Discovering aroma patterns in food products using Latent Dirichlet Allocation and Jensen Shannon divergence

Michael Fitzke and ALICIA OLIVARES
Mars Petcare. Eitzer Str. 215, 27283 Verden, Germany.

Abstract

Aroma Extract Dilution Analysis (AEDA) evaluates volatile compounds most likely contributing to the overall aroma of a food sample by means of flavour dilution (FD) factors. In the food industry, this can be useful to compare aroma-active profiles of raw materials or finished products and to select those that are statistically similar. When multiple samples are analysed, the high number of variables makes it difficult to take conclusions. Principal Component Analysis (PCA) should not be applied to FD values as they are discrete numbers. To our knowledge, there are no appropriate methods available to interpret AEDA results from multiple samples. In this study, a new rapid methodology to interpret AEDA results was developed. Latent Dirichlet Allocation (LDA) was developed in the context of text analysis as a mean of dimensionality reduction and has been successfully applied for the analysis of AEDA outcomes. Furthermore, Jensen Shannon divergence measure was a useful tool to compare the distribution of volatile compounds with similar descriptions ("berries", "cheese" or "fruits") among different samples.

Introduction

Gas chromatography-olfactometry (GC-O) is used to judge the sensory relevance of the volatiles present in foods. In particular, AEDA evaluates the odour activities of the volatiles by sniffing the effluent of a series of dilutions of the original aroma extract. The result is expressed as the flavour dilution (FD) factor that corresponds to the maximum dilution value detected. Compounds with the highest FD are assumed to be most likely contributing to the overall aroma of a food product. AEDA is a time-consuming technique and generally research articles report the analysis of 1-3 samples where it is fairly easy to see differences. However, when multiple samples are analysed, the interpretation of AEDA results becomes challenging. This is because AEDA data set is fairly high-dimensional but sparse and it is difficult to conclude similarity among samples. A common approach in situations like this is to map the data into an adequate lower dimensional sub space where the comparison and clustering is done. When the data is normally distributed, PCA is often used. However, PCA should not be applied to AEDA because the data are discrete. This may be the reason why in other works, the statistical interpretation of AEDA has been claimed to be controversial or even not applicable [1, 2] although the authors did specify the reasons.

Latent Dirichlet Allocation (LDA) was developed in the context of text analysis as a means of dimensionality reduction [3]. For example, LDA can be used to cluster documents where instead of cluster them word by word, they can be clustered by topic (a topic would be described by a distribution over words). In probability theory and statistics, the Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions [4].

The aim of this work was to develop a rapid methodology using LDA and Jensen-Shannon divergence to interpret AEDA results from multiple samples. In particular, the method was used to investigate the similarities in the aroma profile of pet foods.

B. Siegmund & E. Leitner (Eds): Flavour Sci., 2018, Verlag der Technischen Universität Graz
DOI: 10.3217/978-3-85125-593-5-68, CC BY-NC-ND 4.0

Materials and methods

Samples

8 pet foods samples from different brands and varieties were used in this study. 20 g of each sample were suspended with 20 mL H_2O and extracted with 100 mL diethyl ether (distilled before use). The organic layer was separated from the residue and the volatiles were isolated via Solvent Assisted Flavour Evaporation. The distillate was dried over sodium sulphate and concentrated to 200 µL using a Vigreux column.

GC-O analysis

High resolution gas chromatography was performed by means of a Trace GC (Finnigan, Bremen) and a column FFAP (30 m x 0.25 mm x 0.25 µm, J&W Scientific). The samples (1 µl) were injected using "on column" injection technique at 40 °C. After 1 min, the temperature was raised 6 °C/min until 240 °C were reached. The flow rate of the carrier gas (helium) was set on 1.5 mL/min. At the end of the capillary, the effluent was split 1:1 into a flame ionization detector (FID) and a sniffing port by using two deactivated, uncoated fused silica capillaries (20 cm × 0.25 mm). The FID and sniffing port were held at 250 °C. Linear retention indices (LRI) were calculated by the equation given by Kovats. The volatile fraction was diluted stepwise 1+1 with solvent and each dilution step was sniffed until no odourant in the effluent was perceived. The odour extract dilution analysis was performed by two trained panellists. FD factors were expressed in logarithmic scale units.

Statistical analysis

LDA was used to model aroma profiles as random mixtures over latent topics, each topic was characterized as a distribution over aroma compounds and was interpreted as a basic aroma profile.

The following generative process was assumed for each product aroma profile I_n:

1. Choose N ~ *Poisson* (ξ) as the sum of all logarithmized FD-factors. in I_n
2. Choose Θ ~ *Dirichlet* (α)
3. For each of the N:
   (a) Choose topic Z_n ~*Multinomial* (Θ)
   (b) Choose a DF from the aroma compounds from $p(I_n|Z_n,\beta)$, a multinomial probability conditioned on the topic Z_n

Model fitting and inference based on this process was done by Variational Bayes.

To determine the similarity of the aroma profiles of two products, to use information-theoretically motivated measure of distance of two probability distributions $\boldsymbol{P}$ and $\boldsymbol{Q}$ like the Kullback-Leibler divergence $\boldsymbol{D_{KL}(P||Q)} = \sum_i \boldsymbol{P(i)} \cdot \boldsymbol{\log \frac{P(i)}{Q(i)}}$ is appropriate.

Jensen-Shannon Divergence is the symmetric version of Kullback-Leibler divergence and was used a distance metric to describe distances between products, as follows:

$$\boldsymbol{JSD(P||Q)} = \frac{1}{2}\boldsymbol{D_{KL}(P||M)} + \frac{1}{2}\boldsymbol{D_{KL}(Q||M)}$$
$$\textit{Where } \boldsymbol{M} = \tfrac{1}{2}(\boldsymbol{P} + \boldsymbol{Q}).$$

Results and discussion

A total of 77 odour-active compounds was detected in the samples although 10 of them could not be identified (Table 1). The 67 identified compounds include 11 alcohols, 10 aldehydes, 10 acids, 8 ketones, 7 sulphur compounds, 4 esters, 4 pyrazines, 4 lactones,

3 hydrocarbons, 2 pyrrolines, 2 furans and 2 nitrogen compounds. Not all of the flavour active compounds were present in all the samples and for those present in all the samples, the FD values were different in many cases. From the FD factors it was not obvious if samples were statistically different to each other (Figure 1).

Table 1: Volatile compounds in the pet food samples and their odour description.

Compound/chemical class	Odour descriptor	LRI FFAP	Compound/chemical class	Odour descriptor	LRI FFAP
Ketones			*Alcohols*		
2,3-butanedione	butter	967	linalool	floral 1	1529
3-mercapto-2-butanone	catty, blackcurrant	1267	geraniol	rose	1839
1-octen-3-one	mushroom	1294	2-methoxyphenol	smoky	1857
3-mercapto-2-pentanone	catty	1356	2-phenylethanol	honey 1	1900
(Z)-1,5-octadien-3-one	geranium	1367	maltol	caramel 2	1957
3-methyl-2,4-nonandione	minty 2	1706	4-ethyl-2-methoxyphenol	clove 1	2014
β-damascenone	apple	1807	4-methylphenol	barnyard	2083
β-ionone	violet	1920	eugenol	clove 2	2162
Aldehydes			3-/4-ethylphenol	leather	2169
2-/3-methylbutanal	malty	911	2,6-dimethoxyphenol	smoky, clove	2258
hexanal	grassy	1077	isoeugenol	clove 3	2333
(Z)-4-heptenal	fishy	1233	*Pyrrolines*		
octanal	citrus	1289	2-acetyl-1-pyrroline	roasty 1	1328
(E,Z)-2,6-nonadienal	cucumber	1582	2-propionyl-1-pyrroline	roasty 2	1406
phenylacetaldehyde	floral 2	1625	*Terpenes and hydrocarbons*		
(E,E)-2,4-nonadienal	fatty 1	1688	á-pinene	resinous	1007
(E,E)-2,4-decadienal	fatty 2	1800	(E,Z)-1,3,5-undecatriene	pineapple	1378
(E,E,Z)-2,4-6-nonatrienal	oatflakes 1	1860	vanillin	vanilla	2560
tr.-4,5-epoxy-(E)-2-decenal	metallic	1986	*Esters*		
Acids			ethyl-2-methylbutanoate	fruity 1	1038
acetic acid	vinegar	1433	methylhexanoate	fruity 2	1174
propanoic acid	cheese 1	1511	ethyl-3-phenylpropanoate	cinnamon 1	1867
2-methylpropanoic acid	cheese 2	1553	ethylcinnamate	cinnamon 2	2113
butanoic acid	cheese 3	1606	*Nitrogen compounds*		
2-/3-methylbutanoic acid	cheese 4	1656	indol	mothballs 1	2440
pentanoic acid	cheese 5	1724	3-methylindol	mothballs 2	2480
3-/4-methylpentanoic acid	cheese 6	1781	*Terpenes and hydrocarbons*		
hexanoic acid	goat 1	1833	á-pinene	resinous	1007
phenylacetic acid	honey 2	2530	(E,Z)-1,3,5-undecatriene	pineapple	1378
phenylpropionic acid	goat 2	>2600	vanillin	vanilla	2560
Sulfur compounds			*Lactones*		
3-methyl-2-buten-1-thiol	beer	1107	γ-octalactone	coconut	1906
dimethyltrisulfide	cabbage 1	1370	sotolon	seasoning 1	2185
2-fufurylthiol	burnt	1418	δ-dodecalactone	peach	2383
methional	cooked potato	1444	3-hydroxy-2(2H)-pyranone	meaty	1953
benzenemethanthiol	cress, burnt	1616	*Unknowns*		
dimethyltetrasulfide	cabbage 2	1713	unknown 1	sulphurous	1150
2-acetyl-2-thiazolin	roasty 3	1744	unknown 2	caramel 1	1415
Pyrazines			unknown 3	minty 1	1555
2,3,5-trimethylpyrazine	earthy 1	1400	unknown 4	catty, rhubarb	1933
2-ethyl-3,5-dimethylpyrazine	earthy 2	1450	unknown 5	oatflakes 2	1975
2,3-diethyl-5-methylpyrazine	earthy 3	1478	unknown 6	sour	2029
2-vinyl-3,5-dimethylpyrazine	earthy 4	1542	unknown 7	minty 3	2079
Furans			unknown 8	fatty 3	2150
furaneol	caramel 3	2017	unknown 9	foxy	2208
abhexone	seasoning 2	2246	unknown 10	chemical	2300

LDA was used to reduce the dimensions by clustering the odour descriptors into “aroma topics”. The 77 odour-active compounds were narrowed down to 3 aroma topics, each aroma topic being a distribution of odour-active compounds as shown in Figure 2. Aroma topic 1 was mainly defined by compounds having sweet, roasted notes, Aroma Topic 2 by spicy, fruity floral notes and Aroma Topic 3 by stable, fatty and cheese notes.

In Figure 3, the aroma topics per sample are shown. As it can be seen the aroma topic 1 was common to all the samples. It could be argued that it contains the basic flavour active compounds for pet foods. The presence of aroma topics 2 and 3 varied among the samples contributing to the specific notes. It was observed that products 2, 4 and 6 had similar flavour active profiles, as well as products 7 and 8.

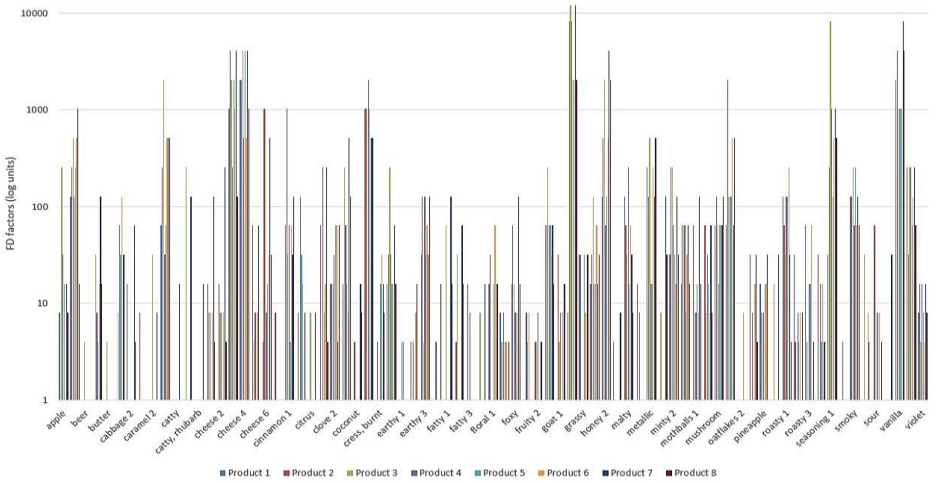

Figure 1: FD factors for the 8 samples analysed and the corresponding descriptors identified for each of the flavour-active compounds.

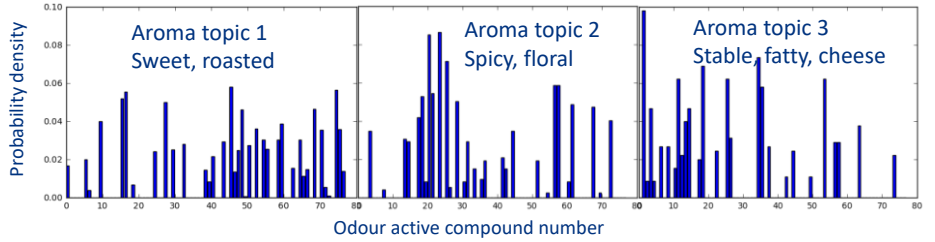

Figure 2: Aroma topics obtained by LDA. Bars represent the distribution of each odour-active compound.

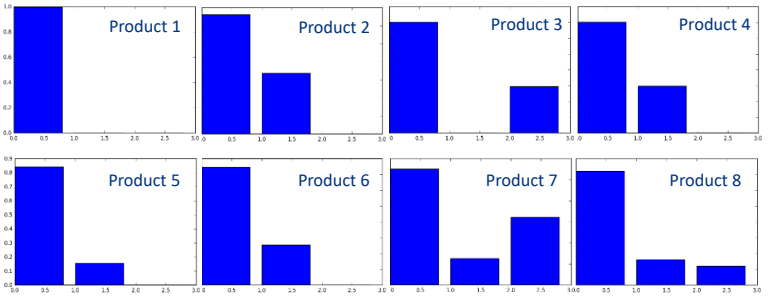

Figure 3: Aroma topics 1, 2 and 3 in the samples (Left, centre and right columns respectively).

The developed method was successfully applied to pet food and could be a useful tool for the food and flavour industry to select raw materials with similar aroma profiles. The correlation between this method and the traditional quantification of compounds could be explored in the future.

References

1. Zellner B., Dugo P., Dugo G., Mondello L. (2008). J.Chromatogr. A, 1186, 123–143.
2. Chin S., Marriott P. J. (2015). Anal. Chim. Acta 854 1–12.
3. Blei D.M., Ng A.Y., Jordan M. I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993-1022.
4. Lin J. (1991). Divergence measures based on the Shannon entropy. IEEE Trans. Inf. theory 37, 1, 145-151.