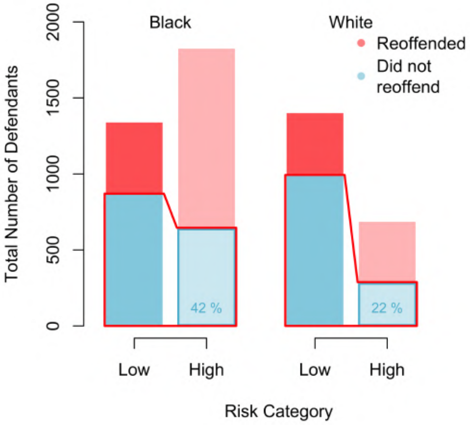
Angwin et al. (2016) Machine Bias
Dressel and Farid (2018)



Motivation (7. Propublica, Actuarial Justice)

From Feller et al. (2016),

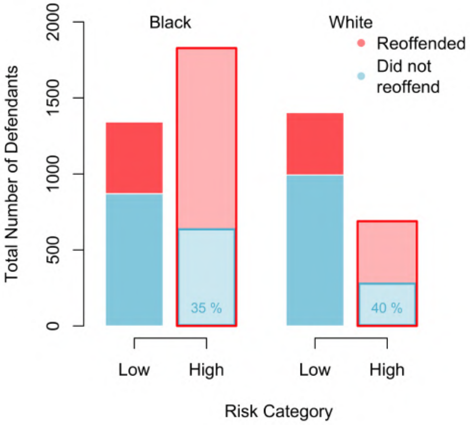
- for White people, among those who did not re-offend, 22% were wrongly classified,
- for Black people, among those who did not re-offend, 42% were wrongly classified,
- problem, since 42% $\gg$ 22%



Motivation (7. Propublica, Actuarial Justice)

From Dieterich et al. (2016),

- for White people, among those who where classified as high risk, 40% did not re-offend,
- for Black people, among those who where classified as high risk, 35% did not re-offend,
- no problem, since 40% $\approx$ 35%



Motivation (7. Propublica, Actuarial Justice)

Formalize the later,

$$\begin{cases} S: & \text{race (binary), black \& white} \\ Y: & \text{re-offense (binary), no \& yes} \\ \widehat{Y}: & \text{classifier (risk category), low \& high} \end{cases}$$

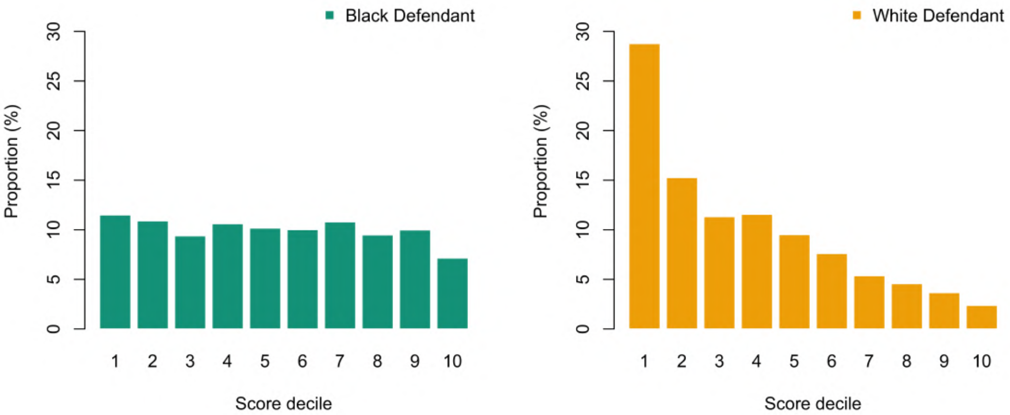
$$\mathbb{P}[\widehat{Y}=\text{high}|Y=\text{no}, \overset{\text{sensitive}}{S=\text{black}}] = 42\% \overset{?}{=} \mathbb{P}[\widehat{Y}=\text{high}|Y=\text{no}, \overset{\text{sensitive}}{S=\text{white}}] = 22\%,$$

false positive rate

$$\mathbb{P}[Y=\text{no}|\widehat{Y}=\text{high}, S=\text{black}] = 35\% \overset{?}{=} \mathbb{P}[Y=\text{no}|\widehat{Y}=\text{high}, S=\text{white}] = 40\%.$$

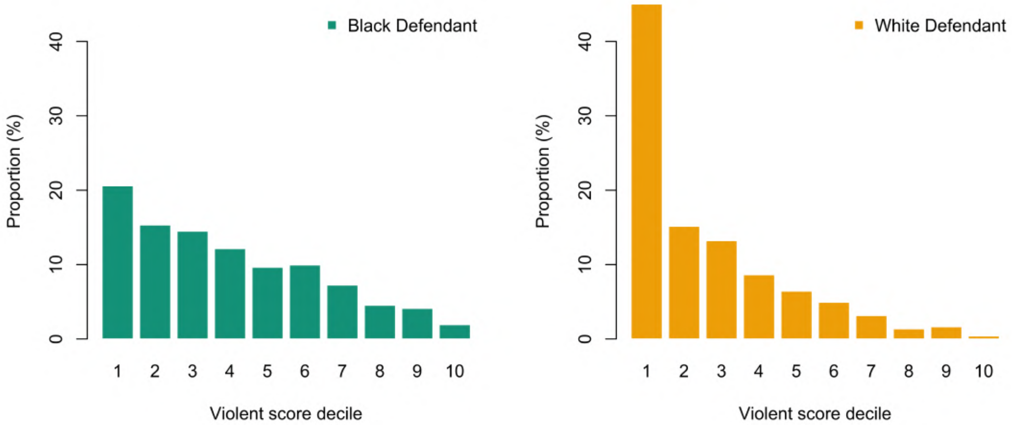
false discovery rate

Motivation (7. Propublica, Actuarial Justice)



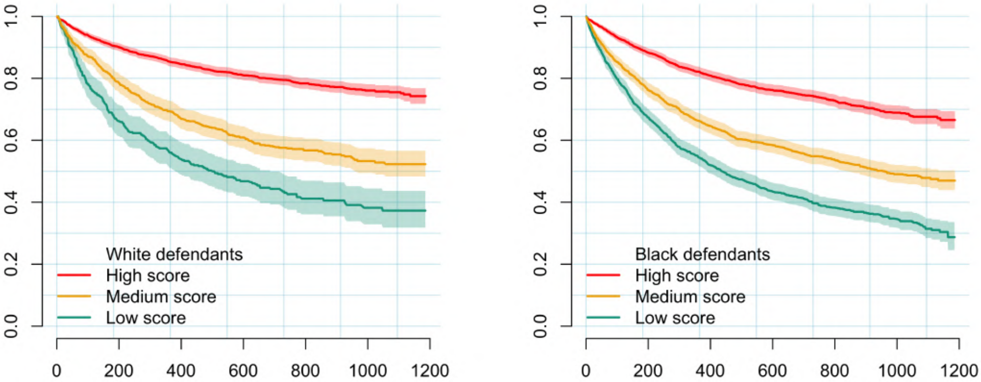
Look at score distributions, black and white defendant, Larson et al. (2016).

Motivation (7. Propublica, Actuarial Justice)



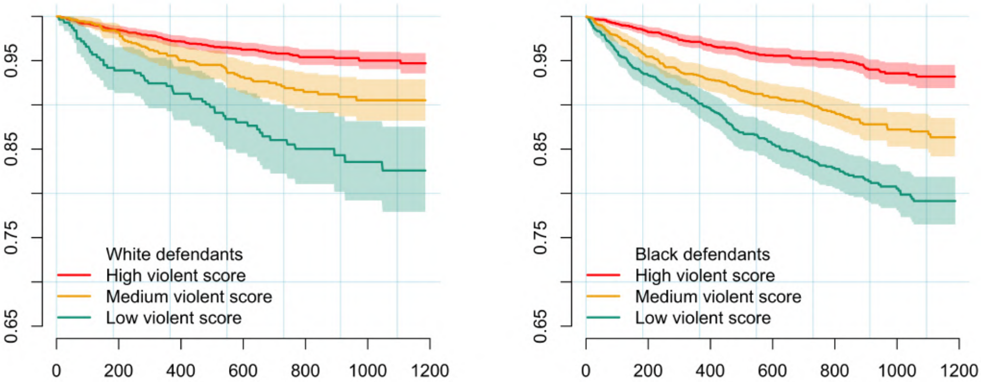
Look at score distributions, black and white defendant, Larson et al. (2016).

Motivation (7. Propublica, Actuarial Justice)



Cox Proportional Hazards model, black and white defendant, Larson et al. (2016).

Motivation (7. Propublica, Actuarial Justice)



Cox Proportional Hazards model, black and white defendant, Larson et al. (2016).

Motivation (8. Intention)

En France, Loi n° 2008-496 du 27 mai 2008

– Article 1 –

Constitue une discrimination indirecte une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifié par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés.

Extention de la "Loi n° 72-546 du 1 juillet 1972", qui supprima l'exigence de l'intention spécifique.

"*Technology is neither good nor bad; nor is it neutral* " , Kranzberg (1986)

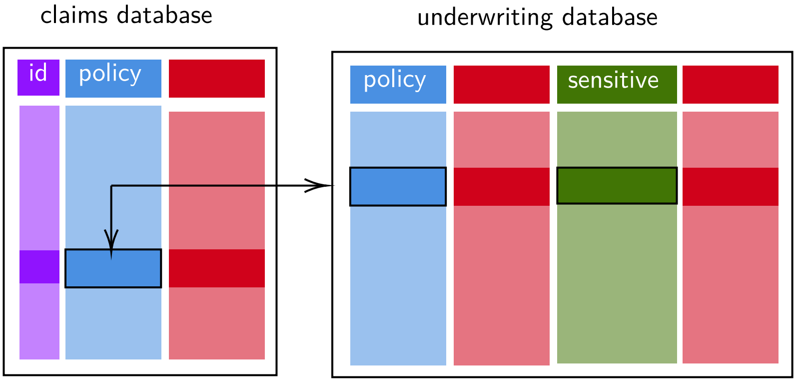
Motivation (9. Biases, biases everywhere...)

underwriters biases

- commercial discounts
- inferred data
- multiple decisions

claims biases

- fraud detection
- sexist mechanic
- ageist manager

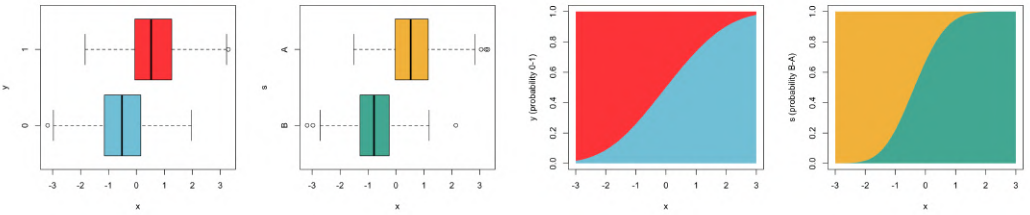


Datasets

`toydata1`

Consider a confounding Gaussian variable X_0, $X_0 \sim \mathcal{N}(0,1)$, and

$$\begin{cases} X = X_0 + \epsilon, \ \ \epsilon \sim \mathcal{N}(0, {}^1\!/{}_2{}^2), \\ S = \mathbf{1}(X_0 + \eta > 0), \ \ \eta \sim \mathcal{N}(0, {}^1\!/{}_2{}^2), \ \ s \in \{\mathtt{A}, \mathtt{B}\}, \\ Y = \mathbf{1}(X_0 + \nu > 0), \ \ \nu \sim \mathcal{N}(0, {}^1\!/{}_2{}^2), \ \ y \in \{0, 1\}. \end{cases}$$

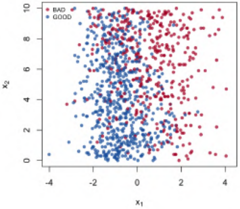


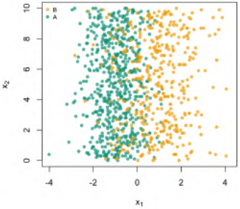
$x \mapsto \mathbb{P}[Y = 0 | X = x]$ (left-hand side) and $x \mapsto \mathbb{P}[S = \mathtt{A} | X = x]$ (right-hand side)

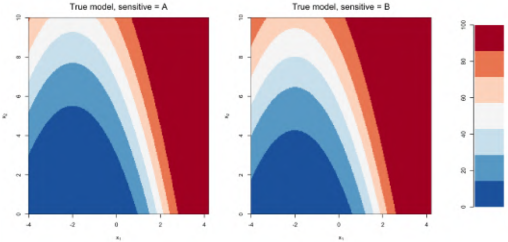
Datasets

`toydata2`

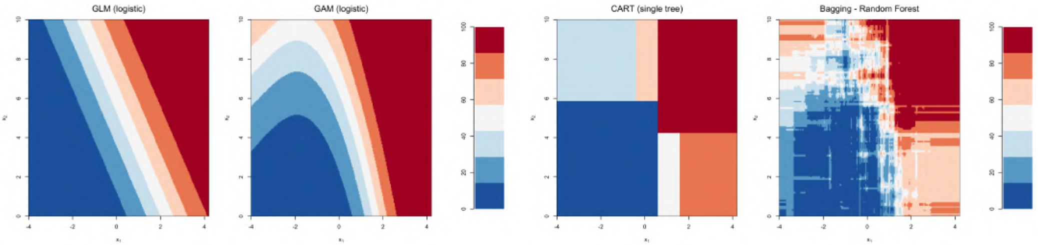
- binary sensitive attribute, $s \in \{\texttt{A}, \texttt{B}\}$, (60% and 40%)
- $(x_1, x_3) \sim \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, $r_{s=\texttt{A}} = 0.4$ and $r_{s=\texttt{B}} = 0.7$
- $x_2 \sim \mathcal{U}([0, 10])$, independent of x_1 and x_3
- $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 \mathbf{1}_{\texttt{B}}(s)$, that does not depend on x_3
- $y \sim \mathcal{B}(p)$ where $p = \exp(\eta)/[1 + \exp(\eta)] = \mu(x_1, x_2, s)$.





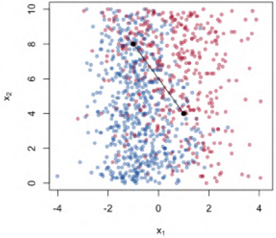


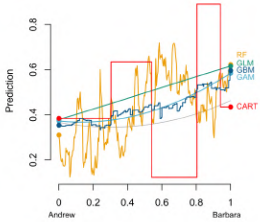
Datasets



Five models are considered

- plain GLM (logistic)
- GAM (cubic splines)
- CART (classification tree)
- RF (random forest)
- GBM (gradient boosting)





Datasets

`GermanCredit`, $m = 1,000$

- binary sensitive attribute, $s \in \{\mathtt{A}, \mathtt{B}\}$, (64% and 36%) corresponding to gender
- y denotes a default (30%)
- $x_1, \cdots, x_k$ denote legitimate credit variables (Duration, Purpose, Credit_amount, Age, Housing, Existing_credits, Foreign_worker, Resident_since, etc)

`FrenchMotor` (policy observe over one year), $n = 12,437$

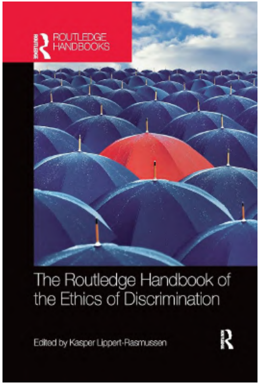
- binary sensitive attribute, $s \in \{\mathtt{A}, \mathtt{B}\}$, (31% and 69%) corresponding to gender
- y denotes the occurrence of a car accident (8.67%, unbalanced data)
- $x_1, \cdots, x_k$ denote legitimate credit variables (MariStat, VehAge, SocioCateg, DrivAge, VehBody, VehEnergy, VehMaxSpeed, Garage, VehUsage, etc)

– Part 1 –

Insurance

Discrimination and Insurance

”*What is unique about insurance is that even statistical discrimination which by definition is absent of any malicious intentions, poses significant moral and legal challenges. Why? Because on the one hand, policy makers would like insurers to treat their insureds equally, without discriminating based on race, gender, age, or other characteristics, even if it makes statistical sense to discriminate* (...) *On the other hand, at the core of insurance business lies discrimination between risky and non-risky insureds. But riskiness often statistically correlates with the same characteristics policy makers would like to prohibit insurers from taking into account.*” Avraham (2017)



Discrimination and Insurance

> **Solidarity**
> The political philosophy of the early twentieth century, condensed into the concept of solidarity, sought to offer both a scientific theory of social interdependence and a moral solution to social problems. According to some scholars, the emergence of this new rationality was made possible by the concept of social risk and the idea and technology of insurance developed to manage it. Social risk is defined as the risk to a group of people, statistically speaking, which is caused in one way or another by their living together and which can be mitigated by a technique of joint and several liability such as insurance.
> The way insurance works requires individuals to take a collective responsibility or the events they feel the need to prepare for. Society can be said to have become 'modern' when insurance becomes social insurance and when, thanks to the techniques and institutions of insurance, the insurance model becomes both a symbolic and a functional basis for the social contract.
> (...)

Discrimination and Insurance

> (...)
> Solidarity and justice are key principles underpinning the insurance system, according to Risto Pelkonen and Timo Somer. In the context of voluntary personal insurance, solidarity means that the insured share the benefits and costs between themselves, while justice means that each insured contributes to the costs according to the actuarial probability. Social insurance, on the other hand, is available to all citizens, regardless of their choice and health status, as the costs are covered by tax revenues and statutory contributions. W

Discrimination and Insurance

Definition 2.1: Mutuality, Wilkie (1997)

Mutuality is considered as the normal form of commercial private insurance, where participants contribute to the risk pool through a premium that relates to their particular risk at the time of the application, i.e., the higher the risk that they bring to the pool, the higher the premium required.

Definition 2.2: Solidarity, Wilkie (1997)

Solidarity is the basis of most national or social insurance schemes. Participation in such state-run schemes is generally compulsory and individuals have no discretion over their level of cover. All participants normally have the same level of cover. In solidarity schemes the contributions are not based on the expected risk of each participant.

Insurance Pricing and Predictive Modeling

"*Humans think in stories rather than facts, numbers or equations - and the simpler the story, the better*," Harari (2018). For insurers, it is often a mixture of both.

For Glenn (2000), insurer's risk selection process has two sides:

- the one presented to regulators and policyholders (numbers, statistics and objectivity),
- the other presented to underwriters (stories, character and subjective judgment).

The rhetoric of insurance exclusion – numbers, objectivity and statistics – forms what Brian Glenn calls "*the myth of the actuary*," "*a powerful rhetorical situation in which decisions appear to be based on objectively determined criteria when they are also largely based on subjective ones*" or "*the subjective nature of a seemingly objective process*".

Insurance Pricing and Predictive Modeling

Glenn (2003) claimed that there are many ways to rate accurately. Insurers can rate risks in many different ways depending on the stories they tell on which characteristics are important and which are not. “*The fact that the selection of risk factors is subjective and contingent upon narratives of risk and responsibility has in the past played a far larger role than whether or not someone with a wood stove is charged higher premiums*.” Going further, “*virtually every aspect of the insurance industry is predicated on stories first and then numbers*.”

“*all models are wrong but some models are useful*,” Box et al. (2011) (in other words, any model is at best a useful fable).

Insurance Pricing and Predictive Modeling

Definition 2.3: Pure premium (homogeneous risks)

Let Y be the non-negative random variable corresponding to the total annual loss associated with a given policy, then the pure premium is $\mathbb{E}[Y]$.

Proposition 2.1: Law of Large Numbers (2)

Consider an infinite collection of i.i.d. random variables $Y, Y_1, Y_2, \cdots, Y_n, \cdots$ in a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$, with finite expected value, then

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} Y_i}_{\text{(empirical) average}} \xrightarrow{\text{a.s.}} \underbrace{\mathbb{E}(Y)}_{\text{expected value}}, \text{ as } n \to \infty.$$

Insurance Pricing and Predictive Modeling

> **Expected Value**
> In probability theory, the expected value is a generalization of the weighted average. Informally, the expected value is the arithmetic mean of the possible values a random variable can take, weighted by the probability of those outcomes W

Following the "*law of the unconscious statistician*," in Schervish and DeGroot (2014), for some g,

$$\mathbb{E}[g(Y)] = \int_{-\infty}^{\infty} g(y) f_Y(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} g(y)\,\mathrm{d}F_Y(y).$$

Insurance Pricing and Predictive Modeling

More realistically, population is heterogeneous (with respect to risks), with some covariates $\boldsymbol{x}$ (legitimate, or not).

Definition 2.4: Pure premium (heterogeneous risks)

Let Y be the non-negative random variable corresponding to the total annual loss associated with a given policy, with covariates $\boldsymbol{x}$, then the pure premium is $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.

In this general setting, $\boldsymbol{x}$ consist in numeric or categorical variables.

Insurance Pricing and Predictive Modeling

Proposition 2.2: Law of Large Numbers (2')

Consider an infinite collection of i.i.d. random pairs $(\boldsymbol{X}, Y)$, $(\boldsymbol{X}_1, Y_1)$, $(\boldsymbol{X}_2, Y_2), \cdots, (\boldsymbol{X}_n, Y_n), \cdots$ in a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$, with finite expected value, then for any $\mathcal{A} \subset \mathcal{X}$ such that $\mathbb{P}[\boldsymbol{X} \in \mathcal{A}] > 0$,

$$\frac{\displaystyle\sum_{i=1}^{n} Y_i \mathbf{1}(\boldsymbol{X}_i \in \mathcal{A})}{\displaystyle\sum_{i=1}^{n} \mathbf{1}(\boldsymbol{X}_i \in \mathcal{A})} = \underbrace{\frac{1}{n_{\mathcal{A}}} \sum_{i \in \mathcal{I}_n(\mathcal{A})} Y_i}_{\text{conditional average}} \xrightarrow{\text{a.s.}} \underbrace{\mathbb{E}(Y | \boldsymbol{X} \in \mathcal{A})}_{\text{conditional expected value}}, \text{ as } n \to \infty,$$

where $\mathcal{I}_n(\mathcal{A}) = \{i : \boldsymbol{X}_i \in \mathcal{A}\} \subset \{1, 2, \cdots, n\}$ and $n_{\mathcal{A}} = \text{Card}(\mathcal{I}_n(\mathcal{A}))$.

Insurance Pricing and Predictive Modeling

Excerpt from the Men and Women life tables in 1720 (source: Struyck (1912)). Mortality, as a function of the age and the gender of the individual.

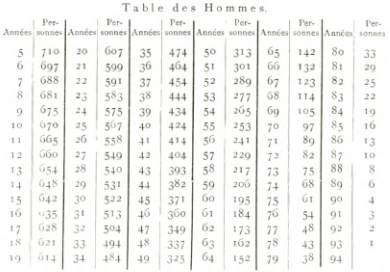


Table des Hommes.

Années	Per-sonnes	Années	Per-sonnes	Années	Per-sonnes	Années	Per-sonnes	Années	Per-sonnes	Années	Per-sonnes
5	710	20	607	35	474	50	313	65	142	80	33
6	697	21	599	36	464	51	301	66	132	81	29
7	688	22	591	37	454	52	289	67	123	82	25
8	681	23	583	38	444	53	277	68	114	83	22
9	675	24	575	39	434	54	265	69	105	84	19
10	670	25	567	40	424	55	253	70	97	85	16
11	665	26	558	41	414	56	241	71	89	86	13
12	660	27	549	42	404	57	229	72	82	87	10
13	654	28	540	43	393	58	217	73	75	88	8
14	648	29	531	44	382	59	206	74	68	89	6
15	642	30	522	45	371	60	195	75	61	90	4
16	635	31	513	46	360	61	184	76	54	91	3
17	628	32	504	47	349	62	173	77	48	92	2
18	621	33	494	48	337	63	162	78	43	93	1
19	614	34	484	49	325	64	152	79	38	94	

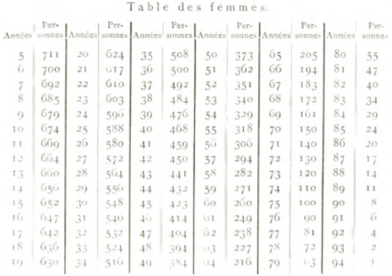


Table des femmes.

Années	Per-sonnes	Années	Per-sonnes	Années	Per-sonnes	Années	Per-sonnes	Années	Per-sonnes	Années	Per-sonnes
5	711	20	624	35	508	50	373	65	205	80	55
6	700	21	617	36	500	51	362	66	194	81	47
7	692	22	610	37	492	52	351	67	183	82	40
8	685	23	603	38	484	53	340	68	172	83	34
9	679	24	596	39	476	54	329	69	161	84	29
10	674	25	588	40	468	55	318	70	150	85	24
11	669	26	580	41	459	56	306	71	140	86	20
12	664	27	572	42	450	57	294	72	130	87	17
13	660	28	564	43	441	58	282	73	120	88	14
14	656	29	556	44	432	59	271	74	110	89	11
15	652	30	548	45	423	60	260	75	100	90	8
16	647	31	540	46	414	61	249	76	90	91	6
17	642	32	532	47	404	62	238	77	81	92	4
18	636	33	524	48	394	63	227	78	72	93	2
19	630	34	516	49	384	64	216	79	63	94	1

Insurance Pricing and Predictive Modeling

Excerpt from the Men and Women life tables in 1720 (source: Struyck (1912))
Mortality, as a function of the age and the gender of the individual.

men						women					
x	L_x	$_5p_x$	x	L_x	$_5p_x$	x	L_x	$_5p_x$	x	L_x	$_5p_x$
0	1000	29.0%	45	371	16.6%	0	1000	28.9%	45	423	11.8%
5	710	5.6%	50	313	19.2%	5	711	5.2%	50	373	14.7%
10	670	4.2%	55	253	22.9%	10	674	3.3%	55	318	18.2%
15	642	5.5%	60	195	27.2%	15	652	4.3%	60	260	21.2%
20	607	6.6%	65	142	31.7%	20	624	5.8%	65	205	26.8%
25	567	7.9%	70	97	37.1%	25	588	6.8%	70	150	33.3%
30	522	9.2%	75	61	45.9%	30	548	7.3%	75	100	45.0%
35	474	10.5%	80	33	51.5%	35	508	7.9%	80	55	56.4%
40	424	12.5%	85	16		40	468	9.6%	85	24	

Insurance Pricing and Predictive Modeling

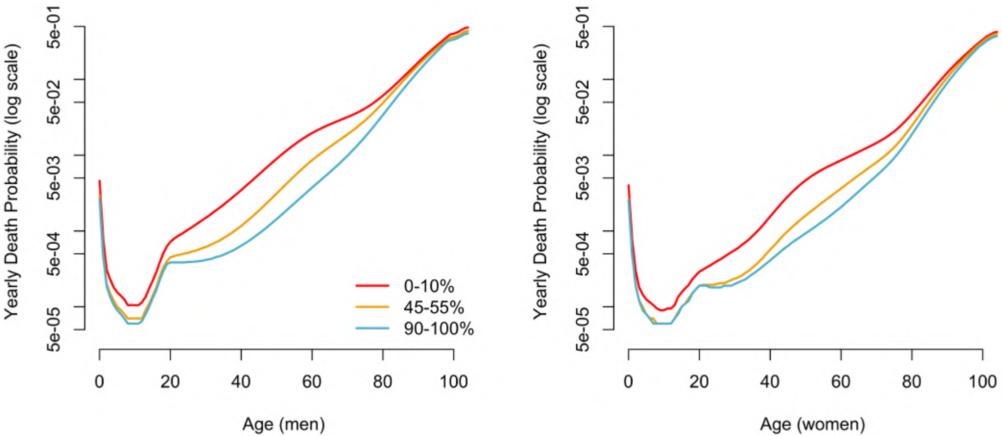
Excerpt from the Men and Women life tables in 2016 (source: Blanpain (2018))

Mortality, as a function of the age, the gender and the wealth of the individual.

men			
x	0-5%	45-50%	95-100%
0	100000	100000	100000
10	99299	99566	99619
20	99024	99396	99469
30	97930	98878	99094
40	95595	98058	98627
50	90031	96172	97757
60	77943	91050	95649
70	59824	79805	90399
80	38548	59103	76115
90	13337	23526	38837
100	530	1308	3231

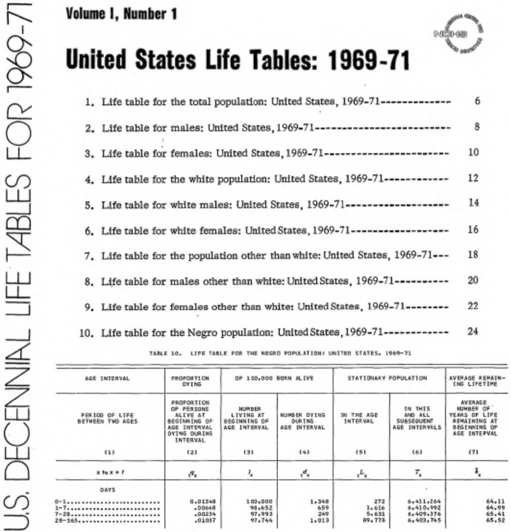
women			
x	0-5%	45-50%	95-100%
0	100000	100000	100000
10	99385	99608	99623
20	99227	99506	99526
30	98814	99302	99340
40	97893	98960	99074
50	95021	97959	98472
60	88786	95543	97192
70	79037	90408	94146
80	63224	79117	85825
90	31190	45750	55918
100	2935	5433	8717

Insurance Pricing and Predictive Modeling



Force of mortality (log scale) for various income quantile, in France, Blanpain (2018).

Insurance Pricing and Predictive Modeling



U.S. DECENNIAL LIFE TABLES FOR 1969-71

Volume I, Number 1

United States Life Tables: 1969-71

1. Life table for the total population: United States, 1969-71 ---- 6
2. Life table for males: United States, 1969-71 ---- 8
3. Life table for females: United States, 1969-71 ---- 10
4. Life table for the white population: United States, 1969-71 ---- 12
5. Life table for white males: United States, 1969-71 ---- 14
6. Life table for white females: United States, 1969-71 ---- 16
7. Life table for the population other than white: United States, 1969-71 ---- 18
8. Life table for males other than white: United States, 1969-71 ---- 20
9. Life table for females other than white: United States, 1969-71 ---- 22
10. Life table for the Negro population: United States, 1969-71 ---- 24

TABLE 10. LIFE TABLE FOR THE NEGRO POPULATION: UNITED STATES, 1969-71

AGE INTERVAL	PROPORTION DYING	OF 100,000 BORN ALIVE		STATIONARY POPULATION		AVERAGE REMAINING LIFETIME
PERIOD OF LIFE BETWEEN TWO AGES (1)	PROPORTION OF PERSONS ALIVE AT BEGINNING OF AGE INTERVAL DYING DURING INTERVAL (2)	NUMBER LIVING AT BEGINNING OF AGE INTERVAL (3)	NUMBER DYING DURING AGE INTERVAL (4)	IN THE AGE INTERVAL (5)	IN THIS AND ALL SUBSEQUENT AGE INTERVALS (6)	AVERAGE NUMBER OF YEARS OF LIFE REMAINING AT BEGINNING OF AGE INTERVAL (7)
x to $x+t$	${}_tq_x$	l_x	${}_td_x$	${}_tL_x$	T_x	$\mathring{e}_x$
DAYS						
0-1	0.01348	100,000	1,348	272	6,411,264	64.11
1-7	.00668	98,652	659	1,616	6,410,992	64.99
7-28	.00254	97,993	249	5,631	6,409,376	65.41
28-365	.01037	97,744	1,013	89,778	6,403,745	65.52

Mortality, gender and "race"

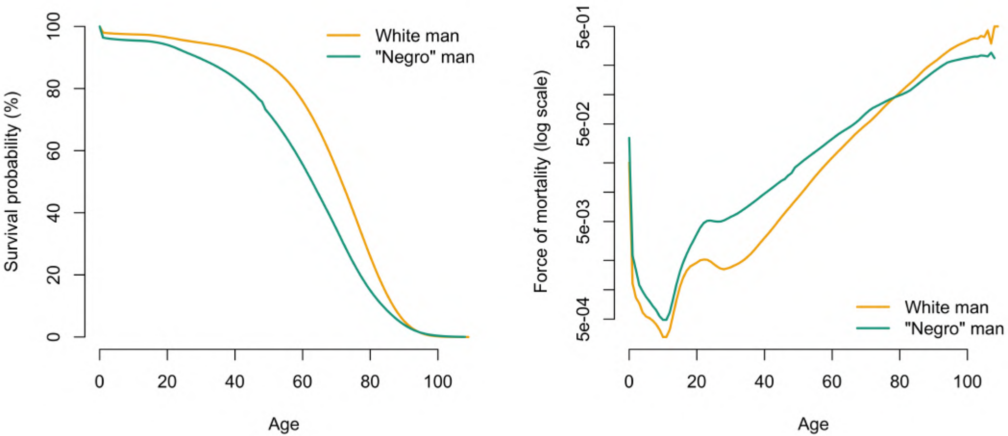
Frederick L. Hoffman

Hoffman (1896, 1918, 1931)

Insurance Pricing and Predictive Modeling

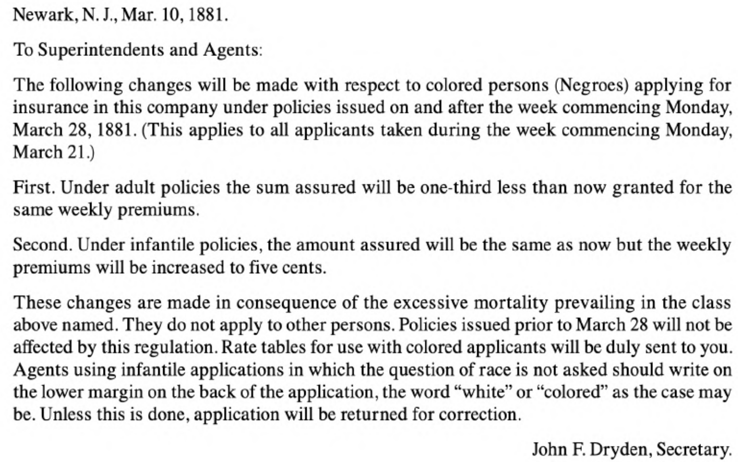
White, men						“Negro”, men					
x	L_x	${}_5p_x$	x	L_x	${}_5p_x$	x	L_x	${}_5p_x$	x	L_x	${}_5p_x$
0	100000	2.3%	55	83001	8.5%	0	100000	4.2%	55	66101	13.1%
5	97671	0.2%	60	75969	12.7%	5	95826	0.3%	60	57457	17.4%
10	97441	0.2%	65	66343	18.4%	10	95497	0.4%	65	47485	22.2%
15	97208	0.7%	70	54138	25.5%	15	95161	1.2%	70	36925	29.8%
20	96480	1.0%	75	40324	35.8%	20	94053	2.3%	75	25921	36.1%
25	95524	0.8%	80	25885	47.7%	25	91904	2.5%	80	16560	41.7%
30	94716	0.9%	85	13527	62.1%	30	89584	3.0%	85	9648	51.3%
35	93843	1.3%	90	5125	75.1%	35	86885	4.0%	90	4696	63.4%
40	92631	2.1%	95	1274	85.2%	40	83441	5.4%	95	1721	71.6%
45	90725	3.3%	100	189	90.5%	45	78976	7.2%	100	489	74.8%
50	87690	5.3%	105	18	100.0%	50	73282	9.8%	105	123	100.0%

Insurance Pricing and Predictive Modeling



Force of mortality (log scale) white men and ”Negro” men, 1968-71, U.S.

Insurance Pricing and Predictive Modeling



Newark, N. J., Mar. 10, 1881.

To Superintendents and Agents:

The following changes will be made with respect to colored persons (Negroes) applying for insurance in this company under policies issued on and after the week commencing Monday, March 28, 1881. (This applies to all applicants taken during the week commencing Monday, March 21.)

First. Under adult policies the sum assured will be one-third less than now granted for the same weekly premiums.

Second. Under infantile policies, the amount assured will be the same as now but the weekly premiums will be increased to five cents.

These changes are made in consequence of the excessive mortality prevailing in the class above named. They do not apply to other persons. Policies issued prior to March 28 will not be affected by this regulation. Rate tables for use with colored applicants will be duly sent to you. Agents using infantile applications in which the question of race is not asked should write on the lower margin on the back of the application, the word “white” or “colored” as the case may be. Unless this is done, application will be returned for correction.

John F. Dryden, Secretary.

Example of “direct discrimination”, from Plater (1997)

Insurance Pricing and Predictive Modeling

Definition 2.5: Balance Property

A pricing function m satisfies the balance property if $\mathbb{E}_{\boldsymbol{X}}[m(\boldsymbol{X})] = \mathbb{E}_Y[Y]$.

Proposition 2.3: Law of total expectations

$$\mathbb{E}_Y[Y] = \mathbb{E}_{\boldsymbol{X}}\big[\mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{X}]\big] = \mathbb{E}_{\boldsymbol{X}}\big[\mu(\boldsymbol{X})\big].$$

Proof Since $\mathbb{E}(Y) = \displaystyle\int y f_y(y)\mathrm{d}y$ and $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \displaystyle\int y f_{y|\boldsymbol{x}}(y|\boldsymbol{x})\mathrm{d}y$,

$$\mathbb{E}(\mathbb{E}(X|Y)) = \int \left(\int x\mathbb{P}[X = x|Y = y]\mathrm{d}x\right)\mathbb{P}[Y = y]\mathrm{d}y = \int\int x\mathbb{P}[X = x, Y = y]\mathrm{d}x\mathrm{d}y$$

$$= \int x\left(\int \mathbb{P}[X = x, Y = y]\mathrm{d}y\right)\mathrm{d}x = \int x\mathbb{P}[X = x]\mathrm{d}x = \mathbb{E}(X).$$

Insurance Pricing and Predictive Modeling

Homogeneous risk sharing

	Policyholder	Insurer
Loss	$\mathbb{E}[Y]$	$Y - \mathbb{E}[Y]$
Average loss	$\mathbb{E}[Y]$	0
Variance	0	$\text{Var}[Y]$

$\mathbb{E}[Y]$ is the premium paid, and Y the total loss,
from De Wit and Van Eeghen (1984) and Denuit and Charpentier (2004)

Insurance Pricing and Predictive Modeling

Heterogeneous risk sharing, with perfect information

	Policyholder	Insurer
Loss	$\mathbb{E}[Y\|\Theta]$	$Y-\mathbb{E}[Y\|\Theta]$
Average loss	$\mathbb{E}[Y]$	0
Variance	$\text{Var}[\mathbb{E}[Y\|\Theta]]$	$\text{Var}[Y-\mathbb{E}[Y\|\Theta]]$

where Θ denotes the heterogeneous risk factor.

The term on the bottom right is $\mathbb{E}[\text{Var}[Y|\Theta]]$, corresponding to the standard variance decomposition (or Pythagoras theorem)

$$\text{Var}[Y]=\text{Var}[\mathbb{E}[Y|\Theta]]+\mathbb{E}[\text{Var}[Y|\Theta]].$$

Insurance Pricing and Predictive Modeling

Proposition 2.4: Variance decomposition (1)

For any measurable random variable Y with finite variance

$$\text{Var}[Y] = \underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\rightarrow \text{ insurer}} + \underbrace{\text{Var}[\mathbb{E}[Y|\Theta]]}_{\rightarrow \text{ policyholder}}.$$

Proof:

$$\begin{aligned}\text{Var}[Y] &= \mathbb{E}\left[Y^2\right] - \text{E}[Y]^2 = \mathbb{E}\left[\text{Var}[Y|\Theta] + \mathbb{E}[Y|\Theta]^2\right] - \mathbb{E}[\mathbb{E}[Y|\Theta]]^2 \\ &= (\mathbb{E}[\text{Var}[Y|\Theta]]) + \left(\mathbb{E}\left[\mathbb{E}[Y|\Theta]^2\right] - \mathbb{E}[\mathbb{E}[Y|\Theta]]^2\right) = \mathbb{E}[\text{Var}[Y|\Theta]] + \text{Var}[\mathbb{E}[Y|\Theta]].\end{aligned}$$

Insurance Pricing and Predictive Modeling

Heterogeneous risk sharing, with imperfect information

	Policyholder	Insurer
Loss	$\mathbb{E}[Y\|\boldsymbol{X}]$	$Y - \mathbb{E}[Y\|\boldsymbol{X}]$
Average loss	$\mathbb{E}[Y]$	0
Variance	$\text{Var}[\mathbb{E}[Y\|\boldsymbol{X}]]$	$\mathbb{E}[\text{Var}[Y\|\boldsymbol{X}]]$

$$\mathbb{E}[\text{Var}[Y|\boldsymbol{X}]] = \underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\text{perfect ratemaking}} + \underbrace{\mathbb{E}\{\text{Var}[\mathbb{E}[Y|\Theta]|\boldsymbol{X}]\}}_{\text{misclassification}}$$

This "misclassification" term (on the right) is called "*subsidierende solidariteit*" in De Pril and Dhaene (1996), or "*subsidiary solidarity*", as opposed to "*kanssolidariteit*" or "*random solidarity*" term (on the left).

Insurance Pricing and Predictive Modeling

Proposition 2.5: Variance decomposition (2)

For any measurable random variable Y with finite variance

$$\text{Var}[Y] = \underbrace{\mathbb{E}[\text{Var}[Y|\boldsymbol{X}]]}_{\to \text{ insurer}} + \underbrace{\text{Var}[\mathbb{E}[Y|\boldsymbol{X}]]}_{\to \text{ policyholder}},$$

where

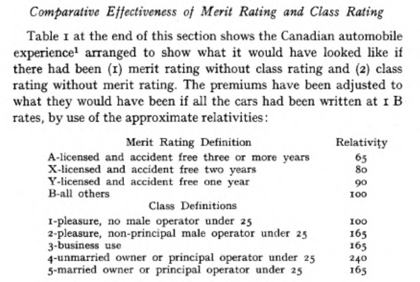
$$\begin{aligned} \mathbb{E}[\text{Var}[Y|\boldsymbol{X}]] &= \mathbb{E}[\mathbb{E}[\text{Var}[Y|\Theta]|\boldsymbol{X}]] + \mathbb{E}[\text{Var}[\mathbb{E}[Y|\Theta]|\boldsymbol{X}]] \\ &= \underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\text{perfect ratemaking}} + \underbrace{\mathbb{E}\{\text{Var}[\mathbb{E}[Y|\Theta]|\boldsymbol{X}]\}}_{\text{misclassification}}. \end{aligned}$$

Clubs, Group and Categories

Groups, or risk classes, are built on the basis of available data, and exist primarily as the product of actuarial models.

For example, as mentioned in Bailey and Simon (1960), in motor insurance five risk classes can be considered,

- “*pleasure, no male operator under 25*,” (reference),
- “*pleasure, non-principal male operator under 25*,” +65%,
- “*business use*,” +65%,
- “*married owner or principal operator under 25*,” +65%,
- “*unmarried owner or principal operator under 25*,” +140%.



Comparative Effectiveness of Merit Rating and Class Rating

Table 1 at the end of this section shows the Canadian automobile experience[1] arranged to show what it would have looked like if there had been (1) merit rating without class rating and (2) class rating without merit rating. The premiums have been adjusted to what they would have been if all the cars had been written at 1 B rates, by use of the approximate relativities:

Merit Rating Definition	Relativity
A-licensed and accident free three or more years	65
X-licensed and accident free two years	80
Y-licensed and accident free one year	90
B-all others	100
Class Definitions	
1-pleasure, no male operator under 25	100
2-pleasure, non-principal male operator under 25	165
3-business use	165
4-unmarried owner or principal operator under 25	240
5-married owner or principal operator under 25	165

There is no “physical basis” for group members to identify other members of *their* group, in the sense that they usually don’t share anything, except some common characteristics, Gandy (2016).

Clubs, Group and Categories

solidarity until 1930 → risk classes 1930-2010 → personalization post 2010

In ancient Rome, a *collegium* (plural *collegia*) was an association, such as military *collegia*, Verboven (2011).

As explained in Ginsburg (1940), upon the completion of his service a veteran had the right to join one of the many *collegia veteranorum* in each legion.

In case of retirement, upon the completion of his term of service, the soldier would received a a lump sum which helped him somewhat to arrange the rest of his life. The membership in a *collegium* gave him a mutual insurance against "*unforeseen risks*." These *collegia*, besides being cooperative insurance companies, had other functions.

Clubs, Group and Categories

In the early 1660th, the Pirate's Code was supposedly written by Portuguese buccaneer Bartolomeu Português.

A section is explicitly dedicated to insurance and benefits: "*a standard compensation is provided for maimed and mutilated buccaneers. Thus they order for the loss of a right arm six hundred pieces of eight, or six slaves; for the loss of a left arm five hundred pieces of eight, or five slaves; for a right leg five hundred pieces of eight, or five slaves; for the left leg four hundred pieces of eight, or four slaves; for an eye one hundred pieces of eight, or one slave; for a finger of the hand the same reward as for the eye*," see Barbour (1911) (or more recently Leeson (2009) and Fox (2013) about this piratical schemes).

Clubs, Group and Categories

In the XIX-th century, in Europe, mutual aid societies involved a group of individuals who made regular payments into a common fund in order to provide for themselves in later, unforeseeable moments of financial hardship or of old age. As mentioned by Garrioch (2011), in 1848, there were in Paris 280 mutual aidsocieties with well over 20,000 members.

For example, the *Société des Arts Graphiques*, was created in 1808. It admitted only men over twenty and under fifty, and it charged much higher admission and annual fees for those who joined at a more advanced age. In return, they received benefits if they were unable to work, reducing over a period of time, but in case of serious illness the Society would pay the admission fee for a hospice. In England, there were "friendly societies," as described in Ismay (2018).

Clubs, Group and Categories

The money collected through contributions came to the rescue of unfortunate workers, who would no longer have any reason to radicalize. It was proposed that insurance should become compulsory (Bismark proposed this in Germany in 1883), but the idea was rejected in favor of giving workers the freedom to contribute, as the only way to moralize the working classes, as Da Silva (2023) explains.

In 1852, of the 236 mutual funds created, 21 were on a professional basis, while the other 215 were on a territorial basis. And from 1870 onwards, mutual funds diversified the professional profile of contributors beyond blue-collar workers, and expanded to include employees, civil servants, the self-employed and artists.

The amount of the premium is not linked to the risk.

Clubs, Group and Categories

As Da Silva (2023) puts it, "*mutual insurers see in the actuarial figure the programmed end of solidarity.*" For mutual funds, solidarity is essential, with everyone contributing according to their means and receiving according to their needs. Around the same time, in France, the first insurance companies appeared, based on risk selection, and the first mathematical approaches to calculating premiums.

Hubbard (1852) advocates the introduction of an "*English-style scientific organization*" in their management. For its members, they had to be able to know "*the probable average of the claims*" that they should cover, like insurance companies. The development of tables should lead insurers to adopt the principle of contributions varying according to the age of entry and the specialization of contributions and funds (health/retirement).

For Stone (1993) and Gowri (2014) the defining feature of "modern insurance" is its reliance on segmenting the risk pool into distinct categories, each receiving a price

Clubs, Group and Categories

corresponding to the particular risk that the individuals assigned to that category are expected to represent (as accurately as can be estimated by actuaries).

Once heterogeneity with respect to the risk was observed in portfolios, insurers have operated by categorizing individuals into risk classes and assigning corresponding tariffs. This ongoing process of categorization ensures that the sums collected, on average, are sufficient to address the realized risks within specific groups.

The aim of risk classification, as explained in Wortham (1986), is to identify the specific characteristics that are supposed to determine an individual's propensity to suffer an adverse event, forming groups within which the risk is (approximately) equally shared. The problem, of course, is that the characteristics associated with various types of risk are almost infinite; as they cannot all be identified and priced in every risk classification system, there will necessarily be unpriced sources of heterogeneity between individuals in a given risk class.

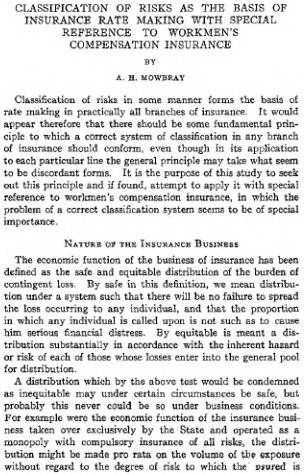
Clubs, Group and Categories

In 1915, as mentioned in Rothstein (2003), the president of the Association of Life Insurance Medical Directors of America noted that the question asked almost universally of the Medical Examiner was "*What is your opinion of the risk? Good, bad, first-class, second-class, or not acceptable?*" Historically, insurance prices were a (finite) collection of prices (maybe more than than the two classes mentioned, "*first-class*" and "*second-class*").

In the early 1920's, Albert Henry Mowbray, who worked for New York Life Insurance Company and later Liberty Mutual (and was also an actuary for state-level insurance commissions in New Carolina and California, and the National Council on Workmen's Insurance) gives his perspective on insurance rate making. See Mowbray (1921).

Clubs, Group and Categories

“*Classification of risks in some manner forms the basis of rate making in practically all branches of insurance. It would appear therefore that there should be some fundamental principle to which a correct system of classification in any branch of insurance should conform* (...) *As long ago as the days of ancient Greece and Rome the gradual transition of natural phenomena was observed and set down in the Latin maxim,* ‘natura non agit per altum’. *If each risk, therefore is to be precisely rated, it would be necessary to recognize very minute differences and precisely measure them.*



CLASSIFICATION OF RISKS AS THE BASIS OF INSURANCE RATE MAKING WITH SPECIAL REFERENCE TO WORKMEN'S COMPENSATION INSURANCE

BY

A. H. MOWBRAY

Classification of risks in some manner forms the basis of rate making in practically all branches of insurance. It would appear therefore that there should be some fundamental principle to which a correct system of classification in any branch of insurance should conform, even though in its application to each particular line the general principle may take what seem to be discordant forms. It is the purpose of this study to seek out this principle and if found, attempt to apply it with special reference to workmen's compensation insurance, in which the problem of a correct classification system seems to be of special importance.

NATURE OF THE INSURANCE BUSINESS

The economic function of the business of insurance has been defined as the safe and equitable distribution of the burden of contingent loss. By safe in this definition, we mean distribution under a system such that there will be no failure to spread the loss occurring to any individual, and that the proportion in which any individual is called upon is not such as to cause him serious financial distress. By equitable is meant a distribution substantially in accordance with the inherent hazard or risk of each of those whose losses enter into the general pool for distribution.

A distribution which by the above test would be condemned as inequitable may under certain circumstances be safe, but probably this never could be so under business conditions. For example were the economic function of the insurance business taken over exclusively by the State and operated as a monopoly with compulsory insurance of all risks, the distribution might be made pro rata on the volume of the exposure without regard to the degree of risk to which the insured is

Clubs, Group and Categories

> “*Since we are not capable of covering a large field fully and at the same time recognizing small differences in all parts of the field, it is natural that we resort to subdivision of the field by means of classification, thereby concentrating our attention on a smaller interval which may again be subdivided by further classification, and the system so carried on to the limit to which we find it necessary or desirable to go. But however far we may go in any system of classification, whether in the field of pure or applied science including the business or insurance, we shall always find difficulties presented by the borderline case, difficulties which arise from the continuous character of natural phenomena which we are attempting to place in more or less arbitrary divisions. While thus acknowledging that classification will never completely solve the problem of recognizing differences between individuals, nevertheless classification seems to be necessary at least as a preliminary step toward such recognition in any field of study. The fact that a complete and final solution cannot be made is, therefore, no justification for completely discarding*

Clubs, Group and Categories

classification as a method of approach. Since it is insurance hazards that we undertake to measure and classify, the preliminary step in studying classification theory may well be to ask what is an insurance hazard and how it may be determined. It must be evident to the members of this Society that an insurance hazard is what is termed "a mathematical expectation," that is a product of a sum at risk and the probability of loss from the conditions insured against, e.g., the destruction of a piece of property by fire, the death of an individual, etc. If the net premiums collected are so determined on the basis of the true natural probability and there is a sufficient spread then the sums collected will just cover the losses and this is what should be," Mowbray (1921)

"*1. The classification should bring together risks which have inherent in their operation the same causes of loss.*

2. The variation from risk to risk in the strength of each cause or at least of the more important should not be greater than can be handled by the formula by

Clubs, Group and Categories

which the classification is subdivided, i.e., the Schedule and / or Experience Rating Plan used.
3. The classification should not cover risks which include, as important elements of their hazard, causes which are not common to all.
4. The classification system and the formula for its extension (Schedule and / or Experience Rating Plans) should be harmonious.
5. The basis throughout should be the outward, recognizable indicia of the presence and potency of the several inherent causes of loss including extent as well as occurrence of loss,” Mowbray (1921).

Clubs, Group and Categories

Several articles and textbooks in sociology tried to understand how classification mechanisms establish symbolic boundaries that reinforce group identities, such as Bourdieu (2018), Massey (2007), Fourcade and Healy (2013).

But here, those "groups" or "classes" do not share any identity, and Simon (1988) or Harcourt (2015) use the term "actuarial classification" (where "actuarial" designates any decision-making technique that relies on predictive statistical methods, replacing more holistic or subjective forms of judgment). In those class-based systems, based on insurance rating table (or grid), results are determined by assigning individuals to a group in which each person is positioned as "average" or "typical".

[Most] "*actuaries cannot think of individuals except as members of groups*" claimed Brilmayer et al. (1979). Each individual is assigned the same value as all other members of the group to which it is assigned.

Clubs, Group and Categories

Simon (1987, 1988), and then Feeley and Simon (1992), defined "actuarialism," that designate the use of statistics to guide "*class-based decision-making*," used to price pensions and insurance. As explained in Harcourt (2015), this "actuarial classification" is the constitution of groups with no experienced social significance for the participants. A person classified as a particular risk by an insurance company shares nothing with the other people so classified, apart from a series of formal characteristics (e.g. age, sex, marital status, etc.).

For Austin (1983) and Simon (1988), categories used by the insurance company when grouping risks are "*singularly sterile*," resulting in inert, immobile and deactivated communities, corresponding to "*artificial*" groups. These are not groups organized around a shared history, common experiences or active commitment, forming some "*aggregates*" – living only in the imagination of the actuary who calculates and tabulates, not in any lived form of human association.

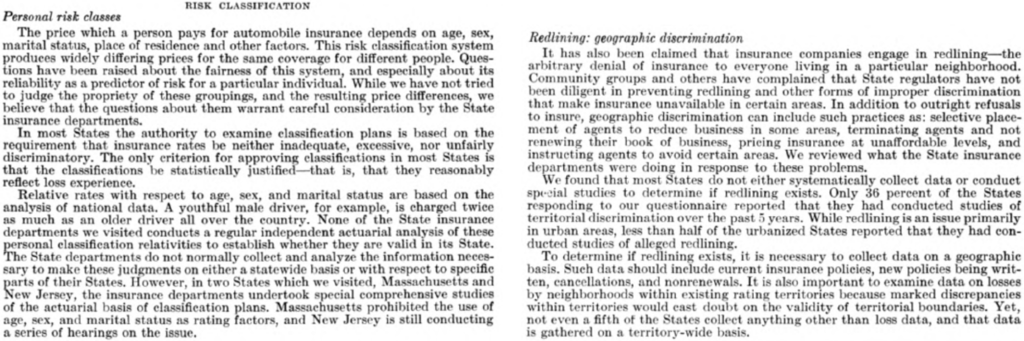
Clubs, Group and Categories

If Hacking (1990) observed that standard classes creates coherent group identities (causing possible stereotypes and discrimination, Simon (1988), provocatively suggests that actuarial classifications can in turn "*undo people's identity*."

As mentioned in Abraham (1986), the goal for actuaries is to create groups, or "classes" made up of individuals who share a series of common characteristics and are therefore presumed to represent the same risk. Following François (2022), we could claim that actuarial techniques reduce individuals to a series of formal roles that have no "*moral density*" and therefore do not grant an "*identity*" that organizes a coherent sense of self. And the inclusion of nominally "*demoralized categories*," such as gender, in class-based rating systems makes their total demoralization difficult to achieve – and is in itself an issue of struggle. Heimer (1985) used the term "*community of fate*."

Rouvroy et al. (2013) and Cheney-Lippold (2017) point out that scoring technologies are continually swapping predictors, "*shuffling the cards*," so that there is no stable basis for constructing group memberships, or a coherent sense.

Clubs, Group and Categories



RISK CLASSIFICATION

Personal risk classes

The price which a person pays for automobile insurance depends on age, sex, marital status, place of residence and other factors. This risk classification system produces widely differing prices for the same coverage for different people. Questions have been raised about the fairness of this system, and especially about its reliability as a predictor of risk for a particular individual. While we have not tried to judge the propriety of these groupings, and the resulting price differences, we believe that the questions about them warrant careful consideration by the State insurance departments.

In most States the authority to examine classification plans is based on the requirement that insurance rates be neither inadequate, excessive, nor unfairly discriminatory. The only criterion for approving classifications in most States is that the classifications be statistically justified—that is, that they reasonably reflect loss experience.

Relative rates with respect to age, sex, and marital status are based on the analysis of national data. A youthful male driver, for example, is charged twice as much as an older driver all over the country. None of the State insurance departments we visited conducts a regular independent actuarial analysis of these personal classification relativities to establish whether they are valid in its State. The State departments do not normally collect and analyze the information necessary to make these judgments on either a statewide basis or with respect to specific parts of their States. However, in two States which we visited, Massachusetts and New Jersey, the insurance departments undertook special comprehensive studies of the actuarial basis of classification plans. Massachusetts prohibited the use of age, sex, and marital status as rating factors, and New Jersey is still conducting a series of hearings on the issue.

Redlining: geographic discrimination

It has also been claimed that insurance companies engage in redlining—the arbitrary denial of insurance to everyone living in a particular neighborhood. Community groups and others have complained that State regulators have not been diligent in preventing redlining and other forms of improper discrimination that make insurance unavailable in certain areas. In addition to outright refusals to insure, geographic discrimination can include such practices as: selective placement of agents to reduce business in some areas, terminating agents and not renewing their book of business, pricing insurance at unaffordable levels, and instructing agents to avoid certain areas. We reviewed what the State insurance departments were doing in response to these problems.

We found that most States do not either systematically collect data or conduct special studies to determine if redlining exists. Only 36 percent of the States responding to our questionnaire reported that they had conducted studies of territorial discrimination over the past 5 years. While redlining is an issue primarily in urban areas, less than half of the urbanized States reported that they had conducted studies of alleged redlining.

To determine if redlining exists, it is necessary to collect data on a geographic basis. Such data should include current insurance policies, new policies being written, cancellations, and nonrenewals. It is also important to examine data on losses by neighborhoods within existing rating territories because marked discrepancies within territories would cast doubt on the validity of territorial boundaries. Yet, not even a fifth of the States collect anything other than loss data, and that data is gathered on a territory-wide basis.

> “*The price which a person pays for automobile insurance depends on age, sex, marital status, place of residence and other factors. This risk classification system produces widely differing prices for the same coverage for different people. Questions have been raised about the fairness of this system, and especially about its reliability as a predictor of risk for a particular individual.*

Clubs, Group and Categories

> *While we have not tried to judge the propriety of these groupings, and the resulting price differences, we believe that the questions about them warrant careful consideration by the State insurance departments. In most States the authority to examine classification plans is based on the requirement that insurance rates are neither inadequate, excessive, nor unfairly discriminatory. The only criterion for approving classifications in most States is that the classifications be statistically justified – that is, that they reasonably reflect loss experience. Relative rates with respect to age, sex, and marital status are based on the analysis of national data. A youthful male driver, for example, is charged twice as much as an older driver all over the country* (...) *It has also been claimed that insurance companies engage in redlining – the arbitrary denial of insurance to everyone living in a particular neighborhood. Community groups and others have complained that State regulators have not been diligent in preventing redlining and other forms of improper discrimination that make insurance unavailable in certain areas. In addition to outright refusals to*

Clubs, Group and Categories

> *insure, geographic discrimination can include such practices as: selective placement of agents to reduce business in some areas, terminating agents and not renewing their book of business, pricing insurance at un-affordable levels, and instructing agents to avoid certain areas. We reviewed what the State insurance departments were doing in response to these problem. To determine if redlining exists, it is necessary to collect data on a geographic oasis. Such data should include current insurance policies, new policies being written, cancellations, and non-renewals. It is also important to examine data on losses by neighborhoods within existing rating territories because marked discrepancies within territories would cast doubt on the validity of territorial boundaries. Yet, not even a fifth of the States collect anything other than loss data, and that data is gathered on a territory-wide basis*," Havens (1979)

Clubs, Group and Categories

> "*On the other hand, the opinion that distinctions based on sex, or any other group variable, necessarily violate individual rights reflects ignorance of the basic rules of logical inference in that it would arbitrarily forbid the use of relevant information. It would be equally fallacious to reject a classification system based on socially acceptable variables because the results appear discriminatory. For example, a classification system may be built on use of car, mileage, merit rating, and other variables, excluding sex. However, when verifying the average rates according to sex one may discover significant differences between males and females. Refusing to allow such differences would be attempting to distort reality by choosing to be selectively blind. The use of rating territories is a case in point. Geographical divisions, however designed, are often correlated with socio-demographic factors such as income level and race because of natural aggregation or forced segregation according to these factors. Again we conclude that insurance companies should be free to delineate territories and assess territorial differences as well as they can. At the*

Clubs, Group and Categories

same time, insurance companies should recognize that it is in their best interest to be objective and use clearly relevant factors to define territories lest they be accused of invidious discrimination by the public. (...) ” Casey et al. (1976)

“*One possible standard does exist for exception to the counsel that particular rating variables should not be proscribed. What we have called ‘equal treatment’ standard of fairness may precipitate a societal decision that the process of differentiating among individuals on the basis of certain variables is discriminatory and intolerable. This type of decision should be made on a specific, statutory basis. Once taken, it must be adhered to in private and public transactions alike and enforced by the insurance regulator. This is, in effect, a standard for conduct that by design transcends and preempts economic considerations. Because it is not applied without economic cost, however, insurance regulators and the industry should participate in and inform legislative deliberations that would ban the, use of particular rating variables as discriminatory.*” Casey et al. (1976)

Price Optimization

Decision theory under uncertainty (see Charpentier (2014)),

$$X \preceq Y \iff \mathcal{R}(X) \leq \mathcal{R}(Y),$$

A classical representation is $\mathcal{R}(Y) = \mathbb{E}[u(\omega - Y)]$, as in Neumann and Morgenstern (1947), where ω is the initial wealth.

u denotes the utility of the agent

Let π denote the premium asked to transfer risk (loss) Y,

$$\begin{cases} u(\omega - \pi) > \mathbb{E}\big[u(\omega - Y)\big] : & \text{purchases insurance} \\ u(\omega - \pi) < \mathbb{E}\big[u(\omega - Y)\big] : & \text{does not purchase insurance} \end{cases}$$

Find π such that $u(\omega - \pi) = \mathbb{E}\big[u(\omega - Y)\big]$.

Price Optimization

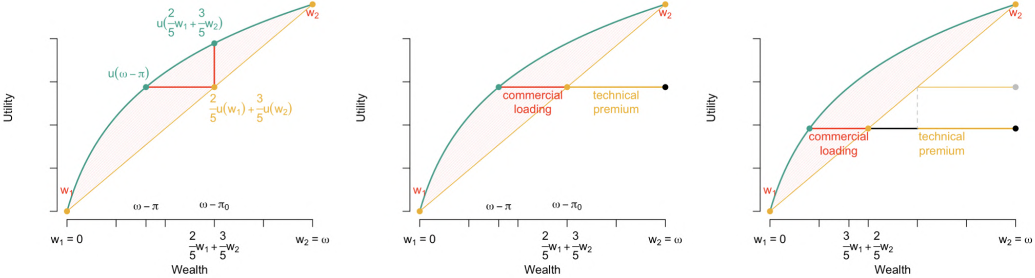
π such that $u(\omega-\pi)=\mathbb{E}\big[u(\omega-Y)\big]$ could be seem as the willingness to pay to transfer the risk.

> **Willingness to Pay**
> In behavioral economics, willingness to pay (WTP) is the maximum price at or below which a consumer will definitely buy one unit of a product.[1] This corresponds to the standard economic view of a consumer reservation price. W

Price Optimization

Definition 2.6: Indifference utility principle

Let Y be the non-negative random variable corresponding to the total annual loss associated with a given policy, for a policyholder with utility u and wealth w, the indifference premium is $\pi = \omega - u^{-1}\big(\mathbb{E}\big[u(\omega - Y)\big]\big)$.



Price Optimization

> **Price Walking**
> Price walking, or the loyalty penalty, is a form of price discrimination whereby longstanding, loyal customers of a service provider are charged higher prices for the same services compared to customers that have just switched to that provider. The pricing strategy is common in the insurance and telecommunications industries. It is used to acquire new customers with artificially low rates or other incentives not available to existing clients, effectively using existing customers to subsidize the prices offered to new clients. W

– Part 2 –

Machine / Statistical Learning

Statistical Learning

Proposition 3.1: Law of Large Numbers (1)

Consider an infinite collection of i.i.d. random variables $Y, Y_1, Y_2, \cdots, Y_n, \cdots$ in a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$, then

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_i \in \mathcal{A})}_{\text{(empirical) frequency}} \xrightarrow{\text{a.s.}} \underbrace{\mathbb{P}(\{Y \in \mathcal{A}\})}_{\text{probability}} = \mathbb{P}[Y \in \mathcal{A}], \text{ as } n \to \infty.$$

"*law of the unconscious statistician*," (Ross (2014) and Casella and Berger (1990)), "*statisticians make liberal use of conditioning arguments to shorten what would otherwise be long proofs*," Proschan and Presnell (1998).

$$\mathbb{P}(Y \in \mathcal{A}|X = x) = \lim_{\epsilon\to 0} \frac{\mathbb{P}(\{Y \in \mathcal{A}\} \cap \{|X - x| \leq \epsilon\})}{\mathbb{P}(\{|X - x| \leq \epsilon\})} = \lim_{\epsilon\to 0} \mathbb{P}(Y \in \mathcal{A}||X - x| \leq \epsilon).$$

Statistical Learning

This frequentist approach is unable to make sense of the probability of a ”single singular event”, as noted by von Mises (1928, 1939).

“*When we speak of the ‘probability of death’, the exact meaning of this expression can be defined in the following way only. We must not think of an individual, but of a certain class as a whole, e.g., ‘all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations’. A probability of death is attached to the class of men or to another class that can be defined in a similar way. We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase ‘probability of death’, when it refers to a single person, has no meaning for us at all.*”

Statistical Learning

Definition 3.1: Loss ℓ

A loss function ℓ is a function defined on $\mathcal{Y} \times \mathcal{Y}$ such that $\ell(y, y') \geq 0$ and $\ell(y, y) = 0$.

Definition 3.2: Risk $\mathcal{R}$

For a fitted model $\widehat{m}$, its risk is

$$\mathcal{R}(\widehat{m}) = \mathbb{E}_{\mathbb{P}}\Big[\ell\,(Y, \widehat{m}(\boldsymbol{X}))\Big] = \int \ell(y, \widehat{m}(\boldsymbol{x}))\mathrm{d}\mathbb{P}(y, \boldsymbol{x}).$$

Statistical Learning

Definition 3.3: Empirical risk $\widehat{\mathcal{R}}_n$

Given a sample $\{(y_i, \boldsymbol{x}_i), i = 1, \cdots, n\}$, define the empirical risk

$$\widehat{\mathcal{R}}_n(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{m}(\boldsymbol{x}_i), y_i\right).$$

Following Vapnik (1991), the "empirical risk minimization principle" states that the learning algorithm $\widehat{m}^*$ is

$$\widehat{m}^* = \underset{\widehat{m}\in\mathcal{M}}{\text{argmin}}\{\widehat{\mathcal{R}}_n(\widehat{m})\}.$$

Statistical Learning

Proposition 3.2: Optimal Decision, "*Bayes decision rule*"

For each $\boldsymbol{x}$ choose the prediction $m^\star_{\boldsymbol{x}}$ that minimizes the conditional expected loss,

$$m^\star_{\boldsymbol{x}} \in \underset{z\in\mathcal{Y}}{\text{argmin}}\left\{\int \ell(y,z)\mathrm{d}\mathbb{P}_{Y|\boldsymbol{X}}(y|\boldsymbol{x})\right\}$$

It is straightforward since $\mathrm{d}\mathbb{P}_{Y,\boldsymbol{X}}(y,\boldsymbol{x}) = \mathrm{d}\mathbb{P}_{Y|\boldsymbol{X}}(y|\boldsymbol{x})\cdot\mathrm{d}\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{x})$,

$$\mathcal{R}(\widehat{m}) = \int\left[\int \ell(y,\widehat{m}(\boldsymbol{x}))\mathrm{d}\mathbb{P}_{Y|\boldsymbol{X}}(y|\boldsymbol{x})\right]\mathrm{d}\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{x}).$$

by definition, $m^\star_{\boldsymbol{x}}$ minimizes the term in blue, i.e., for any $\widehat{m}$

$$\mathcal{R}(\widehat{m}) \geq \int\left[\int \ell(y,m^\star_{\boldsymbol{x}})\mathrm{d}\mathbb{P}_{Y|\boldsymbol{X}}(y|\boldsymbol{x})\right]\mathrm{d}\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{x}) = \mathcal{R}(m^\star).$$

Statistical Learning

It is coined "Bayes decision rule" because the conditional distribution $Y|\boldsymbol{X}$ is sometimes be referred to as the "posterior" distribution of Y given data $\boldsymbol{X}$.

Definition 3.4: Misclassification loss, $\ell_{0/1}$

$$\ell_{0/1}(y,\widehat{y}) = \mathbf{1}(y \neq \widehat{y}).$$

In the case of a binary classifier, observe that

$$\begin{aligned}
\mathcal{R}(\widehat{m}) &= \mathbb{E}[\ell(\widehat{m}(\boldsymbol{X}),Y)] = \mathbb{E}\big[\mathbb{E}[\ell(\widehat{m}(\boldsymbol{X}),Y)\mid \boldsymbol{X}]\big] \\
&= \mathbb{E}\big[\ \ell(\widehat{m}(\boldsymbol{X}),1)\ \cdot\ \mathbb{P}(Y=1\mid \boldsymbol{X})\ +\ \ell(\widehat{m}(\boldsymbol{X}),0)\ \cdot\ \mathbb{P}(Y=0\mid \boldsymbol{X})\big] \\
&= \mathbb{E}\big[\mathbf{1}[\widehat{m}(\boldsymbol{X})\neq 1]\cdot\ \mu(\boldsymbol{X})\ +\ \mathbf{1}[\widehat{m}(\boldsymbol{X})\neq 0]\ \cdot\ (1-\mu(\boldsymbol{X}))\ \big] \\
&= \mathbb{E}\big[\mathbf{1}[\widehat{m}(\boldsymbol{X})\neq 1]\cdot\mu(X) + (1-\mathbf{1}[\widehat{m}(\boldsymbol{X})\neq 1])\cdot(1-\mu(\boldsymbol{X}))\big] \\
&= \mathbb{E}\big[\ \mathbf{1}[\widehat{m}(\boldsymbol{X})\neq 1]\cdot(2\mu(\boldsymbol{X})-1)\ +1-\mu(\boldsymbol{X})\big].
\end{aligned}$$

Statistical Learning

Since $\widehat{m} : \mathcal{X} \to \{0,1\}$, this expectation is minimized by choosing $\widehat{m} = m^\star$, where

$$m^\star(\boldsymbol{x}) = \mathbf{1}(\mu(\boldsymbol{x}) > 1/2) = \begin{cases} 1 \text{ if } \mu(\boldsymbol{x}) > 1/2 \\ 0 \text{ if } \mu(\boldsymbol{x}) \leq 1/2 \end{cases}$$

The optimal risk ("Bayes risk") is $\mathcal{R}(m^\star) = \inf_m \{\mathcal{R}(m)\}$.

Definition 3.5: Excess of risk of $\widehat{m}$

For any model $\widehat{m}$, the excess of risk is $\mathcal{R}(\widehat{m}) - \mathcal{R}(m^\star)$.

For a classifier

$$\mathcal{R}(\widehat{m}) - \mathcal{R}(m^\star) = \mathbb{E}\big[\big|2\mu(\boldsymbol{X}) - 1\big| \cdot \mathbf{1}(\widehat{m}(\boldsymbol{X}) \neq m^\star(\boldsymbol{X}))\big].$$

Since we do not know μ consider a classifier based on $\widehat{m}$ ......

Statistical Learning

Definition 3.6: Plug-in Estimator

Estimate $\widehat{\mu}$ and use, as a classifier, $\mathbf{1}(\widehat{\mu}(\boldsymbol{x}) > {}^1\!/_2)$.

Proposition 3.3

For any model $\widehat{\mu}$, the risk of the plug-in classifier $\widehat{m}(\boldsymbol{x}) = \mathbf{1}(\widehat{\mu}(\boldsymbol{x}) > {}^1\!/_2)$ satisfies

$$\mathcal{R}(\widehat{m}) - \mathcal{R}(m^\star) \leq 2\mathbb{E}\big|\mu(\boldsymbol{X}) - \widehat{\mu}(\boldsymbol{X})\big|.$$

Proof We have seen that

$$\mathcal{R}(\widehat{m}) - \mathcal{R}\left(m^*\right) = \mathbb{E}\big(1[\widehat{m}(\boldsymbol{X}) \neq 1] - 1\left[m^*(\boldsymbol{X}) \neq 1\right]\big) \cdot (2\mu(\boldsymbol{X}) - 1).$$

Statistical Learning

But

$$
\begin{aligned}
&\left(\mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1] - \mathbf{1}\left[m^*(\boldsymbol{X}) \neq 1\right]\right)(2\mu(\boldsymbol{X})-1)\\
&= \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]\left(1[\widehat{m}(\boldsymbol{X}) \neq 1] - 1\left[m^*(\boldsymbol{X}) \neq 1\right]\right)(2\mu(\boldsymbol{X})-1)\\
&= \begin{cases} \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right](2\mu(\boldsymbol{X})-1) & \text{if } 2\mu(\boldsymbol{X})-1>0,\\ \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right](-1)(2\mu(\boldsymbol{X})-1) & \text{if } 2\mu(\boldsymbol{X})-1\leq 0.\end{cases}
\end{aligned}
$$

(from the definition of m^*)

$$= \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]\cdot|2\mu(\boldsymbol{X})-1|,$$

$$\mathcal{R}(\widehat{m}) - \mathcal{R}\left(m^*\right) = \mathbb{E}\left(\mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]\right)\cdot 2|\mu(\boldsymbol{X}) - {}^1\!/\!_2|.$$

If $\widehat{m}(\boldsymbol{x}) \neq m^*(\boldsymbol{x})$, it means that $\widehat{\mu}(\boldsymbol{x})$ and $\mu(\boldsymbol{x})$ lie on opposite sides of ${}^1\!/\!_2$,

$$|\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})| = |\widehat{\mu}(\boldsymbol{x}) - {}^1\!/\!_2| + \underbrace{|{}^1\!/\!_2 - \mu(\boldsymbol{x})|}_{\geq 0} \geq |\widehat{\mu}(\boldsymbol{x}) - {}^1\!/\!_2|$$

Statistical Learning

i.e.

$$|\widehat{\mu}(\boldsymbol{x})-\mu(\boldsymbol{x})| \geq |\widehat{\mu}(\boldsymbol{x})-{}^1\!/\!{}_2|\cdot \mathbf{1}\left[\widehat{m}(\boldsymbol{X})\neq m^*(\boldsymbol{X})\right]$$

which is also valid when $\widehat{m}(\boldsymbol{x})=m^*(\boldsymbol{x})$, thus

$$\mathcal{R}(\widehat{m})-\mathcal{R}\left(m^*\right)=2\mathbb{E}\big(\mathbf{1}\left[\widehat{m}(\boldsymbol{X})\neq m^*(\boldsymbol{X})\right]\big)\cdot|\mu(\boldsymbol{X})-{}^1\!/\!{}_2|\leq 2\mathbb{E}\big[|\widehat{\mu}(\boldsymbol{X})-\mu(\boldsymbol{X})|\big].$$

This $\ell_{0/1}$ loss function may be difficult to directly optimize, as shown in Bartlett et al. (2006). One could consider some surrogate loss $\tilde{\ell}$ which is easier to optimize.

Statistical Learning

Definition 3.7: Elicitation, Brier (1950), Good (1952)

A statistical functional $\mathcal{I}(Y)$ is said to be elicitable if it minimizes expected loss for some loss function s, in the sense that

$$\mathcal{I}(Y) = \underset{y\in\mathbb{R}}{\text{argmin}}\big\{\mathbb{E}\big[s(Y,y)\big]\big\}$$

"*The elicitability of a risk measure means that the risk measure can be obtained by minimizing the expectation of a forecasting objective function. Elicitability is closely related to backtesting, whose objective is to evaluate the performance of a risk forecasting model. If a risk measure is elicitable, then the sample average forecasting error based on the objective function can be used for backtesting the risk measure,*" He et al. (2022).

Loss Functions

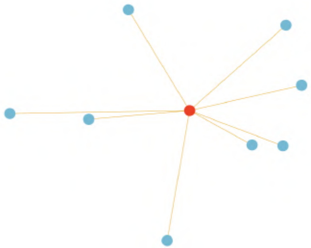
In a regression problem, a quadratic loss function ℓ_2 is used

Definition 3.8: Quadratic loss, ℓ_2

$\ell_2(y, \widehat{y}) = (y - \widehat{y})^2$, and the risk is then $\mathcal{R}_2(\widehat{m}) = \mathbb{E}\Big[(Y - \widehat{m}(\boldsymbol{X}))^2\Big]$.

Observe that

$$\mathbb{E}[Y] = \underset{m\in\mathbb{R}}{\text{argmin}}\{\mathcal{R}_2(m)\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{\mathbb{E}\Big[\ell_2(Y, m)\Big]\Big\}.$$



The expected value is “ellicitable” (for the $s = \ell_2$ loss).

The empirical risk minimizer is the ”least-square” estimate.

Loss Functions

See Huttegger (2013), explaining why the expected value is also called “best estimate”.

Up to a monotonic transformation (the square root function), the distance here is the expectation of the quadratic loss function. With the terminology of Angrist and Pischke (2009), the regression function μ is the function of $\boldsymbol{x}$ that serves as “*the best predictor of y, in the mean-squared error sense*.”

Proposition 3.4: Optimal Decision, ”*Bayes decision rule*”

For the quadratic loss ℓ_2, Bayes decision rule is the (conditional) expected value, $m^\star_{\boldsymbol{x}} = \mathbb{E}\big[Y\big|\boldsymbol{X} = \boldsymbol{x}\big] = \mu(\boldsymbol{x})$.

Loss Functions

Definition 3.9: Inner product

An inner product on $\mathcal{H}$ is the application $(f,g)\mapsto\langle f,g\rangle_{\mathcal{H}}$ (taking value in $\mathbb{R}$) bilinear, symmetric, definite positive:

- $\langle f,g\rangle_{\mathcal{H}}=\langle g,f\rangle_{\mathcal{H}}$
- $\langle \alpha f+\beta g,h\rangle_{\mathcal{H}}=\alpha\langle f,h\rangle_{\mathcal{H}}+\beta\langle g,h\rangle_{\mathcal{H}}$
- $\langle f,f\rangle_{\mathcal{H}}\geq 0$ and $\langle f,f\rangle_{\mathcal{H}}=0$ if and only if $f=0$.

Example : $\mathcal{H}=\mathbb{R}^n$, $\langle \boldsymbol{x},\boldsymbol{y}\rangle=\boldsymbol{x}^\top\boldsymbol{y}$

Example : $\mathcal{H}=\mathbb{R}^n$, let $\boldsymbol{\Sigma}$ denote some symmetric $n\times n$ positive definite matrix. Then

$$\langle \boldsymbol{x},\boldsymbol{y}\rangle_{\boldsymbol{\Sigma}}=\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\text{ is an inner product on }\mathbb{R}^n.$$

Example : $\mathcal{H}=\ell^2=\left\{u:\sum_{i=1}^\infty u_i^2<\infty\right\}$, $\langle u,v\rangle=\sum_{i=1}^\infty u_iv_i$

Loss Functions

Example : $\mathcal{H} = L^2(\mu) = \left\{f : \int f(x)^2 d\mu(x) < \infty\right\}$, $\langle f, g\rangle = \int f(x)g(x)d\mu(x)$

Example : Consider the vector space $\mathcal{V}$ that consists of all real-valued random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Given $k \in [1, \infty)$, define

$$\|X\|_k = \left[\mathbb{E}\left(|X|^k\right)\right]^{1/k}.$$

Loss Functions

A norm $\|\cdot\|$, in $\mathbb{R}^n$, satisfies

- homogeneity, $\|a\vec{\boldsymbol{u}}\| = |a| \cdot \|\vec{\boldsymbol{u}}\|, \forall a$
- triangle inequality, $\|\vec{\boldsymbol{u}} + \vec{\boldsymbol{v}}\| \leq \|\vec{\boldsymbol{u}}\| + \|\vec{\boldsymbol{v}}\|$
- positivity, $\|\vec{\boldsymbol{u}}\| \geq 0$
- definiteness, $\|\vec{\boldsymbol{u}}\| = 0 \iff \vec{\boldsymbol{u}} = \vec{\boldsymbol{0}}$

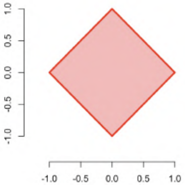
ℓ_1 norm: $\|\boldsymbol{x}\|_{\ell_1} = |x_1| + \cdots + |x_n|$,

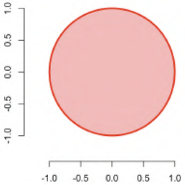
ℓ_2 norm: $\|\boldsymbol{x}\|_{\ell_2} = \sqrt{x_1^2 + \cdots + x_n^2}$,

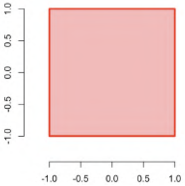
ℓ_p norm: with $p \geq 1$, $\|\boldsymbol{x}\|_{\ell_p} = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$

ℓ_∞ norm: $\|\boldsymbol{x}\|_{\ell_\infty} = \max\{x_i\}$

Unit balls ($\|\boldsymbol{x}\|_{\ell_p} \leq 1$) are convex sets







Loss Functions

Proposition 3.5: Gradient of ℓ_p norms

$$\frac{\partial}{\partial x_j}\|\boldsymbol{x}\|_{\ell_p} = \frac{1}{p}\left(\sum_i |x_i|^p\right)^{\frac{1}{p}-1} \cdot p|x_j|^{p-1}\,\text{sign}(x_j) = \left(\frac{|x_j|}{\|\boldsymbol{x}\|_{\ell_p}}\right)^{p-1}\text{sign}(x_j).$$

$$\begin{aligned}\frac{\partial}{\partial x_j}\|\boldsymbol{x}\|_{\ell_p} &= \frac{\partial}{\partial x_j}\left(\sum_{i=1}^n |x_i|^p\right)^{1/p} = \frac{1}{p}\left(\sum_{i=1}^n |x_i|^p\right)^{(1/p)-1}\frac{\partial}{\partial x_j}\left(\sum_{i=1}^n |x_i|^p\right)\\ &= \left[\left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}\right]^{1-p}\sum_{i=1}^n |x_i|^{p-1}\delta_{ij}\frac{x_i}{|x_i|} = \left(\frac{|x_j|}{\|\boldsymbol{x}\|_{\ell_p}}\right)^{p-1}\text{sign}(x_j).\end{aligned}$$

Loss Functions

Definition 3.10: Quantile loss, $\ell_{\mathbf{q},\alpha}$

The quantile loss $\ell_{\mathbf{q},\alpha}$ for some $\alpha \in (0,1)$ is

$$\ell_{\mathbf{q},\alpha}(y,\widehat{y}) = \max\left\{\alpha(y-\widehat{y}),(1-\alpha)(\widehat{y}-y)\right\} = (y-\widehat{y})(\alpha-\mathbf{1}_{(y<\widehat{y})}).$$

This loss is not symmetric $\ell_{\mathbf{q},\alpha}(y,\widehat{y}) \neq \ell_{\mathbf{q},\alpha}(\widehat{y},y)$ (if $\alpha \neq {}^1\!/_2$).

It is called "quantile" loss since

$$Q(\alpha) = F^{-1}(\alpha) \in \underset{q\in\mathbb{R}}{\text{argmin}}\left\{\mathbb{E}\Big[\ell_{\mathbf{q},\alpha}\left(Y,q\right)\Big]\right\},$$

(quantiles are also "ellicitable" functionals, elicited by
$s(y,\hat{y}) = \alpha(y-\hat{y})_+ + (1-\alpha)(y-\hat{y})_-$)

Loss Functions

Indeed, the first order condition of

$$\min_{q\in\mathbb{R}}\left\{(\alpha-1)\int_{-\infty}^{q}(y-q)dF_Y(y)+\alpha\int_{q}^{\infty}(y-q)dF_Y(y)\right\},$$

can be written, using Leibniz integral rule,

$$(1-\alpha)\int_{-\infty}^{q^\star}dF_Y(y)-\alpha\int_{q^\star}^{\infty}dF_Y(y)=0$$

i.e. $F_Y(q^\star)-\alpha=0$.

Loss Functions

Definition 3.11: Expectile loss, $\ell_{\mathsf{e},\alpha}$

The expectile loss $\ell_{\mathsf{e},\alpha}$, for some $\alpha \in (0,1)$ is

$$\ell_{\mathsf{e},\alpha}(y,\widehat{y}) = (y-\widehat{y})^2 \cdot (\alpha - \mathbf{1}_{(y<\widehat{y})})$$

$$E(\alpha) = \underset{e\in\mathbb{R}}{\text{argmin}}\left\{\mathbb{E}\left[\ell_{\mathsf{e},\alpha}\,(Y,e)\right]\right\},$$

(expectiles are elicited by $s(x,y) = \alpha(y-x)_+^2 + (1-\alpha)(y-x)_-^2$).

"*Expectiles have properties that are similar to quantiles*," Newey and Powell (1987)

Loss Functions

"*The Gaussian Hare and the Laplacian Tortoise*," Portnoy and Koenker (1997)

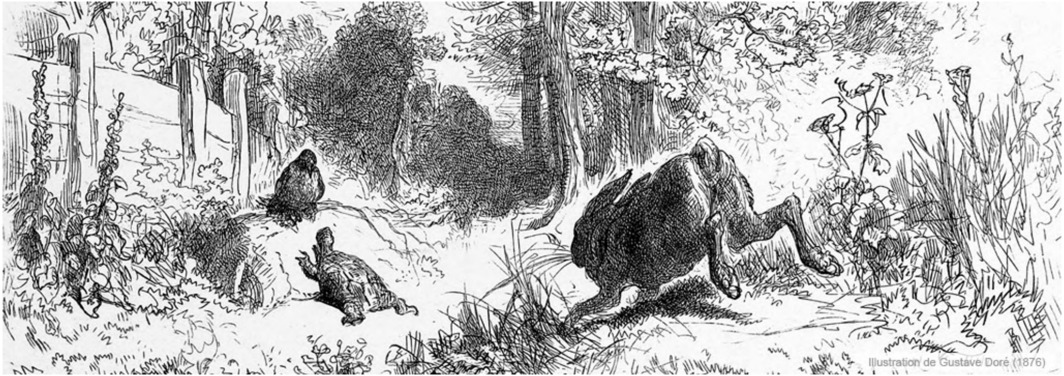


Illustration de Gustave Doré (1876)

Loss and Generalized Linear Models

In GLM, the scaled deviance ($-2\times$ the log-likelihood) of the exponential model is

$$D^\star = \sum_{i=1}^{n} d^\star(y_i, \widehat{y}_i), \text{ where } d^\star(y_i, \widehat{y}_i) = 2\big(\log \mathcal{L}_i(y_i) - \log \mathcal{L}_i(\widehat{y}_i)\big).$$

that can be related to in-sample empirical risk

$$\widehat{\mathcal{R}}_n(\widehat{m}) = \sum_{i=1}^{n} \ell\left(y_i, \widehat{m}(\boldsymbol{x}_i)\right),$$

For the Poisson distribution (with a log-link), the loss would be

$$\ell(y_i, \widehat{y}_i) = \begin{cases} 2\left(y_i \log y_i - y_i \log \widehat{y}_i - y_i + \widehat{y}_i\right) & y_i > 0 \\ 2\widehat{y}_i & y_i = 0, \end{cases}$$

while for a logistic regression, we have the standard binary cross-entropy loss

$$\ell(y_i, \widehat{y}_i) = -\big(y_i \log[\widehat{y}_i] + (1 - y_i)\log[1 - \widehat{y}_i]\big).$$

Distance Between Distributions

Definition 3.12: Distance (or metric)

A distance d on a set E is a function $E \times E \to \mathbb{R}_+$ such that

- d is symmetric, $\forall(a,b) \in E^2$, $d(a,b) = d(b,a)$,
- d is separable, $\forall(a,b) \in E^2$, $d(a,b) = 0 \Leftrightarrow a = b$,
- d satisfies $\forall(a,b,c) \in E^3$, $d(a,c) \leq d(a,b) + d(b,c)$

In a vector space, with norm $\|\cdot\|$ the induced distance is $d(x,y) = \|y - x\|$. Conversely, if

- d invariant by translation, $d(x,y) = d(x+a, y+a)$
- d is homogeneous, $d(\alpha x, \alpha y) = |\alpha| d(x,y)$

then $\|x\| = d(x,0)$ is a norm.

Distance Between Distributions

Proposition 3.6

If d is a distance on E, and if $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function such that $\psi(0) = 0$ and $\psi(t) > 0$ for all $t > 0$. If ψ if subadditive ($\psi(s+t) \leq \psi(s)+\psi(t)$), then $\delta(a, b) = \psi(d(a, b))$ is also a distance on E.

Proposition 3.7: ⚠

If d is a distance on E, then d^2 is not necessarily a distance.

Distance Between Distributions

Consider the Euclidean distance in $E = \mathbb{R}^2$, i.e. $d(\boldsymbol{z}_1, \boldsymbol{z}_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. d^2 is not a distance, see

$$\begin{cases} d^2(-\mathbf{1}, +\mathbf{1}) = 2^2 + 2^2 = 8 \\ d^2(-\mathbf{1}, \mathbf{0}) = 1^2 + 1^2 = 2 \\ d^2(\mathbf{0}, +\mathbf{1}) = 1^2 + 1^2 = 2 \end{cases}$$

i.e. d^2 does not satisfy the triangular inequality

$$d^2(-\mathbf{1}, +\mathbf{1}) > d^2(-\mathbf{1}, \mathbf{0}) + d^2(\mathbf{0}, +\mathbf{1}),$$

while

$$d(-\mathbf{1}, +\mathbf{1}) \leq d(-\mathbf{1}, \mathbf{0}) + d(\mathbf{0}, +\mathbf{1}).$$

(functions that generalize squared distance are sometimes referred to as divergences)

Distance Between Distributions

In addition to "distance", similar terms are used, including "dissimilarity", "deviance", "deviation", "discrepancy", "discrimination", and "divergence"

(... all denoted "d", or "D")

A fundamental problem in statistics and machine learning is to come up with useful measures of "distance" between pairs of probability distributions. Two desirable properties of a distance function are symmetry and the triangle inequality.

Unfortunately, many notions of "distance" between probability distributions do not satisfy these properties. Weaker notions of distance are often used, such as dissimilarity measures and divergences.

See Cha (2007) for a comprehensive list of distances...

Distance Between Distributions

Definition 3.13: Dissimilarity measure

A dissimilarity measure D on a set E is a function $E \times E \to \mathbb{R}_+$ such that D is positive and separable, i.e., $\forall(a,b) \in E^2,\ D(a,b) = 0 \Leftrightarrow a = b$,

Definition 3.14: Divergence on $\mathbb{R}^n$

A divergence D on a set $E \subset \mathbb{R}^n$ is a function $E \times E \to \mathbb{R}_+$ such that

- D is separable, $\forall(\boldsymbol{x},\boldsymbol{y}) \in E^2,\ D(\boldsymbol{x},\boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$,
- D admits development $\forall(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{\epsilon}) \in E^2,\ D(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{\epsilon}) = \dfrac{1}{2}\sum A_{i,j}(\epsilon)\epsilon_i\epsilon_j + O(|\epsilon|^3)$,
  where $A(\epsilon)$ is definite positive.

Distance Between Distributions

Definition 3.15: Scale sensitive divergence, Zolotarev (1976)

A divergence D is scale sensitive (of order $\beta > 0$) if $D(c\boldsymbol{x}, c\boldsymbol{y}) \leq |c|^\beta D(\boldsymbol{x}, \boldsymbol{y})$

Definition 3.16: Bregman Divergence, Bregman (1967)

Let $\psi : \mathcal{X} \to \mathbb{R}$ be a strictly convex function that is continuously differentiable. Then the Bregman divergence $D_\psi(\boldsymbol{x}, \boldsymbol{y})$ is defined as

$$D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) - \langle \nabla\psi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle.$$

If $\psi(\boldsymbol{x}) = \dfrac{1}{2}\|\boldsymbol{x}\|^2$ (strictly convex), then $D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \dfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$.

(recall that $\nabla\|\boldsymbol{x}\|^2 = 2\boldsymbol{x}$)

Distance Between Distributions

Proposition 3.8: Bregman Divergence

Let $\psi : \mathcal{X} \to \mathbb{R}$ be a strictly convex function that is continuously differentiable. Then Bregman divergence $D_\psi(\boldsymbol{x}, \boldsymbol{y})$ is

- strictly convex in $\boldsymbol{x}$,
- (generally) non-convex in $\boldsymbol{y}$,
- non-negative $D_\psi(\boldsymbol{x}, \boldsymbol{y}) \geq 0$,
- separable, $D_\psi(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$,
- (generally) asymmetric.

Distance Between Distributions

If $\mathcal{X}=\mathbb{R}^{\boldsymbol{n}}$, and $\psi(\boldsymbol{x})=\dfrac{1}{2}\sum_{ij}A_{ij}x_ix_j=\dfrac{1}{2}\boldsymbol{x}^\top A(\boldsymbol{x}$ for some $n\times n$ matrix A definite positive, then

$$D_\psi(\boldsymbol{x},\boldsymbol{y})=\frac{1}{2}\sum_{ij}A_{ij}(x_i-y_i)(x_j-y_j)=(\boldsymbol{x}-\boldsymbol{y})^\top A(\boldsymbol{x}-\boldsymbol{y})$$

(see Mahalanobis distance).

If $\mathcal{X}=\mathbb{R}^{\boldsymbol{n}}$, and $\psi(\boldsymbol{x})=-\sum_i\log(x_i)$ then

$$D_\psi(\boldsymbol{x},\boldsymbol{y})=\sum_i\frac{x_i}{y_i}-\log\frac{x_i}{y_i}-1$$

See Banerjee et al. (2005) for more examples.

Distance Between Distributions

We have defined norms
on $\mathbb{R}^n$, e.g.,

$$\|\boldsymbol{x}\|_{\ell_2} = \big(|x_1|^2 + \cdots + |x_n|^2\big)^{1/2} = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

that could be extended
on $\mathbb{R}$-valued random variables, e.g.,

$$\|X\|_2 = \left(\mathbb{E}\left[|X|^2\right]\right)^{1/2} = \left(\sum |x|^2 p(x)\right)^{1/2} = \left(\int |x|^2 f(x)\mathrm{d}x\right)^{1/2}$$

We can also define "distances", "dissimilarity" measures, and "divergences"
on $\mathbb{R}^n$, e.g.,

$$\ell_2(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{x},\boldsymbol{y}) = \big(|x_1-y_1|^2 + \cdots + |x_n-y_n|^2\big)^{1/2} = \left(\sum_{i=1}^{n} |x_i-y_i|^2\right)^{1/2}$$

Distance Between Distributions

that could be extended
on $\mathbb{R}$-valued random variables as components of a random vector, e.g.,

$$D(X,Y)=\left(\mathbb{E}\left[|X-Y|^2\right]\right)^{1/2}=\left(\sum|x-y|^2p(x,y)\right)^{1/2}=\left(\int|x-y|^2f(x,y)\mathrm{d}x\mathrm{d}y\right)^{1/2}$$

where p or f is the joint distribution of (X,Y), e.g., for a Gaussian vector

$$D(X,Y)=(\mu_x-\mu_y)^2+(\sigma_x-\sigma_y)^2+2\sigma_x\sigma_y(1-\rho).$$

on $\mathbb{R}$-valued random variables assuming that random variables are independent, e.g.,

$$D_\perp(X,Y)=\left(\sum|x-y|^2p_x(x)p_y(y)\right)^{1/2}=\left(\int|x-y|^2f_x(x)f_y(y)\mathrm{d}x\mathrm{d}y\right)^{1/2}$$

e.g., for two Gaussian distributions

$$D_\perp(X,Y)=(\mu_x-\mu_y)^2+\sigma_x^2+\sigma_y^2.$$

Distance Between Distributions

and one can consider some distance
on $\mathbb{R}$-valued distributions, e.g.,

$$D(\mathcal{N}(\mu_x, \sigma_x^2), \mathcal{N}(\mu_y, \sigma_y^2)) = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2.$$

In the context of "probabilistic forecasts" (as in Gneiting et al. (2007)), a "distance" on pairs $\mathbb{R} \times \mathbb{R}$-valued distributions, e.g.,

$$D(x, \mathcal{N}(\mu_y, \sigma_y^2)) = (x - \mu_y)^2 + \sigma_y^2.$$

Distance Between Distributions

Definition 3.17: Sum invariant divergence, Zolotarev (1976)

A divergence D is sum invariant if $D(X+Z, Y+Z) \leq D(X,Y)$ whenever $Z \perp\!\!\!\perp X, Y$

Example: if D is 1-scale sensitive, $D(\mathbf{1}_0, \mathbf{1}_1) \leq \dfrac{1}{2} D(\mathbf{1}_0, \mathbf{1}_2)$

Example: if D is sum invariant, $D(\mathbf{1}_0, \mathbf{1}_1) = D(\mathbf{1}_1, \mathbf{1}_2)$

See Bellemare et al. (2017a).

Distance Between Distributions

Consider sample $\{x_1, \cdots, x_n\}$ an i.i.d. sample, with empirical measure $\widehat{p}_n = \dfrac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i}$

Definition 3.18: Divergence based inference

Consider some parametric family $\mathcal{Q} = \{q_\theta, \theta \in \Theta\}$. Given a divergence D, we want to find

$$\theta^\star = \underset{\theta\in\Theta}{\text{argmin}}\{D(p, q_\theta)\}$$

unknown p

$q_\theta \in \mathcal{Q}$

or its empirical version

$$\widehat{\theta}_n = \underset{\theta\in\Theta}{\text{argmin}}\{D(\widehat{p}_n, q_\theta)\}$$

estimated $\widehat{p}_n$

Distance Between Distributions

Definition 3.19: Unbiased sample gradients, Bellemare et al. (2017a)

A divergence D has unbiased sample gradients when the expected gradient of the sample loss equals the gradient of the true loss for all p and n,

$$\mathbb{E}\left(\nabla_\theta D(\widehat{p}_n, q_\theta)\right) = \nabla_\theta D(p, q_\theta).$$

Then D is a proper scoring rule (see Gneiting and Raftery (2007)).

If this is not satisfied, stochastic gradient descent may not converge...

Distance Between Distributions

Definition 3.20: Integral probability metric, Müller (1997)

Integral probability metrics (IPMs) are distances on the space of distributions over a set $\mathcal{X}$, defined by a class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$ as

$$D_{\mathcal{F}}(p,q) = \sup_{f\in\mathcal{F}} \Big|\mathbb{E}[f(\underset{\uparrow\; X\sim p}{X})] - \mathbb{E}[f(\underset{\uparrow\; Y\sim q}{Y})]\Big|.$$

Discussed also in Dedecker and Merlevède (2007)

Distance Between Distributions

Note that it is still possible to define projections with deviance (that will not be "orthogonal" projections since divergence are not related to inner products)

Definition 3.21: Projection, Bregman (1967), Bauschke et al. (1997)

Given a strictly convex function continuously differentiable ψ and the associated Bregman divergence D_ψ, a closed closed convex $K \subset \mathcal{X}$ and a point $\boldsymbol{x} \in \mathcal{X}$. The Bregman projection of $\boldsymbol{x}$ onto K is

$$\boldsymbol{x}^\star = \underset{\boldsymbol{y}\in K}{\text{argmin}}\{D_\psi(\boldsymbol{x}, \boldsymbol{y})\}$$

If $\psi(\boldsymbol{x}) = \|\boldsymbol{x}\|_{\ell_2}^2$, Bregman projection is the standard orthogonal projection onto a convex set,

$$\boldsymbol{x}^\star = \underset{\boldsymbol{y}\in K}{\text{argmin}}\{\|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}^2\}$$

Distance Between Distributions

Definition 3.22: Hellinger distance, Hellinger (1909)

For two discrete distributions p and q, Hellinger distance is

$$d_{\mathrm{H}}(p,q)^2=\frac{1}{2}\sum_i\left(\sqrt{p(i)}-\sqrt{q(i)}\right)^2=1-\sum_i\sqrt{p(i)q(i)}\in[0,1],$$

and for absolutely continuous distributions, if p and q are densities,

$$d_{\mathrm{H}}(p,q)^2=\frac{1}{2}\int_{\mathbb{R}}\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2\mathrm{d}x \text{ or } \frac{1}{2}\int_{\mathbb{R}^k}\left(\sqrt{p(\boldsymbol{x})}-\sqrt{q(\boldsymbol{x})}\right)^2\mathrm{d}\boldsymbol{x}.$$

See Pardo (2018).

Distance Between Distributions

Proposition 3.9: Distance between Beta variables

Consider two Beta distribution, then $d^2_{\mathrm{H}}\left(\mathcal{B}(a_1, b_1), \mathcal{B}(a_2, b_2)\right)$ is

$$1 - \frac{1}{\sqrt{B(a_1, b_1)B(a_2, b_2)}} B\left(\frac{a_1 + a_2}{2}, \frac{b_1 + b_2}{2}\right)$$

Proof

$$1 - \int_0^1 \sqrt{f_1(t)f_2(t)}\mathrm{d}t = 1 - \frac{1}{\sqrt{B(a_1, b_1)B(a_2, b_2)}} \int_0^1 t^{(a_1+a_2)/2-1}(1-t)^{(b_1+b_2)/2-1}\mathrm{d}t,$$

then use $B(a, b) = \mathrm{B}(a, b) = \displaystyle\int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Distance Between Distributions

Proposition 3.10: Distance between Gaussian vectors

Consider two Gaussian distributions, then $d_{\mathrm{H}}^2\left(\mathcal{N}\left(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right),\mathcal{N}\left(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2\right)\right)$ is

$$2-2\frac{\left|\boldsymbol{\Sigma}_1\right|^{\frac{1}{4}}\left|\boldsymbol{\Sigma}_2\right|^{\frac{1}{4}}}{\left|\bar{\boldsymbol{\Sigma}}\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{8}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^\top\bar{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)\right)$$

where $\bar{\boldsymbol{\Sigma}}=\dfrac{1}{2}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)$.

Note that it is a Bregman divergence D_ψ with $\psi(\boldsymbol{x})=\sum_{i=1} x_i{}^2$

Distance Between Distributions

Definition 3.23: Pearson/Neyman χ-square divergences Nielsen and Nock (2013)

For two discrete distributions p and q, Pearson chi-square divergence is

$$d_{\mathrm{P}\chi}(p\|q)^2=\sum_i\frac{\big[p(i)-q(i)\big]^2}{q(i)},$$

while Neyman chi-square divergence is

$$d_{\mathrm{N}\chi}(p\|q)^2=\sum_i\frac{\big[(i)-q(i)\big]^2}{p(i)}=d_{\mathrm{P}\chi}(q\|p),$$

Distance Between Distributions

Note that both are Bregman divergences D_ψ with $\psi_{\mathrm{P}}(\boldsymbol{x}) = -2\sum_{i=1}\sqrt{x_i}$ and $\psi_{\mathrm{N}}(\boldsymbol{x}) = \sum_{i=1} {x_i}^{-1}$.

d_χ can be extended to the case of continuous distributions, e.g.,

$$d_{\mathrm{P}\chi}(p\|q)^2 = \int \left(\frac{p(x)}{q(x)} - 1\right)^2 p(x)\mathrm{d}x$$

Distance Between Distributions

Definition 3.24: Total Variation, Jordan (1881); Rudin (1966)

For two distributions p and q, the total variation distance between p and q is

$$d_{\mathrm{TV}}(p,q)=\sup_{\mathcal{A}}\left\{|p(\mathcal{A})-q(\mathcal{A})|\right\}.$$

Proposition 3.11: Total Variation

For two univariate distributions p and q, the total variation distance between p and q is

$$d_{\mathrm{TV}}(p,q)=\frac{1}{2}\sum_i|p(i)-q(i)|=\frac{1}{2}\|p-q\|_{\ell_1}=\sum_{i:p(i)\geq q(i)}(p(i)-q(i))$$

See Proposition 4.2 in Levin and Peres (2017).

Distance Between Distributions

Equivalently,

$$d_{\mathrm{TV}}(p,q)=\frac{1}{2}\sup_{f:\mathbb{R}^k\to\{0,1\}}\left\{\int f\mathrm{d}p-\int f\mathrm{d}q\right\}$$

(see e.g. https://djalil.chafai.net/blog/, with $f:\mathbb{R}^k\to\{-1,1\}$, $f=\mathbf{1}_{\mathcal{A}}-\mathbf{1}_{\mathcal{A}^c}$)

It is an IPM with $\mathcal{F}=\{f:\mathcal{X}\to\{0,1\}\}$, so that $\mathcal{F}$ is a set of indicator functions for any event.

For Gaussian distributions, the distance has no explicit formula, see, e.g., Devroye et al. (2018).

Distance Between Distributions

Proposition 3.12: Total Variation, Scheffé theorem, Billingsley (2017)

For two distributions p and q on $\mathbb{R}^k$,

$$d_{\mathrm{TV}}(p,q) = \frac{1}{2}\int_{\mathbb{R}^d} |p(\boldsymbol{x}) - q(\boldsymbol{x})|\mathrm{d}\boldsymbol{x},$$

$$d_{\mathrm{TV}}(p,q) = 1 - \int_{\mathbb{R}^d} \min\{p(\boldsymbol{x}) - q(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x},$$

$$d_{\mathrm{TV}}(p,q) = p(\mathcal{A}) - q(\mathcal{A}) \text{ where } \mathcal{A} = \{\boldsymbol{x} : p(\boldsymbol{x}) \geq q(\boldsymbol{x})\}.$$

Distance Between Distributions

In the univariate case, we can restrict $\mathcal{A}$ to half-lines $(-\infty, t]$

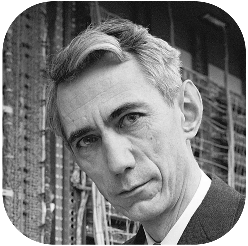
Definition 3.25: Kolmorov-Smirnov, Kolmogorov (1933); Smirnov (1948)

For two distributions p and q, Kolmorov-Smirnov distance between p and q is

$$d_{\mathrm{KS}}(p,q)=\sup_{t\in\mathbb{R}}\left\{|p((-\infty,t])-q((-\infty,t])|\right\}=\sup_{t\in\mathbb{R}}\left\{|F_p(t)-F_q(t)|\right\}=\|F_p-F_q\|_\infty,$$

where F_p and F_q are the respective cumulative distribution functions.

Distance Between Distributions



Definition 3.26: Entropy, Shannon (1948)

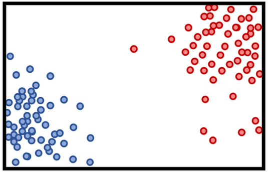
The entropy associated with distribution p is

$$\mathcal{E}_p(p) = -\sum_i p(i)\log p(i) = \mathbb{E}_p\big[-\log p(X)\big].$$

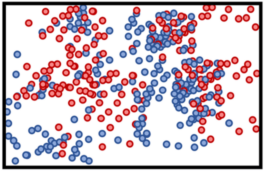
and define cross-entropy (of q relative to p) as

$$\mathcal{E}_q(p) = -\sum_i p(i)\log q(i) = \mathbb{E}_p\big[-\log q(X)\big].$$

See Amari (2016) or Chambert-Loir (2023) for more details.



Low Entropy



High Entropy

Distance Between Distributions

Definition 3.27: Kullback–Leibler, Kullback and Leibler (1951)

For two discrete distributions p and q, Kullback–Leibler divergence of p, with respect to q is

$$D_{\mathrm{KL}}(p\|q)=\sum_i p(i)\log\frac{p(i)}{q(i)},$$

and for absolutely continuous distributions,

$$D_{\mathrm{KL}}(p\|q)=\int_{\mathbb{R}} p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x \text{ or } \int_{\mathbb{R}^k} p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x},$$

in higher dimension.

Also called relative entropy, since $D_{\mathrm{KL}}(p\|q)=\mathcal{E}_q(p)-\mathcal{E}_p(p)$.

Distance Between Distributions

Proposition 3.13: KL divergence for Gaussian vectors

Consider two Gaussian distributions, then $D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\|\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))$ is

$$\frac{1}{2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)-\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}-k\right]$$

where k is the dimension, see Polyanskiy and Wu (2022).

Distance Between Distributions

The entropy of X according to p is smaller than or equal to the cross-entropy of p and q, or equivalently

Proposition 3.14: Gibbs' inequality

$D_{\text{KL}}(p\|q)$ is positive and separable, i.e. $D_{\text{KL}}(p\|q)\geq 0$ and $D_{\text{KL}}(p\|q)=0$ if and only if $p=q$.

Proof: $\displaystyle\sum_{x\in I}p(x)\log\frac{p(x)}{q(x)}\geq 0$ where I is the set of all x for which $p(x)>0$. Recall that $\log x\leq x-1$ (with equality only when $x=1$), thus $\log(1/x)\geq 1-x$, and

$$\sum_{x\in I}p(x)\ln\frac{p(x)}{q(x)}\geq\sum_{x\in I}p(x)\left(1-\frac{q(x)}{p(x)}\right)=\underbrace{\sum_{x\in I}p(x)}_{=1}-\underbrace{\sum_{x\in I}q(x)}_{\leq 1}\geq 0.$$

Distance Between Distributions

Proposition 3.15: Additivity for independence distributions

$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = D_{\mathrm{KL}}(p_x\|q_x) + D_{\mathrm{KL}}(p_y\|q_y)$ if $\boldsymbol{p}(x,y) = p_x(x)p_y(y)$ and $\boldsymbol{q}(x,y) = q_x(x)q_y(y)$.

Proof By definition

$$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = \int_{\mathcal{X}}\int_{\mathcal{Y}} p(x,y)\cdot\log\frac{p(x,y)}{q(x,y)}\,\mathrm{d}y\,\mathrm{d}x\ .$$

and since $\boldsymbol{p}(x,y) = p_x(x)p_y(y)$ and $\boldsymbol{q}(x,y) = q_x(x)q_y(y)$,

$$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = \int_{\mathcal{X}}\int_{\mathcal{Y}} p_1(x)\,p_2(y)\cdot\log\frac{p_1(x)\,p_2(y)}{q_1(x)\,q_2(y)}\,\mathrm{d}y\,\mathrm{d}x\ .$$

Distance Between Distributions

$$\begin{aligned}
D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) &= \int_{\mathcal{X}}\int_{\mathcal{Y}} p_x(x)\,p_y(y)\cdot\left(\log\frac{p_x(x)}{q_x(x)}+\log\frac{p_y(y)}{q_y(y)}\right)\,\mathrm{d}y\,\mathrm{d}x\\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} p_x(x)\,p_y(y)\cdot\log\frac{p_x(x)}{q_x(x)}\,\mathrm{d}y\,\mathrm{d}x+\int_{\mathcal{X}}\int_{\mathcal{Y}} p_x(x)\,p_y(y)\cdot\log\frac{p_y(y)}{q_y(y)}\,\mathrm{d}y\,\mathrm{d}x\\
&= \int_{\mathcal{X}} p_x(x)\cdot\log\frac{p_x(x)}{q_x(x)}\int_{\mathcal{Y}} p_y(y)\,\mathrm{d}y\,\mathrm{d}x+\int_{\mathcal{Y}} p_y(y)\cdot\log\frac{p_y(y)}{q_y(y)}\int_{\mathcal{X}} p_x(x)\,\mathrm{d}x\,\mathrm{d}y\\
&= \int_{\mathcal{X}} p_x(x)\cdot\log\frac{p_x(x)}{q_x(x)}\,\mathrm{d}x+\int_{\mathcal{Y}} p_y(y)\cdot\log\frac{p_y(y)}{q_y(y)}\,\mathrm{d}y\\
&= D_{\mathrm{KL}}(p_x\|q_x)+D_{\mathrm{KL}}(p_y\|q_y).
\end{aligned}$$

But for other distances,

$$\begin{cases} d_{\mathrm{H}}(\boldsymbol{p},\boldsymbol{q})^2 \leq d_{\mathrm{H}}(p_x,q_x)^2+d_{\mathrm{H}}(p_y,q_y)^2\\ d_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{q}) \leq d_{\mathrm{TV}}(p_x,q_x)+d_{\mathrm{TV}}(p_y,q_y).\end{cases}$$

Distance Between Distributions

It is only defined in this way if, for all x, $q(x) = 0$ implies $p(x) = 0$ (“absolute continuity” with respect to p).

Proposition 3.16

The KL divergence has unbiased sample gradients, but is not scale sensitive.

Proof Bellemare et al. (2017b).

In a Bayesian setting, $D_{\mathrm{KL}}(p\|q)$ is a measure of the information gained by revising one’s beliefs from the prior probability distribution q to the posterior probability distribution p (it is the amount of information lost when q is used to approximate p).

If $\psi(\boldsymbol{x}) = \sum x_i \log(x_i)$ (strictly convex), then Bregman divergence is

$$D_{\psi}(\boldsymbol{x},\boldsymbol{y}) = \sum x_i \log \frac{x_i}{y_i} = D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{y})$$

Distance Between Distributions

$$D_{\text{KL}}(\mathcal{B}(p)\|\mathcal{B}(q)) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

$$D_{\text{KL}}(\mathcal{B}(n,p)\|\mathcal{B}(n,q)) = np\log\frac{p}{q} + n(1-p)\log\frac{1-p}{1-q} = nD_{\text{KL}}(\mathcal{B}(p)\|\mathcal{B}(q))$$

$$D_{\text{KL}}(\mathcal{U}([a_1,b_1])\|\mathcal{U}([a_2,b_2])) = \log\frac{b_2-a_2}{b_1-a_1}$$

$$D_{\text{KL}}(\mathcal{N}(\mu_1,\sigma_1^2)\|\mathcal{N}(\mu_2,\sigma_2^2)) = \frac{1}{2}\left[\frac{(\mu_1-\mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \log\frac{\sigma_1^2}{\sigma_2^2} - 1\right]$$

$$D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\|\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) = \frac{1}{2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) + \text{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - \log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} - n\right]$$

Distance Between Distributions

Consider some distribution p_θ, as in Nielsen (2022). Using Taylor expansion,

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta+\mathrm{d}\theta}) = \frac{1}{2}\mathrm{d}\theta^\top I(\theta)\mathrm{d}\theta \approx \frac{1}{2}\mathrm{d}s_\theta^2.$$

Definition 3.28: Jeffreys (symmetric) divergence Jeffreys (1946)

The Jeffrey divergence is a symmetric divergence induced by Kullback-Liebler divergence,

$$D_{\mathrm{J}}(p_1, p_2) = \frac{1}{2}D_{\mathrm{KL}}(p_1 \| p_2) + \frac{1}{2}D_{\mathrm{KL}}(p_2 \| p_1).$$

Distance Between Distributions

Definition 3.29: Jensen-Shannon, Lin (1991)

The Jensen-Shannon divergence is a symmetric divergence induced by Kullback-Liebler divergence,

$$D_{\mathrm{JS}}(p_1,p_2)=\frac{1}{2}D_{\mathrm{KL}}(p_1\|q)+\frac{1}{2}D_{\mathrm{KL}}(p_2\|q),$$

where $q=\dfrac{1}{2}(p_1+p_2)$.

Endres and Schindelin (2003) proved that $\sqrt{D_{\mathrm{JS}}(p_1,p_2)}$ is a proper distance.

See `philentropy` package.

Distance Between Distributions

Definition 3.30: f-divergence, Rényi (1961),Ali and Silvey (1966)

Given a continuous convex function $f : [0, \infty) \to \overline{\mathbb{R}}$, define

$$D_f(p\|q) = \sum_i q(i) \cdot f\left(\frac{p(i)}{q(i)}\right)$$

and for absolutely continuous function

$$D_f(p\|q) = \int_{\mathbb{R}} q(x) f\left(\log \frac{p(x)}{q(x)}\right)\,\mathrm{d}x \text{ or } \int_{\mathbb{R}^k} q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\,\mathrm{d}\boldsymbol{x},$$

$D_f(p\|q)$ is properly defined when $p \ll q$, see also Csiszár (1964, 1967).
If $f(u) = u \log u$, $D_f(p\|q) = D_{\mathrm{KL}}(p, q)$
If $f(u) = |u - 1|$, $D_f(p\|q) = d_{\mathrm{TV}}(p, q)$

Distance Between Distributions

If $f(u)=\dfrac{1}{2}(\sqrt{u}-1)^2$, $D_f(p\|q)=d_{\mathrm{H}}(p,q)^2$

If $f(u)=\dfrac{1}{2}\left(u\log u-(u+1)\log\left(\dfrac{u+1}{2}\right)\right)$, $D_f(p\|q)=d_{\mathrm{JS}}(p,q)$

One can define $D_f(p\|q)$ when $p\not\ll q$: Since f is convex, and $f(1)=0$, the function $\dfrac{f(x)}{x-1}$ must nondecrease, so there exists $f'(\infty):=\lim_{x\to\infty}f(x)/x$, taking value in $(-\infty,+\infty]$. And since for any $p(x)>0$, we have $\lim_{q(x)\to 0}q(x)f\left(\dfrac{p(x)}{q(x)}\right)=p(x)f'(\infty)$.

Proposition 3.17

$D_f(p\|q)$ is linear in f, $D_{af+bg}(p\|q)=aD_f(p,q)+bD_g(p\|q)$.

Distance Between Distributions

Proposition 3.18

$D_f = D_g$ if and only if $f(x) = g(x) + c(x-1)$ for some $c \in \mathbb{R}$.

The only f-divergence that is also a Bregman ψ-divergence is the KL divergence

The only f-divergence that is also an integral probability metric is the total variation.

There is a variational representation of D_f, in Polyanskiy and Wu (2022).

Distance Between Distributions

Since f is convex, let $f^\star$ be the convex conjugate of f. Let $\mathrm{effdom}(f^\star)$ be the effective domain of $f^\star$ (i.e., $\mathrm{effdom}(f^*) = \{y : f^*(y) < \infty\}$)

$$D_f(p; q) = \sup_{g:\Omega\to\mathrm{effdom}(f^\star)} \mathbb{E}_p[g] - \mathbb{E}_q[f^\star \circ g]$$

For example, with the total variation, $f(x) = \dfrac{1}{2}|x-1|$, its convex conjugate is $f^*(x^*) = \begin{cases} x^* \text{ on } [-1/2, 1/2], \\ +\infty \text{ else.} \end{cases}$, and we obtain

$$d_{\mathrm{TV}}(p, q) = \sup_{|g|\leq 1/2} \mathbb{E}_p[g(X)] - \mathbb{E}_q[g(X)].$$

Distance Between Distributions

Extending Rényi entropy of order α, $\mathrm{H}_\alpha(X)=\dfrac{1}{1-\alpha}\log\left(\sum_i p(i)^\alpha\right)$, define

Definition 3.31: Rényi α-divergence, Rényi (1961)

Given $\alpha\in(0,\infty)$, define

$$D_\alpha(p\|q)=\frac{1}{\alpha-1}\log\left(\sum_i\frac{p(i)^\alpha}{q(i)^{\alpha-1}}\right)$$

and for absolutely continuous function

$$D_\alpha(p\|q)=\frac{1}{\alpha-1}\log\left(\int_{\mathbb{R}}\frac{p(x)^\alpha}{q(x)^{\alpha-1}}\mathrm{d}x\right)\text{ or }\frac{1}{\alpha-1}\log\left(\int_{\mathbb{R}^k}\frac{p(\boldsymbol{x})^\alpha}{q(\boldsymbol{x})^{\alpha-1}}\mathrm{d}\boldsymbol{x}\right).$$

Distance Between Distributions

Recall that

$$D_\alpha(p\|q) = \frac{1}{\alpha-1}\log\left(\sum_i \frac{p(i)^\alpha}{q(i)^{\alpha-1}}\right) \text{ when } \alpha \in (0,\infty).$$

One can define limiting cases, $D_0(P\|Q) = -\log Q(\{i : p_i > 0\})$ and $D_\infty(P\|Q) = \log \sup_i \dfrac{p_i}{q_i}$

Observe also that $D_1(p\|q) = D_{\mathrm{KL}}(p\|q)$

Distance Between Distributions

Definition 3.32: Cramér, Cramér (1928a,b) and Székely (2003)

Consider two measures on p and q on $\mathbb{R}$. Then define Cramér distance

$$C_k(p,q) = \left(\int_{-\infty}^{\infty} |F_p(x) - F_q(x)|^k \mathrm{d}x\right)^{1/k}, \text{ for } k \geq 1$$

C_2 is named "energy-distance" in Székely (2003) and Rizzo and Székely (2016), and "continuous ranked probability score" in Gneiting et al. (2007).

It is an Integral Probability Metrics (IPM), since

$$C_{\boxed{k}}(p,q) = \sup_{f\in\boxed{\mathcal{F}_{k'}}} \left|\mathbb{E}[f(\underset{X\sim p}{\boxed{X}})] - \mathbb{E}[f(\underset{Y\sim q}{\boxed{Y}})]\right|.$$

$k^{-1}+k'^{-1}=1$

where $\mathcal{F}_{k'}$ is the set of absolutely continuous functions such that $\|\nabla f\|_{k'} \leq 1$.

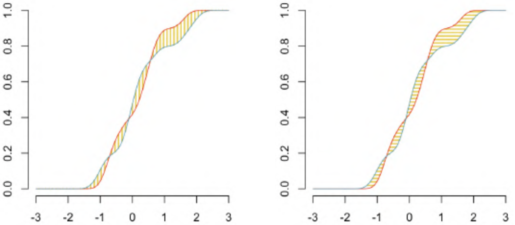
For example, if $k=1$, $\|\nabla f\|_\infty \leq 1$ (corresponding to 1-Lipschitz functions).

Distance Between Distributions

Definition 3.33: Wasserstein, Wasserstein (1969)

Consider two measures on p and q on $\mathbb{R}$. Then define Wasserstein distance

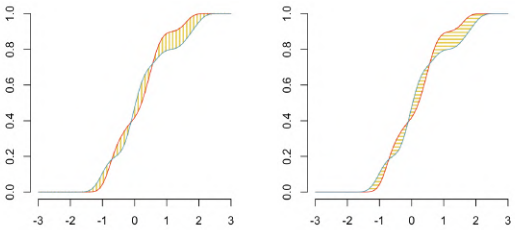
$$W_k(p,q)=\left(\int_0^1 |F_p^{-1}(u)-F_q^{-1}(u)|^k \mathrm{d}u\right)^{1/k},\text{ for } k\geq 1$$



```
> c2 = function(x) (pnorm(x,0,1)-pnorm(x,1,2))^2
> w2 = function(u) (qnorm(u,0,1)-qnorm(u,1,2))^2
> sqrt(integrate(c2,-Inf,Inf)$value)
[1] 0.5167714
> sqrt(integrate(w2,0,1)$value)
[1] 1.414214
```

where F^{-1} denotes the generalized inverse of F, $F^{-1}(u)=\inf_{x\in\mathbb{R}}\{F(x)\geq u\}$.

Distance Between Distributions



```
> c1 = function(x) abs(pnorm(x,0,1)-
    pnorm(x,1,2))
> w1 = function(x) abs(qnorm(x,0,1)-
    qnorm(x,1,2))
> integrate(c1,-Inf,Inf)$value
[1] 1.166631
> integrate(w1,0,1)$value
[1] 1.166636
```

Proposition 3.19: C_1 and W_1

Consider two measures on p and q on $\mathbb{R}$.

$$W_1(p,q)=\int_0^1 |F_p^{-1}(u)-F_q^{-1}(u)|\mathrm{d}u=\int_{-\infty}^{\infty}|F_p(x)-F_q(x)|\mathrm{d}x=C_1(p,q).$$

Proof See Prokhorov (1956), Dall'Aglio (1956) and Vallender (1974).

Distance Between Distributions

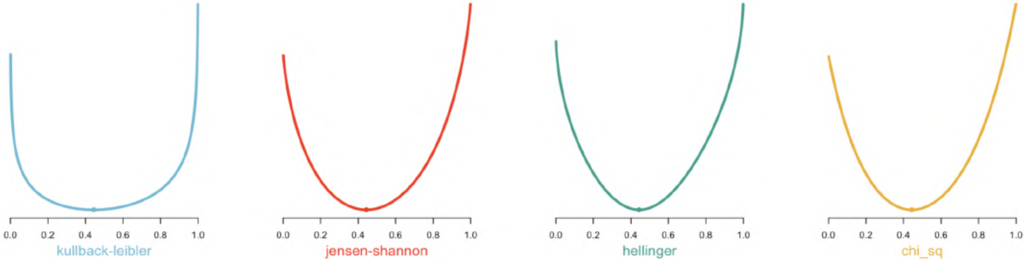
Instead of the geometric proof (see plot above), observe that

$$\begin{aligned}
\int_0^1 |F_p^{-1}(u)-F_q^{-1}(u)|\mathrm{d}u &= \int_0^1\int_{-\infty}^{\infty} g(u,x)\mathrm{d}x\mathrm{d}u,\ \ g(u,x)=1 \text{ if } \begin{cases} x\in[F_p^{-1}(u),F_q^{-1}(u)] \\ x\in[F_q^{-1}(u),F_p^{-1}(u)] \end{cases} \\
&= \int_{-\infty}^{\infty}\int_0^1 h(u,x)\mathrm{d}u\mathrm{d}x,\ \ h(u,x)=1 \text{ if } \begin{cases} u\in[F_p(x),F_q(x)] \\ u\in[F_q(x),F_p(x)] \end{cases} \\
&= \int_{-\infty}^{\infty} |F_p(x)-F_q(x)|\mathrm{d}x
\end{aligned}$$

(see Proposition 2.17 in Santambrogio (2015) for a proper justification)

Distance Between Distributions

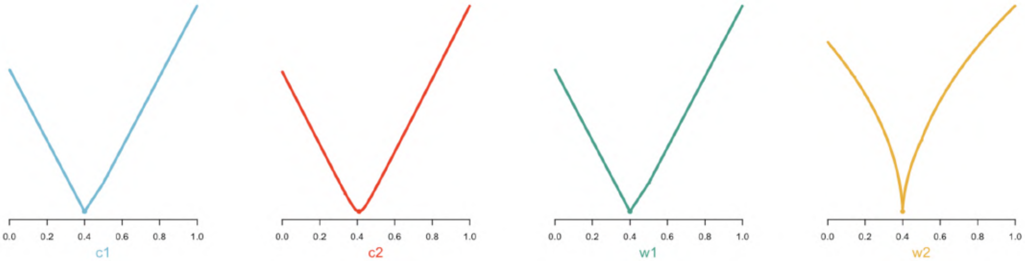
μ: multinomial distribution on $\{0, 1, 10\}$, with $\boldsymbol{p} = (.5, .1, .4)$
ν_θ: binomial type distribution on $\{0, 10\}$, with $\boldsymbol{q}_\theta = (1 - \theta, \theta)$
Let $\theta^\star = \text{argmin}\{d(p, q_\theta)\}$ or $\theta^\star = \text{argmin}\{d(p \| q_\theta)\}$



with $d_{\text{KL}}(p \| q_\theta)$, $d_{\text{JS}}(p, q_\theta)$, $d_{\text{H}}(p, q_\theta)$ and $d_{\text{H}_{\chi^2}}(p \| q_\theta)$.

Distance Between Distributions

μ: multinomial distribution on $\{0,1,10\}$, with $\boldsymbol{p}=(.5,.1,.4)$
ν_θ: binomial type distribution on $\{0,10\}$, with $\boldsymbol{q}_\theta=(1-\theta,\theta)$
Let $\theta^\star=\text{argmin}\{d(p,q_\theta)\}$



with $C_1(p,q_\theta)$, $C_2(p,q_\theta)$, $W_1(p,q_\theta)$ and $W_2(p,q_\theta)$.

Distance Between Distributions

Proposition 3.20

The Wasserstein metric is scale and sum invariant, but does not have unbiased sample gradients.

Proof Bellemare et al. (2017b)

Example If x_i are drawn from a Bernoulli distribution

Non-vanishing minimax bias: $\forall n,\ \exists p, q_\theta,\ \left|\mathbb{E}(\nabla_\theta W_k^k(\widehat{p}_n, q_\theta)) - \nabla_\theta W_k^k(p, q_\theta)\right| \geq 2e^{-2}$

Wrong minimum: in general,

$$\widehat{\theta}_n = \text{argmin}\left\{\mathbb{E}(\left(W_k^k(\widehat{p}_n, q_\theta)\right))\right\} \neq \text{argmin}\left\{W_k^k(\mathbb{P}, \mathbb{Q}_\theta)))\right\} = \theta$$

Distance Between Distributions

Proposition 3.21

The Cramér metric is scale and sum invariant.

$C_k(X+Z, Y+Z) \leq C_k(X, Y)$ whenever $Z \perp\!\!\!\perp X, Y$ and $k \geq 1$, and $C_k(cX, cY) \leq |c|^{1/k} C_k(X, Y)$.

Proposition 3.22

C_2 has unbiased sample gradients (only $k=2$),

$$\mathbb{E}\left(\nabla_\theta C_2(\widehat{p}_n, q_\theta)\right) = \nabla_\theta C_2(p, q_\theta).$$

Distance Between Distributions

Consider first W_1 (earth mover's distance), which was the only distance discussed in Wasserstein (1969). See also Vallender (1974) for an extensive review.

W_1 is an IPM where $\mathcal{F}$ the set of 1-Lipschitz functions, Kantorovich and Rubinstein (1958), i.e., if p and q have bounded support,

$$W_1(p,q) = \sup_{f\in\mathcal{F}} \left\{ \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}(p-q)(x) \right\},$$

$\mathcal{F}$ being the class of 1-Lipschitz functions

Proposition 3.23: W_1 and First Order Dominance

Suppose that $X_1 \preceq X_2$ (first order dominance, $F_2^{-1}(u) \geq F_1^{-1}(u)$, $\forall u \in (0,1)$),

$$W_1(p_1,p_2) = \mathbb{E}[X_2] - \mathbb{E}[X_1].$$

Distance Between Distributions

Proof

$$W_1(p_1,p_2)=\int_0^1 \Big|\overset{\geq 0}{F_2^{-1}(u)-F_1^{-1}(u)}\Big|\mathrm{d}u=\overset{\mathbb{E}[X_2]}{\int_0^1 F_2^{-1}(u)\mathrm{d}u}-\overset{\mathbb{E}[X_1]}{\int_0^1 F_1^{-1}(u)\mathrm{d}u}$$

then (property discussed later)

$$W_1(p_1,p_2)=\inf_C\int\int|x_2-x_1|\mathrm{d}C(F_1(x_1),F_2(x_2))=\inf_C \overset{\mathbb{E}[|X_1-X_2|]}{\int\int|F_2^{-1}(v)-F_1^{-1}(u)|\mathrm{d}C(u,v)}$$

As discussed in Vallender (1974),

$$\begin{aligned}\mathbb{E}[|X_1-X_2|] &= \int\big[\mathbb{P}[X_1<t,X_2\geq t]+\mathbb{P}[X_1\geq t,X_2<t]\big]\mathrm{d}t\\ &= \int\big[\mathbb{P}[X_1<t]+\mathbb{P}[X_2<t]-2\mathbb{P}[X_1<t,X_2<t]\big]\mathrm{d}t\end{aligned}$$

Distance Between Distributions

$$\mathbb{E}[|X_1 - X_2|] = [F_1(t) + F_2(t) - 2C(F_1(t), F_2(t))]\mathrm{d}t$$

From Fréchet-Hoeffding bounds, $C(u, v) \leq M(u, v) = \min\{u, v\}$ and

$$F_1(t) + F_2(t) - 2C(F_1(t), F_2(t)) \geq F_1(t) + F_2(t) - 2M(F_1(t), F_2(t))$$

$$\mathbb{E}[|X_1 - X_2|] \geq \underbrace{\boxed{\int\int |F_2^{-1}(v) - F_1^{-1}(u)|\mathrm{d}M(u, v)}}_{\int |F_2^{-1}(u) - F_1^{-1}(u)|\mathrm{d}u}$$

Example let $p_1 \leq p_2$

$$W_1(\mathcal{B}(p_1), \mathcal{B}(p_2)) = p_2 - p_1.$$

Distance Between Distributions

We can also consider W_2

Proposition 3.24: C_2 and W_2

Consider two measures on p and q on $\mathbb{R}$.

$$W_2(p,q)^2=\int_0^1|F_p^{-1}(u)-F_q^{-1}(u)|^2\mathrm{d}u \text{ while } C_2(p,q)=\int_{-\infty}^{\infty}|F_p(x)-F_q(x)|^2\mathrm{d}x.$$

Distance Between Distributions

Proposition 3.25: W_2 for Gaussian / Bernoulli distributions

Consider two Gaussian distributions, then

$$W_2\big(\mathcal{N}(\mu_1,\sigma_1^2),\mathcal{N}(\mu_2,\boldsymbol{\sigma}_2^2)\big)^2=(\mu_1-\mu_2)^2+(\sigma_1-\sigma_2)^2,$$

and for two Bernoulli distributions, if $p_1 \leq p_2$

$$W_2\big(\mathcal{B}(p_1),\mathcal{B}(p_2)\big)=\sqrt{p_2-p_1}.$$

Distance Between Distributions

Proposition 3.26: Representation for W_2

Consider two measures on p and q on $\mathbb{R}$.

$$W_2(p,q)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(F_p(\min\{x,y\}) - F_q(\max\{x,y\})\right)_+ + \left(F_q(\min\{x,y\}) - F_p(\max\{x,y\})\right)_+ \mathrm{d}x\mathrm{d}y$$

or

$$W_2(p,q)^2 = 2\int_{-\infty}^{\infty}\int_{x}^{\infty} \left[\left(F_p(x) - F_q(y)\right)_+ + \left(F_q(x) - F_p(y)\right)_+\right]\mathrm{d}x\mathrm{d}y$$

Proof Since $W_2(p,q)^2 = \int_0^1 |F_p^{-1}(u) - F_q^{-1}(u)|^2\mathrm{d}u$ observe that

$F_p^{-1}(u) - F_q^{-1}(u) = F_p^{-1}(u) - F_p^{-1}(F_p(F_q^{-1}(u))) = F_p^{-1}(u) - F_p^{-1}(G(u))$ where $G = F_p \circ F_q^{-1}$.

Distance Between Distributions

Since F_q is continuously differentiable, so that $H = F_p' \circ F_p^{-1}$, then

$$F_p^{-1}(u) - F_q^{-1}(u) = \int_{G(u)}^{u} \frac{\mathrm{d}t}{H(t)}$$

and write

$$\big(F_p^{-1}(u) - F_q^{-1}(u)\big)^2 = \int_{G(u)}^{u} \frac{\mathrm{d}t}{H(t)} \frac{\mathrm{d}v}{H(v)}$$

and depending on whether $G(u) \leq u$ or $u \leq G(u)$, we can write

$$\int_0^1 \big(F_p^{-1}(u) - F_q^{-1}(u)\big)^2 \mathrm{d}u = \int_0^1 \int_0^1 \big(G^{-1}(\min\{t,v\}) - \max\{r,v\}\big)_+ + \big(\min\{t,v\} - G^{-1}(\max\{t,v\})\big)_+ \mathrm{d}t\mathrm{d}v.$$

And finally, let $t = F_p(x)$ and $v = F_q(v)$, so that $G^{-1}(t) = F_q(x)$ and $G^{-1}(v) = F_q(y)$, and we get the desired expression.

Distance Between Distributions

We can finally consider W_∞

Proposition 3.27: W_∞

Consider two measures on p and q on $\mathbb{R}$.

$$W_\infty(p,q) = \sup_{u\in(0,1)} \left|F_p^{-1}(u) - F_q^{-1}(u)\right|.$$

Furthermore, $W_\infty(p,q)$ is the infimum over all $h \geq 0$ such that

$$F_q(x-h) \leq F_p(x) \leq F_q(x+h), \text{ for all } x \in \mathbb{R}.$$

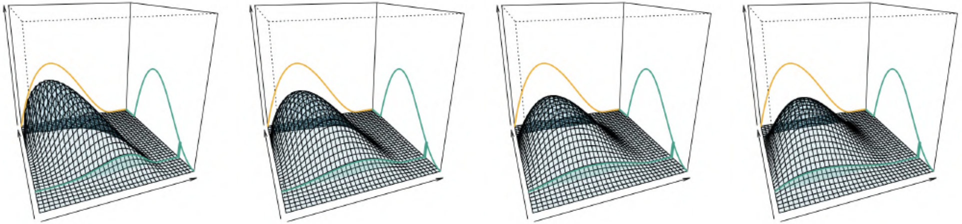
Optimal transport and Wasserstein distance

Definition 3.34: Wasserstein, Wasserstein (1969)

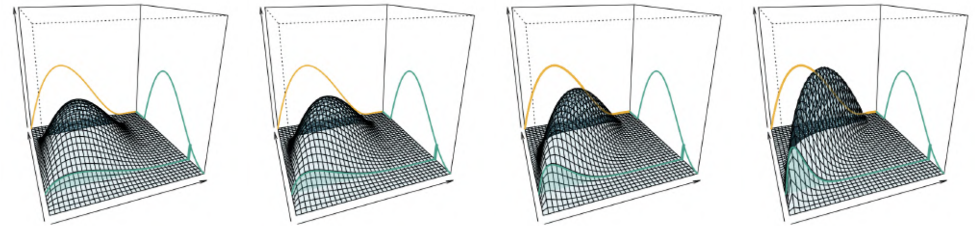
Consider two measures on p and q on $\mathbb{R}^k$, with a norm $\|\cdot\|$ (on $\mathbb{R}^k$). Then define

$$W_k(p,q) = \left(\inf_{\pi \in \Pi(p,q)} \int_{\mathbb{R}^k \times \mathbb{R}^k} \|\boldsymbol{x} - \boldsymbol{y}\|^k \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) \right)^{1/k},$$

where $\Pi(p,q)$ is the set of all couplings of p and q.



Optimal transport and Wasserstein distance



Definition 3.35: Kantorovich Problem

Kantorovich Problem is defined as

$$W_c(p,q)=\inf_{\pi\in\Pi(p,q)}\int_{\mathcal{X}\times\mathcal{Y}}c(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}),$$

for cost function c (or loss function).

Optimal transport and Monge mapping

Definition 3.36: Push-Forward and Transport Map

Given two metric spaces $\mathcal{X}$ and $\mathcal{Y}$, a measurable map $T : \mathcal{X} \to \mathcal{Y}$ and a measure μ on $\mathcal{X}$. The push-forward of μ by T is the measure $\nu = T_{\#}\mu$ on $\mathcal{Y}$ defined by

$$\forall \mathcal{B} \subset \mathcal{Y},\ \ T_{\#}\mu(B) = \mu\big(T^{-1}(\mathcal{B})\big).$$

By the change-of-variable formula

Proposition 3.28: Push-Forward and Transport Map

For all measurable and bounded $\varphi : \mathcal{Y} \to \mathbb{R}$,

$$\int_{\mathcal{Y}} \varphi(y) \mathrm{d} T_{\#}\mu(y) = \int_{\mathcal{X}} \varphi(T(x)) \mathrm{d}\mu(x).$$

Optimal transport and Monge mapping

If $\mathcal{Y}$ is a finite set $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$,

$$T_{\#}\mu = \sum_{i=1}^{n} \mu\big(T^{-1}(\{\boldsymbol{y}_i\})\big) \cdot \delta_{\{\boldsymbol{y}_i\}}$$

If $\mathcal{X}$ is a single atom, $\{\boldsymbol{x}\}$, $\mu = \delta_{\boldsymbol{x}}$ and $T_{\#}\mu(\mathcal{B}) = \mu\big(T^{-1}(\mathcal{B})\big) = \delta_{T(\boldsymbol{x})}$. If Card(support($\nu$)) > 1, there is no transport map.

One solution is to allow mass to split, leading to Kantorovich's relaxation of Monge's problem

Proposition 3.29: Existence of a map

If $\mathcal{X} = \mathcal{Y}$ is a compact subset of $\mathbb{R}^k$, if μ and ν are two measures, and if μ is atomless, then there exists T such that $\nu = T_{\#}\mu$.

see Santambrogio (2015).

Optimal transport and Monge mapping

If $\mathcal{X}$ and $\mathcal{Y}$ are two sets of $\mathbb{R}^k$, and if measures μ and ν are absolutely continuous, with densities f and g (w.r.t. Lebesgue measure),

$$\int_{\mathcal{Y}} \varphi(\boldsymbol{y})g(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \int_{\mathcal{X}} \varphi(T(\boldsymbol{x})) \cdot \underbrace{g(T(\boldsymbol{x}))\det \nabla T(\boldsymbol{x})}_{=f(\boldsymbol{x})} \cdot\mathrm{d}\boldsymbol{x}.$$

Definition 3.37: Monge Problem

Monge problem

$$\inf_{T_{\#}p=q} \int_{\mathcal{X}} c(\boldsymbol{x}, T(\boldsymbol{x}))\mathrm{d}\mathbb{P}_{\mathtt{A}}(\boldsymbol{x}),$$

for cost function c.

Note that the constraint and the objective function are non-convex.

Optimal transport and Monge mapping

Theorem 3.1: Optimal map for continuous univariate distributions

The optimal Monge map $T^\star$ for some strictly convex cost c such that $T^\star_\#\mathbb{P}_{\mathtt{A}} = \mathbb{P}_{\mathtt{B}}$ is $T^\star = F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}$.

$T^\star$ is an increasing mapping.

Example Univariate Gaussian

$$x_{\mathtt{B}} = \mathcal{T}^\star(x_{\mathtt{A}}) = \mu_{\mathtt{B}} + \sigma_{\mathtt{B}}\sigma_{\mathtt{A}}^{-1}(x_{\mathtt{A}} - \mu_{\mathtt{A}}).$$

Optimal transport and Monge mapping

Theorem 3.2: Optimal map for continuous multivariate distributions, Brenier (1991)

With a quadratic cost, the optimal Monge map $T^\star$ is unique, and it is the gradient of a convex function, $T^\star = \nabla\varphi$.

Example Multidimensional Gaussian

$$\boldsymbol{x}_{\texttt{B}} = \mathcal{T}^\star(\boldsymbol{x}_{\texttt{A}}) = \boldsymbol{\mu}_{\texttt{B}} + \boldsymbol{A}(\boldsymbol{x}_{\texttt{A}} - \boldsymbol{\mu}_{\texttt{A}}),$$

where $\boldsymbol{A}$ is a symmetric positive matrix that satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_{\texttt{A}}\boldsymbol{A} = \boldsymbol{\Sigma}_{\texttt{B}}$, which has a unique solution given by $\boldsymbol{A} = \boldsymbol{\Sigma}_{\texttt{A}}^{-1/2}\big(\boldsymbol{\Sigma}_{\texttt{A}}^{1/2}\boldsymbol{\Sigma}_{\texttt{B}}\boldsymbol{\Sigma}_{\texttt{A}}^{1/2}\big)^{1/2}\boldsymbol{\Sigma}_{\texttt{A}}^{-1/2}$, where $\boldsymbol{M}^{1/2}$ is the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008).

Optimal transport and Monge mapping

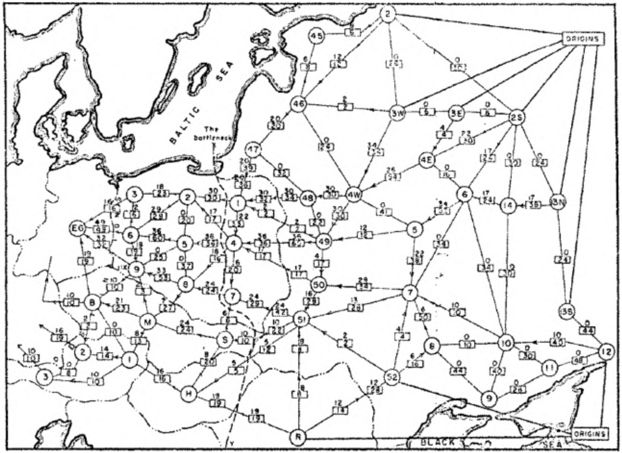
Gangbo (1999) proved, when $\mathcal{X} = \mathcal{Y}$ is a compact subset of $\mathbb{R}$, the infimum in Monge problem and the minimum in Kantorovich problem coincide, if μ is atomless,

Proposition 3.30: Monge/Kantorovich Problems

$\mathcal{X} = \mathcal{Y}$ is a compact subset of $\mathbb{R}^k$ and if μ is atomless,

min{Monge problem, see Def. 3.37} $=$ min{Kantorovich problem, see Def. 3.35}.

Optimal transport (discrete)



(via Harris and Ross (1955))

Optimal transport (discrete)

One can consider optimal transport for empirical measures, $\mathbb{P}=\sum_{i=1}^{n}\omega_i\delta_{\boldsymbol{x}_i}$.

With uniform weights and n points for $\mathbb{P}_{\mathtt{A}}$ and $\mathbb{P}_{\mathtt{B}}$, W_k^k is the optimal matching cost (Hungarian algorithm, Kuhn (1955, 1956)), cast as a linear program

$$W_k(\mathbb{P}_{\mathtt{A}},\mathbb{P}_{\mathtt{B}})=\left(\min_{s\in\mathcal{S}_n}\frac{1}{n}\sum_{i=1}^{n}d(x_i,y_{s(i)})^k\right)^{1/k},$$

where $\mathcal{S}_n$ is the set of permutations on $\{1,2,\cdots,n\}$.

Optimal transport (discrete)

Consider the set of $n \times n$ doubly-stochastic matrices,

$$D_n = \{M \in \mathbb{R}_+^{n\times n} : M\mathbf{1}_n = \mathbf{1}_n \text{ and } M^\top \mathbf{1}_n = \mathbf{1}_n\},$$

and the subset of permutation matrices,

$$U_n = \{M \in \{0,1\}^{n\times n} : M\mathbf{1}_n = \mathbf{1}_n \text{ and } M^\top \mathbf{1}_n = \mathbf{1}_n\}.$$

Let C denote the cost matrix, $C_{i,j} = d(x_i, y_j)^k$, then

$$W_k(\boldsymbol{x}, \boldsymbol{y})^k = \underset{P\in U_n}{\text{argmin}}\Big\{\langle P, C\rangle\Big\}, \text{ where } \langle P, C\rangle = \sum_{i=1}^n \sum_{j=1}^n P_{i,j} C_{i,j} \qquad (1)$$

and “optimal transport” permutation matrix

$$P^* \in \underset{P\in U_n}{\text{argmin}}\Big\{\langle P, C\rangle\Big\} \qquad (2)$$

Optimal transport (discrete)

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

	7	8	9	10	11	12
1	·	·	·	·	1	·
2	·	1	·	·	·	·
3	·	·	1	·	·	·
4	1	·	·	·	·	·
5	·	·	·	1	·	·
6	·	·	·	·	·	1

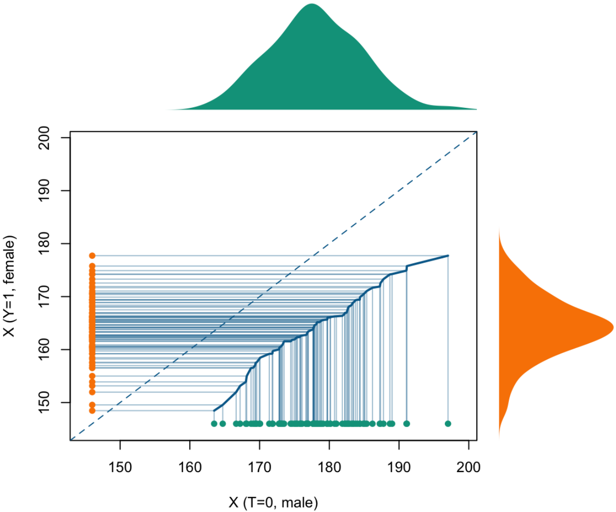
1 ↔ 11
2 ↔ 8
3 ↔ 9
4 ↔ 7
5 ↔ 10
6 ↔ 12

Optimal transport (discrete)

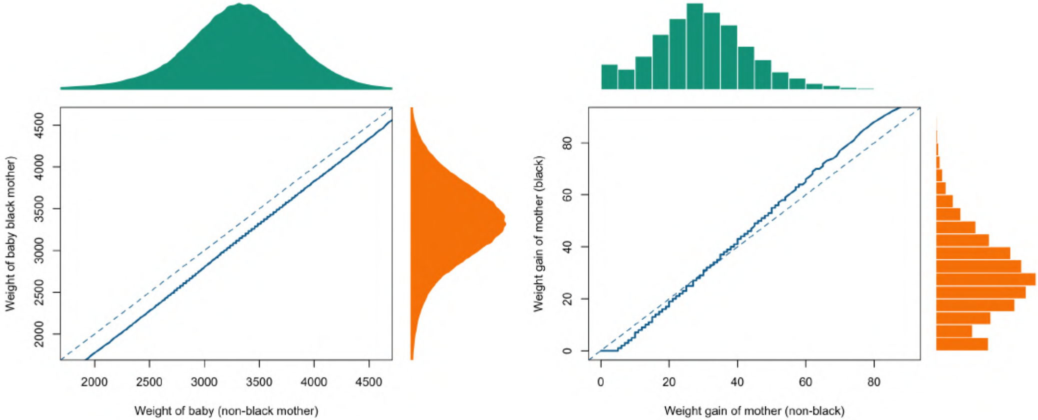
Consider wo samples, with the height of men and women (both groups of size n).

On the following graph, we can visualize the optimal matching of individuals in the two groups.

It is a monotone mapping.



Optimal transport (discrete)



Two groups, with black and non-black mothers, delivering babies (in the U.S.)

$x_1 \leftrightarrow x_1$ (newborn weight) and $x_2 \leftrightarrow x_2$ (weight gain of the mother)

Optimal transport (discrete)

Proposition 3.31: Hardy–Littlewood–Pólya inequality, Hardy et al. (1952)

Given $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$ n pairs of ordered real numbers, for every permutation σ of $\{1, 2, \cdots, n\}$,

$$\sum_{i=1}^{n} x_i y_{n+1-i} \leq \sum_{i=1}^{n} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n} x_i y_i .$$

various implications, e.g. bounds on the covariance, and the correlation, see Proposition 5.1.

This can be extended to more general function $\Phi(x_i, y_j)$.

Optimal transport (discrete)

Definition 3.38: Supermodular, Topkis (1998)

Function $\Phi : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ is supermodular if for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^k$,

$$\Phi(\boldsymbol{z} \wedge \boldsymbol{z}') + \Phi(\boldsymbol{z} \vee \boldsymbol{z}') \geq \Phi(\boldsymbol{z}) + \Phi(\boldsymbol{z}'),$$

where $\boldsymbol{z} \wedge \boldsymbol{z}'$ and $\boldsymbol{z} \vee \boldsymbol{z}'$ denote respectively the maximum and the minimum componentwise. If $-\Phi$ is supermodular, Φ is said to be submodular.

Optimal transport (discrete)

Proposition 3.32: Hardy–Littlewood–Pólya inequality, Hardy et al. (1952)

Given $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$ n pairs of ordered real numbers, and some supermodular function $\Phi : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$, for every permutation σ of $\{1, 2, \cdots, n\}$,

$$\sum_{i=1}^n \Phi(x_i, y_{n+1-i}) \leq \sum_{i=1}^n \Phi(x_i, y_{\sigma(i)}) \leq \sum_{i=1}^n \Phi(x_i, y_i),$$

while if $\Phi : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is submodular,

$$\sum_{i=1}^n \Phi(x_i, y_i) \leq \sum_{i=1}^n \Phi(x_i, y_{\sigma(i)}) \leq \sum_{i=1}^n \Phi(x_i, y_{n+1-i}).$$

Functions $\Phi(x, y) = \gamma(x - y)$ for some concave function $\gamma : \mathbb{R}\to\mathbb{R}$, such as $\Phi(x, y) = -|x - y|^k$ with $k \geq 1$, are supermodular.

Optimal transport (discrete)

```
> permutations = function(n){
+   if(n==1){
+     return(matrix(1))
+   } else {
+   sp = permutations(n-1)
+   p = nrow(sp)
+   A = matrix(nrow=n*p,ncol=n)
+   for(i in 1:n){
+       A[(i-1)*p+1:p,] =
+           cbind(i,sp+(sp>=i))
+         }
+   return(A)
+     }
+ }
```

$\Phi(x,y) = (x-y)^2$, submodular function,

Consider $x_1 \leq \cdots \leq x_n$

```
> Phi = function(x,y) sum((x-y)^2)
> set.seed(1)
> x = sort(x)
> y = y[1:6]
> vect = permutations(6)
> MY = matrix(vect, ncol=6)
> MPhi = function(i) Phi(x, y[MY[i,]])
> S = Vectorize(MPhi)(1:nrow(MY))
> y[MY[which.min(S),]]
[1] 0.046 0.288 0.409 0.788 0.883 0.940
```

Optimal transport (discrete)

In a very general setting (with $n_{\mathtt{A}} \neq n_{\mathtt{B}}$), if $\boldsymbol{a}_{\mathtt{A}} \in \mathbb{R}_+^{n_{\mathtt{A}}}$ and $\boldsymbol{a}_{\mathtt{B}} \in \mathbb{R}_+^{n_{\mathtt{B}}}$ satisfy $\boldsymbol{a}_{\mathtt{A}}^\top 1_{n_{\mathtt{A}}} = \boldsymbol{a}_{\mathtt{B}}^\top 1_{n_{\mathtt{B}}}$ (identical sums), define

$$U(\boldsymbol{a}_{\mathtt{A}}, \boldsymbol{a}_{\mathtt{B}}) = \left\{M \in \mathbb{R}_+^{n_{\mathtt{A}} \times n_{\mathtt{B}}} : M\mathbf{1}_{n_{\mathtt{B}}} = \boldsymbol{a}_{\mathtt{A}} \text{ and } M^\top \mathbf{1}_{n_{\mathtt{A}}} = \boldsymbol{a}_{\mathtt{B}}\right\}.$$

This set of matrices is a convex transportation polytope (see Brualdi (2006)).

In our case, let $U_{n_{\mathtt{A}},n_{\mathtt{B}}}$ denote $U\left(\mathbf{1}_{n_{\mathtt{A}}}, \dfrac{n_{\mathtt{A}}}{n_{\mathtt{B}}}\mathbf{1}_{n_{\mathtt{B}}}\right)$ ($U_{n,n}$ is the set of permutation matrices associated with $\mathcal{S}_n$). Let C denote the cost matrix, $C_{i,j} = d(x_i, y_j)^k$.

$$W_k(\boldsymbol{x}, \boldsymbol{y})^k = \underset{P \in U_{n_{\mathtt{A}},n_{\mathtt{B}}}}{\text{argmin}} \left\{\langle P, C\rangle\right\}, \text{ where } \langle P, C\rangle = \sum_{i=1}^{n_{\mathtt{A}}} \sum_{j=1}^{n_{\mathtt{B}}} P_{i,j} C_{i,j} \qquad (3)$$

and “optimal transport”

$$P^* \in \underset{P \in U_{n_A,n_B}}{\text{argmin}} \left\{\langle P, C\rangle\right\} \qquad (4)$$

Optimal transport (discrete)

	7	8	9	10	11	12	13	14	15	16
1	0.41	0.55	0.22	0.64	0.04	0.25	0.24	0.77	0.74	0.55
2	0.28	0.24	0.73	0.22	0.64	0.80	0.76	0.76	0.12	0.10
3	0.28	0.47	0.32	0.52	0.16	0.37	0.27	0.68	0.63	0.45
4	0.28	0.62	0.81	0.25	0.64	0.85	0.58	0.32	0.51	0.48
5	0.41	0.37	0.89	0.25	0.81	0.97	0.91	0.81	0.05	0.25
6	0.66	0.76	0.21	0.89	0.22	0.14	0.33	0.96	0.99	0.79

	7	8	9	10	11	12	13	14	15	16
1	·	·	1/5	·	3/5	·	1/5	·	·	·
2	·	2/5	·	·	·	·	·	·	·	3/5
3	3/5	·	·	·	·	·	2/5	·	·	·
4	·	·	·	2/5	·	·	·	3/5	·	·
5	·	1/5	·	1/5	·	·	·	·	3/5	·
6	·	·	2/5	·	·	3/5	·	·	·	·

Optimal transport (discrete)

From Kantorovich (1942), one can use the dual linear programming problem

$$W_k(\boldsymbol{a},\boldsymbol{b})^k = \begin{cases} \text{primal}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{C}) = \min_{P\in U_{\boldsymbol{a},\boldsymbol{b}}} \{\langle P,\boldsymbol{C}\rangle\} \\ \quad \text{or} \\ \text{dual}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{C}) = \max_{(\boldsymbol{u},\boldsymbol{v})\in M_{\boldsymbol{C}}} \{\boldsymbol{u}^\top\boldsymbol{a}+\boldsymbol{v}^\top\boldsymbol{b}\} \end{cases}$$

where $M_{\boldsymbol{C}} = \{(\boldsymbol{u},\boldsymbol{v})\in\mathbb{R}^{n_A+n_B} \big| u_i+v_j\leq \boldsymbol{C}_{i,j}\}$.

If $n_A\sim n_B\sim n$, $O(n^3\log(n))$ problem.

Set $\psi_{\boldsymbol{b}}(\boldsymbol{a},\boldsymbol{C}) = \max_{(\boldsymbol{u},\boldsymbol{v})\in M_{\boldsymbol{C}}} \{\boldsymbol{u}^\top\boldsymbol{a}+\boldsymbol{v}^\top\boldsymbol{b}\}$, $\boldsymbol{a}\mapsto\psi_{\boldsymbol{b}}(\boldsymbol{a},\boldsymbol{C})$ is a convex non-smooth map.

The dual optimum $\boldsymbol{u}^\star$ is subgradient of $\boldsymbol{a}\mapsto\psi_{\boldsymbol{b}}(\boldsymbol{a},\boldsymbol{C})$.
If $k=2$ (Euclidean distance), convex quadratic problem.

Optimal transport (discrete)

Given $P \in U_{n_{\mathtt{A}},n_{\mathtt{B}}}$, define the entropy as

$$\mathcal{E}(P) = -\sum_{i=1}^{n_{\mathtt{A}}}\sum_{j=1}^{n_{\mathtt{B}}} P_{i,j}\log P_{i,j} \text{ or } \mathcal{E}'(P) = -\sum_{i=1}^{n_{\mathtt{A}}}\sum_{j=1}^{n_{\mathtt{B}}} P_{i,j}\big[\log P_{i,j} - 1\big]$$

and consider the γ-regularized optimal transport problem

$$P^*_\gamma = \underset{P\in U_{n_{\mathtt{A}},n_{\mathtt{B}}}}{\text{argmin}}\Big\{\langle P, C\rangle - \gamma\mathcal{E}(P)\Big\} \tag{5}$$

since the problem is strictly convex.

The Lagrangian is here

$$\mathcal{L}(P,\lambda_{\mathtt{A}},\lambda_{\mathtt{B}}) = \langle P, C\rangle - \gamma\mathcal{E}(P) - \langle\lambda_{\mathtt{A}}, P\mathbf{1}_{n_{\mathtt{B}}} - \mathbf{1}_{n_{\mathtt{A}}}\rangle - \langle\lambda_{\mathtt{B}}, P^\top\mathbf{1}_{n_{\mathtt{A}}} - \mathbf{1}_{n_{\mathtt{B}}}\rangle$$

Optimal transport (discrete)

and the first order conditions are

$$C_{i,j} + \gamma \log(P_{i,j}) - \lambda_{\mathtt{A},i} - \lambda_{\mathtt{B},j} = 0,$$

i.e.

$$P_{i,j} = \exp[\lambda_{\mathtt{A},i} - C_{i,j} + \lambda_{\mathtt{B},j}] \text{ or } P = D_{\mathtt{A}} \exp[-C] D_{\mathtt{B}}$$

where $D_{\mathtt{A}}$ and $D_{\mathtt{B}}$ are diagonal matrices.

This can be related to the Doubly Stochastic Scaling Problem: let A be some $n \times n$ matrix with positive coefficients, we want to find $D_{\mathtt{A}}$ and $D_{\mathtt{B}}$ two positive diagonal matrices ($n \times n$) such that $D_{\mathtt{A}} A D_{\mathtt{B}}$ is doubly stochastic (see Parlett and Landis (1982))

More generally, this corresponds to the Matrix Scaling Problem: Let A be some $n_{\mathtt{A}} \times n_{\mathtt{B}}$ matrix with positive coefficients, we want to find $D_{\mathtt{A}}$ and $D_{\mathtt{B}}$ two positive diagonal matrices (respectively $n_{\mathtt{A}} \times n_{\mathtt{A}}$ and $n_{\mathtt{B}} \times n_{\mathtt{B}}$) such that $D_{\mathtt{A}} A D_{\mathtt{B}}$ is in $U(\boldsymbol{a}_{\mathtt{A}}, \boldsymbol{a}_{\mathtt{B}})$.

Optimal transport (discrete)

Theorem 3.3: Sinkhorn - Matrix Scaling, Sinkhorn (1962)

For any matrix $\boldsymbol{A}$ $n \times m$ with positive entries, for any $\boldsymbol{a}$ and $\boldsymbol{b}$ in the simplex, there exist unique $\boldsymbol{u} \in \mathbb{R}_+^n$ and $\boldsymbol{v} \in \mathbb{R}_+^m$ such that

$$\mathrm{diag}[\boldsymbol{u}]\ \boldsymbol{A}\ \mathrm{diag}[\boldsymbol{v}] \in U_{\boldsymbol{a},\boldsymbol{b}}.$$

Sinkhorn and Knopp (1967) (extending Sinkhorn (1962, 1964, 1966)) suggested the following algorithm (updating alternatively $D_{\mathtt{A}}$ and $D_{\mathtt{B}}$)

$$\begin{cases} D_{\mathtt{A}}^{(t)} = \mathrm{diag}(\boldsymbol{a}_{\mathtt{A}}/(AD_{\mathtt{B}})^{(t-1)}) \\ D_{\mathtt{B}}^{(t)} = \mathrm{diag}(\boldsymbol{a}_{\mathtt{B}}/(AD_{\mathtt{A}})^{(t)}) \end{cases}$$

(where the division here is element-wise).

Optimal transport (discrete)

An alternative way to write the entropic optimization problem is

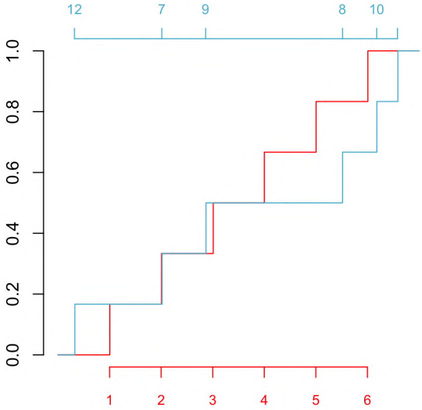
$$P_\gamma^* = \underset{P\in U_{\boldsymbol{a}_{\mathtt{A}},\boldsymbol{a}_{\mathtt{B}}}}{\text{argmin}} \left\{\langle P, C\rangle + \gamma\cdot d_{\text{KL}}(P||\boldsymbol{a}_{\mathtt{A}}\otimes\boldsymbol{a}_{\mathtt{B}})\right\} \tag{6}$$

Using mutual information here makes it easier to extend to the continuous case...

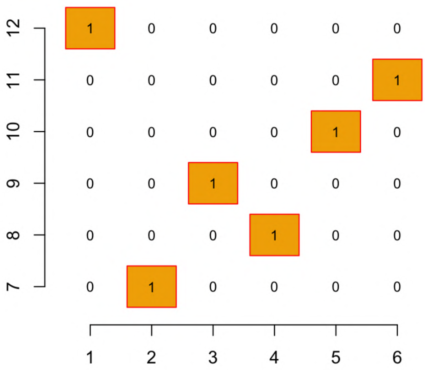
The extension of Sinkhorn algorithm is the coordinate descent/ascent algorithm.

Optimal transport (discrete)

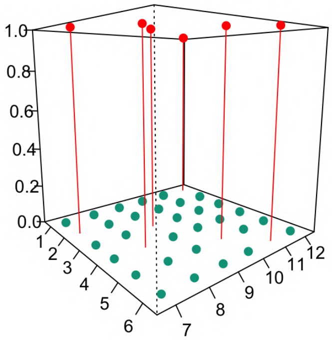
```
> set.seed(123)
> x = (1:6)/7
> y = runif(9)
> x
[1] 0.14 0.29 0.43 0.57 0.71 0.86
> y[1:6]
[1] 0.29 0.79 0.41 0.88 0.94 0.05
> library(T4transport)
> Wxy = wasserstein(x,y[1:6])
> Wxy$plan
```



Optimal transport (discrete)

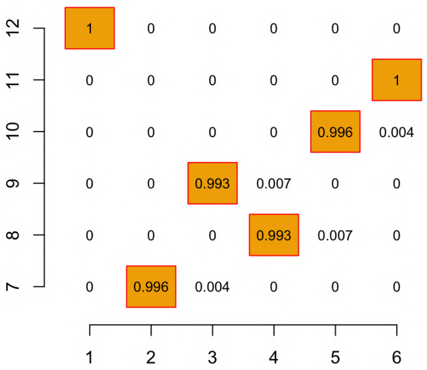


	1	2	3	4	5	6
12	1	0	0	0	0	0
11	0	0	0	0	0	1
10	0	0	0	0	1	0
9	0	0	1	0	0	0
8	0	0	0	1	0	0
7	0	1	0	0	0	0

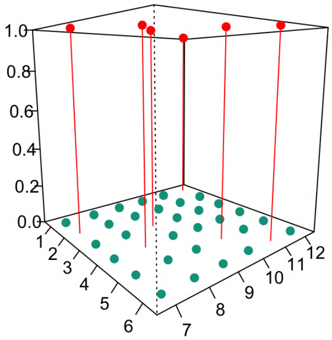


```
> Wxy = wasserstein(x,y[1:6])
> Wxy$plan
```

Optimal transport (discrete)

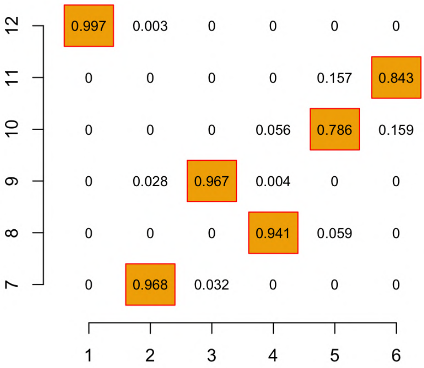


	1	2	3	4	5	6
12	1	0	0	0	0	0
11	0	0	0	0	0	1
10	0	0	0	0	0.996	0.004
9	0	0	0.993	0.007	0	0
8	0	0	0	0.993	0.007	0
7	0	0.996	0.004	0	0	0

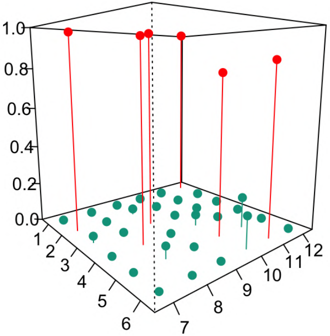


```
> Sxy = sinkhorn(x, y[1:6], p = 2, lambda = 0.001)
> Sxy$plan
```

Optimal transport (discrete)

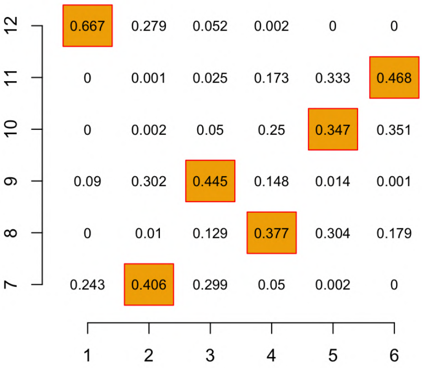


	1	2	3	4	5	6
12	0.997	0.003	0	0	0	0
11	0	0	0	0	0.157	0.843
10	0	0	0	0.056	0.786	0.159
9	0	0.028	0.967	0.004	0	0
8	0	0	0	0.941	0.059	0
7	0	0.968	0.032	0	0	0

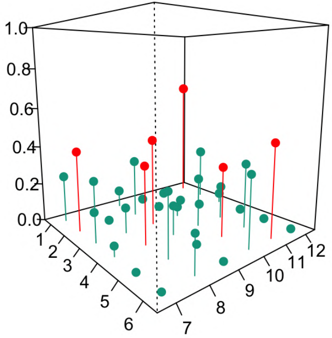


```
> Sxy = sinkhorn(x, y[1:6], p = 2, lambda = 0.005)
> Sxy$plan
```

Optimal transport (discrete)



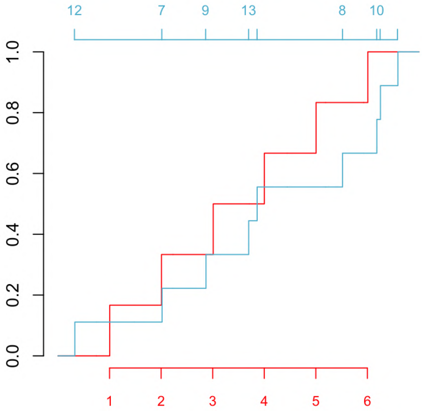
12	0.667	0.279	0.052	0.002	0	0
11	0	0.001	0.025	0.173	0.333	0.468
10	0	0.002	0.05	0.25	0.347	0.351
9	0.09	0.302	0.445	0.148	0.014	0.001
8	0	0.01	0.129	0.377	0.304	0.179
7	0.243	0.406	0.299	0.05	0.002	0
	1	2	3	4	5	6



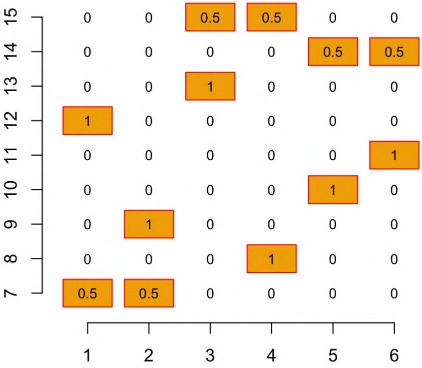
```
> Sxy = sinkhorn(x, y[1:6], p = 2, lambda = 0.05)
> Sxy$plan
```

Optimal transport (discrete)

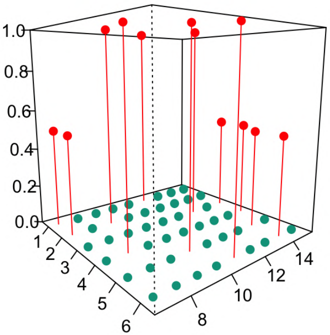
```
> y
[1] 0.29 0.79 0.41 0.88 0.94 0.05
[7] 0.53 0.89 0.55
> library(T4transport)
> Wxy = wasserstein(x,y)
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]  0.5  0.5  0.0  0.0  0.0  0.0
[2,]  0.0  0.0  0.0  1.0  0.0  0.0
[3,]  0.0  1.0  0.0  0.0  0.0  0.0
[4,]  0.0  0.0  0.0  0.0  1.0  0.0
[5,]  0.0  0.0  0.0  0.0  0.0  1.0
[6,]  1.0  0.0  0.0  0.0  0.0  0.0
[7,]  0.0  0.0  1.0  0.0  0.0  0.0
[8,]  0.0  0.0  0.0  0.0  0.5  0.5
[9,]  0.0  0.0  0.5  0.5  0.0  0.0
```



Optimal transport (discrete)

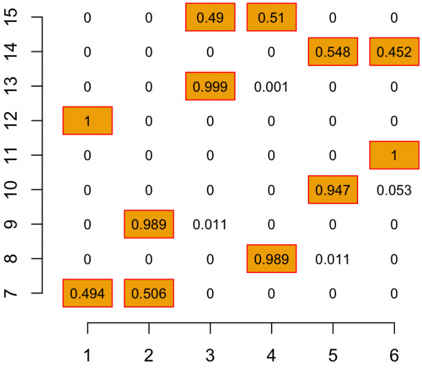


15	0	0	0.5	0.5	0	0
14	0	0	0	0	0.5	0.5
13	0	0	1	0	0	0
12	1	0	0	0	0	0
11	0	0	0	0	0	1
10	0	0	0	0	1	0
9	0	1	0	0	0	0
8	0	0	0	1	0	0
7	0.5	0.5	0	0	0	0
	1	2	3	4	5	6

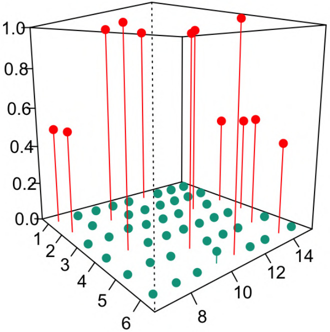


```
> Wxy = wasserstein(x,y)
> Wxy$plan
```

Optimal transport (discrete)

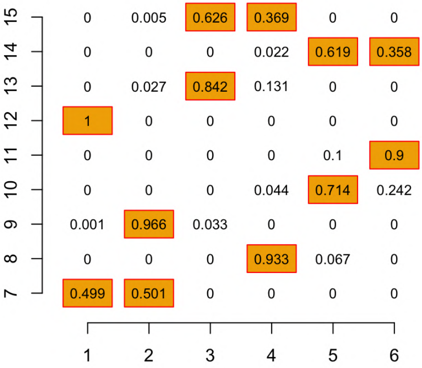


	1	2	3	4	5	6
15	0	0	0.49	0.51	0	0
14	0	0	0	0	0.548	0.452
13	0	0	0.999	0.001	0	0
12	1	0	0	0	0	0
11	0	0	0	0	0	1
10	0	0	0	0	0.947	0.053
9	0	0.989	0.011	0	0	0
8	0	0	0	0.989	0.011	0
7	0.494	0.506	0	0	0	0

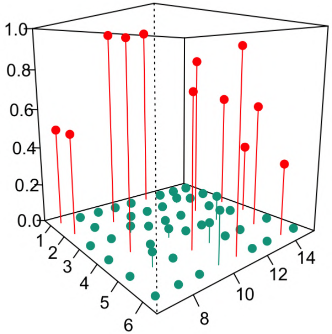


```
> Sxy = sinkhorn(x, y, p = 2, lambda = 0.001)
> Sxy$plan
```

Optimal transport (discrete)

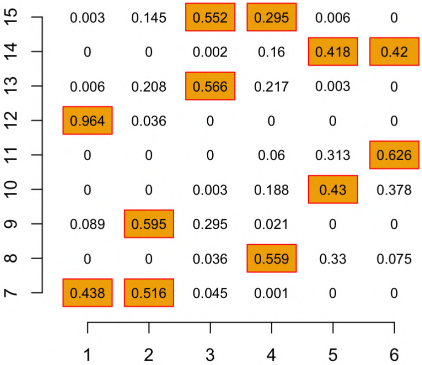


	1	2	3	4	5	6
15	0	0.005	0.626	0.369	0	0
14	0	0	0	0.022	0.619	0.358
13	0	0.027	0.842	0.131	0	0
12	1	0	0	0	0	0
11	0	0	0	0	0.1	0.9
10	0	0	0	0.044	0.714	0.242
9	0.001	0.966	0.033	0	0	0
8	0	0	0	0.933	0.067	0
7	0.499	0.501	0	0	0	0

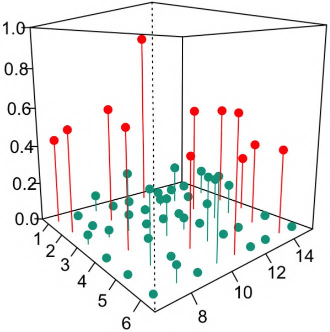


```
> Sxy = sinkhorn(x, y, p = 2, lambda = 0.005)
> Sxy$plan
```

Optimal transport (discrete)

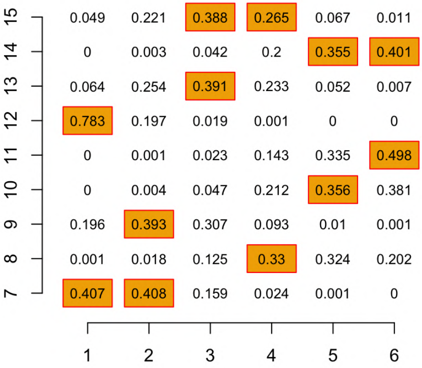


	1	2	3	4	5	6
15	0.003	0.145	0.552	0.295	0.006	0
14	0	0	0.002	0.16	0.418	0.42
13	0.006	0.208	0.566	0.217	0.003	0
12	0.964	0.036	0	0	0	0
11	0	0	0	0.06	0.313	0.626
10	0	0	0.003	0.188	0.43	0.378
9	0.089	0.595	0.295	0.021	0	0
8	0	0	0.036	0.559	0.33	0.075
7	0.438	0.516	0.045	0.001	0	0

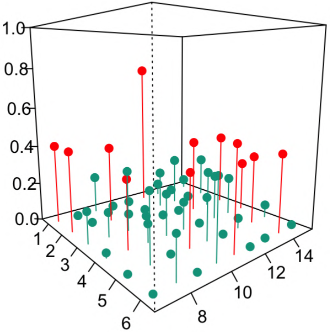


```
> Sxy = sinkhorn(x, y, p = 2, lambda = 0.02)
> Sxy$plan
```

Optimal transport (discrete)



	1	2	3	4	5	6
15	0.049	0.221	0.388	0.265	0.067	0.011
14	0	0.003	0.042	0.2	0.355	0.401
13	0.064	0.254	0.391	0.233	0.052	0.007
12	0.783	0.197	0.019	0.001	0	0
11	0	0.001	0.023	0.143	0.335	0.498
10	0	0.004	0.047	0.212	0.356	0.381
9	0.196	0.393	0.307	0.093	0.01	0.001
8	0.001	0.018	0.125	0.33	0.324	0.202
7	0.407	0.408	0.159	0.024	0.001	0



```
> Sxy = sinkhorn(x, y, p = 2, lambda = 0.05)
> Sxy$plan
```

Optimal transport (discrete)

Theorem 3.4: Optimal transport for discrete univariate distributions

Consider n points each group, on $\mathbb{R}$, $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_n\}$, ordered in the senses that $x_1 \leq x_2 \leq \cdots \leq x_n$ and $y_1 \leq y_2 \leq \cdots \leq y_n$, for any $k \geq 1$,

$$W_k = \left(\frac{1}{n}\sum_{i=1}^{n} \left|x_i - y_i\right|^k\right)^{1/k}$$

Theorem 3.5: Optimal transport for continuous univariate distributions

$$W_k = \left(\int_0^1 |F_x^{-1}(u) - F_y^{-1}(u)|^k du\right)^{1/k}$$

Optimal transport (discrete)

Theorem 3.6: Optimal transport for continuous univariate distributions

Let $\mathbb{P}_{\mathtt{A}}$ and $\mathbb{P}_{\mathtt{B}}$ be two probability measures on $\mathbb{R}$, and suppose that $c(x, y) = h(x-y)$ for some strictly convex function h. The there exists a unique $\pi \in \Pi(\mathbb{P}_{\mathtt{A}}, \mathbb{P}_{\mathtt{B}})$ such that

- π is optimal to Kantorovich problem (3.35)
- π is the comonotone joint distribution with marginals $\mathbb{P}_{\mathtt{A}}$ and $\mathbb{P}_{\mathtt{B}}$.

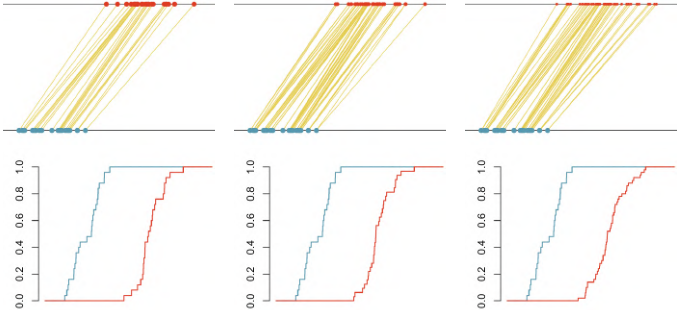
If $c(x, y) = |x - y|$, the optimal transport solution might be non-unique.

Theorem 3.7: Optimal map for continuous univariate distributions

The optimal Monge map $T^\star$ such that $T^\star_\# \mathbb{P}_{\mathtt{A}} = \mathbb{P}_{\mathtt{B}}$ is $T^\star = F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}$.

Optimal transport (discrete)

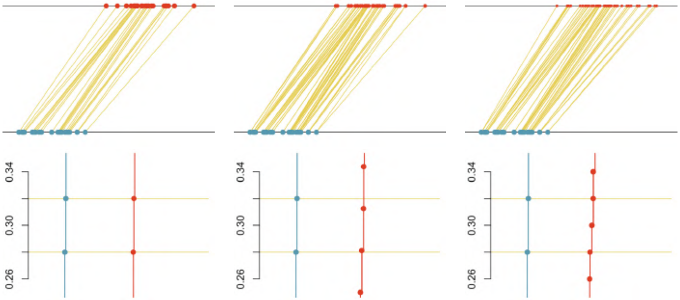
Consider $n_{\text{A}} = 25$ and $n_{\text{B}} = 25$ points in $\mathbb{R}$, $n_{\text{B}} = 32$ and $n_{\text{B}} = 50$



$$\widehat{F}_{n_{\text{A}}}(x) = \frac{1}{n_{\text{A}}} \sum_{i=1}^{n_{\text{A}}} \mathbf{1}(x_i \leq x) \text{ and } \widehat{F}_{n_{\text{B}}}(x) = \frac{1}{n_{\text{B}}} \sum_{i=1}^{n_{\text{B}}} \mathbf{1}(x_i \leq x)$$

Optimal transport (discrete)

Consider $n_{\mathrm{A}}=25$ and $n_{\mathrm{B}}=25$ points in $\mathbb{R}$, $n_{\mathrm{B}}=32$ and $n_{\mathrm{B}}=50$

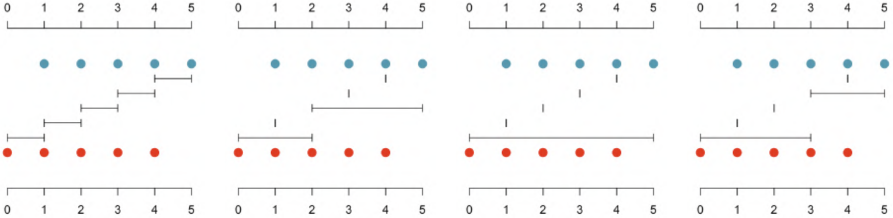


$$\widehat{F}_{n_{\mathrm{A}}}(x)=\frac{1}{n_{\mathrm{A}}}\sum_{i=1}^{n_{\mathrm{A}}}\mathbf{1}(x_i\leq x)\text{ and }\widehat{F}_{n_{\mathrm{B}}}(x)=\frac{1}{n_{\mathrm{B}}}\sum_{i=1}^{n_{\mathrm{B}}}\mathbf{1}(x_i\leq x)$$

Optimal transport (discrete)

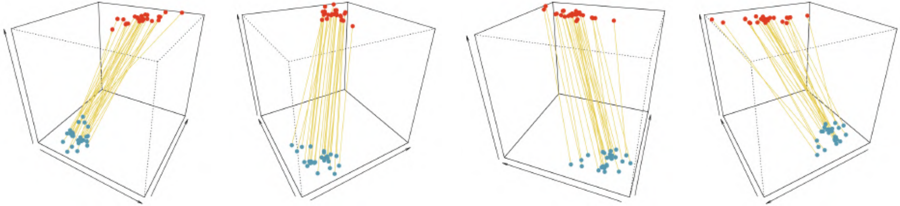
In the univariate case, if $k = 1$,

$$W_1 = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - y_{\sigma(i)}\right|$$

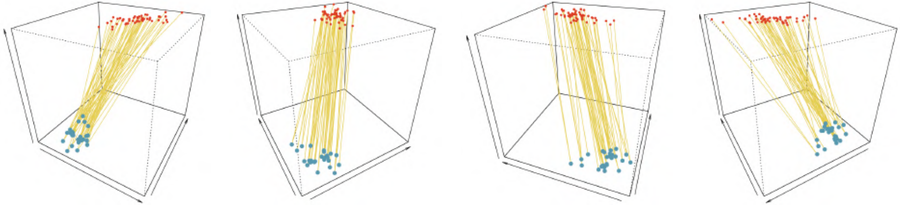


Multivariate Optimal Transport

Consider n and n points in $\mathbb{R}^2$

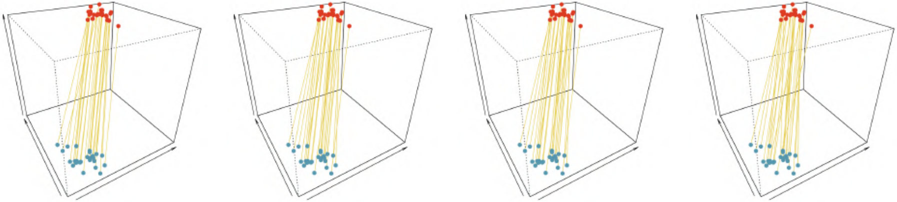


Consider n and $2n$ points in $\mathbb{R}^2$

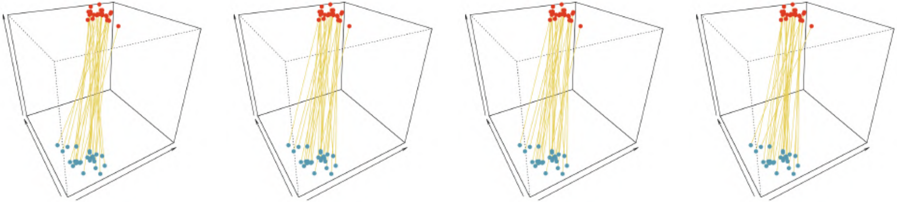


Multivariate Optimal Transport

Consider n and n points in $\mathbb{R}^2$, and $k = 1, 2, 3, 4$, $T_{\#}\mathbb{P}_{\mathtt{A}} = \mathbb{P}_{\mathtt{B}}$



Consider n and n points in $\mathbb{R}^2$, and $p = 1, 2, 3, 4$, $T_{\#}\mathbb{P}_{\mathtt{B}} = \mathbb{P}_{\mathtt{A}}$



Multivariate Optimal Transport

Theorem 3.8: Optimal map for continuous multivariate distributions, Brenier (1991)

With a quadratic cost, the optimal Monge map $T^\star$ is unique, and it is the gradient of a convex function, $T^\star = \nabla\varphi$.

Example Multidimensional Gaussian

$$\boldsymbol{x}_{\mathtt{B}} = \mathcal{T}^\star(\boldsymbol{x}_{\mathtt{A}}) = \boldsymbol{\mu}_{\mathtt{B}} + \boldsymbol{A}(\boldsymbol{x}_{\mathtt{A}} - \boldsymbol{\mu}_{\mathtt{A}}),$$

where $\boldsymbol{A}$ is a symmetric positive matrix that satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_{\mathtt{A}}\boldsymbol{A} = \boldsymbol{\Sigma}_{\mathtt{B}}$, which has a unique solution given by $\boldsymbol{A} = \boldsymbol{\Sigma}_{\mathtt{A}}^{-1/2}\big(\boldsymbol{\Sigma}_{\mathtt{A}}^{1/2}\boldsymbol{\Sigma}_{\mathtt{B}}\boldsymbol{\Sigma}_{\mathtt{A}}^{1/2}\big)^{1/2}\boldsymbol{\Sigma}_{\mathtt{A}}^{-1/2}$, where $\boldsymbol{M}^{1/2}$ is the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008).

Multivariate Optimal Transport

Proposition 3.33: W_2 for Gaussian vectors

Consider two Gaussian distributions, then

$$W_2(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1),\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))^2=\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_2^2+\mathrm{tr}\big(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2-2(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{1/2})^{1/2}\big)$$

Proof: Let $\boldsymbol{X}_1\sim\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$, $\boldsymbol{X}_2\sim\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, and Γ define the covariance matrix of $(\boldsymbol{X}_1,\boldsymbol{X}_2)$,

$$\Gamma=\begin{pmatrix}\Sigma_1 & C\\ C^\top & \Sigma_2\end{pmatrix}$$

where (generally), C is some $n_1\times n_2$ matrix. Recall that $n_1\times n_2$ matrices can have a pseudo-inverse, in the sense that (Penrose conditions)

$$\begin{cases}AA^-A=A\\ A^-AA^-=A^-,\end{cases}\quad\begin{cases}(AA^-)^\top=AA^-\\ (A^-A)^\top=A^-A,\end{cases}$$

Multivariate Optimal Transport

Observe that $\mathbb{E}(\|\boldsymbol{X}_1-\boldsymbol{X}_2\|_{\ell_2}^2)=\text{tr}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2-2C)$. Recall that C must satisfy the Schur complement constraint, $\boldsymbol{\Sigma}_1-C\boldsymbol{\Sigma}_2^{-1}C^\top\succeq 0$, so that we want to solve

$$C^\star=\text{argmin}\{-2\text{tr}(C)\}\text{ s.t. }\boldsymbol{\Sigma}_1-C\boldsymbol{\Sigma}_2^{-1}C^\top\succeq 0,$$

as studied in Olkin and Pukelsheim (1982), where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive ($\succeq 0$) matrices.

Let $\mathcal{G}=\{C,n_1\times n_2:\ \boldsymbol{\Sigma}_1-C\boldsymbol{\Sigma}_2^{-1}C^\top\succeq 0\}$, $\mathcal{S}=\{S:\ SS^-\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}_2\}$, one can prove (standard duality and convexity arguments) that

$$\max_{C\in\mathcal{G}}\{2\text{tr}(C)\}=\max_{S\in\mathcal{S}}\{\text{tr}(\boldsymbol{\Sigma}_1S+\boldsymbol{\Sigma}_2S^-)\}=2\text{tr}\big(\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{1/2}\big)$$

with respective (unique) solutions

$$\begin{cases}C^\star=\boldsymbol{\Sigma}_1S^\star\\ S^\star=\boldsymbol{\Sigma}_2^{1/2}\big[(\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{1/2})^{1/2}\big]\boldsymbol{\Sigma}_2^{1/2}\end{cases}$$

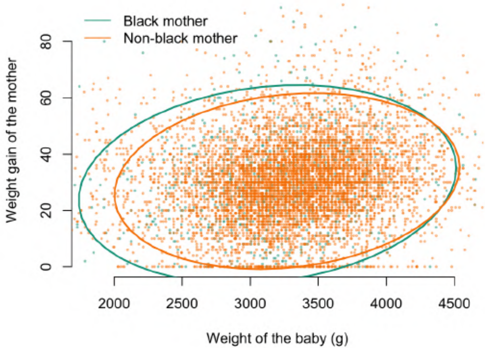
Multivariate Optimal Transport

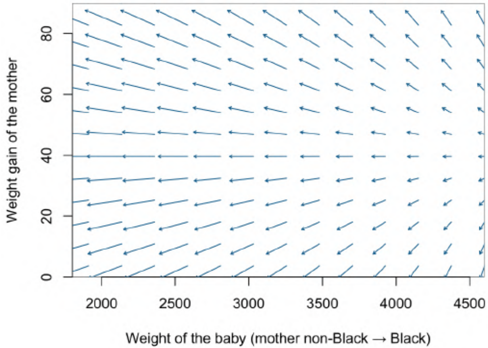
See Olkin and Pukelsheim (1982), Givens and Shortt (1984) and Knott and Smith (1984), or more recently Takatsu (2008) and Takatsu and Yokota (2012), with more geometric interpretations.

To illustrate, consider the previous example, with newborn weight and weight gain of mothers, in the U.S., with Black and non-Black mothers, with here a joint mapping $\mathbb{R}_2^+ \to \mathbb{R}_2^+$.

Multivariate Optimal Transport

$(x_1, x_2) \leftrightarrow (x_1, x_2)$ (newborn weight, weight gain of the mother)





– Part 3 –

Models

Generalized Linear Model

Definition 4.1: Exponential family, McCullagh and Nelder (1989)

The distribution of Y is in the exponential family if its density (with respect to some appropriate measure) is

$$f_{\theta,\varphi}(y) = \exp\left(\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi)\right),$$

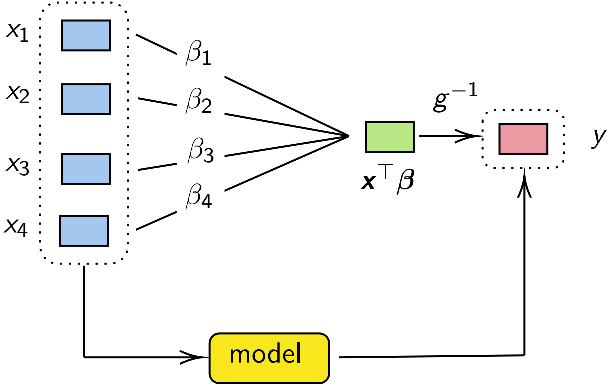
where θ is the canonical parameter, φ is a nuisance parameter, and $b : \mathbb{R} \to \mathbb{R}$ is some $\mathbb{R} \to \mathbb{R}$ function.

Such as the binomial, Poisson, Gaussian, gamma distributions, etc.

Also compound Poisson / Tweedie (from Tweedie (1984)).

Generalized Linear Model

Given some dataset $(y_i, \boldsymbol{x}_i)$, suppose that $\mu(\boldsymbol{x}) = g^{-1}(\boldsymbol{x}^\top \beta)$



OLS, $\mu(\boldsymbol{x}) = \boldsymbol{x}^\top \beta$ and $\widehat{\beta}^{\text{ols}} = \text{argmin}\left\{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \beta)^2\right\} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$.

Generalized Linear Model

Consider problems

$$\begin{array}{l}\min_{\boldsymbol{x}\in\mathbb{R}^k}\{f(\boldsymbol{x})\}\\ \text{under constraint } g(\boldsymbol{x})=\boldsymbol{0}\end{array} \quad \text{or} \quad \begin{array}{l}\min_{\boldsymbol{x}\in\mathbb{R}^k}\{f(\boldsymbol{x})\}\\ \text{under constraint } g(\boldsymbol{x})\leq\boldsymbol{0}\end{array}$$

Karush-Kuhn-Tucker condition is

$$\begin{cases}\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{z}^\star)=\boldsymbol{0}\\ \nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{z}^\star)=\boldsymbol{0}\end{cases}$$

where

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{z})=f(\boldsymbol{x})+\boldsymbol{z}^\top g(\boldsymbol{x})$$

is the Lagrangian problem (parameter $\boldsymbol{z}$ are multipliers)

Generalized Linear Model

Definition 4.2: Ridge Estimator (OLS), Hoerl and Kennard (1970)

$$\widehat{\beta}_\lambda^{\text{ridge}} = \underset{\beta\in\mathbb{R}^k}{\text{argmin}}\left\{\frac{1}{2}\sum_{i=1}^n(y_i-\boldsymbol{x}_i^\top\beta)^2+\lambda\sum_{j=1}^k\beta_j^2\right\}.$$

$$\widehat{\beta}_\lambda^{\text{ridge}} = (\boldsymbol{X}^\top\boldsymbol{X}+\lambda\mathbb{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

Definition 4.3: Ridge Estimator (GLM)

$$\widehat{\beta}_\lambda^{\text{ridge}} = \underset{\beta\in\mathbb{R}^k}{\text{argmin}}\left\{-\sum_{i=1}^n\log f(y_i|\mu_i=g^{-1}(\boldsymbol{x}_i^\top\beta))+\lambda\sum_{j=1}^k\beta_j^2\right\}.$$

Generalized Linear Model

Definition 4.4: LASSO Estimator (OLS), Tibshirani (1996)

$$\widehat{\beta}_{\lambda}^{\text{lasso}} = \text{argmin}\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\beta)^2 + \lambda\sum_{j=1}^{k}|\beta_j|\right\}.$$

Definition 4.5: LASSO Estimator (GLM)

$$\widehat{\beta}_{\lambda}^{\text{lasso}} = \text{argmin}\left\{-\sum_{i=1}^{n}\log f(y_i|\mu_i = g^{-1}(\boldsymbol{x}_i^\top\beta)) + \lambda\sum_{j=1}^{k}|\beta_j|\right\}.$$

Generalized Linear Model

```
> library(glmnet)
> fit_ridge = glmnet(x, y, alpha = 0)
> fit_lasso = glmnet(x, y, alpha = 1)
```

Elastic net

$$\min\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^\top\beta)^2+\lambda_1\sum_{j=1}^{k}|\beta_j|+\frac{\lambda_2}{2}\sum_{j=1}^{k}\beta_j^2\right\},$$

e.g. $\lambda_1=\alpha\lambda$ and $\lambda_2=(1-\alpha)\lambda$ (two parameters — one for the global regularization, one for the trade-off between Ridge (Tikhonov) vs. Lasso)

Accuracy

Consider the case where $y \in \{0, 1\}$, and a score $m(\boldsymbol{x})$ (classically in $[0, 1]$).

E.g., for a logistic regression, $m(\boldsymbol{x}) = \dfrac{\exp[\boldsymbol{x}^\top \beta]}{1 + \exp[\boldsymbol{x}^\top \beta]}$.

> **Receiver operating characteristic**
> A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier model (can be used for multi class classification as well) at varying threshold values. The true-positive rate is also known as sensitivity, recall or probability of detection. The false-positive rate is also known as the probability of false alarm and equals (1 - specificity). W

Accuracy

Definition 4.6: ROC curve

The ROC curve is the parametric curve

$$\{\mathbb{P}[m(\boldsymbol{X}) > t|Y = 0], \mathbb{P}[m(\boldsymbol{X}) > t|Y = 1]\} \text{ for } t \in [0, 1],$$

when the score $m(\boldsymbol{X})$ and Y evolve in the same direction (a high score indicates a high risk).

$$C(t) = \text{TPR} \circ \text{FPR}^{-1}(t),$$

where

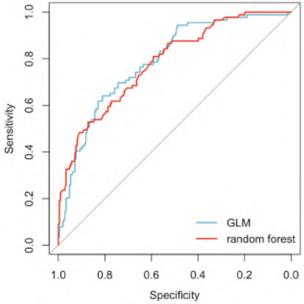
$$\begin{cases} \text{FRP}(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 0] = \mathbb{P}[m_0(\boldsymbol{X}) > t] \\ \text{TPR}(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 1] = \mathbb{P}[m_1(\boldsymbol{X}) > t]. \end{cases}$$

Accuracy

```
> library(ROCR)
> pred = prediction(df$yhat, df$y)
> roc = performance(pred,"tpr","fpr")
> plot(roc)
> auc = performance( pred,"auc")
```

see also

```
> library(pROC)
```

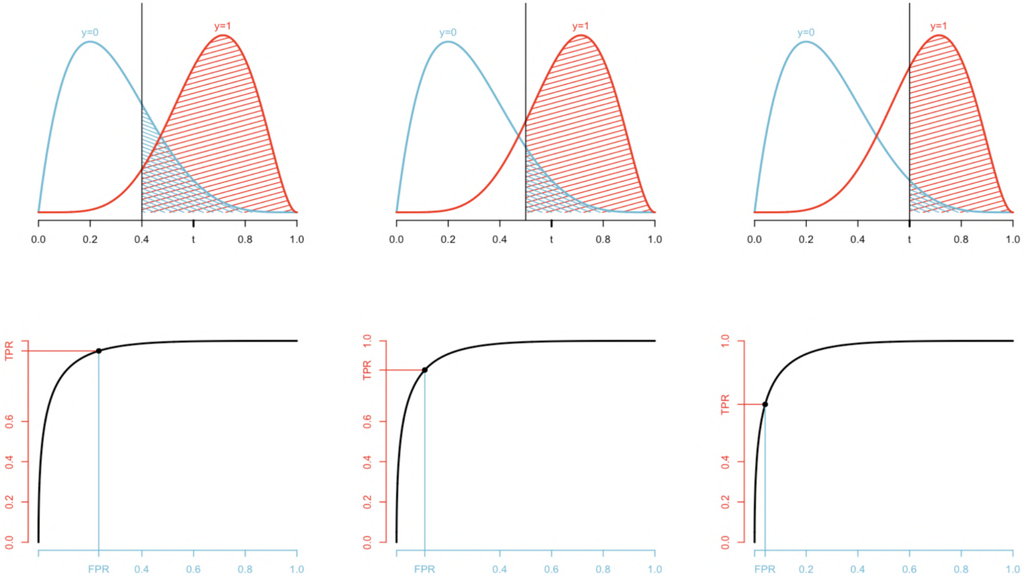


Definition 4.7: AUC, area under the ROC curve

The area under the curve is defined as the area below the ROC curve,

$$\text{AUC} = \int_0^1 C(t)\mathrm{d}t = \int_0^1 \text{TPR} \circ \text{FPR}^{-1}(t)\mathrm{d}t.$$

Accuracy



Calibration

Well-calibration was initially discussed in forecasting

Definition 4.8: Well-calibrated (1), Van Calster et al. (2019), Krüger and Ziegel (2021)

The forecast X of Y is a well-calibrated forecast of Y if $\mathbb{E}(Y|X) = X$ almost surely, or $\mathbb{E}[Y|X = x] = x$, for all x.

one can define "well-calibration" in prediction

Definition 4.9: Well-calibrated (2), Zadrozny and Elkan (2002); Cohen and Goldszmidt (2004)

The prediction $m(\boldsymbol{X})$ of Y is a well-calibrated prediction if $\mathbb{E}[Y|m(\boldsymbol{X}) = \widehat{y}] = \widehat{y}$, for all $\widehat{y}$.

Calibration

"*[S]uppose the Met Office says that the probability of rain tomorrow in your region is 80%. They aren't saying that it will rain in 80% of the land area of your region, and not rain in the other 20%. Nor are they saying it will rain for 80% of the time. What they are saying is there is an 80% chance of rain occurring at any one place in the region, such as in your garden.* [...] *[A] forecast of 80% chance of rain in your region should broadly mean that, on about 80% of days when the weather conditions are like tomorrow's, you will experience rain where you are. [...] If it doesn't rain in your garden tomorrow, then the 80% forecast wasn't wrong, because it didn't say rain was certain. But if you look at a long run of days, on which the Met Office said the probability of rain was 80%, you'd expect it to have rained on about 80% of them.*" McConway (2021)

Calibration

"*Well calibrated classifiers are probabilistic classifiers for which the output can be directly interpreted as a confidence level. For instance, a well calibrated (binary) classifier should classify the samples such that among the samples to which it gave a*[predicted probability] *value close to 0.8, approximately 80% actually belong to the positive class*," scikit learn: Probability calibration

"*Suppose that a forecaster sequentially assigns probabilities to events. He is well calibrated if, for example, of those events to which he assigns a probability 30 percent, the long-run proportion that actually occurs turns out to be 30 percent*," Dawid (1982)

Calibration

"*Out of all the times you said there was a 40 percent chance of rain, how often did rain actually occur? If, over the long run, it really did rain about 40 percent of the time, that means your forecasts were well calibrated*," Silver (2012)

"*we desire that the estimated class probabilities are reflective of the true underlying probability of the sample*," Kuhn and Johnson (2013)

See Murphy and Epstein (1967), Roberts (1968), Gneiting and Raftery (2005) on ensemble methods for weather forecasting, or more generally Lichtenstein et al. (1977), Oakes (1985), Gneiting et al. (2007).

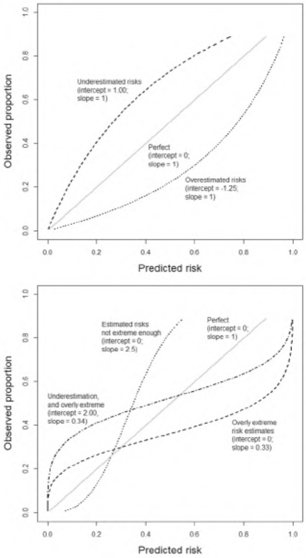
Calibration

As explained in Van Calster et al. (2019), "*among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event*",

- If 40 out of 100 in this group are found to have the disease, the risk is underestimated
- If we observe that in this group, 10 out of 100 have the disease, we have overestimated the risk.

Hosmer-Lemeshow test, from Hosmer Jr et al. (2013) (logistic regression), and Bier score, from Brier (1950) and Murphy (1973)

Function plotted in psychological papers Keren (1991)



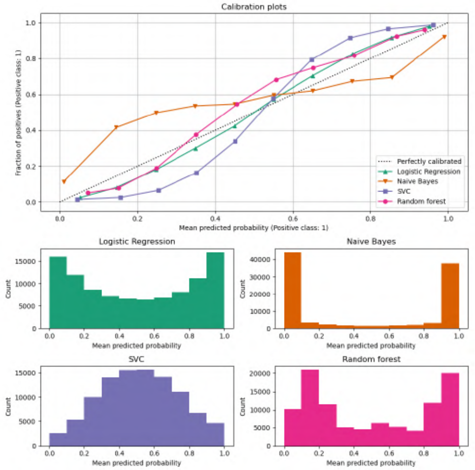
Calibration

- "reliability diagrams", Wilks (1990)

Used in `scikit-learn` (`calibration_curve`) see Pakdaman Naeini et al. (2015) and Kumar et al. (2019), with quantile-based bins (average of y_i's against average of $\hat{m}(\mathbf{x}_i)$'s)

- "local regression", Denuit et al. (2021)

See also Austin and Steyerberg (2019) regression of y_i's against $\hat{m}(\mathbf{x}_i)$'s



Calibration

Definition 4.10: Calibration plot

The calibration plot associated with model m is the function $\widehat{y} \mapsto \mathbb{E}(Y|m(\boldsymbol{X}) = \widehat{y})$. The empirical version is some local regression on $\{y_i, m(\boldsymbol{x}_i)\}$.

Definition 4.11: Globally unbiased model m, Denuit et al. (2021)

Model m is globally unbiased if $\mathbb{E}[Y] = \mathbb{E}[m(\boldsymbol{X})]$.

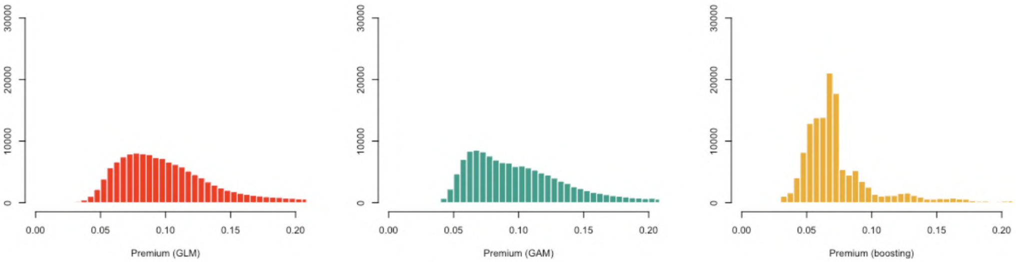
Definition 4.12: Locally unbiased model m, Denuit et al. (2021)

Model m is locally unbiased at $\widehat{y}$ if $\mathbb{E}[Y|m(\boldsymbol{X}) = \widehat{y}] = \widehat{y}$.

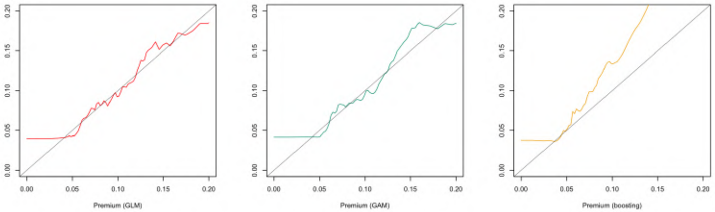
Calibration

Consider claims (annual) frequency, corrected from the exposure, `freMTPL2freq` from `CASDataset` package, as in Denuit et al. (2021).

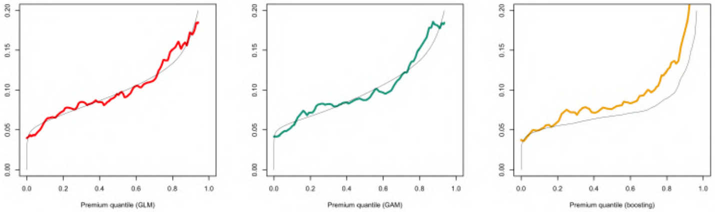
	$\widehat{m}^{\text{glm}}$	$\widehat{m}^{\text{gam}}$	$\widehat{m}^{\text{bst}}$
average $\widehat{m}(\mathbf{x})$'s	0.1087	0.1092	0.0820
10% quantile	0.0605	0.0598	0.0498
90% quantile	0.1682	0.1713	0.1244



Calibration



Evolution of $p \mapsto \mathbb{E}[Y|\widehat{m}(\boldsymbol{X}) = p]$ and $u \mapsto \mathbb{E}[Y|\widehat{m}(\boldsymbol{X}) = F_{\widehat{m}}^{-1}(u)]$



Calibration

For GLM, remember that

$$f(y_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i,\varphi)\right),$$

$$\frac{\partial \log \mathcal{L}_i}{\partial \beta_j} = \frac{\partial \log \mathcal{L}_i}{\partial \theta_i}\cdot\frac{\partial \theta_i}{\partial \mu_i}\cdot\frac{\partial \mu_i}{\partial \eta_i}\cdot\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \log \mathcal{L}_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\varphi}\cdot\frac{1}{V(\mu_i)}\cdot x_{i,j}\cdot\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1}$$

When g is the canonical link ($g_\star = b'^{-1}$ or $\eta_i = \boldsymbol{x}_i^\top\beta = \theta_i$)

$$\nabla \log \mathcal{L} = \boldsymbol{X}^\top(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{0}$$

Proposition 4.1: Calibration of GLM

In the GLM framework with the canonical link function, $\widehat{m}(\boldsymbol{x}) = g_\star^{-1}(\boldsymbol{x}_i^\top\widehat{\beta})$ is globally unbiased (on the training dataset), but possibly locally biased.

Calibration

Otherwise

$$\nabla \log \mathcal{L} = \boldsymbol{X}^\top \boldsymbol{\Omega}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{0},$$

where $\boldsymbol{\Omega}$ is a diagonal matrix ($\boldsymbol{\Omega} = \boldsymbol{W}\boldsymbol{\Delta}$, where $\boldsymbol{W} = \text{diag}\big((V(\mu_i)g'(\mu_i)^2)^{-1}\big)$ and $\boldsymbol{\Delta} = \text{diag}\big(g'(\mu_i)\big)$, so that we recognize Fisher information - corresponding to the Hessian matrix (up to a negative sign) – $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$).

	training data					validation data				
	$\overline{y}$	GLM	CART	GAM	RF	$\overline{y}$	GLM	CART	GAM	RF
$\widehat{m}(\boldsymbol{x}, s)$	8.73	8.73	8.73	8.73	8.27	8.55	9.05	9.03	8.84	8.70
$\widehat{m}(\boldsymbol{x})$	8.73	8.73	8.73	8.73	8.29	8.55	9.05	9.03	8.84	8.73

Calibration

Definition 4.13: Brier score (binary classifier) Brier (1950)

Brier score is the mean squared error of probability estimate,

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{m}(\boldsymbol{x}_i), y_i\big)^2$$

Consider "confidence" value given by `Picpurify`, using pictures generate by a GAN (a generative adversarial network, used in Hill and White (2020)).

Calibration

female (0.984)
male (0.016)

female (0.983)
male (0.017)

female (0.982)
male (0.018)

female (0.960)
male (0.040)

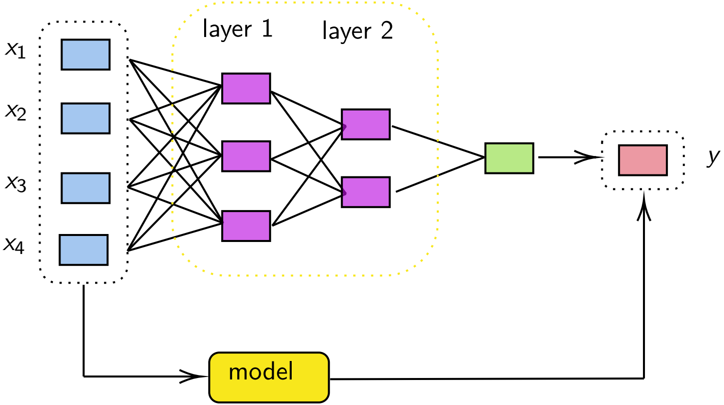
female (0.009)
male (0.991)

female (0.013)
male (0.987)

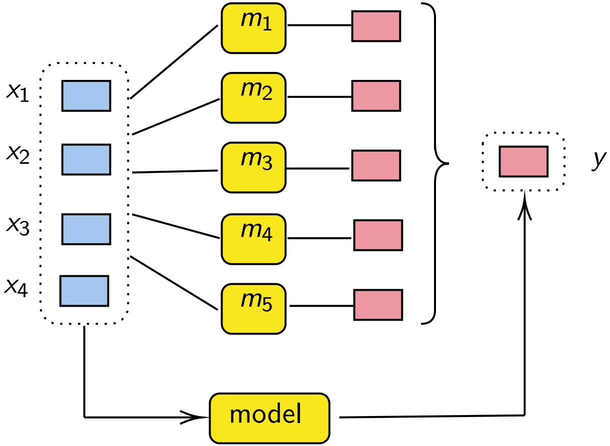
female (0.014)
male (0.986)

female (0.015)
male (0.985)

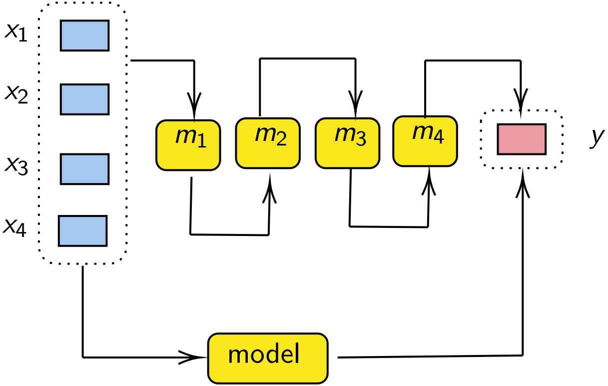
Standard modeling architecture



Standard modeling architecture



Standard modeling architecture

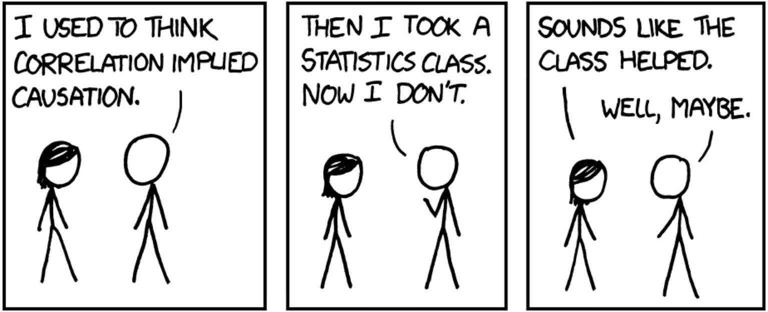


– Part 4 –

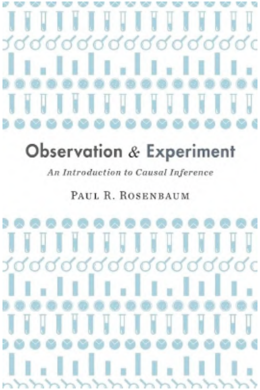
Data

Data (the two types)

"*It is often said, 'You cannot prove causality with statistics.' One of my professors, Frederick Mosteller, liked to counter, 'You can only prove causality with statistics.'* (...) *The title, 'Observation and Experiment,' marks the modern distinction between randomized experiments and observational studies,*" Rosenbaum (2018)



Correlation, Randall Munroe, 2009 https://xkcd.com/552/



Data (the three rung ladder)

"Ladder of causation" from Pearl et al. (2009)

3. **Counterfactuals**

(Imagining, "*what if I had done...*")

2. **Intervention**

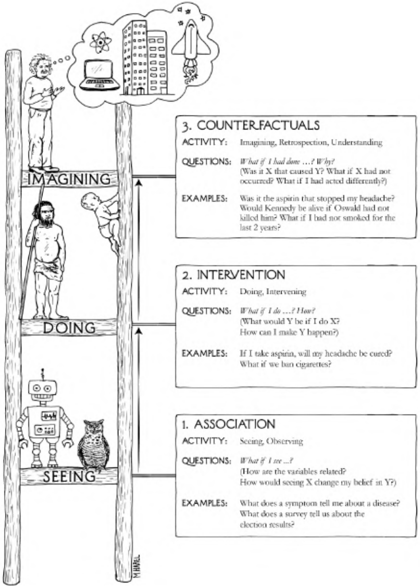
(Doing, "*what if I do...*")

1. **Association**

(Seeing, "*what if I see...*")

Picture source: Pearl and Mackenzie (2018)

What would be the impact of a treatment T on a variable of interest Y ?



3. COUNTERFACTUALS
ACTIVITY: Imagining, Retrospection, Understanding
QUESTIONS: *What if I had done ...? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION
ACTIVITY: Doing, Intervening
QUESTIONS: *What if I do ...? How?*
(What would Y be if I do X?
How can I make Y happen?)
EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)
EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

Proxy

"*OK, let's not use race, but should we use zip code, which of course is a proxy for race in our segregated society?*," O'Neil (2016).

Definition 5.1: Proxy, Merriam-Webster (2022)

A proxy is a person authorized to act for another (from a contracted form of the Middle English word *procuracie* (from French "procuration")).

Definition 5.2: Perfect proxy, Datta et al. (2017)

A variable X is a perfect proxy for Z if there exist functions $\varphi : \mathcal{X} \to \mathcal{Z}$ and $\psi : \mathcal{Z} \to \mathcal{Y}$ such that

$$\mathbb{P}[X = \psi(Z)] = \mathbb{P}[\varphi(X) = Z] = 1.$$

Proxy

Definition 5.3: Comonotonicity, Hoeffding (1940); Fréchet (1951)

Variables X and Y are comonotonic if $(X,Y) = (F_x^{-1}(U), F_y^{-1}(U))$ for some $U \sim \mathcal{U}([0,1])$.

> **Comonotonicity**
> In probability theory, comonotonicity mainly refers to the perfect positive dependence between the components of a random vector, essentially saying that they can be represented as increasing functions of a single random variable. W

See also Dhaene et al. (2002a,b) on comonotonic vectors.

See also Prince and Schwarcz (2019), or Tschantz (2022) for discrimination by proxy.

Range of possible situation between independence and perfect proxy.

Independence

> **Independence**
> Independence is a fundamental notion in probability theory, as in statistics and the theory of stochastic processes. Two events are independent if, informally speaking, the occurrence of one does not affect the probability of occurrence of the other or, equivalently, does not affect the odds. W

Definition 5.4: Independence (dimension 2)

X and Y are independent, denoted $X \perp\!\!\!\perp Y$, if for any sets $\mathcal{A}, \mathcal{B} \subset \mathbb{R}$,

$$\mathbb{P}\big[X \in \mathcal{A}, Y \in \mathcal{B}\big] = \mathbb{P}\big[X \in \mathcal{A}\big] \cdot \mathbb{P}\big[Y \in \mathcal{B}\big].$$

Definition 5.5: Linear Independence (dimension 2)

Consider two random variables X and Y. $X \perp Y$ if and only if $\text{Cov}[X, Y] = 0$.

Independence

> **Correlation**
> in the broadest sense, ”correlation” may indicate any type of association, in statistics it usually refers to the degree to which a pair of variables are linearly related. W

Definition 5.6: Correlation (dimension 2), Pearson (1895)

X and Y are two random variables

$$\text{Corr}\,[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}}.$$

where $\text{Cov}[X, Y] = \mathbb{E}\big[\,(X - \mathbb{E}[X])\,(Y - \mathbb{E}[Y])\,\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Independence

From Cauchy-Schwarz theorem, $-1 \leq \text{Corr}[X, Y] \leq +1$ but those bounds are rarely sharp,

Proposition 5.1: Correlation bounds (dimension 2)

For any random variables X and Y (with finite variances), $r_{\text{min}} \leq \text{Corr}[X, Y] \leq r_{\text{max}}$, where

$$r_{\text{min}} = \frac{\text{Cov}[F_x^{-1}(U), F_y^{-1}(1-U)]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}} \text{ and } r_{\text{max}} = \frac{\text{Cov}[F_x^{-1}(U), F_y^{-1}(U)]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}}$$

Maximal correlation is obtained when X and Y are comonotonic (minimal correlation when X and $-Y$ are comonotonic).

Related to optimal transport, see also Knott and Smith (1984).

Independence

Proposition 5.2

Consider two random variables X and Y. $X \perp\!\!\!\perp Y$ if and only if for any functions $\varphi : \mathbb{R} \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ (such that the expected values below exist and are well-defined) $\text{Cov}[\varphi(X), \psi(Y)] = 0$, i.e.,

$$\mathbb{E}[\varphi(X) \cdot \psi(Y)] = \mathbb{E}[\varphi(X)] \cdot \mathbb{E}[\psi(Y)].$$

Definition 5.7: Maximal Correlation, HGR

Consider two random variables X and Y,

$$r^\star(X, Y) = \max_{\varphi,\psi}\{\text{Corr}[\varphi(X), \psi(Y)]\}.$$

Independence

HGR because of Hirschfeld (1935), Gebelein (1941) and Rényi (1959) (also Sarmanov (1958a,b)).

$$r^\star(X,Y)=\max_{\varphi\in\mathcal{F}_x,\ \psi\in\mathcal{G}_y}\mathbb{E}[\varphi(X)\psi(Y)],$$

where

$$\begin{cases}\mathcal{F}_x=\{\varphi:\mathcal{X}\to\mathbb{R}:\mathbb{E}[\varphi(X)]=0\text{ and }\mathbb{E}[\varphi^2(X)]=1\}\\ \mathcal{G}_y=\{\psi:\mathcal{Y}\to\mathbb{R}:\mathbb{E}[\psi(Y)]=0\text{ and }\mathbb{E}[\psi^2(Y)]=1\}\end{cases}$$

See either `ccaPP` or `acepack` package,

```
> ccaPP::maxCorProj(x = x, y = y, method = "pearson")
> corstar = acepack::ace(x = x, y = y)
> cor(corstar$tx, corstar$ty)
```

Independence

Proposition 5.3

Consider two random variables X and Y. $X \perp\!\!\!\perp Y$ if and only if $r^\star(X,Y)=0$.

Proof: Given a random variable X, its characteristic function is $\phi_X(t)=\mathbb{E}[e^{itX}]$. Recall that

$$\begin{cases}\phi_X(t)=\phi_Y(t),\ \forall t\in\mathbb{R} \text{ if and only if } X\overset{\mathcal{L}}{=}Y\\ \phi_{X,Y}(s,t)=\mathbb{E}[e^{i(sX+tY)}]=\phi_X(s)\cdot\phi_Y(t),\ \forall s,t\in\mathbb{R} \text{ if and only if } X\perp\!\!\!\perp Y\end{cases}$$

If $r^\star(X,Y)=0$, let $s,t\in\mathbb{R}$ and consider $\varphi(x)=\phi_X(x)=\mathbb{E}[e^{ixX}]$ and $\psi(y)=\phi_Y(y)=\mathbb{E}[e^{isY}]$, then $\text{Cov}[e^{isX},e^{itY}]=\text{Cov}[X'_s,Y'_t]=0$, i.e. $\mathbb{E}[X'_sY'_t]=\mathbb{E}[X'_s]\mathbb{E}[Y'_t]$,

$$\underbrace{\mathbb{E}[e^{i(sX+tY)}]}_{\phi_{XY}(s,t)}=\underbrace{\mathbb{E}[e^{isX}]\cdot\mathbb{E}[e^{itY}]}_{\phi_X(s)\cdot\phi_Y(t)},\ \forall s,t\in\mathbb{R} \text{ i.e. } X\perp\!\!\!\perp Y.$$

Independence

Proposition 5.4

Consider two random variables X and Y such that (X, Y) is a Gaussian vector. Then $r^\star(X, Y) = |\text{Corr[X,Y]}|$.

See Lancaster (1957, 1958), and Gauss-Hermite decomposition

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2\left[1-\rho^2\right]}\right) = \phi(x)\phi(y)\cdot\sum_{i=0}^{\infty} r^i H_i(x)H_i(y)$$

where H_i's are Hermite polynomial.

Independence

Instead of

$$r^\star(X,Y) = \max_{\varphi\in\mathcal{F}_x,\ \psi\in\mathcal{G}_y} \mathbb{E}[\varphi(X)\psi(Y)],$$

where

$$\begin{cases}\mathcal{F}_x = \{\varphi : \mathcal{X}\to\mathbb{R} : \mathbb{E}[\varphi(X)] = 0 \text{ and } \mathbb{E}[\varphi^2(X)] = 1\}\\ \mathcal{G}_y = \{\psi : \mathcal{Y}\to\mathbb{R} : \mathbb{E}[\psi(Y)] = 0 \text{ and } \mathbb{E}[\psi^2(Y)] = 1\}\end{cases}$$

Definition 5.8: Constrained Maximal Correlation, Bach and Jordan (2002), Gretton et al. (2005)

Consider two random variables X and Y, as well as some Hilbert spaces $\bar{\mathcal{F}}_x \subset \mathcal{F}_x$ and $\bar{\mathcal{G}}_y \subset \mathcal{G}_y$,

$$\bar{r}^\star(X,Y) = \max_{\varphi\in\bar{\mathcal{F}}_x,\psi\in\bar{\mathcal{G}}_y} \{\text{Corr}[\varphi(X),\psi(Y)]\}.$$

Independence

Kimeldorf and Sampson (1978) and Kimeldorf et al. (1982) suggested to consider for $\bar{\mathcal{F}}_x$ and $\bar{\mathcal{G}}_y$ as subsets of monotone functions.

$$\begin{cases}\bar{\mathcal{F}}x = \{\varphi \in \mathcal{F}_x : \varphi \text{ monotone}\}\\ \bar{\mathcal{G}}_y = \{\psi \in \mathcal{G}_y : \psi \text{ monotone}\}\end{cases}$$

See Mourier (1953), Hannan (1961), Jensen and Mayer (1977) and Lin (1987).

Independence

Definition 5.9: Linear Independence

In a general context, consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively. $\boldsymbol{X} \perp \boldsymbol{Y}$ if and only if for any $\boldsymbol{a} \in \mathbb{R}^{d_x}$ and $\boldsymbol{b} \in \mathbb{R}^{d_y}$

$$\text{Cov}[\boldsymbol{a}^\top \boldsymbol{X}, \boldsymbol{b}^\top \boldsymbol{Y}] = 0.$$

Definition 5.10: Independence

In a general context, consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ if and only if for any $\mathcal{A} \subset \mathbb{R}^{d_x}$ and $\mathcal{B} \subset \mathbb{R}^{d_y}$,

$$\mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\} \cap \{\boldsymbol{Y} \in \mathcal{B}\}] = \mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\}] \cdot \mathbb{P}[\{\boldsymbol{Y} \in \mathcal{B}\}].$$

Independence

Proposition 5.5: Independence

Consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ if and only if for any functions $\varphi : \mathbb{R}^{d_x} \to \mathbb{R}$ and $\psi : \mathbb{R}^{d_y} \to \mathbb{R}$ (such that the expected values below exist and are well-defined)

$$\mathbb{E}[\varphi(\boldsymbol{X})\psi(\boldsymbol{Y})] = \mathbb{E}[\varphi(\boldsymbol{X})] \cdot \mathbb{E}[\psi(\boldsymbol{Y})],$$

or equivalently

$$\text{Cov}[\varphi(\boldsymbol{X}), \psi(\boldsymbol{Y})] = 0.$$

Independence

Definition 5.11: Mutual Independence

Let $\boldsymbol{Y} = (Y_1, \cdots, Y_k)$ denote some random vector. All components of $\boldsymbol{Y}$ are (mutually) independent if for any $\mathcal{A}_1, \cdots, \mathcal{A}_k \subset \mathbb{R}$

$$\mathbb{P}\left[\{(Y_1, \cdots, Y_k) \in \bigcap_{i=1}^{k} \mathcal{A}_i\}\right] = \prod_{i=1}^{k} \mathbb{P}[\{Y_i \in \mathcal{A}_i\}].$$

Definition 5.12: Conditional Independence (dimension 2)

X and Y are independent conditionally on Z, denoted $X \perp\!\!\!\perp Y \mid Z$, if for any sets $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathbb{R}$,

$$\mathbb{P}\big[X \in \mathcal{A}, Y \in \mathcal{B} \big| Z \in \mathcal{C}\big] = \mathbb{P}\big[X \in \mathcal{A} \big| Z \in \mathcal{C}\big] \cdot \mathbb{P}\big[Y \in \mathcal{B} \big| Z \in \mathcal{C}\big].$$

Independence

Definition 5.13: Conditional Independence

In a general context, consider three random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$. $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y})|\boldsymbol{Z}$ if and only if for any $\mathcal{A} \subset \mathbb{R}^{d_x}$, $\mathcal{B} \subset \mathbb{R}^{d_y}$ and $\mathcal{C} \subset \mathbb{R}^{d_z}$,

$$\mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\} \cap \{\boldsymbol{Y} \in \mathcal{B}\}|\boldsymbol{Z} \in \mathcal{C}] = \mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\}|\boldsymbol{Z} \in \mathcal{C}] \cdot \mathbb{P}[\{\boldsymbol{Y} \in \mathcal{B}\}|\boldsymbol{Z} \in \mathcal{C}].$$

Proposition 5.6

Consider three random variables X, Y, and Z. If $X \perp Z$ and $Y \perp Z$, then $aX + bY \perp Z$, for any $a, b \in \mathbb{R}$.

Independence

Proposition 5.7: $X \perp Z,\ Y \perp Z \not\Longrightarrow \psi(X,Y) \perp Z$

Consider three random variables X, Y, and Z. If $X \perp Z$ and $Y \perp Z$, it does not imply that $\psi(X,Y) \perp Z$, for any $\psi : \mathbb{R}^2 \to \mathbb{R}$.

$$(X,Y,Z) = \begin{cases} (0,0,0) & \text{with probability } 1/4, \\ (0,1,1) & \text{with probability } 1/4, \\ (1,0,1) & \text{with probability } 1/4, \\ (1,1,0) & \text{with probability } 1/4. \end{cases}$$

Proposition 5.8

Consider a random vector $\boldsymbol{X}$ in $\mathbb{R}^k$, and a random variable Z.
$\boldsymbol{X} \perp Z$ does not imply that $\psi(\boldsymbol{X}) \perp Z$, for any $\psi : \mathbb{R}^k \to \mathbb{R}$.

Independence

Proposition 5.9

Consider three random variables X, Y, and Z. Even if $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$, it does not imply either that $\psi(X,Y) \perp Z$ or that $\psi(X,Y) \perp\!\!\!\perp Z$, for any $\psi:\mathbb{R}^2 \to \mathbb{R}$.

Proposition 5.10

Consider a random vector $\boldsymbol{X}$ in $\mathbb{R}^k$, and a random variable Z.
$\boldsymbol{X} \perp\!\!\!\perp Z$ does not imply either that $\psi(\boldsymbol{X}) \perp Z$ **or** $\psi(\boldsymbol{X}) \perp\!\!\!\perp Z$, for any $\psi:\mathbb{R}^k \to \mathbb{R}$.

Causation

Definition 5.14: Common cause, Reichenbach (1956)

If X and Y are non-independent, $X \not\perp\!\!\!\perp Y$, then, either

$$\begin{cases} X \text{ causes } Y \\ Y \text{ causes } X \\ \text{there exists } Z \text{ such that } Z \text{ causes both } X \text{ and } Y. \end{cases}$$

See also Bollen and Pearl (2013)

SCM, Goldberger (1972), Duncan (1975) or Bollen (1989)

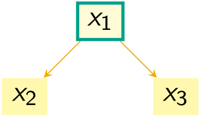
Bayesian network, Pearl (1985), Henrion (1988), Charniak (1991)

Causal path diagrams and probabilistic DAGs, Spirtes et al. (1993)

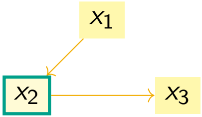
Causation

Sewall Wright (see Wright (1921, 1934)) use directed graphs to represent probabilistic cause and effect relationships among a set of variables, and developed path diagrams and path analysis

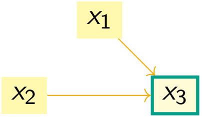
(a)
confounder



(b)
mediator



(c)
collider



Causation

Definition 5.15: Path

A path π from a node x_i to another node x_j is a sequence of nodes and edges starting at x_i and ending at x_j.

Definition 5.16: d-separation

A set of nodes $\boldsymbol{x}_i$ is said to be *d*-separated with another set of nodes $\boldsymbol{x}_j$ by $\boldsymbol{x}_c$ whenever every path from any $x_i \in \boldsymbol{x}_i$ to any $x_j \in \boldsymbol{x}_j$ is blocked by $\boldsymbol{x}_c$. We will simply denote $\boldsymbol{x}_i \perp_{\mathcal{G}} \boldsymbol{x}_j \mid \boldsymbol{x}_c$.

Proposition 5.11

Two nodes x_i and x_j are *d*-separated by $\boldsymbol{x}_c$ if and only members of $\boldsymbol{x}_c$ block all paths from x_i to x_j.

Causation

$$\text{Chain rule}: \begin{cases} \mathbb{P}[x_1,x_2,x_3,x_4]=\mathbb{P}[x_1]\times\mathbb{P}[x_2|x_1]\times\mathbb{P}[x_3|x_1,x_2]\times\mathbb{P}[x_4|x_1,x_2,x_3] \\ \mathbb{P}[x_1,x_2,x_3,x_4]=\mathbb{P}[x_4]\times\mathbb{P}[x_3|x_4]\times\mathbb{P}[x_2|x_3,x_4]\times\mathbb{P}[x_1|x_2,x_3,x_4] \end{cases}$$

Definition 5.17: Directed acyclic graph, DAG (or causal graph)

A directed acyclic graph (DAG) $\mathcal{G}$ is a directed graph with no directed cycles.

Definition 5.18: Markov Property

Given a causal graph $\mathcal{G}$ with nodes $\boldsymbol{x}$, the joint distribution of $\boldsymbol{X}$ satisfies the (global) Markov property with respect to $\mathcal{G}$ if, for any disjoints $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_c$

$$\boldsymbol{x}_1 \perp_{\mathcal{G}} \boldsymbol{x}_2 \mid \boldsymbol{x}_c \;\Rightarrow\; \boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 \mid \boldsymbol{X}_c.$$

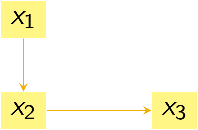
Causation

Proposition 5.12: Probabilistic graphical model

If $\boldsymbol{X}$ satisfies the (global) Markov property with respect to $\mathcal{G}$

$$\mathbb{P}[x_1, \cdots, x_n] = \prod_{i=1}^{n} \mathbb{P}[x_i | \text{parents}(x_i)]$$

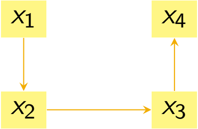
where parents(x_i) are nodes with edges directed towards x_i



Path from x_1 to x_3 is blocked by x_2, i.e., $x_1 \perp_{\mathcal{G}} x_3 \mid x_2$, or $X_1 \perp\!\!\!\perp X_3 \mid X_2$. From the chain rule,

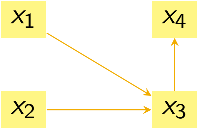
$$\mathbb{P}[x_1, x_2, x_3] = \mathbb{P}[x_1] \times \mathbb{P}[x_2 | x_1] \times \underbrace{\mathbb{P}[x_3 | x_2, x_1]}_{\mathbb{P}[x_3 | x_2]}$$

Causation



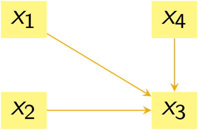
From the chain rule, for the causal graph on the left (top),

$$\mathbb{P}[x_1,x_2,x_3,x_4]=\mathbb{P}[x_1]\times\mathbb{P}[x_2|x_1]\times\mathbb{P}[x_3|x_2]\times\mathbb{P}[x_4|x_3]$$



From the chain rule, for the causal graph on the left (middle),

$$\mathbb{P}[x_1,x_2,x_3,x_4]=\mathbb{P}[x_1]\times\mathbb{P}[x_2]\times\mathbb{P}[x_3|x_1,x_2]\times\mathbb{P}[x_4|x_3]$$



From the chain rule, for the causal graph on the left (bottom),

$$\mathbb{P}[x_1,x_2,x_3,x_4]=\mathbb{P}[x_1]\times\mathbb{P}[x_2]\times\mathbb{P}[x_3|x_1,x_2,x_4]\times\mathbb{P}[x_4]$$

Intervention

$\mathbb{P}[Y \in \mathcal{A} | X = x]$: how $Y \in \mathcal{A}$ is likely to occur if X happened to be equal to x
Therefore, it is an observational statement.

$P[Y \in \mathcal{A} | \text{do}(X = x)]$: how $Y \in \mathcal{A}$ is likely to occur if X is set to x
It is here an intervention statement.

Using causal graphs, intervention $\text{do}(X = x)$ means that all incoming edges to x are cut.

If $P[Y \in \mathcal{A} | \text{do}(X = x)] \neq \mathbb{P}[Y \in \mathcal{A} | X = x]$, it means that X and Y are confounded, see Pearl (2009).

Intervention

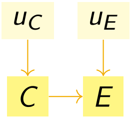
Definition 5.19: Structural Causal Models (SCM)

In a simple causal graph, with two nodes C (the cause) and E (the effect), the causal graph is $C \to E$, and the mathematical interpretation can be summarized in two assignments

$$\begin{cases} C = h_c(U_C) \\ E = h_e(C, U_E), \end{cases}$$

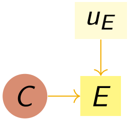
where U_C and U_E are two independent random variables, $U_C \perp\!\!\!\perp U_E$.

(a) observation



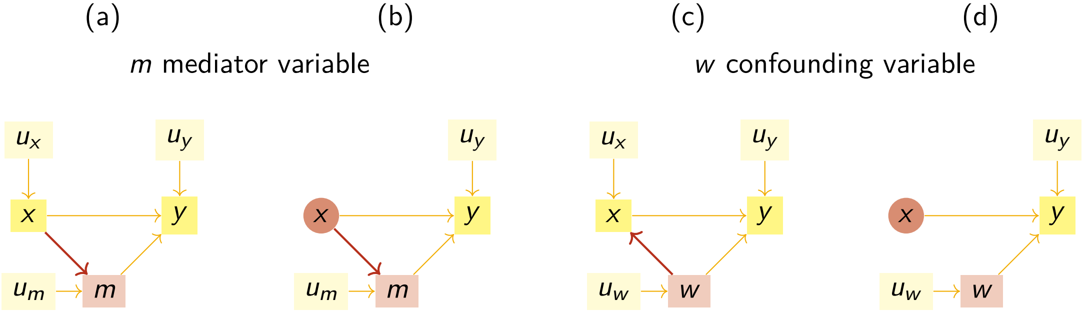
$$\begin{cases} C = h_c(U_C) \\ E = h_e(C, U_E) \end{cases}$$

(b) intervention



$$\begin{cases} C = c \quad (\text{or } \text{do}(C = c)) \\ E_c^\star = h_e(c, U_E) \end{cases}$$

Intervention



$$\begin{cases}\text{mediator}: & \mathbb{P}[Y^\star_x=1]=\mathbb{P}[Y=1|\text{do}(X=x)]=\mathbb{P}[Y=1|X=x]\\ \text{confusion}: & \mathbb{P}[Y^\star_x=1]=\mathbb{P}[Y=1|\text{do}(X=x)]\neq\mathbb{P}[Y=1|X=x].\end{cases}$$

Intervention

In fact, in the presence of a confounding factor, $\mathbb{P}[Y_x^\star = 1]$ which corresponds to $\mathbb{P}[Y = 1|\text{do}(X = x)]$ should be written

$$\sum_w \mathbb{P}[Y = 1|W = w, X = x] \cdot \mathbb{P}[W = w] = \mathbb{E}(\mathbb{P}[Y = 1|W, X = x]).$$

Causal Inference and counterfactuals

Define potential outcomes to quantify the treatment effect, $\text{TE} = y^\star_{i,T\leftarrow 1} - y^\star_{i,T\leftarrow 0}$

$$\begin{cases}\text{observation} & : \; y^\star_{i,T\leftarrow 1} \text{ when } t_i = 1 \text{ is observed, and } \boldsymbol{x}_i \\ \text{counterfactual} & : \; y^\star_{i,T\leftarrow 0} \text{ when } t_i = 1 \text{ is observed, and } \boldsymbol{x}_i\end{cases}$$

Here we want to observe counterfactuals $y^\star_{i,T\leftarrow t'}$ at the individual level.

	Gender	Name	Treatment			Outcome (Weight)				Height	⋯
			t_i	0	1	y_i	$y^\star_{i,T\leftarrow 0}$	$y^\star_{i,T\leftarrow 1}$	TE	x_i	⋯
1	H	Alex	0	☑	☐	75	75	64	11	172	⋯
2	F	Betty	1	☐	☑	52	67	52	15	161	⋯
3	F	Beatrix	1	☐	☑	57	71	57	14	163	⋯
4	H	Ahmad	0	☑	☐	78	78	61	17	183	⋯

Different notations are used $y(1)$ and $y(0)$ in Imbens and Rubin (2015), y^1 and y^0 in Cunningham (2021), or $y_{t=1}$ and $y_{t=0}$ in Pearl and Mackenzie (2018).

Causal Inference and counterfactuals

Define potential outcomes to quantify the treatment effect, TE $= y^\star_{i,T\leftarrow 1} - y^\star_{i,T\leftarrow 0}$

$$\begin{cases} \text{observation} & : \; y^\star_{i,T\leftarrow 1} \text{ when } t_i = 1 \text{ is observed, and } \boldsymbol{x}_i \\ \text{counterfactual} & : \; y^\star_{i,T\leftarrow 0} \text{ when } t_i = 1 \text{ is observed, and } \boldsymbol{x}_i \end{cases}$$

Here we want to observe counterfactuals $y^\star_{i,T\leftarrow t'}$ at the individual level.

	Gender	Name	Treatment			Outcome (Weight)				Height	···
			t_i	0	1	y_i	$y^\star_{i,T\leftarrow 0}$	$y^\star_{i,T\leftarrow 1}$	TE	x_i	···
1	H	Alex	0	☑	☐	75	75	?	?	172	···
2	F	Betty	1	☐	☑	52	?	52	?	161	···
3	F	Beatrix	1	☐	☑	57	?	57	?	163	···
4	H	Ahmad	0	☑	☐	78	78	?	?	183	···

Different notations are used $y(1)$ and $y(0)$ in Imbens and Rubin (2015), y^1 and y^0 in Cunningham (2021), or $y_{t=1}$ and $y_{t=0}$ in Pearl and Mackenzie (2018).

Causal Inference and counterfactuals

Definition 5.20: Average Treatment Effect, Holland (1986)

Given a treatment T, the average treatment effect on outcome Y is

$$\tau = \text{ATE} = \mathbb{E}\big[Y^\star_{t\leftarrow 1} - Y^\star_{t\leftarrow 0}\big].$$

Definition 5.21: Conditional Average Treatment Effect, Wager and Athey (2018)

Given a treatment T, the conditional average treatment effect on outcome Y, given some covariates $\boldsymbol{X}$,is

$$\tau(\boldsymbol{x}) = \text{CATE}(\boldsymbol{x}) = \mathbb{E}\big[Y^\star_{t\leftarrow 1} - Y^\star_{t\leftarrow 0}\big|\boldsymbol{X} = \boldsymbol{x}\big].$$

Causal Inference and counterfactuals

Definition 5.22: Individual Average Treatment Effect

Given a treatment T, the conditional average treatment effect on outcome Y, for individual i, given covariates $\boldsymbol{X}_i$, is

$$\text{IATE}(i) = \mathbb{E}\big[Y^\star_{i,t\leftarrow(1-t_i)} - Y^\star_{i,t\leftarrow t_i}\big].$$

– Part 5 –

Sensitive Variables and Proxies

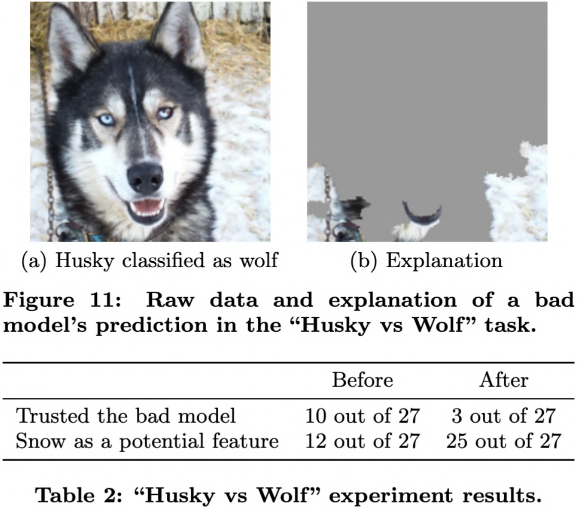
Context

There exists list of variables considered (by law) as sensitive (e.g., in Québec)

- race,
- color,
- sex,
- gender identity or expression,
- pregnancy,
- sexual orientation,
- civil status,
- age,
- religion,
- political convictions,
- language,
- ethnic or national origin,
- social condition,
- disability.

Explainability

“*On a collection of additional 60 images, the classifier predicts “Wolf” if there is snow (or light background at the bottom), and “Husky” otherwise, regardless of animal color, position, pose, etc.*”, Ribeiro et al. (2016)



(a) Husky classified as wolf (b) Explanation

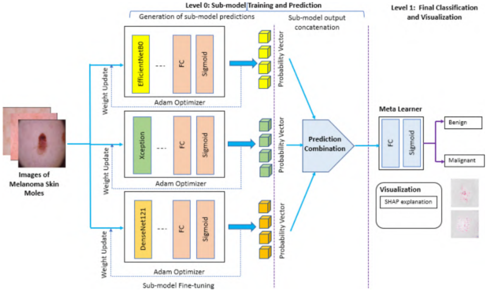
Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

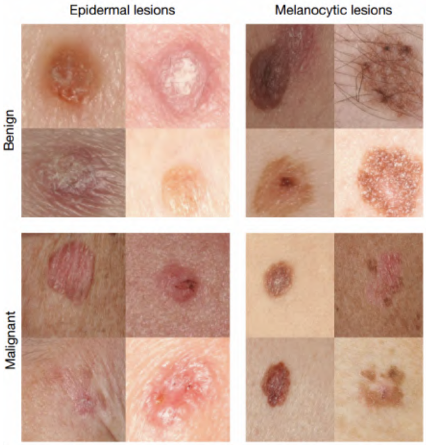
	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: “Husky vs Wolf” experiment results.

Explainability

Esteva et al. (2017) and Winkler et al. (2019) use deep-classifiers to detect skin cancer





“So in the set of biopsy images, if an image had a ruler in it, the algorithm was more likely to call a tumor malignant, because the presence of a ruler correlated with an increased likelihood a lesion was cancerous,” Patel (2017)

Racial Discrimination

Racial Discrimination

Definition 6.1: Racism, Merriam-Webster (2022)

A belief that race is a fundamental determinant of human traits and capacities and that racial differences produce an inherent superiority of a particular race; *also* behavior or attitudes that reflect and foster this belief.

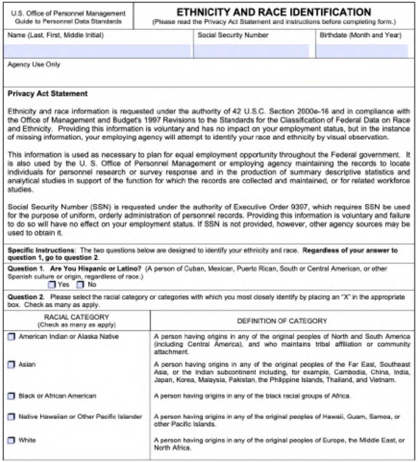
Du Bois (1899)

Racial Discrimination

Gannon (2016) "*race is a social construct*"

In the U.S., "*an individual's response to the race question is based upon self-identification*"

- White American, European American, or Middle Eastern American (59.3%)
- "Hispanic or Latino Americans (18.9%)"
- Black or African American (12.6%)
- American Indian or Alaska Native (0.7%)
- Asian American (5.9%)
- Native Hawaiian or Other Pacific Islander (0.2%)



U.S. Office of Personnel Management
Guide to Personnel Data Standards

ETHNICITY AND RACE IDENTIFICATION
(Please read the Privacy Act Statement and instructions before completing form.)

Name (Last, First, Middle Initial)	Social Security Number	Birthdate (Month and Year)

Agency Use Only

Privacy Act Statement

Ethnicity and race information is requested under the authority of 42 U.S.C. Section 2000e-16 and in compliance with the Office of Management and Budget's 1997 Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. Providing this information is voluntary and has no impact on your employment status, but in the instance of missing information, your employing agency will attempt to identify your race and ethnicity by visual observation.

This information is used as necessary to plan for equal employment opportunity throughout the Federal government. It is also used by the U. S. Office of Personnel Management or employing agency maintaining the records to locate individuals for personnel research or survey response and in the production of summary descriptive statistics and analytical studies in support of the function for which the records are collected and maintained, or for related workforce studies.

Social Security Number (SSN) is requested under the authority of Executive Order 9397, which requires SSN be used for the purpose of uniform, orderly administration of personnel records. Providing this information is voluntary and failure to do so will have no effect on your employment status. If SSN is not provided, however, other agency sources may be used to obtain it.

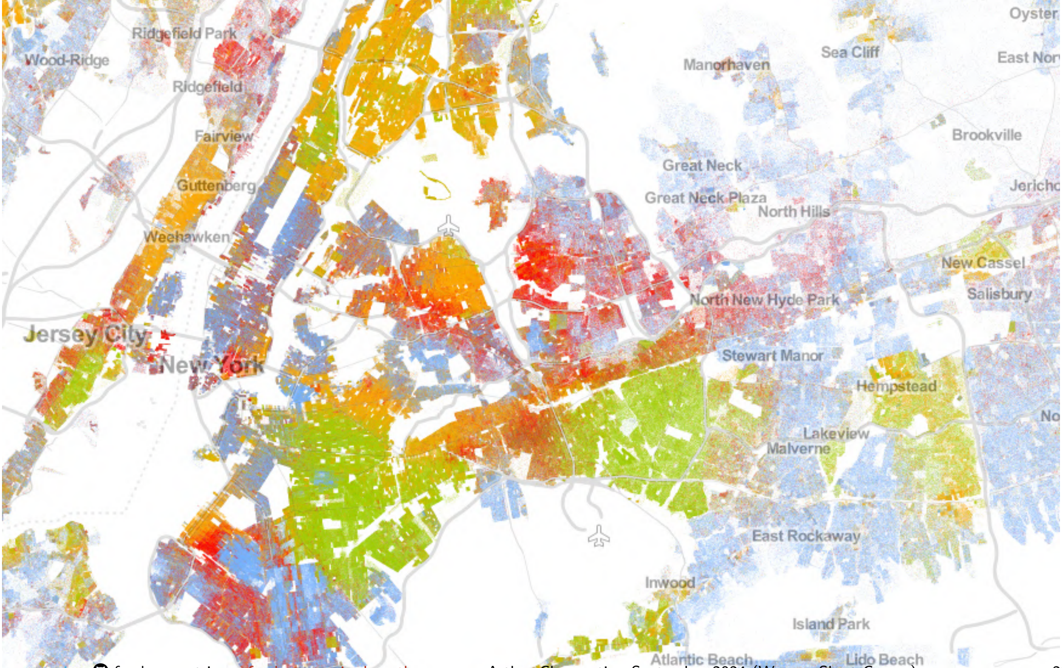
Specific Instructions: The two questions below are designed to identify your ethnicity and race. **Regardless of your answer to question 1, go to question 2.**

Question 1. Are You Hispanic or Latino? (A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race.)
☐ Yes ☐ No

Question 2. Please select the racial category or categories with which you most closely identify by placing an "X" in the appropriate box. Check as many as apply.

RACIAL CATEGORY (Check as many as apply)	DEFINITION OF CATEGORY
☐ American Indian or Alaska Native	A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
☐ Asian	A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
☐ Black or African American	A person having origins in any of the black racial groups of Africa.
☐ Native Hawaiian or Other Pacific Islander	A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.
☐ White	A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

See maps on https://www.arcgis.com/apps/mapviewer/index.html


Oyster
Ridgefield Park
Wood-Ridge
Sea Cliff
Manorhaven
East Nor
Ridgefield
Brookville
Fairview
Great Neck
Guttenberg
Jerich
Great Neck Plaza
North Hills
Weehawken
New Cassel
Salisbury
North New Hyde Park
Jersey City
New York
Stewart Manor
Hempstead
No
Lakeview
Malverne
East Rockaway
Inwood
Island Park
Atlantic Beach
Lido Beach

Racial Discrimination

By comparing skull anatomy and skin color, "*generis humani varietates quinae principes, species vero unica*" (one species, and five principle varieties of humankind), Blumenbach (1775)

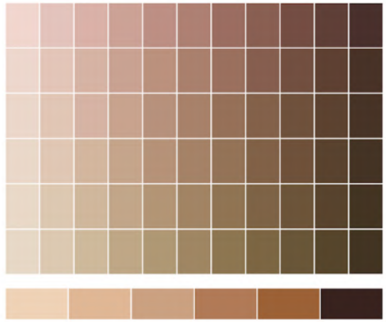
- the "*Caucasian*" (or white race, for Europeans, including Middle Easterners and South Asians in the same category),
- the "*Mongolian*" (or yellow race, including all East Asians)
- the "*Malayan*" (or brown race, including Southeast Asians and Pacific Islanders)
- the "*Ethiopian*" (or black race, including all sub-Saharan Africans)
- the "*American*" (or red race, including all Native Americans)

Racial Discrimination

Definition 6.2: Colourism, Merriam-Webster (2022)

Prejudice or discrimination especially within a racial or ethnic group favoring people with lighter skin over those with darker skin.

Fitzpatrick Skin Scale (six levels), Telles (2014).



Racial Discrimination

In the context of insurance, several reference in the late XIX-th Century

"*industrial insurers operated a high-volume business; so to simplify sales they charged the same nickel to everyone. The home office then calculated benefits according to actuarially defensible discrimination, by age initially and then by race. In November 1881, Metropolitan decided to mimic Prudential, allowing policies to be sold to African Americans once again, but with the understanding that black policyholders' survivors only received two-thirds of the standard benefit,*" Bouk (2015)

1884, Massachusetts state legislature passed the Act to Prevent Discrimination by Life Insurance Companies Against People of Color

See Frederick L. Hoffman (1896) (discussed earlier)

Racial Discrimination

In auto insurance, Heller (2015) observed that African American neighbourhood pay 70% more, on average, for auto insurance premiums than other neighbourhoods.

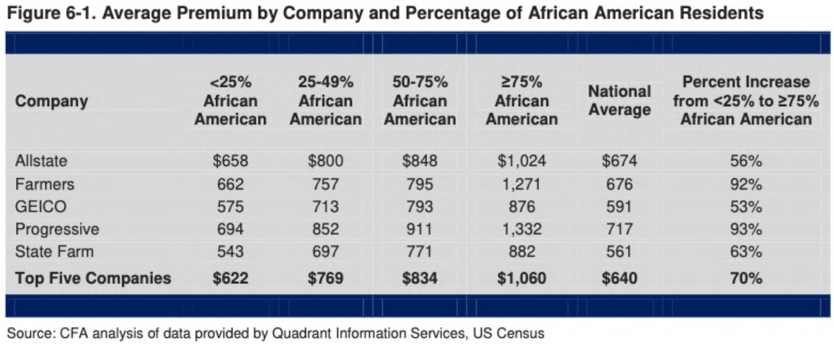


Figure 6-1. Average Premium by Company and Percentage of African American Residents

Company	<25% African American	25-49% African American	50-75% African American	≥75% African American	National Average	Percent Increase from <25% to ≥75% African American
Allstate	$658	$800	$848	$1,024	$674	56%
Farmers	662	757	795	1,271	676	92%
GEICO	575	713	793	876	591	53%
Progressive	694	852	911	1,332	717	93%
State Farm	543	697	771	882	561	63%
Top Five Companies	**$622**	**$769**	**$834**	**$1,060**	**$640**	**70%**

Source: CFA analysis of data provided by Quadrant Information Services, US Census

via https://www.michiganautolaw.com/wp-content/uploads/2017/08/Consumer-Federation-of-America-High-Price-of-Mandatory-Auto-Insurance-in-Predominantly...

Racial Discrimination

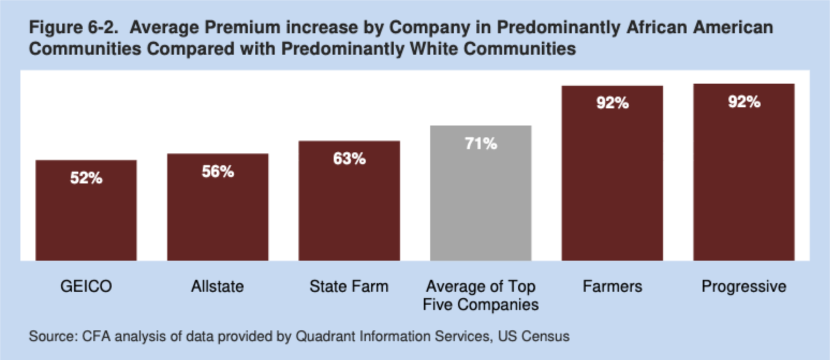


Figure 6-2. Average Premium increase by Company in Predominantly African American Communities Compared with Predominantly White Communities

Source: CFA analysis of data provided by Quadrant Information Services, US Census

The Property Casualty Insurers Association of America responded that "*insurance rates are color-blind and solely based on risk*."

via https://www.pciaa.net/pciwebsite/cms/content/viewpage?sitePageId=43349

Sex and Gender Discrimination

You mean a woman can open it ?

Easily—without a knife blade, a bottle opener, or even a husband! All it takes is a dainty grasp, an easy, two-finger twist—and the catsup is ready to pour.

We call this safe-sealing bottle cap the Alcoa HyTop. It is made of pure, food-loving Alcoa Aluminum. It spins off—and back on again—without muscle power because an exclusive Alcoa process tailors it to each bottle's threads after it is on the bottle. By vacuum sealing both top and sides, the HyTop gives purity a double guard.

You'll recognize the attractive, tractable HyTop when you see it on your grocer's shelf. It's long, it's white, it's grooved—and it's on the most famous and flavorful brands. Put the bottle that wears it in your basket . . . save fumbling, fuming and fingers at opening time with the most cooperative cap in the world—the Alcoa HyTop Closure.

Alcoa Aluminum

ALUMINUM COMPANY OF AMERICA
Pittsburgh, Pa

Is it always illegal to KILL a woman?

PITNEY-BOWES
Postage Meter

show her
it's a man's world

Van Heusen

man's world
ties

For men only! . . . brand new man-talking, power-packed patterns that tell her it's a man's world . . . and make her so happy it is. And man! . . . how that Van Heusen sewmanship makes the fine fabrics hold their shape.
And for Christmas . . . here's the Christmas tie that is *really* different. **$2.00**

Other Van Heusen ties from $1.00 to $2.50

Phillips-Jones Corp., N. Y. 1, N. Y., Makers of Van Heusen Shirts • Sport Shirts • Ties • Pajamas • Handkerchiefs • Collars.

To Gladden Hearts and Lighten Labor

DOWMETAL · · · THE WORLD'S LIGHTEST STRUCTURAL ALLOY

THE DOW CHEMICAL COMPANY, MIDLAND, MICHIGAN

DOW
CHEMICALS INDISPENSABLE TO INDUSTRY

Sex and Gender Discrimination

See slides with life tables per gender (exist since 1720, see Struyck (1912))

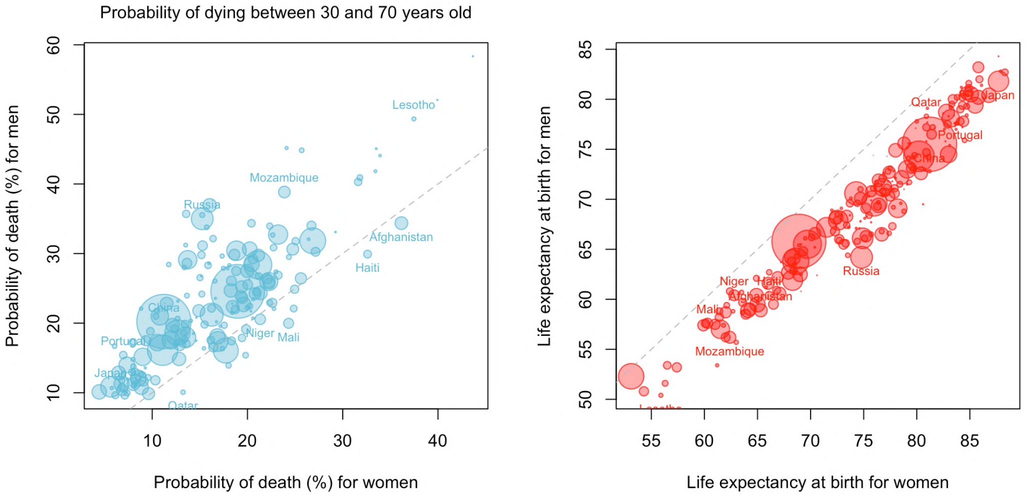
Definition 6.3: Sexism, Merriam-Webster (2022)

Prejudice or discrimination based on sex especially, discrimination against women; *also* behavior, conditions, or attitudes that foster stereotypes of social roles based on sex.

Martin (1977), Hedges (1977) and Myers (1977) in the U.S. In Los Angeles, Department of Water and Power vs. Manhart, the Supreme Court considered a pension system in which female employees made higher contributions than males for the same monthly benefit because of longer life expectancy.

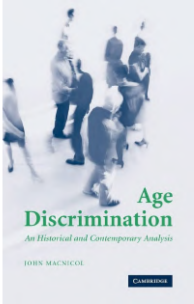
See slides about the "Gender Directive" in Europe (and Thiery and Van Schoubroeck (2006)).

Sex and Gender Discrimination



Data Ortiz-Ospina and Beltekian (2018).

Age-based Discrimination



Age is not a club in which one enters at birth, and it will change with time, Macnicol (2006)

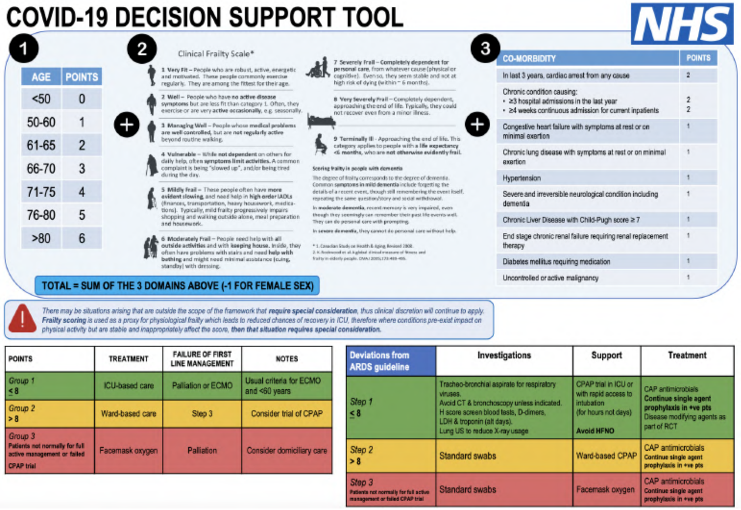
"*If you are not already part of a group disadvantaged by prejudice, just wait a couple of decades—you will be*," Robbins (2015).

Definition 6.4: Ageism, Merriam-Webster (2022)

Prejudice or discrimination against a particular age-group and especially the elderly.

COVID-19 Decision Support Tool used in England, in March 2020, provided by the NHS (National Health System).
https://www.nhsdghandbook.co.uk/wp-content/uploads/2020/04/COVID-Decision-Support-Tool.pdf

Age-based Discrimination



COVID-19 DECISION SUPPORT TOOL

NHS

1

AGE	POINTS
<50	0
50-60	1
61-65	2
66-70	3
71-75	4
76-80	5
>80	6

\+

2

Clinical Frailty Scale*

1 Very Fit – People who are robust, active, energetic and motivated. These people commonly exercise regularly. They are among the fittest for their age.

2 Well – People who have **no active disease symptoms** but are less fit than category 1. Often, they exercise or are very **active occasionally**, e.g. seasonally.

3 Managing Well – People whose **medical problems are well controlled**, but are **not regularly active** beyond routine walking.

4 Vulnerable – While **not dependent** on others for daily help, often **symptoms limit activities.** A common complaint is being "slowed up", and/or **being** tired during the day.

5 Mildly Frail – These people often have **more evident slowing**, and need help in **high order IADLs** (finances, transportation, heavy housework, medications). Typically, mild frailty progressively impairs shopping and walking outside alone, meal preparation and housework.

6 Moderately Frail – People need help with **all outside activities** and with **keeping house.** Inside, they often have problems with stairs and need **help with bathing** and might need minimal assistance (cuing, standby) with dressing.

7 Severely Frail – **Completely dependent for personal care**, from whatever cause (physical or cognitive). Even so, they seem stable and not at high risk of dying (within ~ 6 months).

8 Very Severely Frail – Completely dependent, approaching the end of life. Typically, they could not recover even from a minor illness.

9 Terminally Ill - Approaching the end of life. This category applies to people with a **life expectancy <6 months**, who are **not otherwise evidently frail.**

Scoring frailty in people with dementia

The degree of frailty corresponds to the degree of dementia. Common **symptoms in mild dementia** include forgetting the details of a recent event, though still remembering the event itself, repeating the same question/story and social withdrawal.

In **moderate dementia**, recent memory is very impaired, even though they seemingly can remember their past life events well. They can do personal care with prompting.

In **severe dementia**, they cannot do personal care without help.

* 1. Canadian Study on Health & Aging, Revised 2008.
2. K. Rockwood et al. A global clinical measure of fitness and frailty in elderly people. CMAJ 2005;173:489-495.

\+

3

CO-MORBIDITY	POINTS
In last 3 years, cardiac arrest from any cause	2
Chronic condition causing: • ≥3 hospital admissions in the last year • ≥4 weeks continuous admission for current inpatients	 2 2
Congestive heart failure with symptoms at rest or on minimal exertion	1
Chronic lung disease with symptoms at rest or on minimal exertion	1
Hypertension	1
Severe and irreversible neurological condition including dementia	1
Chronic Liver Disease with Child-Pugh score ≥ 7	1
End stage chronic renal failure requiring renal replacement therapy	1
Diabetes mellitus requiring medication	1
Uncontrolled or active malignancy	1

TOTAL = SUM OF THE 3 DOMAINS ABOVE (-1 FOR FEMALE SEX)

There may be situations arising that are outside the scope of the framework that ***require special consideration****, thus clinical discretion will continue to apply.* ***Frailty scoring*** *is used as a proxy for physiological frailty which leads to reduced chances of recovery in ICU, therefore where conditions pre-exist impact on physical activity but are stable and inappropriately affect the score,* ***then that situation requires special consideration.***

POINTS	TREATMENT	FAILURE OF FIRST LINE MANAGEMENT	NOTES
Group 1 ≤ 8	ICU-based care	Palliation or ECMO	Usual criteria for ECMO and <60 years
Group 2 > 8	Ward-based care	Step 3	Consider trial of CPAP
Group 3 **Patients not normally for full active management or failed CPAP trial**	Facemask oxygen	Palliation	Consider domiciliary care

Deviations from ARDS guideline	Investigations	Support	Treatment
Step 1 ≤ 8	Tracheo-bronchial aspirate for respiratory viruses. Avoid CT & bronchoscopy unless indicated. H score screen blood tests, D-dimers, LDH & troponin (alt days). Lung US to reduce X-ray usage	CPAP trial in ICU or with rapid access to intubation (for hours not days) **Avoid HFNO**	CAP antimicrobials **Continue single agent prophylaxis in +ve pts** Disease modifying agents as part of RCT
Step 2 > 8	Standard swabs	Ward-based CPAP	CAP antimicrobials **Continue single agent prophylaxis in +ve pts**
Step 3 **Patients not normally for full active management or failed CPAP trial**	Standard swabs	Facemask oxygen	CAP antimicrobials **Continue single agent prophylaxis in +ve pts**

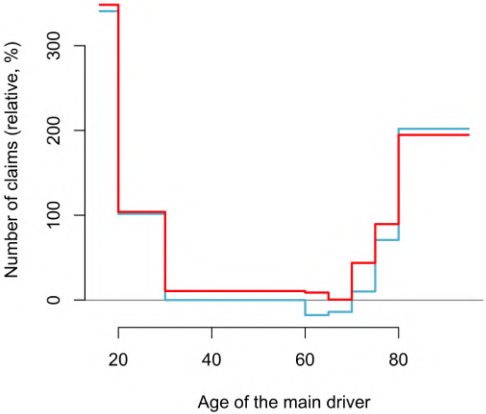
Age-based Discrimination

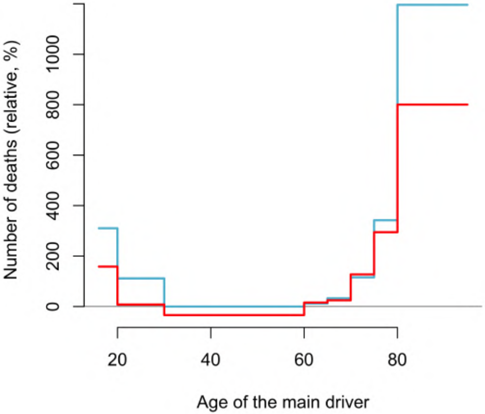
"*on the grounds of age do not constitute discrimination* (...) *if age is a determining factor in the assessment of risk for the service in question and this assessment is based on actuarial principles and relevant and reliable statistical data*," of the European Union (2018)

"*a society that relentlessly discriminates against people because of their age can still treat them equally throughout their lives. Everyone's turn* [to be discriminated against]*is coming*," Gosseries (2014)

Number of crashes (left) and number of fatalities (right), per million miles driven, for both males and females (males in blue and females in red), by driver age. The reference (0) are men aged 30-60 years. The number of accidents is three times higher (+200%) for those over 85, and the number of deaths more than ten times higher (+900%). (data source: Li et al. (2003))

Age-based Discrimination



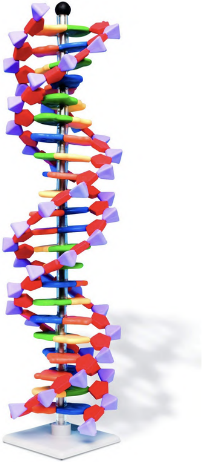


Genetic or Social Identity

Definition 6.5: Genetic discrimination, Ajunwa (2014, 2016)

Genetic discrimination should be defined as when an individual is subjected to negative treatment, not as a result of the individual's physical manifestation of disease or disability, but solely because of the individual's genetic composition

Related to "*genetic determinism*" (as defined in de Melo-Martín (2003) and Harden (2023)) or more recently "*genetic essentialism*" (as in Peters (2014)).



Genetic or Social Identity

According to Rawls (1999), the starting point for each person in society is the result of a social lottery (the political, social, and economic circumstances in which each person is born) and a natural lottery (the biological potentials with which each person is born)

"*Those suffering from disease, a genetic defect, or disability on the basis of a natural lottery should not be penalized in insurance,*" Wortham (1986)

Social identity refers to a person's membership in a social group. The common groups that make up a person's social identity are age, ability, ethnicity, race, gender, sexual orientation, socioeconomic status and religion, as discussed by Tajfel (1978) and Tajfel et al. (1986).

Names, Text and Language

Icelandic surnames are different from most other naming systems in the modern Western world by being patronymic or occasionally matronymic, as mentioned in Willson (2009) and Johannesson (2013): they indicate the father (or mother) of the child and not the historic family lineage. Generally, with few exceptions, a person's last name indicates the first name of their father (patronymic) or in some cases mother (matronymic) in the genitive, followed by –**son** "(son") or –**dóttir** ("daughter").

For instance, in 2017, Iceland's national Women's soccer team players were Agla Maria Alberts**dóttir**, Sigridur Gardars**dóttir**, Ingibjorg Sigurdar**dóttir**, Glodis Viggos**dóttir**, Dagny Brynjars**dóttir**, Sara Bjork Gunnars**dóttir**, Fanndis Fridriks**dóttir**, Hallbera Gisla**dóttir**, Gudbjorg Gunnars**dóttir**, Sif Atla**dóttir** or Gunnhildur Jons**dóttir**. In the national Men's soccer team, players were Hákon Rafn Valdimars**son**, Patrik Gunnars**son**, Höskuldur Gunnlaugs**son**, Júlíus Magnús**son**, Viktor Örlygur Andra**son** or Kristall Máni Inga**son**.

From Gaddis (2017), (data from US Census (2012)

Names, Text and Language

Name	Rank	White (%)	Black (%)	Hispanic (%)
Washington	138	5.2%	89.9%	1.5%
Jefferson	594	18.7%	75.2%	1.6%
Booker	902	30.0%	65.6%	1.5%
Banks	278	41.3%	54.2%	1.5%
Jackson	18	41.9%	53.0%	1.5%
Becker	315	96.4%	0.5%	1.4%
Meyer	163	96.1%	0.5%	1.6%
Walsh	265	95.9%	1.0%	1.4%
Larsen	572	95.6%	0.4%	1.5%
Orozco	690	3.9%	0.1%	95.1%
Velazquez	789	4.0%	0.5%	94.9%
Gonzalez	23	4.8%	0.4%	94.0%
Hernandez	15	4.6%	0.4%	93.8%

Names, Text and Language

As discussed in Riach and Rich (1991) and Rorive (2009), a popular technique to test for discrimination (in a real life context) is to use "*practice testing*" or "*situation testing*". This started probably in the 60's in the U.K., with Daniel et al. (1968)

In France, Top 3 first names by sex and generations in France, according to the origin (Southern Europe or Maghreb) of grandparents, Coulmont and Simon (2019)

	immigrants	children	grandchildren
Southern Europe	José, Antonio, Manuel Maria, Marie, Ana	Jean, David, Alexandre Marie, Sandrine, Sandra	Thomas, Lucas, Enzo Laura, Léa, Camille
Maghreb	Mohamed, Ahmed, Rachid Fatima, Fatiha, Khaduja	Mohamed, Karim, Mehdi Sarah, Nadia, Myriam	Yanis, Nicolas, Mehdi Sarah, Ines, Lina

Names, Text and Language

White	Black	Asian	Hispanic
Cost estimators	Postal service	Manicurists	Drywall installers
Farmers, ranchers	Nursing assistants	Medical scientists	Roofers
Construction	Security guards	Software developers	Carpet installers
Surveying	Probation officers	Computer engineers	Painters and paperhangers
Heavy vehicle	Orderlies aides	Database administrators	Maids-housekeeping cleaners
Property appraisers	Bus drivers	Computer programmers	Construction laborers
Floral designers	Vocational nurses	Chemists	Cement masons
Electrical installers	Barbers	Pharmacists	Brickmasons
Logging workers	Shuttle drivers	Supervisors of personal care	Pipelayers
Brickmasons	Home health aides	Other physicians	Landscaping workers
Aircraft pilots	Social workers	Taxi drivers	Agricultural workers

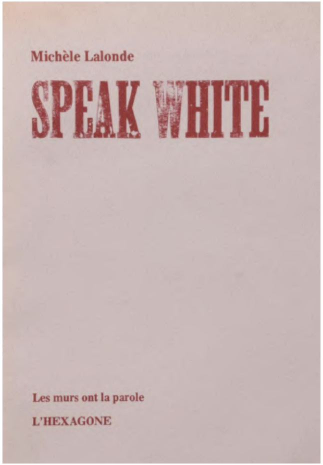
https://flowingdata.com/2024/01/31/occupation-and-race/

Names, Text and Language

Jobs can also be related to gender (see https://translate.google.com/) in Turkish

		2017	2023
o bir öğretmen	›	she is a teacher	he is a teacher
o bir hemşire	›	she is a nurse	she is a nurse
o bir doktor	›	he is a doctor	she is a doctor
o bir Şarkıcı	›	she is a singer	he is a singer
o bir sekreter	›	she is a secretary	she is a secretary
o bir dişçi	›	he is a dentist	he is a dentist
o bir çiçekçi	›	she is a florist	she is a florist
o çalışkan	›	he is hard working	he is hard working
o tembel	›	she is lazy	he is lazy
o güzel	›	she is beautiful	she is beautiful
o çirkin	›	he is ugly	he is ugly

Names, Text and Language

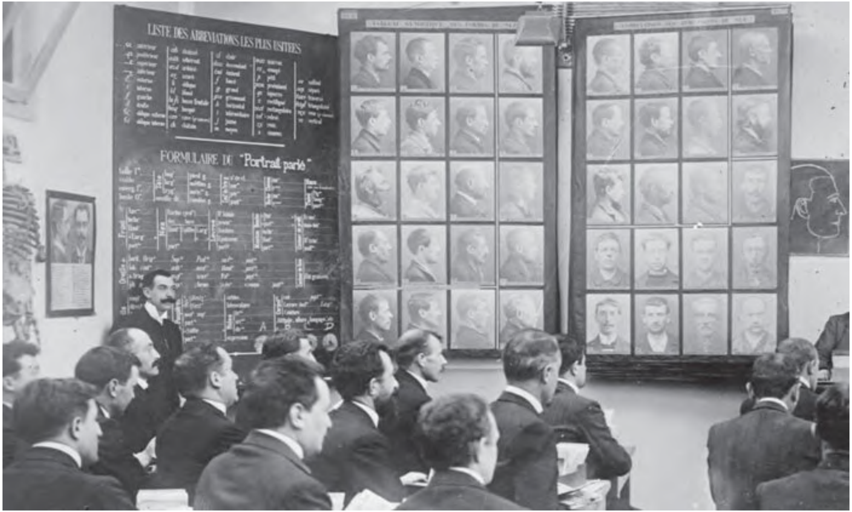


"*Speak White is the protest of white Negroes in America. Language here is the equivalent of colour for the American Negro. The French language is our black colour*,"
Michèle Lalonde, author of the 1968 poem "*Speak White*" (reported by Dostie (1974))

"*phonostyle discrimination*," Léon (1993), or of "*diastratic variation*," with differences between usages by gender, age and social background (in the broad sense), in Gadet (2007).

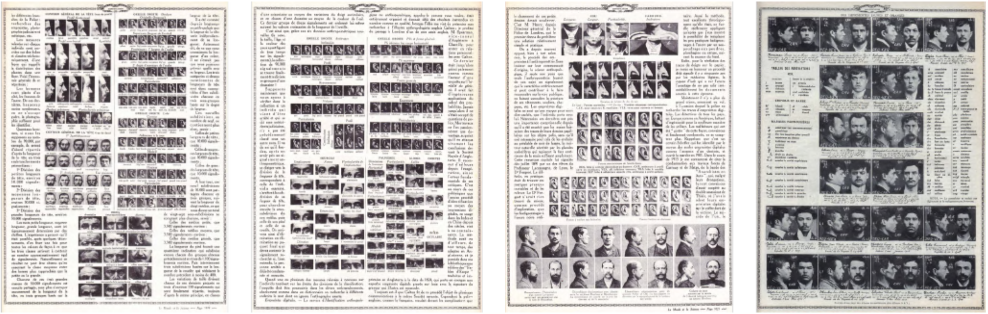
"*linguistic profiling*," (identification of a person's race from the sound of their voice),Squires and Chadwick (2006)

Pictures


LISTE DES ABREVIATIONS LES PLUS USITEES
FORMULAIRE DU "Portrait parlé"

Pictures

More than a century ago, first Lombroso (1876), and then Bertillon and Chervin (1909), laid the foundations of phrenology and the "*born criminal*" theory, which assumes that physical characteristics are correlated with psychological traits and criminal inclinations ("*prima facie*").



Pictures

Faces generated by Karras et al. (2020). Gender and age were provided by gender.toolpie, facelytics, picpurify with a “confidence,” cloud.google, howolddoyoulook and facialage

female, age: 38
female (0.997)
age: 34
joy (74%)

female, age: 23
female (0.989)
age: 20
joy (85%)

male, age: 37
male (0.967)
age: 27
joy (81%)

male, age: 53
male (0.985)
age: 38
joy (73%)

Pictures

Faces generated by Karras et al. (2020). Gender and age were provided by gender.toolpie, facelytics, picpurify with a “confidence,” cloud.google, howolddoyoulook and facialage

female, age: 30
female (0.985)
age: 28
joy (82%)

male, age: 27
male (0.983)
age:33
joy (69%)

male, age: 43
male (0.984)
age: 38
joy (78%)

male, age: 37
male (0.996)
age: 38
joy (56%)

Pictures

Faces generated by Karras et al. (2020). Gender and age were provided by gender.toolpie, facelytics, picpurify with a "confidence," cloud.google, howolddoyoulook and facialage

male, age: 24
male (0.944)
age: 26
joy (70%)

male, age: 33
male (0.981)
age: 32
joy (81%)

male, age: 34
female (0.905)
age: 34
joy (82%)

male, age: 48
male (0.989)
age: 48
joy (83%)

Spatial Information

“*Geographic location is a well-established variable in many lines of insurance*,” Bender et al. (2022).

Spatial Information

"*Geographic information is crucial for estimating the future costs of an insurance contract*," Blier-Wong et al. (2021).

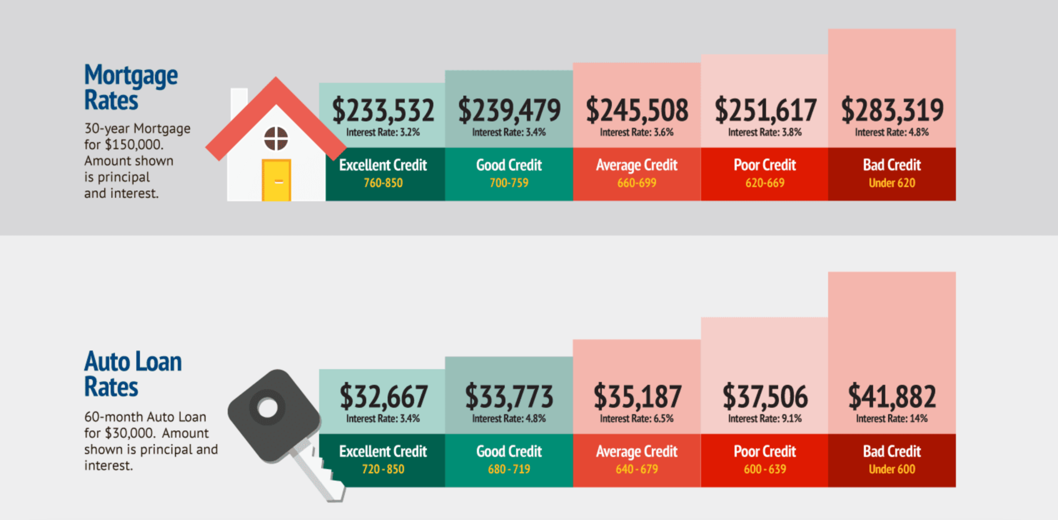
Credit Scoring

"*Credit scoring is one of the most successful applications of statistical and operations research modeling in finance and banking*," Thomas et al. (2002).

In the brief section "*how insurers determine your premium*," in the National Association of Insurance Commissioners (2011, 2022) reports, it is explained that "*most insurers use the information in your credit report to calculate a credit-based insurance score. They do this because studies show a correlation between this score and the likelihood of filing a claim. Credit-based insurance scores are different from other credit scores.*"

As shown in Dean and Nicholas (2018) and Dean et al. (2018), "*credit scores are increasingly used to understand health outcomes*."

Credit Scoring



Mortgage Rates

30-year Mortgage for $150,000. Amount shown is principal and interest.

$233,532	$239,479	$245,508	$251,617	$283,319
Interest Rate: 3.2%	Interest Rate: 3.4%	Interest Rate: 3.6%	Interest Rate: 3.8%	Interest Rate: 4.8%
Excellent Credit	Good Credit	Average Credit	Poor Credit	Bad Credit
760-850	700-759	660-699	620-669	Under 620

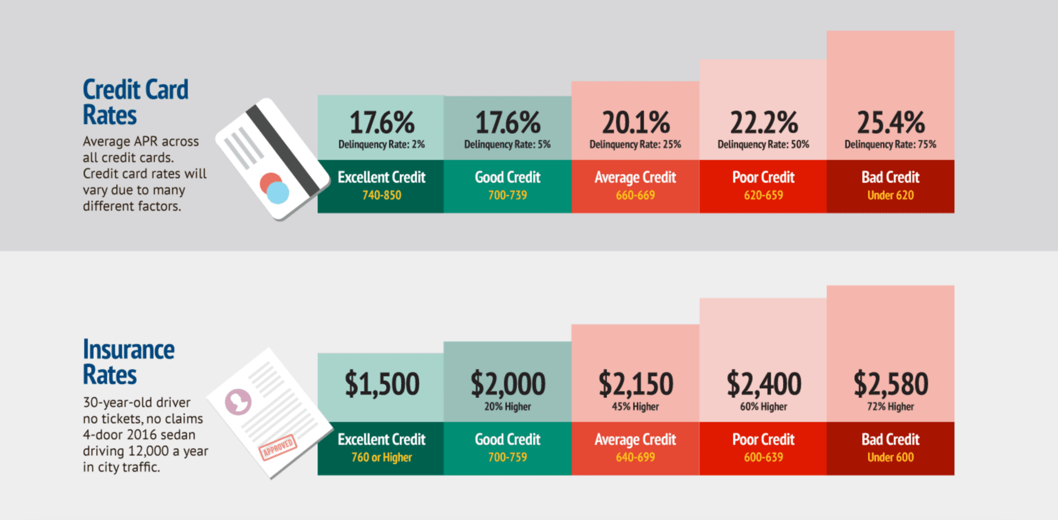
Auto Loan Rates

60-month Auto Loan for $30,000. Amount shown is principal and interest.

$32,667	$33,773	$35,187	$37,506	$41,882
Interest Rate: 3.4%	Interest Rate: 4.8%	Interest Rate: 6.5%	Interest Rate: 9.1%	Interest Rate: 14%
Excellent Credit	Good Credit	Average Credit	Poor Credit	Bad Credit
720-850	680-719	640-679	600-639	Under 600

https://www.incharge.org/debt-relief/credit-counseling/credit-score-and-credit-report/

Credit Scoring



Credit Card Rates

Average APR across all credit cards. Credit card rates will vary due to many different factors.

17.6%	17.6%	20.1%	22.2%	25.4%
Delinquency Rate: 2%	Delinquency Rate: 5%	Delinquency Rate: 25%	Delinquency Rate: 50%	Delinquency Rate: 75%
Excellent Credit	Good Credit	Average Credit	Poor Credit	Bad Credit
740-850	700-739	660-669	620-659	Under 620

Insurance Rates

30-year-old driver no tickets, no claims 4-door 2016 sedan driving 12,000 a year in city traffic.

$1,500	$2,000	$2,150	$2,400	$2,580
	20% Higher	45% Higher	60% Higher	72% Higher
Excellent Credit	Good Credit	Average Credit	Poor Credit	Bad Credit
760 or Higher	700-759	640-699	600-639	Under 600

https://www.incharge.org/debt-relief/credit-counseling/credit-score-and-credit-report/

Networks

"*Network and data analyses compound and reflect discrimination embedded within society*," Bernstein (2007).

"*You apply for a loan and your would-be lender somehow examines the credit ratings of your Facebook friends. If the average credit rating of these members is at least a minimum credit score, the lender continues to process the loan application. Otherwise, the loan application is rejected*," Bhattacharya (2015)

Homophily principle (in the sense of McPherson et al. (2001)), because as popular saying goes, "*birds of a feather flock together*."

"*Insurance companies can base premiums on all insured drivers in your household, including those not related by blood, such as roommates,*" National Association of Insurance Commissioners (2011, 2022)

but there are a few things to bear in mind when using network data...

Networks

Definition 6.6: Network

A (directed) network $\mathcal{G} = (V, E)$, where, as a convention, $V = \{1, \cdots, n\}$ denote either nodes, or vertices, and $E \in \{0, 1\}^{n \times n}$ represents the relationships.

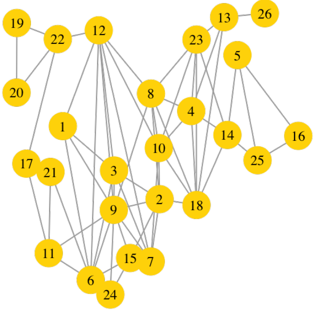
Definition 6.7: Adjacency Matrix

$A_{ij} \in \{0, 1\}$, and $A_{ij} = 1$ if and only if i and j are linked,

$$A_{ij} = \begin{cases} 1 \text{ if } (i, j) \in E \\ 0 \text{ otherwise} \end{cases}$$

There are no self-loops, i.e. $A_{i,i} = 0$. If the matrix is symmetric ($A_{ij} = A_{ji}$), the network is undirected.

Networks



```
> library(igraph)
> g
IGRAPH 8d07103 U--- 26 61 --
+ attr: gender (v/c)
+ edges from 99d7971:
 [1] 1--12 1-- 3 1-- 9 1-- 6 2-- 3
 [6] 2-- 8 2--10 2--18 2-- 7 2--15
[11] 2-- 9 3--12 3-- 9 3-- 6 3-- 7
[16] 4--18 4--10 4-- 8 4--23 4--13
[21] 4--14 5--14 5--25 5--16 6--11
[26] 6--21 6--12 6-- 9 6--15 6--24
[31] 7-- 9 7--12 7--10 8--12 8--23
[36] 8--10 8-- 9 8--18 9--12 9--15
+ ... omitted several edges
```

Networks

Definition 6.8: Neighbors

(immediate) neighbors of node i are

$$N_i = \{j \in V : (i,j) \in E\}.$$

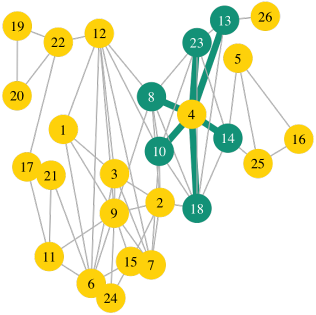
Proposition 6.1: Neighbors

$$N_i = \{j \in V : A_{i,j} > 0\}.$$

Definition 6.9: Extended Neighborhood

(immediate) extended neighbors of node i are

$$\overline{N}_i = N_i \cup \{i\}$$



```
> neighbors(g, 4)
+ 6/26 vertices
[1]  8 10 13 14 18 23
```

Networks

Definition 6.10: Neighbors of neighbors

Neighbors of neighbors of node i are

$$N_i^{(2)} = \{j \in V : (A^2)_{i,j} > 0\}.$$

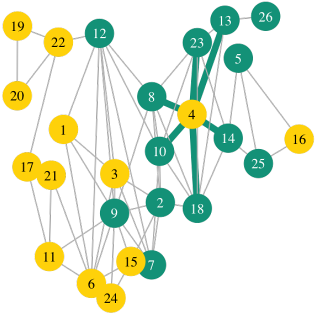
where classically, $(A^2)_{i,j} = \sum_{k=1}^{n} A_{i,k} A_{k,j}$

Definition 6.11: 2-Neighbors

Neighbors of order 2 of node i are

$$\overline{N}_2(i) = \{j \in V : \exists k \leq 2,\ (A^k)_{i,j} > 0\}.$$

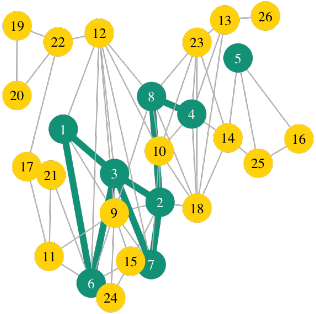
Note that $\overline{N}_2(i) = N_i \cup N_i^{(2)}$.



Networks

Definition 6.12: Subgraph of $\mathcal{G}$

Given two networks $\mathcal{G} = (E, V)$ and $\mathcal{G}' = (E', V')$, $\mathcal{G}'$ is a subgraph of $\mathcal{G}$ (denoted $\mathcal{G}' \subset \mathcal{G}$) if $E' \subset E$ and $V' \subset V$.



Networks

> **Induced subgraph**
> an induced subgraph of a graph is another graph, formed from a subset of the vertices of the graph and all of the edges, from the original graph, connecting pairs of vertices in that subset W

Definition 6.13: Induced subgraph of $\mathcal{G}$

Given a network $\mathcal{G} = (V, E)$ and a subset of vertices $V' \subset V$. The induced subgraph $\mathcal{G}_{V'} = (V', E')$ is the graph whose vertex set is V' and whose edge set consists of all of the edges in E that have both endpoints in V' (denoted E').

Networks

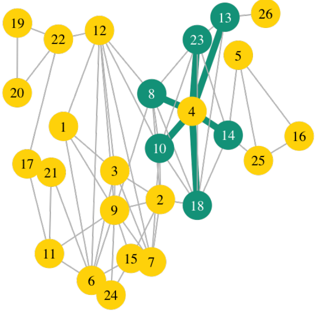
Set $E_i = \overline{N}_i$ and

$$V_i = \{(i,j) \in E, \text{ where } j \in N_i\}.$$

Definition 6.14: Induced subgraph of neighbors

Given a node i in a network (E,V), the induced subgraph of node i is $\mathcal{G}_{\overline{N}_i}$, also denoted $\mathcal{G}_i = (E_i, V_i)$.

E.g. $\mathcal{G}_4 = (E_4, V_4)$



Networks

Definition 6.15: Degrees

Row i contains list of vertices connected to vertex i,

$$d_i = \sum_{j=1}^{n} A_{i,j} = \boldsymbol{A}_{i,\cdot}^{\top}\mathbf{1} = \#N_i.$$

Let $\boldsymbol{d} = (d_i)$ denote the vector of degrees, and $\boldsymbol{D} = \text{diag}(\boldsymbol{d})$.

Definition 6.16: Normalized Adjacency Matrix

$\boldsymbol{A}_0 = \boldsymbol{D}^{-1}\boldsymbol{A} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$ is the normalized adjacency matrix.

Networks

(for directed networks, this corresponds to “out degrees”)

Definition 6.17: Walk

A walk from node i to node j is a sequence of edges, (i, v_1), (v_1, v_2), (v_2, v_3), $\cdots$, (v_{k-1}, v_k), (v_k, j)

Definition 6.18: Path

A walk where all the vertices are distinct is a path.

Definition 6.19: Connected graph

There exists a path that connects very pair of nodes in the network.

Networks

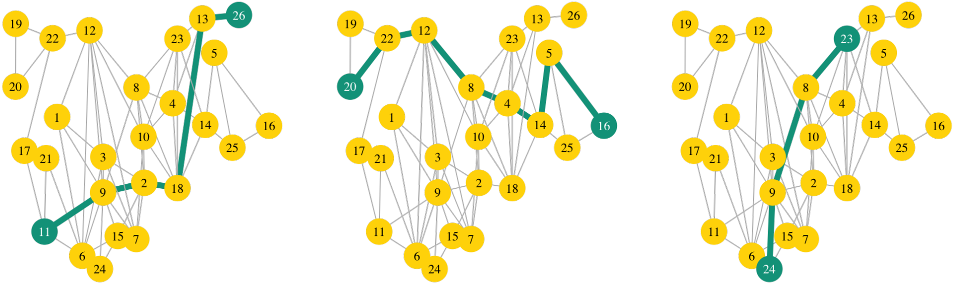
Definition 6.20: Shortest path

A geodesic between nodes i and j is a "shortest path" (i.e., with minimum number of edges) between these nodes. $d_{sp}(i,j)$ is the distance between nodes i and j.

Conveniently suppose that the set of vertices V is $\mathcal{I}_n = \{1, 2, \cdots, n\}$.

Networks

Exemples of (shortest) paths.



```
> shortest_paths(g,from=11,to=26)
$vpath
$vpath[[1]]
+ 6/26 vertices,
[1] 11  9  2 18 13 26
```

```
> shortest_paths(g,from=20,to=16)
$vpath
$vpath[[1]]
+ 8/26 vertices,
[1] 20 22 12  8  4 14  5 16
```

Networks

Definition 6.21: Random walk

Random walk with transition matrix $\boldsymbol{P} = \text{diag}(\boldsymbol{d})^{-1}\boldsymbol{A}$.

Let x_t denote the node reached at time t, and $\boldsymbol{p}(t) \in \mathcal{S}_n \subset \mathbb{R}_+^n$ the probability vector associated with $\{x_t = i\}$. Then

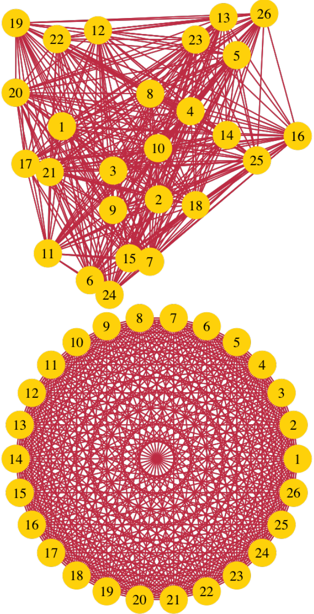
$$\boldsymbol{p}_{t+1} = \text{diag}(\boldsymbol{d})^{-1}\boldsymbol{A}\boldsymbol{p}_t.$$

The stationary distribution is $\boldsymbol{\pi} = \lim_{t\to\infty} \boldsymbol{p}_t$.

Proposition 6.2: Unique Stationnary Distribution

$\boldsymbol{\pi}$ exists and is unique if the network is connected and aperiodic.

Random Graphs: Regular Graph (Dirac)

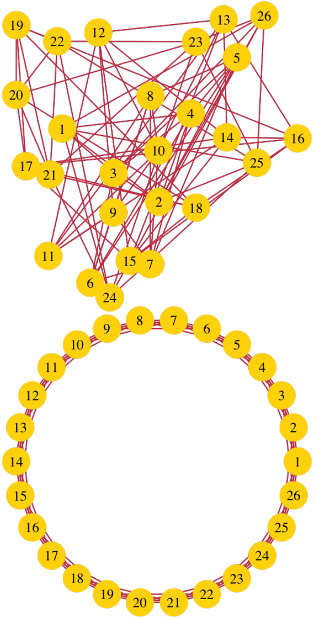


Definition 6.22: Complete graph

A complete graph is a simple undirected graph in which every pair of distinct vertices is connected

Here $d_i = (n-1), \forall i \in \{1, \cdots, n\}$

Random Graphs: Regular Graph (Dirac)

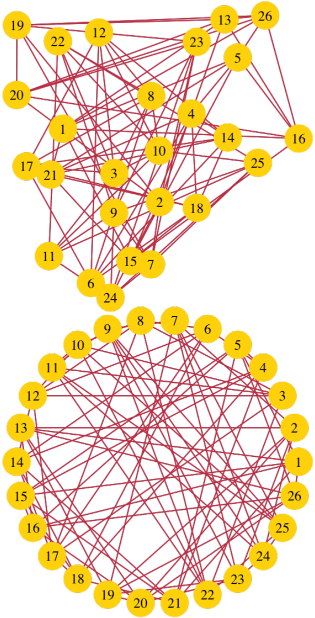


Definition 6.23: (r) Regular graph

a regular graph is a graph where each vertex has the same number of neighbors; i.e. every vertex has the same degree.

Here $d_i = r$, $\forall i \in \{1, \cdots, n\}$

Random Graphs: Regular Graph (Dirac)



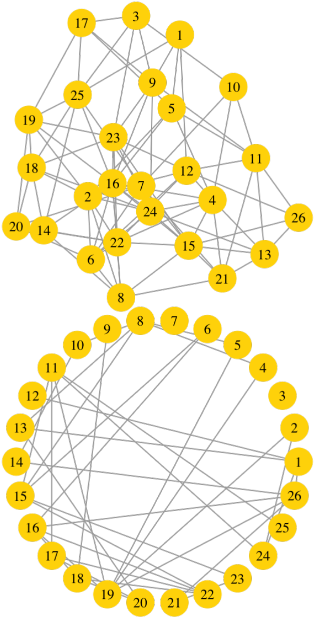
Definition 6.24: (r) Regular graph

a regular graph is a graph where each vertex has the same number of neighbors; i.e. every vertex has the same degree.

Here $d_i = r, \forall i \in \{1, \cdots, n\}$

See Bollobás (1998) for regular random graphs

Random Graphs: Erdös-Rényi (Binomial-Poisson)



From Gilbert (1959), $d_i \leftarrow D_i \sim \mathcal{B}(n-1,p)$

Definition 6.25: Erdös-Rényi graph

$A_{i,j} = A_{j,i} \leftarrow X_{i,j}$ where $X_{i,j}$ are i.i.d. $\mathcal{B}(p)$ random variables (each edge has a fixed probability of being present or absent, independently of the other edges).

$$\mathbb{P}(D_i = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k},$$

$$\mathbb{P}(D_i = k) \rightarrow \frac{(np)^k \mathrm{e}^{-np}}{k!} \quad \text{as } n \rightarrow \infty \text{ and } np = \text{constant}.$$

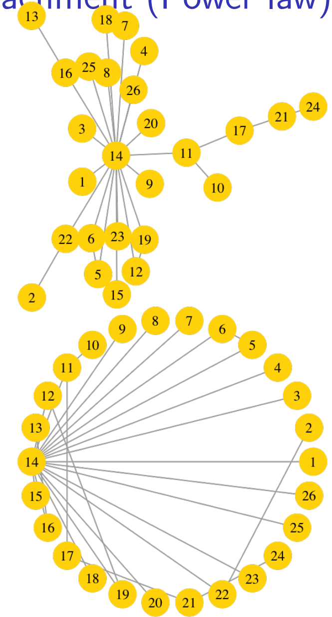
Random Graphs: Barabási–Albert, preferential attachment (Power law)

From Barabási and Albert (1999),

Definition 6.26: Barabási–Albert

Let $m \geq 1$. The network initializes with a network of $m_0 \geq m$ nodes. At each step, add 1 new node, then sample m existing vertices from the network, with a probability that is proportional to the number of links that the existing nodes already have.

(heavily linked nodes ("hubs") tend to quickly accumulate even more links)



Networks Generation

> **Havel–Hakimi algorithm**
> The Havel–Hakimi algorithm is an algorithm in graph theory solving the graph realization problem. That is, it answers the following question: *Given a finite list of nonnegative integers in non-increasing order, is there a simple graph such that its degree sequence is exactly this list?* A simple graph contains no double edges or loops. W

Suppose that the sum of degrees is even, random networks can then be generated with the algorithm of Havel (1955) and Hakimi (1962) (see also Viger and Latapy (2005)).

```
> degs = sort(round(1+rexp(100, 1/10)), decreasing=TRUE)
> if (sum(degs) %% 2 != 0) {
+   degs[1] <- degs[1] + 1
+ }
> g = realize_degseq(degs, allowed.edge.types = "all")
```

Networks Centrality

Definition 6.27: Degree Centrality

Degree centrality of node i is $c_d(i) = d_i$, and $\boldsymbol{c}_d = \boldsymbol{d}$.

Definition 6.28: Eigenvector Centrality

Eigenvector centrality of node i is solution of $c_e(i) = \dfrac{1}{\lambda}\sum_{j=1}^{n} A_{i,j} c_e(j)$, or $\boldsymbol{c}_e = \dfrac{1}{\lambda}\boldsymbol{A}^\top \boldsymbol{c}_e$, for some fixed constant $\lambda > 0$.

Equation $\boldsymbol{A}^\top \boldsymbol{c}_e = \lambda \boldsymbol{c}_e$ means that $\boldsymbol{c}_e$ is some eigenvector associates with $\boldsymbol{A}^\top$ (or $\boldsymbol{A}$ if $\mathcal{G}$ is undirected).

Networks Centrality

Definition 6.29: PageRank Centrality

PageRank centrality of node i is solution of $c_p(i) = \alpha \sum_{j=1}^{n} A_{i,j} \frac{c_p(j)}{d_j} + \beta$, or $\boldsymbol{c}_p = \alpha \boldsymbol{A}^\top \boldsymbol{D}^{-1} \boldsymbol{c}_p + \beta \boldsymbol{1}$, for some fixed constant α and β.

```
> eigen_centrality(g)
 [1] 0.527 0.821 0.732 0.544 0.060
 [6] 0.702 0.671 0.833 1.000 0.788
[11] 0.310 0.864 0.286 0.298 0.458
[16] 0.019 0.075 0.657 0.030 0.030
[21] 0.162 0.160 0.544 0.345 0.060
[26] 0.046
> eigen(t(get.adjacency(g1)))
    $vectors[,1]
```

```
> page_rank(g)
 [1] 0.030 0.049 0.043 0.045 0.033
 [6] 0.060 0.036 0.049 0.070 0.049
[11] 0.036 0.058 0.036 0.044 0.031
[16] 0.024 0.023 0.052 0.026 0.026
[21] 0.020 0.044 0.045 0.025 0.033
[26] 0.014
>
>
```

Networks Centrality

Definition 6.30: Closeness Centrality

Closeness centrality of node i is $c_c(i) = \dfrac{n}{\displaystyle\sum_{j=1}^{n} d_{sp}(i,j)}$.

```
> closeness(g1)
 [1] 0.014 0.018 0.017 0.017 0.011
 [6] 0.016 0.017 0.019 0.019 0.019
[11] 0.014 0.019 0.013 0.014 0.015
[16] 0.009 0.012 0.018 0.011 0.011
[21] 0.012 0.014 0.017 0.014 0.011
[26] 0.010
```

```
>
```

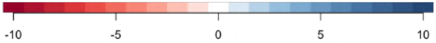
See Freeman et al. (1979)

Networks Centrality

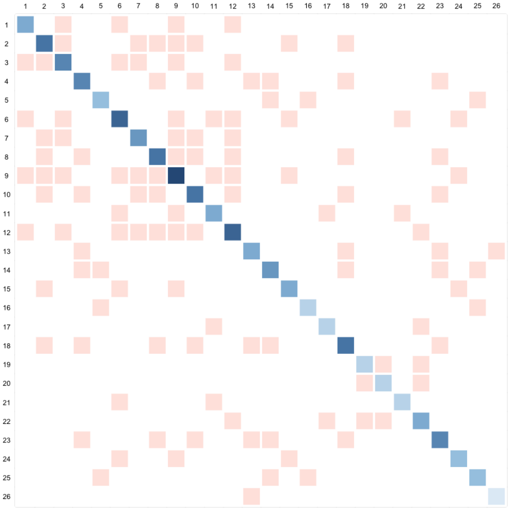
Definition 6.31: Laplacian

$$\boldsymbol{L} = \text{diag}(\boldsymbol{d}) - \boldsymbol{A},$$

$$L_{i,j} := \begin{cases} d_i & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } (i,j) \in E \\ 0 & \text{otherwise,} \end{cases}$$



```
> L = laplacian_matrix(g)
```



Networks Centrality

Proposition 6.3: Alternative expression for $\boldsymbol{L}$

Let $e_i = (0, \cdots, 0, 1, , 0 \cdots, 0) \in \{0,1\}^n$,

$$\boldsymbol{L} = \sum_{(i,j)\in E} \overset{\ell_{i,j}}{\overbrace{(e_i - e_j)(e_i - e_j)^\top}}$$

$$\ell_{i,j}\ n \times n \text{ matrix}, \ell_{i,j} = \begin{array}{cc} \begin{array}{ccccc} & i & & j & \end{array} & \\ \begin{pmatrix} (0) & \vdots & (0) & \vdots & (0) \\ \cdots & 1 & \cdots & -1 & \cdots \\ (0) & \vdots & (0) & \vdots & (0) \\ \cdots & -1 & \cdots & 1 & \cdots \\ (0) & \vdots & (0) & \vdots & (0) \end{pmatrix} & \begin{array}{c} \\ i \\ \\ j \\ \\ \end{array} \end{array}$$

Sidenote on quadratic forms

Definition 6.32: Normalized Laplacian Matrix

$\boldsymbol{L}_0 = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2} = \mathbb{I} - \boldsymbol{A}_0$ is the normalized adjacency matrix.

$\boldsymbol{L}$ and $\boldsymbol{L}_0$ are symmetric positive semidefinite matrices.

Proposition 6.4: Laplacian and quadratic form

$\boldsymbol{L} = \text{diag}(\boldsymbol{d}) - \boldsymbol{A}$,

$$\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x} = \frac{1}{2}\sum_{(i,j)\in E}(x_i - x_j)^2 = \frac{1}{2}\sum_{i,j=1}^{n} A_{i,j}(x_i - x_j)^2$$

Sidenote on quadratic forms

Proof.

$$\sum_{i,j=1}^{n} A_{i,j}(x_i - x_j)^2 = \sum_{i,j=1}^{n} A_{i,j}(x_i^2 - 2x_ix_j + x_j^2) = \sum_{i=1}^{n} \overbrace{A_{i,\cdot}}^{A_{i,\cdot} = \sum_{j=1}^{n} A_{i,j} = d_i} x_i^2 - \sum_{i,j=1}^{n} 2A_{i,j}x_ix_j + \sum_{j=1}^{n} A_{\cdot,j}x_j^2$$

$$\sum_{i,j=1}^{n} A_{i,j}(x_i - x_j)^2 = 2\underbrace{\sum_{i=1}^{n} d_ix_i^2}_{\boldsymbol{x}^\top\boldsymbol{D}\boldsymbol{x}} + 2\underbrace{\sum_{i,j=1}^{n} A_{i,j}x_ix_j}_{\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}} = 2\boldsymbol{x}^\top(\boldsymbol{D} - \boldsymbol{A})\boldsymbol{x} = 2\boldsymbol{x}^\top\boldsymbol{L}\boldsymbol{x}$$

□

Since $\boldsymbol{x}^\top\boldsymbol{L}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x}$, $\boldsymbol{L}$ is symmetric positive semidefinite matrices.

Sidenote on quadratic forms

Let $\lambda_n \geq \lambda_{n-1} \geq \cdots \geq \lambda_2 \geq \lambda_1 \geq 0$ denote $\boldsymbol{L}$'s eigenvalues.

Proposition 6.5: Spectrum of $\boldsymbol{L}$ and λ_1

The n-vector of one's, $\mathbf{1}$, is an eigenvector of $\boldsymbol{L}$ associated with eigenvalue $\lambda_1 = 0$.

Proof.

$$\boldsymbol{L}\mathbf{1} = \sum_{(i,j)\in E} (\mathbf{1}_i - \mathbf{1}_j) \underbrace{(\mathbf{1}_i - \mathbf{1}_j)^\top \mathbf{1}}_{0} = \sum_{(i,j)\in E} (\mathbf{1}_i - \mathbf{1}_j)0 = 0.$$

$\square$

Proposition 6.6: Spectrum of $\boldsymbol{L}$ and λ_2

Network $\mathcal{G} = (E, V)$ is disconnected in two groups if and only if $\lambda_2 = 0$.

Sidenote on quadratic forms

Proposition 6.7: Spectrum of L and λ_2

Network $\mathcal{G} = (E, V)$ is disconnected in at least k groups if and only if $\lambda_k = 0$.

Proposition 6.8: Laplacian and quadratic form

$\boldsymbol{L}_0 = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$,

$$\boldsymbol{x}^\top \boldsymbol{L}_0 \boldsymbol{x} = \frac{1}{2}\sum_{(i,j)\in E}\left(\frac{x_i}{d_i} - \frac{x_j}{d_j}\right)^2 = \frac{1}{2}\sum_{i,j=1}^{n} A_{i,j}\left(\frac{x_i}{d_i} - \frac{x_j}{d_j}\right)^2$$

Networks Homophily and Assortative Mixing

Definition 6.33: Homophily

Homophily is the tendency of individuals to form relations with others similar to them.

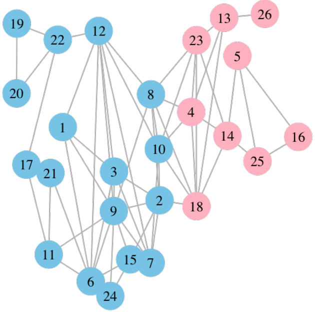
Definition 6.34: Community, Newman and Girvan (2004), Newman (2018)

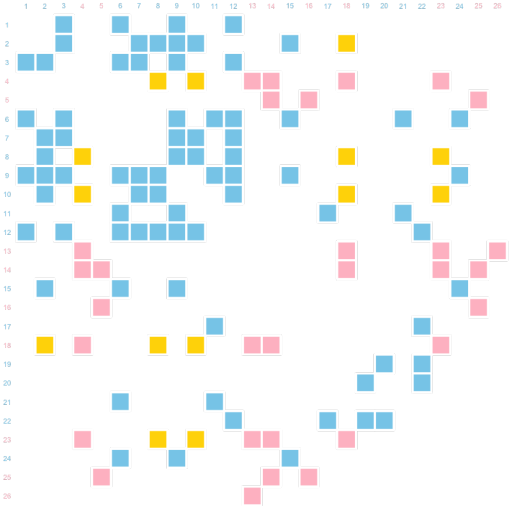
Communities are partitions of nodes.

The total number of edges that run between nodes of the same type is

$$\sum_{(i,j)\in E}\delta(c_i,c_j)=\frac{1}{2}\sum_{i,j}A_{i,j}\delta(c_i,c_j)\text{ where }\delta(c_i,c_j)=\begin{cases}1 \text{ if } c_i=c_j\\0 \text{ otherwise.}\end{cases}$$

Networks Homophily and Assortative Mixing





Networks Homophily and Assortative Mixing

The expected number of edges between nodes if edges are placed at random is

$$\frac{1}{2}\sum_{i,j}\frac{d_i d_j}{2m}\delta(c_i, c_j)$$

and the difference between the actual and expected number of edges in the network that join nodes of the same type is mQ where Q is the modularity measure,

Definition 6.35: Modularity measure, Newman (2003)

The modularity measure of a partition (c) of a network (E, V) is

$$Q = \frac{1}{2m}\sum_{i,j}\overbrace{\left(A_{i,j} - \frac{d_i d_j}{2m}\right)}^{B_{i,j}}\delta(c_i, c_j)$$

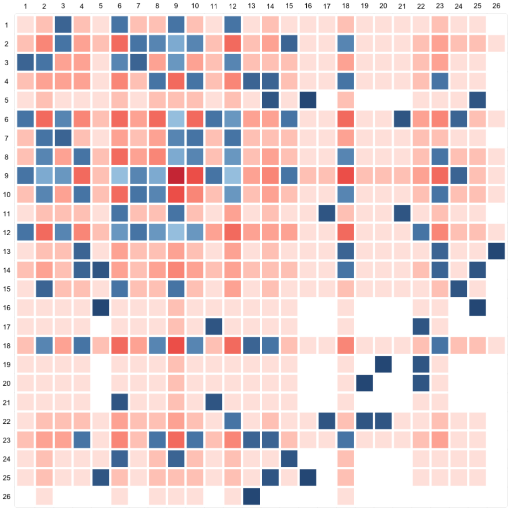
where m is the total number of links. $\boldsymbol{B}$ is coined “modularity matrix”.

Networks Homophily and Assortative Mixing

A network is said to be assortative if a significant portion of its links are between nodes that belong to the same community

$$\boldsymbol{B} = \boldsymbol{A} - \frac{\boldsymbol{d}^\top \boldsymbol{d}}{2m}, \text{ i.e. } B_{i,j} = A_{i,j} - \frac{d_i d_j}{2m},$$

```
> A = get.adjacency(g)
> m = sum(A)/2
> d = apply(A,1,sum)
> B = A - d %*% t(d)/(2*m)
```



Networks Homophily and Assortative Mixing

Since $B_{i,j} = A_{i,j} - \dfrac{d_i d_j}{2m}$, and

$$\sum_{i=1}^{n} B_{i,j} = \overset{d_j}{\boxed{\sum_{i=1}^{n} A_{i,j}}} - \frac{d_j}{2m}\overset{2m}{\boxed{\sum_{i=1}^{n} d_i}} = d_j - \frac{d_j}{2m}2m = 0$$

$$\sum_{j=1}^{n} B_{i,j} = \underset{d_i}{\boxed{\sum_{j=1}^{n} A_{i,j}}} - \frac{d_i}{2m}\underset{2m}{\boxed{\sum_{j=1}^{n} d_j}} = d_i - \frac{d_i}{2m}2m = 0$$

In the case where were two communities, A and B, set

$$s_i^A = \begin{cases} +1 \text{ if } i \in A \\ -1 \text{ if } i \in B \end{cases} \quad \text{and } s_i^B = -s_i^A = \begin{cases} +1 \text{ if } i \in B \\ -1 \text{ if } i \in A \end{cases} .$$

Networks Homophily and Assortative Mixing

then

$$\delta(c_i, c_j) = \frac{1}{2}(s_i s_j + 1)$$

so that

$$Q = \frac{1}{2m}\sum_{i,j} B_{i,j}\delta(c_i, c_j) = \frac{1}{4m}\sum_{i,j} B_{i,j}(s_i s_j + 1) = \frac{1}{4m}\sum_{i,j} B_{i,j} s_i s_j = \frac{1}{4m}\boldsymbol{s}^\top \boldsymbol{B}\boldsymbol{s},$$

(whatever the reference group).

Proposition 6.9

The modularity measure can be written

$$Q = \frac{1}{4m}\boldsymbol{s}^\top \boldsymbol{B}\boldsymbol{s},\ \text{where } \boldsymbol{s} = \boldsymbol{1}_A - \boldsymbol{1}_B,\ \text{i.e. } s_i^A = \begin{cases} +1 \text{ if } i \in A \\ -1 \text{ if } i \in B \end{cases}$$

Networks Homophily and Assortative Mixing

When is Q maximal (in $\boldsymbol{s}$) ? see "modularity maximization," in Newman (2012)

Recall that $\boldsymbol{s} \in \{\pm 1\}^n$, so that $\boldsymbol{s}^\top \boldsymbol{s} = n$. Our problem is

$$\max_{\boldsymbol{s} \in \{\pm 1\}^n} \{\boldsymbol{s}^\top \boldsymbol{B} \boldsymbol{s}\}, \text{ subject to } \boldsymbol{s}^\top \boldsymbol{s} = n.$$

Using the Legrangian, our optimization problem has the following first order condition

$$\frac{\partial}{\partial \boldsymbol{s}} (\boldsymbol{s}^\top \boldsymbol{B} \boldsymbol{s} + \lambda(n - \boldsymbol{s}^\top \boldsymbol{s})) = \boldsymbol{0}$$

i.e.

$$\frac{\partial}{\partial s_k} \Big(\sum_{i,j} B_{i,j} s_i s_j + \lambda \Big(n - \sum_j s_j^2 \Big) \Big) = \sum_{i=1}^n B_{i,k} s_i - \lambda s_k = 0, \ \forall k$$

or, with matrix notations, $\boldsymbol{B} \boldsymbol{s}_\star = \lambda \boldsymbol{s}_\star$, i.e. $\boldsymbol{s}_\star$ is an eigenvector of $\boldsymbol{B}$. Thus

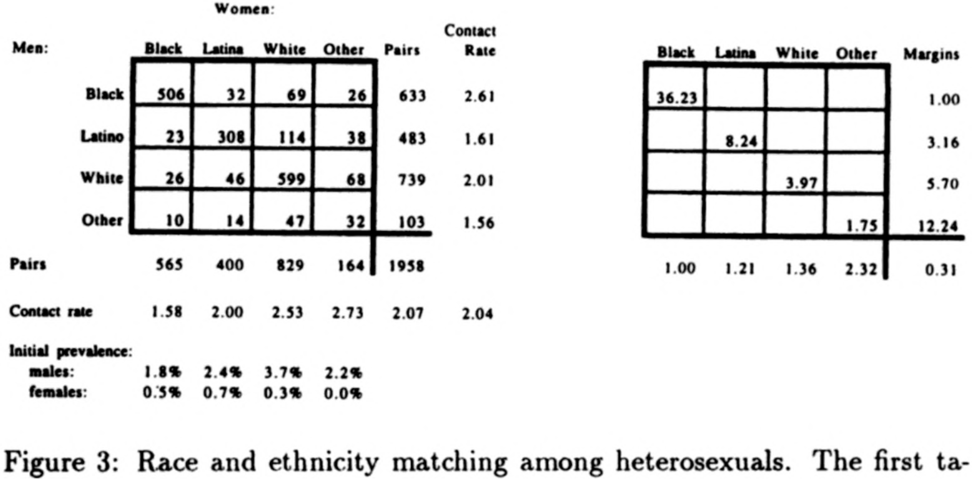
$$Q^\star = \frac{1}{4m} \boldsymbol{s}_\star^\top \boldsymbol{B} \boldsymbol{s}_\star = \frac{1}{4m} \boldsymbol{s}_\star^\top \lambda \boldsymbol{s}_\star = \frac{n}{4m} \lambda.$$

Networks Homophily and Assortative Mixing

```
> modularity(g1, 1+(V(g1)$gender=="female"))
[1] 0.3078474
```

Networks, without networks

Following Morris (1995), from AMEN (AIDS in Multi-Ethnic Neighborhoods) Study



Men:	Women: Black	Latina	White	Other	Pairs	Contact Rate
Black	506	32	69	26	633	2.61
Latino	23	308	114	38	483	1.61
White	26	46	599	68	739	2.01
Other	10	14	47	32	103	1.56
Pairs	565	400	829	164	1958	
Contact rate	1.58	2.00	2.53	2.73	2.07	2.04
Initial prevalence:						
males:	1.8%	2.4%	3.7%	2.2%		
females:	0.5%	0.7%	0.3%	0.0%		

Black	Latina	White	Other	Margins
36.23				1.00
	8.24			3.16
		3.97		5.70
			1.75	12.24
1.00	1.21	1.36	2.32	0.31

Figure 3: Race and ethnicity matching among heterosexuals. The first ta-

Networks, without networks

Here we have 4 sensitive groups, on a bipartite network (heterosexual relationships)
Consider a discrete copula representation

$e_{i,j}$	Black	Hispanic	White	Other	a_i
Black	0.258	0.016	0.035	0.013	0.323
Hispanic	0.012	0.157	0.058	0.019	0.247
White	0.013	0.023	0.306	0.035	0.377
Other	0.005	0.007	0.024	0.016	0.053
b_j	0.289	0.204	0.423	0.084	

$$a_i = \sum_{j=1}^{K} e_{i,j} = \boldsymbol{E}_{i,\cdot}^\top \mathbf{1} \text{ and } b_j = \sum_{i=1}^{K} e_{i,j} = \boldsymbol{E}_{\cdot,j}^\top \mathbf{1}$$

Further

$$\boldsymbol{a}^\top \boldsymbol{b} = \sum_{k=1}^{K} a_k b_k = \sum_{k=1}^{K} \Big(\sum_{i=1}^{K} e_{i,k} \Big) \Big(\sum_{j=1}^{K} e_{k,j} \Big) = \sum_{i,j} (\boldsymbol{E}^2)_{i,j} = \|\boldsymbol{E}^2\|$$

Networks, without networks

Thus, we recover the coefficient introduced in Gupta et al. (1989),

Definition 6.36: Assortativity coefficient, Gupta et al. (1989)

With K communities

$$r=\frac{\displaystyle\sum_{k=1}^{K}e_{k,k}-\sum_{k=1}^{K}a_kb_k}{\displaystyle 1-\sum_{k=1}^{K}a_kb_k}=\frac{\text{trace}[\boldsymbol{E}]-\|\boldsymbol{E}^2\|}{1-\|\boldsymbol{E}^2\|}$$

More generally, when dealing with data with a network topology, we should be careful...

Statistics with a Network Topology

sample data $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{S})$

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{\mathbf{1}^\top \boldsymbol{y}}{\mathbf{1}^\top \mathbf{1}}$$

→ sample version of $\mathbb{E}[Y]$.

$$\overline{y}_s = \frac{1}{n_s}\sum_{j=1}^{n} \mathbf{1}(s_j = s) y_j = \frac{\mathbf{1}_s^\top \boldsymbol{y}}{\mathbf{1}_s^\top \mathbf{1}}$$

→ sample version of $\mathbb{E}[Y|S = s]$.

network data $(V, E, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{S})$

for a node $i \in V$,

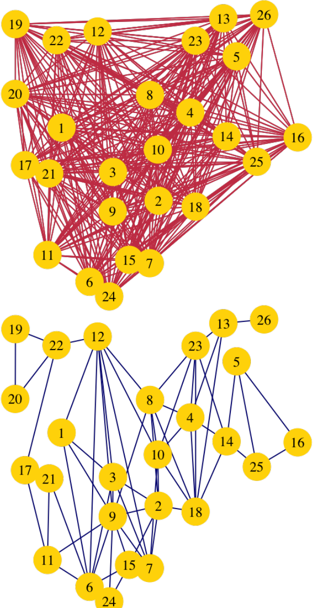
$$\overline{y}(i) = \frac{1}{d_i}\sum_{j\in N_i} y_j = \frac{1}{d_i}\sum_{i=1}^{n} A_{i,j} y_j = \frac{\boldsymbol{A}_{i\cdot}^\top \boldsymbol{y}}{\boldsymbol{A}_{i\cdot}^\top \mathbf{1}}$$

→ sample version of $E_i[Y]$.

$$\overline{y}_s(i) = \frac{1}{d_{i:s}}\sum_{j\in N_i} \mathbf{1}(s_j = s) y_j = \frac{(\boldsymbol{A}_{i\cdot} \cdot \mathbf{1}_s)^\top \boldsymbol{y}}{(\boldsymbol{A}_{i\cdot} \cdot \mathbf{1}_s)^\top \mathbf{1}}$$

→ sample version of $E_i[Y|S = s]$.
where $\boldsymbol{a} \cdot \boldsymbol{b}$ is the element-wise product.

Statistics with a Network Topology



Given sample $\{x_1, \cdots, x_n\}$, the empirical variance,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \text{, where } \overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

could be written as a *U*-stat, Lee (2019)

$$\sigma^2 = \frac{1}{2n^2}\boxed{\sum_{i,j=1}^{n}(x_i - x_j)^2}.$$

On a network, with adjacency matrix *A*,

$$\sigma^2_{\mathcal{G}} = \frac{1}{4e}\boxed{\sum_{i,j=1}^{n}A_{i,j}(x_i - x_j)^2}\text{, where } 2e = \sum_{i,j=1}^{n}A_{i,j}$$

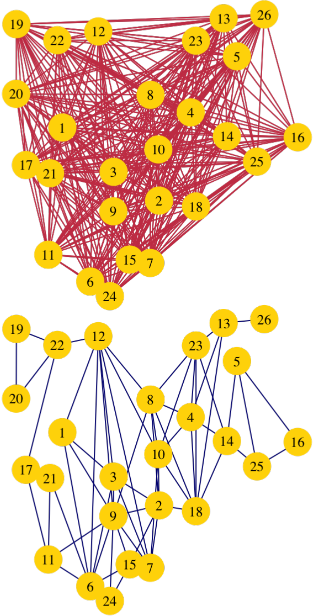
Statistics with a Network Topology

Given sample $\{(x_1,y_1),\cdots,(x_n,y_n)\}$, the empirical covariance could be written as a U-stat,

$$cv = \frac{1}{2n^2}\sum_{i,j=1}^{n}(x_i-x_j)(y_i-y_j)$$

and if observations are nodes on a network

$$cv_{\mathcal{G}} = \frac{1}{4e}\sum_{i,j=1}^{n}A_{i,j}(x_i-x_j)(y_i-y_j)$$

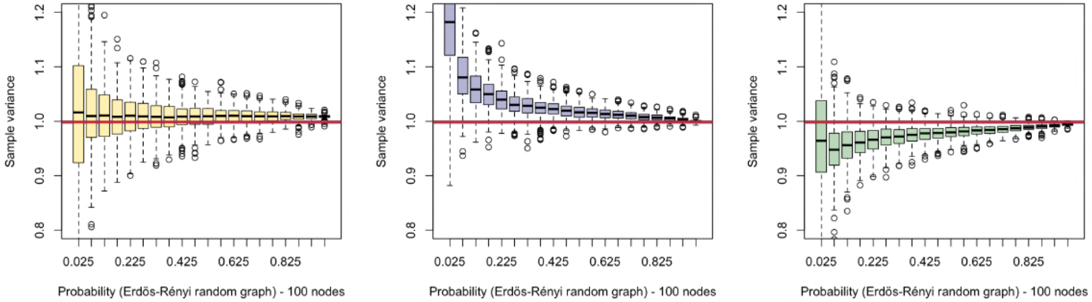


Statistics with a Network Topology

If x is independent of the topology of the network (summarized by A),

$$\boxed{\sigma^2} = \frac{1}{2n^2}\boxed{\sum_{i,j=1}^{n}(x_i - x_j)^2} \approx \frac{1}{4e}\boxed{\sum_{i,j=1}^{n}A_{i,j}\,(x_i - x_j)^2} = \boxed{\sigma^2_{\mathcal{G}}}$$

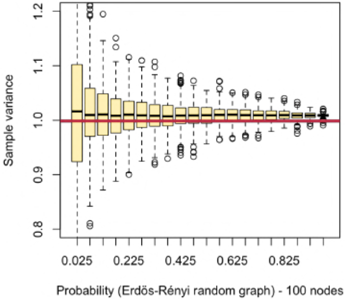
otherwise, the topology of the network is not neutral...



Statistics with a Network Topology

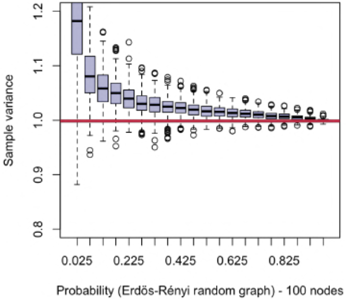
$\sigma^2_{\mathcal{G}} \approx \sigma^2$

“$\text{cor}(A,x) \approx 0$”



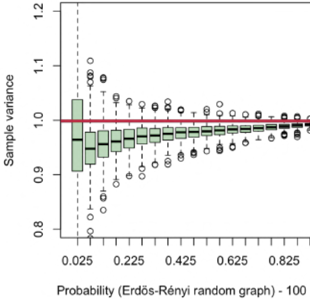
$\sigma^2_{\mathcal{G}} \geq \sigma^2$

“$\text{cor}(A,x) \geq 0$”



$\sigma^2_{\mathcal{G}} \leq \sigma^2$

“$\text{cor}(A,x) \leq 0$”



Erdös-Rényi network with $n = 100$ nodes, probability p (drawn randomly in $[0,1]$)

Statistics with a Network Topology

Following Hall (1970), write

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_j + x_j - \bar{x})^2$$

$$\Longrightarrow (n-1)s^2 = \sum_{i=1}^{n}(x_i - x_j)^2 + 2\sum_{i=1}^{n}(x_i - x_j)(x_j - \bar{X}) + \sum_{i=1}^{n}(x_j - \bar{x})^2.$$

$$\Longrightarrow n(n-1)s^2 = \sum_{j=1}^{n}\sum_{i=1}^{n}(x_i - x_j)^2 + 2\sum_{j=1}^{n}\sum_{i=1}^{n}(x_i - x_j)(x_j - \bar{x}) + \sum_{j=1}^{n}\sum_{i=1}^{n}(x_j - \bar{x})^2.$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = -2\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(x_j - \bar{x}) = 2\sum_{i=1}^{n}\sum_{j=1}^{n}(x_j - \bar{x} + \bar{x} - x_i)(x_j - \bar{x})$$

$$\Longrightarrow \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = 2\sum_{i=1}^{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 + 2\sum_{i=1}^{n}(\bar{x} - x_i)\underbrace{\sum_{j=1}^{n}(x_j - \bar{x})}_{0} = 2n(n-1)s^2.$$

Statistics with a Network Topology

Thus, for any m, write

$$2n(n-1)s^2 = \sum_{i,j=1}^{n}(x_i - x_j)^2 = \sum_{i,j=1}^{n}\Big(\underbrace{x_i - m}_{u_i} - \underbrace{x_j - m}_{u_j}\Big)^2 = \sum_{i,j=1}^{n} u_i^2 + u_j^2 + 2u_iu_j$$

If $m = \overline{x}$,

$$\sum_{i,j=1}^{n} u_i^2 = n\sum_{i=1}^{n} u_i^2 = n(n-1)S^2 \text{ and therefore } \sum_{i,j=1}^{n} u_iu_j = 0.$$

Hence,

$$\frac{1}{2n^2}\sum_{i,j=1}^{n}(x_i-\overline{x})(x_j-\overline{x}) = 0 \text{ but possibly } \frac{1}{4e}\sum_{i,j=1}^{n}A_{i,j}(x_i-\overline{x})(x_j-\overline{x}) \neq 0.$$

Paradoxes in Networks

"*on average your friends have more friends than you do.*"

Proposition 6.10: Friendship Paradox

The average number of friends of the collection of friends of individuals in a social network will be higher than the average number of friends of the collection of the individuals themselves. More formally

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{d_i}\sum_{j=1}^{n}A_{ij}d_j\right)\geq\frac{1}{n}\sum_{i=1}^{n}d_i.$$

Define differences Δ_i's between the average of its neighbours' degrees and its own degree, in the sense that

$$\Delta_i=\frac{1}{d_i}\sum_{j=1}^{n}A_{ij}d_j-d_i.$$

Paradoxes in Networks

Write the average as

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{d_i}\sum_{j=1}^{n}A_{ij}d_j-d_i\right)=\frac{1}{n}\sum_{ij=1}^{n}\left(A_{ij}\frac{d_j}{d_i}-A_{ij}\right),$$

that yields

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i=\frac{1}{n}\sum_{ij=1}^{n}A_{ij}\left(\frac{d_j}{d_i}-1\right)\text{ but also }\frac{1}{n}\sum_{ij=1}^{n}A_{ij}\left(\frac{d_i}{d_j}-1\right),$$

by exchanging the summation indices, and because $\boldsymbol{A}$ is a symmetric matrix. By adding the two, we can write

$$\frac{2}{n}\sum_{i=1}^{n}\Delta_i=\frac{1}{n}\sum_{ij}A_{ij}\left(\frac{d_j}{d_i}+\frac{d_i}{d_j}-2\right)=\frac{1}{2n}\sum_{ij}A_{ij}\left(\sqrt{\frac{d_j}{d_i}}-\sqrt{\frac{d_i}{d_j}}\right)^2\geq 0.$$

(the exact equality holds only when $d_i=d_j$ for all pairs of neighbors)

Attributed Networks

Definition 6.37: Attributed Network

An attributed (directed) network $\mathcal{G}_{\boldsymbol{x}} = (V, E, \boldsymbol{X})$ is a network (V, E) where $\boldsymbol{X}$ is a node attributes matrix, $n \times k$, where each row is a feature vectors, for each node in $V = \mathcal{I}_n$.

If $\boldsymbol{X} = (x_1, \cdots, x_n)$, the classical average is

$$\mu(\boldsymbol{x}) = \overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Given an attributed (directed) network $\mathcal{G}_{\boldsymbol{x}} = (V, E, \boldsymbol{X})$, where $\boldsymbol{X} = (x_1, \cdots, x_n)$,

$$\mu_{\mathcal{G}}(\boldsymbol{x}) = \frac{1}{\displaystyle\sum_{i,j} A_{i,j}} \sum_{i,j} A_{i,j} x_i = \frac{1}{2m}\sum_{i=1}^{n} d_i x_i$$

Attributed Networks

Similarly, the variance of $\boldsymbol{X} = (x_1, \cdots, x_n)$ is

$$\text{Var}(\boldsymbol{x}) = \frac{-1}{n-1} \sum_{i \neq j} (x_i - \mu(\boldsymbol{x}))(x_j - \mu(\boldsymbol{x}))$$

while variance over edges

$$\text{Var}_{\mathcal{G}}(\boldsymbol{x}) = \frac{1}{\displaystyle\sum_{i,j} A_{i,j}} \sum_{i,j} A_{i,j}(x_i - \mu_{\mathcal{G}}(\boldsymbol{x}))(x_j - \mu_{\mathcal{G}}(\boldsymbol{x})) = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \frac{d_i d_j}{2m} \right) x_i x_j.$$

This leads to an other modularity measure, after anoter renormalization, so that it takes the value 1 in a network with perfect assortative mixing—one in which all edges fall between nodes with precisely equal values of x_i,

$$\overline{Q} = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} x_i^2 - \frac{d_i d_j}{2m} x_i x_j \right) = \frac{1}{2m} \sum_{i,j} \left(d_i \mathbf{1}_{i=j} - \frac{d_i d_j}{2m} \right) x_i x_j$$

Attributed Networks

Definition 6.38: Modularity measure for attributed networks

For some categorical variable x, the modularity measure is

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \frac{d_i d_j}{2m} \right) \delta(x_i, x_j).$$

If x is a numerical variable, a different normalization is considered

$$Q = \frac{1}{\kappa} \sum_{i,j} \left(A_{i,j} - \frac{d_i d_j}{2m} \right) x_i x_j, \text{ where } \kappa = \sum_{k,l} \left(d_k \mathbf{1}_{k=l} - \frac{d_k d_l}{2m} \right) x_k x_l$$

also coined “assortativity coefficient”.

Attributed Networks

"*you apply for a loan and your would-be lender somehow examines the credit ratings of your Facebook friends. If the average credit rating of these members is at least a minimum credit score, the lender continues to process the loan application. Otherwise, the loan application is rejected,*" Bhattacharya (2015)

"*il ne faut jamais juger les gens sur leurs fréquentations. Tenez, Judas, par exemple, il avait des amis irréprochables,*" Paul Verlaine

For the generalized friendship paradox, which considers attributes other than degree, as in Cantwell et al. (2021), one can define an analogous quantity, $\Delta_i^{(x)}$, for some attribute x (such as the wealth) is defined as

$$\Delta_i^{(x)} = \frac{1}{d_i}\sum_j A_{ij}x_j - x_i,$$

Attributed Networks

which measures the difference between the average of the attribute for node i's neighbours and the value for i itself. When the average of this quantity over all nodes is positive one may say that the generalized friendship paradox holds. In contrast to the case of degree, this is not always true – the value of $\Delta_i^{(x)}$ can be zero or negative – but we can write the average as

$$\frac{1}{n}\sum_i \Delta_i^{(x)} = \frac{1}{n}\sum_i \left(\frac{1}{d_i} \sum_j A_{ij} x_j - x_i \right) = \frac{1}{n}\sum_i \left(x_i \sum_j \frac{A_{ij}}{d_j} - x_i \right),$$

where the second line again follows from interchanging summation indices. Defining the new quantity

$$\delta_i = \sum_j \frac{A_{ij}}{d_j},$$

and noting that

$$\frac{1}{n}\sum_i \delta_i = \frac{1}{n}\sum_{ij} \frac{A_{ij}}{d_j} = \frac{1}{n}\sum_j \frac{1}{d_j}\sum_i A_{ij} = 1,$$

Attributed Networks

we can then write

$$\frac{1}{n}\sum_i \Delta_i^{(x)} = \frac{1}{n}\sum_i x_i\delta_i - \frac{1}{n}\sum_i x_i \frac{1}{n}\sum_i \boldsymbol{\delta}_i = \text{Cov}(\boldsymbol{x},\boldsymbol{\delta}).$$

Thus, we will have a generalized friendship paradox in the sense defined here if (and only if) $\boldsymbol{x}$ and $\boldsymbol{\delta}$ are positively correlated. But this is not always the case

$$\left.\begin{array}{l}\text{Cov}(\boldsymbol{d},\boldsymbol{\delta})\geq 0\\ \text{Cov}(\boldsymbol{x},\boldsymbol{\delta})\geq 0\end{array}\right\} \not\Longrightarrow \text{Cov}(\boldsymbol{d},\boldsymbol{x})\geq 0.$$

Network Centric Fairness Perception

Definition 6.39: d-Neighbors

Given $d \in \mathbb{N}_\star$, let $N_d : V \to \mathcal{P}(V)$ defined as $\overline{N}_d(i) = \{j \in V : \exists k \leq d,\ (A^k)_{i,j} > 0\}$. $N_1(i) = N_i$ corresponds to (standard) neighbors of node i.

Definition 6.40: d-centered subgraph

Given $d \in \mathbb{N}_\star$, and a node i, the subgraph centered on node i (of order d) is $\mathcal{G}_i^d = (\overline{N}_d(i), E_d(i))$ where $E_d(i) = \{(j,j') \in E\ :\ j,j' \in \overline{N}_d(i)\}$.

Suppose that y is binary, $y_i \in \{0,1\}$.

Instead of a "model" $m : \mathcal{X} \to [0,1]$, consider a decision function $h : V \to [0,1]$ decision function.

Network Centric Fairness Perception

Definition 6.41: Isomorphic Networks

Two subgraphs $\mathcal{G}_1 = (V_1, E_1)$ and $\mathcal{G}_2 = (V_2, E_2)$ of $\mathcal{G}$ are isomorphic with respect to $h : V \to \mathbb{R}$ if there exists a one-to-one mapping $\psi : V_1 \to V_2$ such that

- $\forall (k, l) \in E_1$, $(\psi(k), \psi(l)) \in E_2$,
- $\forall k \in V_1$, $h(k) = h(\psi(k))$.

Network Centric Fairness Perception

Definition 6.42: Isomorphic Attributed Networks

Two attributed subgraphs $\mathcal{G}_1 = (V_1, E_1, \boldsymbol{X}_1)$ and $\mathcal{G}_2 = (V_2, E_2, \boldsymbol{X}_2)$ of $\mathcal{G}$ are isomorphic with respect to $h : V \to \mathbb{R}$ if there exists a one-to-one mapping $\psi : V_1 \to V_2$ such that

- $\forall (k, l) \in E_1$, $(\psi(k), \psi(l)) \in E_2$,
- $\forall k \in V_1$, $h(k) = h(\psi(k))$,and $\boldsymbol{x}_k = \mathbf{1}, \boldsymbol{x}_{2,\psi(k)}$.

Network Centric Fairness Perception

Definition 6.43: Fairness Perception Function

$\mathcal{F}(i,h)$ associate with decision h, for some node i (on a given network $\mathcal{G}$), "fairness perception function" if

- local axiom, if $h(i) = h'(i)$ and $\forall j \in N(i)$, $h(j) = h'(j)$, then $\mathcal{F}(i,h) = \mathcal{F}(i,h')$,
- monotonicty axiom, if $h(i) < h'(i)$ and $\forall j \in N(i)$, $h(j) = h'(j)$, then $\mathcal{F}(i,h) \leq \mathcal{F}(i,h')$,
- neighborhood expectation axiom, if $h(i) = h'(i)$ and $\forall j \in N(i)$, $h(j) \leq h'(j)$, then $\mathcal{F}(i,h) \geq \mathcal{F}(i,h')$,
- homogeneity axiom, let $\mathcal{G}_i = (E_i, V_i)$ and $\mathcal{G}_j = (E_j, V_j)$ be two subgraphs, if $\mathcal{G}_i$ and $\mathcal{G}_j$ are isomorphic with decision function h, then $\mathcal{F}(i,h) = \mathcal{F}(j,h)$

Network Centric Fairness Perception

Definition 6.44: Neighborhood Peer Expectation

Given an network $\mathcal{G}$, a decision function $h : V \to [0,1]$, and a node i

$$E_i[h] = \frac{y_i}{\displaystyle\sum_{j\in N_i} y_j} \sum_{j\in N_i} y_j h(j) + \frac{1-y_i}{\displaystyle\sum_{j\in N_i} 1-y_j} \sum_{j\in N_i} (1-y_j)h(j)$$

where actually, if $y_i = 1$, $E_i[h] = \dfrac{1}{\displaystyle\sum_{j\in N_i} y_j} \displaystyle\sum_{j\in N_i} y_j h(j)$,

while if $y_i = 0$, $E_i[h] = \dfrac{1}{\displaystyle\sum_{j\in N_i} 1-y_j} \displaystyle\sum_{j\in N_i} (1-y_j)h(j)$.

Network Centric Fairness Perception

The Neighborhood Peer Expectation considers the average decision of all neighbors with the same output y.

$$E_i[h] = \frac{y_i}{\displaystyle\sum_{j\in N_i} y_j} \sum_{j\in N_i} y_j h(j) + \frac{1-y_i}{\displaystyle\sum_{j\in N_i} 1-y_j} \sum_{j\in N_i} (1-y_j)h(j)$$

can we extended when considered larger networks, with $d \geq 1$,

$$E_{i,d}[h] = \frac{y_i}{\displaystyle\sum_{j\in \overline{N}_d(i)} y_j} \sum_{j\in \overline{N}_d(i)} y_j h(j) + \frac{1-y_i}{\displaystyle\sum_{j\in \overline{N}_d(i)} 1-y_j} \sum_{j\in \overline{N}_d(i)} (1-y_j)h(j)$$

Network Centric Fairness Perception

Proposition 6.11: Network-Centric Fairness Perception

Given a network $\mathcal{G} = (V, E)$, and a decision function h, the network-centric fairness perception function is defined as

$$\mathcal{F}(i,h) = \begin{cases} 1 \text{ if } E_i[h] \leq h(i) \\ 0 \text{ otherwise} \end{cases}$$

satisfies the locality, monotonicity, neighborhood expectation, and homogeneity axioms, i.e. it is a fairness perception function.

More generally, function $E_i[h]$ should satisfy

- if $\forall j \in N_i$, such that $h(j) = h'(j)$, then $E_j[h] = E_j[h']$,
- if $\forall j \in N_i$, such that $h(j) \leq h'(j)$, then $E_j[h] \leq E_j[h']$,
- if $\mathcal{G}_i$ and $\mathcal{G}_j$ are isomorphic, with respect to h, $E_i[h] = E_j[h]$

Network Centric Fairness Perception

Consider an attributed network $\mathcal{G}_{\boldsymbol{s}} = (V, E, \boldsymbol{S})$

Definition 6.45: Fairness Visibility

Let $V_{\boldsymbol{s}} = \{i \in V : \boldsymbol{S}_i = \boldsymbol{s}\}$, then fairness visibility of h for group $\boldsymbol{s}$ is

$$\overline{\mathcal{F}}_d(\boldsymbol{s}, h) = \frac{1}{\# V_{\boldsymbol{s}}} \sum_{i \in V_{\boldsymbol{s}}} \mathcal{F}_d(i, h)$$

Definition 6.46: Fairness Visibility Parity

h satisfies fairness visibility parity, with respect to $\boldsymbol{S}$, if

$$\overline{\mathcal{F}}_d(\boldsymbol{s}, h) = \overline{\mathcal{F}}_d(\boldsymbol{s}', h).$$

Network Centric Fairness Perception

Consider some binary decision rule $h : V \to \{0, 1\}$,

Proposition 6.12: Asymptotic Fairness Visibility

Assuming the network graph is connected, and the decision function h has non-zero true positive and false positive rates, the fairness visibility of group $V_{\boldsymbol{s}}$, based on the neighborhood peer expectation, converges to the acceptance probability for $V_{\boldsymbol{s}}$ as the d-neighborhood size increases,

$$\overline{\mathcal{F}}_d(\boldsymbol{s}, h) = \frac{1}{\# V_{\boldsymbol{s}}} \sum_{i \in V_{\boldsymbol{s}}} \mathcal{F}(i, h) \to \mathbb{P}[h(i) = 1 | i \in V_{\boldsymbol{s}}), \text{ as } d \to \infty.$$

Heuristically, since the graph is connected, $N_i^{(d)} \to V$ as d increases.

For any i, ultimately, $\begin{cases} \mathcal{F}_d(i, h) = 1 \text{ if } h(i) = 1 \\ \mathcal{F}_d(i, h) = 0 \text{ if } h(i) = 0 \end{cases}$, thus consider only $i \in V_{\boldsymbol{s}}$

Network Centric Fairness Perception

For non-relational data, standard definition of demographic parity is

Definition 6.47: Demographic Parity

Decision function h satisfies demographic parity if

$$\mathbb{P}[h(i)=1|i\in V_{\boldsymbol{s}}) = \mathbb{P}[h(i)=1|i\in V_{\boldsymbol{s}'}).$$

acceptance probability for group s

Again, this definition ignores the neighborhood structure of a node.

Network Centric Fairness Perception

Proposition 6.13: Local vs. Asymptotic Fairness Visibility

Even if decision function h satisfies demographic parity,

$$\mathbb{P}[h(i) = 1 | i \in V_{\boldsymbol{s}}) = \mathbb{P}[h(i) = 1 | i \in V_{\boldsymbol{s}'}),$$

there can still be non-parity w.r.t. fairness visibility, for some d,

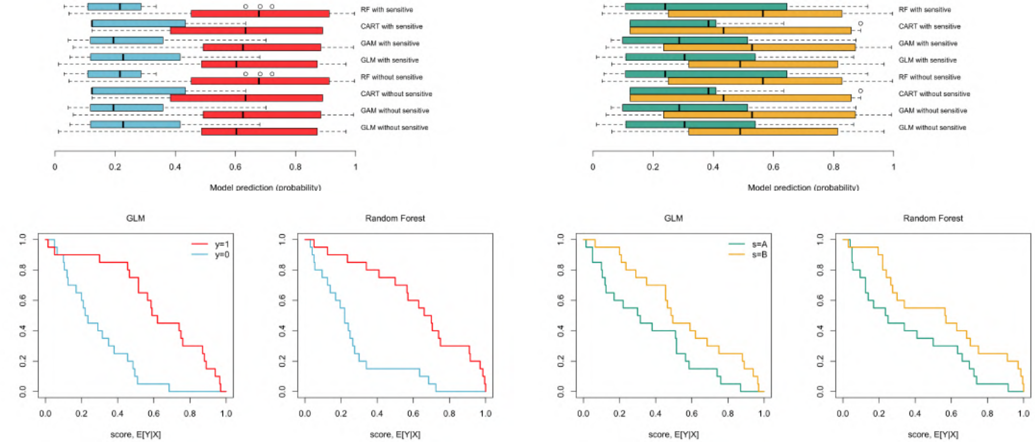
$$\overline{\mathcal{F}}_d(\boldsymbol{s}, h) \neq \overline{\mathcal{F}}_d(\boldsymbol{s}', h).$$

– Part 6 –

Group Fairness

Group Fairness

Back on `toydata2`, distributions of scores, $\widehat{m}(\boldsymbol{x}_i)$'s conditional on y_i and s_i



Group Fairness

Definition 8.1: Fairness through unawareness, Dwork et al. (2012)

A model m satisfies the fairness through unawareness criteria, with respect to sensitive attribute $s \in \mathcal{S}$ if $m : \mathcal{X} \to \mathcal{Y}$.

by Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel,

Group Fairness

See introduction about the gender directive,

"*institutional messages of color blindness may therefore artificially depress formal reporting of racial injustice. Color-blind messages may thus appear to function effectively on the surface even as they allow explicit forms of bias to persist*," Apfelbaum et al. (2010)

Definition 8.2: Aware and unaware regression functions μ

The aware regression function is $\mu(\boldsymbol{x}, s) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, S = s]$ and the unaware regression function is $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.

Historical Perspective: "Cultural Fairness" and "Statistical Discrimination"

Definition 8.3: Four definitions of cultural fairness, Darlington (1971)

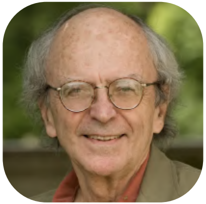
A test ($\widehat{y}$) is considered "culturally fair" if it fits the appropriate equation

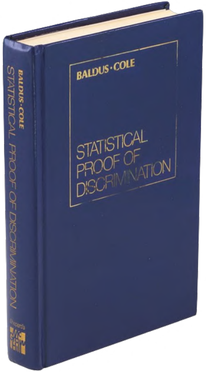
$$\begin{cases}\text{Cor}[S,\widehat{Y}]=\text{Cor}[S,Y]/\text{Cor}[Y,\widehat{Y}]\\ \text{Cor}[S,\widehat{Y}]=\text{Cor}[S,Y]\\ \text{Cor}[S,\widehat{Y}]=\text{Cor}[S,Y]\cdot\text{Cor}[Y,\widehat{Y}]\\ \text{Cor}[S,\widehat{Y}]=0\end{cases}$$

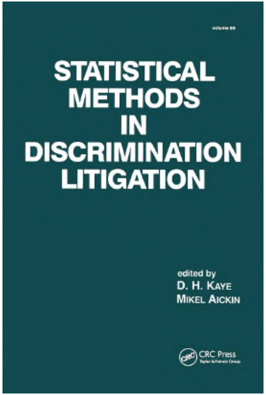
See also Thorndike (1971), Linn and Werts (1971), following Cleary (1968).

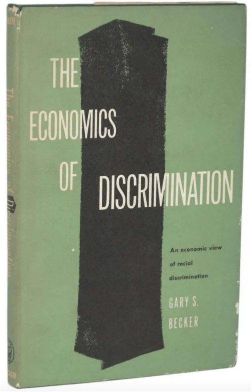
"Economics of Discrimination" and "Statistical Discrimination"

See Becker (1957) or Baldus and Cole (1980), among (many) others.









Historical Perspective: Decomposition

$$\begin{cases} y_{\text{A}:i} = \boldsymbol{x}_{\text{A}:i}^\top \beta_{\text{A}} + \varepsilon_{\text{A}:i} & (\text{group A}),\ \overline{y}_{\text{A}} = \overline{\boldsymbol{x}}_{\text{A}}^\top \widehat{\beta}_{\text{A}} \\ y_{\text{B}:i} = \boldsymbol{x}_{\text{B}:i}^\top \beta_{\text{B}} + \varepsilon_{\text{B}:i} & (\text{group B}),\ \overline{y}_{\text{B}} = \overline{\boldsymbol{x}}_{\text{B}}^\top \widehat{\beta}_{\text{B}}. \end{cases}$$

Using ordinary least squares estimates

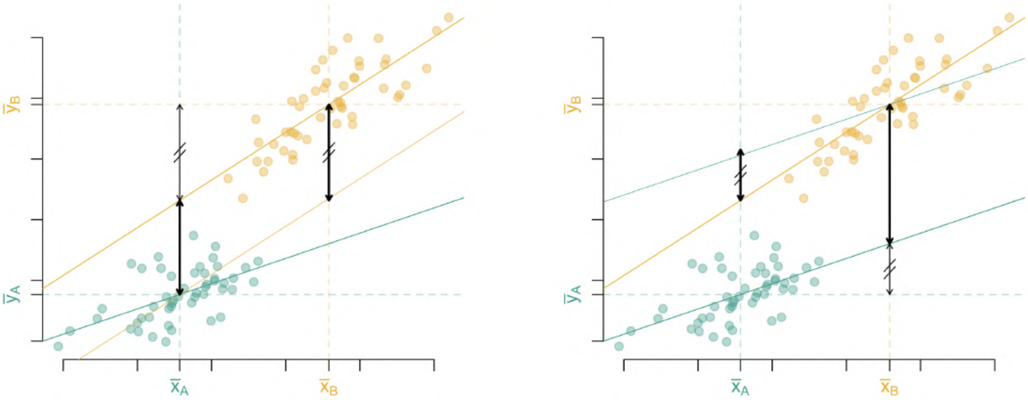
Definition 8.4: Kitagawa (1955), Oaxaca (1973), Blinder (1973)

$$\overline{y}_{\text{A}} - \overline{y}_{\text{B}} = \underbrace{\left(\overline{\boldsymbol{x}}_{\text{A}} - \overline{\boldsymbol{x}}_{\text{B}}\right)^\top \widehat{\beta}_{\text{B}}}_{\text{characteristics}} + \underbrace{\overline{\boldsymbol{x}}_{\text{A}}^\top \left(\widehat{\beta}_{\text{A}} - \widehat{\beta}_{\text{B}}\right)}_{\text{coefficients}}, \tag{7}$$

$$\overline{y}_{\text{A}} - \overline{y}_{\text{B}} = \underbrace{\left(\overline{\boldsymbol{x}}_{\text{A}} - \overline{\boldsymbol{x}}_{\text{B}}\right)^\top \widehat{\beta}_{\text{A}}}_{\text{characteristics}} + \underbrace{\overline{\boldsymbol{x}}_{\text{B}}^\top \left(\widehat{\beta}_{\text{A}} - \widehat{\beta}_{\text{B}}\right)}_{\text{coefficients}}. \tag{8}$$

Also Brown et al. (1980) and Conway and Roberts (1983).

Historical Perspective: Decomposition



$x_{A}(\widehat{\beta}_{A}-\widehat{\beta}_{B})$ and $(\overline{x}_{A}-\overline{x}_{B})\widehat{\beta}_{B}$ (as in Equation 7) on the left
$x_{B}(\widehat{\beta}_{A}-\widehat{\beta}_{B})$ and $(\overline{x}_{A}-\overline{x}_{B})\widehat{\beta}_{A}$ (as in Equation 8) on the right.

Independence and Demographic Parity

Definition 8.5: Independence, Barocas et al. (2017)

A model m satisfies the independence property if $m(\boldsymbol{Z}) \perp\!\!\!\perp S$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.

by Solon Barocas, Moritz Hardt and Arvind Narayanan

For classifiers, one might ask for independence $\widehat{Y} \perp\!\!\!\perp S$ (where $\widehat{y}$ is a class), as Darlington (1971).

Independence and Demographic Parity

Definition 8.6: Demographic Parity, Calders and Verwer (2010), Corbett-Davies et al. (2017)

A decision function $\widehat{y}$ – or a classifier m_t, taking values in $\{0,1\}$ – satisfies demographic parity, with respect to some sensitive attribute S if (equivalently)

$$\begin{cases} \mathbb{P}[\widehat{Y}=1|S=\mathtt{A}]=\mathbb{P}[\widehat{Y}=1|S=\mathtt{B}]=\mathbb{P}[\widehat{Y}=1] \\ \mathbb{E}[\widehat{Y}|S=\mathtt{A}]=\mathbb{E}[\widehat{Y}|S=\mathtt{B}]=\mathbb{E}[\widehat{Y}] \\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{A}]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{B}]=\mathbb{P}[m_t(\boldsymbol{Z})=1]. \end{cases}$$

by Toon Calders, Sicco Verwer, Sam Corbett-Davies, Emma Pierson, Sharad Goel, etc

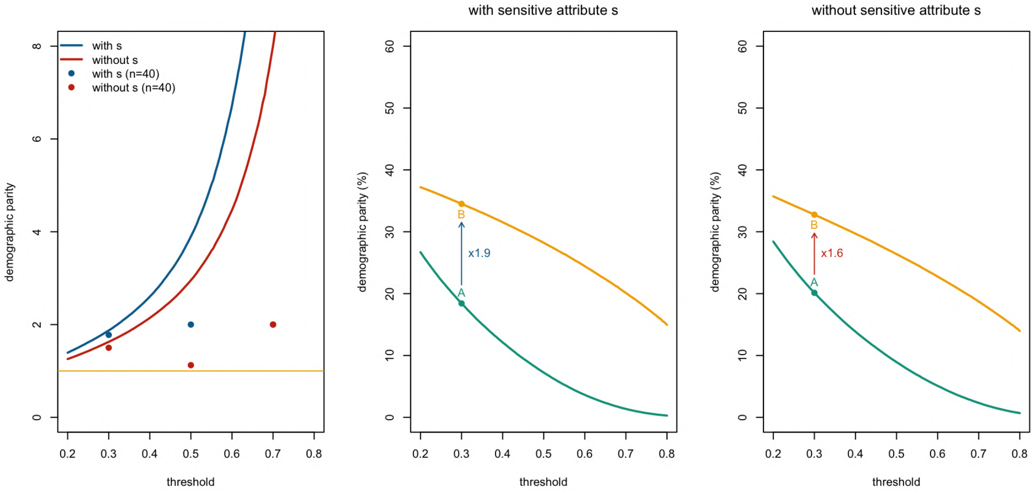
Independence and Demographic Parity

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
$n = 1000$, various t, ratio $\mathbb{P}[\widehat{Y} = 1 \vert S = \texttt{B}]/\mathbb{P}[\widehat{Y} = 1 \vert S = \texttt{A}]$								
$t = 30\%$	1.652	1.519	1.235	1.559	1.918	1.714	1.235	1.798
$t = 50\%$	1.877	2.451	2.918	2.404	2.944	3.457	2.918	2.180
$t = 70\%$	6.033	8.711	26.000	4.621	7.917	19.333	26.000	4.578

(`dem_parity` from R package `fairness`)

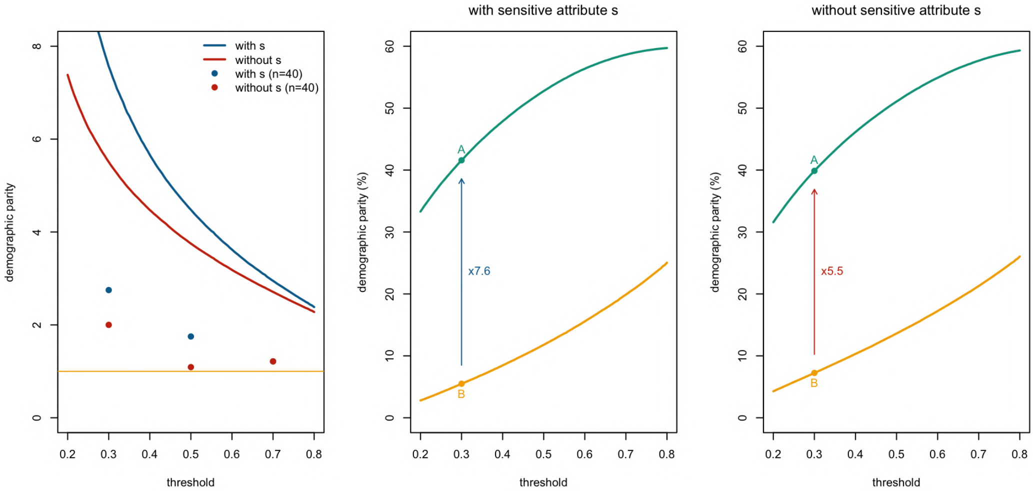
On the left-hand side, evolution of the ratio ratio $\mathbb{P}[\widehat{Y} = 1|S = \texttt{B}]/\mathbb{P}[\widehat{Y} = 1|S = \texttt{A}]$.
The horizontal line (at $y = 1$) corresponds to perfect demographic parity.
In the middle $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) > t|S = \texttt{B}]$ and $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) > t|S = \texttt{A}]$ on the model with s, and on the right-hand side without s.

Independence and Demographic Parity



On the left-hand side, evolution of the ratio ratio $\mathbb{P}[\widehat{Y}=1|S=\texttt{B}]/\mathbb{P}[\widehat{Y}=1|S=\texttt{A}]$.

Independence and Demographic Parity



On the left-hand side, evolution of the ratio ratio $\mathbb{P}[\widehat{Y}=0|S=\texttt{A}]/\mathbb{P}[\widehat{Y}=0|S=\texttt{B}]$

Independence and Demographic Parity

Definition 8.7: Weak Demographic Parity

A decision function $\widehat{y}$ satisfies weak demographic parity if

$$\mathbb{E}[\widehat{Y}|S=\texttt{A}]=\mathbb{E}[\widehat{Y}|S=\texttt{B}].$$

Definition 8.8: Strong Demographic Parity

A decision function $\widehat{y}$ satisfies demographic parity if $\widehat{Y}\perp\!\!\!\perp S$, i.e., for all A,

$$\mathbb{P}[\widehat{Y}\in\mathcal{A}|S=\texttt{A}]=\mathbb{P}[\widehat{Y}\in\mathcal{A}|S=\texttt{B}],\ \forall\mathcal{A}\subset\mathcal{Y}.$$

Independence and Demographic Parity

Proposition 8.1

A model *m* satisfies the strong demographic parity property if and only if

$$d_{\mathrm{TV}}(\mathbb{P}_{m|\mathtt{A}},\mathbb{P}_{m|\mathtt{B}}) = d_{\mathrm{TV}}(\mathbb{P}_{\mathtt{A}},\mathbb{P}_{\mathtt{B}}) = 0.$$

$d_{\mathrm{TV}}(\mathbb{P}_{m|\mathtt{A}},\mathbb{P}_{m|\mathtt{B}})$ could be seen as a measure of “unfairness”, but for a non-binary sensitive attribute, a more general definition is necessary (see Denis et al. (2021)).

Independence and Demographic Parity

Definition 8.9: Conditional demographic parity, Corbett-Davies et al. (2017)

We will have a conditional demographic parity if (at choice) for all $\boldsymbol{x}$,

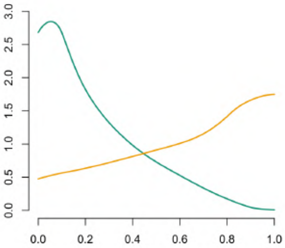
$$\begin{cases}\mathbb{P}[\widehat{Y}=1|\boldsymbol{X}_L=\boldsymbol{x},S=\texttt{A}]=\mathbb{P}[\widehat{Y}=1|\boldsymbol{X}_L=\boldsymbol{x},S=\texttt{B}],\ \forall y\in\{0,1\}\\ \mathbb{E}[\widehat{Y}|\boldsymbol{X}_L=\boldsymbol{x},S=\texttt{A}]=\mathbb{E}[\widehat{Y}|\boldsymbol{X}_L=\boldsymbol{x},S=\texttt{B}],\\ \mathbb{P}[\widehat{Y}\in\mathcal{A}|\boldsymbol{X}_L=\boldsymbol{x},S=\texttt{A}]=\mathbb{P}[\widehat{Y}\in\mathcal{A}|\boldsymbol{X}_L=\boldsymbol{x},S=\texttt{B}],\ \forall\mathcal{A}\subset\mathcal{Y},\end{cases}$$

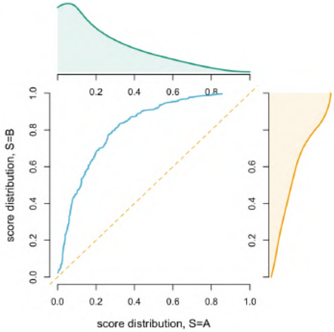
where L denotes a "legitimate" subset of unprotected covariates.

Independence and Demographic Parity

Proposition 8.2

A model m satisfies is strongly fair if and only if $W_2(\mathbb{P}_{\texttt{A}}, \mathbb{P}_{\texttt{B}}) = 0$.

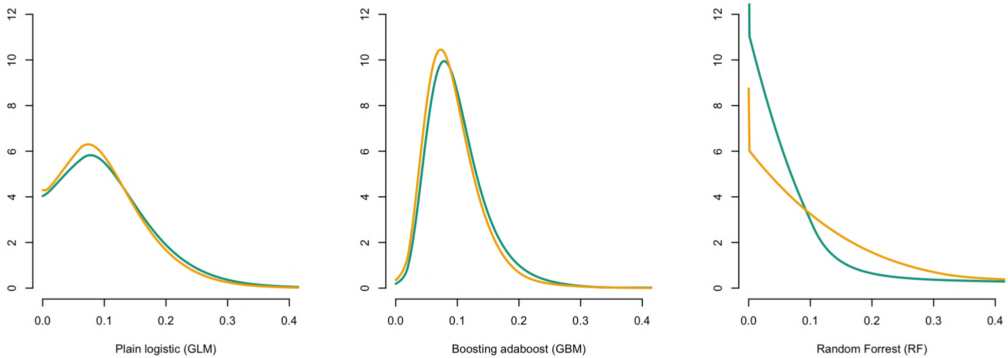




```
> model_glm = glm(y~x1+x2+x3, data=toydata2, family=binomial)
> pred_y_glm = predict(model_glm, type="response")
> sA = pred_y_glm[toydata2$sensitive=="A"]
> library(transport)
> wasserstein1d(sA,sB)
[1] 0.3860795
```

Independence and Demographic Parity

On the `FrenchMotor` dataset, consider GLM, GBM and RF for claim occurence



```
> wasserstein1d(lA,lB)
[1] 0.007220468
```

```
> wasserstein1d(bA,bB)
[1] 0.008895917
```

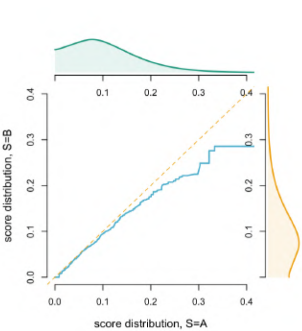
```
> wasserstein1d(fA,fB)
[1] 0.01001088
```

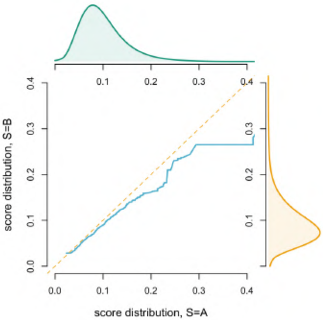
Independence and Demographic Parity

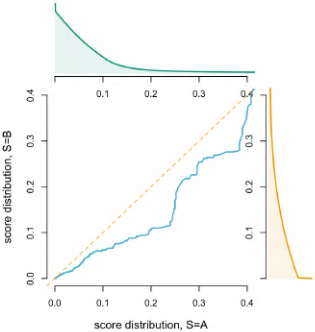
```
> wasserstein1d(lA,lB)
[1] 0.007220468
```

```
> wasserstein1d(bA,bB)
[1] 0.008895917
```

```
> wasserstein1d(fA,fB)
[1] 0.01001088
```







Independence and Demographic Parity

Definition 8.10: Unfairness, Denis et al. (2021); Chzhen and Schreuder (2022)

Given a model m, let $\mathbb{P}_m$ denote the distribution of $m(\boldsymbol{X}, S)$ and $\mathbb{P}_{m|s}$ denote the conditional distribution of $m(\boldsymbol{X}, S)$ given $S = s$, define

$$\begin{cases} \mathcal{U}_{\text{TV}}(m) = \max\limits_{s\in\{\texttt{A},\texttt{B}\}} \{d_{\text{TV}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \text{ or } \sum\limits_{s\in\{\texttt{A},\texttt{B}\}} d_{\text{TV}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \\ \mathcal{U}_{\text{KS}}(m) = \max\limits_{s\in\{\texttt{A},\texttt{B}\}} \{d_{\text{KS}}(\mathbb{P}_m, \mathbb{P}_{m|s})\} \text{ or } \sum\limits_{s\in\{\texttt{A},\texttt{B}\}} d_{\text{KS}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \\ \mathcal{U}_{\text{W}_k}(m) = \max\limits_{s\in\{\texttt{A},\texttt{B}\}} \{W_k(\mathbb{P}_m, \mathbb{P}_{m|s})\} \text{ or } \sum\limits_{s\in\{\texttt{A},\texttt{B}\}} W_k(\mathbb{P}_m, \mathbb{P}_{m|s}) \end{cases}$$

In the original version, Chzhen and Schreuder (2022) suggested to use the one on the right.

Independence and Demographic Parity

Those measures characterize strong demographic parity,

Proposition 8.3: Strong Demographic Parity

A model m is strongly fair if and only if $\mathcal{U}(m) = 0$.

Separation and Equalized Odds

Definition 8.11: Separation, Barocas et al. (2017)

A model $m : \mathcal{Z} \to \mathcal{Y}$ satisfies the separation property if $m(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.

by Solon Barocas, Moritz Hardt and Arvind Narayanan

Separation and Equalized Odds

Definition 8.12: True positive equality, (Weak) Equal Opportunity, Hardt et al. (2016)

A decision function $\widehat{y}$ – or a classifier $m_t(\cdot)$, taking values in $\{0,1\}$ – satisfies equal opportunity, with respect to some sensitive attribute S if

$$\begin{cases}\mathbb{P}[\widehat{Y}=1|S=\mathtt{A},Y=1]=\mathbb{P}[\widehat{Y}=1|S=\mathtt{B},Y=1]=\mathbb{P}[\widehat{Y}=1|Y=1]\\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{A},Y=1]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{B},Y=1]=\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=1],\end{cases}$$

which corresponds to parity of true positives, in the two groups, $\{\mathtt{A},\mathtt{B}\}$.

Separation and Equalized Odds

Definition 8.13: Strong Equal Opportunity

A classifier $m(\cdot)$, taking values in $\{0,1\}$, satisfies equal opportunity, with respect to some sensitive attribute S if

$$\mathbb{P}[m(\boldsymbol{X},S)\in\mathcal{A}|S=\texttt{A},Y=1]=\mathbb{P}[m(\boldsymbol{X},S)\in\mathcal{A}|S$$

for all $\mathcal{A}\subset[0,1]$.

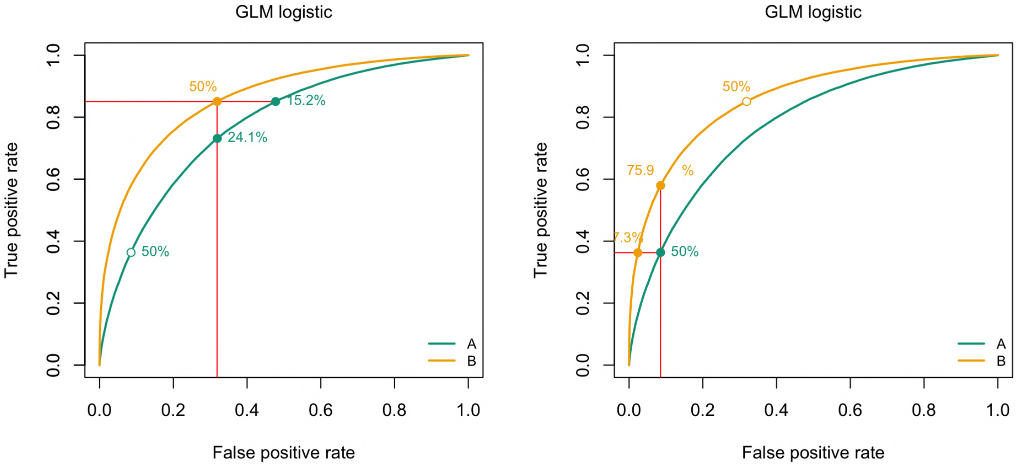
Separation and Equalized Odds

Definition 8.14: False positive equality, Hardt et al. (2016)

A decision function $\widehat{y}$ – or a classifier $m_t(\cdot)$, taking values in $\{0,1\}$ – satisfies parity of false positives, with respect to some sensitive attribute s, if

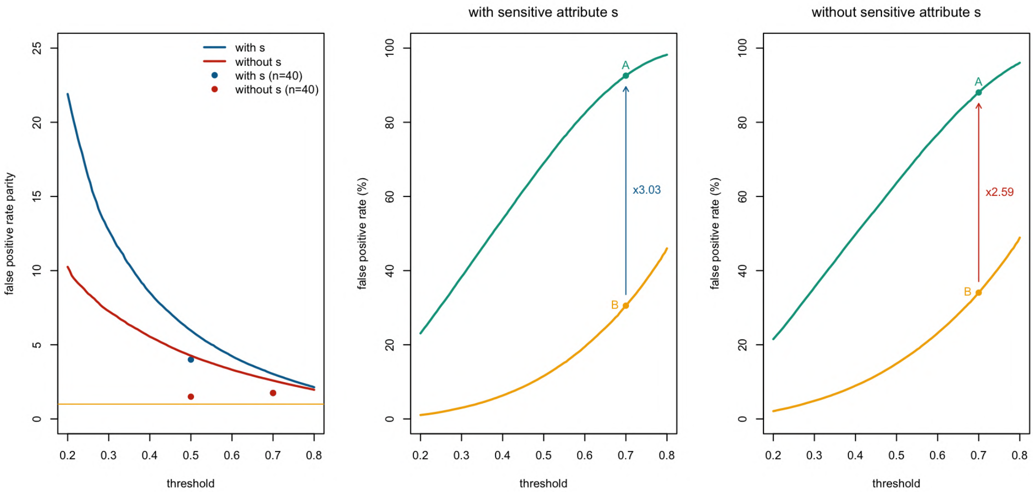
$$\begin{cases}\mathbb{P}[\widehat{Y}=1|S=\mathtt{A},Y=0]=\mathbb{P}[\widehat{Y}=1|S=\mathtt{B},Y=0]=\mathbb{P}[\widehat{Y}=1|Y=0]\\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{A},Y=0]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{B},Y=0]=\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=0].\end{cases}$$

Separation and Equalized Odds



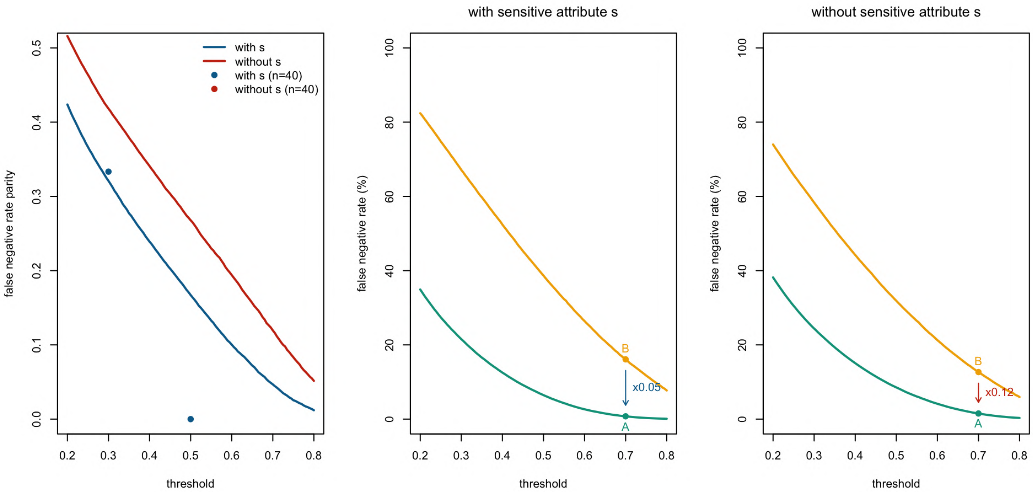
ROC curves (TPR against FPR) for the logistic regression on `toydata2`.

Separation and Equalized Odds



Evolution of the false positive rates, `fpr_parity` from `fairness`.

Separation and Equalized Odds



Evolution of the false negative rates, `fnr_parity` from `fairness`.

Separation and Equalized Odds

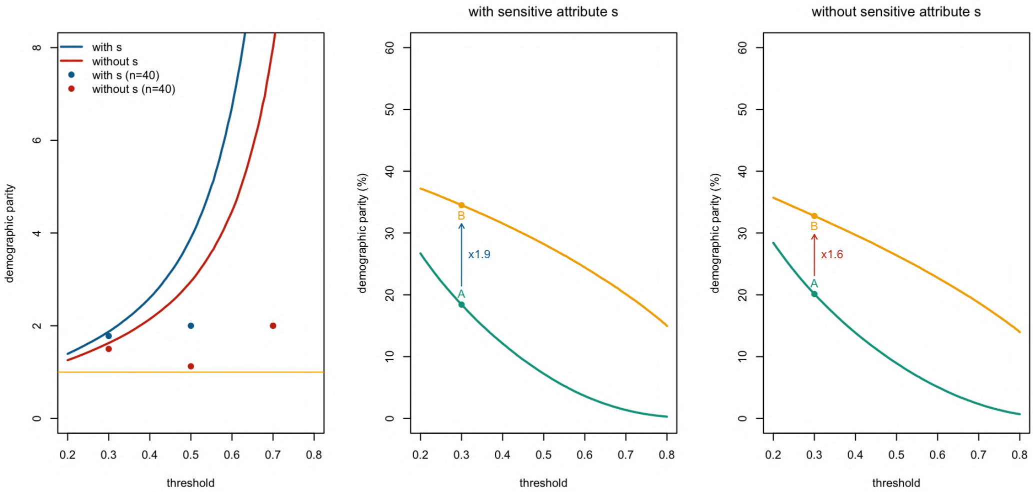
Definition 8.15: Equalized Odds, Hardt et al. (2016)

A decision function $\widehat{y}$ – or a classifier $m_t(\cdot)$ taking values in $\{0,1\}$ – satisfies equal odds constraint, with respect to some sensitive attribute S, if

$$\begin{cases}\mathbb{P}[\widehat{Y}=1|S=\texttt{A},Y=y]=\mathbb{P}[\widehat{Y}=1|S=\texttt{B},Y=y]=\mathbb{P}[\widehat{Y}=1|Y=y],\ \forall y\in\{0,1\}\\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\texttt{A},Y=y]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\texttt{B},Y=y],\ \forall y\in\{0,1\},\end{cases},$$

which corresponds to parity of true positive and false positive, in the two groups.

Separation and Equalized Odds



Evolution of the equalized odds metrics

Separation and Equalized Odds

One can also consider any kind of standard metrics on confusion matrices, such as ϕ (introduced in Yule (1912)), usually named ”Matthews correlation coefficient”

Definition 8.16: ϕ-fairness, Chicco and Jurman (2020)

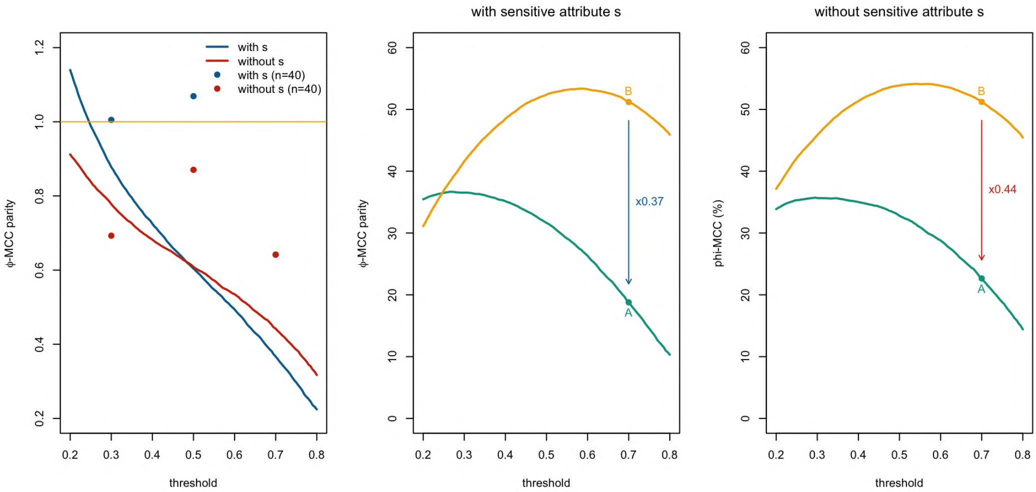
We will have ϕ-fairness if $\phi_{\mathtt{A}} = \phi_{\mathtt{B}}$, where ϕ_s denotes Matthews correlation coefficient for the s group,

$$\phi_s = \frac{\text{TP}_s \cdot \text{TN}_s - \text{FP}_s \cdot \text{FN}_s}{\sqrt{(\text{TP}_s + \text{FP}_s)(\text{TP}_s + \text{FN}_s) \cdot (\text{TN}_s + \text{FP}_s)(\text{TN}_s + \text{FN}_s)}}, \; s \in \{\mathtt{A}, \mathtt{B}\}.$$

but one could consider the F_1-score (as defined in Van Rijsbergen (1979)), Fowlkes–Mallows or Jaccard indices (in Fowlkes and Mallows (1983) or Jaccard (1901)).

.. or AUC as we will considered later on.

Separation and Equalized Odds



Evolution of the ϕ-fairness metric

Separation and Equalized Odds

Definition 8.17: Class Balance, Kleinberg et al. (2016)

We will have class balance in the weak sense if

$$\mathbb{E}[m(\boldsymbol{X})|Y=y,S=\texttt{A}]=\mathbb{E}[m(\boldsymbol{X})|Y=y,S=\texttt{B}],\ \forall y\in\{0,1\},$$

or in the strong sense if

$$\mathbb{P}[m(\boldsymbol{X})\in\mathcal{A}|Y=y,S=\texttt{A}]=\mathbb{P}[m(\boldsymbol{X})\in\mathcal{A}|Y=y,S=\texttt{B}],\ \forall\mathcal{A}\subset[0,1],\ \forall y\in\{0,1\}.$$

Separation and Equalized Odds

Definition 8.18: Similar Mistreatement, Zafar et al. (2019)

We will have similar mistreatment, or "*lack of disparate mistreatment*," if

$$\begin{cases}\mathbb{P}[\widehat{Y}=Y|S=\texttt{A}]=\mathbb{P}[\widehat{Y}=Y|S=\texttt{B}]=\mathbb{P}[\widehat{Y}=Y]\\ \mathbb{P}[m_t(\boldsymbol{X})=Y|S=\texttt{A}]=\mathbb{P}[m_t(\boldsymbol{X})=Y|S=\texttt{B}]=\mathbb{P}[m_t(\boldsymbol{X})=Y].\end{cases}$$

Definition 8.19: Equality of ROC curves, Vogel et al. (2021)

Let $\text{FRP}_s(t)=\mathbb{P}[m(\boldsymbol{X})>t|Y=0,S=s]$ and $\text{TPR}_s(t)=\mathbb{P}[m(\boldsymbol{X})>t|Y=1,S=s]$, where $s\in\{\texttt{A},\texttt{B}\}$. Set $\Delta_{TPR}(t)=\text{TPR}_{\texttt{B}}\circ\text{TPR}_{\texttt{A}}^{-1}(t)-t$ et $\Delta_{FRP}(t)=\text{FPR}_{\texttt{B}}\circ\text{FPR}_{\texttt{A}}^{-1}(t)-t$. We will have fairness with respect to ROC curves if $\|\Delta_{TPR}\|_\infty=\|\Delta_{FPR}\|_\infty=0$.

Separation and Equalized Odds

Definition 8.20: AUC Fairness, Borkan et al. (2019)

We will have AUC fairness if $\text{AUC}_{\mathtt{A}} = \text{AUC}_{\mathtt{B}}$, where AUC_s is the AUC associated with model m within the s group.

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
ratio of AUC	0.837	0.839	0.913	0.768	0.857	0.860	0.913	0.763

Sufficiency and Calibration

Inspired by Cleary (1968), define

Definition 8.21: Sufficiency, Barocas et al. (2017)

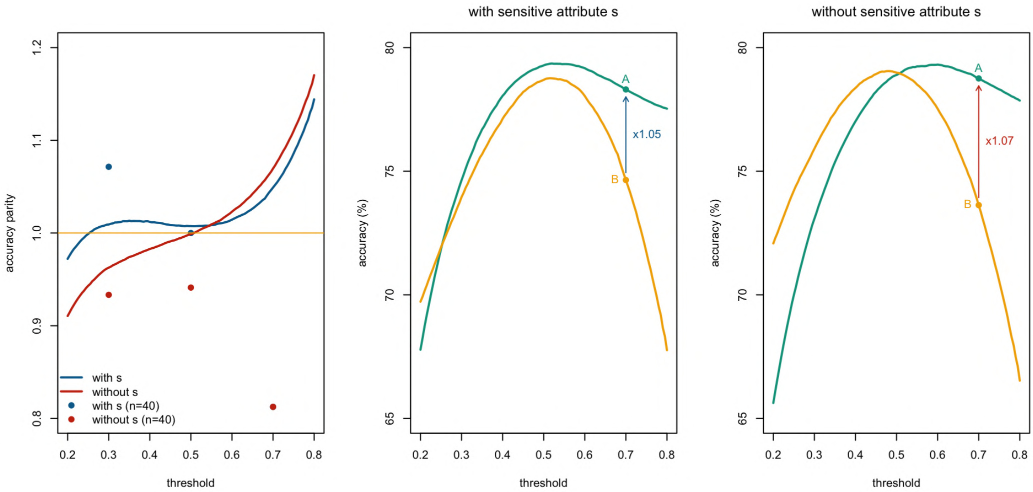
A model $m : \mathcal{Z} \to \mathcal{Y}$ satisfies the sufficiency property if $Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z})$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.

Definition 8.22: Calibration Parity, Accuracy Parity, Kleinberg et al. (2016), Zafar et al. (2019)

Calibration parity is met if

$$\mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{A}] = \mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{B}], \ \forall t \in [0, 1].$$

Sufficiency and Calibration



Evolution of accuracy, in groups A and B.

Sufficiency and Calibration

Definition 8.23: Good Calibration, Kleinberg et al. (2017), Verma and Rubin (2018)

Fairness of good calibration is met if

$$\mathbb{P}[Y=1|m(\boldsymbol{X})=t,S=\texttt{A}]=\mathbb{P}[Y=1|m(\boldsymbol{X})=t,S=\texttt{B}]=t,\ \forall t\in[0,1].$$

Definition 8.24: Non-Reconstruction of Protected Attribute, Kim (2017)

If we cannot tell from the result ($\boldsymbol{x}$, $m(\boldsymbol{x})$, y and $\widehat{y}$) whether the subject was a member of a protected group or not, we will talk about fairness by non-reconstruction of the protected attribute

$$\mathbb{P}[S=\texttt{A}|\boldsymbol{X},m(\boldsymbol{X}),\widehat{Y},Y]=\mathbb{P}[S=\texttt{B}|\boldsymbol{X},m(\boldsymbol{X}),\widehat{Y},Y].$$

Relaxation and Approximate Fairness

Definition 8.25: Disparate Impact, Feldman et al. (2015)

A decision function $\widehat{Y}$ has a disparate impact, for a given threshold τ, if,

$$\min\left\{\frac{\mathbb{P}[\widehat{Y}=1|S=\texttt{A}]}{\mathbb{P}[\widehat{Y}=1|S=\texttt{B}]},\frac{\mathbb{P}[\widehat{Y}=1|S=\texttt{B}]}{\mathbb{P}[\widehat{Y}=1|S=\texttt{A}]}\right\}<\tau \text{ (usually 80\%).}$$

The 80% rule was suggested by the "Technical Advisory Committee on Testing", from the State of California Fair Employment Practice Commission (FEPC) in 1971, or the 1978 "Uniform Guidelines on Employee Selection Procedures", a document used by the U.S. Equal Employment Opportunity Commission (EEOC), see Biddle (2017).

Relaxation and Approximate Fairness

We have defined (Definition 8.10) unfairness as

$$\mathcal{U}_k(m) = \max_{s \in \{\mathtt{A},\mathtt{B}\}} \left\{ W_k(\mathbb{P}_m, \mathbb{P}_{m|s}) \right\},$$

so that m is (strongly) fair if and only if $\mathcal{U}_k(m) = 0$.

Chzhen and Schreuder (2022) introduced the notion of Relative Improvement

Definition 8.26: ε-Approximate Fairness

Model m is ε-approximately fair if $\mathcal{U}_k(m) \leq \varepsilon \cdot \mathcal{U}_k(m^\star)$, where $m^\star$ is Bayes regressor, for some $\epsilon \geq 0$.

Three different concepts ?

$$\begin{cases} \text{Independence (Definition 8.5)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \\ \text{Separation (Definition 8.11)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y \\ \text{Sufficiency (Definition 8.21)}: & Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z}) \end{cases}$$

- Independence assumes no differences among groups, regardless of accuracy
- Separation minimizes differences among groups by not trying to maximize accuracy
- Sufficiency maximizes accuracy by not trying to minimize differences among groups

See Kleinberg et al. (2016) or Chouldechova (2017).

Impossibility theorems

Unless very specific properties are assumed on $\mathbb{P}$, there is no prediction function $m(\cdot)$ that can satisfy at the same time two fairness criteria.

$$\begin{cases} \text{Independence (Definition 8.5)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \\ \text{Separation (Definition 8.11)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y \\ \text{Sufficiency (Definition 8.21)}: & Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z}) \end{cases}$$

Proposition 8.4

Suppose that a model m satisfies the independence condition (8.5) and the sufficiency property (8.21), with respect to a sensitive attribute s, then necessarily, $Y \perp\!\!\!\perp S$.

Therefore, unless the sensitive attribute s has no impact on the outcome y, there is no model m which satisfies independence and sufficiency simultaneously.

Impossibility theorems

From the sufficiency property , $S \perp\!\!\!\perp Y \mid m(\boldsymbol{Z})$, then, for $s \in \mathcal{S}$ and $\mathcal{A} \subset \mathcal{Y}$,

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\big[\mathbb{P}[S = s, Y \in \mathcal{A} | m(\boldsymbol{Z})]\big],$$

can be written

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\big[\mathbb{P}[S = s | m(\boldsymbol{Z})] \cdot \mathbb{P}[Y \in \mathcal{A} | m(\boldsymbol{Z})]\big].$$

And from the independence property (8.21), $m(\boldsymbol{Z}) \perp\!\!\!\perp S$, we can write the first component $\mathbb{P}[S = s | m(\boldsymbol{Z})] = \mathbb{P}[S = s]$, almost surely, and therefore

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\big[\mathbb{P}[S = s] \cdot \mathbb{P}[Y \in \mathcal{A} | m(\boldsymbol{Z})]\big] = \mathbb{P}[S = s] \cdot \mathbb{P}[Y \in \mathcal{A}],$$

for all $s \in \mathcal{S}$ and $\mathcal{A} \subset \mathcal{Y}$, corresponding to the independence between S and Y.

Impossibility theorems

Proposition 8.5

Consider a classifier m_t taking values in $\mathcal{Y} = \{0, 1\}$. Suppose that m_t satisfies the independence condition (8.5) and the separation property (8.11), with respect to a sensitive attribute s, then necessarily either $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$ or $Y \perp\!\!\!\perp S$ (possibly both).

Because m_t satisfies the independence condition (8.5), $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S$, and the separation property (8.11), $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$, them, for $\widehat{y} \in \mathcal{Y}$ and for $s \in \mathcal{S}$,

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}|S = s] = \mathbb{E}\big[\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}|Y, S = s]\big],$$

that we can write

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \sum_y \mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y = y, S = s\big] \cdot \mathbb{P}\big[Y = y|S = s\big],$$

Impossibility theorems

or

$$\mathbb{P}[m_t(\boldsymbol{Z})=\widehat{y}]=\sum_y \mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}|Y=y\big]\cdot\mathbb{P}\big[Y=y\big|S=s\big],$$

almost surely. Furthermore, we can also write

$$\mathbb{P}[m_t(\boldsymbol{Z})=\widehat{y}]=\sum_y \mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}|Y=y\big]\cdot\mathbb{P}\big[Y=y\big],$$

so that, if we combine the two expressions, we get

$$\sum_y \mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}|Y=y\big]\cdot\Big(\mathbb{P}\big[Y=y\big|S=s\big]-\mathbb{P}\big[Y=y\big]\Big)=0,$$

almost surely. And since we assumed that y was a binary variable, $\mathbb{P}[Y=0]=1-\mathbb{P}[Y=1]$, as well as $\mathbb{P}[Y=0|S=s]=1-\mathbb{P}[Y=1|S=s]$, and therefore

$$\mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}|Y=1\big]\cdot\Big(\mathbb{P}\big[Y=1\big|S=s\big]-\mathbb{P}\big[Y=1\big]\Big)$$

Impossibility theorems

or

$$-\mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}|Y=0\big]\cdot\Big(\mathbb{P}\big[Y=0\big|S=s\big]-\mathbb{P}\big[Y=0\big]\Big)$$

can be written

$$\mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}|Y=0\big]\cdot\Big(\mathbb{P}\big[Y=1\big|S=s\big]-\mathbb{P}\big[Y=1\big]\Big).$$

Thus, either $\mathbb{P}\big[Y=1\big|S=s\big]-\mathbb{P}\big[Y=1\big]$ almost surely, or $\mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}|Y=0\big]=\mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}|Y=1\big]$ (or both).
Of course, the previous proposition holds only when y is a binary variable.

Impossibility theorems

Proposition 8.6

Consider a classifier m_t taking values in $\mathcal{Y} = \{0, 1\}$. Suppose that m_t satisfies the sufficiency condition (8.21) and the separation property (8.11), with respect to a sensitive attribute s, then necessarily either $\mathbb{P}[m_t(\boldsymbol{Z}) = 1 | Y = 1] = 0$ or $Y \perp\!\!\!\perp S$ or $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$.

Suppose that m_t satisfies the sufficiency condition (8.21) and the separation property (8.11), respectively $Y \perp\!\!\!\perp S \mid m_t(\boldsymbol{Z})$ and $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$. For all $s \in \mathcal{S}$, we can write, using Bayes formula

$$\mathbb{P}[Y = 1 | S = s, m_t(\boldsymbol{Z}) = 1] = \frac{\mathbb{P}[m_t(\boldsymbol{Z}) = 1 | Y = 1, S = s] \cdot \mathbb{P}[Y = 1 | S = s]}{\mathbb{P}[m_t(\boldsymbol{Z}) = 1 | S = s]},$$

Impossibility theorems

i.e.,

$$\mathbb{P}[Y=1|S=s,m_t(\boldsymbol{Z})=1]=\frac{\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=1]\cdot\mathbb{P}[Y=1|S=s]}{\sum_{y\in\{0,1\}}\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=y]\cdot\mathbb{P}[Y=1|S=s]},$$

that should not depend on s (from the sufficiency property). So a similar property holds if $S=s'$. Observe further that $\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=1]$ is the *true positive rate* (TPR) while $\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=0]$ is the *false positive rate* (TPR). Let $p_s=\mathbb{P}[Y=1|S=s]$, so that

$$\mathbb{P}[Y=1|S=s,m_t(\boldsymbol{Z})=1]=\frac{\text{TPR}}{p_s\cdot\text{TPR}+(1-p_s)\cdot\text{FPR}}.$$

Impossibility theorems

Suppose that Y and S are not independent (otherwise $Y \perp\!\!\!\perp S$ as stated in the proposition), i.e., there are s and s' such that $p_s = \mathbb{P}[Y = 1|S = s] \neq \mathbb{P}[Y = 1|S = s'] = p_{s'}$. Hence, $p_s \neq p_{s'}$, but at the same time

$$\frac{\text{TPR}}{p_s \cdot \text{TPR} + (1 - p_s) \cdot \text{FPR}} = \frac{\text{TPR}}{p_{s'} \cdot \text{TPR} + (1 - p_{s'}) \cdot \text{FPR}}.$$

Supposes that $\text{TPR} \neq 0$ (otherwise $\text{TPR} = \mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1] = 0$ as stated in the proposition), then

$$(p_s - p_{s'}) \cdot \text{TPR} = (p_s - p_{s'}) \cdot \text{FPR} \neq 0,$$

and therefore $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$.

Group fairness, wrap-up

independence, $\widehat{Y} \perp\!\!\!\perp S$, (Definition 8.5)

statistical parity	Dwork et al. (2012)	$\mathbb{P}[\widehat{Y}=1\vert S=s]=\text{cst},\ \forall s$
conditional statistical parity	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y}=1\vert S=s,X=x]=\text{cst}_x,\ \forall s,y$

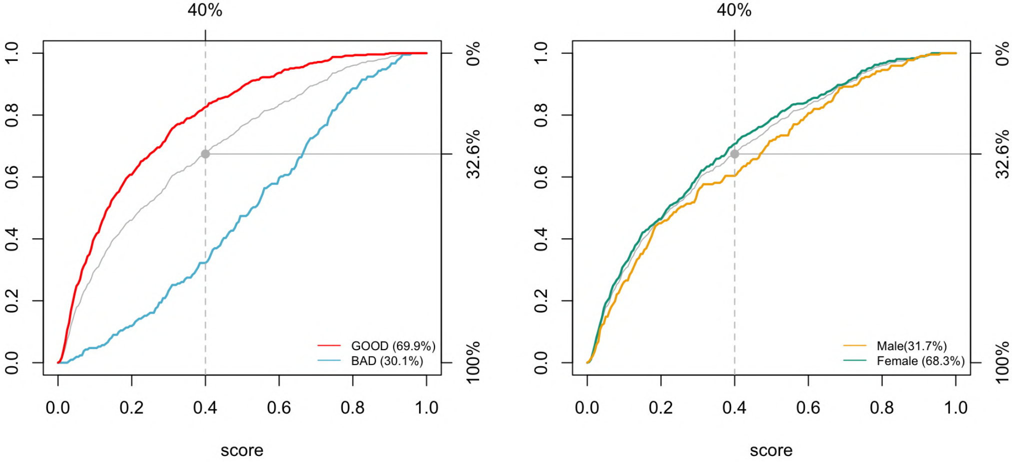
separation, $\widehat{Y} \perp\!\!\!\perp S \mid Y$, (Definition 8.11)

equalized odds	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y}=1\vert S=s,Y=y]=\text{cst}_y,\ \forall s,y$
equalized opportunity	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y}=1\vert S=s,Y=1]=\text{cst},\ \forall s$
predictive equality	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y}=1\vert S=s,Y=0]=\text{cst},\ \forall s$
balance	Kleinberg et al. (2016)	$\mathbb{E}[M\vert S=s,Y=1]=\text{cst}_y,\ \forall s$,y

sufficiency, $Y \perp\!\!\!\perp S \mid \widehat{Y}$, (Definition 8.21)

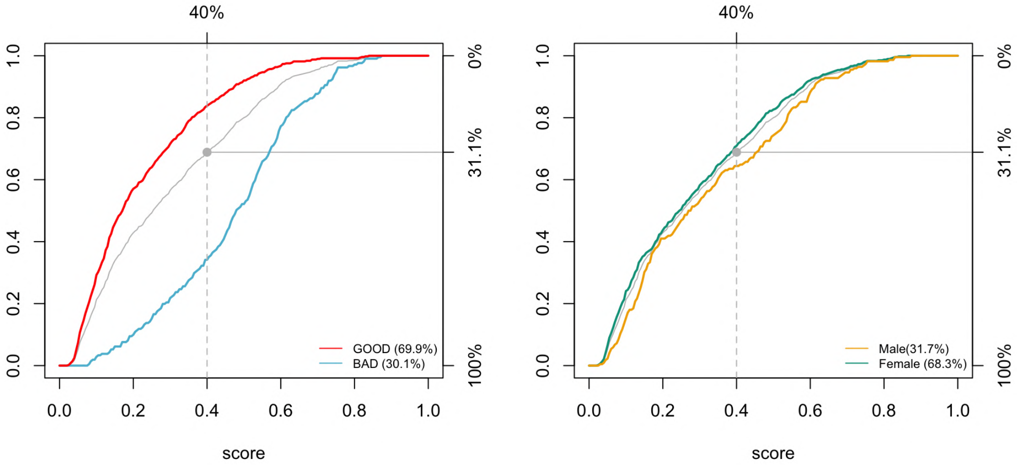
disparate mistreatment	Zafar et al. (2019)	$\mathbb{P}[Y=y\vert S=s,\widehat{Y}=y]=\text{cst}_y,\ \forall s,y$
predictive parity	Chouldechova (2017)	$\mathbb{P}[Y=1\vert S=s,\widehat{Y}=1]=\text{cst},\ \forall s$
calibration	Chouldechova (2017)	$\mathbb{P}[Y=1\vert M=m,S=s]==\psi(m),\ \forall m,s$
well-calibration	Chouldechova (2017)	$\mathbb{P}[Y=1\vert M=m,S=s]=m,\ \forall m,s$

Numerical examples



Conditional distributions of scores on `GermanCredit`, logistic regression.

Numerical examples



Conditional distributions of scores on `GermanCredit`, boosting model.

Numerical examples

	with sensitive				without sensitive			
	GLM	tree	boosting	bagging	GLM	tree	boosting	bagging
$\mathbb{P}[m(\boldsymbol{X}) > t]$	51.7%	28.0%	54.7%	61.7%	50.7%	28.0%	56.0%	60.7%
Predictive Rate Parity	0.992	1.190	0.992	1.050	0.957	1.190	1.041	1.037
Demographic Parity	0.998	1.091	1.159	1.027	1.213	1.091	1.112	1.208
FNR Parity	1.398	0.740	1.078	1.124	1.075	0.740	1.064	0.970
Proportional Parity	0.922	1.008	1.071	0.949	1.121	1.008	1.027	1.116
Equalized odds	0.816	1.069	0.947	0.888	0.956	1.069	0.953	1.031
Accuracy Parity	0.843	1.181	0.912	0.904	0.896	1.181	0.943	0.966
FPR Parity	1.247	0.683	1.470	0.855	2.004	0.683	0.962	1.069
NPV Parity	0.676	1.141	0.763	0.772	0.735	1.141	0.799	0.823
Specificity Parity	0.941	1.439	0.930	1.028	0.851	1.439	1.007	0.990
ROC AUC Parity	0.928	1.162	0.997	1.108	0.926	1.162	1.004	1.090
MCC Parity	0.604	2.013	0.744	0.851	0.639	2.013	0.884	0.930

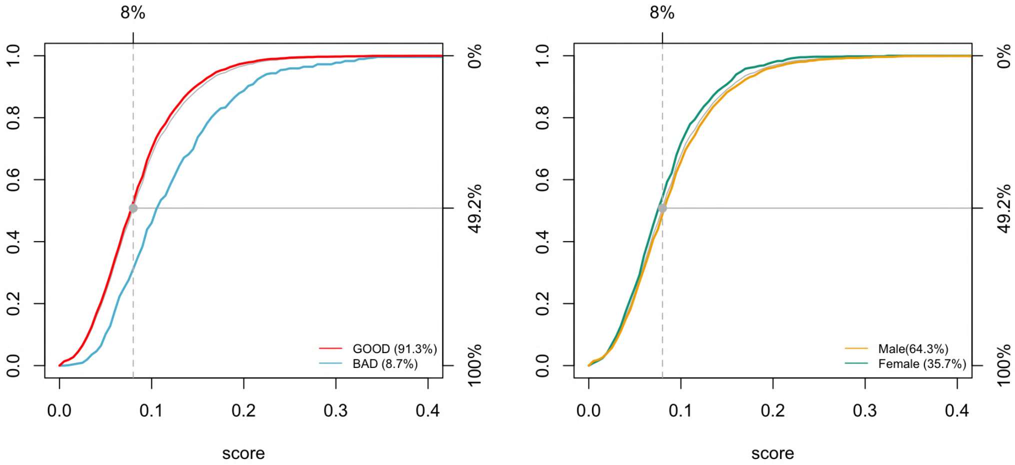
Fairness metrics on `GermanCredit`, with threshold at 20%.

Numerical examples

	with sensitive				without sensitive			
	GLM	tree	boosting	bagging	GLM	tree	boosting	bagging
$\mathbb{P}[m(\boldsymbol{X}) > t]$	30.3%	26.0%	27.7%	25.7%	30.7%	26.0%	28.0%	27.0%
Predictive Rate Parity	1.030	1.179	1.110	1.182	1.034	1.179	1.111	1.200
Demographic Parity	1.090	1.062	1.074	1.069	1.108	1.062	1.044	1.019
FNR Parity	1.533	0.851	1.110	0.781	1.342	0.851	1.322	0.962
Proportional Parity	1.007	0.981	0.992	0.987	1.024	0.981	0.964	0.942
Equalized odds	0.925	1.032	0.982	1.041	0.944	1.032	0.955	1.008
Accuracy Parity	0.949	1.154	1.054	1.164	0.963	1.154	1.038	1.159
FPR Parity	1.118	0.703	0.820	0.653	1.118	0.703	0.784	0.641
NPV Parity	0.738	1.080	0.890	1.108	0.766	1.080	0.848	1.082
Specificity Parity	0.935	1.470	1.169	1.480	0.935	1.470	1.203	1.652
ROC AUC Parity	0.928	1.162	0.997	1.108	0.926	1.162	1.004	1.090
MCC Parity	0.745	1.817	1.105	1.754	0.779	1.817	1.056	2.055

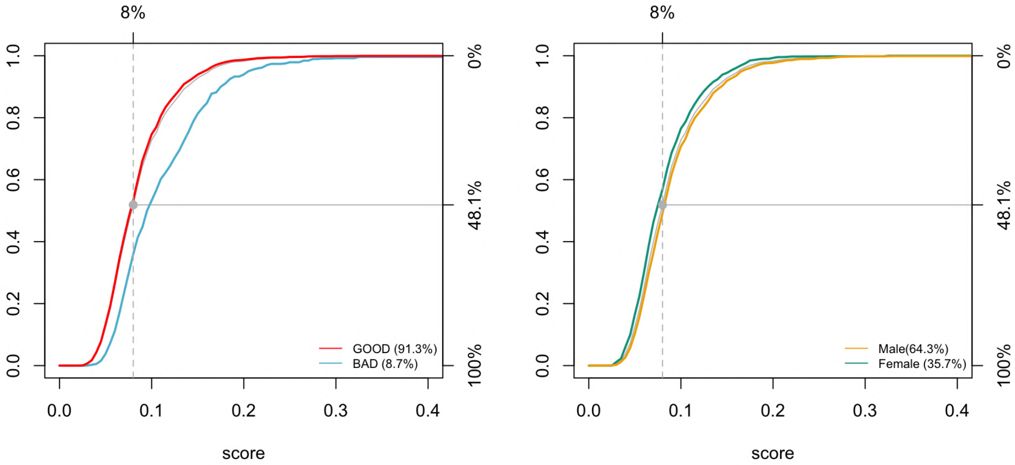
Fairness metrics on `GermanCredit`, with threshold at 40%.

Numerical examples



Conditional distributions of scores on `FrenchMotor`, from the logistic regression.

Numerical examples



Conditional distributions of scores on `FrenchMotor`, from a boosting classification.

– Part 7 –

Individual Fairness

Individual Fairness

Definition 10.1: Similarity Fairness, Luong et al. (2011), Dwork et al. (2012)

Consider two metrics, one on $\mathcal{Y} \times \mathcal{Y}$ (or for a classifier $[0,1]$ and not $\{0,1\}$) noted D_y, and one on $\mathcal{X}$ noted D_x, such that we will have similarity fairness on a database of size n if we have the following property (called Lipschitz property)

$$D_y(m(\boldsymbol{x}_i, s_i), m(\boldsymbol{x}_j, s_j)) \leq L \cdot D_x(\boldsymbol{x}_i, \boldsymbol{x}_j),\ \forall i, j = 1, \cdots, n,$$

for some $L < \infty$.

Individual Fairness

Definition 10.2: Local individual fairness, Petersen et al. (2021)

Consider two metrics, one on $\mathcal{Y}$ ($[0,1]$ for a classifier and not $\{0,1\}$) noted D_y, and one on $\mathcal{X}$ noted D_x, model m is locally individually fair if

$$\mathbb{E}_{(\boldsymbol{X},S)}\left[\limsup_{\boldsymbol{x}':D_x(\boldsymbol{X},\boldsymbol{x}')\to 0}\frac{D_y(m(\boldsymbol{X},S),m(\boldsymbol{x}',S))}{D_x(\boldsymbol{X},\boldsymbol{x}')}\right]\leq L<\infty.$$

Individual Fairness

Definition 10.3: Proxy Based Fairness, Kilbertus et al. (2017)

A decision making process $\widehat{y}$ exhibits no proxy discrimination with respect to sensitive attribute s if

$$\mathbb{E}\big[\widehat{Y}\big|\text{do}(S=\texttt{A})\big]=\mathbb{E}\big[\widehat{Y}\big|\text{do}(S=\texttt{B})\big].$$

Definition 10.4: Fairness on Average Treatment Effect, Kusner et al. (2017)

We achieve fairness on average treatment effect (counterfactual fairness on average)

$$\text{ATE}=\mathbb{E}\big[Y^{\star}_{S\leftarrow\texttt{A}}-Y^{\star}_{S\leftarrow\texttt{B}}\big]=0.$$

Individual Fairness

A decision satisfies counterfactual fairness if ”*had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same*.”

Definition 10.5: Counterfactual Fairness, Kusner et al. (2017)

We achieve counterfactual fairness for an individual with characteristics $\boldsymbol{x}$ if

$$\text{CATE}(\boldsymbol{x}) = \mathbb{E}\big[Y^{\star}_{S\leftarrow \mathtt{A}} - Y^{\star}_{S\leftarrow \mathtt{B}} \big| \boldsymbol{X} = \boldsymbol{x}\big] = 0.$$

Individual Fairness

Definition 10.6: Path-Specific Counterfactual Effect, Wu et al. (2019)

Given a causal diagram, a factual condition (denoted $\mathcal{F}$), and a path π some s to y, the π-effect of a change of s from $\mathtt{B}$ to $\mathtt{A}$ on y is

$$\text{PCE}_\pi(\mathtt{B}\rightarrow\mathtt{A}|\mathcal{F})=\mathbb{E}[Y|\text{do}_\pi(S=\mathtt{A}),\mathcal{F}]-\mathbb{E}[Y|S=\mathtt{B},\mathcal{F}].$$

Counterfactual Fairness

If the protected variable is considered as the treatment, individual fairness is close a measuring a **treatment effect**.

What does "*other things being equal*" really mean ?

It is possible to suppose that the protected attribute s could affect some explanatory variables $\boldsymbol{x}$ in a non-discriminatory way, Kilbertus et al. (2017) (concept of "revolving variable").

See **ceteris paribus** and **mutatis mutandis CATE**, in Charpentier et al. (2023a)

$$\begin{cases} \text{"ceteris paribus CATE"} : \mathbb{E}[Y^\star(B)|\boldsymbol{X}=\boldsymbol{x}] - \mathbb{E}[Y^\star(A)|\boldsymbol{X}=\boldsymbol{x}] \\ \text{"mutatis mutandis CATE"} : \mathbb{E}[Y^\star(B)|\boldsymbol{X}=\boldsymbol{x}^\star(B)] - \mathbb{E}[Y^\star(A)|\boldsymbol{X}=\boldsymbol{x}] \end{cases}$$

suggested also in **?**,**?** and **?**. We need to transport $\boldsymbol{X}|S=A$ to $\boldsymbol{X}|S=B$ (multivariate transport).

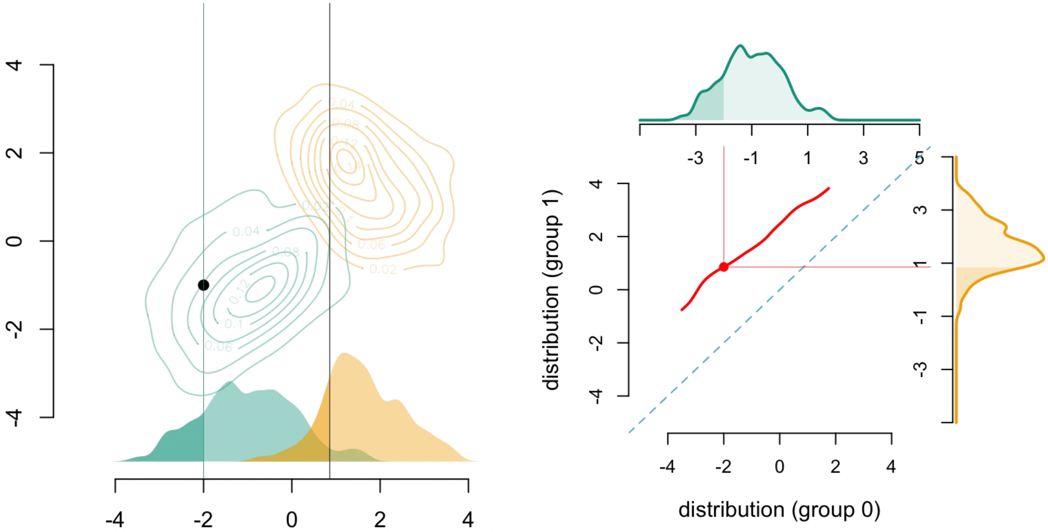
Counterfactual Fairness

As explained in Villani (2003); Carlier et al. (2010); Bonnotte (2013), the Knothe-Rosenblatt rearrangement is directly inspired by the Rosenblatt chain rule, from Rosenblatt (1952), and some extensions obtained on general measures by Knothe (1957). The Knothe-Rosenblatt rearrangement is

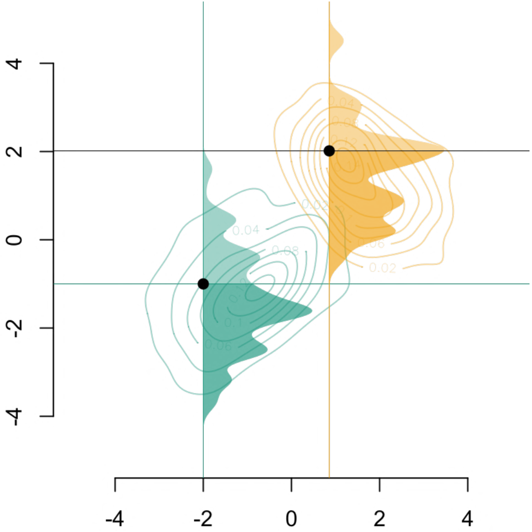
$$T_{\overline{kr}}(x_1,\cdots,x_d)=\begin{pmatrix}T^\star_{\overline{1}}(x_1|x_2,\cdots,x_d)\\T^\star_{\overline{2}}(x_2|x_3,\cdots,x_d)\\\vdots\\T^\star_{\overline{d-1}}(x_{d-1}|x_d)\\T^\star_{\overline{d}}(x_d)\end{pmatrix}\text{ or }T_{\underline{kr}}(x_1,\cdots,x_d)=\begin{pmatrix}T^\star_{\underline{1}}(x_1)\\T^\star_{\underline{2}}(x_2|x_1)\\\vdots\\T^\star_{\underline{d-1}}(x_{d-1}|x_1,\cdots,x_{d-2})\\T^\star_{\underline{d}}(x_d|x_1,\cdots,x_{d-1})\end{pmatrix}.$$

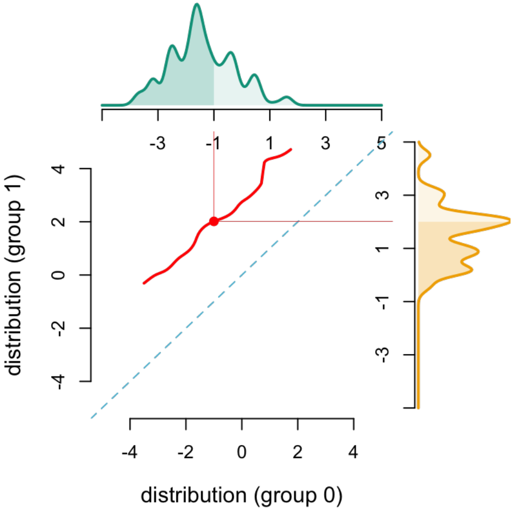
the “monotone lower triangular map,” defined in Bogachev et al. (2005).

Counterfactual Fairness

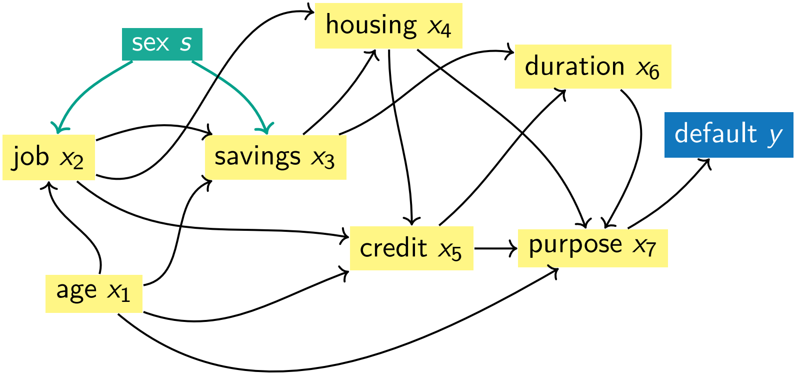


Counterfactual Fairness





Counterfactual Fairness



Causal graph in the German Credit dataset from Watson et al. (2021), or DAG.
(acyclical probablistic graphical models)

Counterfactual Fairness

The joint distribution of $\boldsymbol{X}$ satisfies the (global) **Markov property** w.r.t. $\mathcal{G}$:

$$\mathbb{P}[x_1, \cdots, x_d] = \prod_{j=1}^{d} \mathbb{P}[x_j | \text{parents}(x_j)],$$

where parents(x_i) are nodes with edges directed towards x_i, in $\mathcal{G}$.

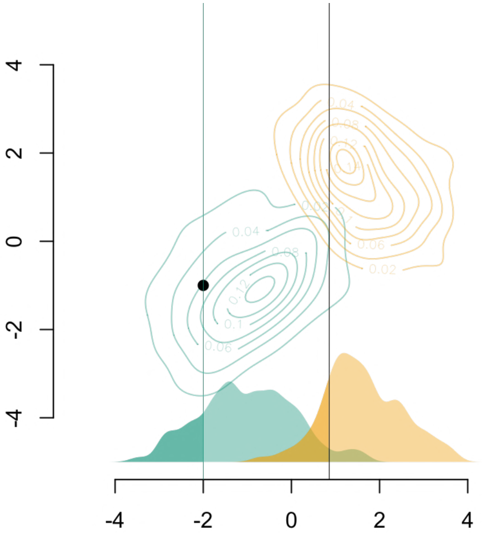
Counterfactual Fairness

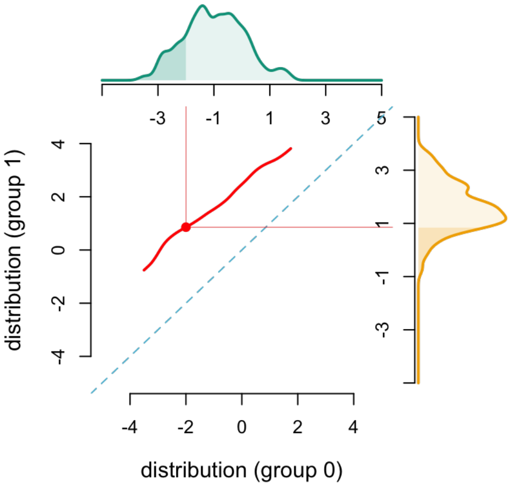
Consider some acyclical causal graph $\mathcal{G}$ on $(s, \boldsymbol{x})$ where variables are topologically sorted, where $s \in \{A, B\}$ is a binary variable , defining two measures μ_A and μ_B on $\mathbb{R}^d$, by conditioning on $s = A$ and $s = B$, respectively, factorized according to $\mathcal{G}$. Define

$$T^\star_{\mathcal{G}}(x_1, \cdots, x_d) = \begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2 |\text{ parents}(x_2)) \\ \vdots \\ T^\star_{d-1}(x_{d-1} |\text{ parents}(x_{d-1})) \\ T^\star_d(x_d |\text{ parents}(x_d)) \end{pmatrix}.$$

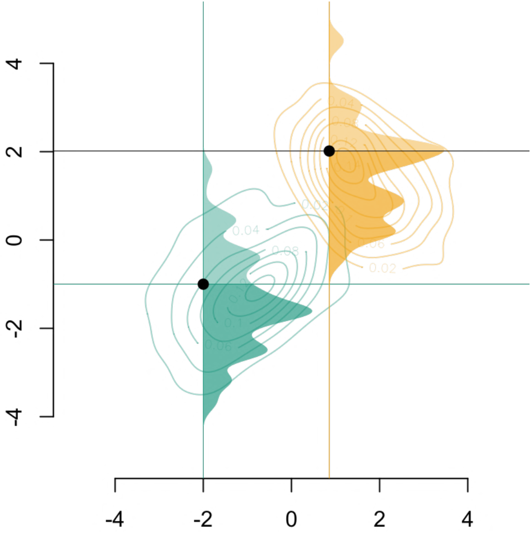
This mapping will be called "sequential conditional transport on the graph $\mathcal{G}$."
The counterfactual value will be obtained by propagating "downstream" the causal graph (following the topological order), when s changes from A to B.

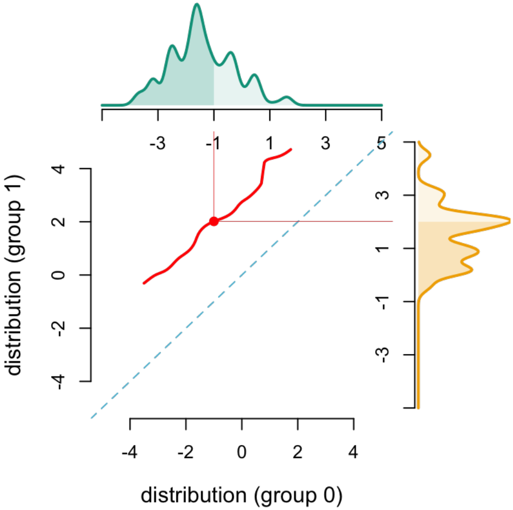
Counterfactual Fairness



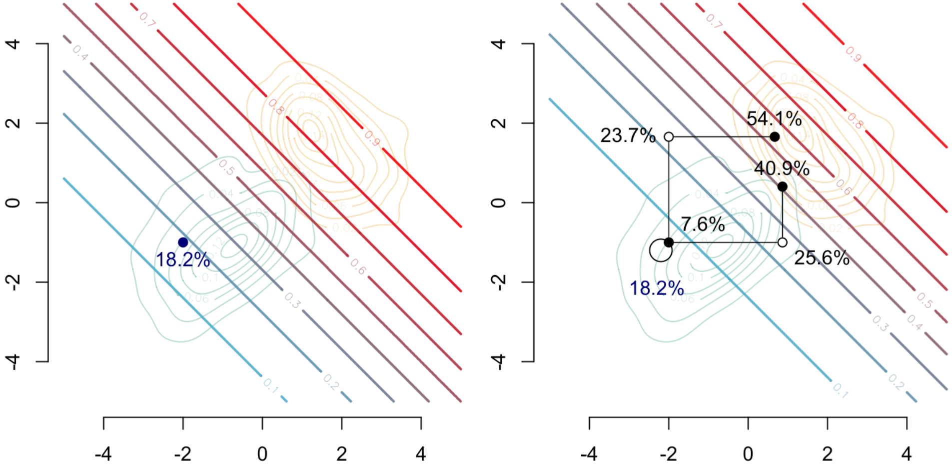


Counterfactual Fairness





Counterfactual Fairness



Counterfactual Fairness

The mutatis mutandis difference $m(s=1,x_1^\star,x_2^\star)-m(s=0,x_1,x_2)$, i.e., $+22.70\%$, is:

$$\begin{array}{rll} & m(s=1,x_1,x_2)-m(s=0,x_1,x_2) & :-10.65\% \\ + & m(s=1,x_1^\star,x_2)-m(s=1,x_1,x_2) & :+17.99\% \\ + & m(s=1,x_1^\star,x_2^\star)-m(s=1,x_1^\star,x_2) & :+15.37\%. \end{array}$$

or $m(s=1,x_1^\star,x_2^\star)-m(s=0,x_1,x_2)$, i.e., $+35.82\%$, is:

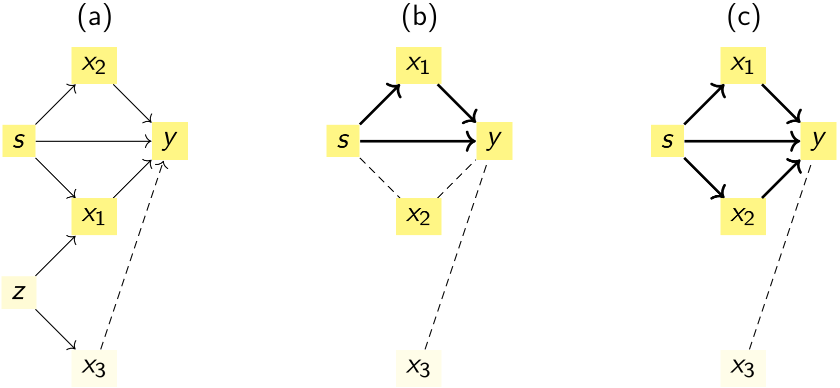
$$\begin{array}{rll} & m(s=1,x_1,x_2)-m(s=0,x_1,x_2) & :-10.66\% \\ + & m(s=1,x_1,x_2^\star)-m(s=1,x_1,x_2) & :+16.07\% \\ + & m(s=1,x_1^\star,x_2^\star)-m(s=1,x_1,x_2^\star) & :+30.41\%. \end{array}$$

The "treatment effect" depends on the causal structure.

Counterfactual Fairness

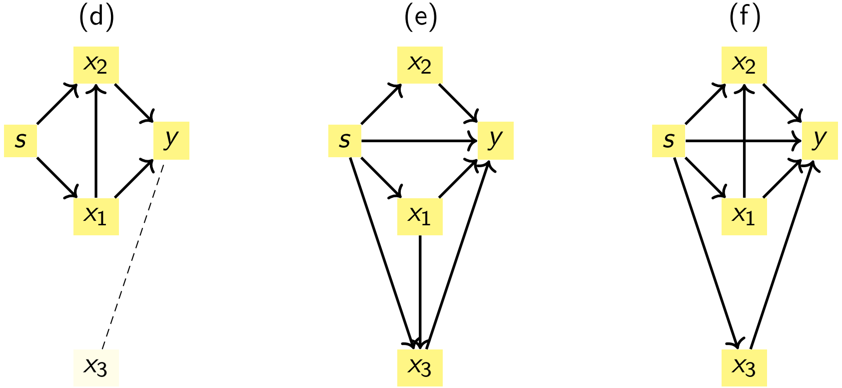
Similarity Fairness (*Lipschitz*)	Dwork et al. (2012)	$D_y(\widehat{y}_i, \widehat{y}_j) \leq D_x(\boldsymbol{x}_i, \boldsymbol{x}_j),\ \forall i$
Proxy Based Fairness,	Kilbertus et al. (2017)	$\mathbb{E}\big[Y \big\vert \text{do}(S = \texttt{A})\big] = \mathbb{E}\big[Y \big\vert \text{do}$
Fairness on Average Treatment Effect	Kusner et al. (2017)	$\mathbb{E}\big[Y^\star_{S\leftarrow\texttt{A}}\big] = \mathbb{E}\big[Y^\star_{S\leftarrow\texttt{B}}\big]$
Counterfactual Fairness,	Kusner et al. (2017)	$\mathbb{E}\big[Y^\star_{S\leftarrow\texttt{A}} \big\vert \boldsymbol{X} = \boldsymbol{x}\big] = \mathbb{E}\big[Y^\star_{S\leftarrow}$
Path-Specific Effect	Avin et al. (2005)	$\mathbb{E}[Y \vert \text{do}_\pi(S = \texttt{A})] = \mathbb{E}[Y \vert \text{d}$
Path-Specific Counterfactual Effect	Wu et al. (2019)	$\mathbb{E}[Y \vert \text{do}_\pi(S = \texttt{A}), \mathcal{F}] = \mathbb{E}[$
Mutatis Mutandis Counterfactual	Kusner et al. (2017)	$\mathbb{E}[Y^\star_{S\leftarrow\texttt{A}} \vert \boldsymbol{X} = \mathcal{T}(\boldsymbol{x})] = \mathbb{E}[$

Counterfactual Fairness



(a) Causal graph used to generate variables in `toydata2`.
(b) Causal graph, where s might cause y, either directly, or indirectly, through x_1.
(c) Causal graph, where s might cause y, either directly or indirectly, via with two possible paths and two mediator variables, x_1 and x_2.

Counterfactual Fairness



(d) Causal graph with no direct impact of s on y, but two mediators, and possibly, x_1 might cause x_2.
(e) similar to (c) with an additional indirect connection from x_1 to y, via mediator x_3.
(f) similar to (d) with an additional indirect connection from x_1 to y, via mediator x_3.

Counterfactual Fairness

Original data

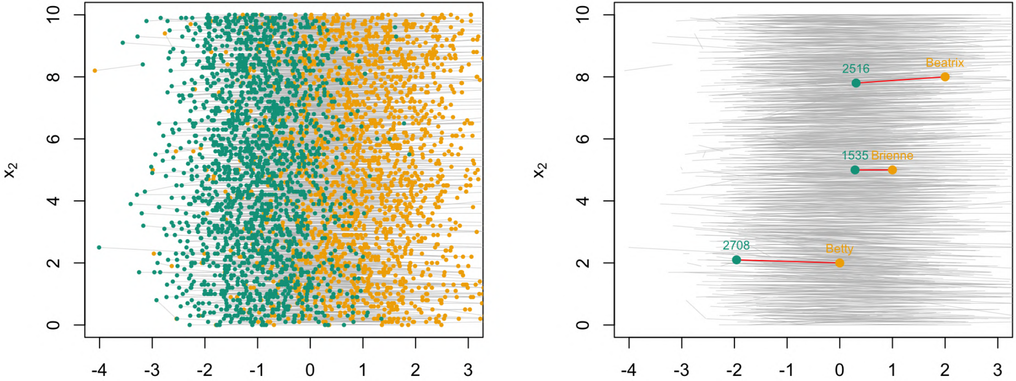
	s	x_1	x_2	x_3	$\widehat{m}_{\text{glm}}(\boldsymbol{x})$	$\widehat{m}_{\text{glm}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{gam}}(\boldsymbol{x})$	$\widehat{m}_{\text{gam}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{rf}}(\boldsymbol{x})$	$\widehat{m}_{\text{rf}}(\boldsymbol{x},s)$
Betty	B	0	2	0	18.22%	24.06%	13.23%	17.63%	17.4%	29.6%
Brienne	B	1	5	1	67.19%	70.47%	66.18%	67.09%	63.60%	61.80%
Beatrix	B	2	8	2	94.95%	94.73%	97.53%	97.58%	96.60%	98.40%
Alex	A	0	2	0	18.22%	13.71%	13.23%	10.05%	17.40%	9.20%
Ahmad	A	1	5	1	67.19%	54.48%	66.18%	50.49%	63.60%	64.40%
Anthony	A	2	8	2	94.95%	90.02%	97.53%	90.51%	96.60%	68.20%

Counterfactual Fairness

Counterfactual

	s	x_1	x_2	x_3	$\widehat{m}_{\text{glm}}(\boldsymbol{x})$	$\widehat{m}_{\text{glm}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{gam}}(\boldsymbol{x})$	$\widehat{m}_{\text{gam}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{rf}}(\boldsymbol{x})$	$\widehat{m}_{\text{rf}}(\boldsymbol{x},s)$
adjusted data, using marginal quantiles										
Betty	A	-1.68	2.1	-1.68	3.51%	3.58%	4.78%	4.85%	10.40%	10.80%
Brienne	A	-0.98	5.1	-0.96	19.39%	17.65%	16.64%	16.13%	29.00%	41.00%
Beatrix	A	-0.27	7.9	-0.26	59.83%	53.65%	51.89%	46.37%	53.60%	49.00%
adjusted data, using optimal transport, (c)										
Betty	A	-1.96	2.1	-1.9	2.62%	2.82%	4.65%	4.81%	0.00%	0.00%
Brienne	A	0.29	5	0.25	48.24%	38.92%	40.04%	32.14%	21.40%	12.20%
Beatrix	A	0.31	7.8	0.21	72.83%	65.1%	67.5%	58.83%	20.80%	15%
adjusted data, using Gaussian transport, (c)										
Betty	A	-1.58	2.15	-1.59	3.95%	3.96%	4.96%	4.99%	0.40%	0.40%
Brienne	A	-0.98	4.96	-0.99	18.47%	16.84%	15.84%	15.40%	19.80%	27.20%
Beatrix	A	-0.38	7.79	-0.38	55.71%	50.05%	47.86%	43.16%	51.80%	63.60%

Counterfactual Fairness



Optimal matching, of individuals in group `B` to individuals in group `A`, on right, where points ● are Betty, Brienne and Beatrix, and ● their counterfactual version in group `A`.

Counterfactual Fairness

Counterfactual

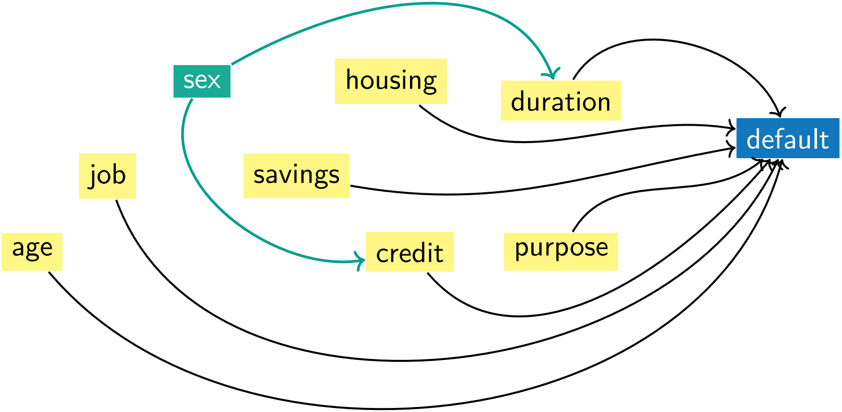
	s	x_1	x_2	x_3	$\widehat{m}_{\text{glm}}(\boldsymbol{x})$	$\widehat{m}_{\text{glm}}(\boldsymbol{x}, s)$	$\widehat{m}_{\text{gam}}(\boldsymbol{x})$	$\widehat{m}_{\text{gam}}(\boldsymbol{x}, s)$	$\widehat{m}_{\text{rf}}(\boldsymbol{x})$	$\widehat{m}_{\text{rf}}(\boldsymbol{x}, s)$
adjusted data, with `fairAdapt`, Figure (e)										
Betty	`A`	-1.65	2	-1.32	3.63%	3.54%	4.72%	4.60%	14.60%	8.00%
Brienne	`A`	-0.97	4.55	-0.94	16.57%	14.96%	13.96%	13.51%	2.20%	5.20%
Beatrix	`A`	-0.33	7.72	-0.44	56.3%	50.71%	48.49%	43.74%	70.60%	74.80%
adjusted data, with `fairAdapt`, Figure (f)										
Betty	`A`	-1.75	2.28	-1.68	3.5%	3.6%	5.03%	5.13%	7.20%	7.00%
Brienne	`A`	-0.96	5.3	-0.96	20.9%	19.05%	17.91%	17.34%	5.80%	8.40%
Beatrix	`A`	-0.24	8.12	-0.34	62.31%	56.43%	54.8%	49.3%	45.60%	39.20%

Numerical illustrations

Original data

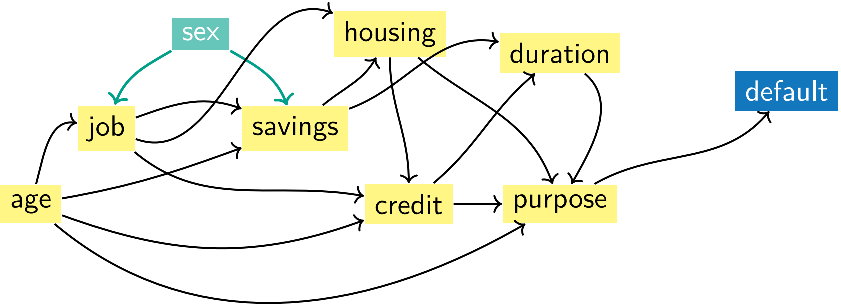
	s	x_1	x_2	x_3	$\widehat{m}_{\text{glm}}(\boldsymbol{x})$	$\widehat{m}_{\text{glm}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{gam}}(\boldsymbol{x})$	$\widehat{m}_{\text{gam}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{rf}}(\boldsymbol{x})$	$\widehat{m}_{\text{rf}}(\boldsymbol{x},s)$
Betty	B	0	2	0	18.22%	24.06%	13.23%	17.63%	17.4%	29.6%
Brienne	B	1	5	1	67.19%	70.47%	66.18%	67.09%	63.60%	61.80%
Beatrix	B	2	8	2	94.95%	94.73%	97.53%	97.58%	96.60%	98.40%
Alex	A	0	2	0	18.22%	13.71%	13.23%	10.05%	17.40%	9.20%
Ahmad	A	1	5	1	67.19%	54.48%	66.18%	50.49%	63.60%	64.40%
Anthony	A	2	8	2	94.95%	90.02%	97.53%	90.51%	96.60%	68.20%

Numerical illustrations



Simple causal graph on the `GermanCredit` dataset,

Numerical illustrations



Causal graph on the `germancredit` dataset, from Watson et al. (2021)

Numerical illustrations

	Alex	Ahmad	Anthony	Betty	Brienne	Beatrix
s (gender)	`M`	`M`	`M`	`F`	`F`	`F`
x_1 `Duration`	12	18	30	12	18	30
$u = F_{1\|s}(x_1)$	36%	57%	86%	34%	50%	79%
$\mathcal{T}(x_1) = F^{-1}_{1\|s=\texttt{M}}(u)$	12	18	30	12	18	24
x_2 `Credit`	1262	2319	4720	1262	2319	4720
$u = F_{2\|s}(x_2)$	25%	55%	82%	17%	45%	76%
$\mathcal{T}(x_2) = F^{-1}_{2\|s=\texttt{M}}(u)$	1262	2319	4720	1074	1855	3854

Numerical illustrations

On the `GermanCredit` dataset

Firstname	s	Firstname	s	Job	Savings	Housing	
Alex	M	Betty	F	highly qualified employee	100 DM	rent	radio
Ahmad	M	Brienne	F	skilled employee	100<=...<500 DM	own	f
Anthony	M	Beatrix	F	unskilled - resident	no savings	for free	c

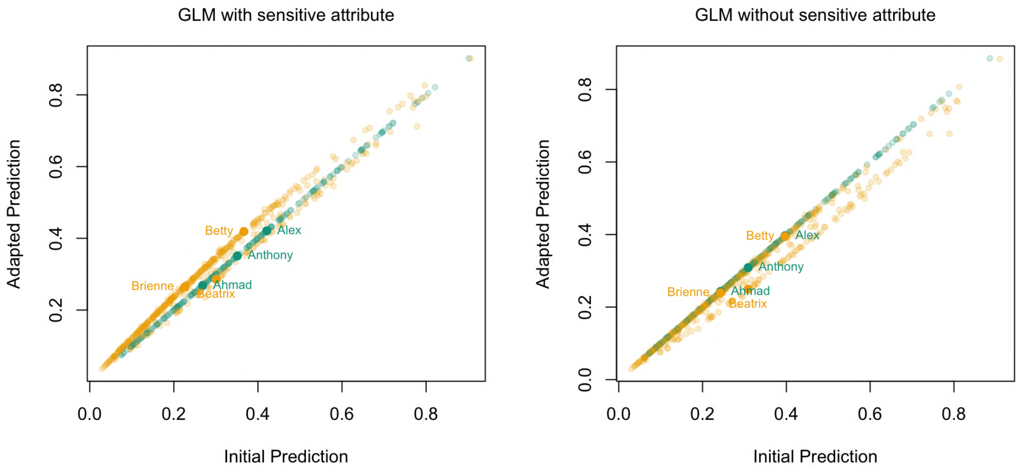
Original data

	s	Age	Duration	Credit	$\widehat{m}_{\text{glm}}(\boldsymbol{x})$	$\widehat{m}_{\text{glm}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{gbm}}(\boldsymbol{x})$	$\widehat{m}_{\text{gbm}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{cart}}(\boldsymbol{x})$	$\widehat{m}_{\text{ca}}$
Betty	F	26	12	1262	39.69%	36.66%	42.30%	43.26%	31.75%	3
Brienne	F	33	18	2320	24.30%	22.61%	23.88%	21.08%	21.31%	2
Beatrix	F	45	30	4720	30.88%	30.08%	28.49%	30.42%	15.38%	1
Alex	M	26	12	1262	39.69%	42.10%	42.30%	44.86%	31.75%	3
Ahmad	M	33	18	2320	24.30%	26.84%	23.88%	22.18%	21.31%	2
Anthony	M	45	30	4720	30.88%	35.08%	28.49%	31.82%	15.38%	1

Numerical illustrations

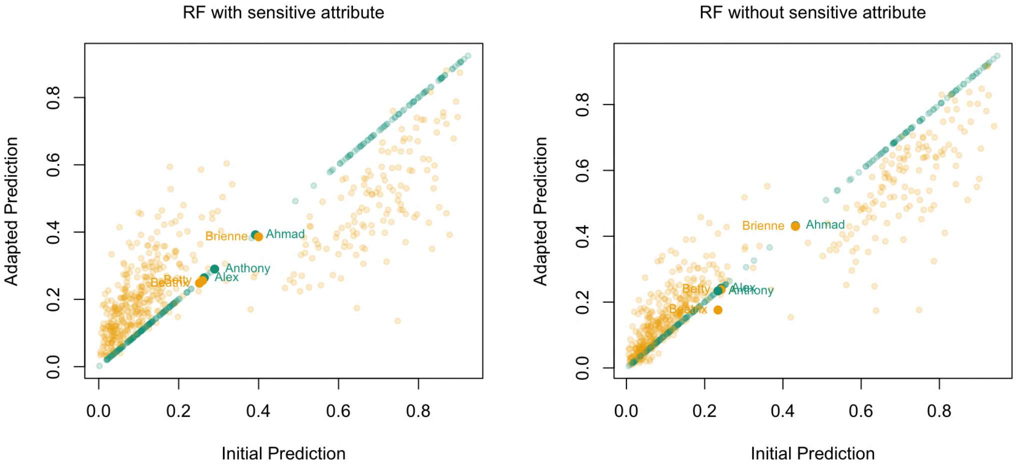
Original data	s	Age	Duration	Credit	$\widehat{m}_{\text{glm}}(\boldsymbol{x})$	$\widehat{m}_{\text{glm}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{gbm}}(\boldsymbol{x})$	$\widehat{m}_{\text{gbm}}(\boldsymbol{x},s)$	$\widehat{m}_{\text{cart}}(\boldsymbol{x})$	$\widehat{m}_{\text{car}}$
Betty	M	26	12	1074	39.51%	41.90%	40.69%	44.86%	31.75%	31
Brienne	M	33	18	1855	23.95%	26.46%	23.88%	22.18%	21.31%	21
Beatrix	M	45	24	3854	24.91%	28.58%	20.55%	20.31%	21.31%	21
adjusted data, with `fairAdapt`, causal graph from Figure **??**										
Betty	M	26	12	1110	42.73%	45.18%	44.24%	46.64%	31.75%	31
Brienne	M	33	18	1787	23.90%	26.40%	23.88%	22.18%	21.31%	21
Beatrix	M	45	24	3990	25.01%	28.70%	22.17%	23.60%	21.31%	21
adjusted data, with `fairAdapt`, causal graph from Figure **??**										
Betty	M	26	18	1778	52.23%	54.03%	40.05%	46.81%	21.31%	21
Brienne	M	33	15	1864	32.25%	35.85%	31.60%	25.97%	21.31%	21
Beatrix	M	45	21	3599	39.70%	43.16%	28.36%	28.90%	21.31%	21
adjusted data, with `fairAdapt`, causal graph from Figure **??**										
Betty	M	26	15	1882	49.05%	50.86%	35.32%	40.12%	21.31%	21
Brienne	M	33	18	1881	50.76%	53.49%	43.00%	38.77%	21.31%	21
Beatrix	M	45	24	3234	24.20%	26.23%	14.63%	16.84%	21.31%	21

Numerical illustrations



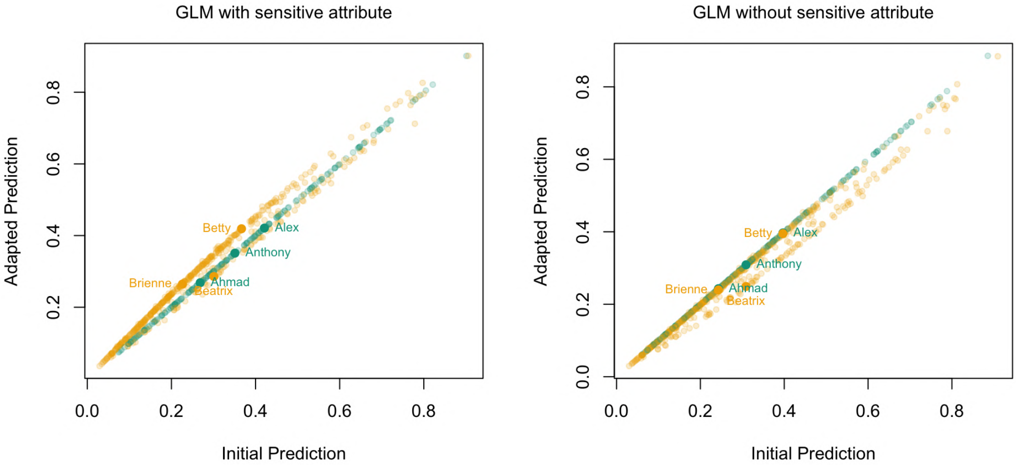
Scatterplot $(m(\boldsymbol{x}_i), m(\mathcal{T}^\star(\boldsymbol{x}_i)))$ for individuals in groups `M` and `F`.

Numerical illustrations



Scatterplot $(m(\boldsymbol{x}_i), m(\mathcal{T}^\star(\boldsymbol{x}_i)))$ for individuals in groups $\texttt{M}$ and $\texttt{F}$.

Numerical illustrations



Scatterplot $(m(\boldsymbol{x}_i), m(\mathcal{T}^\star(\boldsymbol{x}_i)))$ for individuals in groups `M` and `F`.

– Part 8 –

Mitigating Discrimination

Achieving a Fair Prediction

Mitigating discrimination is usually seen as paradoxical, because in order to avoid discrimination, we must create another discrimination. More precisely, Supreme Court Justice Harry Blackmun stated, in 1978, "*in order to get beyond racism, we must first take account of race. There is no other way. And in order to treat some persons equally, we must treat them differently*," cited in Knowlton (1978), as mentioned in Lippert-Rasmussen (2020)).

More formally, an argument in favor of affirmative action – called "*the present-oriented anti-discrimination argument*" – is simply that justice requires that we eliminate or at least mitigate (present) discrimination by the best morally permissible means of doing so, which corresponds to affirmative action. Freeman (2007) suggested a "*time-neutral anti-discrimination argument*," in order to mitigate past, present, or future discrimination.

Achieving a Fair Prediction

But there are also arguments against affirmative action, corresponding to "*the reverse discrimination objection*," as defined in Goldman (1979): some might consider that there is an absolute ethical constraint against unfair discrimination (including affirmative action). To quote another Supreme Court Justice, in 2007, John G. Roberts of the US Supreme Court submits: "*The way to stop discrimination on the basis of race is to stop discriminating on the basis of race*" (Turner (2015) and Sabbagh (2007)).

The arguments against affirmative action are usually based on two theoretical moral claims, according to Pojman (1998). The first denies that groups have moral status (or at least meaningful status). According to this view, individuals are only responsible for the acts they perform as specific individuals and, as a corollary, we should only compensate individuals for the harms they have specifically suffered. The second asserts that a society should distribute its goods according to merit.

Achieving a Fair Prediction

We have defined the risk of a model $m \in \mathcal{M}$ as $\mathcal{R}(m) = \mathbb{E}\big[\ell(Y, m(\boldsymbol{X}))\big]$.
Define the classes of fair models,

$$\begin{cases} \mathcal{M}_{\text{DP}} = \{m \in \mathcal{M} \text{ s.t. } m(\boldsymbol{X}) \perp\!\!\!\perp S\} \\ \mathcal{M}_{\text{EO}} = \{m \in \mathcal{M} \text{ s.t. } m(\boldsymbol{X}) \perp\!\!\!\perp S \mid Y\} \end{cases}$$

Fairness is achieved by projection onto a fair subspace

$$\widehat{m}_{\text{fair}} \in \underset{m \in \mathcal{M}_{\text{fair}}}{\text{argmin}} \left\{\widehat{\mathcal{R}}_n(m)\right\}$$

Definition 12.1: Price of fairness

Given a risk $\mathcal{R}$, a class $\mathcal{M}$ and the fair subclass $\mathcal{M}_{\text{fair}}$, the price of fairness

$$\mathcal{E}_{\text{fair}}(\mathcal{M}) = \min_{m \in \mathcal{M}_{\text{fair}}} \left\{\mathcal{R}(m)\right\} - \min_{m \in \mathcal{M}} \left\{\mathcal{R}(m)\right\}.$$

Achieving a Fair Prediction

Recall that Bayes estimator is the best model,

$$\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] \text{ and set } \begin{cases} \mu_{\mathtt{A}}(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, S = \mathtt{A}] \\ \mu_{\mathtt{B}}(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, S = \mathtt{B}] \end{cases}$$

From the definition of Wasserstein distance,

$$W_2(p, q) = \left(\inf_{\pi \in \Pi(p,q)} \int |x - y|^2 d\pi(x, y) \right)^{1/2}$$

Thus,

$$\mathbb{E}\big[|m(\boldsymbol{X}, S) - \mu_S(\boldsymbol{X})|^2 \big| S = s\big] \geq W_2(\mathbb{P}_m, \mathbb{P}_s)^2$$

Achieving a Fair Prediction

Proposition 12.1: Price of fairness and Wasserstein Barycenter

$$\mathcal{E}_{\text{fair}}(\mathcal{M}) = \min_{m\in\mathcal{M}_{\text{fair}}}\{\mathcal{R}(m)\} - \min_{m\in\mathcal{M}}\{\mathcal{R}(m)\} \geq \min_{g\in\mathcal{M}}\left\{\mathbb{E}\left(W_2(\mathbb{P}_S,\mathbb{P}_{S,g})^2\right)\right\}$$

where $\mathbb{P}_S$ is the condition distribution of $\mu(\boldsymbol{X},S)$, given S, and $\mathbb{P}_{S,g}$ is the condition distribution of $g(\boldsymbol{X},S)$, given S. Moreover, if $\mathcal{M}_{\text{fair}} = \mathcal{M}_{\text{DP}}$, and if $\mathbb{P}_s$ is absolutely continuous (w.r.t. Lebesgue measure),

$$\mathcal{E}_{\text{DP}}(\mathcal{M}) = \min_{g\in\mathcal{M}}\left\{\mathbb{E}\left(W_2(\mathbb{P}_S,\mathbb{P}_{S,g})^2\right)\right\} = \min_{g\in\mathcal{M}}\left\{\sum_s \mathbb{P}[S=s]\cdot W_2(\mathbb{P}_s,\mathbb{P}_{s,g})^2\right\}$$

See Gouic et al. (2020).

The minimum is reached at the Wasserstein barycenter of $\mathbb{P}_S$'s.

Pre-Processing

Write the $n \times k$ matrix $\boldsymbol{S}$ as a collection of k vectors in $\mathbb{R}^n$, $\boldsymbol{S} = \begin{pmatrix} \boldsymbol{s}_1 & \cdots & \boldsymbol{s}_k \end{pmatrix}$, that will correspond to k sensitive attributes. The orthogonal projection on variables $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_k\}$ is associate to matrix $\Pi_S = \boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{S})^{-1}\boldsymbol{S}^\top$, while the projection on the orthogonal of $\boldsymbol{S}$ is $\Pi_{S^\perp} = \mathbb{I} - \Pi_S$ (Gram-Schmidt orthogonalization,).

Let $\widetilde{\boldsymbol{S}}$ denote the collection of centered vectors (using matrix notations, $\widetilde{\boldsymbol{S}} = \boldsymbol{HS}$ where $\boldsymbol{H} = \mathbb{I} - (\mathbf{1}\mathbf{1}^\top)/n$).

Write the $n \times p$ matrix $\boldsymbol{X}$ as a collection of p vectors in $\mathbb{R}^n$, $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_p \end{pmatrix}$. For any $\boldsymbol{x}_j$, define

$$\boldsymbol{x}_j^\perp = \Pi_{\widetilde{S}^\perp}\boldsymbol{x}_j = \boldsymbol{x}_j - \widetilde{\boldsymbol{S}}(\widetilde{\boldsymbol{S}}^\top \widetilde{\boldsymbol{S}})^{-1}\widetilde{\boldsymbol{S}}^\top \boldsymbol{x}_j.$$

One can easily prove that $\boldsymbol{x}_j^\perp$ is then orthogonal to any $\boldsymbol{s}$, since

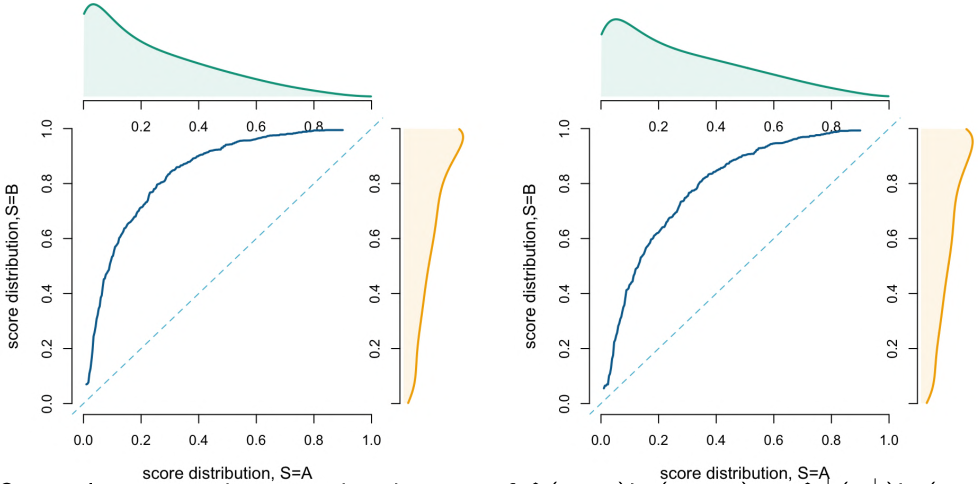
$$\text{Cov}(\boldsymbol{s}, \boldsymbol{x}_j^\perp) = \frac{1}{n}\boldsymbol{s}^\top \boldsymbol{H}\boldsymbol{x}_j^\perp = \frac{1}{n}\widetilde{\boldsymbol{s}}^\top \Pi_{\widetilde{S}^\perp}\boldsymbol{x}_j = 0.$$

Pre-Processing

And similarly the centered version of $\boldsymbol{x}_j^\perp$ is then also orthogonal to any $\boldsymbol{s}$. From an econometric perspective, $\boldsymbol{x}_j^\perp$ can be seen as the residual of the regression of $\boldsymbol{x}_j$ against $\boldsymbol{s}$'s, obtained from least square estimation

$$\boldsymbol{x}_j = \widetilde{\boldsymbol{s}}^\top \widehat{\beta}_j + \boldsymbol{x}_j^\perp.$$

Pre-Processing



Optimal transport between distributions of $\widehat{m}(\boldsymbol{x}_i, s_i)$'s ($x$-axis) to $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s ($y$-axis), for individuals in group `A` on the left-hand side, and in group `B` on the right-hand side.

Pre-Processing

Consider the linear model $\boldsymbol{y} = \boldsymbol{S}\boldsymbol{\alpha} + \boldsymbol{X}^{\perp}\beta + \varepsilon$
Consider the fairness constraint

$$R^2_{\text{fair}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\text{Var}[\boldsymbol{S}\boldsymbol{\alpha}]}{\text{Var}[\boldsymbol{S}\boldsymbol{\alpha} + \boldsymbol{X}^{\perp}\beta]} = \frac{\boldsymbol{\alpha}^{\top}\text{Var}[\boldsymbol{S}]\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\top}\text{Var}[\boldsymbol{S}]\boldsymbol{\alpha} + \boldsymbol{\beta}^{\top}\text{Var}[\boldsymbol{X}^{\perp}]\boldsymbol{\beta}}$$

Then solve

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{\mathbb{E}\left[\|\boldsymbol{y} - \boldsymbol{S}\boldsymbol{\alpha} - \boldsymbol{X}^{\perp}\beta\|^2\right]\right\} \text{ s.t. } R^2_{\text{fair}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq r^2 \ (\in \mathbb{R}_+).$$

Pre-Processing

An alternative was considered in Komiyama and Shimao (2017), with a Ridge penalty

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left\{\mathbb{E}\big[\|\boldsymbol{y}-\boldsymbol{S}\boldsymbol{\alpha}-\boldsymbol{X}^{\perp}\beta\|_{\ell_2}^2\big]+\lambda\|\boldsymbol{\alpha}\|_{\ell_2}^2\right\}$$

The penalty is on $\boldsymbol{\alpha}$ only because (by construction) there is no discriminating information in $\boldsymbol{X}^{\perp}$. There is a closed form solution

$$\begin{pmatrix}(\boldsymbol{S}^{\top}\boldsymbol{S}+\lambda\mathbb{I})^{-1}\boldsymbol{S}^{\top}\boldsymbol{y}\\(\boldsymbol{X}^{\perp\top}\boldsymbol{X}^{\perp})^{-1}\boldsymbol{X}^{\perp}\boldsymbol{y}\end{pmatrix}$$

In-Processing

In a linear regression problem, $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$. Zafar et al. (2017) suggested

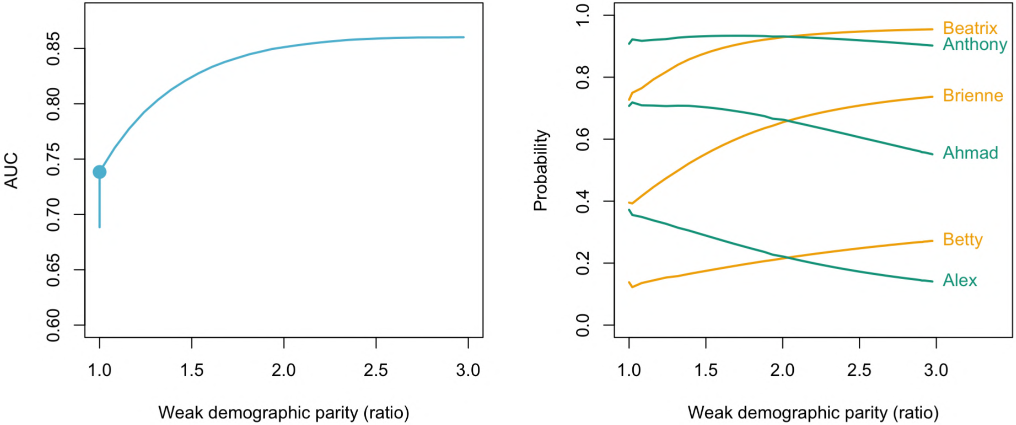
$$\beta^{\star} = \min_{\beta}\left\{\mathbb{E}\big[\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2\big]\right\} \text{ s.t. } \big|\text{Cov}[\boldsymbol{X}\beta, S]\big| \leq c \ (\in \mathbb{R}_+).$$

	$\widehat{m}(\boldsymbol{x}, s)$, aware					$\widehat{m}(\boldsymbol{x})$, unaware			
	← less fair			more fair →		← less fair		more fair →	
$\widehat{\beta}_0$ (Intercept)	-2.55	-2.29	-1.97	-1.51	-1.03	-2.14	-1.98	-1.78	-1.63
$\widehat{\beta}_1$ (x_1)	0.88	0.88	0.85	0.77	0.62	1.01	0.84	0.57	0.26
$\widehat{\beta}_2$ (x_2)	0.37	0.37	0.35	0.32	0.25	0.37	0.35	0.31	0.24
$\widehat{\beta}_3$ (x_3)	0.02	0.02	0.02	0.02	0.03	0.15	0.02	-0.15	-0.29
$\widehat{\beta}_{\text{B}}$ ($\mathbf{1}_{\text{B}}$)	0.82	0.44	-0.03	-0.70	-1.31	-	-	-	-

In-Processing

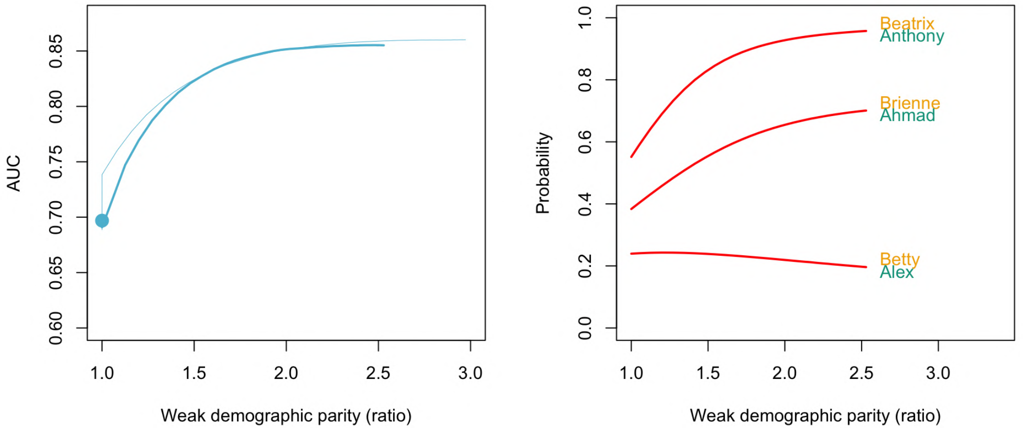
	$\widehat{m}(\boldsymbol{x}, s)$, aware					$\widehat{m}(\boldsymbol{x})$, unaware			
	← less fair			more fair →		← less fair		more fair →	
Betty	0.27	0.25	0.22	0.17	0.14	0.20	0.22	0.24	0.24
Brienne	0.74	0.71	0.66	0.54	0.40	0.70	0.66	0.55	0.38
Beatrix	0.95	0.95	0.93	0.87	0.73	0.96	0.93	0.82	0.55
Alex	0.14	0.17	0.22	0.29	0.37	0.20	0.22	0.24	0.24
Ahmad	0.55	0.61	0.66	0.70	0.71	0.70	0.66	0.55	0.38
Anthony	0.90	0.92	0.93	0.93	0.91	0.96	0.93	0.82	0.55
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \texttt{A}]$	0.23	0.26	0.31	0.36	0.42	0.25	0.30	0.37	0.41
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \texttt{B}]$	0.67	0.65	0.61	0.53	0.42	0.64	0.61	0.54	0.41
(ratio)	×2.97	×2.49	×2.01	×1.46	×1.00	×2.53	×2.02	×1.48	×1.00
AUC	0.86	0.86	0.85	0.82	0.74	0.86	0.85	0.82	0.70

In-Processing



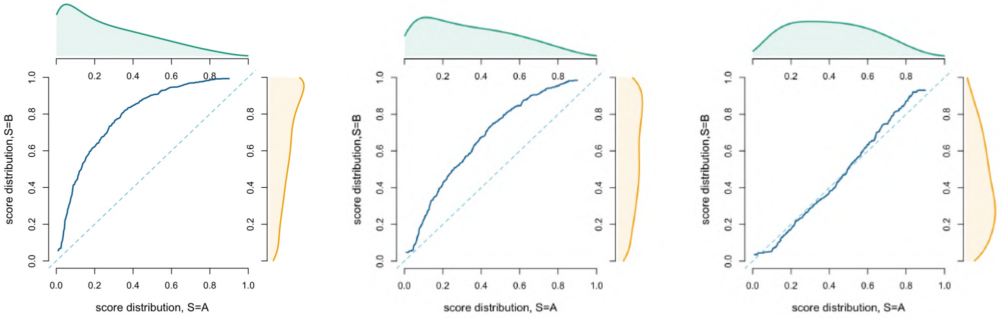
AUC of $\widehat{m}_{\widehat{\beta}_\lambda}$ and evolution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ (with a logistic regression)

In-Processing



AUC of $\widehat{m}_{\widehat{\beta}_\lambda}$ and evolution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i)$ (with a logistic regression)

In-Processing

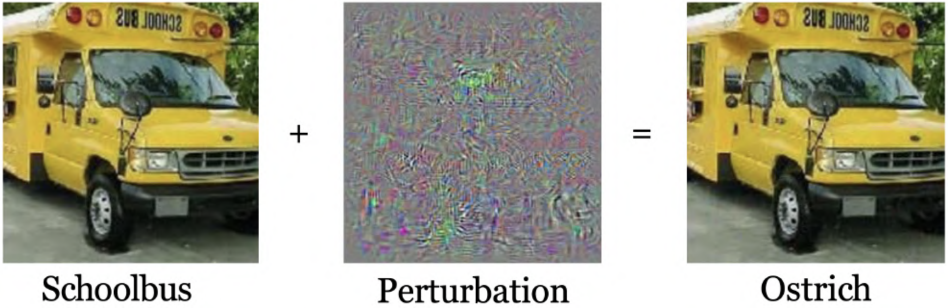


Optimal transport between distributions of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group `A` and in `B`, for different values of λ (low value on the left-hand side and high value on the right-hand side), associated with a demographic parity penalty criteria

In-Processing

Adversarial learning has to do with robustness of learning algorithm, Szegedy et al. (2013) ("*are neural network stables?*").

"*Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake*", Bengio et al. (2017)



In-Processing

Adversarial learning deals with the problem that the distributions we obtain IRL are not the ones we train the model on... and we try to quantify what can go wrong

Popular in pictures (what happens if we rotate an object, add glasses to people, etc). Brittleness of ML algorithms...

Problem of data pollution (add outliers) and problems of adversarial examples.

Machine learning perspective

$$\min_{\boldsymbol{\theta}}\left\{\mathbb{E}_{(\boldsymbol{X},Y)\sim\mathbb{P}}\left[\ell(m_{\boldsymbol{\theta}}(\boldsymbol{X}),Y)\right]\right\}$$

Adversarial perspective

$$\max_{\boldsymbol{\varepsilon}\in\mathcal{E}}\left\{\mathbb{E}_{(\boldsymbol{X},Y)\sim\mathbb{P}}\left[\ell(m_{\boldsymbol{\theta}}(\boldsymbol{X}+\boldsymbol{\varepsilon}),Y)\right]\right\}$$

leads to robust learning...

In-Processing

$$\overbrace{\min_{\boldsymbol{\theta}}\left\{\underbrace{\max_{\varepsilon\in\mathcal{E}}\left\{\mathbb{E}_{(\boldsymbol{X},Y)\sim\mathbb{P}}\left[\ell(m_{\boldsymbol{\theta}}(\boldsymbol{X}+\varepsilon),Y)\right]\right\}}_{\text{creating an adversarial example}}\right\}}^{\text{training a robust classifier}}$$

Approaches based on robust optimization, Ben-Tal et al. (2009), e.g., Danskin's Theorem, Danskin (1967),

$$\nabla_{\boldsymbol{\theta}}\max_{\varepsilon\in\mathcal{E}}\left\{\ell(m_{\boldsymbol{\theta}}(\boldsymbol{X}+\varepsilon),Y)\right\}=\nabla_{\boldsymbol{\theta}}\ell(m_{\boldsymbol{\theta}}(\boldsymbol{X}+\varepsilon^\star),Y)$$

where $\varepsilon^\star=\underset{\varepsilon\in\mathcal{E}}{\text{argmax}}\{\ell(m_{\boldsymbol{\theta}}(\boldsymbol{X}+\varepsilon),Y)\}$.

Recall the minimax theorem from von Neumann (1928)

In-Processing

Proposition 12.2: Nash equilibrium and Minimax

Let A be some $m \times n$ real-valued matrix, there is a Nash equilibrium $(\boldsymbol{x}_\star, \boldsymbol{y}_\star)$ associated with A if

$$\boldsymbol{y}_\star^\top A\boldsymbol{x}_\star = \max_{\boldsymbol{x}\in\mathcal{S}_m} \min_{\boldsymbol{y}\in\mathcal{S}_n} \{\boldsymbol{y}^\top A\boldsymbol{x}\} = \min_{\boldsymbol{y}\in\mathcal{S}_n} \max_{\boldsymbol{x}\in\mathcal{S}_m} \{\boldsymbol{y}^\top A\boldsymbol{x}\}.$$

In-Processing

Consider a Minimax games: given that the discriminator will try to do the best job it can, the generator is set to make the discriminator as wrong as possible

$$\min_{\boldsymbol{\theta}_g} \max_{\boldsymbol{\theta}_d} \left\{ \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}\left[\log(m_{\boldsymbol{\theta}_d}(\boldsymbol{x}))\right] + \mathbb{E}_{\boldsymbol{Z}\sim\mathbb{Q}}\left[\log(1 - m_{\boldsymbol{\theta}_d}(G_{\boldsymbol{\theta}_d}(\boldsymbol{z})))\right] \right\}$$

where $\boldsymbol{X} \sim \mathbb{P}$ denotes data sampled from the training data, while $\boldsymbol{Z} \sim \mathbb{Q}$ are sampled by the opponent

See Wadsworth et al. (2018), Xu et al. (2021), Lima et al. (2022) for achieving fairness through adversarial learning

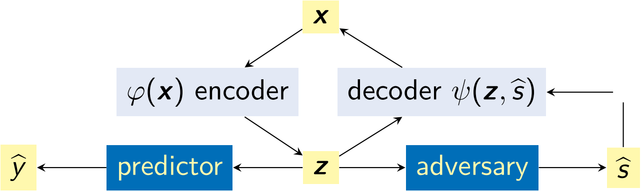
In-Processing

FairGAN, Xu et al. (2018)

Pre-processing approach actually, with demographic parity (DP)

Other algorithms are in-processing approaches, with demographic parity (DP) and equalized odds (EO)

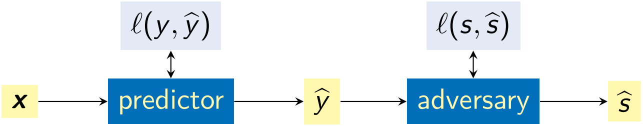
Learning adversarially fair and transferable representations, Madras et al. (2018)



Adversarially learning fair representations, Beutel et al. (2017)

Fair Adversarial Debiasing Approach, Zhang et al. (2018)

In-Processing



Following Zhang et al. (2018)
the predictor predicts y given $\boldsymbol{x}$,
the adversary tries to predict s bases on the output of the predictor

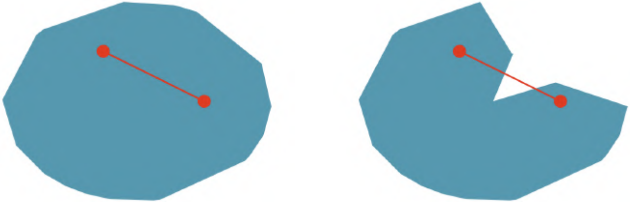
the predictor targets to increase its prediction accuracy
and tries to increase the adversary's loss

Barycenter

Several approaches can be considered to define means, averages, centroids, barycenters (etc.), as discussed in Fréchet (1948) and Grove and Karcher (1973),

- convex properties (from Möbius (1827) and Rockafellar (1970))
- axiomatization (from Nagumo (1930), Kolmogorov (1930) and Aczél (1948))
- optimization (from Hey (1814), Nathan (1952) and Agueh and Carlier (2011))

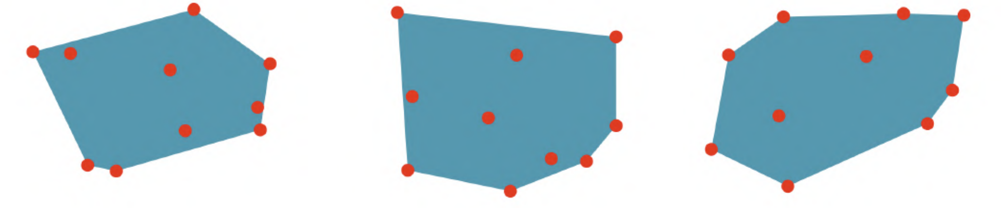
$C \subset \mathbb{R}^n$ is convex if $\boldsymbol{x}, \boldsymbol{y} \in C \Longrightarrow t\boldsymbol{x} + (1-t)\boldsymbol{y} \in C$ for all $t \in [0,1]$



Barycenter

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k \in \mathbb{R}^n$, then a convex combination is any linear combination $\omega_1 \boldsymbol{x}_1 + \cdots + \omega_k \boldsymbol{x}_k$ with $(\omega_1, \cdots, \omega_k) \in \mathcal{S}_k \subset \mathbb{R}_+$.

The convex hull of a set C is the set of all convex combinations of elements of C.



The geometric centroid of a convex object always lies in the object.

Barycenter

Define the barycenter for two points, with equal weights as a function $M : E \times E \rightarrow E$

- Reflexivity: $M(\boldsymbol{x}, \boldsymbol{x}) = \boldsymbol{x}$,
- Symmetry: $M(\boldsymbol{x}_1, \boldsymbol{x}_2) = M(\boldsymbol{x}_2, \boldsymbol{x}_1)$,
- Continuity: $M(\cdot, \cdot)$ is continuous,
- Bisymmetry: $M(M(\boldsymbol{x}_{11}, \boldsymbol{x}_{12}), M(\boldsymbol{x}_{21}, \boldsymbol{x}_{22})) = M(M(\boldsymbol{x}_{11}, \boldsymbol{x}_{21}), M(\boldsymbol{x}_{12}, \boldsymbol{x}_{22}))$

Then (see Aczél (1948)), there is f such that

$$M(\boldsymbol{x}_1, \boldsymbol{x}_2) = f^{-1}\left(\frac{1}{2}f(\boldsymbol{x}_1) + \frac{1}{2}f(\boldsymbol{x}_2)\right).$$

If $E \subset \mathbb{R}^k$, consider means on each coordinate axis independently.

A natural extension is

$$B_f(\boldsymbol{x}, \boldsymbol{\omega}) = f^{-1}\left(\sum_{i=1}^{n} \omega_i f(x_i)\right).$$

Barycenter

For the optimisation approach, given a distance d on E, set

$$B_d(\boldsymbol{x},\boldsymbol{\omega})=\underset{z\in E}{\text{argmin}}\left\{\sum_{i=1}^n \omega_i d(x_i,z)\right\}$$

Consider some points $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k\}$ in a metric space $\mathbb{R}^2$
The mean is

$$\overline{\boldsymbol{x}}=\frac{\boldsymbol{x}_1+\boldsymbol{x}_2+\cdots+\boldsymbol{x}_k}{k}=\frac{1}{k}\sum_{i=1}^k \boldsymbol{x}_i,$$

or equivalently

$$\overline{\boldsymbol{x}}=\underset{\boldsymbol{x}}{\text{argmin}}\left\{\frac{1}{k}\sum_{i=1}^k \|\boldsymbol{x}-\boldsymbol{x}_i\|_{\ell_2}^2\right\}.$$

But they can be defined using any distance/divergence/discrepancy

Barycenter

Instead of points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k\}$ in the metric space $\mathbb{R}^2$, we can consider some measures $\{\mathbb{P}_1, \mathbb{P}_2, \cdots, \mathbb{P}_k\}$.

The Euclidean mean is

$$\overline{\mathbb{Q}} = \underset{\mathbb{Q}}{\text{argmin}} \left\{ \frac{1}{k} \sum_{i=1}^{k} \Delta^2(\mathbb{Q}, \mathbb{P}_i) \right\},$$

where $\Delta^2(\mathbb{Q}, \mathbb{P}_i) = \displaystyle\int_{\mathbb{R}^2} \big(d\mathbb{Q} - d\mathbb{P}_i\big)^2$.

But any discrepancy function can be considered

One can consider Wasserstein discrepancy

Barycenter

Definition 12.2: Wasserstein W_2 Barycenter, Agueh and Carlier (2011)

$$\overline{\mathbb{Q}} = \underset{\mathbb{Q}}{\text{argmin}} \left\{ \sum_{i=1}^{k} \omega_i W_2(\mathbb{Q}, \mathbb{P}_i)^2 \right\},$$

This can be seen as a multi-marginal optimal transport problem.
Recall that the "*push-forward*" measure is

$$\mathbb{P}_1(\mathcal{A}) = \mathcal{T}_{\#}\mathbb{P}_0(\mathcal{A}) = \mathbb{P}_0(\mathcal{T}^{-1}(\mathcal{A})),\ \forall \mathcal{A} \subset \mathbb{R}.$$

An optimal transport $\mathcal{T}^\star$ (in Brenier's sense, from Brenier (1991), see Villani (2009) or Galichon (2016)) from $\mathbb{P}_0$ towards $\mathbb{P}_1$ will be solution of

$$\mathcal{T}^\star \in \underset{\mathcal{T}:\mathcal{T}_{\#}\mathbb{P}_0=\mathbb{P}_1}{\text{arginf}} \left\{ \int_{\mathbb{R}^k} \ell(\boldsymbol{x}, \mathcal{T}(\boldsymbol{x})) d\mathbb{P}_0(\boldsymbol{x}) \right\},$$

Barycenter

and for univariate distributions, the optimal transport $\mathcal{T}^\star$ is the monotone transformation.

$$\mathcal{T}^\star : x_0 \mapsto x_1 = F_1^{-1} \circ F_0(x_0).$$

Given a reference measure, say $\mathbb{P}_1$, it is possible to write the barycenter as the "*average push-forward*" transformation of $\mathbb{P}_1$: if $\mathbb{P}_i = \mathcal{T}_\#^{1\to i}\mathbb{P}_1$ (with the convention that $\mathcal{T}_\#^{1\to 1}$ is the identity),

Proposition 12.3: Wasserstein W_2 Barycenter,

$$\overline{\mathbb{Q}} = \left(\sum_{i=1}^{k} \omega_i \mathcal{T}^{1\to i}\right)_\# \mathbb{P}_1.$$

Barycenter

Proposition 12.4: Wasserstein W_2 Barycenter,

$$\overline{\mathbb{Q}} = \left(\sum_{i=1}^{k} \omega_i \mathcal{T}^{1\to i}\right)_{\#} \mathbb{P}_1 .$$

Computation of barycenters can be computationnaly difficult, Altschuler and Boix-Adsera (2021)

For univariate distributions, there is a simple expression, $\mathcal{T}^{1\to i}$ is simply a rearrangement, defined as $\mathcal{T}^{1\to i} = F_i^{-1} \circ F_1$, where $F_i(t) = \mathbb{P}_i((-\infty, t])$ and F_i^{-1} is its generalized inverse

Barycenter

Proposition 12.5: Wasserstein W_2 Barycenter, univariate distributions

$\mathcal{T}^{1\to i}$ is simply a rearrangement, defined as $\mathcal{T}^{1\to i} = F_i^{-1} \circ F_1$, where $F_i(t) = \mathbb{P}_i((-\infty, t])$, and

$$\overline{\mathbb{Q}} = \left(\sum_{i=1}^{n} k\omega_i \mathcal{T}^{1\to i}\right)_{\#} \mathbb{P}_1.$$

Proposition 12.6: Wasserstein W_2 Barycenter, univariate distributions

Given two scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the "fair barycenter score" is

$$\begin{cases} m^\star(\boldsymbol{x}, s = \mathtt{A}) = \mathbb{P}[S = \mathtt{A}] \cdot m(\boldsymbol{x}, s = \mathtt{A}) + \mathbb{P}[S = \mathtt{B}] \cdot F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}(m(\boldsymbol{x}, s = \mathtt{A})) \\ m^\star(\boldsymbol{x}, s = \mathtt{B}) = \mathbb{P}[S = \mathtt{A}] \cdot F_{\mathtt{A}}^{-1} \circ F_{\mathtt{B}}(m(\boldsymbol{x}, s = \mathtt{B})) + \mathbb{P}[S = \mathtt{B}] \cdot m(\boldsymbol{x}, s = \mathtt{B}). \end{cases}$$

Barycenter

that is generally numerically intractable (computing one subgradient requires solving k optimal transports)

In the discrete case, if we consider a fixed grid (so that C can be computed once only)

$$\min_{a}\left\{\sum_{i=1}^{k}\min_{P_i\in U_{a,b_i}}\left\{\langle P_i, C\rangle\right\}\right\},$$

We can write this as a large linear program

$$\min_{\mathbb{Q}}\left\{\min_{P_1,\cdots,P_k,a}\sum_{i=1}^{k}\left\{\langle P_i, C\rangle\right\}\right\},\text{ where }\begin{cases}P_1^\top\mathbf{1}_n=\boldsymbol{b}_1\\ \quad\vdots\\ P_k^\top\mathbf{1}_n=\boldsymbol{b}_k\\ P_1\mathbf{1}_n=\cdots=P_k\mathbf{1}_n=\boldsymbol{a}\end{cases}$$

Barycenter

Theorem 12.1: Variance $\boldsymbol{\Sigma}$

If $k = 2$, $\boldsymbol{\Sigma}$ satisfies

$$\boldsymbol{\Sigma} = \omega_1 \left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{1/2}\right)^{1/2} + \omega_2 \left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}^{1/2}\right)^{1/2}$$

and the explicit expression is

$$\boldsymbol{\Sigma} = \omega_1^2\boldsymbol{\Sigma}_1 + \omega_2^2\boldsymbol{\Sigma}_2 + \omega_1\omega_2\Big(\boldsymbol{\Sigma}_1^{\frac{1}{2}}(\boldsymbol{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{\frac{1}{2}})^{\frac{1}{2}}\boldsymbol{\Sigma}_1^{-\frac{1}{2}} + \boldsymbol{\Sigma}_1^{-\frac{1}{2}}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{\frac{1}{2}})^{\frac{1}{2}}\boldsymbol{\Sigma}_1^{\frac{1}{2}}\Big)$$

Proposition 12.7: Variance $\boldsymbol{\Sigma}$

$\displaystyle\sum_{i=1}^{k} \omega_i \boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}$ is a positive matrix.

Barycenter

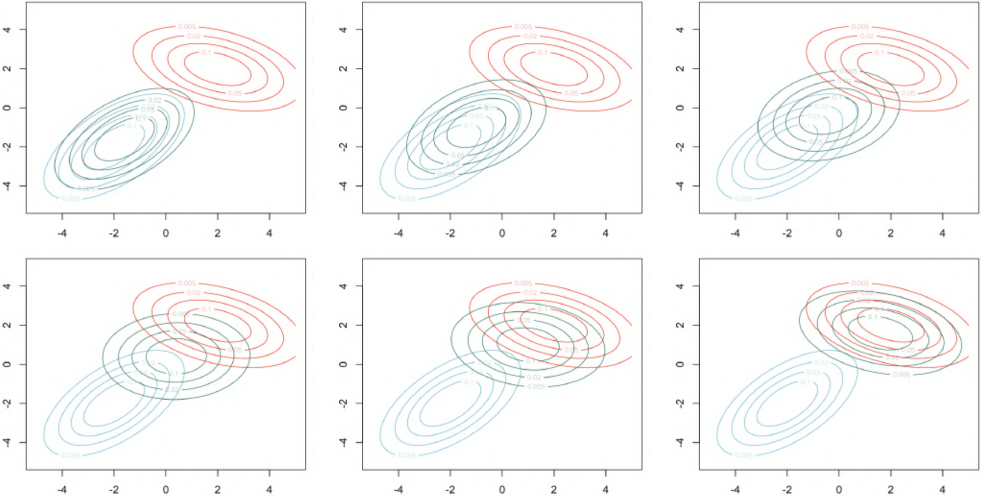
Proposition 12.8: Variance $\boldsymbol{\Sigma}$

If $\boldsymbol{\Sigma}_i = \boldsymbol{P}\boldsymbol{\Delta}_i\boldsymbol{P}^\top$, then $\boldsymbol{\Sigma} = \boldsymbol{P}\left(\sum_{i=1}^{k}\omega_i\boldsymbol{\Delta}_i^{\frac{1}{2}}\right)^2\boldsymbol{P}^\top$

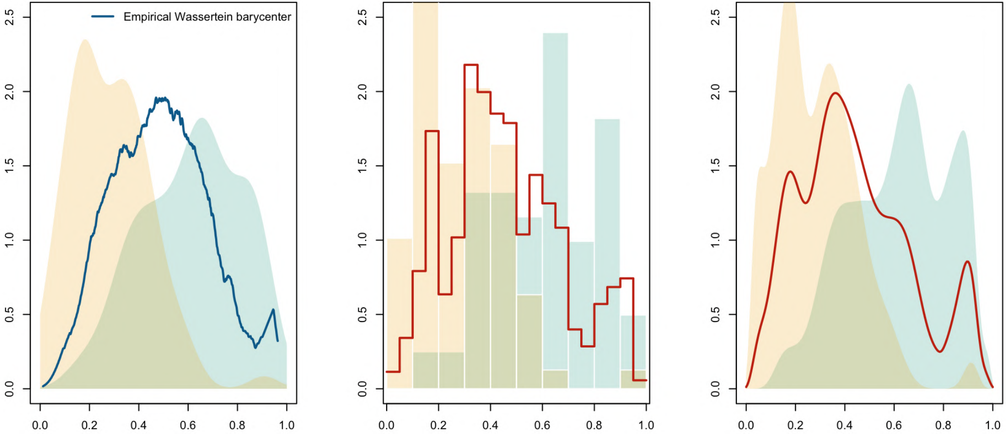
Consider two Gaussian distributions, $\mathcal{N}(\boldsymbol{\mu}_{\texttt{A}}, \boldsymbol{\Sigma}_{\texttt{A}})$ and $\mathcal{N}(\boldsymbol{\mu}_{\texttt{B}}, \boldsymbol{\Sigma}_{\texttt{B}})$, and weights $\omega_{\texttt{A}} = t$ and $\omega_{\texttt{B}} = 1 - t$, with $t \in [0, 1]$.

Barycenter

Barycenter of two bivariate Gaussian distribution ($t = 0.1, 0.25, 0.4, 0.6, 0.75, 0.9$)



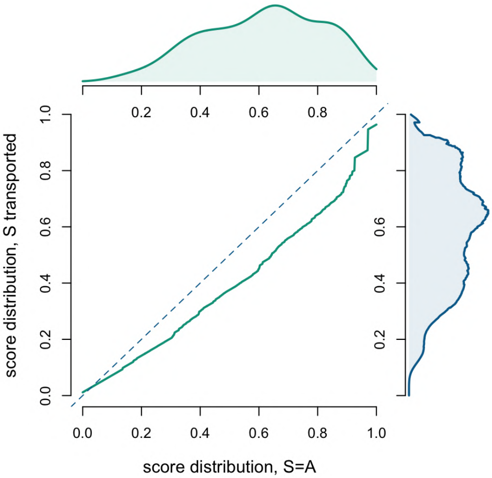
Application to `toydata1`



Application to `toydata1`

Given scores $m(\boldsymbol{x}, s = \texttt{A})$ and $m(\boldsymbol{x}, s = \texttt{B})$, the "fair barycenter score" is

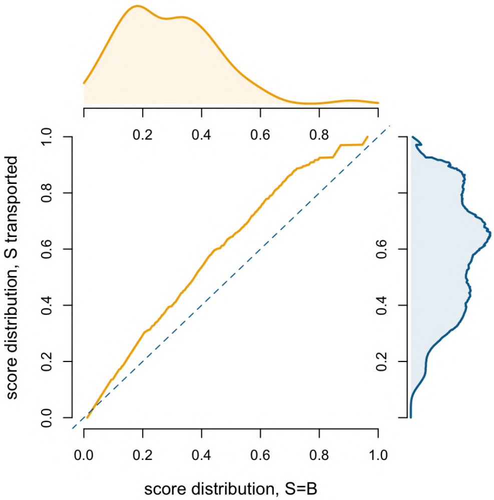
$$\begin{aligned} & m^{\star}(\boldsymbol{x}, s = \texttt{A}) \\ = \; & \mathbb{P}[S = \texttt{A}] \cdot m(\boldsymbol{x}, s = \texttt{A}) \\ + \; & \mathbb{P}[S = \texttt{B}] \cdot F_{\texttt{B}}^{-1} \circ F_{\texttt{A}}\big(m(\boldsymbol{x}, s = \texttt{A})\big) \end{aligned}$$



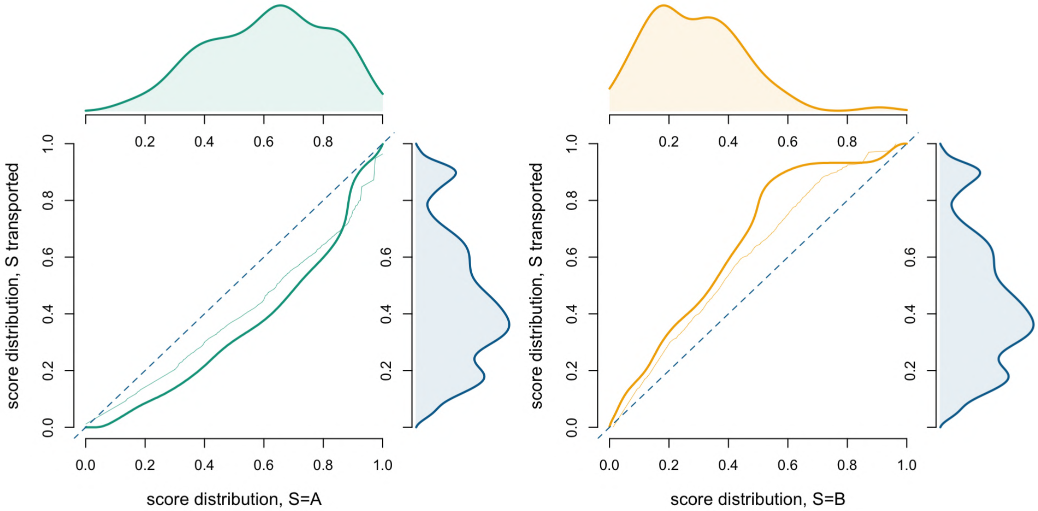
Application to `toydata1`

Given scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the "fair barycenter score" is

$$
\begin{aligned}
& m^{\star}(\boldsymbol{x}, s = \mathtt{B}) \\
= \quad & \mathbb{P}[S = \mathtt{A}] \cdot F_{\mathtt{A}}^{-1} \circ F_{\mathtt{B}}\big(m(\boldsymbol{x}, s = \mathtt{B})\big) \\
+ \quad & \mathbb{P}[S = \mathtt{B}] \cdot m(\boldsymbol{x}, s = \mathtt{B})
\end{aligned}
$$



Application to `toydata1`



Application to `toydata1`

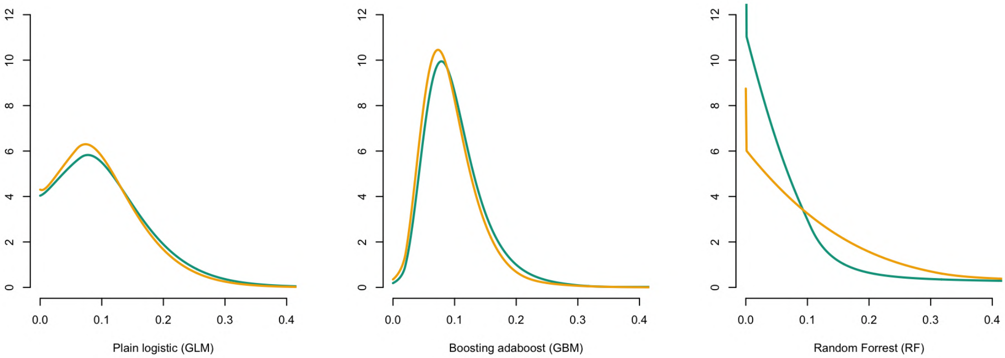
	x	s	$\overline{y}$	$\widehat{m}(x,s)$	$\widehat{m}(x)$	$\widehat{m}^*_{\text{w}}(x)$	$\widehat{m}^*_{\text{jkl}}(x)$
Alex	-1	`A`	0.475	0.250	0.219	0.154	0.094
Betty	-1	`B`	0.475	0.205	0.219	0.459	0.357
Ahmad	0	`A`	0.475	0.490	0.465	0.341	0.279
Brienne	0	`B`	0.475	0.426	0.465	0.719	0.692
Anthony	+1	`A`	0.475	0.734	0.730	0.571	0.521
Beatrix	+1	`B`	0.475	0.681	0.730	0.842	0.932

Application to `FrenchMotor`

If the two models are balanced, $m^\star$ is also balanced.

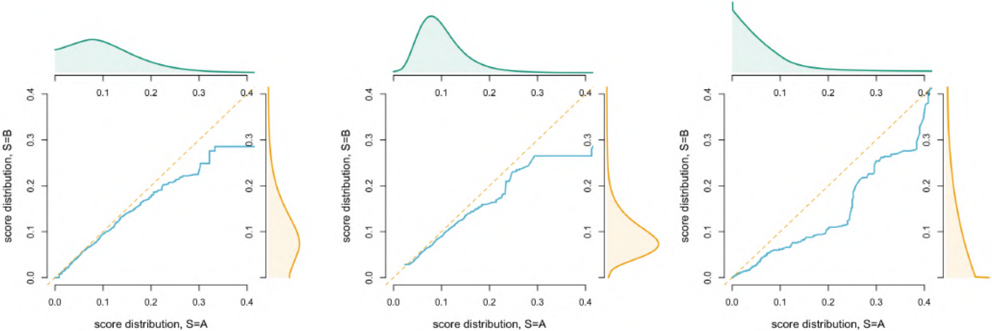
Annual claim occurrence (motor insurance, Charpentier et al. (2023b))

Three models (plain GLM, GBM, Random Forest)



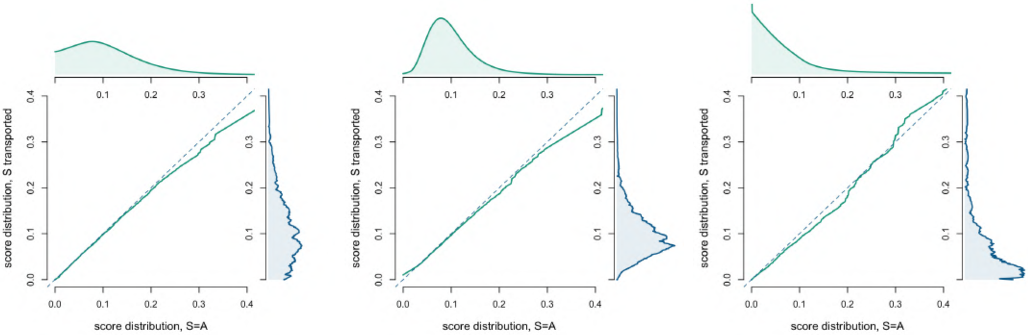
Application to FrenchMotor

Predictions are different for men $(= \texttt{A})$ and women $(S = \texttt{B})$



since $W_2 \neq 0$ consider post processing mitigation

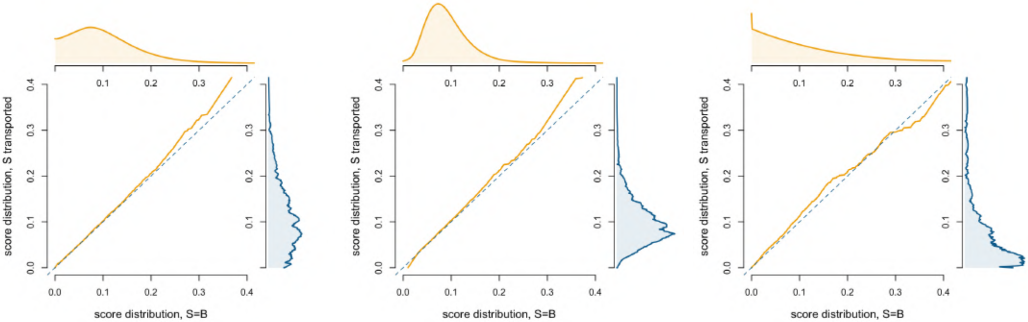
Application to FrenchMotor



Given scores $m(\boldsymbol{x}, s = \texttt{A})$ and $m(\boldsymbol{x}, s = \texttt{B})$, the "fair barycenter score" is

$$m^\star(\boldsymbol{x}, s = \texttt{A}) = \mathbb{P}[S = \texttt{A}] \cdot m(\boldsymbol{x}, s = \texttt{A}) + \mathbb{P}[S = \texttt{B}] \cdot F_{\texttt{B}}^{-1} \circ F_{\texttt{A}}\big(m(\boldsymbol{x}, s = \texttt{A})\big)$$

Application to FrenchMotor

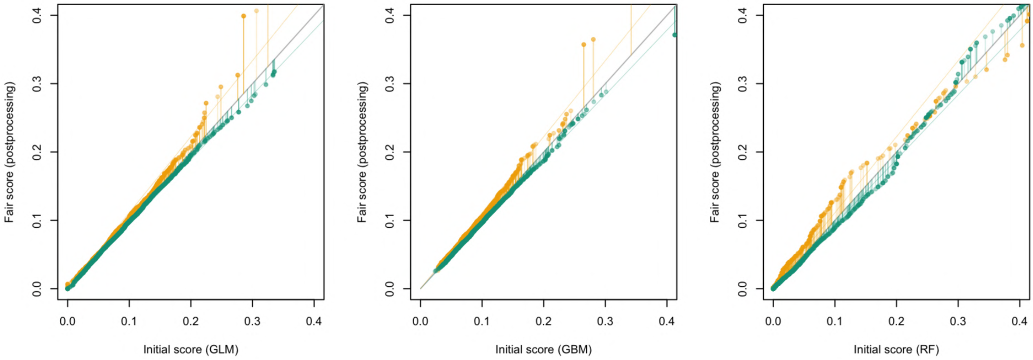


Given scores $m(\boldsymbol{x}, s=\texttt{A})$ and $m(\boldsymbol{x}, s=\texttt{B})$, the “fair barycenter score” is

$$m^\star(\boldsymbol{x}, s=\texttt{B}) = \mathbb{P}[S=\texttt{A}] \cdot F_{\texttt{A}}^{-1} \circ F_{\texttt{B}}(m(\boldsymbol{x}, s=\texttt{B})) + \mathbb{P}[S=\texttt{B}] \cdot m(\boldsymbol{x}, s=\texttt{B})$$

Application to `FrenchMotor`

We can plot $\{(m(\boldsymbol{x}_i, \texttt{A}), m^\star(\boldsymbol{x}_i, \texttt{A})\}$ and $\{(m(\boldsymbol{x}_i, \texttt{B}), m^\star(\boldsymbol{x}_i, \texttt{B})\}$



Application to `FrenchMotor`

Numerical values, for initial occurence probability of 5%, 10% and 20%, we have

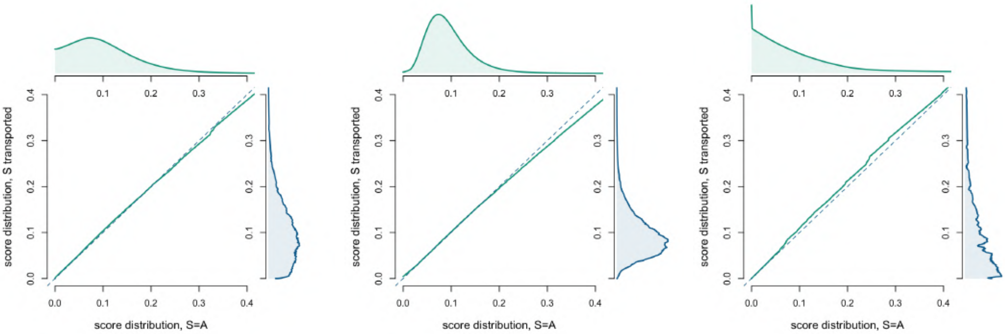
	`A` (men)				`B` (women)			
	×0.94	GLM	GBM	RF	×1.11	GLM	GBM	RF
$m(\boldsymbol{x}) = \ \ 5\%$	4.73%	4.94%	4.80%	4.42%	5.56%	5.16%	5.25%	6.15%
$m(\boldsymbol{x}) = 10\%$	9.46%	9.83%	9.66%	8.92%	11.12%	10.38%	10.49%	12.80%
$m(\boldsymbol{x}) = 20\%$	18.91%	19.50%	18.68%	18.26%	22.25%	20.77%	21.63%	21.12%

Application to `FrenchMotor`

We can do the same for discrimination against "old" drivers.

	`A` (younger < 65)				`B` (old > 65)			
	$\times 1.01$	GLM	GBM	RF	$\times 0.94$	GLM	GBM	RF
$m(\boldsymbol{x}) = 5\%$	5.05%	5.17%	5.10%	5.27%	4.71%	3.84%	3.84%	3.96%
$m(\boldsymbol{x}) = 10\%$	10.09%	10.37%	10.16%	11.00%	9.42%	7.81%	9.10%	6.88%
$m(\boldsymbol{x}) = 20\%$	20.19%	19.98%	19.65%	21.26%	18.85%	19.78%	23.79%	12.54%

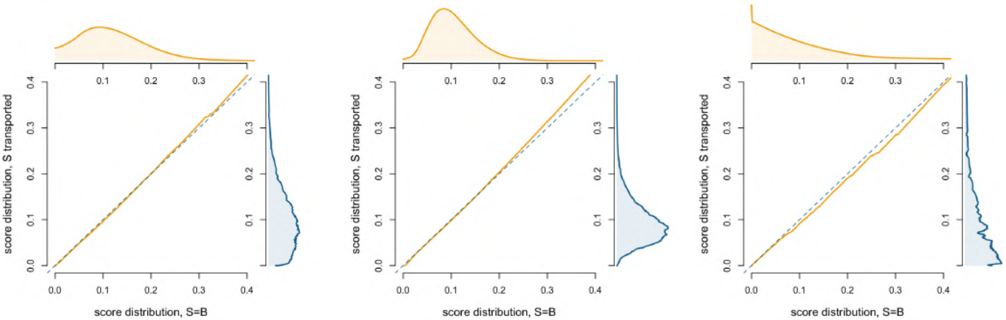
Application to FrenchMotor



Given scores $m(\boldsymbol{x}, s = \texttt{A})$ and $m(\boldsymbol{x}, s = \texttt{B})$, the "fair barycenter score" is

$$m^{\star}(\boldsymbol{x}, s = \texttt{A}) = \mathbb{P}[S = \texttt{A}] \cdot m(\boldsymbol{x}, s = \texttt{A}) + \mathbb{P}[S = \texttt{B}] \cdot F_{\texttt{B}}^{-1} \circ F_{\texttt{A}}\big(m(\boldsymbol{x}, s = \texttt{A})\big)$$

Application to `FrenchMotor`

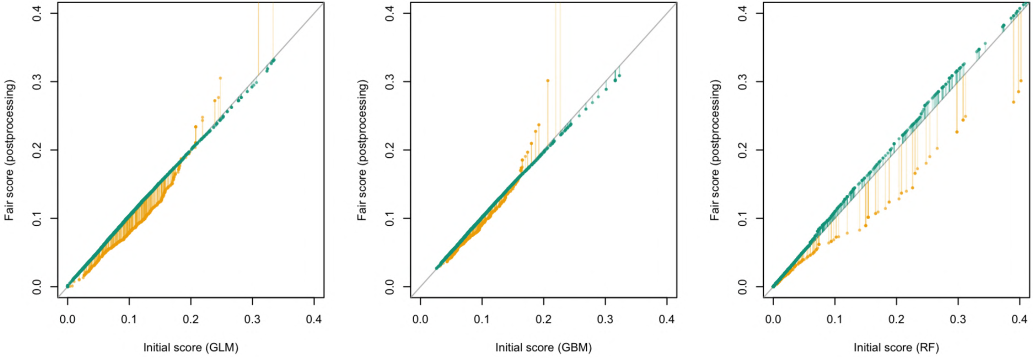


Given scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the “fair barycenter score” is

$$m^\star(\boldsymbol{x}, s = \mathtt{B}) = \mathbb{P}[S = \mathtt{A}] \cdot F_{\mathtt{A}}^{-1} \circ F_{\mathtt{B}}\big(m(\boldsymbol{x}, s = \mathtt{B})\big) + \mathbb{P}[S = \mathtt{B}] \cdot m(\boldsymbol{x}, s = \mathtt{B})$$

Application to `FrenchMotor`

We can plot $\{(m(\boldsymbol{x}_i,\texttt{A}),m^\star(\boldsymbol{x}_i,\texttt{A})\}$ and $\{(m(\boldsymbol{x}_i,\texttt{B}),m^\star(\boldsymbol{x}_i,\texttt{B})\}$



– Part 10 –

Non-Observed Sensitive Attributes

Bayesian Surname Geocoding

Method	First name	Last name	Geo location	Other	Reference
GO	☐	☐	☑	☐	Fiscella and Fremont (2006)
SA	☐	☑	☐	☐	Lauderdale and Kestenbaum (2000)
CSG	☐	☑	☑	☐	Fiscella and Fremont (2006)
BSG	☐	☑	☑	☐	Elliott et al. (2008)
BISG	☐	☑	☑	☐	Elliott et al. (2009)
MBISCG	☑	☑	☑	☑	Martino et al. (2013)
BIFSG	☑	☑	☑	☐	Voicu (2018)
Regression	☑	☑	☑	☑	Xue et al. (2019)

GO (Geocoding Only); SA (Surname Analysis); CSG (Categorical Surname and Geocoding); BSG (Bayesian Surname Geocoding); BISG (Improved BSG); MBISCG (Medicare BISG); BIFSG(BISG with First Name)

Bayesian Surname Geocoding

First names and their associated race/ethnicity prevalences, Tzioumis (2018)
comprehensive list of 4,250 first names
Census 2010 surname list, Word et al. (2008)
160,000 surnames, covering about 90 percent of the U.S. population
Decennial Census 2010 SF1 datase
GO, SA, CSG, pre-Bayesian methods

- GO, Fiscella and Fremont (2006), Elliott et al. (2008)
  Krieger et al. (2002)
- SA, Elliott et al. (2008), Word and Perkins (1996)
- CSG, Fiscella and Fremont (2006)

Bayesian Surname Geocoding

Bayes's Theorem, $\mathbb{P}[A|B] = \dfrac{\mathbb{P}[B|A]\cdot\mathbb{P}[A]}{\mathbb{P}[B]} = \dfrac{\mathbb{P}[B|A]\cdot\mathbb{P}[A]}{\mathbb{P}[B|A]\cdot\mathbb{P}[A] + \mathbb{P}[B|\overline{A}]\cdot\mathbb{P}[\overline{A}]}$

BSG, $\mathbb{P}[\text{race} = r|\text{surname} = s]$ is

$$\frac{\mathbb{P}[\text{surname} = s|\text{race} = r]\cdot\mathbb{P}[\text{race} = r]}{\mathbb{P}[\text{surname} = s|\text{race} = r]\cdot\mathbb{P}[\text{race} = r] + \mathbb{P}[\text{surname} = s|\text{race} \neq r]\cdot\mathbb{P}[\text{race} \neq r]}$$

BIFSG, $\mathbb{P}[\text{race} = r|\text{first name} = f, \text{surname} = s, \text{geolocalisation} = g]$, Voicu (2018)

$$\mathbb{P}[r|f,s,g] = \frac{\mathbb{P}[r|s]\cdot\mathbb{P}[g|r]\cdot\mathbb{P}[f|r]}{\sum_t \mathbb{P}[R = t|s]\cdot\mathbb{P}[g|R = t]\cdot\mathbb{P}[f|R = t}$$

$$\text{Assumption : } \begin{cases} \mathbb{P}[g|r,s] = \mathbb{P}[g|r] \text{ or } G \perp\!\!\!\perp R \mid S \\ \mathbb{P}[f|r,s,g] = \mathbb{P}[f|r] \end{cases}$$

"Given the race, the geolocation is not informative about the surname"

Bayesian Surname Geocoding

$\mathbb{P}[r|s]$ is the probability that a person is of race/ethnicity r, given that the person has surname s, (i.e., the surname-based probability described above); $\mathbb{P}[f|r]$ is the probability that a person has first name f, given that the person is of race/ethnicity r (i.e., the aforementioned first-name-based probability);$\mathbb{P}[g|r]$ is the probability that a person resides in geographic area g, given that the person is of race/ethnicity r (i.e., the aforementioned geography-based probability); and the summation in the denominator occurs over the six race/ethnicity categories defined previously

Bayesian Surname Geocoding

BIFSG Bayesian First Name Surname Geocode
BISG = Bayesian Improved Surname and Geocoding
BISG computes the probability of race given a voter's surname and geographic location, $\mathbb{P}(R=r|S=s,G=g)$, using Bayes theorem. Assuming $G \perp\!\!\!\perp S|R$,

$$\mathbb{P}(R=r|S=s,G=g) \propto \mathbb{P}(G=g|R=r)\cdot\mathbb{P}(R=r|S=s)$$

The probability $\mathbb{P}(G=g|R=r)$ can be obtained from Census summary tables by taking the number of people of race/ethnicity $R=r$ in neighborhood $G=g$ divided by the total number of people of race/ethnicity r.
The probability of race given surname, $\mathbb{P}(R=r|S=s)$, comes directly from the Census Bureau's surname lists which contain the proportion of all Decennial Census respondents with each surname in each racial-ethnic category
Decter-Frain (2022)

Bayesian Surname Geocoding

Fully Bayesian Improved Surname Geocoding (fBISG)
BISG suffers from two data problems regarding minorities:

- the census often contains zero counts
  ↪ fBISG uses a measurement error model so that zero values mean low probability instead of nonexistence
- many surnames are missing from the census data
  ↪ fBISG also supplemens the surname list with additional data from voter files from six Southern states

BISG Elliott et al. (2009)

$$\mathbb{P}(R_i|S_i,G_i)\propto\mathbb{P}(S_i|R_i)\cdot\mathbb{P}(R_i|G_i)$$

$\mathbb{P}(R_i=r|G_i=g)\propto N_{rg}$, obtained from US census data.

Bayesian Surname Geocoding

fBISG Imai and Khanna (2016)

$$\mathbb{P}(R_i|S_i, G_i) \propto \mathbb{P}(R_i|S_i) \cdot \mathbb{P}(G_i|R_i)$$

$\mathbb{P}(R_i = r|G_i = g, R_{-i}) \propto n_{rg}^{-i} + N_{rg} + 1 > 0$, with:

- the term $+1$ arises from the assumption of a Dirichlet prior distribution over the race distribution for geolocation g,
- n_{rg}^{-i} is obtained using Gibbs sampling on the dataset of individuals whose race is being predicted, by conditioning on the race of other individuals R_{-i} in geolocation g.
- Minorities continue to be underestimated. They are absorbed by the majority
- How can we give more power to the minorities?

– Appendix –

Additional Results

Appendix: Sensitive attributes in insurance (in the U.S.)

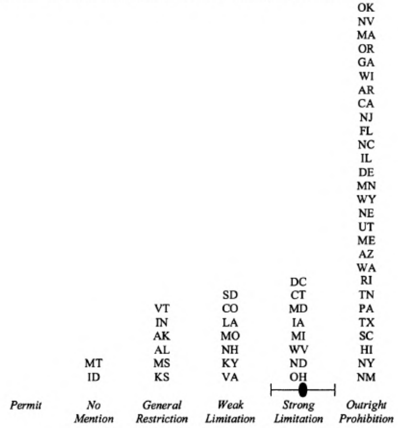
From Avraham et al. (2013),

Expressly Permit (-1) - The state has a statute expressly or impliedly permitting insurers to take the characteristic into account.

No Law on Point (0) - The state laws are silent with respect to the particular characteristic.

General Restriction (1) - The state has a statute that generally prohibits "unfair discrimination," either across all lines of insurance or in some lines of insurance, but that statute does not provide any explanation as to what constitutes unfair discrimination and does not single out any particular trait for limitation.

FIGURE 1a. Distribution of States' Scores for Race, in Auto Insurance



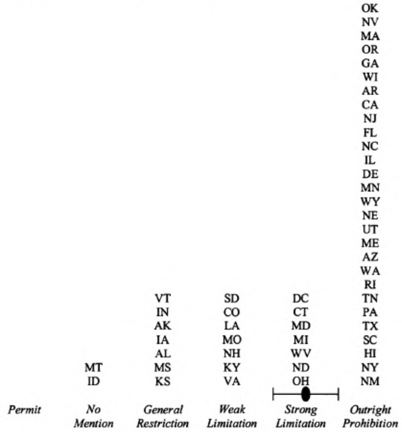
Appendix: Sensitive attributes in insurance (in the U.S.)

Characteristic-Specific Weak Limitation (2) - The state has a statute that limits the use of a particular characteristic in either issuance, renewal, or cancellation.

Characteristic-Specific Strong Limitation (3) - The state has a statute that prohibits the use of a particular characteristic when the policy is either issued, renewed, or cancelled, or the state has a statute that limits but does not completely prohibit the use of a particular characteristic in rate setting.

Characteristic-Specific Prohibition (4) - The state has a statute the expressly prohibits insurers from taking into account a specific characteristic in setting rates.

FIGURE 1b. Distribution of States' Scores for National Origin, in Auto Insurance



Appendix: Sensitive attributes in insurance (in the U.S.)

"*Race, national origin, and religion have a special place in this country's history; and, as discussed above, discrimination on the basis of these three characteristics has been subject to stricter scrutiny in American law than have other characteristics*," Avraham et al. (2013)

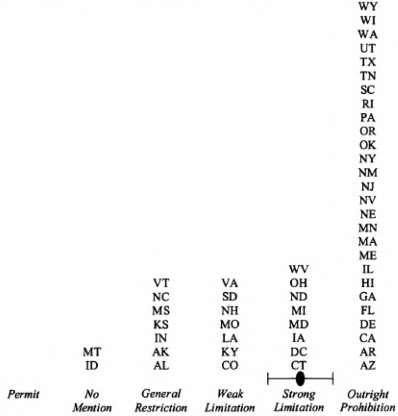
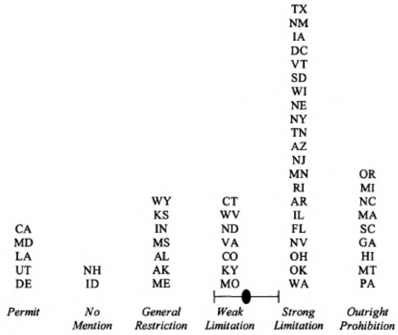


FIGURE 1c. Distribution of States' Scores for Religion, in Auto Insurance

Appendix: Sensitive attributes in insurance (in the U.S.)

"*Gender-based discrimination in insurance has long been controversial. And differential treatment on the basis of gender is, of course, in many contexts widely considered unacceptable or illegal. Nevertheless, there does not seem to be the same level of agreement-as there is for race, religion, and national origin-that drawing gender-based distinctions is always wrong. Federal constitutional law treats gender as only a quasi-suspect classification; as a result, laws that discriminate on the basis of gender are subject to an intermediate level of scrutiny.*" Avraham et al. (2013)

FIGURE 3a. Distribution of States' Scores for Gender, in Auto Insurance

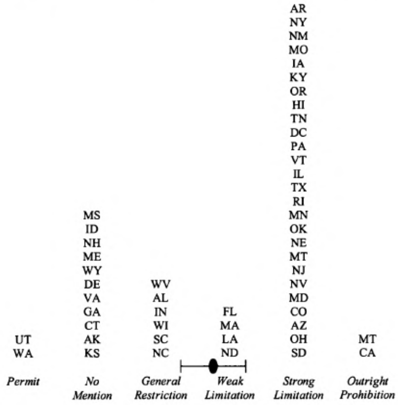


Permit	*No Mention*	*General Restriction*	*Weak Limitation*	*Strong Limitation*	*Outright Prohibition*
CA MD LA UT DE	NH ID	WY KS IN MS AL AK ME	CT WV ND VA CO KY MO	TX NM IA DC VT SD WI NE NY TN AZ NJ MN RI AR IL FL NV OH OK WA	OR MI NC MA SC GA HI MT PA

Appendix: Sensitive attributes in insurance (in the U.S.)

"*With respect to life insurance, we predict that the laws regulating gender discrimination will be on average relatively weak, since adverse selection in the life insurance market is especially problematic.*" Avraham et al. (2013)

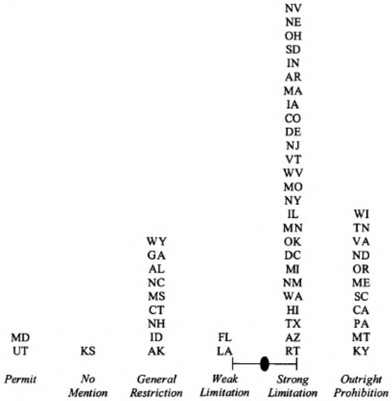
FIGURE 3c. Distribution of States' Scores for Gender, in Disability Insurance



Appendix: Sensitive attributes in insurance (in the U.S.)

"*Regarding property/casualty insurance, as there seems to be no conceivable correlation between those risks and gender, we predict either states will cluster around no regulation, or, alternatively, states will cluster around forbidding the use of gender in property/casualty insurance on symbolic or expressive grounds.*" Avraham et al. (2013)

FIGURE 3d. Distribution of States' Scores for Gender, in Property/Casualty Insurance



Appendix: Sensitive attributes in insurance (in the U.S.)

"*The gender discrimination will be more strictly regulated on average for health insurance (where gender-rated policies often result in higher premiums for women) than for auto insurance (where gender-rated policies result in higher premiums for men).*" Avraham et al. (2013)

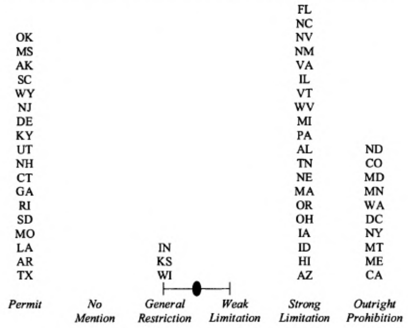
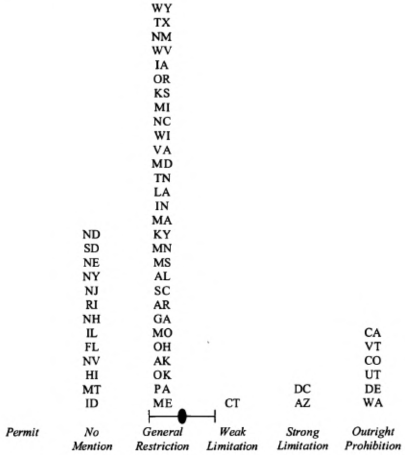


FIGURE 3e. Distribution of States' Scores for Gender, in Health Insurance

Appendix: Sensitive attributes in insurance (in the U.S.)

"*Unlike with race, national origin, religion, and gender, legal classifications on the basis of an individual's sexual orientation have not clearly been identified by the Supreme Court as deserving special scrutiny. In addition, unlike race, national origin, and gender, there are no federal laws forbidding discrimination on the basis of sexual orientation in employment.*" Avraham et al. (2013)

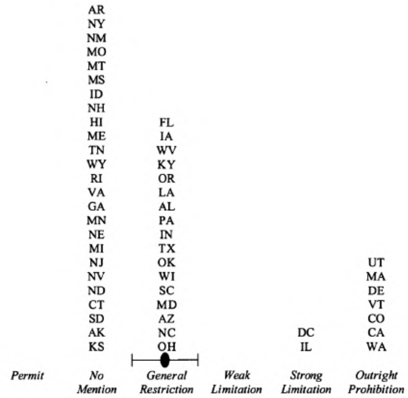
FIGURE 4a. Distribution of States' Scores for Sexual Orientation, in Auto Insurance



Appendix: Sensitive attributes in insurance (in the U.S.)

"*However, there are state laws that forbid discrimination on the basis of sexual orientation, and some lower courts have held that sexual orientation should be a suspect or quasi-suspect characterisation.*" Avraham et al. (2013)

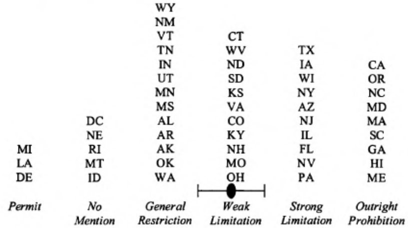
FIGURE 4c. Distribution of States' Scores for Sexual Orientation, in Disability Insurance



Appendix: Sensitive attributes in insurance (in the U.S.)

"*We expect that* age *will have the lowest average regulatory score of all the risk characteristics we are studying. First, age is not a suspect classification, at least not by constitutional standards. Second, age tends to correlate causally with several important areas of risk (mortality, health, and perhaps disability risks), thereby increasing the perceived fairness of rating on that basis.*" Avraham et al. (2013)

FIGURE 5a. Distribution of States' Scores for Age, in Auto Insurance

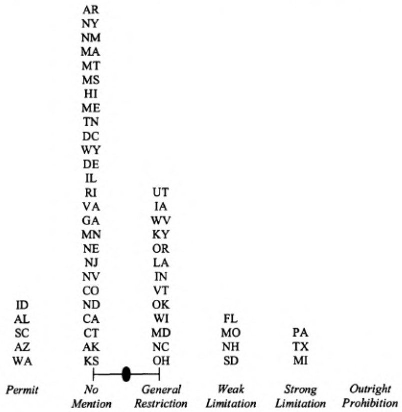


Permit	No Mention	General Restriction	Weak Limitation	Strong Limitation	Outright Prohibition
		WY			
		NM			
		VT	CT		
		TN	WV	TX	
		IN	ND	IA	CA
		UT	SD	WI	OR
		MN	KS	NY	NC
		MS	VA	AZ	MD
	DC	AL	CO	NJ	MA
	NE	AR	KY	IL	SC
MI	RI	AK	NH	FL	GA
LA	MT	OK	MO	NV	HI
DE	ID	WA	OH	PA	ME

Appendix: Sensitive attributes in insurance (in the U.S.)

"*Third, age can present serious adverse selection problems for insurers if they are forbidden from taking it into account, since individual insureds know their own age and the associated risks. Fourth, social solidarity arguments with respect to age are relatively weak, since individuals can spread risk over their lifetime through various income smoothing products*." Avraham et al. (2013)

FIGURE 5c. Distribution of States' Scores for Age, in Disability Insurance



Appendix: Sensitive attributes in insurance (in the U.S.)

Avraham et al. (2013) suggested to visualize the distribution of scores (Expressly Permit (-1) / No Law on Point (0) / General Restriction (1) / $\cdots$ / Characteristic-Specific Prohibition (4))

FIGURE 6. Distribution of States' Scores for Age, by Line of Insurance

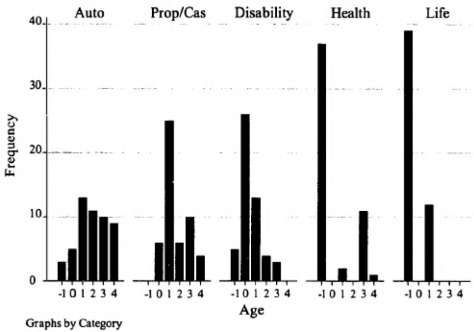
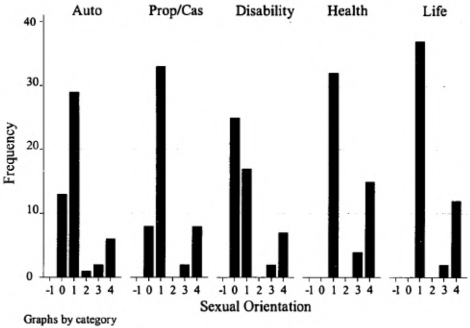


FIGURE 7. Distribution of States' Scores for Sexual Orientation, by Line of Insurance



Appendix: Sensitive attributes in insurance (in the U.S.)

FIGURE 8. Distribution of States' Scores for Zip Code, by Line of Insurance

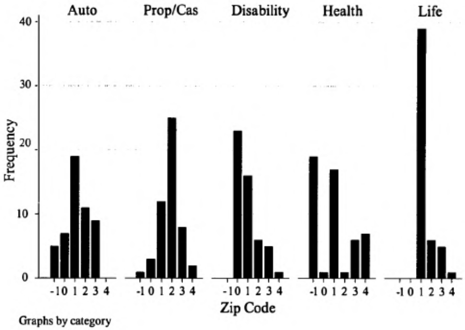
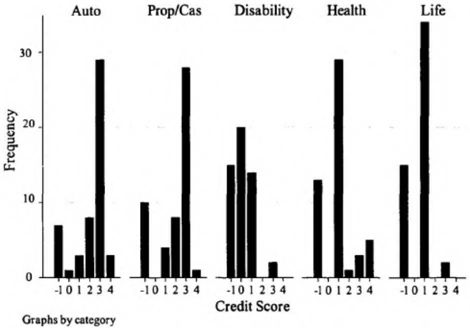


FIGURE 9. Distribution of States' Scores for Credit Score, by Line of Insurance



"*Credit score and zip code are not, by themselves, socially suspect characteristics. However, some commentators have argued that credit score and zip code are used by auto and home insurers as proxies for potentially socially suspect characteristics.*"

References

Abraham, K. (1986). *Distributing risk: Insurance, legal theory and public policy*. Yale University Press,.

Aczél, J. (1948). On mean values. *Bulletin of the American Mathematical Society*, 54(4):392–400.

Agueh, M. and Carlier, G. (2011). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Ajunwa, I. (2014). Genetic testing meets big data: Tort and contract law issues. *Ohio State Law Journal*, 75:1225.

Ajunwa, I. (2016). Genetic data and civil rights. *Harvard Civil Rights-Civil Liberties Law Review*, 51:75.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.

Altschuler, J. M. and Boix-Adsera, E. (2021). Wasserstein barycenters can be computed in polynomial time in fixed dimension. *The Journal of Machine Learning Research*, 22(1):2000–2018.

Amari, S.-i. (2016). *Information geometry and its applications*, volume 194. Springer.

Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 249–260.

References

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*, May 23.

Apfelbaum, E. P., Pauker, K., Sommers, S. R., and Ambady, N. (2010). In blind pursuit of racial equality? *Psychological science*, 21(11):1587–1592.

Austin, P. C. and Steyerberg, E. W. (2019). The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, 38:4051 – 4065.

Austin, R. (1983). The insurance classification controversy. *University of Pennsylvania Law Review*, 131(3):517–583.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. *IJCAI International Joint Conference on Artificial Intelligence*, pages 357–363.

Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.

Avraham, R., Logue, K. D., and Schwarcz, D. (2013). Understanding insurance antidiscrimination law. *Southern California Law Review*, 87:195.

References

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

Bailey, R. A. and Simon, L. J. (1960). Two studies in automobile insurance ratemaking. *ASTIN Bulletin: The Journal of the IAA*, 1(4):192–217.

Baldus, D. C. and Cole, J. W. (1980). *Statistical proof of discrimination*. McGraw-Hill.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(10).

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Barbour, V. (1911). Privateers and pirates of the west indies. *The American Historical Review*, 16(3):529–566.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.

Barry, L. and Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society*, 7(1):2053951720935143.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

References

Bauschke, H. H., Borwein, J. M., et al. (1997). Legendre functions and the method of random bregman projections. *Journal of convex analysis*, 4(1):27–67.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

Bellemare, M. G., Dabney, W., and Munos, R. (2017a). A distributional perspective on reinforcement learning. *arXiv:1707.06887*.

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. (2017b). The cramer distance as a solution to biased wasserstein gradients. *arXiv:1705.10743*.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*, volume 28. Princeton university press.

Bender, M., Dill, C., Hurlbert, M., Lindberg, C., and Mott, S. (2022). Understanding potential influences of racial bias on p&c insurance: Four rating factors explored. CAS Research Paper Series on Race and Insurance Pricing.

Bengio, Y., Goodfellow, I., and Courville, A. (2017). *Deep learning*, volume 1. MIT press Cambridge, MA, USA.

Bernstein, E. (2007). *Temporarily Yours: Intimacy, Authenticity, and the Commerce of Sex*. University of Chicago Press.

References

Bertillon, A. and Chervin, A. (1909). *Anthropologie métrique: conseils pratiques aux missionnaires scientifiques sur la manière de mesurer, de photographier et de décrire des sujets vivants et des pièces anatomiques*. Imprimerie nationale.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv*, 1707.00075.

Bhattacharya, A. (2015). Facebook patent: Your friends could help you get a loan - or not. *CNN Business*, 2015/08/04.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404.

Biddle, D. (2017). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.

Billingsley, P. (2017). *Probability and measure*. John Wiley & Sons.

Blanpain, N. (2018). L'espérance de vie par niveau de vie-méthode et principaux résultats. *INSEE Document de Travail*, F1801.

Blier-Wong, C., Cossette, H., Lamontagne, L., and Marceau, E. (2021). Geographic ratemaking with spatial embeddings. *ASTIN Bulletin: The Journal of the IAA*, pages 1–31.

References

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455.

Blumenbach, J. F. (1775). *De generis humani varietate nativa*. Vandenhoek & Ruprecht.

Bogachev, V. I., Kolesnikov, A. V., and Medvedev, K. V. (2005). Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309.

Bollen, K. A. (1989). *Structural equations with latent variables*, volume 210. John Wiley & Sons.

Bollen, K. A. and Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pages 301–328. Springer.

Bollobás, B. (1998). *Random graphs*. Springer.

Bonnotte, N. (2013). From Knothe's rearrangement to Brenier's optimal transport map. *SIAM Journal on Mathematical Analysis*, 45(1):64–87.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Bouk, D. (2015). *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual*. The University of Chicago Press.

References

Bourdieu, P. (2018). Distinction a social critique of the judgement of taste. In *Inequality Classic Readings in Race, Class, and Gender*, pages 287–318. Routledge.

Box, G. E., Luceño, A., and del Carmen Paniagua-Quinones, M. (2011). *Statistical control by monitoring and adjustment*, volume 700. John Wiley & Sons.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Brilmayer, L., Hekeler, R. W., Laycock, D., and Sullivan, T. A. (1979). Sex discrimination in employer-sponsored insurance plans: A legal and demographic analysis. *University of Chicago Law Review*, 47:505.

Britz, G. (2008). *Einzelfallgerechtigkeit versus Generalisierung: verfassungsrechtliche Grenzen statistischer Diskriminierung*. Mohr Siebeck.

References

Brown, R. S., Moon, M., and Zoloth, B. S. (1980). Incorporating occupational attainment in studies of male-female earnings differentials. *Journal of Human Resources*, pages 3–28.

Brualdi, R. A. (2006). *Combinatorial matrix classes*, volume 13. Cambridge University Press.

Calders, T. and Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292.

Cantwell, G. T., Kirkley, A., and Newman, M. E. (2021). The friendship paradox in real and model networks. *Journal of Complex Networks*, 9(2):cnab011.

Carlier, G., Galichon, A., and Santambrogio, F. (2010). From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Duxbury Advanced Series.

Casey, B., Pezier, J., and Spetzler, C. (1976). *The Role of Risk Classification in Property and Casualty Insurance: A Study of the Risk Assessment Process : Final Report*. Stanford Research Institute.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions, intl. *Journal of Mathematical Models and Methods in Applied Sciences, Issue*, 4.

Chambert-Loir, A. (2023). *Information Theory: Three Theorems by Claude Shannon*, volume 144. Springer Nature.

References

Charniak, E. (1991). Bayesian networks without tears. *AI magazine*, 12(4):50–50.

Charpentier, A. (2014). Mesures de risque. In Droesbeke, J.-J. and Saporta, G., editors, *Approches statistiques du risque*. Éditions Technip.

Charpentier, A. (2023). *Insurance: biases, discrimination and fairness*. Springer Verlag.

Charpentier, A., Flachaire, E., and Gallic, E. (2023a). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, *Optimal Transport Statistics for Economics and Related Topics*. Springer Verlag.

Charpentier, A., Hu, F., and Ratz, P. (2023b). Mitigating discrimination in insurance with wasserstein barycenters. *BIAS, 3rd Workshop on Bias and Fairness in AI, International Workshop of ECML PKDD*.

Cheney-Lippold, J. (2017). We are data. In *We Are Data*. New York University Press.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Chzhen, E. and Schreuder, N. (2022). A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442.

References

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2):115–124.

Cohen, I. and Goldszmidt, M. (2004). Properties and benefits of calibrated classifiers. In *8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 3202, pages 125–136. Springer.

Conway, D. A. and Roberts, H. V. (1983). Reverse regression, fairness, and employment discrimination. *Journal of Business & Economic Statistics*, 1(1):75–85.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.

Coulmont, B. and Simon, P. (2019). Quels prénoms les immigrés donnent-ils à leurs enfants en france? *Population Societes*, (4):1–4.

Cramér, H. (1928a). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Cramér, H. (1928b). On the composition of elementary errors: second paper: statistical applications. *Scandinavian Actuarial Journal*, 1928(1):141–180.

Crossney, K. B. (2016). Redlining. *https://philadelphiaencyclopedia.org/essays/redlining/*.

References

Csiszár, I. (1964). Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8:85–108.

Csiszár, I. (1967). On information-type measure of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

Cunningham, S. (2021). *Causal inference*. Yale University Press.

Da Silva, N. (2023). *La bataille de la Sécu: une histoire du système de santé*. La fabrique éditions.

Dall'Aglio, G. (1956). Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 10(1-2):35–74.

Daniel, W. W. et al. (1968). *Racial discrimination in England: based on the PEP report*. Penguin Books.

Danskin, J. M. (1967). *The theory of max-min and its application to weapons allocation problems*. Springer.

Darlington, R. B. (1971). Another look at "cultural fairness". *Journal of educational measurement*, 8(2):71–82.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. (2017). Proxy non-discrimination in data-driven systems. *arXiv*, 1707.08120.

References

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

De Baere, G. and Goessens, E. (2011). Gender differentiation in insurance contracts after the judgment in case c-236/09, Association Belge des Consommateurs Test-Achats asbl v. conseil des ministres. *Colum. J. Eur. L.*, 18:339.

de La Fontaine, J. (1668). *Fables*. Barbin.

de Melo-Martín, I. (2003). When is biology destiny? biological determinism and social responsibility. *Philosophy of science*, 70(5):1184–1194.

De Pril, N. and Dhaene, J. (1996). Segmentering in verzekeringen. *DTEW Research Report 9648*, pages 1–56.

De Wit, G. and Van Eeghen, J. (1984). Rate making and society's sense of fairness. *ASTIN Bulletin: The Journal of the IAA*, 14(2):151–163.

Dean, L. T. and Nicholas, L. H. (2018). Using credit scores to understand predictors and consequences of disease.

References

Dean, L. T., Schmitz, K. H., Frick, K. D., Nicholas, L. H., Zhang, Y., Subramanian, S., and Visvanathan, K. (2018). Consumer credit as a novel marker for economic burden and health after cancer in a diverse population of breast cancer survivors in the usa. *Journal of Cancer Survivorship*, 12(3):306–315.

Decter-Frain, A. (2022). How should we proxy for race/ethnicity? comparing bayesian improved surname geocoding to machine learning methods. *arXiv*, 2206.14583.

Dedecker, J. and Merlevède, F. (2007). The empirical distribution function for dependent variables: asymptotic and nonasymptotic results in. *ESAIM: Probability and Statistics*, 11:102–114.

Denis, C., Elie, R., Hebiri, M., and Hu, F. (2021). Fairness guarantee in multi-class classification. *arXiv*, 2109.13642.

Denuit, M. and Charpentier, A. (2004). *Mathématiques de l'assurance non-vie: Tome I Principes fondamentaux de théorie du risque*. Economica.

Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics & Economics*.

Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional gaussians with the same mean. *arXiv*, 1810.08693.

References

Dhaene, J., Denuit, M., Goovaerts, M. J., Kaas, R., and Vyncke, D. (2002a). The concept of comonotonicity in actuarial science and finance: applications. *Insurance: Mathematics and Economics*, 31(2):133–161.

Dhaene, J., Denuit, M., Goovaerts, M. J., Kaas, R., and Vyncke, D. (2002b). The concept of comonotonicity in actuarial science and finance: theory. *Insurance: Mathematics and Economics*, 31(1):3–33.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.

Dostie, G. (1974). Entrevue de michèle lalonde. *Le Journal*, 1er juin 1974.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

Du Bois, W. E. B. (1899). *The philadelphia negro*. Cosimo, Inc.

Duncan, O. D. (1975). *Introduction to structural equation models*. Academic Press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

References

Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., and Lurie, N. (2008). A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health services research*, 43(5p1):1722–1736.

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., and Lurie, N. (2009). Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9:69–83.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.

Feeley, M. and Simon, J. (1994). Actuarial justice: The emerging new criminal law. *The futures of criminology*, 173:174.

Feeley, M. M. and Simon, J. (1992). The new penology: Notes on the emerging strategy of corrections and its implications. *Criminology*, 30(4):449–474.

References

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, volume 1412.3756, pages 259–268.

Feller, A., Pierson, E., Corbett-Davies, S., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, October 17.

Fiscella, K. and Fremont, A. M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health services research*, 41(4p1):1482–1500.

Fourcade, M. and Healy, K. (2013). Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society*, 38(8):559–572.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569.

Fox, E. T. (2013). *'Piratical Schemes and Contracts': Pirate Articles and Their Society 1660-1730*. PhD Thesis, University of Exeter.

François, P. (2022). Catégorisation, individualisation. retour sur les scores de crédit. *hal*, 03508245.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310.

References

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, 3e serie, Sciences, Sect. A*, 14:53–77.

Freeman, L. C. et al. (1979). Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge*, 1:238–263.

Freeman, S. (2007). *Rawls*. Routledge.

Frees, E. W. and Huang, F. (2021). The discriminating (pricing) actuary. *North American Actuarial Journal*, pages 1–23.

Gaddis, S. M. (2017). How black are lakisha and jamal? racial perceptions from names used in correspondence audit studies. *Sociological Science*, 4:469–489.

Gadet, F. (2007). *La variation sociale en français*. Editions Ophrys.

Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.

Gandy, O. H. (2016). *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*. Routledge.

Gangbo, W. (1999). The monge mass transfer problem and its applications. *Contemporary Mathematics*, 226:79–104.

Gannon, M. (2016). Race is a social construct, scientists argue. *Scientific American*, 5:1–11.

References

Garrioch, D. (2011). Mutual aid societies in eighteenth-century paris. *French History & Civilization*, 4.

Gebelein, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Ginsburg, M. (1940). Roman military clubs and their social functions. In *Transactions and Proceedings of the American Philological Association*, volume 71, pages 149–156. JSTOR.

Givens, C. R. and Shortt, R. M. (1984). A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240.

Glenn, B. J. (2000). The shifting rhetoric of insurance denial. *Law and Society Review*, pages 779–808.

Glenn, B. J. (2003). Postmodernism: the basis of insurance. *Risk Management and Insurance Review*, 6(2):131–143.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, 310(5746):248–249.

References

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica*, pages 979–1001.

Goldman, A. (1979). *Justice and Reverse Discrimination*. Princeton University Press.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114.

Gosseries, A. (2014). What makes age discrimination special: A philosophical look at the ecj case law. *Netherlands Journal of Legal Philosophy*, 43:59–80.

Gouic, T. L., Loubes, J.-M., and Rigollet, P. (2020). Projection to fairness in statistical learning. *arXiv*, 2005.11720.

Gowri, A. (2014). *The Irony of Insurance: Community and Commodity*. PhD thesis, University of Southern California.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. (2005). Kernel constrained covariance for dependence measurement. In *International Workshop on Artificial Intelligence and Statistics*, pages 112–119. PMLR.

References

Grove, K. and Karcher, H. (1973). How to conjugate c 1-close group actions. *Mathematische Zeitschrift*, 132(1):11–20.

Gupta, S., Anderson, R. M., and May, R. M. (1989). Networks of sexual contacts: implications for the pattern of spread of hiv. *Aids*, 3(12):807–818.

Hacking, I. (1990). *The taming of chance*. Number 17. Cambridge University Press.

Hakimi, S. L. (1962). On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):496–506.

Hall, K. M. (1970). An r-dimensional quadratic placement algorithm. *Management science*, 17(3):219–229.

Hannan, E. J. (1961). The general theory of canonical correlation and its relation to functional analysis. *Journal of the Australian Mathematical Society*, 2(2):229–242.

Harari, Y. N. (2018). *21 Lessons for the 21st Century*. Random House.

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, 27(4):237–243.

Harden, K. P. (2023). Genetic determinism, essentialism and reductionism: semantic clarity for contested science. *Nature Reviews Genetics*, 24(3):197–204.

References

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Hardy, G. H., Littlewood, J. E., Pólya, G., Pólya, G., et al. (1952). *Inequalities*. Cambridge university press.

Harris, T. and Ross, F. (1955). Fundamentals of a method for evaluating rail net capacities. Technical report.

Havel, V. (1955). A remark on the existence of finite graphs. *Časopis pro pěstování' matematiky*, 80:477–480.

Havens, H. V. (1979). Issues and needed improvements in state regulation of the insurance business. *U.S. General Acounting Office*.

He, X. D., Kou, S., and Peng, X. (2022). Risk measures: robustness, elicitability, and backtesting. *Annual Review of Statistics and Its Application*, 9:141–166.

Hedges, B. A. (1977). Gender discrimination in pension plans: Comment. *The Journal of Risk and Insurance*, 44(1):141–144.

Heimer, C. A. (1985). *Reactive Risk and Rational Action*. University of California Press.

Heller, D. (2015). High price of mandatory auto insurance in predominantly african american communities. Technical report, Consumer Federation of America.

References

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271.

Henrion, M. (1988). Propagating uncertainty in bayesian networks by probabilistic logic sampling. In *Machine intelligence and pattern recognition*, volume 5, pages 149–163. Elsevier.

Hey, R. (1814). Xviii. propositions containing some properties of tangents to circles; and of trapeziums inscribed in circles, and non-inscribed. together with propositions on the elliptic representations of circles, upon a plane surface, by perspective. *Philosophical Transactions of the Royal Society of London*, (104):348–396.

Higham, N. J. (2008). *Functions of matrices: theory and computation*. SIAM.

Hill, K. and White, J. (2020). Designed to deceive: do these people look real to you? *The New York Times*, 11(21).

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524.

Hoeffding, W. (1940). Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts und Instituts fur Angewandte Mathematik der Universitat Berlin*, 5:181–233.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

References

Hoffman, F. L. (1896). *Race traits and tendencies of the American Negro*, volume 11. American Economic Association.

Hoffman, F. L. (1918). *Mortality from respiratory diseases in dusty trades (inorganic dusts)*. Number 231. US Government Printing Office.

Hoffman, F. L. (1931). Cancer and smoking habits. *Annals of surgery*, 93(1):50.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Hubbard, G. N. (1852). *De l'organisation des sociétés de bienfaisance ou de secours mutuels et des bases scientifiques sur lesquelles elles doivent être établies*. Paris, Guillaumin.

Huttegger, S. M. (2013). In defense of reflection. *Philosophy of Science*, 80(3):413–433.

Imai, K. and Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2):263–272.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

References

Ismay, P. (2018). *Trust among strangers: friendly societies in modern Britain*. Cambridge University Press.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise de Sciences Naturelles*, 37:547–579.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Jensen, D. and Mayer, L. (1977). Some variational results and their applications in multiple inference. *The Annals of Statistics*, pages 922–931.

Johannesson, G. T. (2013). *The history of Iceland*. ABC-CLIO.

Jordan, C. (1881). Sur la serie de fourier. *Camptes Rendus Hebdomadaires de l'Academie des Sciences*, 92:228–230.

Kantorovich, L. and Rubinstein, G. (1958). On the space of completely additive functions. *Vestnic Leningrad Univ., Ser. Mat. Mekh. i Astron.*, 13(7):52–59. In Russian.

Kantorovich, L. V. (1942). On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201.

References

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta psychologica*, 77(3):217–273.

Kerner, O. (1968). *Report of The National Advisory Commission on Civil Disorder*. Bantam Books.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *arXiv*, 1706.02744.

Kim, P. T. (2017). Auditing algorithms for discrimination. *University of Pennsylvania Law Review*, 166:189.

Kimeldorf, G., May, J. H., and Sampson, A. R. (1982). Concordant and discordant monotone correlations and their evaluation by nonlinear optimization. *Studies in the Management Sciences*, 19:117–130.

Kimeldorf, G. and Sampson, A. R. (1978). Monotone dependence. *The Annals of Statistics*, pages 895–903.

References

Kitagawa, E. M. (1955). Components of a difference between two rates. *Journal of the american statistical association*, 50(272):1168–1194.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv*, 1609.05807.

Knothe, H. (1957). Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52.

Knott, M. and Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49.

Knowlton, R. E. (1978). Regents of the university of california v. bakke. *Arkansas Law Review*, 32:499.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91.

Kolmogorov, A. N. (1930). *Sur la notion de la moyenne*. G. Bardi, tip. della R. Accad. dei Lincei.

Komiyama, J. and Shimao, H. (2017). Two-stage algorithm for fairness-aware machine learning. *arXiv*, 1710.04924.

References

Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.

Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M.-J., Subramanian, S., and Carson, R. (2002). Geocoding and monitoring of us socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? the public health disparities geocoding project. *American journal of epidemiology*, 156(5):471–482.

Krüger, F. and Ziegel, J. F. (2021). Generic conditions for forecast dominance. *Journal of Business & Economic Statistics*, 39(4):972–983.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Kuhn, H. W. (1956). Variants of the hungarian method for assignment problems. *Naval research logistics quarterly*, 3(4):253–258.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

References

Kumar, A., Liang, P. S., and Ma, T. (2019). Verified uncertainty calibration. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, 44(1/2):289–292.

Lancaster, H. O. (1958). The Structure of Bivariate Distributions. *The Annals of Mathematical Statistics*, 29(3):719 – 736.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. *ProPublica*, 23-05.

Lauderdale, D. S. and Kestenbaum, B. (2000). Asian american ethnic identification by surname. *Population Research and Policy Review*, 19:283–300.

Lee, A. J. (2019). *U-statistics: Theory and Practice*. CRC Press.

Leeson, P. T. (2009). The calculus of piratical consent: the myth of the myth of social contract. *Public Choice*, 139:443–459.

Léon, P. R. (1993). *Précis de phonostylistique: parole et expressivité/Pierre R. Léon,..* Nathan.

References

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Li, G., Braver, E. R., and Chen, L.-H. (2003). Fragility versus excessive crash involvement as determinants of high death rates per vehicle-mile of travel among older drivers. *Accident Analysis & Prevention*, 35(2):227–235.

Li, K. C.-W. (1996). The private insurance industry's tactics against suspected homosexuals: redlining based on occupation, residence and marital status. *American Journal of Law & Medicine*, 22(4):477–502.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1977). Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, pages 275–324.

Lima, L. F. F. P. d., Ricarte, D. R. D., and Siebra, C. d. A. (2022). An overview on the use of adversarial learning strategies to ensure fairness in machine learning models. In *XVIII Brazilian Symposium on Information Systems*, pages 1–8.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Lin, P.-E. (1987). Measures of asociation between vectors. *Communications in Statistics-Theory and Methods*, 16(2):321–338.

References

Linn, R. L. and Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, 8(1):1–4.

Lippert-Rasmussen, K. (2020). *Making sense of affirmative action*. Oxford University Press.

Lombroso, C. (1876). *L'uomo delinquente*. Hoepli.

Luong, B. T., Ruggieri, S., and Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510.

Macnicol, J. (2006). *Age discrimination: An historical and contemporary analysis*. Cambridge University Press.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR.

Martin, G. D. (1977). Gender discrimination in pension plans: Author's reply. *The Journal of Risk and Insurance*, 44(1):145–149.

Martino, S. C., Weinick, R. M., Kanouse, D. E., Brown, J. A., Haviland, A. M., Goldstein, E., Adams, J. L., Hambarsoomian, K., Klein, D. J., and Elliott, M. N. (2013). Reporting cahps and hedis data by race/ethnicity for medicare beneficiaries. *Health services research*, 48(2pt1):417–434.

References

Massey, D. S. (2007). *Categorically unequal: The American stratification system*. Russell Sage Foundation.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall.

Mcdonald, S. (2015). Indirect gender discrimination and the 'test-achats ruling': an examination of the uk motor insurance market. In *Royal Economic Society Conf., Manchester*.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Merriam-Webster (2022). *Dictionary*. .

Möbius, A. F. (1827). *Der barycentrische Calcul, ein Hülfsmittel zur analytischen Behandlung der Geometrie (etc.)*. Leipzig: J.A. Barth.

Morris, M. (1995). Data driven network models for the spread of infectious disease. *Epidemic models: their structure and relation to data*, 5:302–22.

Mourier, E. (1953). Eléments aléatoires dans un espace de banach. In *Annales de l'institut Henri Poincaré*, volume 13, pages 161–244.

Mowbray, A. (1921). Classification of risks as the basis of insurance rate making with special reference to workmen's compensation. *Proceedings of the Casualty Actuarial Society*.

References

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600.

Murphy, A. H. and Epstein, E. S. (1967). Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology and Climatology*, 6(5):748–755.

Myers, R. J. (1977). Gender discrimination in pension plans: Further comment. *The Journal of Risk and Insurance*, 44(1):144–145.

Nagumo, M. (1930). Über eine klasse der mittelwerte. In *Japanese journal of mathematics*, volume 7, pages 71–79. The Mathematical Society of Japan.

Nathan, A. (1952). *College Geometry: An Introduction to the Modern Geometry of the Triangle and the Circle*. Barnes & Noble.

National Association of Insurance Commissioners (2011). A consumer's guide to auto insurance. *NAIC Reports*.

National Association of Insurance Commissioners (2022). A consumer's guide to auto insurance. *NAIC Reports*.

References

Neumann, J. v. and Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

Newman, M. (2018). *Networks*. Oxford university press.

Newman, M. E. (2003). Mixing patterns in networks. *Physical review E*, 67(2):026126.

Newman, M. E. (2012). Communities, modules and large-scale structure in networks. *Nature physics*, 8(1):25–31.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Nielsen, F. (2022). The many faces of information geometry. *Notices of the American Mathematical Society*, 69(1):36–45.

Nielsen, F. and Nock, R. (2013). On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13.

Oakes, D. (1985). Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390):339–339.

References

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709.

of the European Union, C. (2018). Proposal for a council directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disablility, age or sexual orientation. *Proceedings of Council of the European Union*, 11015/08.

Olkin, I. and Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Ortiz-Ospina, E. and Beltekian, D. (2018). Why do women live longer than men? *Our World in Data*.

Pakdaman Naeini, M., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1):2901–2907.

Pardo, L. (2018). *Statistical inference based on divergence measures*. CRC press.

Parlett, B. and Landis, T. (1982). Methods for scaling to doubly stochastic form. *Linear Algebra and its Applications*, 48:53–79.

References

Patel, N. V. (2017). Why doctors aren't afraid of better, more efficient ai diagnosing cancer. *The Daily Beast*, 11 Dec.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*, pages 15–17.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the royal society of London*, 58(347-352):240–242.

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Peters, T. (2014). *Playing God?: Genetic determinism and human freedon*. Routledge.

Petersen, F., Mukherjee, D., Sun, Y., and Yurochkin, M. (2021). Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955.

References

Plater, M. A. (1997). African-american insurance enterprises: An early vehicle for economic and social development. *Journal of Management History*, 3(1):42–58.

Pojman, L. P. (1998). The case against affirmative action. *International Journal of Applied Philosophy*, 12(1):97–115.

Polyanskiy, Y. and Wu, Y. (2022). *Information theory: From coding to learning*. Cambridge University Press.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

Prince, A. E. and Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, 105:1257.

Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214.

Proschan, M. A. and Presnell, B. (1998). Expect the unexpected from conditional expectation. *The American Statistician*, 52(3):248–252.

Rawls, J. (1999). *A theory of justice: Revised edition*. Harvard university press.

Reichenbach, H. (1956). *The direction of time*, volume 65. University of California Press.

References

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press.

Rhynhart, R. (2020). Mapping the legacy of structural racism in philadelphia. *Philadelphia, Office pf the Controller*.

Riach, P. A. and Rich, J. (1991). Measuring discrimination by direct experimental methods: Seeking gunsmoke. *Journal of Post Keynesian Economics*, 14(2):143–150.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rizzo, M. L. and Székely, G. J. (2016). Energy distance. *wiley interdisciplinary reviews: Computational statistics*, 8(1):27–38.

Robbins, L. A. (2015). The pernicious problem of ageism. *Generations*, 39(3):6–9.

Roberts, H. V. (1968). On the meaning of the probability of rain. In *first national conference on statistical meteorology*.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

References

Rorive, I. (2009). Proving discrimination cases: The role of situation testing. *Centre for Equal Rights and Migration Policy Group*.

Rosenbaum, P. (2018). *Observation and experiment*. Harvard University Press.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Rothstein, W. G. (2003). *Public health and the risk factor: A history of an uneven medical revolution*, volume 3. Boydell & Brewer.

Rouvroy, A., Berns, T., and Carey-Libbrecht, L. (2013). Algorithmic governmentality and prospects of emancipation. *Réseaux*, 177(1):163–196.

Rudin, W. (1966). *Real and Complex Analysis*. McGraw-hill New York.

Sabbagh, D. (2007). *Equality and transparency: A strategic perspective on affirmative action in American law*. Springer.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*, volume 55. Springer.

Sarmanov, O. (1958a). Maximum correlation coefficient (non-symmetrical case). *Doklady Akademii Nauk SSSR*, 121(1):52–55.

References

Sarmanov, O. V. (1958b). The maximum correlation coefficient (symmetrical case). *Doklady Akademii Nauk SSSR*, 120(4):715–718.

Schanze, E. (2013). Injustice by generalization: notes on the Test-Achats decision of the european court of justice. *German Law Journal*, 14(2):423–433.

Schauer, F. (2006). *Profiles, probabilities, and stereotypes*. Harvard University Press.

Schervish, M. J. and DeGroot, M. H. (2014). *Probability and statistics*, volume 563. Pearson Education London, UK:.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Simon, J. (1987). The emergence of a risk society-insurance, law, and the state. *Socialist Review*, (95):60–89.

Simon, J. (1988). The ideological effects of actuarial practices. *Law & Society Review*, 22:771.

Sinkhorn, R. (1962). On the factor spaces of the complex doubly stochastic matrices. *Notices of the American Mathematical Society*, 9:334–335.

References

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879.

Sinkhorn, R. (1966). A relationship between arbitrary positive matrices and stochastic matrices. *Canadian Journal of Mathematics*, 18:303–306.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281.

Spirtes, P., Glymour, C. N., and Scheines, R. (1993). *Causation, prediction, and search*. Springer Verlag.

Squires, G. D. (2003). Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs*, 25(4):391–410.

Squires, G. D. and Chadwick, J. (2006). Linguistic profiling: A continuing tradition of discrimination in the home insurance industry? *Urban Affairs Review*, 41(3):400–415.

Squires, G. D. and Velez, W. (1988). Insurance redlining and the process of discrimination. *The Review of Black Political Economy*, 16(3):63–75.

References

Stone, D. A. (1993). The struggle for the soul of health insurance. *Journal of Health Politics, Policy and Law*, 18(2):287–317.

Struyck, N. (1912). *Les oeuvres de Nicolas Struyck (1687-1769): qui se rapportent au calcul des chances, à la statistique général, la statistique des décès et aux rentes viagèter*. Société générale néerlandaise d'assurances sur la vie et de rentes viagères.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv*, 1312.6199.

Székely, G. J. (2003). E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18.

Tajfel, H. (1978). *Differentiation between social groups: Studies in the social psychology of intergroup relations.* Academic Press.

Tajfel, H., Turner, J. C., Worchel, S., Austin, W. G., et al. (1986). Psychology of intergroup relations. *Chicago: Nelson-Hall*, pages 7–24.

Takatsu, A. (2008). On wasserstein geometry of the space of gaussian measures. *arXiv*, 0801.2250.

Takatsu, A. and Yokota, T. (2012). Cone structure of l2-wasserstein spaces. *Journal of Topology and Analysis*, 4(02):237–253.

Telles, E. (2014). *Pigmentocracies: Ethnicity, race, and color in Latin America.* UNC Press Books.

References

The Zebra (2022). Car insurance rating factors by state. *https://www.thezebra.com/*.

Thiery, Y. and Van Schoubroeck, C. (2006). Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 31(2):190–211.

Thomas, L., Crook, J., and Edelman, D. (2002). *Credit scoring and its applications*. SIAM.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2):63–70.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Topkis, D. M. (1998). *Supermodularity and complementarity*. Princeton university press.

Tschantz, M. C. (2022). What is proxy discrimination? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1993–2003.

Turner, R. (2015). The way to stop discrimination on the basis of race. *Stanford Journal of Civil Rights & Civil Liberties*, 11:45.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. *Statistics: applications and new directions (Calcutta, 1981)*, pages 579–604.

Tzioumis, K. (2018). Demographic aspects of first names. *Scientific data*, 5(1):1–9.

References

US Census (2012). Frequently occurring surnames from census 2000, census report data file a: Top 1000 names. *Genealogy Data*.

Vallender, S. (1974). Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7.

Van Gerven, G. (1993). Case c-109/91, Gerardus Cornelis Ten Oever v. Stichting bedrijfspensioenfonds voor het glazenwassers-en schoonmaakbedrijf. *EUR-Lex*, 61991CC0109.

Van Rijsbergen, C. (1979). Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, volume 79.

Vapnik, V. (1991). Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.

Verboven, K. (2011). Introduction: Professional collegia: Guilds or social clubs? *Ancient Society*, pages 187–195.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE.

References

Viger, F. and Latapy, M. (2005). Fast generation of random connected graphs with prescribed degrees. *arXiv*, 0502085.

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Society.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Vogel, R., Bellet, A., Clémen, S., et al. (2021). Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 784–792. PMLR.

Voicu, I. (2018). Using first name information to improve race and ethnicity classification. *Statistics and Public Policy*, 5(1):1–13.

von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*. Springer.

von Mises, R. (1939). *Probability, statistics and truth*. Macmillan.

von Neumann, J. (1928). Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320.

Wadsworth, C., Vera, F., and Piech, C. (2018). Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv*, 1807.00199.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

References

Wasserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72.

Watson, D. S., Gultchin, L., Taly, A., and Floridi, L. (2021). Local explanations via necessity and sufficiency: Unifying theory and practice. *Uncertainty in Artificial Intelligence*, pages 1382–1392.

Wilkie, D. (1997). Mutuality and solidarity: assessing risks and sharing losses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1357):1039–1044.

Wilks, D. S. (1990). On the combination of forecast probabilities for consecutive precipitation periods. *Weather and Forecasting*, 5(4):640–650.

Willson, K. (2009). Name law and gender in iceland. *UCLA: Center for the Study of Women*.

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., et al. (2019). Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141.

Word, D. L., Coleman, C. D., Nunziata, R., and Kominski, R. (2008). Demographic aspects of surnames from census 2000. *Unpublished manuscript*.

References

Word, D. L. and Perkins, R. C. (1996). *Building a Spanish Surname List for the 1990's–: A New Approach to an Old Problem*, volume 13. Population Division, US Bureau of the Census Washington, DC.

Wortham, L. (1986). The economics of insurance classification: The sound of one invisible hand clapping. *Ohio State Law Journal*, 47:835.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20.

Wright, S. (1934). The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215.

Wu, Y., Zhang, L., Wu, X., and Tong, H. (2019). Pc-fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems*, 32.

Xu, D., Yuan, S., Zhang, L., and Wu, X. (2018). Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE.

Xu, H., Liu, X., Li, Y., Jain, A., and Tang, J. (2021). To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pages 11492–11501. PMLR.

Xue, Y., Harel, O., and Aseltine Jr, R. H. (2019). Imputing race and ethnic information in administrative health data. *Health Services Research*, 54(4):957–963.

References

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *arXiv*, 1507.05259.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Zolotarev, V. M. (1976). Metric distances in spaces of random variables and their distributions. *Mathematics of the USSR-Sbornik*, 30(3):373.