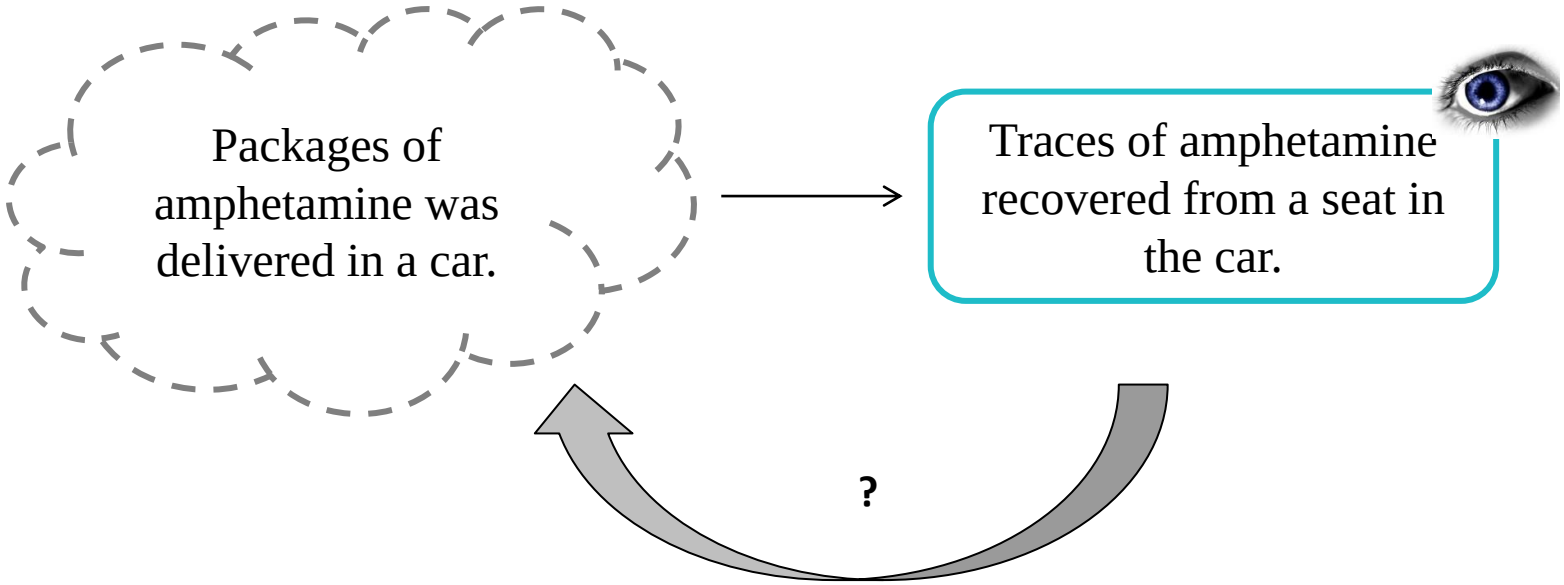
Meeting 16
Forensic applications, part I

(*Model(s) used at the Swedish National Forensic Centre, NFC*)

Evidence evaluation…

… is an *inductive* inferential process.

”Draw conclusions about what has happened from observed consequences of what happened – what we observe afterwards”



The two modes of the forensic process

- Investigative mode
- Evaluative mode

Investigative mode:

- Formulation of hypotheses regarding the activities (that have taken place at the crime scene)
- Definition of criteria for subsequent recovery of traces

Evaluative mode:

- Specific questions (formulated hypotheses) about recovered traces and their possible links to suspects or seized goods are treated by use of probabilistic reasoning

…but the two modes come interchangeably in course of the investigation:

More about the investigative mode

- In the investigative mode *several hypotheses are formulated* about what might have happened at a crime scene, a scene of fire, a finding-place,…
- Assessment and interpretation of detailed observations lead up to a ranking of the hypotheses formulated
- The assessment is made by deeming how expected detailed observations are under each of the hypotheses formulated. – and falsify such hypotheses under which the observations are considered improbable

This is where the evaluative mode may enter…

- The finally retained hypotheses would form a context that serve as the explanation delivered about what happened at the scene – a kind of giant hypothesis, supported by the investigation, but to be challenged in court

More about the evaluative mode

- Some questions (hypotheses) cannot be assessed – completely or directly – at the scene (of crime, of fire)
- E.g. hypotheses linking recovered traces to subsequently identified suspects or seized material

  - Was it this shoe that made the recovered footwear mark?
  - Does the blood on the floor originate from the dead person found in the villa?
  - Were the glass fragments recovered from the suspect's jacket transferred from the smashed window at the crime scene?
  - Did somebody burn hazardous waste here?
  - Were the two seized bags of amphetamine parts of the same manufacturing batch?
  - …

Common for these types of questions is that they have the ultimate answers Yes and No.

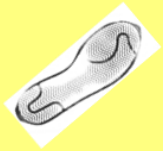
- Conclusions from the evaluative mode may
  - Sort things out at a crime scene/finding-place that helps in deciding along which path the subsequent investigation should continue…

    … should the conclusion be interpreted as a Yes or a No by the CSI
  - Be used as support for an hypothesis about a certain course at the crime scene/finding-place
  - constitute a self-standing piece of forensic evidence that links a recovered material to an individual or another control material, or a specific class/category

Is it then possible to always conclude with a Yes or No?

Source level attributions

A recovered footwear mark and a pair of shoes seized with a suspect.



Blood recovered from a garment and DNA from swabbing a suspect.

Two seizures of amphetamine – same origin?

Forensic investigation – and evaluation…

Two seizures of amphetamine – same origin?

Findings:

- Both seizures (materials) have only caffeine as cutting agent
- The dry concentration of amphetamine is about 40 % in both seizures
- The two seizures show similarities in their impurity profiles (presence of small amounts of other substances than amphetamine – bi-products in the manufacturing)

What do these findings signify?

Same origin with 100 % certainty? 90 % certainty? 50 % certainty?

Two seizures of amphetamine – same origin?

The question (most probably) put by the commissioner…

> Do the two seizures have a common origin (come from the same batch of manufacturing)?

…can be reformulated into a ***main hypothesis***

> $\boldsymbol{H_m}$ **:** The two seizures have a common origin

The main hypothesis is a *statement* that constitutes *one* explanation – but not necessarily a good one – to the findings obtained.

- Caffeine as single cutting agent in both
- Similar dry concentrations
- Similarities between impurity profiles

The "forensic contribution" of this statement consists of the belief in the statement – and its relevance for the current alleged activity.

N.B! H_m can only be true or false. It is the uncertainty about its truth that is the subject of discussion.

Two seizures of amphetamine – same origin?

Focusing on the belief (or what would be a proper expression)…

When we cannot categorically state that we are 100% (or 0%) certain about the truth of H_m we must use probability calculus.

Is it then possible to directly estimate the probability that

> ***H_m* :** The two seizures have a common origin

is true?

> Answer: No.
>
> This probability is deemed on by ***combining*** the forensic evaluation with other (non-forensic) information from the investigation (supporting or non-supporting H_m).

Two seizures of amphetamine – same origin?

Other information

Forensic evaluation

H_m **:** The two seizures have a common origin

Final probability of H_m being true

Other information

Common platform of evaluation

Forensic evaluation

Alternative hypothesis

e.g. $\boldsymbol{H_a}$ **:** The two seizures have different origins

Should be chosen to cover all relevant alternatives to the main hypothesis.

H_m : The two seizures have a common origin

H_a : **The two seizures have different origins**

Forensic evaluation

How expected/probable are…

- Both seizures (materials) have only caffeine as cutting agent
- The dry concentration of amphetamine is about 40 % in both seizures
- The two seizures show similarities in their impurity profiles (presence of small amounts of other substances than amphetamine – bi-products in the manufacturing)

…if ***the main hypothesis*** is true? $\Rightarrow P(Findings|H_m)$

…if ***the alternative hypothesis*** is true? $\Rightarrow P(Findings|H_a)$

$$\text{Forensic value of evidence} = V = \frac{P(Findings|H_m)}{P(Findings|H_a)}$$

$V > 1 \Rightarrow$ The findings are V times more probable if H_m is true compared to if H_a is true

$V < 1 \Rightarrow$ The findings are $1/V$ times more probable if H_a is true compared to if H_m is true

H_m : The two seizures have a common origin

H_a : **The two seizures have different origins**

Other information

How probable – prior to the forensic investigation – is…

H_m : The two seizures have a common origin ? $\Rightarrow P(H_m)$

…and how probable – prior to the forensic investigation – is …

H_a : The two seizures have different origins ? $\Rightarrow P(H_a)$

$$\text{Prior odds} = O = \frac{P(H_m)}{P(H_a)}$$

H_m : The two seizures have a common origin

H_a : The two seizures have different origins

Other information

$$O = \frac{P(H_m)}{P(H_a)}$$

$$V = \frac{P(Findings|H_m)}{P(Findings|H_a)}$$

Forensic evaluation

Bayes'theorem:

$$\frac{P(H_m|Findings)}{P(H_a|Findings)} = V \times O$$

Posterior odds

$$P(H_m|Findings) = \frac{V \times O}{V \times O + 1}$$

Final probability of/degree-of-belief in H_m

H_m : The two seizures have a common origin

H_a : **The two seizures have different origins**

Probability of whom?...

A source level attribution is generally in Sweden a forensic investigation with the prosecutor as ”destination”.

It is the prosecutor (via the police leader of the preliminary investigation) who (at least in theory) …

- is involved with formulating the alternative hypothesis
- has to deem on the magnitude of the prior odds
- must consider if…

…the probability of the main hypothesis is sufficiently high to bring this hypothesis as evidence to the indictment

Hence, a decision rule without a utility/loss function?

Bayes' theorem both in terms of a mathematical formula and as a graphical description

$$\frac{P(H_h)}{P(H_a)} \times \frac{P(E|H_h)}{P(E|H_a)} = \frac{P(H_h|E)}{P(H_a|E)}$$

Prior odds
Measures how certain/uncertain the commissioner (prosecutor, police, judge) is about the truth of H_m before considering the outcome of the forensic investigation.

Forensic value of evidence, V
States how much more (or less) probable the forensic findings/evidence E are if H_m is true compared to if H_a is true.

Often a ***Likelihood ratio***, but could also be a general ***Bayes factor***

Posterior odds
Measures how certain/uncertain the commissioner is about the truth of H_m upon considering the outcome of the forensic investigation

$$\frac{P(H_h)}{P(H_a)} \times \frac{P(E|H_h)}{P(E|H_a)} = \frac{P(H_h|E)}{P(H_a|E)}$$

Prior odds

Posterior odds

high

1

low

Forensic value of evidence, V

high

1

low

Prosecutor

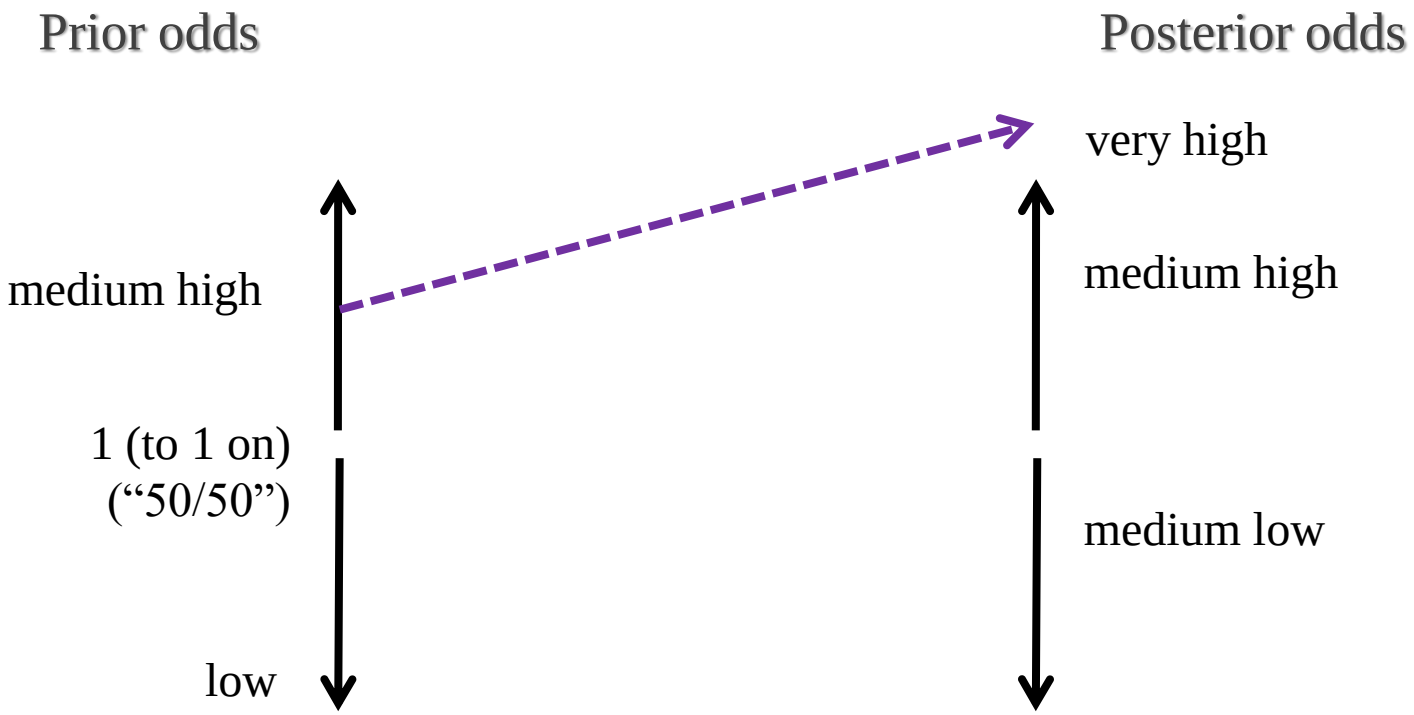
Defence

NFC

Prosecutor

Defence

- With the same pair of main and alternative hypothesis the forensic findings always have the same forensic value of evidence (likelihood ratio) i.e. the arrow has the same angle.
- The posterior odds can on the other hand differ depending on the magnitude of the prior odds



Estimation/calculation – in practice – of the magnitude of the forensic value of evidence (*V*)

In most forensic subject fields today there are no validated mathematical models to support the calculation of the forensic values of evidence.

Lack of background/reference data is the main explanation.

A forensic laboratory should however have uniform standards for reporting their values of evidence.

When models and data are lacking, the components of the value of evidence (i.e. the probabilities in the numerator and denominator of *V*) must be assigned based on (subjective and/or collective) experience and subject knowledge.
$\Rightarrow$ Fairly rough estimates of the magnitudes

$\Rightarrow$ All reporting of evidence from NFC – with or without using data bases and/or mathematical models – are made using a common ordinal scale of conclusions!

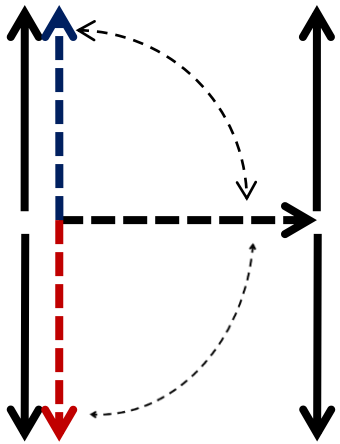
The scale of conclusion used at NFC:

Scale level	Magnitude of *V*	”Explanation”
		The findings are…
+4	at least one million	…at least one million times more probable…
+3	between 6000 and one million	…at least 6000 times more probable…
+2	between 100 and 6000	…at least 100 times more probable…
+1	between 6 and 100	…at least 6 times more probable…
0	between 1/6 and 6	… equally probable …
		…if the main hypothesis is true compared to if the alternative hypothesis is true
–1	between 1/100 and 1/6	…at least 6 times more probable…
–2	between 1/6000 and 1/100	…at least 100 times more probable…
–3	between 1/(one million) and 1/6000	…at least 6000 times more probable…
–4	at most 1/(one million)	…at least one million times more probable…
		…if the alternative hypothesis is true compared to if the main hypothesis is true

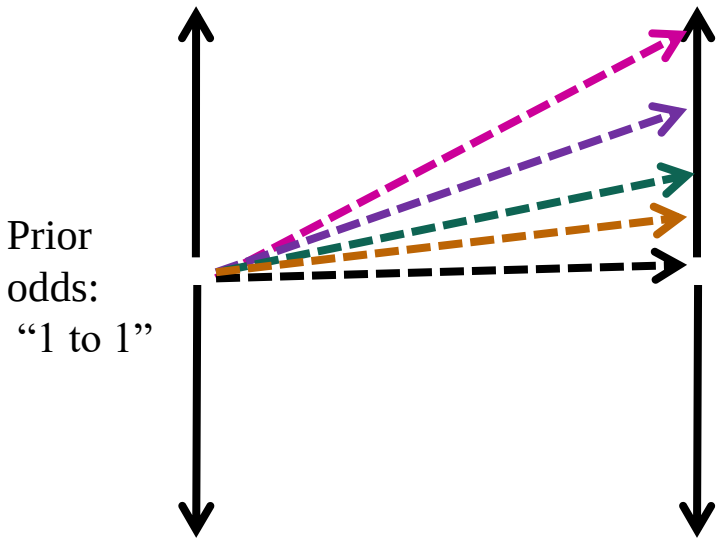
How were the levels of the NFC scale derived?

- The forensic value of evidence is a ratio of two probabilities (or a ratio of two probability density functions evaluated at the same point, or…a Bayes factor)
- In theory this value can vary from
  - zero (exclusion of the ***main hypothesis***)

  to
  - infinity (exclusion of the ***alternative hypothesis***)



- …but for a scale to be useful in practice the number of levels must be limited – decomposition of an infinitely long interval into a finite number of intervals
- We chose a set of levels *symmetrically* spread around the value 1, with four supporting levels (positive) and four non-supporting levels (negative), each level corresponding to an interval of values
- The lower limits of the intervals on the supporting side were chosen so that the (final) posterior probability would be acceptably high with respect to each level when the prior odds are equal to one.



Scale level	Posterior probability $P(H_h \mid E)$	Lower limit for V
+4 :	**> 0.999999**	**$10^6 \leq V$**
+3:	**> 0.9998**	**$6000 \leq V$**
+2:	**> 0.99**	**$100 \leq V$**
+1:	**> 0.86**	**$6 \leq V$**
0:	**between 0.14 och 0.86**	**($1/6 < V < 6$)**

Level + 2: Probability 0,99 (= 99%) is generally anticipated among lawyers as sufficiently high to ”confirm” a hypothesis (for bringing a specific evidence to the prosecution)

Level + 4: 10^6 was early chosen as a reasonable limit for the highest level for the evidentiary strength of DNA evidence in Sweden

Lecels +1 and + 3: The interval limits have been chosen so that the intervals successively increase in length regularly from a mathematical point of view (close to logarithmic increase). Probabilities 0,9998 and 0,86 became automatic consequences.

Nordgaard A., Ansell R., Drotz W. & Jaeger L.: ”Scale of conclusions for the value of evidence”. *Law, Probability and Risk* 11(1): 1-24.

Is this something specific for NFC Sweden?

Paternity index (Gürtler (1956)) – Early introduction of the *Likelihood Ratio*

Lindley (1977) ”A problem in forensic science”

Forensic Science Service, England and Wales: ”Case Assesment and Interpretation”, 1990s

Ecole des Sciences Criminelles, Université de Lausanne, CH

LGC Forensics Ltd, England

ESR, New Zealand

NFI, The Netherlands

NFC (SKL), Sweden

Guardia Civil, Spain

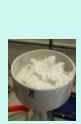
INCC, Belgium

IES, Krakow, Poland

EFE, Ireland

Eliciting probabilities in the amphetamine seizures case

Two seizures of amphetamine – same origin?



H_m : The two seizures have a common origin

H_a : The two seizures have different origins

Conditional probabilities of the findings assuming H_m is true

E_1: Both seizures (materials) have only caffeine as cutting agent
E_2: The dry concentration of amphetamine is about 40 % in both seizures
E_3: The two seizures show similarities in their impurity profiles (presence of small amounts of other substances than amphetamine – bi-products in the manufacturing)

All findings are *consistent* with H_m which means the conditional probability of obtaining them if H_m is true should be close to one.

Reasons for the probability not to equal one may be

- findings are expected to be even more consistent (e.g. exactly the same dry concentration in both seizures)
- more is expected as findings (than what has been obtained)

H_m : The two seizures have a common origin

H_a : **The two seizures have different origins**

Conditional probabilities of the findings assuming H_a is true

If we assume the two seizures consist of amphetamine from different batches of manufacturing generally …

… how probable do we deem …

E_1: Both seizures (materials) have only caffeine as cutting agent
E_2: The dry concentration of amphetamine is about 40 % in both seizures
E_3: The two seizures show similarities in their impurity profiles (presence of small amounts of other substances than amphetamine – bi-products in the manufacturing) ?

E_1: Both seizures (materials) have only caffeine as cutting agent?
E_2: The dry concentration of amphetamine is about 40 % in both seizures?

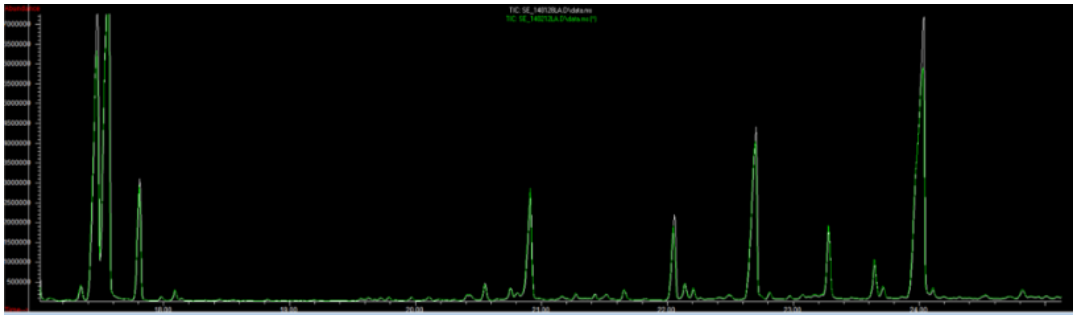
Use information from historical cases with seizures of amphetamine

- How often is caffeine the single cutting agent in amphetamine powder material? ⇒ guides the assignment of $P(E_1 \mid H_a)$
- How common is 40% dry concentration? ⇒ guides the assignment of $P(E_2 \mid H_a)$

H_m : The two seizures have a common origin

H_a : The two seizures have different origins

***E_3*:** The two seizures show similarities in their impurity profiles (presence of small amounts of other substances than amphetamine – bi-products in the manufacturing) ?



More complex!

Experience and knowledge based (subjective) assignment

- The profiles (set of peak areas) must first be interpreted one at a time
- The peak areas are continuous-valued – a set has a multivariate continuous probability distribution
- It is expected that there are slight discrepancies between the profiles (due to their continuous nature)
- Empirical data must be available from which it should be possible to study the variation among seizures with a common origin (within-variation) [H_m true] and the variation among seizures with different origins (between-variation) [H_a true]

How empirical data data may look like…

H_m : The two seizures have a common origin

H_a : **The two seizures have different origins**

Training data: 74 materials with 4-9 replicate analyses on each material

			TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS18	TS19	TS20	TS21	TS22	TS23	TS24	TS25	TS26	TS27	TS28	TS29	TS30
Manufacturing batch	Sample Multiplier	Inner standard	Ketoxime 1	Ketoxime 2	4-Methyl-5-phenylpyrimidine	Unknown	N-Benzylpyrimidine	N-Acetylamphetamine	N-Formylamphetamine	1,2-Diphenylethylamine	N,N-Dibenzylamine	1,2-Diphenylethylhanone	Benzylamphetamine	DPPA	DPIA 1	DPIA 2	alfa-Methyldiphenetyletylamine	DPIMA 1	DPIMA 2	Unknown A2	Naphthalene 1	Unknown A3	Naphthalene 2	N-Benzoylamphetamine	Unknown B2	2-Oxo	2,6-Dimethyl-3,5-diphenylpyridine	2,4-Dimetyl-3,5-diphenylpyridine	Pyridine 7 and 14	2,6-Diphenyl-3,4-dimethylpyridine	DPIF 1	DPIF 2
1	25	3013810	16476.74	5743.792	73655551.9	0	19605541.9	26975.65	87782.06	13687.44	0	57478.5	4241024	0	312960094.5	0	1002481	3031821	2092618	1451857	619155.3	0	78968.59	39242.94	2639141.39	0	444501	247555.2	1284954	255470.5	3113537.611	1555577
1	25	3041807	14647.12	6180.482	70972473.2	0	19014426.5	25421.87	87877.86	15871.02	0	55061.52	4099645	0	299165990.7	0	972134.5	2920446	1998073	1406672	600259.7	0	76315.09	38561.96	2515551.16	0	426041	229865.3	1245866	249647	2968307.003	1490150
1	25	2953134	14305.01	6220.258	69591541	0	18603912.3	27185.12	94006.3	14528.86	0	50755.59	3977849	0	290492463	0	936249.6	2842872	1962398	1305926	585274.8	0	76609.67	36961.51	2313236.55	0	417432.9	233694.8	1211059	242617.2	2833923.16	1365461
1	25	2987421	14060.76	5049.846	69969199.1	0	18694664.6	25039.16	84376.91	13780.97	0	51941.6	3945832	0	282162353.9	0	943215.4	2897387	2003740	1342803	601940.4	0	76087.32	36726.86	2455721.21	0	427800.8	233039.6	1232473	236088.8	2919125.854	1463175
1	25	3016062	13945.42	5786.284	70397076.3	0	18837813.5	25138.61	85836.93	12957.78	0	52974.17	4018744	0	295889196.3	0	943355.3	2852884	1947757	1352582	595472.4	0	76249.4	37179.79	2426612.46	0	419281.2	225739.6	1205542	233699.6	2862987.054	1419028
1	20	3031551	216117.3	100238.2	2131672.7	0	786369.673	293466	94173.32	0	0	14663.85	2204719	0	291092096.2	0	684171.8	2002366	1343762	1739857	501488.1	0	77609.63	544246.9	3826524	0	523745.7	357879.5	1581245	380597.5	4774159.742	2442870
1	20	3056269	215690.4	97407.6	2258413.22	0	829676.709	275575.5	94023.11	0	0	16570.79	2214260	0	282455295.1	0	665452.8	1927273	1280830	1689038	485353.8	0	71723.27	527817.8	3714198.65	0	520590.6	350905.6	1531466	374475.1	4726833.757	2412960
1	5	2846569	223754.6	115411.4	462763.166	0	203162.577	448899.6	78368.2	12562.88	0	40149.94	541765.2	0	234254300.6	0	595854.2	1880780	1251785	5063921	540045.4	724296.2	127213.9	2184442	12496394.8	43751.86	1358373	898749.5	4156964	1149792	15663212.08	8213142
1	5	2887200	198264.8	101397.5	449267.33	0	191566.429	400046.3	76392.33	12420.26	0	37813.36	442926.2	0	212362374.7	0	552614.5	1617674	1081968	5174910	463241.4	712926.6	112759.4	2130383	12317762.6	42971.28	1264927	868448.2	3867105	1093061	14906590.96	7859105
2	25	3018222	17235.38	6273.184	73795105.4	0	19884851.6	31318.66	91299.77	14805.19	0	42614.29	4262590	0	310292921.5	0	1006233	3085157	2099866	1349110	635707.6	0	82821	41819.85	2502275.93	0	442165.4	242220	1295051	254409.5	3062769.471	1535756
2	25	3032803	16486.07	6588.997	69989151.9	0	18808925.3	31242.22	87923.64	14027.28	0	42727.26	4000821	0	295475405.6	0	949630.8	2870445	1960282	1293844	587420.9	0	76199.06	38823.08	2277234.65	0	421240.5	234206.7	1202510	241997.1	2861701.334	1421529
2	25	3093308	17334.19	6658.91	71332017.7	0	19114878.4	32870.71	95246.09	14415.99	0	43932.69	4136092	0	299072632.2	0	975972.4	2893586	1229466	1996587	565350.2	0	80193.82	37429.97	2327861.52	0	434360.5	236498.9	1255750	248931.3	2981716.115	1495762
2	25	3011433	16603.03	6018.898	70676469.4	0	18967759.7	31139.8	89539.75	13580.49	0	43627.72	4087477	0	298390830.9	0	966309.3	2939138	2023813	1355112	603248.6	0	75225.9	38629.29	2379279.46	0	434714	231513.5	1239024	244545.1	2942339.146	1463058
2	25	3059922	15722.31	5606.733	70905652.3	0	18928398.8	29165.71	86859.77	14137.84	0	43067.68	4076822	0	298719208.4	0	976948.4	2944149	2038768	1282194	597795.7	0	77894.46	37511.44	2202687.2	0	427023.7	229608.7	1236216	239138.7	2920589.525	1449480
2	40	3077662	178896	75889.71	71814424.2	0	15542103.7	187158.6	87911.72	14248.71	0	0	3184093	0	318566574.6	0	747023.4	1790711	1215665	1139194	426881.3	0	58359.41	249335.6	2100986.47	0	309043.4	196261.6	911329.9	208454.1	2382829.131	1225717
2	40	3275898	165542.3	72158.51	65975170.7	0	14644070.2	189549.9	85298.39	13544.59	0	0	3192797	0	322008447.5	0	749451.3	1857280	1262796	1179858	427100.9	0	58957.3	274632.2	2166807.11	0	306692.9	206207.7	904196.8	212105.7	2442386.069	1280992
2	40	2858661	173236.3	77108.35	47721502.8	0	11507987.4	203291.6	85753.25	12221.95	0	17992.9	2868951	0	285494707.6	0	679980.5	1626189	1085313	1035338	376839.1	0	51921.54	239207.5	1923159.93	0	286706.9	185862.1	855203.6	193823.4	2275978.484	1165615
2	40	2847073	157448.7	73347.42	35982682.3	0	9041385.15	177208.3	77365.75	9922.968	0	16977.06	2505916	0	251327048.1	0	593684.8	1413638	954316.4	953965.5	324549.1	0	44531.81	233229.2	1778910.14	0	266918.6	169995	771503.5	176002.7	2151603.139	1103499
3	25	3024978	15607.52	5833.815	59645680.9	0	16015474.4	53950.45	67241.06	9779.744	0	0	3579062	0	279276553.7	0	854365.2	2597974	1779715	1063657	533248.5	0	67806.59	49673.13	2016302.7	0	376795.3	215574.6	1104076	213645	2545781.282	1267058
3	25	2995500	16215.85	6413.923	57174318.8	0	15234840	52817.96	75399.23	12213.11	0	0	3433564	0	263578179.3	0	819306.6	2469062	1687622	1043238	505854	0	66045	48345.05	1939895.72	0	365183.6	200097	1050345	204744	2451076.604	1220952
3	25	3032406	17079.95	6694.254	58666625.9	0	15739273.6	53317.58	75970.87	11812.23	0	0	3539252	0	271711669.9	0	831251.8	2542547	1746838	1089627	531437.9	0	68687.11	52513.96	1977221.34	0	374291.9	210147.1	1087723	208897.3	2486091.871	1231996
3	25	2952422	15556.18	6017.646	57887709.7	0	15416450	51059.19	72203.99	10403.01	0	0	3401931	0	264203741.3	0	805165.6	2435965	1680434	1025507	513742.1	0	65493.74	47121.08	1823383.87	0	354211.8	197743.3	1008050	207999.9	2430919.031	1196256
3	25	3027947	15140.22	6185.819	56180439.6	0	15145942.1	51805.34	70626.69	12047.86	0	0	3421193	0	266249064.7	0	811393.3	2471560	1726829	1105805	517060.9	0	66215.68	48943.09	1879516.06	0	362179.2	198102.6	1048641	200843	2397721.845	1188657
|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
74	6	1865214	0	0	83250724.5	0	29063838.4	362015.8	268600.4	122649.1	0	0	3024261	0	217270428.8	0	26728193	1.26E+08	84078822	4961858	194871.9	0	106852.2	450913	7487676.86	482016.6	402237.8	2296577	1329297	908851.4	78080710.63	48598815
74	6	1821220	0	0	79105948.7	0	27696046.1	339782.3	250356.5	119647.6	0	0	2874472	0	206274177.8	0	25272801	1.2E+08	80499916	4746880	180781.9	0	99057.5	434049.7	7277979.9	462626.5	375928.3	2142432	1268401	864014.3	74449963.53	45884385
74	6	1808019	0	0	78977011.9	0	27569888.2	345568	255667.9	117992.3	0	0	2870990	0	206115314.4	0	25291400	1.19E+08	80935945	4720790	186857	0	100145.5	420449.4	7303971.36	462712.8	381782.8	2155029	1249315	867676.8	74379382.6	46129454
74	6	1838779	0	0	80329906.7	0	28068889.2	348891.8	254045.4	121521.8	0	0	2918690	0	210766488.9	0	25714842	1.22E+08	82295184	4851210	184289.5	0	105446.8	434654.8	7449699.68	478072.8	384772	2186955	1270928	884023	75746352.76	46671158
74	6	1814855	0	0	78636244.2	0	27455903.8	342626.1	250075	117475.9	0	0	2883142	0	206593937.2	0	25252017	1.2E+08	80182295	4840192	185783	0	98698.24	427460.6	7280252.38	465360.2	375721.1	2154055	1251466	867590	75037389.84	46402684

TS5	TS6	TS7	TS8
N-Benzylpyrimidine	N-Acetylamphetamine	N-Formylamphetamine	1,2-Diphenyletylamine
19605541.9	26975.65	87782.06	13687.44
19014426.5	25421.87	87877.86	15871.02
18603912.3	27185.12	94006.3	14528.86
18694664.6	25039.16	84376.91	13780.97
18837813.5	25138.61	85836.93	12957.78
786369.673	293466	94173.32	0
829676.709	275575.5	94023.11	0
203162.577	448899.6	78368.2	12562.88
191566.429	400046.3	76392.33	12420.26
19884851.6	31318.66	91299.77	14805.19
18808925.3	31242.22	87923.64	14027.28

Peak areas of 30 impurities

H_m : The two seizures have a common origin

H_a : **The two seizures have different origins**

Hence, the finding E_3 might be

$$E_{3,1} = \boldsymbol{y}_1 = \begin{pmatrix} y_{1,1,1} & y_{1,1,2} & \cdots & y_{1,1,30} \\ y_{1,2,1} & y_{1,2,2} & \cdots & y_{1,2,30} \\ y_{1,3,1} & y_{1,3,2} & \cdots & y_{1,3,30} \end{pmatrix}$$

3 replicate analyses (3 × 30 peak areas) on material 1

$$E_{3,2} = \boldsymbol{y}_2 = \begin{pmatrix} y_{2,1,1} & y_{2,1,2} & \cdots & y_{2,1,30} \\ y_{2,2,1} & y_{2,2,2} & \cdots & y_{2,2,30} \\ y_{2,3,1} & y_{2,3,2} & \cdots & y_{2,3,30} \end{pmatrix}$$

3 replicate analyses (3 × 30 peak areas) on material 2

but at the laboratory it is rare to have more than one analysis made on each material

What would the forensic value of this finding be?

$$\frac{P(\boldsymbol{H_h})}{P(\boldsymbol{H_a})} \times B(E_3) = \frac{P(\boldsymbol{H_h}|E_3)}{P(\boldsymbol{H_a}|E_3)}$$

Bayes factor

$$\boldsymbol{y}_1 = \begin{pmatrix} y_{1,1,1} & y_{1,1,2} & \cdots & y_{1,1,30} \\ y_{1,2,1} & y_{1,2,2} & \cdots & y_{1,2,30} \\ y_{1,3,1} & y_{1,3,2} & \cdots & y_{1,3,30} \end{pmatrix}$$

$$\boldsymbol{y}_2 = \begin{pmatrix} y_{2,1,1} & y_{2,1,2} & \cdots & y_{2,1,30} \\ y_{2,2,1} & y_{2,2,2} & \cdots & y_{2,2,30} \\ y_{2,3,1} & y_{2,3,2} & \cdots & y_{2,3,30} \end{pmatrix}$$

$\boldsymbol{H_m}$ **:** The two seizures have a common origin

$\boldsymbol{H_a}$ **: The two seizures have different origins**

The Bayes factor can be shown to be (Lindley, Biometrika, 1977):

$$B(E_3) = \frac{\int f(\bar{\boldsymbol{y}}_1|\boldsymbol{\theta}) \cdot f(\bar{\boldsymbol{y}}_2|\boldsymbol{\theta}) \cdot g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\bar{\boldsymbol{y}}_1|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \times \int f(\bar{\boldsymbol{y}}_2|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

where $f(\bar{\boldsymbol{y}}_1|\boldsymbol{\theta})$ and $f(\bar{\boldsymbol{y}}_2|\boldsymbol{\theta})$ are conditional probability density functions of the mean vector of $E_{3,x}$ and $E_{3,y}$ respectively given the true mean $\boldsymbol{\theta}$ of the peak areas of material 1 and material 2 respectively (under $\boldsymbol{H_m}$ the two materials are assumed to have the same mean), and $g(\boldsymbol{\theta})$ is the probability density function of the variation in mean between materials with different origins.

$$\boldsymbol{y}_1 = \begin{pmatrix} y_{1,1,1} & y_{1,1,2} & \cdots & y_{1,1,30} \\ y_{1,2,1} & y_{1,2,2} & \cdots & y_{1,2,30} \\ y_{1,3,1} & y_{1,3,2} & \cdots & y_{1,3,30} \end{pmatrix}$$

$$\boldsymbol{y}_2 = \begin{pmatrix} y_{2,1,1} & y_{2,1,2} & \cdots & y_{2,1,30} \\ y_{2,2,1} & y_{2,2,2} & \cdots & y_{2,2,30} \\ y_{2,3,1} & y_{2,3,2} & \cdots & y_{2,3,30} \end{pmatrix}$$

$\boldsymbol{H_m}$ **:** The two seizures have a common origin

$\boldsymbol{H_a}$ **: The two seizures have different origins**

$$B(E_3) = \frac{\int f(\bar{\boldsymbol{y}}_1|\boldsymbol{\theta}) \cdot f(\bar{\boldsymbol{y}}_2|\boldsymbol{\theta}) \cdot g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\bar{\boldsymbol{y}}_1|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \times \int f(\bar{\boldsymbol{y}}_2|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

Aitken & Lucy (Applied statistics, 2004) showed that with $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ assumed equally and normal distributed and $g(\boldsymbol{\theta})$ estimated by a kernel density function with a Gaussian kernel $B(E_3)$ could be written

$$B(E_3) = \frac{f_n(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2 | p, m, n_1, n_2, \boldsymbol{U}, \boldsymbol{C})}{f_d(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2 | p, m, n_1, n_2, \boldsymbol{U}, \boldsymbol{C})}$$

where $\boldsymbol{U}$ and $\boldsymbol{C}$ are the within-material and between-material covariance matrices of the vectors of p peak areas in the population of materials (two-level random effects model); m is the number of materials in the training data set and n_1 and n_2 are the number of replicate analyses in material 1 and 2 respectively.

$$\boldsymbol{y}_1 = \begin{pmatrix} y_{1,1,1} & y_{1,1,2} & \cdots & y_{1,1,30} \\ y_{1,2,1} & y_{1,2,2} & \cdots & y_{1,2,30} \\ y_{1,3,1} & y_{1,3,2} & \cdots & y_{1,3,30} \end{pmatrix}$$

$$\boldsymbol{y}_2 = \begin{pmatrix} y_{2,1,1} & y_{2,1,2} & \cdots & y_{2,1,30} \\ y_{2,2,1} & y_{2,2,2} & \cdots & y_{2,2,30} \\ y_{2,3,1} & y_{2,3,2} & \cdots & y_{2,3,30} \end{pmatrix}$$

$\boldsymbol{H_m}$ **:** The two seizures have a common origin

$\boldsymbol{H_a}$ **: The two seizures have different origins**

$$B(E_3) = \frac{f_n(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2 | p, m, n_1, n_2, \boldsymbol{U}, \boldsymbol{C})}{f_d(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2 | p, m, n_1, n_2, \boldsymbol{U}, \boldsymbol{C})}$$

Explicit expressions for f_n and f_d were shown to be

$$f_n(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2 | p, m, n_1, n_2, \boldsymbol{U}, \boldsymbol{C}) =$$

$$= (2\pi)^{-p} \left|\frac{\boldsymbol{U}}{n_1}\right|^{-1/2} \left|\frac{\boldsymbol{U}}{n_2}\right|^{-1/2} |\boldsymbol{C}|^{-1/2} (mh^p)^{-1/2} \left| \left(\frac{\boldsymbol{U}}{n_1}\right)^{-1} + \left(\frac{\boldsymbol{U}}{n_2}\right)^{-1} + (h^2\boldsymbol{C})^{-1} \right|^{-1/2}$$

$$\times \exp\left\{ -\frac{1}{2} (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)' \left(\frac{\boldsymbol{U}}{n_1} + \frac{\boldsymbol{U}}{n_2}\right)^{-1} (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)' \right\}$$

$$\times \sum_{i=1}^{m} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}^* - \bar{\boldsymbol{x}}_i)' \left[\left[\left(\frac{\boldsymbol{U}}{n_1}\right)^{-1} + \left(\frac{\boldsymbol{U}}{n_2}\right)^{-1} \right]^{-1} + h^2\boldsymbol{C} \right]^{-1} (\boldsymbol{w} - \bar{\boldsymbol{x}}_i) \right\}$$

where h is a chosen bandwidth for the kernel density estimate;

$\boldsymbol{y}^* = \left[\left(\frac{\boldsymbol{U}}{n_1}\right)^{-1} + \left(\frac{\boldsymbol{U}}{n_2}\right)^{-1} \right]^{-1} \left(\left(\frac{\boldsymbol{U}}{n_1}\right)^{-1} \bar{\boldsymbol{y}}_1 + \left(\frac{\boldsymbol{U}}{n_2}\right)^{-1} \bar{\boldsymbol{y}}_2 \right)$; and $\bar{\boldsymbol{x}}_i$ is the mean vector of peak areas of the replicate analyses of material *i*in the training set.

$$\boldsymbol{y}_1 = \begin{pmatrix} y_{1,1,1} & y_{1,1,2} & \cdots & y_{1,1,30} \\ y_{1,2,1} & y_{1,2,2} & \cdots & y_{1,2,30} \\ y_{1,3,1} & y_{1,3,2} & \cdots & y_{1,3,30} \end{pmatrix}$$

$$\boldsymbol{y}_2 = \begin{pmatrix} y_{2,1,1} & y_{2,1,2} & \cdots & y_{2,1,30} \\ y_{2,2,1} & y_{2,2,2} & \cdots & y_{2,2,30} \\ y_{2,3,1} & y_{2,3,2} & \cdots & y_{2,3,30} \end{pmatrix}$$

$\boldsymbol{H_m}$ **:** The two seizures have a common origin

$\boldsymbol{H_a}$ **: The two seizures have different origins**

$$B(E_3) = \frac{f_n(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2 | p, m, n_1, n_2, \boldsymbol{U}, \boldsymbol{C})}{f_d(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2 | p, m, n_1, n_2, \boldsymbol{U}, \boldsymbol{C})}$$

$$f_n(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2 | p, m, n_1, n_2, \boldsymbol{U}, \boldsymbol{C}) =$$

$$= (2\pi)^{-p} |\boldsymbol{C}|^{-1} (mh^p)^{-1/2} \prod_{k=1}^{2} \left[\left| \frac{\boldsymbol{U}}{n_k} \right|^{-1/2} \cdot \left| \left(\frac{\boldsymbol{U}}{n_k} \right)^{-1} + (h^2 \boldsymbol{C})^{-1} \right|^{-1/2} \times \cdots \right.$$
$$\left. \cdots \times \exp\left\{ -\frac{1}{2} (\bar{\boldsymbol{y}}_k - \bar{\boldsymbol{x}}_i)' \left(\frac{\boldsymbol{U}}{n_k} + h^2 \boldsymbol{C} \right)^{-1} (\bar{\boldsymbol{y}}_k -)' \right\} \right]$$

Estimates of $\boldsymbol{U}$ and $\boldsymbol{C}$ (both $p \times p$) are made on the training $\bar{\boldsymbol{x}}_i$ data.

Here $p = 30$ and $m = 74$, $n_1 = n_2 = 3$
and $\bar{\boldsymbol{x}}_i$ is based on 4-9 replicate analyses ($i = 1, 2, \ldots,$ m

Problem?

Dimension reduction via graphical modelling (*again!*)

For a multivariate random vector with correlation matrix $\boldsymbol{R} = (r_{ij})$ the matrix of partial correlation coefficients can be obtained as follows:

Compute the inverse of $\boldsymbol{R} \Rightarrow \boldsymbol{R}^{-1} = \boldsymbol{Q} = (q_{ij})$

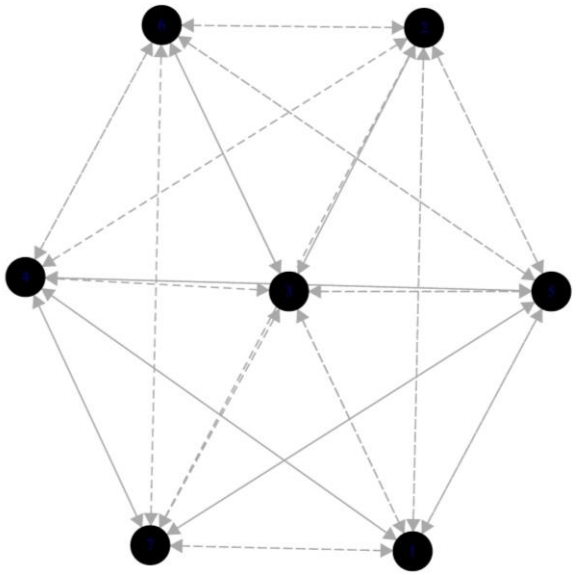
The partial correlation matrix is then $\boldsymbol{P} = (p_{ij})$ where $p_{ij} = \frac{-q_{ij}}{\sqrt{q_{ii} \cdot q_{jj}}}$

The partial correlation between two components (marginal variables) of a random vector is the degree of linear dependence that is unique between them, i.e. when all dependencies via the other components have been taken out.

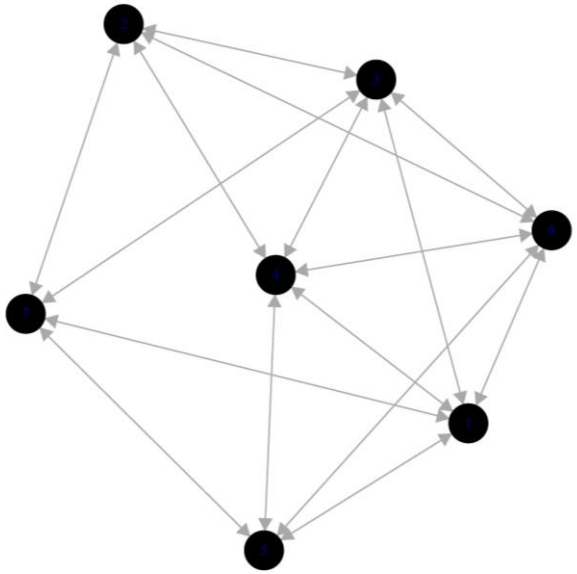
A graphical model of a random vector can be defined as a graphical model where the links (edges) between two components exist provided their partial correlation exceeds a chosen threshold.

Example Random vector with 7 components, all partial correlations are > 0.

Full model ($p_{ij} > 0$):

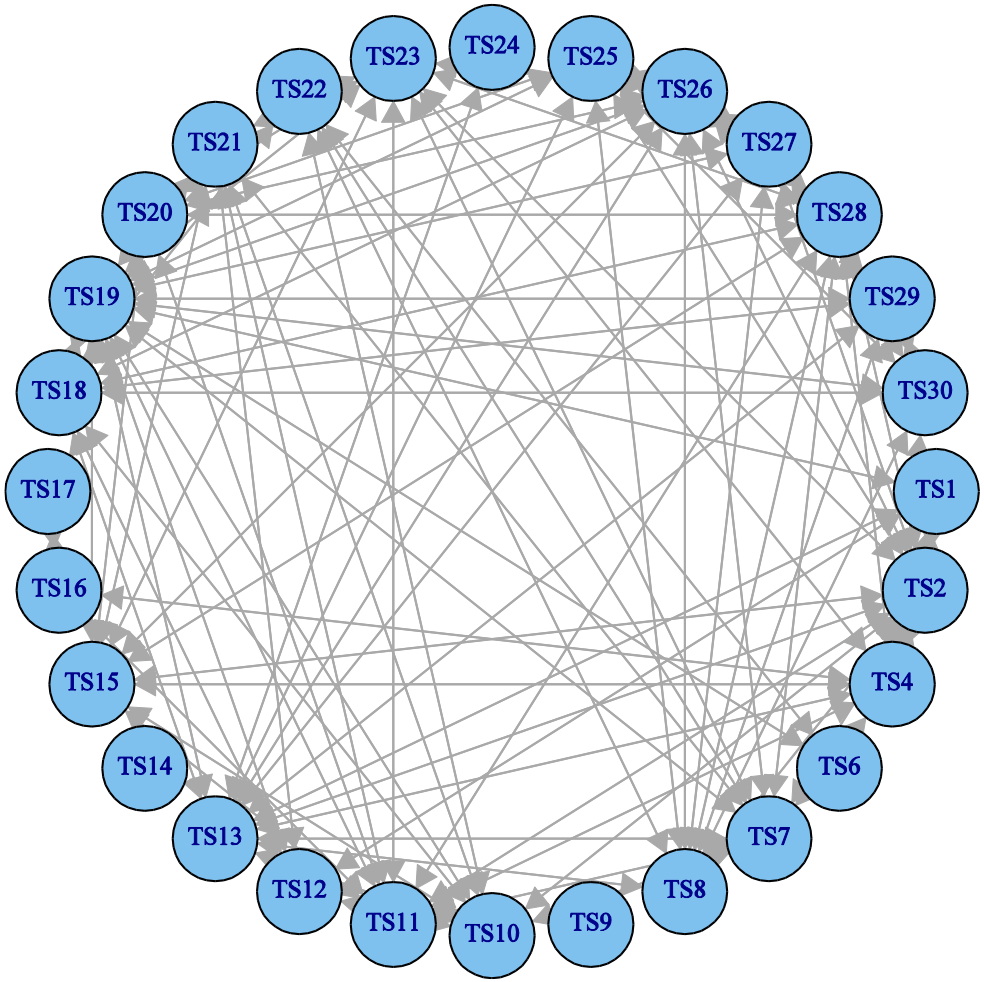


Reduced model ($p_{ij} > 0.5$):



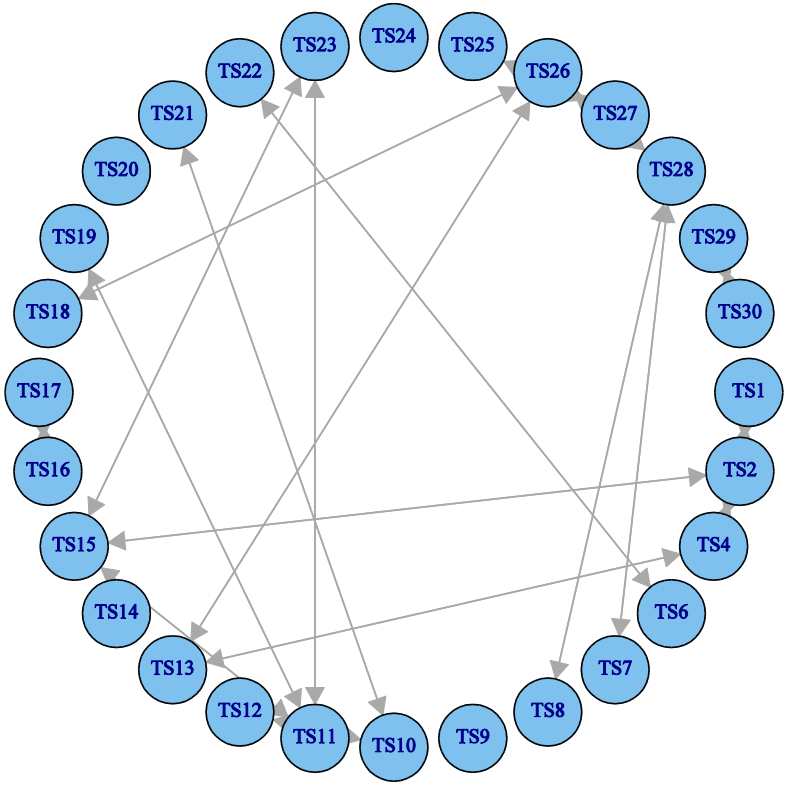
For the data with amphetamine impurities we name the impurities TS1, TS2, …, TS30 (Target Substance)

A graphical model based on partial correlations ≥ 0.2 becomes

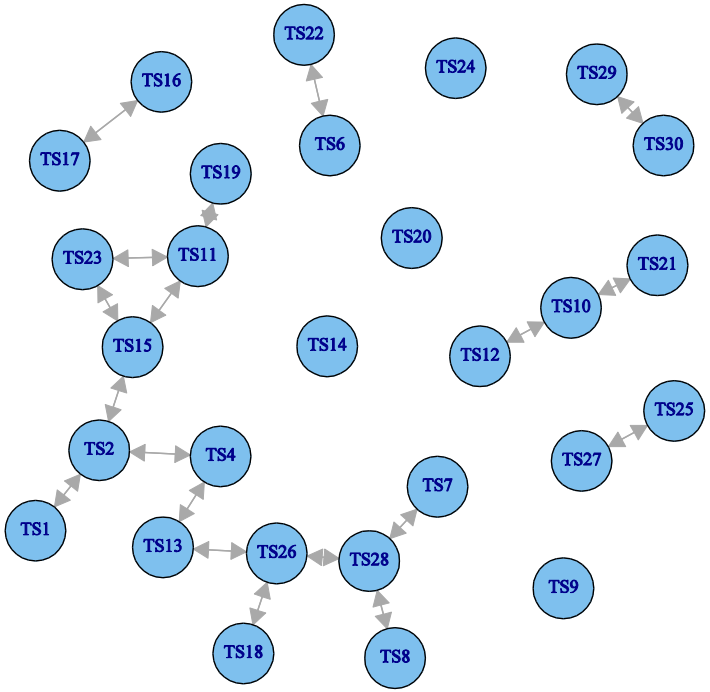


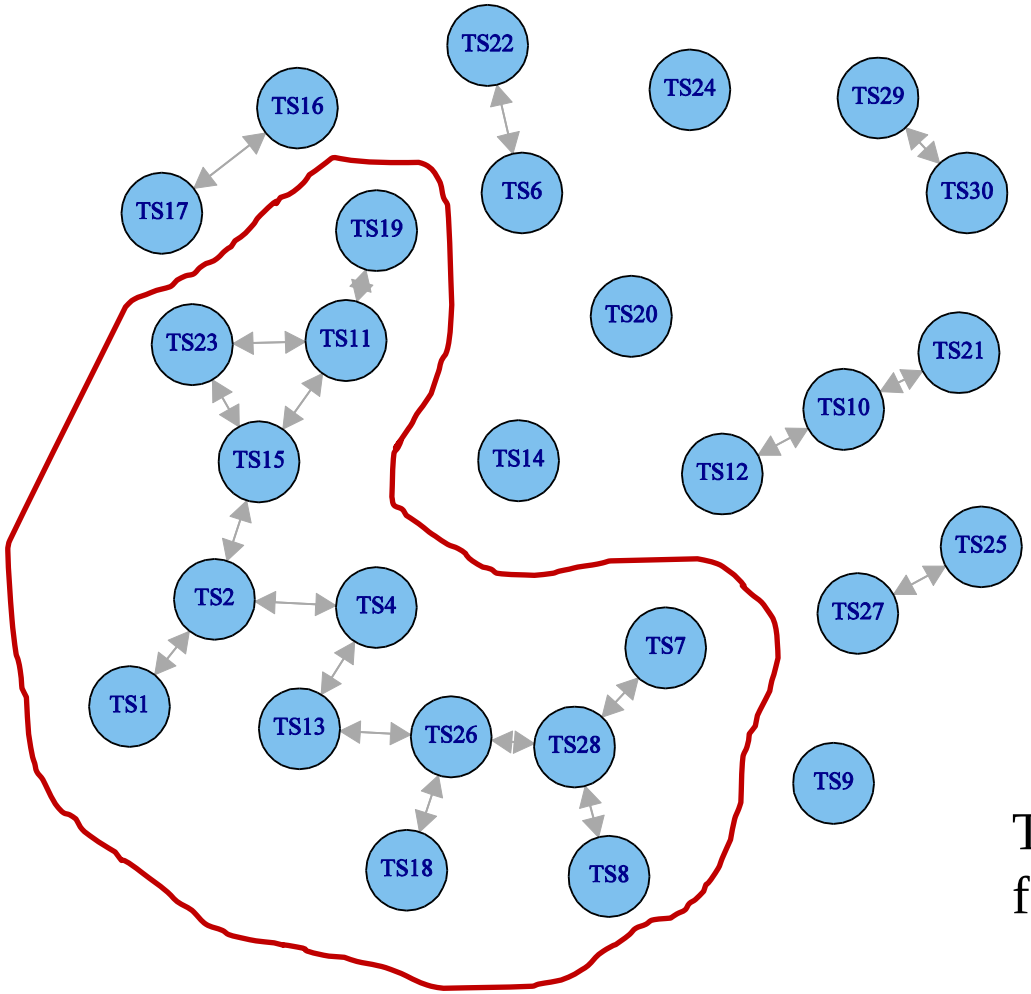
Chemical considerations about the substances gives that 28 of the 30 impurities should be retained (TS3 and TS5 are taken out).

Then, a graphical model based on partial correlations ≥ 0.4 becomes



with another layout:





If we know assume that partial correlations less than 0.4 can be considered as noise, we have 10 approximately uncorrelated graphs instead of 1 single graph with correlated components.

The largest graph has 13 nodes – 13 correlated variables.

So we have reduced the dimension from 28 to 13.

The Bayes factor may then be factorized into 10 factors:

$$B(E_3) = B_1 \cdot B_2 \cdot B_3 \cdot B_4 \cdot B_5 \cdot B_6 \cdot B_7 \cdot B_8 \cdot B_9 \cdot B_{10}$$

By using ***junction trees*** we can (most often) factorize the probability density function of the largest graph and so reduce the dimension even more.