Exceptional Model Mining

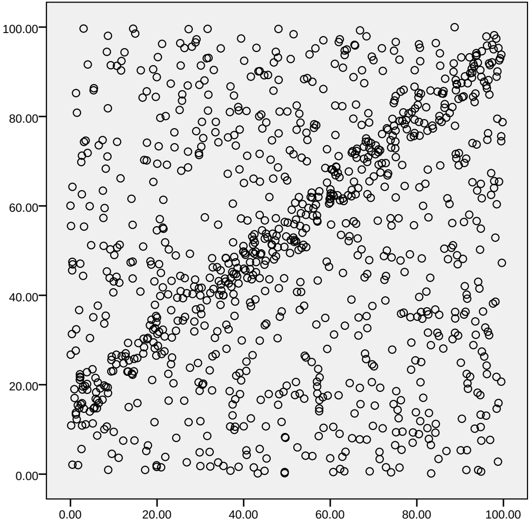
Identifying Deviations in Data

Wouter Duivesteijn

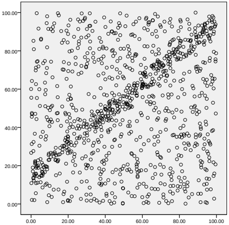
March 14, 2013

Universiteit Leiden

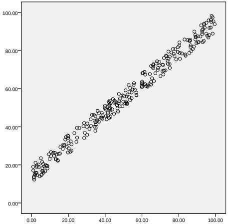
Mixture of distributions (1/2)



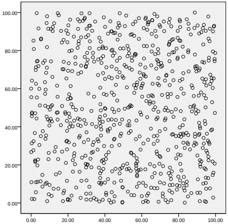
Mixture of distributions (2/2)



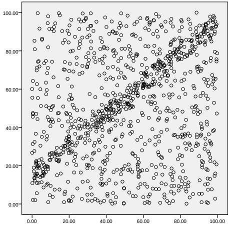
$\Rightarrow$



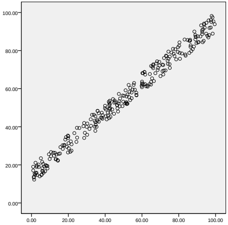
$\cup$



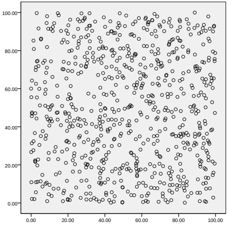
Mixture of distributions (2/2)



$\Rightarrow$



$\cup$



- For each datapoint it is unclear whether it belongs to G or $\overline{G}$
- Description of exceptional subgroup G?
- Model class unknown
- Model parameters unknown

Solution: extend Subgroup Discovery

- Use other information than X and Y: object descriptions D
- Use Subgroup Discovery to scan subsets of the data in terms of D

Subgroup Discovery: find subgroups of the database where the target attribute shows an unusual distribution.

Solution: extend Subgroup Discovery

- Use other information than X and Y: object descriptions D
- Use Subgroup Discovery to scan subsets of the data in terms of D
- Model over subgroup becomes target of SD

Subgroup Discovery: find subgroups of the database where the target attribute shows an unusual distribution.

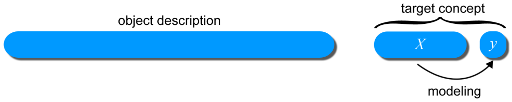
Exceptional Model Mining: find subgroups of the database where the target attributes show an unusual distribution, by means of modeling over the target attributes.

Table of Contents

Exceptional Model Mining

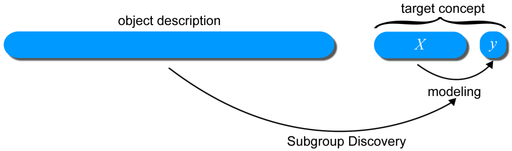
- Define a target concept (X and y)

Exceptional Model Mining



- Define a target concept (X and y)
- Choose a model class C
- Define a quality measure φ over C

Exceptional Model Mining



- Define a target concept (X and y)
- Choose a model class C
- Define a quality measure φ over C
- Use Subgroup Discovery to find exceptional subgroups G and associated models M

Managing the candidate space in SD and EMM

SD and EMM are exploratory techniques. Find subsets of the dataset $\Rightarrow |\text{candidates}| = 2^N$.

Managing the candidate space in SD and EMM

SD and EMM are exploratory techniques. Find subsets of the dataset $\Rightarrow$ $|\text{candidates}| = 2^N$.

In SD, with only nominal attributes: exhaustive algorithm (anti-montonicity).

In EMM, with numeric attributes and general quality measure: no such property. Instead: beam search.

Build up candidate subgroups level-wise, imposing one constraint on one attribute at a time.

Model classes

Investigated model classes:

- correlation coefficient between two numeric variables x_1 and x_2
- simple linear regression model $y_i = a + bx_i + e_i$
- classification model with discrete y and arbitrary $x_1, \ldots, x_k$
- Bayesian networks on discrete variables $x_1, \ldots, x_k$
- general linear regression model $Y = X\beta + \varepsilon$
- ...

1963, Makah Bay, Washington state, USA

Robert T. Paine investigates marine ecosystem with 15 species.

1963, Makah Bay, Washington state, USA

Robert T. Paine investigates marine ecosystem with 15 species.

Pisaster is the top carnivore in the system.

Pisaster ochraceus

1963, Makah Bay, Washington state, USA

Robert T. Paine investigates marine ecosystem with 15 species.

Pisaster is the top carnivore in the system.

With *Pisaster* removed, 8 species remained. A foodchain between a sponge and a nudibranch was displaced, and the anemone population reduced in density. *Pisaster* does not eat any of these species.

Pisaster ochraceus

2013, Dortmund, Germany

Suppose we want to find meteorological conditions coinciding with large ecosystem displacements.

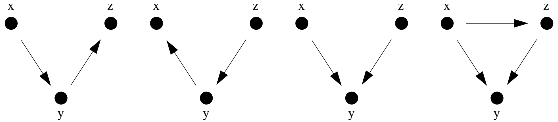
Detect indirect influences such as the dependence between the sponge and the nudibranch that is conditional on *Pisaster*.

Solution: an Exceptional Model Mining instance designed to find exceptional (conditional) dependence relations between attributes.

Pisaster ochraceus

Bayesian networks

- Capture interdependecies between discrete variables $x_1, \ldots, x_k$
- Model conditional dependency relations between these target variables: Bayesian network
- Fit network BN_Ω w.r.t. whole dataset
- For each subgroup G: fit network BN_G w.r.t. the records in G. Then determine difference in structure between BN_G and BN_Ω



(a) (b) (c) (d)

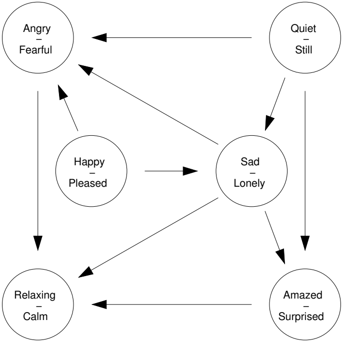
Edit distance for Bayesian networks

Verma & Pearl (1990): two Bayesian networks are equivalent $\Leftrightarrow$ they have the same skeleton and v-structures

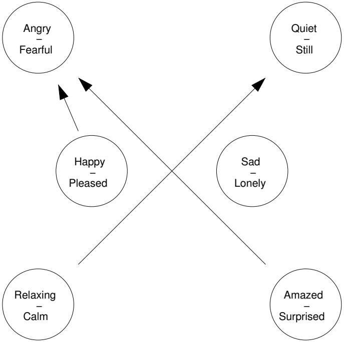
- Let BN_1 and BN_2 be Bayesian networks, S_1 and S_2 the edge sets of their skeletons, and M_1 and M_2 the edge sets of their moralized graphs
- Compute $\ell = \Big|[S_1 \oplus S_2] \ \cup \ [M_1 \oplus M_2]\Big|$
- We let $\mathrm{d}(BN_1, BN_2) = \frac{2\ell}{k(k-1)}$

Subgroup quality: $\varphi_{\text{ed}}(G) = \mathrm{d}(BN_\Omega, BN_G)$

Edit distance between BNs fitted on the *emotions* dataset

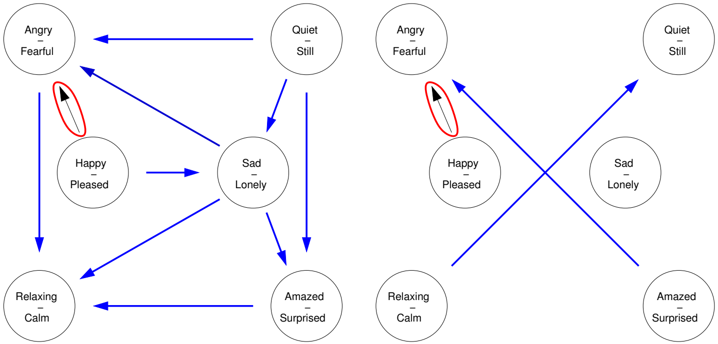


Whole dataset



$\mathit{STD\ MFCC\ 7} \leq 0.203 \wedge$
$\mathit{Mean\ Centroid} \geq 0.066$

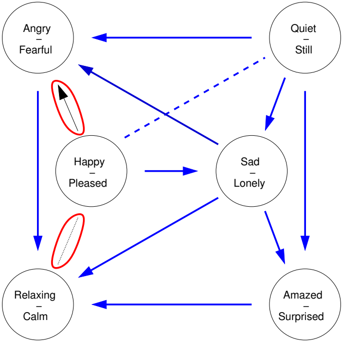
Edit distance between BNs fitted on the *emotions* dataset



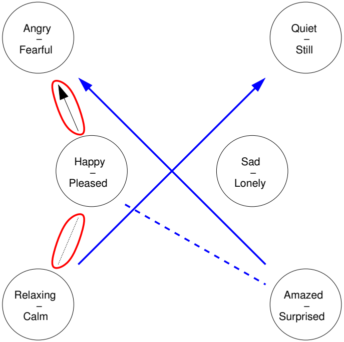
Whole dataset

STD MFCC 7 $\leq$ 0.203 $\wedge$
Mean Centroid $\geq$ 0.066

Edit distance between BNs fitted on the *emotions* dataset

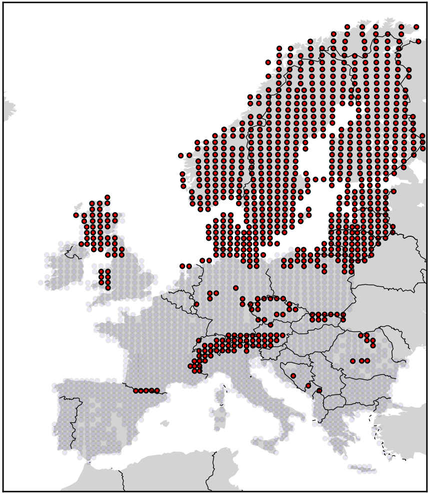


Whole dataset



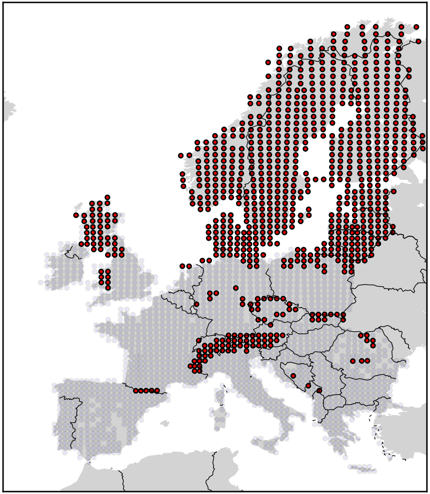
STD MFCC 7 $\leq 0.203 \wedge$
Mean Centroid ≥ 0.066

Subgroup found on *Mammals* dataset



$\mathit{max\ temp\ mar} \leq 7.97 \quad \wedge \quad \mathit{max\ temp\ sep} \leq 17.65$

Subgroup found on *Mammals* dataset



max temp mar $\leq 7.97 \quad \wedge \quad$ *max temp sep* ≤ 17.65

1895, Scotland

1895, Scotland

R. Giffen

1895, Scotland

"[. . .] as Sir R. Giffen has pointed out, a rise in the price of bread makes so large a drain on the resources of the poorer labouring families and raises so much the marginal utility of money to them, that they are forced to curtail their consumption of meat and the more expensive farinaceous foods: and, bread being still the cheapest food which they can get and will take, they consume more, and not less of it."

Alfred Marshall, *Principles of Economics*

R. Giffen

2008, Hunan, China

"This paper provides the first real-world evidence of Giffen behavior, i.e., upward sloping demand. Subsidizing the prices of dietary staples for extremely poor households in two provinces of China, we find strong evidence of Giffen behavior for rice in Hunan, and weaker evidence for wheat in Gansu."

Robert Jensen and Nolan Miller, *American Economic Review*

EMM meets linear regression

Given N records r^i of the form $\left\{a_1^i, \ldots, a_k^i, x_1^i, \ldots, x_{p-1}^i, y^i\right\}$

Use $a_1 \ldots, a_k$ for describing subgroups.

Fit linear regression model $Y = X\beta + \varepsilon$, where

$$X = \begin{pmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_{p-1}^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_{p-1}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^N & x_2^N & \cdots & x_{p-1}^N \end{pmatrix} \qquad Y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{pmatrix}$$

Ordinary least squares $\Rightarrow$ estimate $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$

How to mine for subgroups G with deviating $\hat{\beta}_G$?

Universiteit Leiden

Cook's distance

Cook [1977]: according to normal theory, the $(1-\alpha) \times 100\%$ confidence ellipsoid for β is given by all β^* satisfying:

$$\frac{\left(\beta^* - \hat{\beta}\right)^\top X^\top X \left(\beta^* - \hat{\beta}\right)}{ps^2} \leq F\left(p, N-p, 1-\alpha\right)$$

Cook's distance

Cook [1977]: according to normal theory, the $(1-\alpha)\times 100\%$ confidence ellipsoid for β is given by all β^* satisfying:

$$\frac{\left(\beta^*-\hat{\beta}\right)^\top X^\top X\left(\beta^*-\hat{\beta}\right)}{ps^2}\leq F\left(p,N-p,1-\alpha\right)$$

- compute OLS-estimate $\hat{\beta}$ on whole dataset;
- compute OLS-estimate $\hat{\beta}_G$ on data covered by subgroup G;
- use Cook's distance D_G as quality of subgroup:

$$D_G=\frac{\left(\hat{\beta}_G-\hat{\beta}\right)^\top X^\top X\left(\hat{\beta}_G-\hat{\beta}\right)}{ps^2}$$

Giffen behavior data – global model

Global model:

$$\%\Delta\mathrm{staple}_{i,t} = \alpha + \beta\%\Delta p_{i,t} + \sum \gamma\%\Delta Z_{i,t} + \sum \delta\mathrm{C}\times\mathrm{T}_{i,t} + \Delta\varepsilon_{i,t}$$

$\%\Delta\mathrm{staple}_{i,t} =$ change in household i's consumption of rice;

$\%\Delta p_{i,t} =$ change in rice price due to subsidy;

$\%\Delta Z_{i,t} =$ changes in income, household size, . . . ;

$\mathrm{C}\times\mathrm{T}_{i,t} =$ dummy variables: county factors changing over time.

$\beta > 0 \Rightarrow$ Giffen behavior.

Giffen behavior data – Jensen and Miller's field experiment

Extremely poor $\Rightarrow$ no Giffen behavior; they consume rice almost exclusively anyway.

Measured by Initial Staple Calorie Share (ISCS).

Jensen and Miller manually selected ISCS thresholds:

Group	$\hat{\beta}_G$	Giffen behavior
ISCS > 0.8	-0.585	No
ISCS ≤ 0.8	0.466	Yes

Giffen behavior data – found subgroups

Ran EMM on dataset, with ISCS as descriptive attribute.

On complete ($N = 1254$) dataset: $\hat{\beta} = 0.22$.

Giffen behavior data – found subgroups

Ran EMM on dataset, with ISCS as descriptive attribute.

On complete ($N = 1254$) dataset: $\hat{\beta} = 0.22$.

Best subgroup found ($n = 106$):

$$\text{ISCS} \geq 0.87 \quad \left(\text{with} \;\; \hat{\beta}_G = -0.96\right)$$

Giffen behavior data – found subgroups

Ran EMM on dataset, with ISCS as descriptive attribute.

On complete ($N = 1254$) dataset: $\hat{\beta} = 0.22$.

Best subgroup found ($n = 106$):

$$\text{ISCS} \geq 0.87 \quad \left(\text{with} \quad \hat{\beta}_G = -0.96\right)$$

Other subgroups:

Group	$\hat{\beta}_G$	Giffen behavior
Income per capita ≤ 64.67	-0.41	No
Income per capita ≥ 803.75	0.79	Yes (strong!)

EAEF data – global model and found subgroup

National Longitudinal Survey of Youth 1979.

Model fitted on complete dataset ($N = 2714$):

$$\text{Earnings} = -29.15 + 2.78 \times \text{YrsOfSchool} + 0.63 \times \text{YrsWorkExp}$$

EAEF data – global model and found subgroup

National Longitudinal Survey of Youth 1979.

Model fitted on complete dataset ($N = 2714$):

$$\text{Earnings} = -29.15 + 2.78 \times \text{YrsOfSchool} + 0.63 \times \text{YrsWorkExp}$$

4th-ranked subgroup: COLLBARG $= 1$. Model fitted on subgroup ($n = 533$):

$$\text{Earnings} = -8.93 + 1.57 \times \text{YrsOfSchool} + 0.43 \times \text{YrsWorkExp}$$

EAEF data – subgroup deviation rationale

Extra dollars earned per	YrsOfSchool	YrsWorkExp
Complete dataset	\$2.78	\$0.63
COLLBARG = 1	\$1.57	\$0.43

Consistent with finding that unions tend to equalize income distribution, particularly between skilled and unskilled workers.

See also T. Aidt and Z. Tzannatos, Unions and Collective Bargaining, The World Bank, 2002.

Managing the candidate space in regression-EMM

SD and EMM are exploratory techniques. Find subsets of the dataset $\Rightarrow$ $|\text{candidates}| = 2^N$.

Managing the candidate space in regression-EMM

SD and EMM are exploratory techniques. Find subsets of the dataset $\Rightarrow$ $|\text{candidates}| = 2^N$.

Additionally, for each candidate we must compute OLS-estimate for $\hat{\beta}_G$, which is rather expensive!

Upper bounds on Cook's distance (1/2)

$$D_G = \frac{\left(\hat{\beta}_G - \hat{\beta}\right)^\top X^\top X \left(\hat{\beta}_G - \hat{\beta}\right)}{ps^2}$$

- rewrite in terms of error vector e_G and hat matrix V_G;
- use spectral decomposition of hat matrix (rewriting in terms of eigenvalues $\lambda_1 \leq \ldots \leq \lambda_m$ and eigenvectors):

$$D_G \leq \frac{\lambda_m}{(1-\lambda_m)^2} \cdot \frac{\sum_{i \in G} e_i^2}{ps^2}$$

- prevent eigenvalue computation by approximation:

$$D_G \leq \frac{\operatorname{tr}(V_G)}{(1-\operatorname{tr}(V_G))^2} \cdot \frac{\sum_{i \in G} e_i^2}{ps^2} \qquad (6)$$

Upper bounds on Cook's distance (2/2)

Bound (6): potentially different for each G. Varying G over subgroups of fixed size m, we can compute:

$$R^2 = \max_G \left(\sum_{i \in G} e_i^2 \right) \qquad T = \max_G \left(\sum_{i \in G} v_{ii} \right)$$

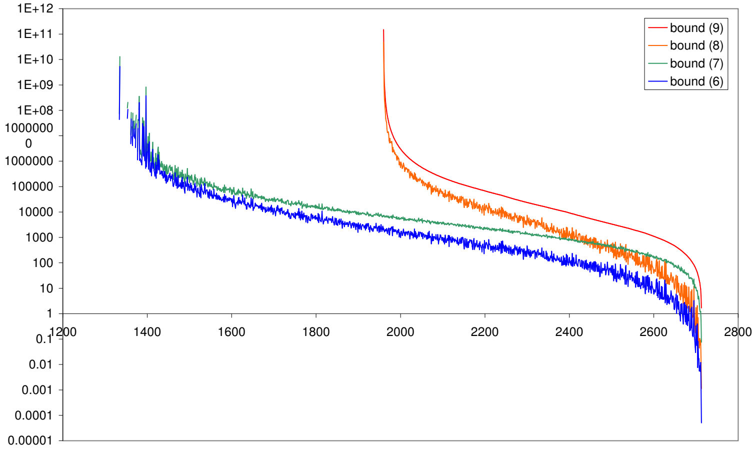
in order to obtain:

$$D_G \leq \frac{\operatorname{tr}(V_G)}{(1 - \operatorname{tr}(V_G))^2} \cdot \frac{R^2}{ps^2} \tag{7}$$

$$D_G \leq \frac{T}{(1 - T)^2} \cdot \frac{\sum_{i \in G} e_i^2}{ps^2} \tag{8}$$

$$D_G \leq \frac{T}{(1 - T)^2} \cdot \frac{R^2}{ps^2} \tag{9}$$

Bound behavior for varying subgroup size – EAEF data



Pruning the candidate space with the bounds

Per beam search level:

- determine number S of subgroups we want to retain;
- enumerate candidates in decreasing order of a bound – consider subgroups in this order
- for each subgroup G:
  - compute bounds (9), (8), (7), and (6);
  - check whether any bound has lower value than D_{G_S} of S^{th} best evaluated subgroup so far;
  - if yes: discard subgroup; if no: compute D_G.

Pruning results

Dataset	N	k	$\lvert\mathcal{C}\rvert$	p	$\frac{\lvert\text{bounded } \mathcal{C}\rvert}{\lvert\mathcal{C}\rvert}$	$\frac{\lvert\text{pruned } \mathcal{C}\rvert}{\lvert\mathcal{C}\rvert}$
Ames Housing	2930	77	980	3	0.419	0.393
Auction	1225	3	40	7	0.350	0.225
EAEF	2714	32	204	3	0.407	0.176
Giffen Behavior	1254	6	100	16	0.010	0.010
PC486	6259	3	6	7	0.333	0.167
Wine	5000	6	56	4	0.464	0.304

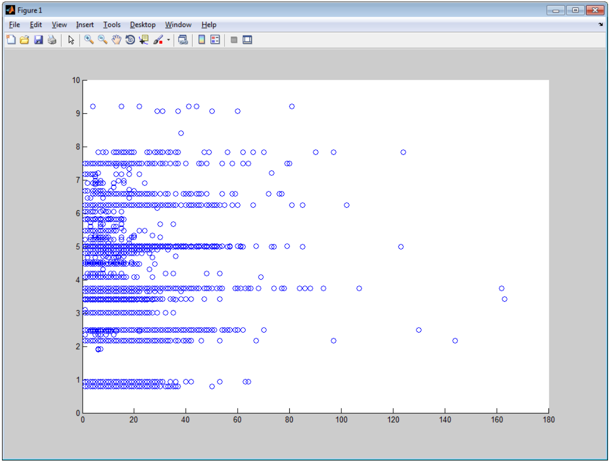
All datasets are publicly available (three of which from Journal of Applied Econometrics).

Nice technique. Why use it?

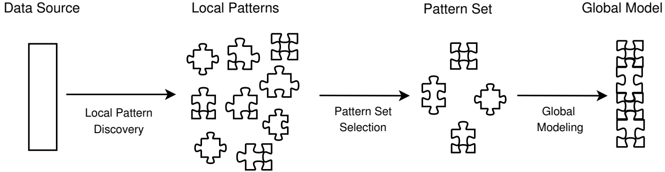
We have explored the “how” of EMM, but not the “why”. Three answers:

1. we learn things about our data;
2. useful for metalearning;
3. improve global modeling.

EMM for metalearning



LeGo: connect with multi-label classification



Joint work with TU Darmstadt: using the found subgroups as constructed binary features to enhance multi-label classifiers.

Tested on three datasets with SVM classifiers: Friedman test with post-hoc Nemenyi test indicate significant better rank when adding top 100 subgroups to dataset.

Does not work well with decision trees.

Ongoing research: improve regression goodness-of-fit

Incorporating regression-EMM subgroups as dummy variables in regression model might improve goodness-of-fit.

Given a binary subgroup indicator variable $D(i)$, instead of fitting

$$y^i = \beta_0 + \beta_2 \cdot x^i + \varepsilon^i$$

we can fit

$$y^i = \beta_0 + \beta_1 \cdot D(i) + \beta_2 \cdot x^i + \beta_3 \cdot (D(i) \cdot x^i) + \varepsilon^i$$

Hence

$$\begin{array}{lcccccl} y^i = & \beta_0 & + & \beta_2 & \cdot x^i + \varepsilon^i & \text{if} & D(i) = 0 \\ y^i = & (\beta_0 + \beta_1) & + & (\beta_2 + \beta_3) & \cdot x^i + \varepsilon^i & \text{if} & D(i) = 1 \end{array}$$

TODO: see if adjusted R-squared increases.

Universiteit Leiden

Statistical trouble in EMM

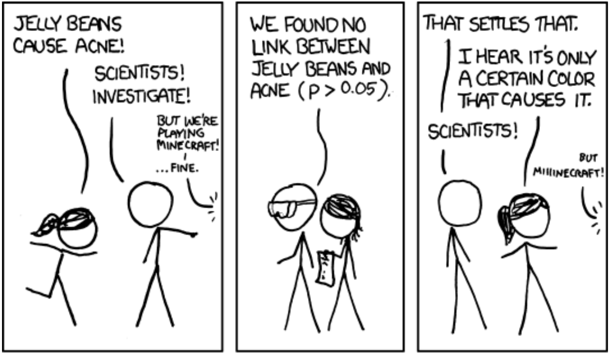
EMM results in list of subgroups ranked by quality. However, are these true discoveries?

Nontrivial to determine exactly how large the search space is:

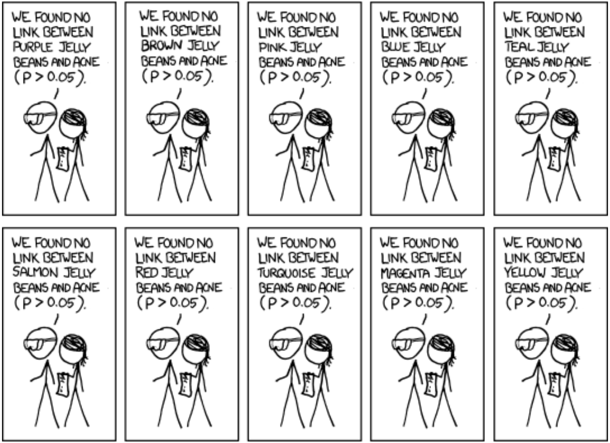
- depends on selected search strategy;
- depends on user-induced constraints.

However, even in quite shallow searches: $\#\{\text{candidates}\} > 10,000$.

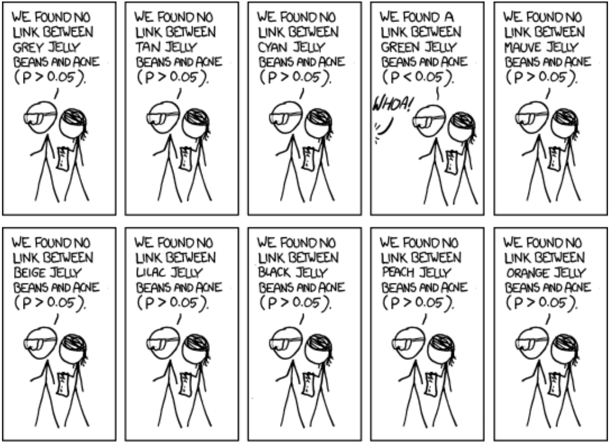
Single jelly bean hypothesis



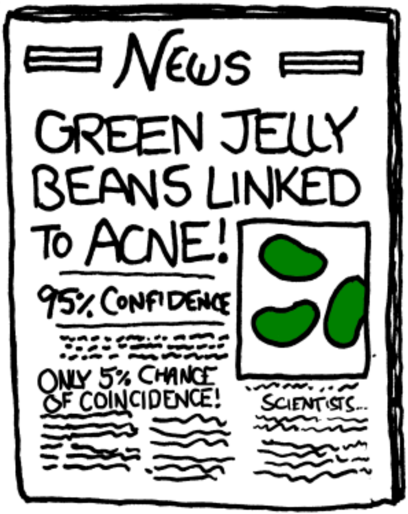
Multiple jelly bean hypotheses (1/2)



Multiple jelly bean hypotheses (2/2)



Breaking news!!!



© Randall Munroe

xkcd.com

The Multiple Comparisons Problem (MCP)

This statistical problem is called:

The Multiple Comparisons Problem
Hypotheses
Testing

We choose to work with the first name.

EMM-specific MCP approach

Suppose a dataset Ω, and a set of subgroups $\mathcal{S}$ found through EMM using some quality measure (QM) φ.

New EMM-specific MCP approach:

1. generate artificial false discoveries;
2. build a statistical model;
3. validate found subgroups by refuting that they stem from the FD model.

Generating false discoveries

We generate n copies $D_1, \ldots, D_n$ of Ω. In each copy, we *swap randomize* the target attributes.

Generating false discoveries

We generate n copies $D_1, \ldots, D_n$ of Ω. In each copy, we *swap randomize* the target attributes.

We run EMM on each new dataset, using same parameters and constraints as when discovering $\mathcal{S}$. Result: sets of false discoveries $\mathcal{R}_1, \ldots, \mathcal{R}_n$.

Generating false discoveries

We generate n copies $D_1, \ldots, D_n$ of Ω. In each copy, we *swap randomize* the target attributes.

We run EMM on each new dataset, using same parameters and constraints as when discovering $\mathcal{S}$. Result: sets of false discoveries $\mathcal{R}_1, \ldots, \mathcal{R}_n$.

From each false discovery set, we select the pattern with the highest quality. Result: *independent* false discoveries $R_1, \ldots, R_n$.

Distribution of False Discoveries

Assuming n is sufficiently large, we can invoke the central limit theorem:

Since $\varphi(R_1), \ldots, \varphi(R_n)$ are i.i.d. random variables, their mean follows a normal distribution.

Let μ and σ denote the sample mean and standard deviation. Then $\mathcal{N}(\mu, \sigma^2)$ is the *Distribution of False Discoveries* (DFD).

Distribution of False Discoveries

Assuming n is sufficiently large, we can invoke the central limit theorem:

Since $\varphi(R_1), \ldots, \varphi(R_n)$ are i.i.d. random variables, their mean follows a normal distribution.

Let μ and σ denote the sample mean and standard deviation. Then $\mathcal{N}(\mu, \sigma^2)$ is the *Distribution of False Discoveries* (DFD).

To validate subgroups $S \in \mathcal{S}$, compute a p-value testing whether $\varphi(S)$ is generated by the DFD.

Experiments on validating subgroups: results

Dataset	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 1\%$	Dataset	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 1\%$
Adult	1.000	1.000	1.000	Ionosphere	1.000	1.000	1.000
Balance-scale	0.561	0.554	0.548	Iris	0.902	0.879	0.834
Car	0.650	0.591	0.518	Labor	0.628	0.567	0.401
CMC	0.506	0.484	0.445	Mushroom	0.967	0.966	0.964
Contact-lenses	0.069	0.069	0.052	Pima-indians	1.000	1.000	1.000
Credit-a	1.000	1.000	1.000	Soybean	0.724	0.713	0.689
Dermatology	0.838	0.808	0.761	Tic-tac-toe	0.493	0.446	0.311
Glass	0.738	0.675	0.562	Wisconsin	1.000	1.000	1.000
Haberman	0.427	0.392	0.327	Yeast	0.687	0.673	0.647
Hayes-roth	0.388	0.313	0.210	Zoo	0.600	0.582	0.524

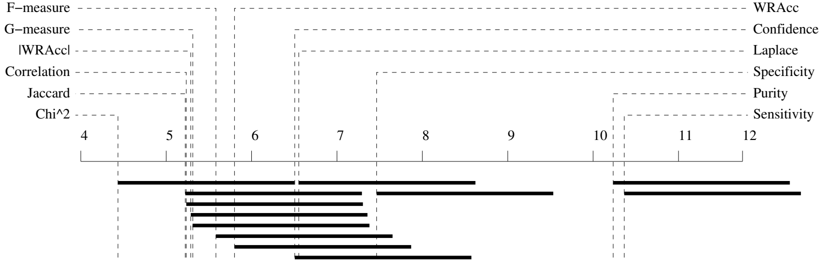
For some datasets no subgroups can be refuted. Why?

Metalearning

Dataset	N	# attributes disc	num	$\lvert\ell\rvert$	$\alpha = 1\%$	Dataset	N	# attributes disc	num	$\lvert\ell\rvert$	$\alpha = 1\%$
Adult	48842	8	6	2	1.000	Ionosphere	351	0	34	2	1.000
Balance-scale	625	0	4	3	0.548	Iris	150	0	4	3	0.834
Car	1728	6	0	4	0.518	Labor	57	8	8	2	0.401
CMC	1473	7	2	3	0.445	Mushroom	8124	22	0	2	0.964
Contact-lenses	24	4	0	3	0.052	Pima-indians	768	0	8	2	1.000
Credit-a	690	9	6	2	1.000	Soybean	683	35	0	19	0.689
Dermatology	366	33	1	6	0.761	Tic-tac-toe	958	9	0	2	0.311
Glass	214	0	9	6	0.562	Wisconsin	699	0	9	2	1.000
Haberman	306	1	2	2	0.327	Yeast	1484	1	7	10	0.647
Hayes-roth	132	0	4	3	0.210	Zoo	101	16	1	7	0.524

These datasets all have more than five numeric attributes.

Comparing 12 QMs for $k = 1$



Conclusions

- We identify all kinds of exceptional models in data:
  - exceptional correlation;
  - exceptional classification;
  - exceptional conditional dependencies;
  - exceptional regression slope;
  - ...
  - enter your personal favorite exceptionality here.
- useful for metalearning;
- fruitful for enhancing global modeling;
- also introduced method to weed out the false discoveries.

Conclusions