Statistical techniques for fraud detection, prevention, and evaluation

David J. Hand
Imperial College London
Research group:

Niall Adams, Adam Brentnall, Martin Crowder, Nick Heard, Dave Weston, Chris Whitrow, Piotr Juszczak, Kiriaki Platanioti, Dimitris Tasoulis, Nicos Pavlidis, Matt Turnbull, James Bentham, Iding Wu, Fanyin Zhou, Christofohos Anagnostopoulos, Daniel Balabanoff, Ed Tricker, Gordon Blunt, Marc Henrion, Gordon Ross, Asif Johar, ...
I: Background
II: How big is fraud?
III: Fraud in banking
IV: Fraud in science
V: Conclusions
Context

*By statistics I mean ‘greater statistics’, in the sense of John Chambers: ‘everything related to learning from data’.*
I: Background

What is fraud?

*Criminal deception; the use of false representations to gain an unjust advantage*

Concise Oxford Dictionary

Older than humanity itself.
- even animals are known to try to deceive others
- camouflage
Fraud occurs in all areas of human endeavour

But the motivation is not always the same

**Motivation 1: money**
- Banking fraud
- Telecommunications fraud
- Insurance fraud
- Health care fraud
- Click fraud

**Motivation 2: power, peer regard, appreciation,...**
- Scientific fraud
- Terrorism fraud - ‘higher’ motivation? Short term financial?
Social aspects of fraud management:

*Unwillingness to admit to being defrauded:* ‘*We have no fraud*’, said to me by a banker at a conference

*Belief that a bank has a very good system deters fraud* .... and the converse ...
The economic imperative

1) Not worth spending $200m to stop $20m fraud
   e.g. Letter from London Times, August 13, 2007
   “Sir, I was recently the victim of an internet fraud. The sum involved was several hundred pounds. My local police refused to investigate, stating that their policy was to investigate only for sums over £5000.”

2) The Pareto principle
   the first 50% of fraud is easy to stop; next 25% takes the same effort; next 12.5% takes the same effort; ...

3) Resources available for fraud detection are always limited
   - in the UK around 3% of police resources go on fraud
   - this will not significantly increase
If we cannot outspend the fraudsters we must out-think them

and bring sophisticated advanced technologies to bear to stop them

such as the technologies being discussed at this meeting
General problems in fraud detection

- may be huge data sets: both $d$ and $n$
- most variables will be irrelevant
- most cases not fraud: classic DM needle in haystack problem
- evolutionary arms race
- leapfrog of prevention and detection
- leapfrog of operations and exploration
- may involve complex data types (images, signals, text, networks)
- the role of data fusion: integrating data from multiple sources
- complicates disclosure risk - example of identity theft
- makes the large $d$ problem even worse!
- introduces additional risks of errors

Which data sources to use in detection depend on type of fraud

- numberplate recognition
- face recognition
- gait recognition
- credit card spend patterns
- money transfer patterns
- travel patterns
- medical records

_Civil liberties issues_
II: How big is fraud?

“Participants in our study estimate U.S. organizations lose 5% of their annual revenues to fraud. Applied to the estimated 2006 United States Gross Domestic Product, this 5% figure would translate to approximately $652 billion in fraud losses.”

Association of Certified Fraud Examiners
<table>
<thead>
<tr>
<th>Category</th>
<th>Cases</th>
<th>Convictions</th>
<th>Recoveries</th>
<th>Fines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corporate Fraud</td>
<td>490</td>
<td>124</td>
<td>42</td>
<td>14</td>
</tr>
<tr>
<td>Securities and Commodities Fraud</td>
<td>1,165</td>
<td>164</td>
<td>21</td>
<td>81</td>
</tr>
<tr>
<td>Health Care Fraud</td>
<td>2,423</td>
<td>534</td>
<td>1,600</td>
<td>173</td>
</tr>
<tr>
<td>Mortgage Fraud</td>
<td>818</td>
<td>204</td>
<td>1</td>
<td>231</td>
</tr>
<tr>
<td>Identity Theft</td>
<td>1,255</td>
<td>405</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Insurance Fraud</td>
<td>233</td>
<td>54</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>Mass Marketing Fraud</td>
<td>147</td>
<td>44</td>
<td>-</td>
<td>87</td>
</tr>
<tr>
<td>Asset Forfeiture/Money Laundering</td>
<td>473</td>
<td>95</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>7,004</strong></td>
<td><strong>1,624</strong></td>
<td><strong>1,674</strong></td>
<td><strong>587</strong></td>
</tr>
</tbody>
</table>

Source: U. S. Department of Justice, Federal Bureau of Investigation, Financial Crimes Report to the Public, Fiscal Year 2006
Cost of fraud

= immediate direct loss due to fraud
+ cost of fraud prevention and detection
+ cost of lost business (when replacing card)
+ opportunity cost of fraud prevention/detection
+ deterrent effect on spread of e-commerce
Does this matter to you personally?

**Example 1: Identity theft**

Fraudsters uses your name and identifying information to
- obtain credit cards
- phone and telecoms
- bank loans
- mortgages
- rent appartments
- give as identity if stopped for speeding, charged with crime, etc.

leaving you with the debts and problems
Identity theft in the USA

10 million victims in 2003
Average individual loss $\approx 5,000$
Total loss to individuals and businesses in 2003 $\approx 50 \text{ bn}$
(Federal Trade Commission survey)

+ time to sort out

⇒ Americans spent nearly 300 million hours resolving ID theft issues in 2003

Typically takes up to two years to sort out the problems, reinstate credit rating, reputation, etc, after detection
Example 2: Advance free fraud (the 419 scam)

Global Top 10 Nigerian 419 AFF Loss

How large are fraud datasets?

- My group deals with datasets involving m’ns or b’ns of transactions
- But one big terrorism case against one or two suspects can involve analysing
  - 6000 Gb of data
  - 8000 CDs
  - 200 phones
  - 70 premises on 3 continents
  - 400 computers

(Source: New Scientist, 4th August 2007, attributed to UK Home Office)
III: Fraud in banking

Banking fraud has many aspects, including:

- money laundering
- identity theft
- employee/staff fraud (sleepers)
- against individuals
- against organisations
- etc
My main focus here is retail or consumer banking fraud
- personal banking
- credit cards
- home mortgages
- car finance
- personal loans
- current accounts
- savings accounts
Nature of plastic card fraud data

- many transactions - billions - algorithms must be efficient
- mixed variable types (generally not text, image)
- large number of variables
- incomprehensible variables, irrelevant variables
- different misclassification costs
- many ways of committing fraud
- unbalanced class sizes (c. 0.1% transactions fraudulent)
- delay in labelling
- mislabelled classes
- random transaction arrival times
- (reactive) population drift
Credit card data (70-80 variables per transaction):

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transaction ID</td>
<td>Acquiring institution ID</td>
</tr>
<tr>
<td>Transaction type</td>
<td>Transaction authorisation code</td>
</tr>
<tr>
<td>Date and time of transaction (to nearest second)</td>
<td>Online authorisation performed</td>
</tr>
<tr>
<td>Amount</td>
<td>New card</td>
</tr>
<tr>
<td>Currency</td>
<td>Transaction exceeds floor limit</td>
</tr>
<tr>
<td>Local currency amount</td>
<td>Number of times chip has been accessed</td>
</tr>
<tr>
<td>Merchant category</td>
<td>Merchant city name</td>
</tr>
<tr>
<td>Card issuer ID</td>
<td>Chip terminal capability</td>
</tr>
<tr>
<td>ATM ID</td>
<td>Chip card verification result</td>
</tr>
<tr>
<td>POS type</td>
<td>. . . . . . . .</td>
</tr>
<tr>
<td>Cheque account prefix</td>
<td></td>
</tr>
<tr>
<td>Savings account prefix</td>
<td></td>
</tr>
</tbody>
</table>
A commercial example of fraud data

US Patent 5,819,226 (see USPTO website) on *Fraud detection and modeling*, (HNC Software in 1992) lists the following variables:

- Customer usage pattern profiles representing time-of-day and day-of-week profiles;
- Expiration date for the credit card;
- Dollar amount spent in each SIC (Standard Industrial Classification) merchant group category during the current day;
- Percentage of dollars spent by a customer in each SIC merchant group category during the current day;
- Number of transactions in each SIC merchant group category during the current day;
- Percentage of number of transactions in each SIC merchant group category during the current day;
- Categorization of SIC merchant group categories by fraud rate (high, medium, or low risk);
- Categorization of SIC merchant group categories by customer types (groups of customers that most frequently use certain SIC categories);
- Categorization of geographic regions by fraud rate (high, medium, or low risk);
- Categorization of geographic regions by customer types;
- Mean number of days between transactions;
- Variance of number of days between transactions;
- Mean time between transactions in one day;
- Variance of time between transactions in one day;
- Number of multiple transaction declines at same merchant;
- Number of out-of-state transactions;
- Mean number of transaction declines;
- Year-to-date high balance;
- Transaction amount;
- Transaction date and time;
- Transaction type.
“Additional fraud-related variables which may also be considered are listed below”
Current Day Cardholder Fraud Related Variables: 
- bweekend: current day boolean indicating current datetime considered weekend.
- cavapvdl: current day mean dollar amount for an approval.
- cavaudl: current day mean dollars per auth across day.
- ccoscdoni: current day cosine of the day of month i.e. cos(day of month / month).
- ccoscdow: current day cosine of the day of week.
- ccoscmoy: current day cosine of the month.
- cdom: current day day of month.
- cdow: current day day of week.
- chibal: current day high balance.
- chidcapv: current day highest dollar amt on a single cash approve.
- chidcdec: current day highest dollar amt on a single cash advance decline.
- chidmapv: current day highest dollar amt on a single merch approve.
- chidmdec: current day highest dollar amt on a single merch decline.
- currbal: current day current balance.
- cvraud1: current day variance of dollars per auth across day.

Decline Data: 
- czrate1: day zip risk group `Zip very high fraud rate`.
- czrate2: SIC factor group 03.

Transaction Data by SIC and Fraud Rate Group: 
- ctdsfa04: current day total dollars of transactions in SIC factor group 04.
- ctdsfa05: current day total dollars of transactions in SIC factor group 05.
- ctnsfa01: current day total number of transactions in SIC factor group 01.
- ctnsfa02: current day total number of transactions in SIC factor group 02.
- ctnsra06: current day total number of transactions in SIC fraud rate group 06.

Transaction Data by Week: 
- raudymdy: 7 day ratio of auth days over number of days in the window.
- ravapvdl: 7 day mean dollar amount for an approval.
- ravaudl: 7 day mean dollars per auth across window.
- rddapv: 7 day mean dollars per day of approvals.
- rddau: 7 day mean dollars per day of auths.

Imperial College
London

NATO ASI: Mining Massive Data sets for Security

25
Unbalanced classes

Detector correctly identifies 99 in 100 legitimate transactions
and correctly identifies 99 in 100 fraudulent transactions

Pretty good?

But suppose only 1 in 1000 transactions are fraudulent
<table>
<thead>
<tr>
<th>True class</th>
<th>Legit</th>
<th>Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted class</td>
<td>Legit</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td>Fraud</td>
<td>1%</td>
</tr>
<tr>
<td>Numbers</td>
<td>999</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>True class</th>
<th>Legit</th>
<th>Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted class</td>
<td>Legit</td>
<td>989.01</td>
</tr>
<tr>
<td></td>
<td>Fraud</td>
<td>9.99</td>
</tr>
<tr>
<td>Numbers</td>
<td>999</td>
<td>1</td>
</tr>
</tbody>
</table>

0.99 / (9.99 + 0.99) = 0.09
91% of suspected frauds are in fact legitimate

This matters because:

- operational decisions must be made (stop card?)
- good customers must not be irritated

Customers are pleased you care

up to a point
Delay in learning class labels

- if fraud alarm is raised, then true class quickly known

- if no alarm, then not detected until statement

This makes it different from the standard supervised classification paradigm

- different fraud strategies:
  - isolated transactions, hope account not notice
  - spend as much as possible as quickly as possible

- banks cannot always say for sure when a fraud commences
Mislabeled classes

Not all fraudulent transactions are labelled as fraud
(account holder fails to check carefully)

Not all legitimate transactions are labelled as legitimate

There may be subtleties

e.g. account holder makes transactions and then claims card was stolen

Such transactions are fraudulent because the holder declares them as such
Reactive population drift

- banks implement detection/prevention strategies
- fraudsters don’t generally give up! but change strategies
- there are many different fraud strategies
- each may have many variants
- each requiring different solutions
  - phishing
  - skimming
  - shoulder surfing
  - lebanese loop
  - false fronts
  - counterfeit
  - advance fee fraud (419 scam, Nigerian Money Transfer fraud, etc)
e.g. variants of the 419 scam

Lottery scam
Counterfeit Postal draft scam
Over invoiced contract scam
eBay check (over) payment and refund scam
Unclaimed inheritance scam
Unclaimed bank account scam
Counterfeit Check scam
Dating-romance scam
Black (defaced) currency scam
Gold dust scam
Diamond scam
Fake bank scam
Housing scam
Anti-terrorist certificate scam
Disaster relief fund scam
Financial representative in your country scam

Work permit scam
Payment for art scam
Deceased next of kin scam
Construction sub contractor scam
Lower priced crude oil scam
SWIFT transfer scam
Antique export payments scam
University study place scam
Money from former ruler scam
Relative of holocaust victim scam
Identity theft
Jobs for professionals scam
Dead millionaire funds for charities or disaster relief scam
Very low interest loans for relatively small advance fees scam
Hotel bookings and refund
United Nations loan approval scam
Death threat scam
Recall: Plastic card fraud in the UK (Gordon Blunt)

The waterbed effect

Source: APACS

Imperial College
NATO ASI: Mining Massive Data sets for Security
London
Reactive population drift example 1: *Chip and PIN*

**Chip and PIN intended/predicted to end card fraud**  
After UK rollout on 14 Feb 06, CC fraud in UK did decline  
How much was a consequence of the publicity?

*but*  
- predicted to lead to increase in identity theft

*and*  
- Lloyds TSB observed increase in fraudulent use of UK cards in Europe (no C&P – mag stripe still counterfeited)
- observed increase in ATM and cardholder not present fraud
- in fact, crooks installed data skimmers into C&P terminals (such devices can be purchased for < £100), over £1m stolen from Shell gas stations
Reactive population drift example 2: passwords

*The trouble with passwords:*

- tell them to others
- write them down
- send them in email
- log onto remote servers and eavesdropped
- often easy to guess!
So they invented one-time passwords:

(i) algorithm generates a new password for each use
(ii) password based on time synchronisation between cardholder and authentication server
(iii) password based on a challenge from the server

Test !:
if you were a fraudster, how would you find a way round such a system?
Our project:
- four major banks providing us with data
- hundreds of millions of transactions

**Phase 1:** develop appropriate criteria for measuring performance of detection algorithm

Completed

**Phase 2:** develop, evaluate, refine detection algorithms

**Phase 3:** the future: implement in collaboration with the banks
What is a good system?

‘Classifies fraudulent transactions as fraudulent, and legitimate transactions as legitimate’?

But: no method is perfect
Need: criteria for assessing effectiveness

Timeliness: speed of classification is vital in fraud detection

Standard two class classification criteria inadequate:
- misclassification rate: treats two types of misclassification equally
- Gini coefficient (AUC): averages over all misclassification cost ratios
- Kolmogorov-Smirnov statistic: data driven cost ratio
Different performance criteria may lead to different models

Optimum error rate: top-left to bottom-right
Optimum Gini: bottom-left to top right   Benton (2001): ionosphere data

So it is sensible to use the same criterion for both
(a) parameter estimation and model choice
(b) performance assessment
Distinguish between

1) performance on a particular training set
   - if we have a set of data and wish to build a rule

2) likely future performance, unconditional on any particular training set
   - to choose which rule to use before collecting data
A very well known consumer credit organisation evaluates fraud using the two ratios

\[ R_1 = \frac{A}{A + C} \quad (= \text{Sensitivity} = \text{Recall}) \]
\[ R_2 = \frac{B}{A + B} \quad (= \text{False positive rate} = 1- \text{Precision} = \text{FDR}) \]

(Note: terms are not always used consistently by all authors)
In itself, this would appear to be fine

But in fact, the units of assessment they use are accounts

An account is flagged as potentially fraudulent if at least one transaction is so flagged

**Problem 1:** This means that one can make the probability of flagging an account as fraudulent as near to 1 as one wishes by examining enough transactions

**Problem 2:** Fails to include timeliness in the measure
A superior measure

An **epoch** is a sequence of transactions ending with either
(i) a *fraud flag* on a true fraud
Or
(ii) or end of observed sequence

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fraud</td>
</tr>
<tr>
<td>Fraud</td>
<td>$m_{f/f}$</td>
</tr>
<tr>
<td>Legitimate</td>
<td>$m_{f/n}$</td>
</tr>
<tr>
<td>Predicted class</td>
<td>True class</td>
</tr>
<tr>
<td>----------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>Fraud</td>
</tr>
<tr>
<td>Fraud</td>
<td>1</td>
</tr>
<tr>
<td>Legitimate</td>
<td>3</td>
</tr>
</tbody>
</table>

This matrix includes *timeliness* in the count $m_{f/n}$.
<table>
<thead>
<tr>
<th>Predicted class</th>
<th>True class</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fraud</td>
<td>Legitimate</td>
</tr>
<tr>
<td>Fraud</td>
<td>$m_{f/f}$</td>
<td>$m_{n/f}$</td>
</tr>
<tr>
<td>Legitimate</td>
<td>$m_{f/n}$</td>
<td>$m_{n/n}$</td>
</tr>
</tbody>
</table>

Overall performance measure for given threshold:

$$T_1 = \frac{m_{f/f} + m_{n/f} + km_{f/n}}{km_f + m_n}$$

where $k$ is the estimated relative cost of misclassifying a fraud as legitimate compared to misclassifying a legitimate as fraud.

Or, if the bank can afford to investigate $C$ cases

$$T_2: \text{minimise } m_{f/n} \text{ subject to } \left( m_{f/f} + m_{n/f} \right) = C$$
Performance plots

ROC plots

\[ \frac{m_{n/n}}{m_n} \text{ against } \frac{m_{f/n}}{m_f} \]

An alternative (equivalent) more relevant to fraud is to plot

\[ \left( \frac{m_{n/f} + m_{f/f}}{m_n + m_f} \right) \text{ against } \frac{m_{f/n}}{m_f} \]
Constructing suspicion scores

Core approaches:
- rule-based methods
- supervised classification
- anomaly detection
- change point detection
- multilevel methods (transaction/account/merchant)
- link analysis - networks

Activity records: sacrifice immediacy?
- not necessarily
- and have potential for more accuracy
Different approaches have different strengths and weaknesses

rule-based:
   need expert knowledge of past fraud behaviour
   highly effective at detecting known fraud types
   ineffective at novel types

supervised methods:
   need examples of past fraud
   can be effective at detecting similar occurrences
   ineffective at novel types

anomaly detection:
   good for new kinds of deviations (but behaviour may change)
   not good for known types
Some evidence for these things, but should be careful of generalising too freely
Rule-based methods

Rules from expert knowledge:

- two near simultaneous transactions using the same card at geographically dispersed locations
- small time between attempts to withdraw maximum amount
- excessively small transactions
- multiple small electrical items
- ....

Rules from analysis of past frauds - supervised
Supervised classification

Basic principle:

Given a set of known fraudulent and legitimate transactions/accounts,

along with descriptive variables for each,

condense these to a rule enabling correct classification of new transactions/accounts

using only their descriptive variables
Methods developed in several areas, including *statistics, pattern recognition, machine learning, data mining*

linear discriminant analysis, quadratic discriminant analysis, regularised discriminant analysis, naive Bayes, logistic discriminant analysis, perceptrons, neural networks, radial basis function methods, vector quantization methods, nearest neighbour and kernel nonparametric methods, tree classifiers such as CART and C4.5, support vector machines, rule-based methods, random forests, etc. etc. etc.
Example: Bank A: (Chris Whitrow)

- 175 million transactions: 1st August 05 to 30th Nov 05
- 16.8 million accounts
- 5,946 accounts with fraud at POS terminals
- 76 raw variables per transaction; mostly categorical
- rolling window activity records - 0, 1, 3, 7 days
- activity records $\Rightarrow$ 87 variables per transaction
Classification methods used in this study:

- logistic regression
- quadratic discrimination
- naive Bayes classifier
- decision tree
- k-nearest neighbour
- SVMs with radial basis kernels
- random forests
Two explorations:

1: *Random*:
   Train on random 70%, test on remainder
   Unrealistic?: - unchanging distributional assumption
   - a baseline?

2: *Prediction*:
   Train up to 30 Oct 05, test after 30 Oct 05
   To allow for population drift
   OK for illustrative purposes
   But question about what exactly it tells us
Random performance

![Graph showing performance comparison between different models: Random Forest, Logistic Regression, Support Vector Machine, Naive Bayes, QDA, CART, and KNN. The $y$-axis represents the Loss Function ($T_1$). Each model is compared across different categories (ix, 1, 3, 7). The graph indicates varying performance levels for each model category.]
Limitations of such comparative studies

Real world vs laboratory conditions

What is meant by ‘method’ in such studies?
- do we include data preprocessing/transformation
- do we include the variable selection method?
- do we include the method of choosing parameters?
- do we include method for handling missing values?

Who will use the method
An expert will obtain different results from a naive user

What is meant by ‘best’ at a higher level:
Method A usually ranks first, but sometimes last
Method B always ranks second or third
(May vary between application domains)
It is meaningless to evaluate methods out of context
One class modelling: outliers

Basic principle: *build a model for the ‘norm’ for this customer and detect when it deviates*

‘Norm’ can be based on
- this customer compared with self at previous times (jamjarring)
- this customer compared with other customers
- life stage card usage patterns
- segmentation into customer types
- a combination of these

Basic advantage of one-class approach
- can detect new kinds of anomalies, not seen before
- more power in dynamic fraud environment?
Modelling the norm

1) Assume distributional form for ‘normal behaviour’

\text{e.g. multivariate normal} \Rightarrow \text{Mahalanobis distance}

- accurate probability estimates
- relatively small sample sizes (can be important in fraud)

- sensitive to assumed form (e.g. if true distribution is skewed)
- and to outliers
- so can robustify
2) measure relative distance from ‘centre’ in any direction using Chebychev’s inequality: 
\[ P\left( \left| x - E(x) \right| > t\sigma^2 \right) < t^{-2} \]

3) nonparametric approaches (e.g. kernel density estimation)

Computer science work in this area is typically described in terms of using *distance based methods*
There can be subtle complications

which can be overcome by comparing the density estimate at a point with that at nearby points.
Example: Bank B: (Piotr Juszczak)
- 44,637 accounts
- 2,374,311 transactions
- 3,742 fraudulent accounts
- 53,844 fraudulent transactions
- 3 months data

77 raw variables, from which we used
- size of transaction
- difference between current and previous transaction size
- sum of current and previous transaction sizes
- product of current and previous transaction sizes
- time of transaction
- time between current and previous transaction
- merchant category code (MCC)
- ATM ID code
Preprocessing the categorical variables (MCC and ATM)

\[ A(j,i) = \text{no. times ATM } j \text{ is accessed from account } i \]

\[ ATM(j) = (A(j,1), A(j,2), \ldots, A(j,K))^T \]

⇒ dissimilarity matrix between ATMs
⇒ reduce dimensionality of ATMs using MDS
⇒ combine with continuous variables
Similar for MCCs
Used several methods for building the pdfs:

- Parzen kernel
- Naive Bayes with Parzen kernel for each variable
- Single multivariate Gaussian
- Mixture of multivariate Gaussian
- 1-nearest neighbour
- Support Vector Data Description
- Self-Organising map
- Minimum spanning tree data description
- Minimax probability machine
Data set 1

\[
\frac{m(f/f) + m(n/f)}{m(n)+m(f)}
\]
Anomaly precursors

Geographic locations (Piotr Juszczak and Gordon Blunt)
There are already segmentations of financial behaviour
e.g. the FRuitTs system

Can also try to segment frauds
- to define fraud behaviour types
Change point detection

Cumulative amount spent, £

Weeks

0 10 20 30 40 50

0 500 1000 1500 2000

Imperial College

NATO ASI: Mining Massive Data sets for Security

London
Example 3: Peer group analysis as a compromise between individual accounts and entire groups

Dave Weston

next talk
Link analysis/networks/graphical approaches (Kiriaki Platanioti, Nick Heard)

*Between people:* Fraudsters don’t work in isolation (e.g. credit cards stolen or cloned and passed on). Networks.

*Between fraud types:* a gang which carries out one kind of fraud probably also carries out others.

Hidden Markov models for state changes
Other related work in the group:

*Dimitris Tasoulis*: dynamic multivariate streaming data subject to random asynchronous and partial delays

*Matt Turnbull*: dynamic multivariate streaming data subject to MAR and NMAR missing values

*Christoforos Anagnostopoulos*: prediction using dynamic multivariate streaming data when bandwidths for measurements are limited

*Nicos Pavlidis*: choice of action with dynamic multivariate streaming data
Intervention (Iding Wu):

Is this account at high risk of default, fraud, etc?
Past data, with outcomes available
Looks like a standard two class supervised classification problem

But suppose the aim is to take some action depending on the class:
  If no, continue
  If yes, intervene (stop card, contact customer, etc.)

*But intervention changes the outcome*
*
The predictive model no longer applies*

How to build a model which will predict the outcome after intervention?
Ideal solution:
Randomised controlled trial: randomly assign customers to the intervene (A) / don’t intervene (B) groups and compare the outcomes

Not usually permissible in banking contexts

We have a data set subject to selectivity bias
Action A taken on those accounts thought likely to default
Action B taken on those accounts thought unlikely to default

Perhaps Action B would be more effective on those thought likely to default
One solution:

1) Heckman method or likelihood method
   (i) distributional assumption for unobserved selection variables
       (e.g. normality)
   (ii) estimate probability of selection for each x
   (iii) estimate mean of unobserved part of z distribution
   (iv) estimate regression coefficients

2) Assign new cases to A or B according as \( E(r_A) > E(r_B) \)

3) Observe outcome of new data points and update regressions
Actions A and B have response functions

\[ r_A = \beta_{0A} + \beta_{xA}x + \beta_{zA}z + \varepsilon_A \]
\[ r_B = \beta_{0B} + \beta_{xB}x + \beta_{zB}z + \varepsilon_B \]

(w.l.g. can assume \( z \) is independent of \( x \) and has zero mean)

But we observe only the \( x \) vector and the response \( z \) is unobserved and may be different for each individual and may not even exist (e.g. subjective assignment)
This is no problem if the assignment to actions uses only $x$

Since then

$$E(r_A \mid x) = \beta_{0A} + \beta_{xA}x + \beta_{zA}E(z \mid x)$$

$$= \beta_{0A} + \beta_{xA}x$$

And we can then choose action for future cases on the basis of which of

$$E(r_A \mid x) = \beta_{0A} + \beta_{xA}x \quad E(r_B \mid x) = \beta_{0B} + \beta_{xB}x$$

is greater

[In fact, interested only in the difference: $E(r_A \mid x) - E(r_B \mid x)$]
**But** suppose the assignment uses the unobserved $z$

e.g. assign to A if in $\Omega_A = \{ \gamma x + z > c \}$ and to B otherwise.

Then

$$\hat{E}(r_A | x) = \beta_{0A} + \beta_{xA} x + \beta_{zA} E_{\Omega_A}(z | x)$$

$$\neq \beta_{0A} + \beta_{xA} x$$
One resolution:

Make assumption about distribution of the unobserved $z$: $z \sim f(z)$

e.g. normal

Then can find $\Omega_A = \{ \gamma x + z > c \}$ using logistic regression on $x$

Then $E_{\Omega_A} (z | x) = \frac{\int_{\Omega_A} z f(z) \, dz}{\int_{\Omega_A} f(z) \, dz} = h(x)$
So that
\[
\hat{E}(r_A \mid x) = \beta_{0A} + \beta_{x_A} x + \beta_{z_A} E_{\Omega_A}(z \mid x)
\]
\[
= \beta_{0A} + \beta_{x_A} x + \beta_{z_A} h(x)
\]
which can be estimated purely from the responses and \(x\)

That is, we obtain estimates of \(\beta_{0A}, \beta_{x_A}, \beta_{0B},\) and \(\beta_{xB}\) which can be used to give
\[
E(r_A \mid x) - E(r_B \mid x) = (\beta_{0A} + \beta_{x_A} x) - (\beta_{0B} + \beta_{xB} x)
\]
Higher level descriptions (e.g. Benford’s law)

Distribution of first digits of the population sizes of US counties in 2000 (Adrien Jamain, 2001)

\[ P(D_1 = d) = \log \left(1 + \frac{1}{d}\right) \]
Distribution of first digits of transaction sizes of one account in an investment bank
Banking fraud summary

- dynamic, reactive
- high dimensions, messy data, large data sets
- unbalanced classes
- mislabelled classes, delay in labelling
- construct ‘suspicion score’
- many approaches
- how to measure performance
- Pareto Principle
- accept some degree of fraud
Note:

The various methods are not *alternatives*

They are to be used in conjunction
Fraud management requires a holistic approach, blending tactical and strategic solutions with the state-of-the-art technology solutions and best practice in fraud strategy and operations.

James Gilmour, Editor Credit Risk International, 2003
IV: Fraud in science
**Jon Subdø**, Radium Hospital, Olso
Common pain relievers (eg ibuprofen)

Papers in *Lancet* and *New England Journal of Medicine*

- NEJM editors found duplicated figures in the 2001 paper
- Subdø admitted duplicating some data, 900 fictitious patients
- ‘every patient in the study by Jon Subdø and colleagues had been invented,’ *Lancet*

Independent commission: ‘the bulk of Subdo’s 30-plus publications were invalid because of the fabrication and manipulation of data’

Subdø resigned, stripped of degrees
Misuse of public funds: criminal charges?
Jan Hendrik Schön, Lucent’s Bell Labs, New Jersey
- Superconductivity, molecular crystals, molecular electronics
- On track for Nobel Prize
- 1998-2001 one paper every 8 days
- A claim too far: suspicions aroused
- Others noted that different experiments had identical noise

Investigation showed be had falsified and fabricated experimental data, 1998-2001, on at least 16 occasions

8 papers withdrawn by Science
7 papers withdrawn by Nature

Experiments repeating the work failed to obtain the same results

Schön fired
Woo Suk Hwang, Seoul National University

Stem cell research, cloning

Ethical questions: women paid to donate eggs, eggs from junior researchers in the lab

- Junior colleague admitted faking data to please Hwang
- Colleagues claim Hwang admitted to faking data
- Identical photos described as of different kinds of cells
- Peculiar data traces, possibly suggesting manipulation
- University decided his test results were fabricated
- Science retracted his papers

Charged with fraud, embezzlement, violation of bioethics laws

Resigned from SNU, but still continuing with animal cloning experiments
Xiaowu Li, University of California at San Francisco
Falsified three images in a published paper

Jason W. Lilly, Boyce Thompson Institute at Cornell University
Electronically replicated the image of a single genetic assay and altered the copies

Charles N. Rudick, Northwestern University
Used a photo-altering program to change pictures of recorded nerve signals.

Shinichi Fujima, archaeologist
Faked all 168 sites he dug, burying artefacts before discovering them

Luk van Parijs, MIT
Fabricated data in papers and grant applications

Eric Poehlman, University of Vermont
Fabricated data in grant applications

And many others: Robert Gullis, Michael Briggs, Robert Slutsky, Roger Pauson,...
Who commits fraud?

_The naive innocent_, who cleans and selects data without being aware of what they are doing - many?

_The rogue scientist_, who distorts and fabricates data deliberately to support a position - few?

_The pyramid_ - Woo Suk Hwang down to PhD students fabricating results:

- most PhD dissertations are read by 3 people if you are lucky

_The slippery slope_
How much fraud is there in science?

US Office of Research Integrity Report, 2006

New cases

Year


60 70 80 90 100
How is scientific fraud committed?

Fabricated data

Adjusted data

Selected data

Changed photographs

Claimed photographs were of things they were not

Substituted chemical, biological samples
“our evidence suggests that mundane ‘regular’ misbehaviours present greater threats to the scientific enterprise than those caused by high-profile misconduct cases such as fraud.”

(Martinson et al)
Where’s the line?

X is conducting measurements of the concentration of a chemical end product of a reaction, in ten replications, each of which is hugely expensive to make. Which of the following observations should you include in the analysis?

1) X bangs into one of the test tubes, and breaks it, so that the last observation X has is just prior to breaking it.

2) X notices that the temperature setting was not quite correct on one of the test tubes.

3) X observes that the final value for one of the test tubes is an order of magnitude larger than the others.
4) X observes that the final value for one of the test tubes is a little larger than the predicted value.

5) X believes that, as he has observed in the past, the measurement instrument is introducing slight systematic bias, so he makes a standard adjustment.

6) X dropped the chemicals on the floor, so he invented the observations in accordance with the expected results, plus some random measurement error.

7) X didn’t have time to collect the data, so he used measurements he had obtained in a previous experiment.

8) X overslept, so he invented the data.
- You won’t include all the observations in the analysis
- You have to decide which to include
- Leading to risk of **selectivity bias**

- You won’t include raw data: data cleaning is normally a significant part of any data analysis
- You have to decide how to clean it.
- Leading to risk of **data distortion**

Is the statistical tool the right one?  The art of statistics

The most appropriate statistical analysis is a complex multivariate analysis of variance, including both within and between subjects factors, and allowing for correlation over time

but the experimenter won’t understand: add up the numbers and do a t-test
Is the statistician the right one? (Statistician selectivity bias)

c.f. approaching different expert witnesses until you find one prepared to give the opinion you want

You should be suspicious about any scientist who brings you perfect data for analysis (Scientist selectivity bias)
Three more cases

John Darsee, Harvard University

Leader in research in interventions to aid recovery from heart attacks

- Fellow workers became suspicious and went to head of lab
- No evidence, but investigated, asking Darsee for his data
- Darsee started creating data, recording measurements taken on the same day as if they were different days

Darsee admitted fabricating data
Stripped of Fellowship, but continued working in lab

Later investigation showed much of his earlier work also involved fabricated data, as far back as undergraduate level
William Summerlin

Claimed to be able to transplant without rejection

Proof: white mice with black skin patches transplanted onto them

Lab assistant noticed patches looked odd
- and could be washed off with alcohol....

Paul Kammerer

Inheritance of acquired characteristics

Proof: a toad which developed black pads on its feet

But these turned out to be injections of Indian Ink
Detecting scientific fraud

Small suspicions by colleagues
  Data too good to be true
  Difficulty of fabricating realistic data

A data mining challenge

Very occasionally by journal editors

Small initial suspicions trigger investigation, and investigation of small cracks reveal gaping crevasses
Contrast with financial and other fraud

1) Intention
   - in science, data adjustments, with the best of intentions;
     (the aim is not to spread a false idea, but rather, to spread information
     they ‘know’ is right, without getting the evidence and checking it)
   - in finance, deliberate attempt to deceive?

2) Individual
   - in science, an individual
   - in finance, typically a gang

3) Aim
   - in science, peer regard
   - in finance, steal money
V: Conclusions

Fraud detection problems
- may involve high dimensions, messy data, large $n$
- typically have unbalanced classes
- often have mislabelled classes, delay in labelling
- may involve dynamic, reactive data distributions

There are
- many approaches / different aspects
- issues of how to measure performance
- differences between laboratory vs life comparisons
- ....
Other, deeper questions

The economic imperative

About methodology
   How much do we learn from ad hoc comparisons of methods on particular data sets?

About society
   Is society changing?
   Accepting some degree of fraud?

Different domains
   Pose different problems
   Different kinds of data
   Require different solutions
Like the poor, fraud is always with us
END

d.j.hand@imperial.ac.uk

http://stats.ma.ic.ac.uk/djhand/public_html/
One time passwords:

Man-in-the-middle attacks:
Fraudster creates a false bank web site and entices user to log on, sending access information directly onto real bank site.

If logon is successful, then disconnects user.

 Trojan attacks:
Software installed on user’s computer, which then piggybacks on a banking session
