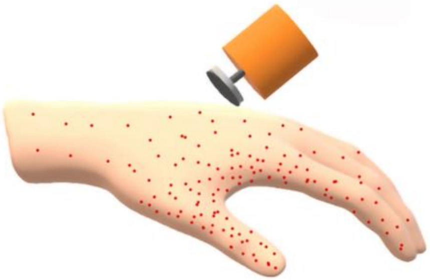
Meeting 19
Forensic applications, part II

Example:

Upon a shooting incident a person is apprehended, suspected of being the shooter.

His hands and clothes are sampled for searching so-called *gunshot residues* (GSR) [or *firearm discharge residues* (FDR), equal things].

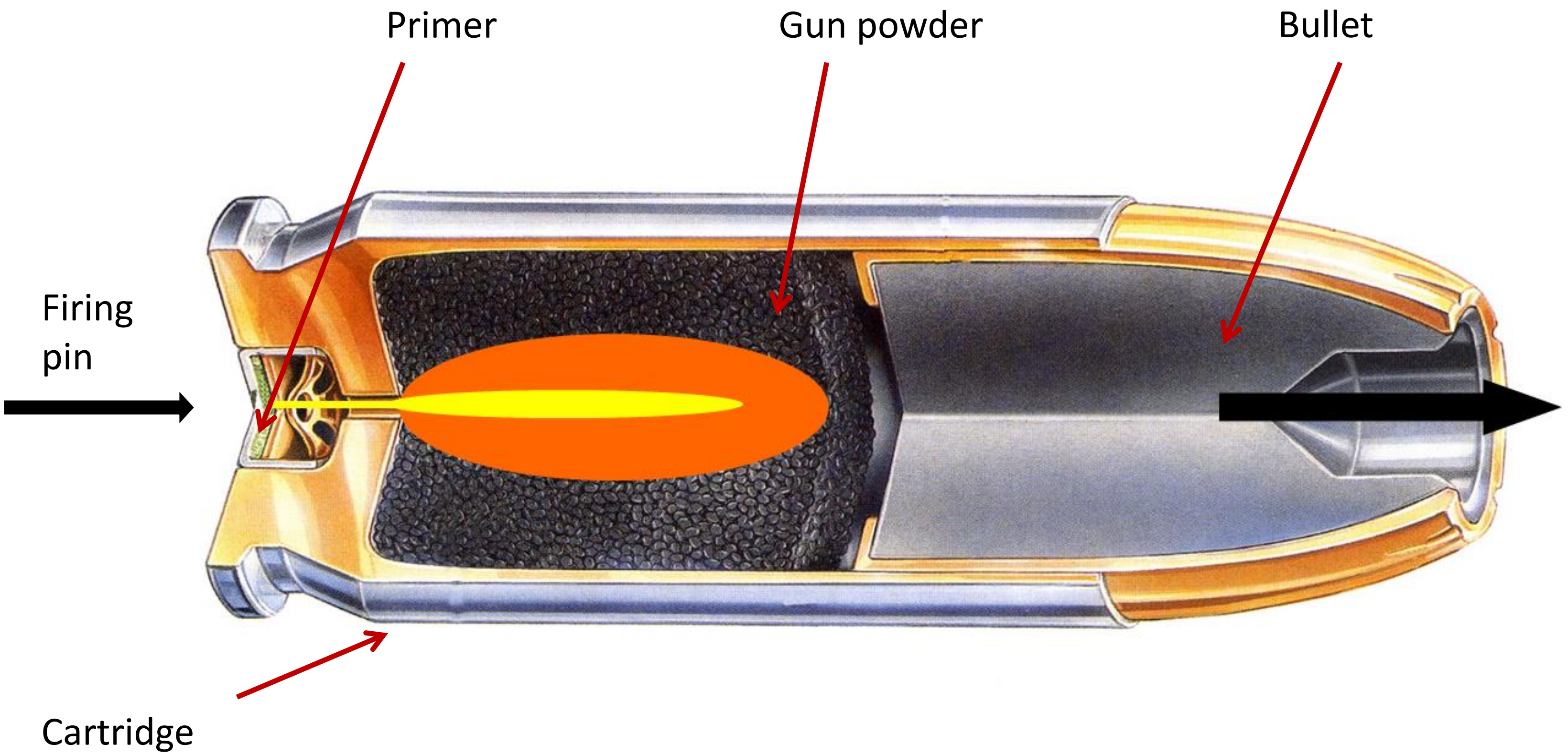


Findings of GSR is expected to give evidence for the suspect being the shooter.

What are GSR?

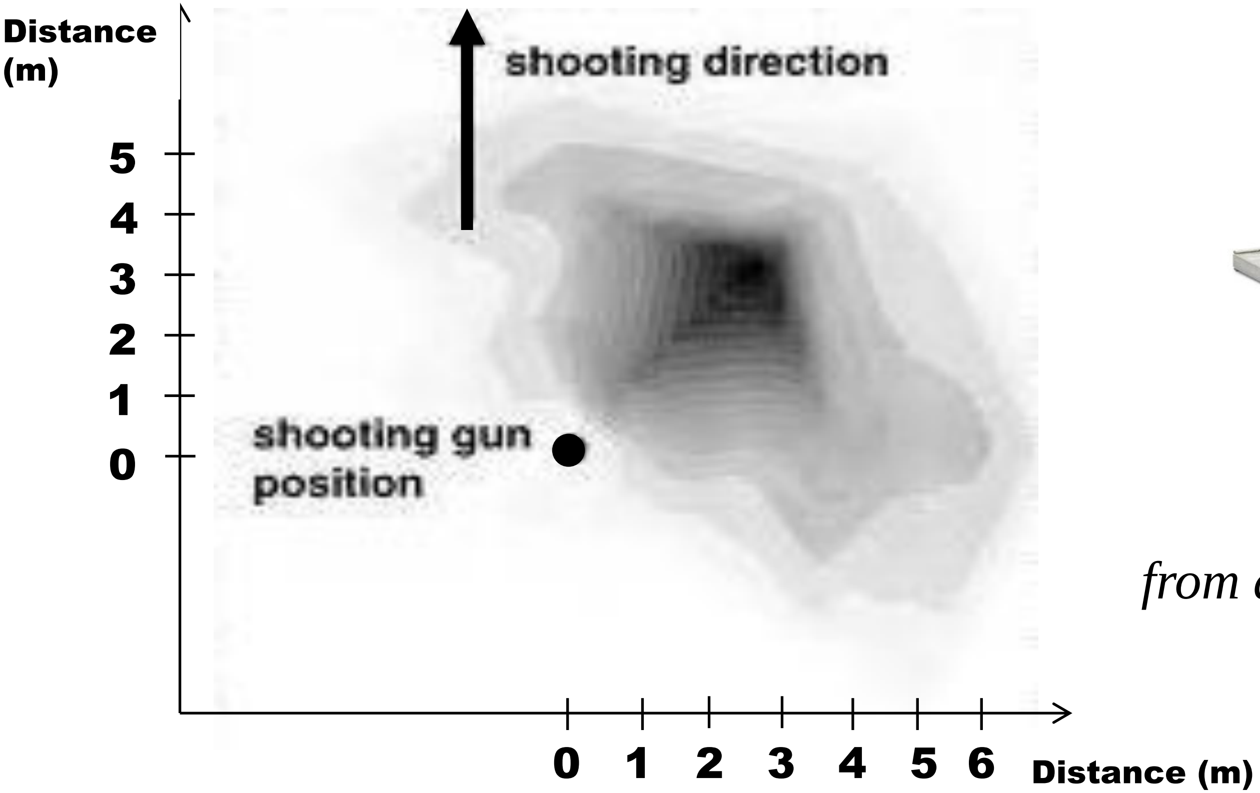
GSR are very small metallic/metalloid particles that come from the *explosive primer* of a cartridge. When the firing pin hits the explosive primer, it explodes and lightens the powder in the cartridge making the bullet to eject.

When exploding, the primer is fragmented into these very small particles.



The GSR are spread around the firearm that was discharged.

A typical pattern with shooting *indoors* with a pistol is:



from a Czech study

Patterns with shooting outdoors are of course affected by the weather conditions.

GSR are volatile.

Drop off garments and body parts quite quickly after deposition – half-life on hands is about 60 minutes, on gloves about 80 minutes

99% vanished after 6 hours.

Very sensitive to washing-off, sensitive to adverse weather (rain, wind).

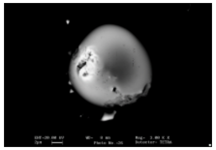
Risk of contamination from other persons (e.g. upon apprehension by the police) or materials (e.g. contact with firearms).

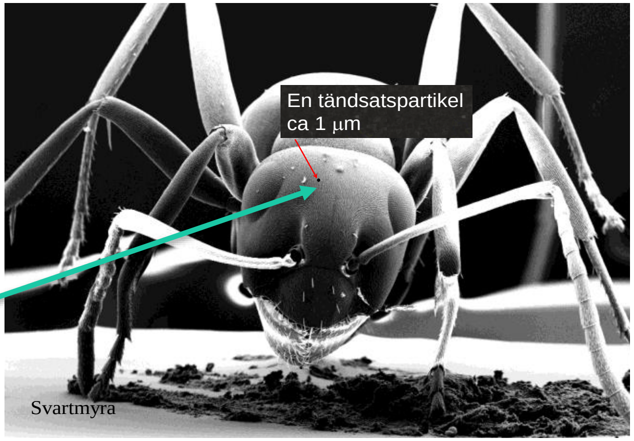
Hence, search for GSR must be done as early as possible after a shooting incident.

GSR are not visible to the human eye.

Size is about 1 µm

They can be observed using Scanning Electron Microscopy (SEM) technique.





GSR have low degree of polymorphism (the way they are analysed today).

Characteristic elemental compositions:

- Type 1 (lead, barium and antimony)
- Type 2 (lead, barium, antimony and tin)
- Type 3 (lead, barium, antimony and aluminium)

Non-characteristic compositions:

- Type 4 (lead, barium, calcium, silicon and tin)
- Type 5 (antimony, tin, potassium and clorine)

Such small variation makes it difficult to attribute GSR to a specific source.

The forensic hypotheses

The main hypothesis:

Since it is not meaningful to try to attribute GSR to a specific source, the main hypothesis can only address a shooting activity. Moreover, since the risk of contamination is high, it is not meaningful to limit the hypothesis to a shooting activity.

$\boldsymbol{H_m}$: The suspect has recently discharged a firearm or been in contact with firearm-related material.

The alternative hypothesis:

$\boldsymbol{H_a}$: The suspect has neither recently discharged a firearm nor been in contact with firearm-related material.

Note that these hypotheses are about activities.

$\boldsymbol{H_m}$: The suspect has recently discharged a firearm or been in contact with firearm-related material.

$\boldsymbol{H_a}$: The suspect has neither recently discharged a firearm nor been in contact with firearm-related material.

The evidence

Assume that **4** GSR were recovered from the taping of the sleeves of the suspect’s jacket ($\boldsymbol{E}$) (recovered using SEM).

Additional information:

The shooting took place around 10 p.m. on April 15.

The weather during the evening and night on April 15 was fair (no precipitation)

The suspect was apprehended about 4 hours after the shooting incident.

$\boldsymbol{H_m}$: The suspect has recently discharged a firearm or been in contact with firearm-related material.
$\boldsymbol{H_a}$: The suspect has neither recently discharged a firearm nor been in contact with firearm-related material.

$\boldsymbol{E}$: 4 recovered GSR from the sleeves of the suspect’s jacket.

Evaluation:

There are no data bases that can assist in eliciting probabilities of the evidence.

$P(\boldsymbol{E}|\boldsymbol{H_h})$: It is expected to recover this amount of GSR if $\boldsymbol{H_h}$ is true given the additional information, hence $P(\boldsymbol{E}|\boldsymbol{H_h}) \approx 1$

$P(\boldsymbol{E}|\boldsymbol{H_a})$: Experience with the expert and studies made gives that if $\boldsymbol{H_a}$ is true, recovering 4 GSR is quite rare. The probability $P(\boldsymbol{E}|\boldsymbol{H_a})$ is in the range 0.01 to 0.1

$\Rightarrow$ The Bayes factor $V = \dfrac{P(\boldsymbol{E}|\boldsymbol{H_h})}{P(\boldsymbol{E}|\boldsymbol{H_a})} \geq \dfrac{1}{0.1} = 10$

The forensic findings are at least 10 times more probable if H_m is true compared to if H_a is true.

What if the suspect says he visited a shooting range that evening?

Continuous data and validation of calculated values of evidence.

In forensic chemistry, most of the data used for evidence evaluation is continuously-valued

Example: Comparison of glass

Typically fragment(s) of glass are recovered from somebody suspected to have broken a glass object (window (burglary), container (assault) etc.).

Forensic hypotheses (at source level):

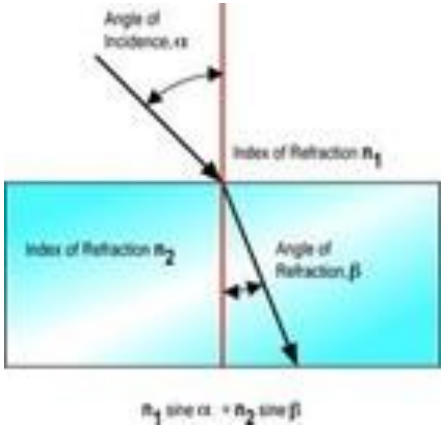
$\boldsymbol{H_m}$: The fragment(s) originate(s) from the broken glass object

$\boldsymbol{H_a}$: The fragment(s) originate(s) from another glass object

$\boldsymbol{H_m}$: The fragment(s) originate(s) from the broken glass object

$\boldsymbol{H_a}$: The fragment(s) originate(s) from another glass object

Using univariate data – measurements of refractive index, RI


Angle of Incidence, α
Index of Refraction n_1
Index of Refraction n_2
Angle of Refraction, β
n_1 sine α = n_2 sine β

Evidence, $\boldsymbol{E}$ (per fragment)

y =Measured RI on recovered fragment

x =Measure RI on broken glass object

How data looks like

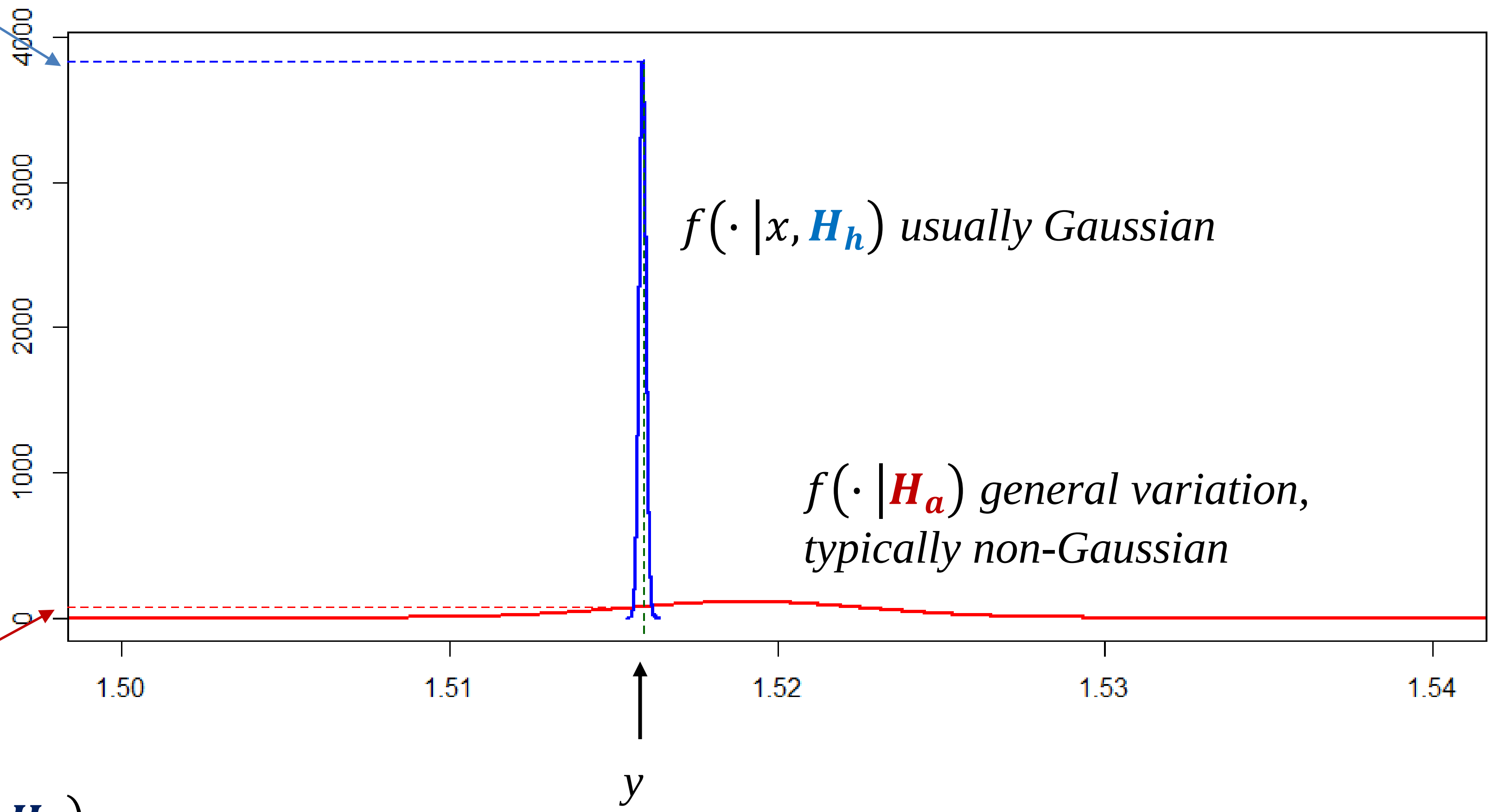
Material	RI
Glass 1	1.51854
Glass 2	1.52289
Glass 3	1.52282
Glass 4	1.52280
Glass 5	1.51625

Bayes factor: $$V = \frac{f(y|x, \boldsymbol{H_h})}{f(y|\boldsymbol{H_a})}$$

$\boldsymbol{H_m}$: The fragment(s) originate(s) from the broken glass object

$\boldsymbol{H_a}$: The fragment(s) originate(s) from another glass object

$f(y|x, \boldsymbol{H_h})$



$f(\cdot\,|x, \boldsymbol{H_h})$ *usually Gaussian*

$f(\cdot\,|\boldsymbol{H_a})$ *general variation, typically non-Gaussian*

$f(y|\boldsymbol{H_a})$

$$V = \frac{f(y|x, \boldsymbol{H_h})}{f(y|\boldsymbol{H_a})}$$

$\boldsymbol{H_m}$: The fragment(s) originate(s) from the broken glass object

$\boldsymbol{H_a}$: The fragment(s) originate(s) from another glass object

Using multivariate data – elemental composition

Weight percentages of element – deduced by *Scanning Electron Microscopy* or *Inductively Coupled Plasma Mass Spectrometry*

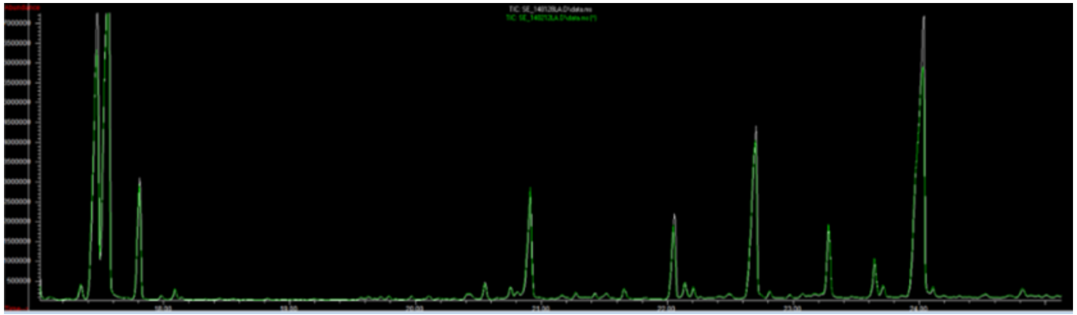
Material	Na	Mg	Al	Si	S	K	Ca	Fe	O
Glass 1	9.28	2.52	0.29	34.68	0.15	0.16	5.65	0.08	47.19
Glass 2	9.27	2.47	0.29	34.70	0.10	0.11	5.72	0.18	47.15
Glass 3	9.22	2.48	0.32	34.65	0.19	0.17	5.71	0.04	47.21
Glass 4	9.32	2.45	0.29	34.66	0.13	0.16	5.80	0.05	47.15
Glass 5	9.33	2.47	0.29	34.72	0.13	0.13	5.70	0.03	47.20

Compositional data (sum to 100%).

Normalise by the weight percent of one element (usually Oxygene (O)) and take natural logarithms.

Example Comparison of seizures of illicit drugs

Gas-chromatographic analysis



Overlaid chromatograms of two amphetamine materials, one in green and one in violet.

The peaks in a chromatogram correspond to specific substances in the material analysed.

Besides the active substance (that makes it a classified drug) a number of impurities are monitored.

These arise in a "random" fashion at or after the stage of manufacturing/preparation – *chemical fingerprint*.

Example of analytical data for precipitated amphetamine powder:

			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
Manufcaturing batch	Sample Multiplier	Inner standard	Ketoxime 1	Ketoxime 2	4-Methyl-5-phenylpyrimidine	Unknown C	N-Benzylpyrimidine	N-Acetylamphetamine	N-Formylamphetamine	1,2-Diphenyletylamine	N,N-Dibenzylamine	1,2-Diphenylethanone	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	96246.09	14116.99	0	43932.69	4136092	0	299072632.2	0	975972.4	2893586	1996587	1229466	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

The forensic hypotheses for comparing two seizures of a drug:

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

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

Case data (generic format):

$$\boldsymbol{E}_1 = \boldsymbol{y}_1 = \begin{pmatrix} y_{1,1,1} & y_{1,1,2} & \cdots & y_{1,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,m_1,1} & y_{1,m_1,2} & \cdots & y_{1,m_1,p} \end{pmatrix}$$

m_1 replicate analyses ($n_1 \times p$ peak areas) on material 1

$$\boldsymbol{E}_2 = \boldsymbol{y}_2 = \begin{pmatrix} y_{1,1,1} & y_{1,1,2} & \cdots & y_{1,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,m_2,1} & y_{1,m_2,2} & \cdots & y_{1,m_2,p} \end{pmatrix}$$

m_2 replicate analyses ($n_2 \times p$ peak areas) on material 2

Numbers of replicate analyses are usually very small (1, 2 or 3).

How to use such data to obtain a Bayes factor, *V* ?

1. Feature-based evaluation

$$\boldsymbol{E_1} = \boldsymbol{y}_1 = \begin{pmatrix} y_{1,1,1} & y_{1,1,2} & \cdots & y_{1,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,m_1,1} & y_{1,m,2} & \cdots & y_{1,m_1,p} \end{pmatrix}$$

$$\boldsymbol{E_2} = \boldsymbol{y}_2 = \begin{pmatrix} y_{1,1,1} & y_{1,1,2} & \cdots & y_{1,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,m_2,1} & y_{1,m_2,2} & \cdots & y_{1,m_2,p} \end{pmatrix}$$

Model the probability distributions of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$.

Normally distributed data ⇒ sufficient to model the distributions of $\overline{\boldsymbol{y}}_1$ and $\overline{\boldsymbol{y}}_2$.

Always strong attempts from chemists to transform their data to be Gaussian.

The following probability densities will be involved:

$f(\overline{\boldsymbol{y}}_1|\boldsymbol{\theta}), f(\overline{\boldsymbol{y}}_2|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the unknown mean vector of the peak areas

$g(\boldsymbol{\theta})$ the (prior) distribution of $\boldsymbol{\theta}$ – empirically deduced

The Bayes factor is then

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

(Lindley, Biometrika, 1977):

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

Learning density functions from multivariate distributions is always a challenge. Even if data shows Gaussian behaviour, the covariance structures needs a lot of data to be accurately estimated.

Training data with known ground truth: Usually limited: "n" > p, but not sufficiently larger.

Dimension reduction?

Principal components?

Removal of "unimportant" dimensions?

Dimension reduction via graphical modelling

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

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

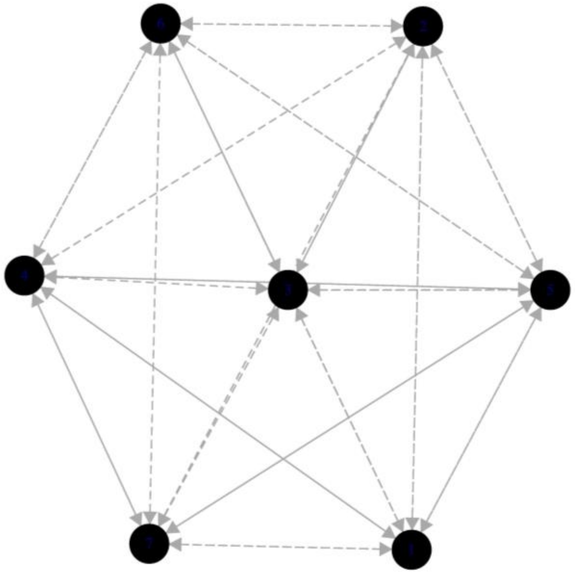
The partial correlation matrix is then $\boldsymbol{P} = \left(p_{ij}\right)$ where $p_{ij} = \frac{-q_{ij}}{\sqrt{q_{ii} \cdot q_{jj}}}$

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.

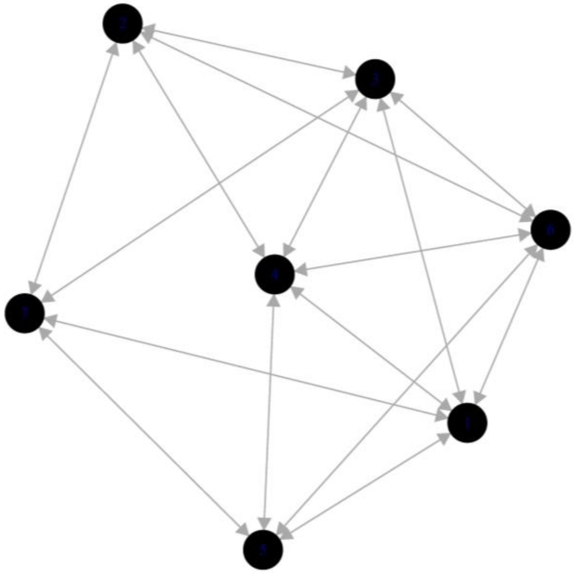
A graphical model of a random vector can be defined as a graphical model where the links (edges) between two components exist provided their partial correlation exceeds a chosen threshold.

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

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

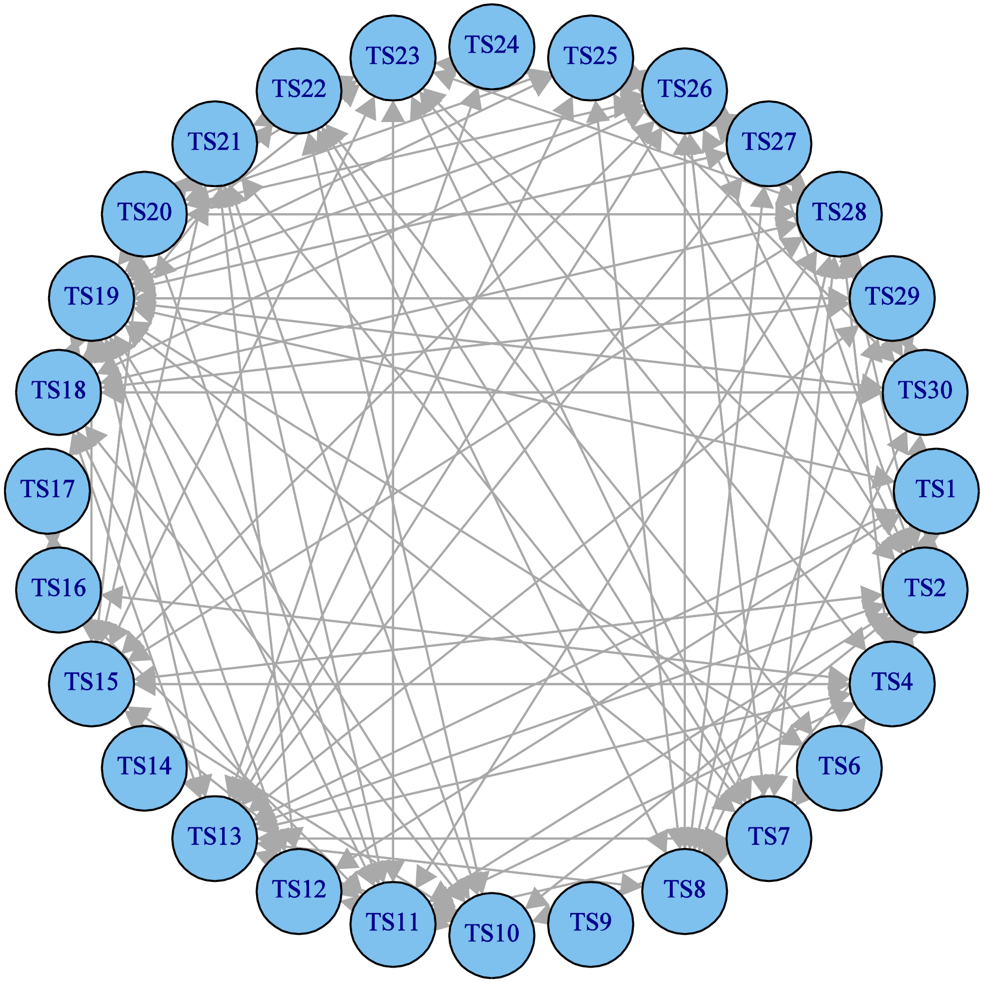


Reduced model ($p_{ij} > 0.5$):



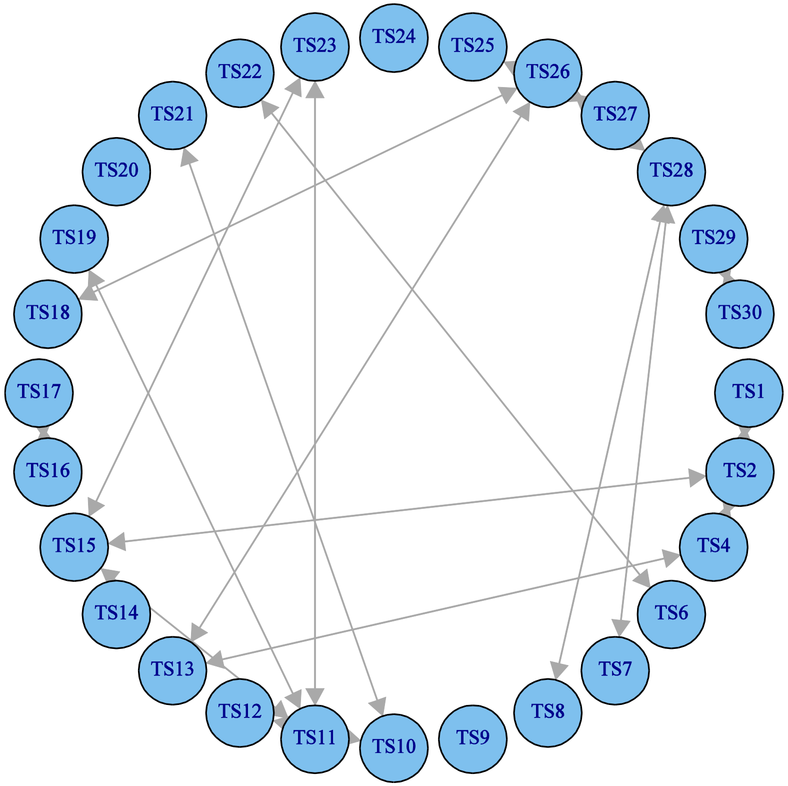
Example: For training data with amphetamine impurities we name the impurities TS1, TS2, …, TS30 (Target Substance)

A graphical model based on partial correlations ≥ 0.2 becomes

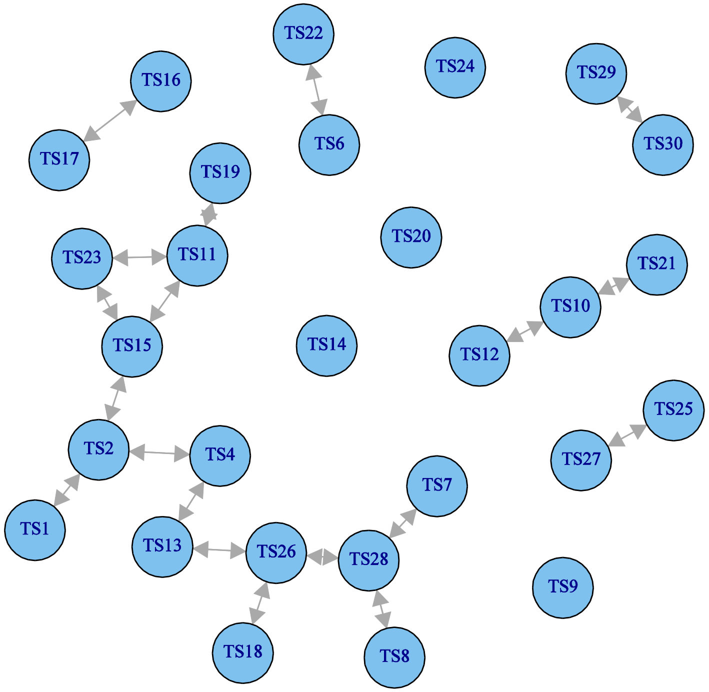


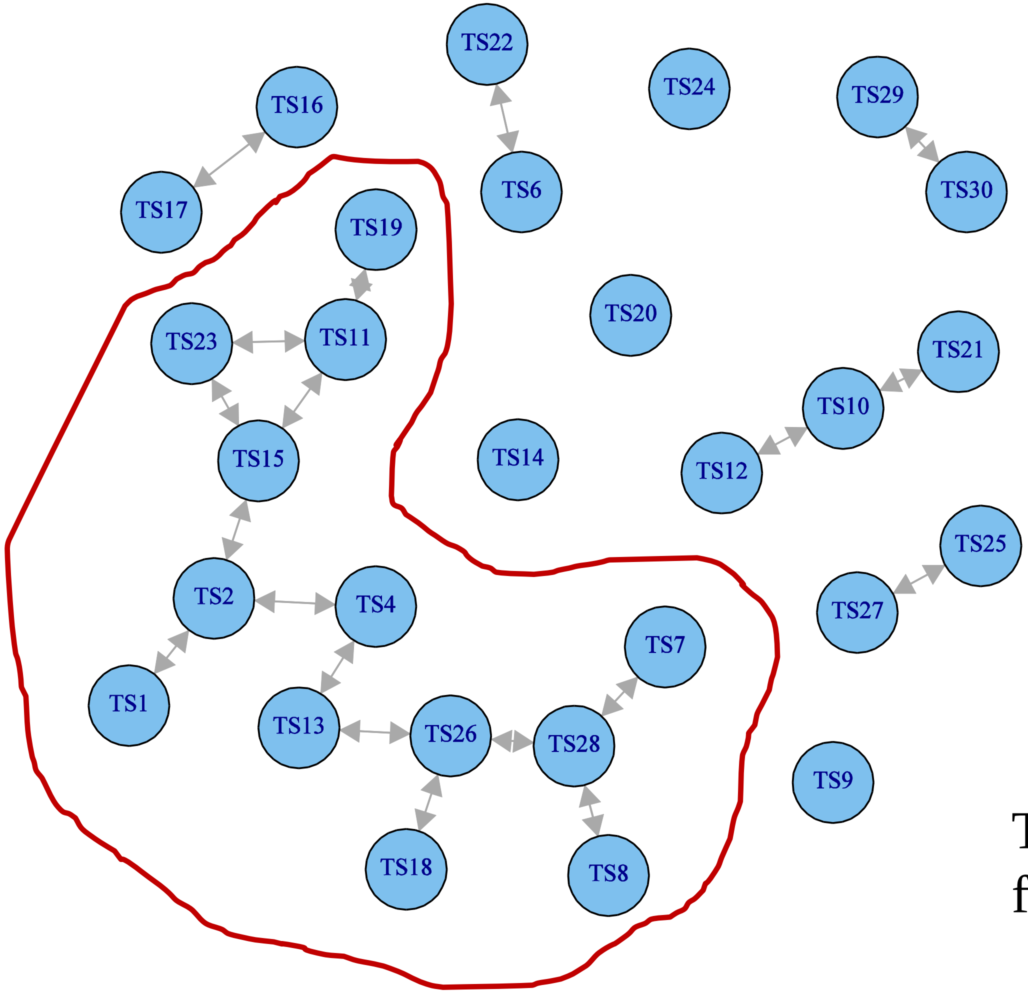
Chemical considerations about the substances gives that 28 of the 30 impurities should be retained (TS3 and TS5 are taken out).

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.

Thus, we have reduced the dimension from 28 to 13.

The Bayes factor may then be factorized into 10 factors:

$$V = V_1 \cdot V_2 \cdot V_3 \cdot V_4 \cdot V_5 \cdot V_6 \cdot V_7 \cdot V_8 \cdot V_9 \cdot V_{10}$$

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

1. Score-based evaluation

Instead of modelling the data from the two seizures, we can compare the data and use a measure of distance or similarity them between.

Examples:

- Euclidean distance

$$D(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2) = \sqrt{\sum_j \left(\bar{y}_{1\cdot j} - \bar{y}_{2\cdot j}\right)^2}$$

- City-block distance

$$D(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2) = \sum_j \left|\bar{y}_{1\cdot j} - \bar{y}_{2\cdot j}\right|$$

- Canberra distance

$$D(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2) = \sum_j \frac{\left|\bar{y}_{1\cdot j} - \bar{y}_{2\cdot j}\right|}{\left|\bar{y}_{1\cdot j}\right| + \left|\bar{y}_{2\cdot j}\right|}$$

- Pearson correlation “distance”

$$D(\bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2) = 1 - \frac{\sum_j \left(\bar{y}_{1\cdot j} - \bar{y}_{1\cdot\cdot}\right)\left(\bar{y}_{2\cdot j} - \bar{y}_{2\cdot\cdot}\right)}{\sqrt{\sum_j \left(\bar{y}_{1\cdot j} - \bar{y}_{1\cdot\cdot}\right)^2 \cdot \sum_j \left(\bar{y}_{1\cdot j} - \bar{y}_{1\cdot\cdot}\right)^2}}$$

From a score to a Bayes factor

Training data:

- N_1 pairs of materials with the same origin
- N_2 pairs of materials with different origins

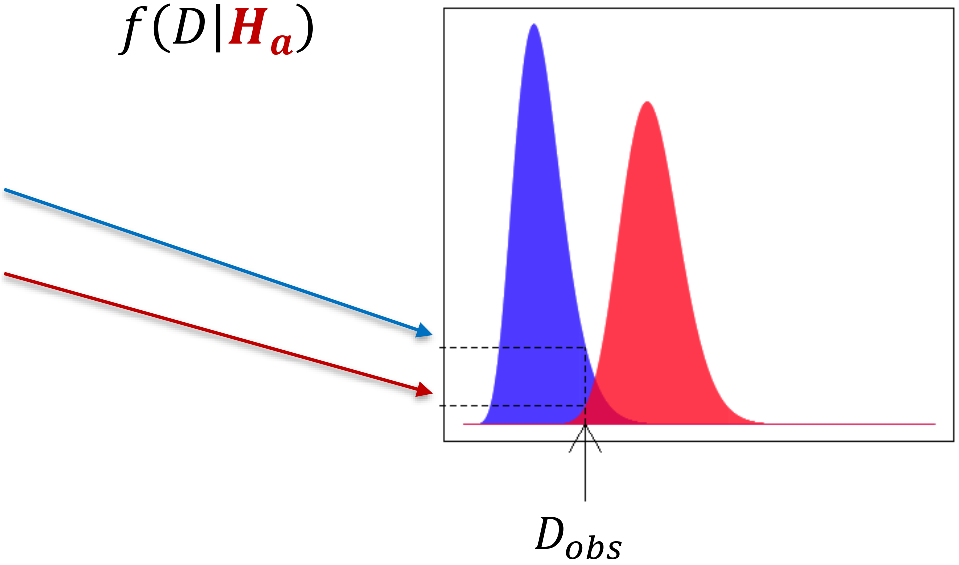
Fit the distribution of D for the pairs with same origin

$\Rightarrow$

Score density for same origin: $f(D|\boldsymbol{H}_{\boldsymbol{m}})$

Score density for different origins: $f(D|\boldsymbol{H}_{\boldsymbol{a}})$

Bayes factor: $$V = \frac{f(D_{obs}|\boldsymbol{H}_{\boldsymbol{m}})}{f(D_{obs}|\boldsymbol{H}_{\boldsymbol{a}})}$$



Are the methods of finding Bayes factors valid?

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

	$\boldsymbol{H_h}$ true	$\boldsymbol{H_a}$ true
$V > 1$	Basically valid	Not valid
$V < 1$	Not valid	Basically valid

But is it sufficient with V not giving support in the wrong direction?

When do we expect V to reflect strong and weak evidence for a hypothesis?

Validation using Empirical Cross-Entropy (ECE)

Entropy of a random variable, X: $H(X) = -\mathbb{E}\{\log\big(f(X)\big)\}$

$\mathbb{E}$ is the expectation operator

Classical *Shannon entropy* for finite discrete probability distribution: $H = -\sum_1^N p_i \cdot \log_2(p_i)$

Cross-entropy between two probability distributions with the same support:

$$H(X,Y) = -\mathbb{E}_X\{\log\big(f_Y(X)\big)\}$$

Validation data set (for assessing Bayes factors for comparisons)

S_m = Data from comparisons of samples with common origin
N_m = Number of comparisons of samples with common origin
S_a = Data from comparisons of samples with different origins
N_a = Number of comparisons of samples with different origins

$\mathcal{X} = (\boldsymbol{H_m}, \boldsymbol{H_a})$ can be seen as a bivariate random variable (usually with probability distribution $(p, 1-p)$)

$\mathcal{Y} = (\boldsymbol{H_m}|\boldsymbol{E}, \boldsymbol{H_a}|\boldsymbol{E})$ is another bivariate random variable with the same support as $\mathcal{X}$ (and analogously with probability distribution $(q, 1-q)$)

It can be shown that the expected entropy of $\mathcal{Y}$ over all possible instances of $\boldsymbol{E}$ cannot be lower than the entropy of $\mathcal{X}$.

S_m = Data from comparisons of samples with common origin
N_m = Number of comparisons of samples with common origin
S_a = Data from comparisons of samples with different origins
N_a = Number of comparisons of samples with different origins

Empirical Cross-Entropy:

$$ECE = -\sum_{i\in S_m} \log_2 P(H_m|E_i) \cdot \frac{P(H_m)}{N_m} - \sum_{j\in S_a} \log_2 P\left(H_a\middle|E_j\right) \cdot \frac{P(H_a)}{N_a}$$

$$= -\sum_{i\in S_m} \log_2\left(\frac{V_i \cdot \frac{P(H_m)}{P(H_a)}}{1 + V_i \cdot \frac{P(H_m)}{P(H_a)}}\right) \cdot \frac{P(H_m)}{N_m} - \sum_{j\in S_a} \log_2\left(\frac{1}{1 + V_j \cdot \frac{P(H_m)}{P(H_a)}}\right) \cdot \frac{P(H_a)}{N_a}$$

$$= -\sum_{i\in S_m} \log_2\left(\frac{1}{1 + \frac{1}{V_i \cdot \frac{P(H_m)}{P(H_a)}}}\right) \cdot \frac{P(H_m)}{N_m} - \sum_{j\in S_a} \log_2\left(\frac{1}{1 + V_j \cdot \frac{P(H_m)}{P(H_a)}}\right) \cdot \frac{P(H_a)}{N_a}$$

$$= \sum_{i\in S_m} \log_2\left(1 + \frac{1}{V_i \cdot \frac{P(H_m)}{P(H_a)}}\right) \cdot \frac{P(H_m)}{N_m} + \sum_{j\in S_a} \log_2\left(1 + V_j \cdot \frac{P(H_m)}{P(H_a)}\right) \cdot \frac{P(H_a)}{N_a}$$

The *ECE* plot

S_m = Data from comparisons of samples with common origin
N_m = Number of comparisons of samples with common origin
S_a = Data from comparisons of samples with different origins
N_a = Number of comparisons of samples with different origins

$$ECE = \sum_{i \in S_m} \log_2 \left(1 + \frac{1}{V_i \cdot \frac{P(H_m)}{P(H_a)}} \right) \cdot \frac{P(\boldsymbol{H_m})}{N_m}$$
$$+ \sum_{j \in S_a} \log_2 \left(1 + V_j \cdot \frac{P(H_m)}{P(H_a)} \right) \cdot \frac{P(\boldsymbol{H_a})}{N_a}$$

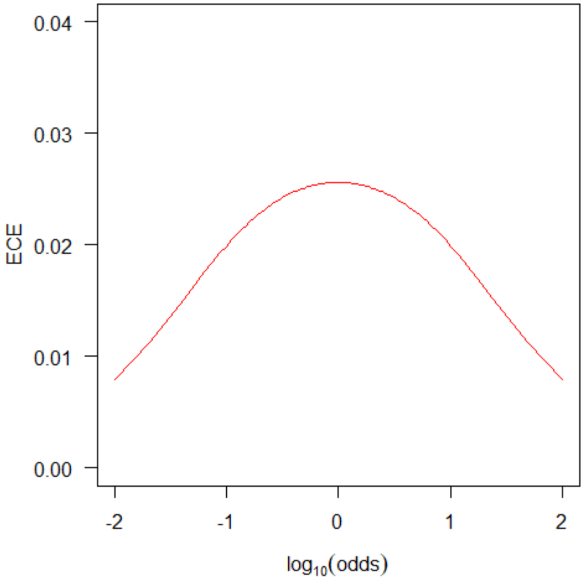
The Bayes factors V_i, V_j are calculated in S_m and S_a respectively, but the *ECE* depends on the prior odds $\frac{P(H_m)}{P(H_a)}$ (and/or the prior probability $P(H_m) = 1 - P(H_a)$)

The validity of the set of Bayes factors can therefore be assessed by plotting *ECE* against the prior odds.

The entropy of $\mathcal{X} = (\boldsymbol{H_m}, \boldsymbol{H_a})$ is as highest when the prior odds are 1, and thus the *ECE* should reach its maximum at that point with basically valid Bayes factors.

The further from 1 the prior odds are the lower the cross-entropy should be.

Example:



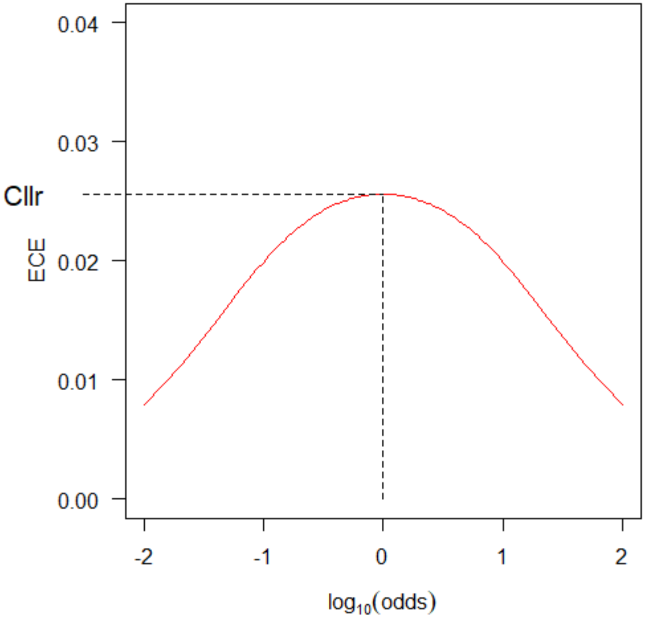
Symmetric shape around prior odds=1 (i.e. $\log_{10}(odds) = 0$)

$\Rightarrow$ Basically valid, but how good are the Bayes factors?

Measure of performance:

$$C_{llr} = ECE(\text{prior odds} = 1)$$

"Cost of log-likelihood ratio"

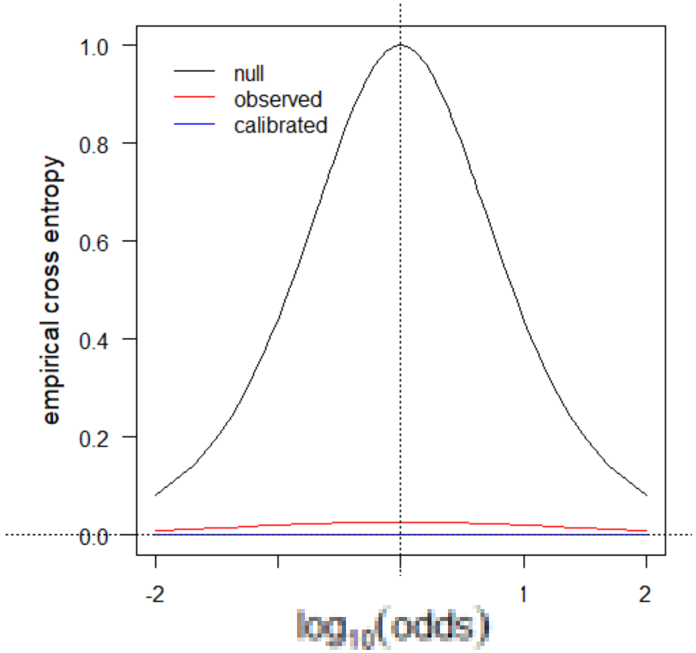


Can be used to compare different methods of calculating Bayes factors.

The *ECE* curve can be compared to a curve constructed such that all Bayes factors are equal to 1 (all-over neutral evidence).

If the *ECE* curve stretches above the neutral curve this mean that one would do worse using Bayes factors calculated with the assessed method than to just base decisions on the prior odds.

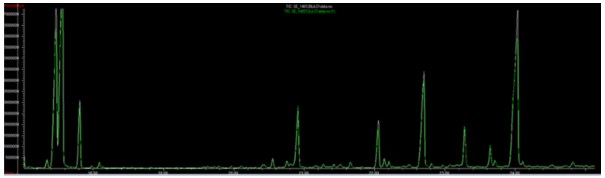
Moreover, since the ground truth is known for the validation set it is possible to calibrate the calculated Bayes factors using the PAV algorithm to values that are the best that could be reached (with this validation set). The corresponding curve is thus the optimal *ECE* curve.

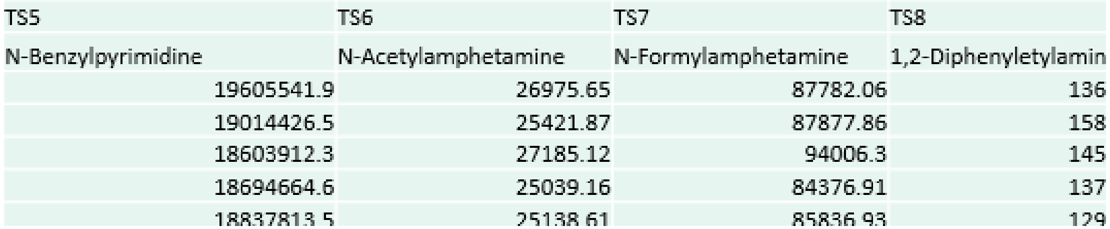


This shows that the method of calculating Bayes factors is very good. The red curve is close to the optimal blue curve, and far away from the neutral (null) black curve.

Example: Back to the comparison of amphetamine seízures

Trying to calculated a feature-based Bayes factor from 12 of the 30 impurities monitored





TS5	TS6	TS7	TS8
N-Benzylpyrimidine	N-Acetylamphetamine	N-Formylamphetamine	1,2-Diphenyletylamin
19605541.9	26975.65	87782.06	136
19014426.5	25421.87	87877.86	158
18603912.3	27185.12	94006.3	145
18694664.6	25039.16	84376.91	137
18837813.5	25138.61	85836.93	129

The Bayes factor $V = \dfrac{\int f(\bar{\boldsymbol{y}}_1|\boldsymbol{\theta}) \cdot f(\bar{\boldsymbol{y}}_2|\boldsymbol{\theta}) \cdot g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\bar{\boldsymbol{y}}_1|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \times \int f(\bar{\boldsymbol{y}}_2|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}$

can be approximated by replacing the prior distributions of means and covariances with estimates from the training set and using normal distributions for $f(\bar{\boldsymbol{y}}_1|\boldsymbol{\theta})$ and $f(\bar{\boldsymbol{y}}_2|\boldsymbol{\theta})$ and a multivariate kernel density (Gaussian kernel) for $g(\boldsymbol{\theta})$.

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

with

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

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

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

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

where

$\boldsymbol{U}$ = within-material covariance matrix

$\boldsymbol{C}$ = between-material covariance matrix

$\bar{\boldsymbol{x}}_i$ = mean vector of peak areas of the replicate analyses from material *i* in training set

$$\boldsymbol{y}^* = \left[\left(\frac{\boldsymbol{U}}{n_1}\right)^{-1} + \left(\frac{\boldsymbol{U}}{n_2}\right)^{-1} \right]^{-1} \left(\left(\frac{\boldsymbol{U}}{n_1}\right)^{-1} \bar{\boldsymbol{y}}_1 + \left(\frac{\boldsymbol{U}}{n_2}\right)^{-1} \bar{\boldsymbol{y}}_2 \right)$$

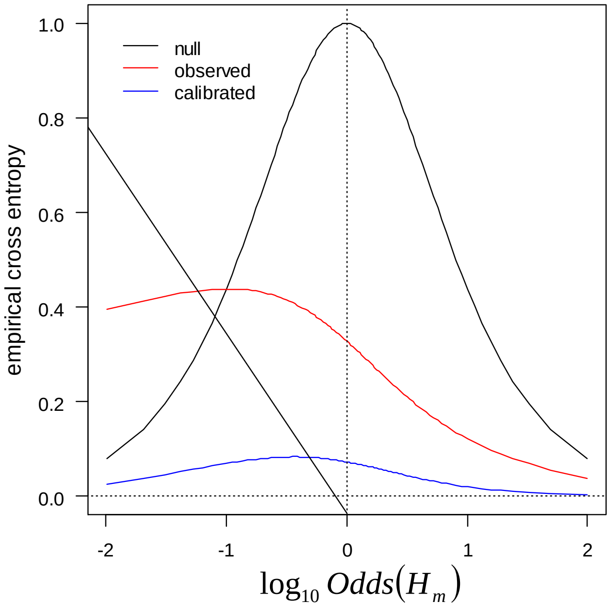
h = bandwidth of kernel density estimate

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

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

(Aitken & Lucy, *JRSS* **C,** 2004)

ECE plot:



Not so good!!