Unsupervised Statistical Tools for Anomaly Detection: The Case of Healthcare Frauds

Fabrizio Ruggeri

Istituto di Matematica Applicata e Tecnologie Informatiche

Consiglio Nazionale delle Ricerche

Via Alfonso Corti 12, I-20133, Milano, Italy, European Union

fabrizio@mi.imati.cnr.it

www.mi.imati.cnr.it/fabrizio/

OUTLINE OF THE TALK

- Motivating example: healthcare frauds
- Anomaly detection via concentration function
- Structural topic modelling
- Multivariate and time varying features
- Anomaly detection via ranks
- Bayesian co-clustering

MOTIVATING EXAMPLE: HEALTHCARE FRAUDS

- Defining fraudulent behaviour, detecting fraudulent cases and measuring fraud losses in healthcare industry are difficult tasks and expensive (audit by licensed professionals)
- Medical data classified as practitioners data, clinical instance data and medical claims data
- Interest in medical claims data, actually insurance claims
- Data containing attributes of patients, providers and claims
  - Patient: gender, age, medical history
  - Provider: type (M.D./hospital), specialty and location
  - Claim: prescription details, monetary and paid amounts
- Public data prepared by CMS (The Centers for Medicare & Medicaid Services), a U.S. federal agency
- *Provider Utilization and Payment Data Physician and Other Supplier Public Use File*

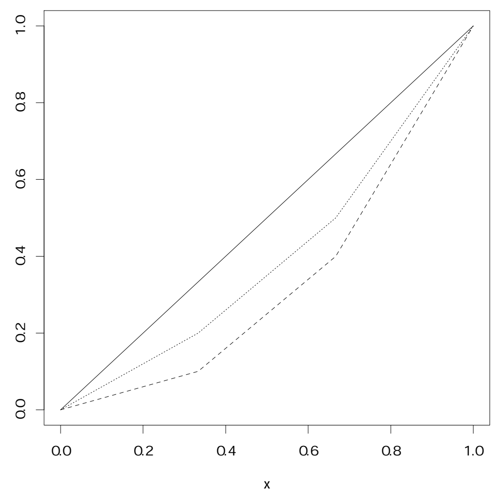
STATISTICAL TOOLS

- Statistical tools can only identify *possible frauds*, subject to further investigations
- Possible use of ”black boxes” like in most machine learning approaches
- Our research is aimed to provide tools which are statistically sound, based on easily understandable concepts, visually self-explaining
  - Supervised methods (decision trees, neural networks, Bayesian networks, logistic regression) mostly used for detecting previously known patterns of fraud
  - Unsupervised methods (Bayesian co-clustering) useful for unlabelled medical data
  - Outlier detection (concentration function, Lorenz curve, Gini and Pietra indices, structural topic modelling and ranks)

LORENZ CURVE

- n individuals with wealth $x_i, i = 1, \ldots, n \Rightarrow$ ordered $x_{(1)} \leq \ldots \leq x_{(n)}$
- $(k/n, S_k/S_n), k = 0, \ldots, n$, $S_0 = 0$ and $S_k = \sum_{i=1}^{k} x_{(i)}$ (Lorenz curve)
- Comparison of discrete p.m.'s with uniform

  Example: $(0.2, 0.3, 0.5)$ & $(0.1, 0.3, 0.6)$ vs. $(1/3, 1/3, 1/3)$



Comparison of two p.m.'s on same $(\Omega, \mathcal{F}) \Rightarrow$ concentration function (c.f.)

CONCENTRATION FUNCTION

- Probability measures $\mathbf{\Pi}$ and $\mathbf{\Pi}_0$ assigning probabilities $\underline{p} = (p_1, \ldots, p_n)$ and $\underline{q} = (q_1, \ldots, q_n)$, respectively, to the same outcomes $(x_1, \ldots, x_n)$
- C.f. constructed adding probabilities of x_i's more unlikely under $\mathbf{\Pi}$ than under $\mathbf{\Pi}_0$

  (Lorenz curve constructed adding income of individuals x_i starting from the poorest)
- For each $i, i = 1, \ldots, n$, compute the (likelihood) ratios $r_i = p_i/q_i$ and order the x_i's according to ascending values of r_i
- Order the outcomes from the ones where $\mathbf{\Pi}$ assigns much less probability than $\mathbf{\Pi}_0$ towards the ones where $\mathbf{\Pi}$ assigns much more probability than $\mathbf{\Pi}_0$
- Ordered outcomes $x_{(1)}, \ldots, x_{(n)}$, with probabilities $q_{(1)}, \ldots, q_{(n)}$ and $p_{(1)}, \ldots, p_{(n)}$
- Similar to the Lorenz curve, we plot the curve connecting the points (Q_k, P_k),

  $k = 0, \ldots, n$, where $Q_0 = P_0 = 0$, $Q_k = \sum_{i=1}^{k} q_{(i)}$ and $P_k = \sum_{i=1}^{k} p_{(i)}$
- $\Rightarrow$ Convex, increasing function: *concentration function of* $\mathbf{\Pi}$ *w.r.t.* $\mathbf{\Pi}_0$

CONCENTRATION FUNCTION

- Basic assumption (although not completely realistic): group of providers with similar characteristics (age, specialty, years in the area, etc.)
  $\Rightarrow$ similar services to patients with similar distribution of age, income, gender, etc,
- Warnings about providers with different patterns about prescriptions and charges
- Non necessarily fraud: maybe abuse or waste, or even a legitimate behaviour!
- Use of concentration function to observe anomalous behaviours:
  - Outcomes x_i's: prescriptions
  - Probabilities p_i's: percentages for each prescription by a provider
  - Probabilities q_i's: percentages for each prescription by the group of providers
- Unsupervised method which adapts to evolving (fraud) patterns (more later)

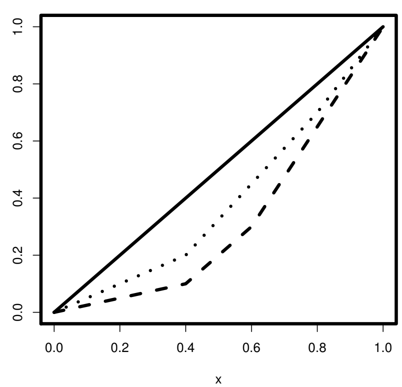
CONCENTRATION FUNCTION

- Homogeneous providers in homogeneous region prescribing only 3 tests: blood (20%), urine (40%) and ECG (40%)

- Interest in two providers: A and B (with A more anomalous than B w.r.t. group)

- Percentages for A (dashed): 20%, 70% and 10%, respectively

- Percentages for B (dotted): 30%, 50% and 20%, respectively



- Ordered l.r.'s for A: ECG ($0.25 = \frac{10}{40}$), Blood ($1 = \frac{20}{20}$), Urine ($1.75 = \frac{70}{40}$)

CONCENTRATION FUNCTION

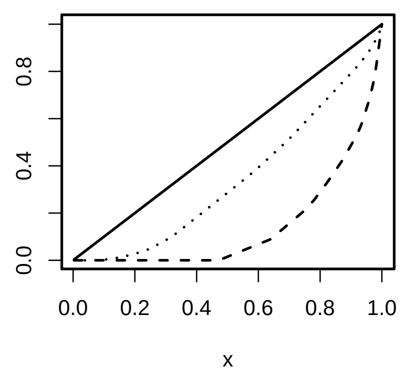
- Comparison also through summarising indices
- Gini's area of concentration (1914)
  - Twice area between Lorenz curve and straight line
  - Lorenz curve $\Rightarrow (n+1)/n - (2/n)\sum_{1\le k\le n} S_k/S_n$
  - This c.f. $\Rightarrow 1 - \sum_{1\le k\le n}(P_k + P_{k-1})(Q_k - Q_{k-1})$
- Pietra index (1915)
  - Maximum distance between Lorenz curve and straight line
  - Lorenz curve $\Rightarrow \sup_{1\le k\le n-1}(k/n - S_k/S_n)$
  - This c.f. $\Rightarrow \sup_{1\le k\le n-1}(Q_k - P_k)$
  - C.f. for two probability measures $\Rightarrow$ total variation distance
- Gini index: 0.36 for A and 0.22 for B
- Pietra index: 0.3 for A and 0.2 for B

CONCENTRATION FUNCTION

- Data: Group of 30 MDs in Diagnostic Radiology in Vermont and percentages of their billings for 61 prescribed services
- Prescribed services include X-rays, Computed Tomography, Magnetic Resonance Imaging for different parts of the human body
- Interest in two MD's: MD1 and MD2
- First warning based on high values (e.g. larger than 5) of likelihood ratios (ratio between prescriptions for a procedure by an MD and the group)
- Large ratios for MD1: Computed Tomography of the abdomen and pelvis (9.26) and X-ray exam of abdomen (9.65), accounting for 3% and 19% of his/her charges, respectively
- $\Rightarrow$ Serious warning about X-ray also because of huge number of prescriptions

CONCENTRATION FUNCTION

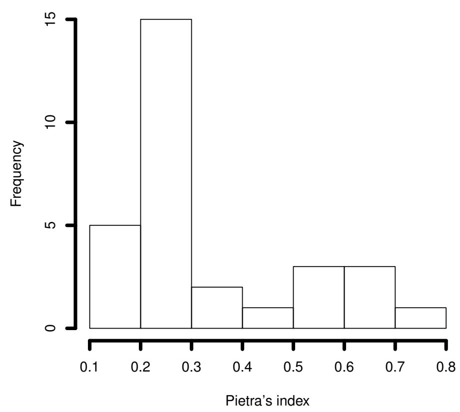
- C.f. for MD1 (dashed) and MD2 (dotted) w.r.t. population



- Anomalous behaviour (and possible fraud) of MD1 w.r.t. population:
  - Flat line from 0 to almost $0.5 \Rightarrow$ no prescriptions for procedures accounting for almost 50% of the billings by the group
  - Sharp increase around $1 \Rightarrow$ excess of charges w.r.t. the group (mostly due to X-ray exam of abdomen)
  - Gini's index: 0.7160 (0.3186 for MD2)
  - Pietra's index: 0.5504 (0.2198 for MD2)

CONCENTRATION FUNCTION

- Behaviour of all providers through summarising indices
- $\Rightarrow$ histogram of Pietra's index values for the 30 MDs in the group
- 20 MDs have their index in the first two bins $\Rightarrow$ substantial concordance among themselves
- 7 MDs with values exceeding 0.5 (possible threshold) $\Rightarrow$ possible subjects of further investigations



DYNAMIC AND MULTIVARIATE FEATURES

- So far comparisons of just one feature (percentages of billings) at a given time
- Now evolution over time accounting for changes in prescription patterns
  $\Rightarrow$ concentration functions evolving over time
- Providers' activities characterised by more features (e.g. percentage of billings and charged amount)
  - Multivariate (actually bivariate) plots of indices
  - Graphical decision frontiers accounting for multiple criteria
- Complexity in extracting data from documents

STRUCTURAL TOPIC MODELLING

- Topic Models: unsupervised hierarchical probabilistic methods to find groups within a set of documents

- Latent Dirichlet Allocation (LDA): most famous topic model in which documents are modelled as a mixture of latent topics, and the topics as a mixture over words where each word within a given document belongs to all topics with varying probabilities

- Structural Topic Model: generative statistical latent variable model allowing correlation among topics and including (unlike LDA) document-level covariates (such as author, source and date)

- Each of D documents consists of words from a vocabulary of V terms

- Predetermined number K of topics: intermediate level between words and document

- Goal: determine topic proportions for each document and frequency of words over all topics

STRUCTURAL TOPIC MODELLING

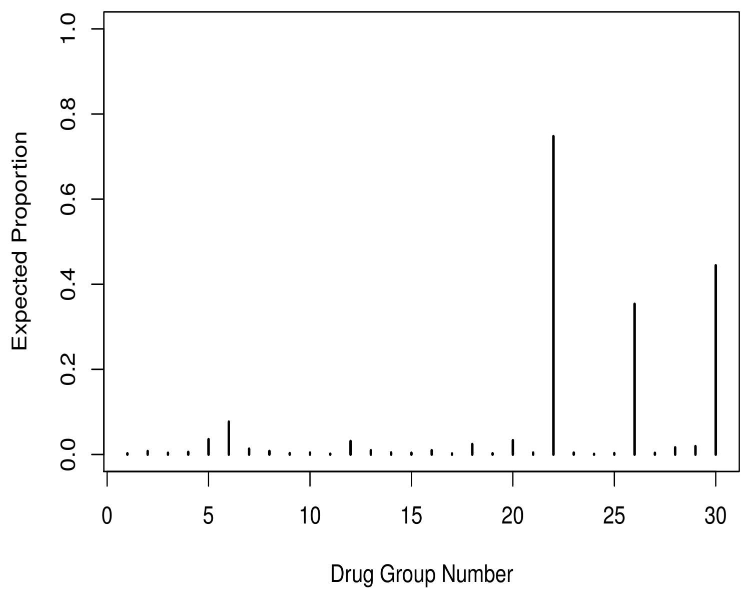
- D MDs prescribing drugs from a list of V drugs over a period T
- One document for each MD, with name of each billed drug repeated as many times as present in different claims
- Example: MD submitting 6 claims with prescriptions of two drugs (A in 4 claims, B in 2) $\Rightarrow$ document $\{A, A, A, A, B, B\}$
- Covariates: time, medical specialty, average beneficiary risk (aggregate health level of patients based on age, sex, prior medical diagnoses, and other criteria)
- Structural topic model $\Rightarrow$
  - Identification of K *large* groups of drugs (somewhat related to e.g. opioids, antibiotics, etc.) out of the many documents
  - Probabilistic membership (proportion) to any of the K groups for each MD
  - Probabilistic membership (proportion) to any of the K groups for each drug (e.g. morphine, methadone)
- Proportions: input to methods to retrieve suspicious hidden prescription patterns

STRUCTURAL TOPIC MODELLING

- Medicare Part D prescriber data in New Hampshire over 5 years, from 2013 to 2017
- New Hampshire: small state with one of the highest drug diffusion rate and opioid overdose death rate in the U.S.
- Filtered data: Top 20 specialties for opioid claims and cost per beneficiaries
- 1,617 providers submitting over 11 million claims for 981 distinct drugs
- Number of distinct prescribed drugs ranges from 1 to 243 with median of 20
- Each document contains the list of all drugs in the claims of an individual provider
- Total number of claims (i.e., document sizes) ranges from 11 to 23,270 with median of 598
- Structural topic model to extract information from documents and obtain groups of drugs and the corresponding number and charges of prescriptions for each provider

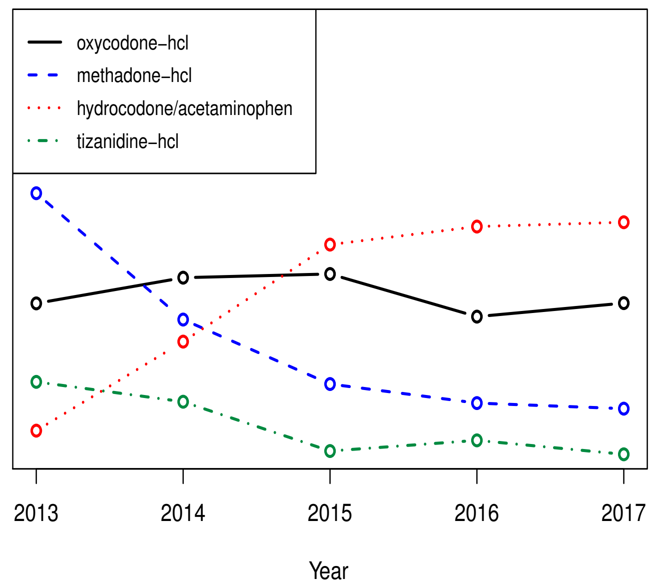
STRUCTURAL TOPIC MODELLING

- Expected opioid frequencies for each of the 30 drug groups averaged over years
- $\Rightarrow$ Group 22 mostly representing opioid drugs



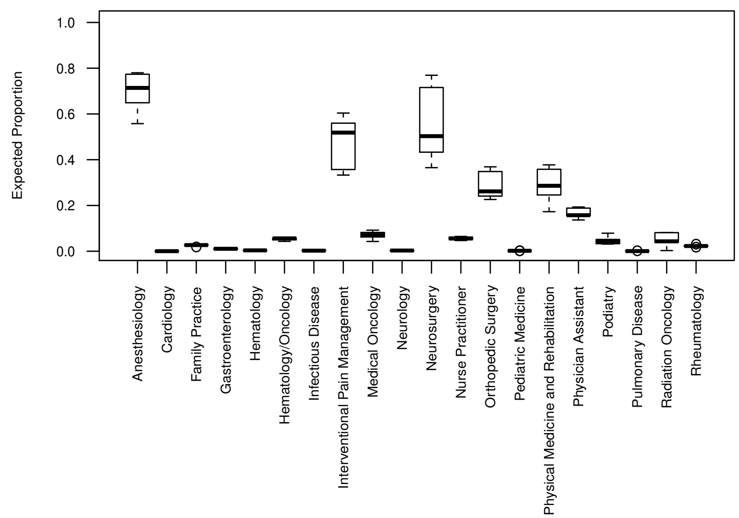
STRUCTURAL TOPIC MODELLING

- Evolution trajectory of the most frequently billed drugs within drug group 22



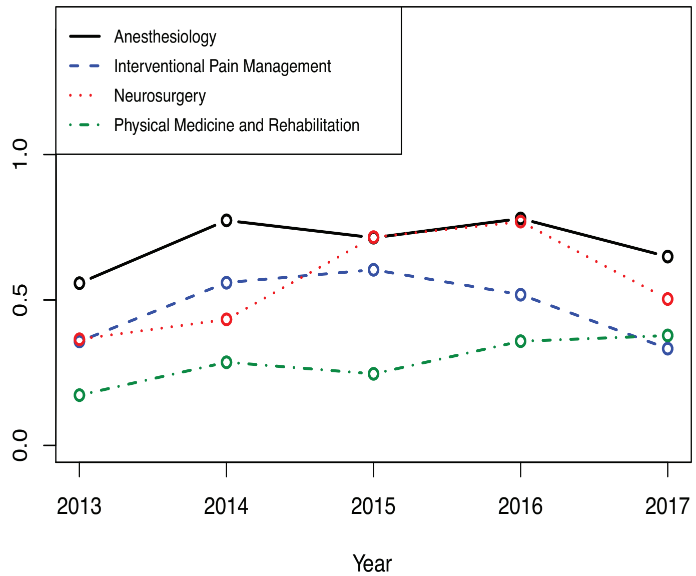
STRUCTURAL TOPIC MODELLING

- Expected billing proportion of each medical specialty from group 22 across years
- Anesthesiologists: higher billing proportions with a relatively smaller variability
- Interventional pain management doctors and neurosurgeons with higher variabilities
- Median expected billing proportion for neurosurgeons lower than mean $\Rightarrow$ some providers billing relatively high amounts compared to their peers



STRUCTURAL TOPIC MODELLING

- Billing proportion trends of several medical specialties from drug group 22
- Upward trend of expected billing by neurosurgeons from 2013 to 2016

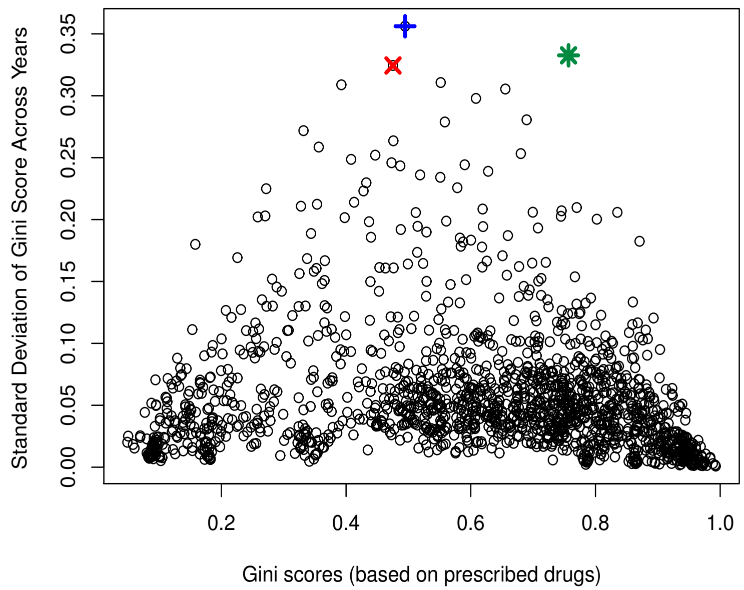


COMBINATION OF MULTIPLE CRITERIA

- N possible criteria (e.g., number of prescriptions over N years, or their number and amount of billings)
- $\Rightarrow$ C.f., Gini and Pietra indices for all of them but very impractical for large N
- Linear combination of indices $G_1, \ldots, G_N$ for each criterion (computed for a given provider w.r.t. the group)
  - $G = \sum_{i=1}^{N} \lambda_i G_i$, with $\lambda_i \geq 0, 1; i = 1, \ldots, N$ and $\sum_{i=1}^{N} \lambda_i = 1$
  - Weights λ_i denoting relative importance, assigned by auditors, of each criterion
  - Gini index: $0 \leq G_i \leq 1$ for each $i \Rightarrow 0 \leq G \leq 1$
  - Threshold $R < 1 \Rightarrow$ further investigations suggested for providers if $G \geq R$
  - Set $R_1 < \ldots < R_n$ of thresholds denoting an increasing level of risks
- In the next we consider Gini indices

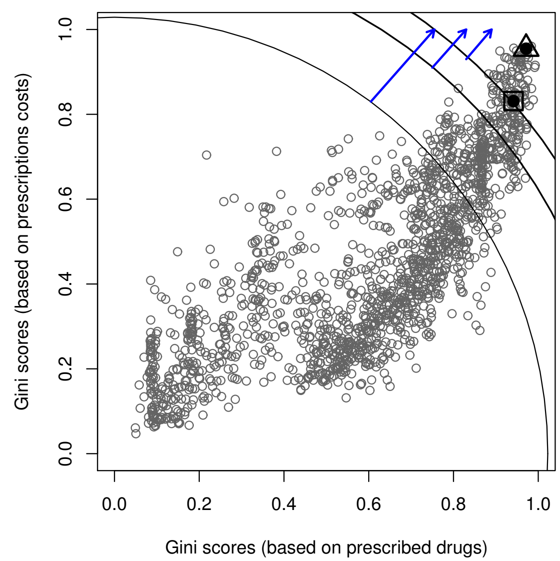
COMBINATION OF MULTIPLE CRITERIA

- Average of Gini indices over 5 years vs. their standard deviation for all MDs
- MDs in lower right corner: consistently different from the others over the years (maybe legitimate because of specialised practice)
- MDs with large standard deviations (with symbols $+$, $\times$, $\star$)
  MD with $+$: very different w.r.t. group in 2013 but very similar in 2017



COMBINATION OF MULTIPLE CRITERIA

- Plot of pairs of Gini indices (number and charges of prescriptions) in space (X, Y)
- Decision frontiers given by $x^2 + y^2 = R_i$, with R_i's thresholds leaving inside 75%, 90%, 95%, 99% of the points
- Generalised to $(1 - w)x^2 + wy^2 = R_i$ to weigh more a
- We could also consider pairs of quantiles



MULTIVARIATE FEATURES: RANKS

- Two MDs (A and B) prescribing four drugs with the following percentages:
  A: (30%, 27%, 23%, 20%) vs. B: (20%, 23%, 27%, 30%)
  $\Rightarrow$ no significant difference using previous approaches but opposite ordering!

- Two populations of MDs, A and B, sharing n groups and their m characteristics
  - n possible tests (e.g., blood, X-rays, urine, chest, etc.)
  - m characteristics (e.g., number of billings and their cost) for each test

- Compare ranking differences for m characteristics over n groups between A and B

- Data: $\left\{X_1^k, \ldots, X_m^k\right\}_{k=1}^n$ and $\left\{Z_1^k, \ldots, Z_m^k\right\}_{k=1}^n$ for A and B, respectively

- X_m^k (Z_m^k): m^{th} characteristic for group k in A (B)

- $r(Y)$: rank of r.v. Y within a population

- Sum of absolute rank differences: $SARD = \sum_{k=1}^{n} \sum_{j=1}^{m} \left| r(X_j^k) - r(Z_j^k) \right|$

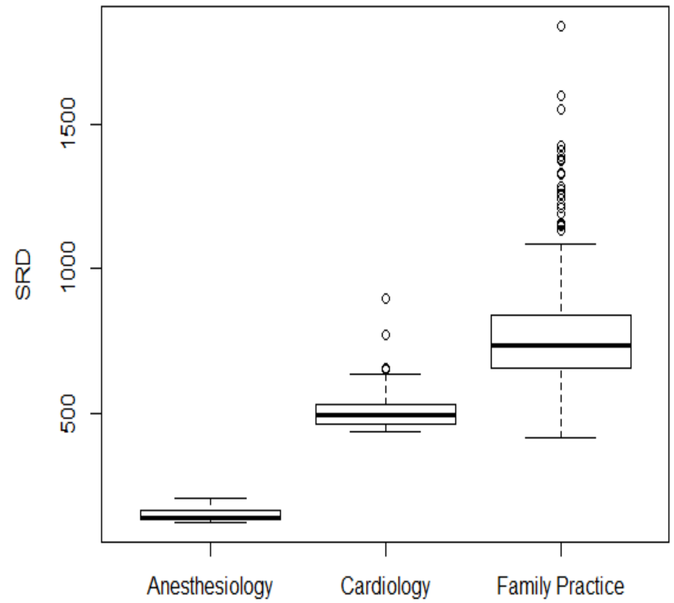
MULTIVARIATE FEATURES: RANKS

- Yearly billing ratios of 1,617 MDs across 30 drug groups from 2013 to 2017
- MDs from 20 different medical specialties
- Each MD compared to a reference population of MDs with similar medical specialty
- $m = 5$ characteristics (i.e., years) and $n = 30$ drug groups
- Billing ratios of given MD for the 30 drug groups over 5 years

ID	**Year**	**Specialty**	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**...**	**G30**
MD1	2013	Family Practice	0.1%	0.0%	3.0%	23.8%	1.1%	0.0%	...	1.4%
MD1	2014	Family Practice	1.6%	0.0%	0.5%	19.2%	0.2%	0.0%	...	6.2%
MD1	2015	Family Practice	1.2%	4.0%	3.7%	19.4%	0.1%	0.0%	...	4.7%
MD1	2016	Family Practice	2.0%	3.7%	1.8%	22.0%	0.1%	0.0%	...	2.6%
MD1	2017	Family Practice	4.2%	4.9%	2.8%	25.6%	6.9%	0.0%	...	4.7%

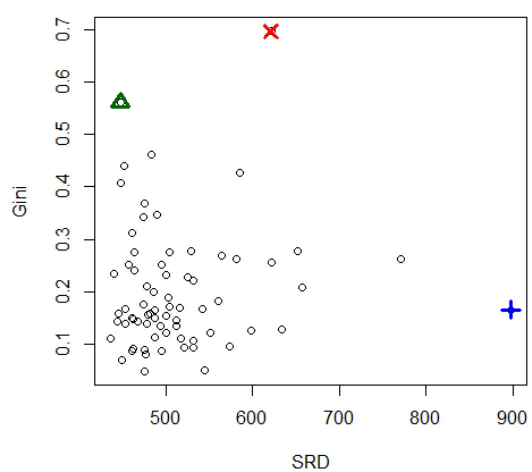
MULTIVARIATE FEATURES: RANKS

- Comparison of different medical specialties through SARD values of all their MDs to identify specialties more prone to higher differences in medical billing patterns



MULTIVARIATE FEATURES: RANKS

- Ranks provide different information w.r.t. Lorenz curve $\Rightarrow$ interest in comparing SARD and Gini indices (from earlier analyses) via scatter plot
- Cardiologist with blue plus: Highest SARD and moderate Gini
- Cardiologist with red cross: Highest Gini and quite large SARD
- Cardiologist with green triangle: Second highest Gini but very low SARD



MULTIVARIATE FEATURES: RANKS

- Findings from more thorough exploration about the three MDs marked in the plot
- Cardiologist with blue plus: Highest SARD and moderate Gini. Similar behaviour w.r.t. population for his/her top drug groups but large rank differences in less commonly billed drug groups
- Cardiologist with red cross: Highest Gini and quite large SARD. Definitely anomalous behaviour!
- Cardiologist with green triangle: Second highest Gini but very low SARD. Group 24 most prescribed by all MDs (average of 68% over 5 years) but group 17 the most prescribed for this MD in the first 4 years (84%) but only 1% the fifth year

BAYESIAN CO-CLUSTERING

- Co-clustering: data mining tool to analyse dyadic data connecting two entities
- Dyadic data: matrix with rows and columns representing each entity respectively
- Earlier works:
  - Hartigan (1972) on simultaneous clustering of rows and columns of a data matrix
  - Binder (1978) on Bayesian cluster analysis
  - Shan and Banerjee (2008) on Bayesian co-clustering in data mining and machine learning
  - Lin et al. (2008) on first application of clustering on medical data to segment practice patterns of general practitioners
- Co-clustering useful to discover the structure of data and predict missing values exploiting the relationships between two entities
- Interest in dyadic relationships among providers and procedure codes

BAYESIAN CO-CLUSTERING

- Co-clustering: fixed K clusters of providers and L clusters of procedures
- I healthcare providers billing for J unique procedures
- X_{ij}: binary value representing if provider i bills for procedure code j
- $\mathbf{X} = \{X_{ij};\ i = 1, \ldots, I,\ j = 1, \ldots, J\}$: data matrix of size $I \times J$
- Membership probabilities s.t. $\sum_{k=1}^{K} \pi_{1k} = \sum_{l=1}^{L} \pi_{2l} = 1$
  - $\pi_{1k}; k = 1, \ldots, K$ for row clusters
  - $\pi_{2l};\ l = 1, \ldots, L$ for column clusters
- Latent variables Z_{1i} and $Z_{2j},\ i = 1, \ldots, I,\ j = 1, \ldots, J$: cluster membership
  - for each provider (row): $Z_{1i} \in \{1, \ldots, K\}$
  - for each procedure (column): $Z_{2j} \in \{1, \ldots, L\}$
- Given $\boldsymbol{\pi}_1 = (\pi_{1k}; k = 1, \ldots, K)$ and $\boldsymbol{\pi}_2 = (\pi_{2l}; l = 1, \ldots, L)$
  $\Rightarrow Z_{1i}$ and Z_{2j} independent discrete random variables

BAYESIAN CO-CLUSTERING

- Stochastic model for data generation: $(X_{ij}|Z_{1i} = k, Z_{2j} = l, \theta_{kl}) \sim Ber(\theta_{kl})$
  - θ_{kl}: probability of billing of a procedure from l^{th} cluster by a provider in k^{th} cluster
- $\Rightarrow$ Assignment of each X_{ij} to a co-cluster defined by latent $(Z_{1i},\, Z_{2j})$
- Independent priors for parameters $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ and $\boldsymbol{\theta} = (\theta_{kl}; k = 1, \ldots, K,\, l = 1, \ldots, L)$
  - $\boldsymbol{\pi}_1 \sim Dir(\alpha_{1k}; k = 1, \ldots, K)$
  - $\boldsymbol{\pi}_2 \sim Dir(\alpha_{2l}; l = 1, \ldots, L)$
  - $\theta_{kl} \sim Beta(a_{kl}, b_{kl}),\ k = 1, \ldots, K,\ l = 1, \ldots, L$
- Given $\boldsymbol{X} = \{X_{ij}; i = 1, \ldots, I,\, j = 1, \ldots, J\} \Rightarrow$ posterior via Gibbs sampling

BAYESIAN CO-CLUSTERING

- Full conditionals of θ_{kl}'s, $k = 1, \ldots, K$, $l = 1, \ldots, L$: (conditionally) independent

$$\begin{aligned}\theta_{kl}|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{X} \sim Beta\Big(\quad & a_{kl} & + \sum_{i,j} X_{ij}\mathbf{I}(Z_{1i} = k, Z_{2j} = l), \\ & b_{kl} & + \sum_{i,j} (1 - X_{ij})\mathbf{I}(Z_{1i} = k, Z_{2j} = l)\Big)\end{aligned}$$

with $\boldsymbol{Z}_1 = \{Z_{1i}; i = 1, \ldots, I\}$, $\boldsymbol{Z}_2 = \{Z_{2j}; j = 1, \ldots, J\}$, $\mathbf{I}(\bullet)$ indicator function

- Full conditionals of $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$: (conditionally) independent

$$\begin{aligned}\boldsymbol{\pi}_1|\boldsymbol{Z}_1 &\sim Dir\Big(\alpha_{1k} + \sum_{i,j} \boldsymbol{I}(Z_{1i} = k); k = 1, \ldots, K\Big), \\ \boldsymbol{\pi}_2|\boldsymbol{Z}_2 &\sim Dir\Big(\alpha_{2l} + \sum_{i,j} \boldsymbol{I}(Z_{2j} = l); l = 1, \ldots, L\Big)\end{aligned}$$

- Full conditionals of (Z_{1i}, Z_{2j}):

$$p(Z_{1i} = k, Z_{2j} = l|\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\theta}, X_{ij}) = \frac{\theta_{kl}^{X_{ij}}\,(1 - \theta_{kl})^{1-X_{ij}}\pi_{1k}\,\pi_{2l}}{\sum_{r=1}^{K}\sum_{c=1}^{L}\theta_{rc}^{X_{ij}}\,(1 - \theta_{rc})^{1-X_{ij}}\pi_{1r}\,\pi_{2c}}$$

BAYESIAN CO-CLUSTERING

- CMS (Centers for Medicare & Medicaid Services, a US federal agency) data on billings by anesthesiologists in Texas

- Consider only providers who billed at least 10 procedures and only procedures billed by at least 20 providers $\Rightarrow$ 94 procedures billed by 376 providers

- Number of clusters set as $K = 3$ and $L = 2$

- MCMC with uniform priors: $2,000$ samples after burn-in of $18,000$ iterations

- Most frequent occurrences for provider-procedure pair found in co-cluster $(3, 1)$

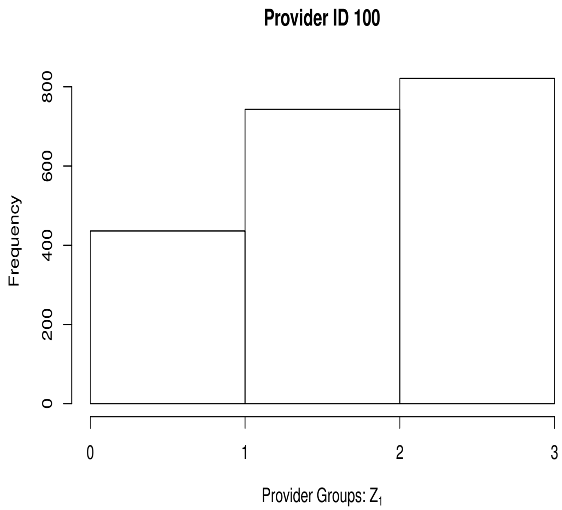
- $\Rightarrow$ Largest cluster with providers behaving similarly in terms of procedures they bill

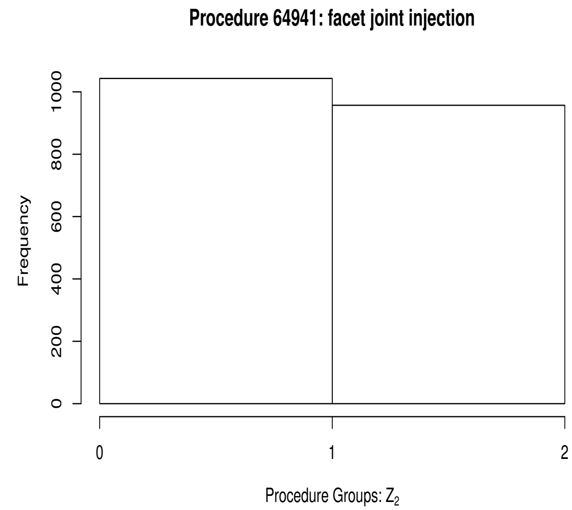
- Under the assumption that the majority of providers behave correctly, such cluster might (but not 100% sure!) correspond to legitimate billings

- Other clusters correspond to less likely procedures by less providers and might lead to potential investigation of the providers there

BAYESIAN CO-CLUSTERING

- Procedure 64941 (facet joint injection) by Provider with ID 100





- Posterior modes: $Z_{1,100} = 3$ and $Z_{2,64941} = 1$

- Co-cluster with highest association $\Rightarrow$ probably a legitimate billing

- More suspicious if posterior mode $Z_{1,100} = 2$

REFERENCES

- Ekin, T., Ieva, F., Ruggeri, F. and Soyer, R. (2018), Statistical Medical Fraud Assessment: Exposition to an Emerging Field. *International Statistical Review*, 86, 379-402.

- Ekin, T., Ieva, F., Ruggeri, F. and Soyer, R. (2017), On the Use of the Concentration Function in Medical Fraud Assessment. *The American Statistician*, 71, 236-241.

- Zafari, B. and Ekin, T. (2019), Topic Modelling for Medical Prescription Fraud and Abuse Detection. *Journal of the Royal Statistical Society, Series C*, 68, 751-769.

- Zafari, B., Ekin, T. and Ruggeri, F. (2022), Multicriteria Decision Frontiers for Prescription Anomaly Detection Over Time. *Journal of Applied Statistics*, 49, 3638–3658.

- Zafari, B., Ekin, T. and Ruggeri, F. (2022), Unsupervised Rank-based Outlier Detection with Feature Importance. *Submitted*.

- Ekin, T., Ieva, F., Ruggeri, F. and Soyer, R. (2013), Applications of Bayesian Methods in Detection of Healthcare Frauds. *Chemical Engineering Transactions*, 33, 151-156.

- Ekin, T. (2019). *Statistics and Health Care Fraud*. Chapman and Hall/CRC, New York, NY, USA.