Moscow Institute of Physics and Technology
(State University)

Keyphrase Extraction

Anton Karazeev
Group 493

2018

Outline

- TF-IDF
- PageRank (TextRank)
- Closed KL and Variational KL

TF-IDF

TF-IDF

$$tf(t,d) = \frac{n_t}{\sum_k n_k}$$

TF-IDF

$$tf(t,d) = \frac{n_t}{\sum_k n_k}$$

$$idf(t,D) = log\frac{|D|}{|\{d_i \in D | t \in d_i\}|}$$

TF-IDF

$$tf(t,d) = \frac{n_t}{\sum_k n_k}$$

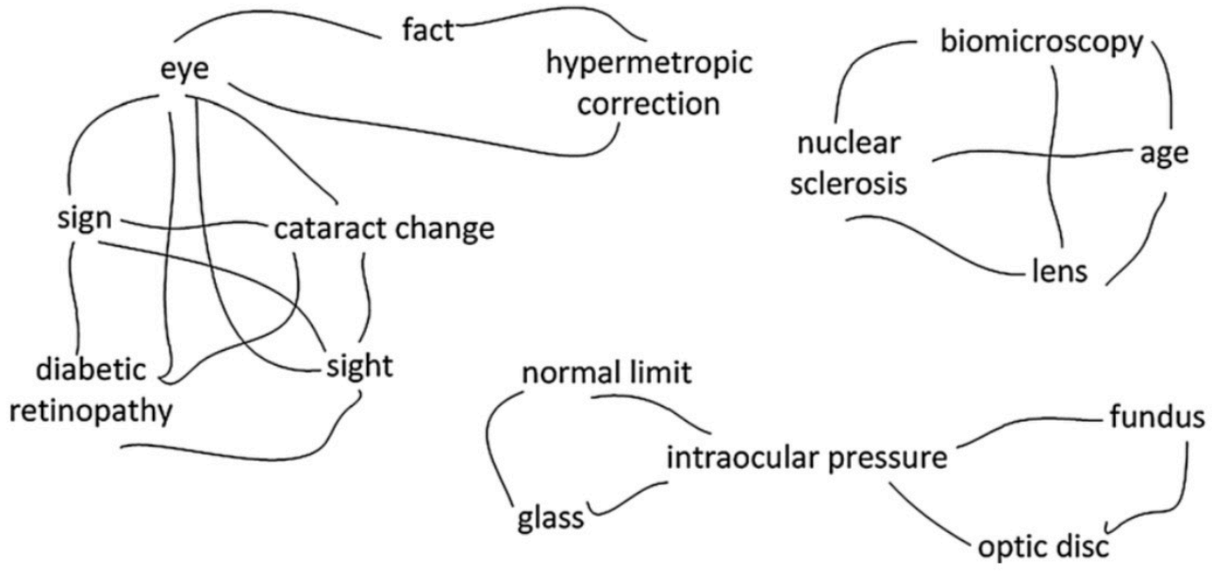
$$idf(t,D) = log\frac{|D|}{|\{d_i \in D | t \in d_i\}|}$$

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$$

TextRank

$$WS(v_i) = (1-d) + d \times \sum_{v_j \in in(v_i)} \frac{w_{ji}}{\sum_{v_k \in out(v_j)} w_{jk}} WS(v_j)$$

Letter 1
He does in fact achieve barely 6/12 unaided, but this improves to 6/6 in each eye separately with a hypermetropic correction. Biomicroscopy showed some nuclear sclerosis in the lens which are quite clear for his age. His intraocular pressures were normal and optic discs and fundi appeared healthy.
Letter 2
Fortunately he still shows no sign of diabetic retinopathy, but is starting to show cataract changes in both eyes even thought this has not affected his sight adversely. His own glasses gave him right 6/9+ left 6/6 and his intraocular pressures were well within normal limits.



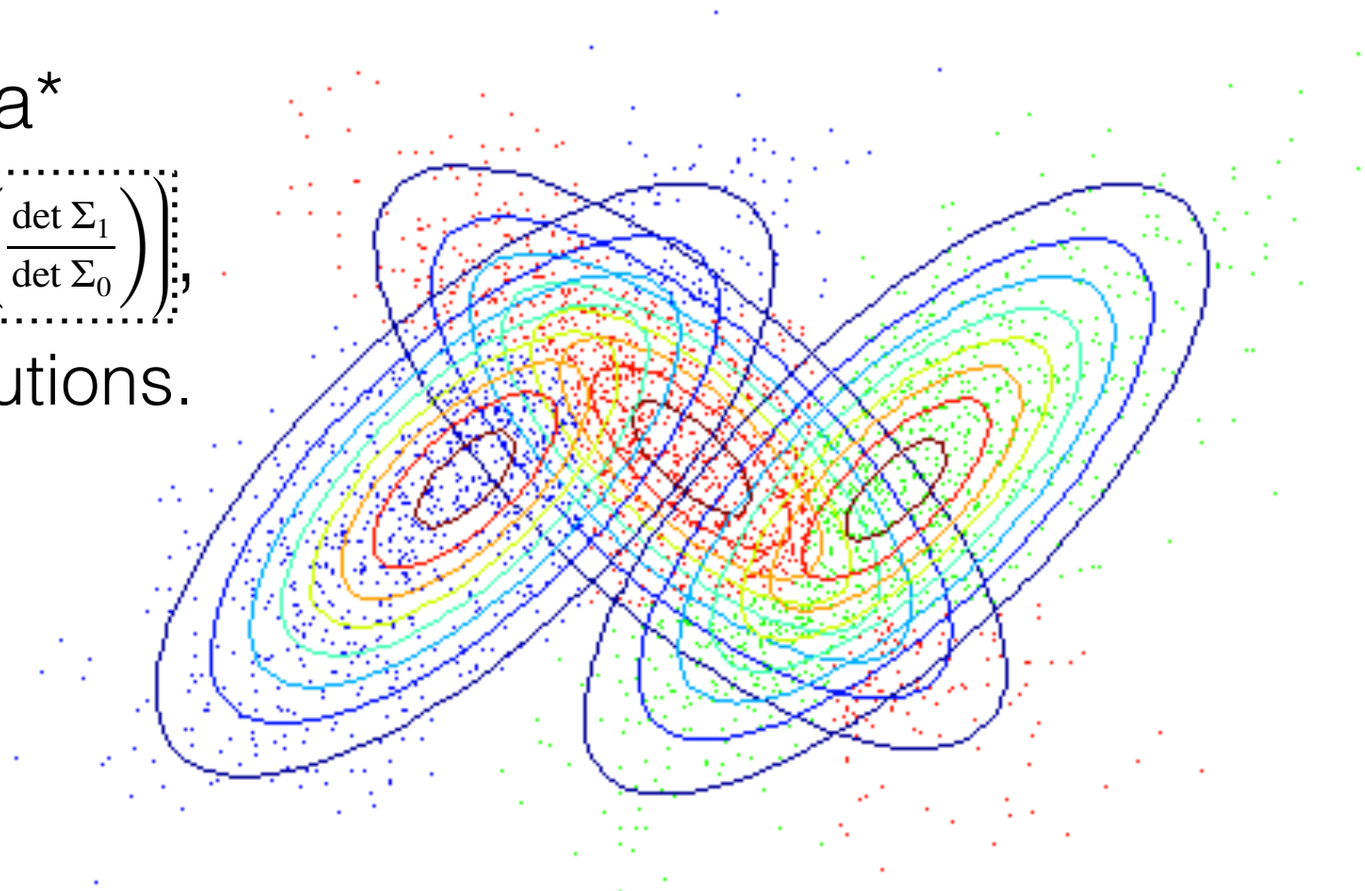
W. Liu, A genetic algorithm enabled ensemble for unsupervised medical term extraction from clinical letters, 2015.

Closed KL Divergence

Closed KL divergence formula*

$$D_{KL}(\mathcal{N}_0 \,||\, \mathcal{N}_1) = \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k + \ln\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right),$$

where k is dimensionality of distributions.



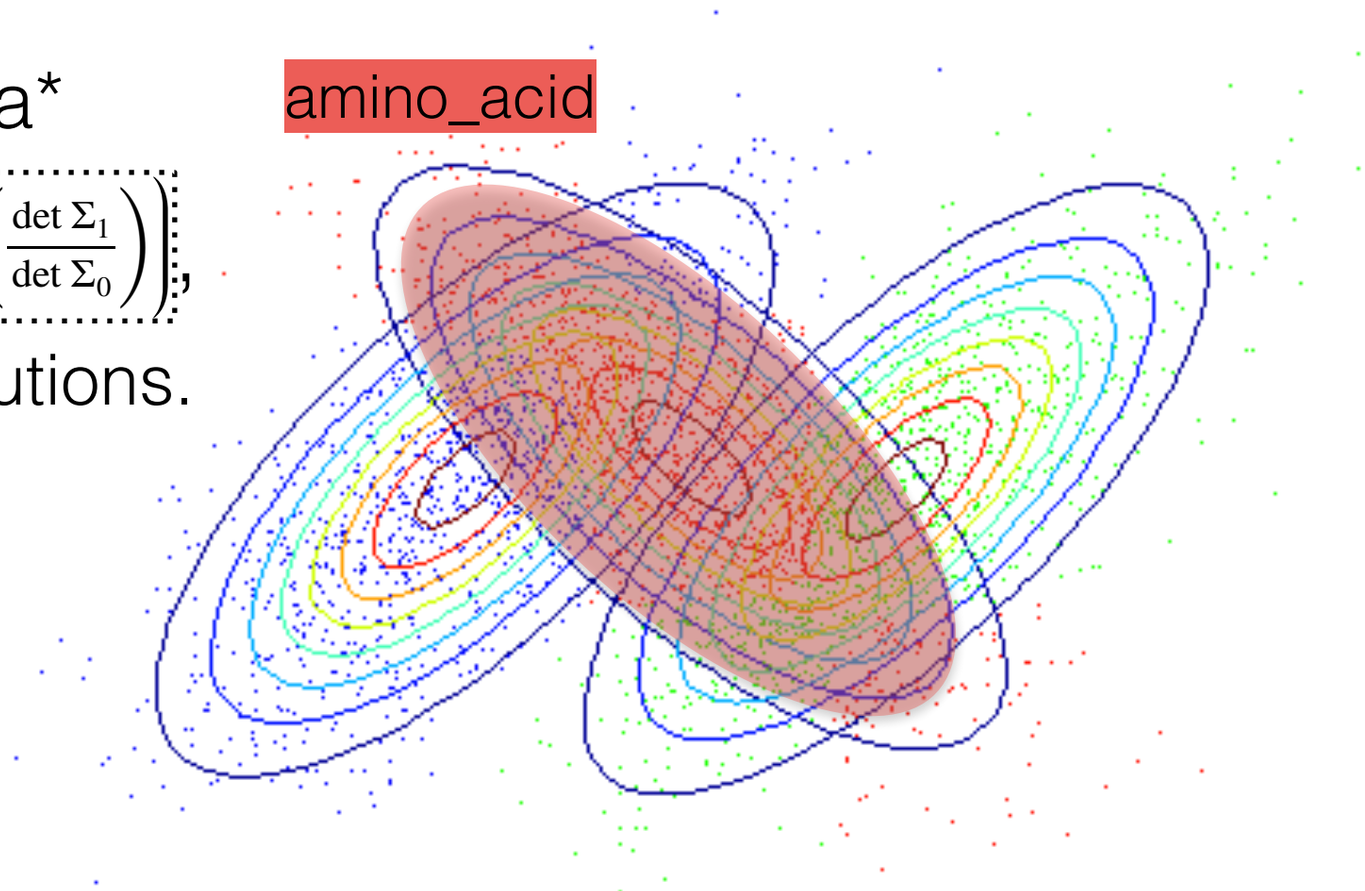
* - https://en.wikipedia.org/wiki/Kullback–Leibler_divergence

Closed KL Divergence

Closed KL divergence formula*

$$D_{KL}(\mathcal{N}_0 \,||\, \mathcal{N}_1) = \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k + \ln\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right),$$

where k is dimensionality of distributions.



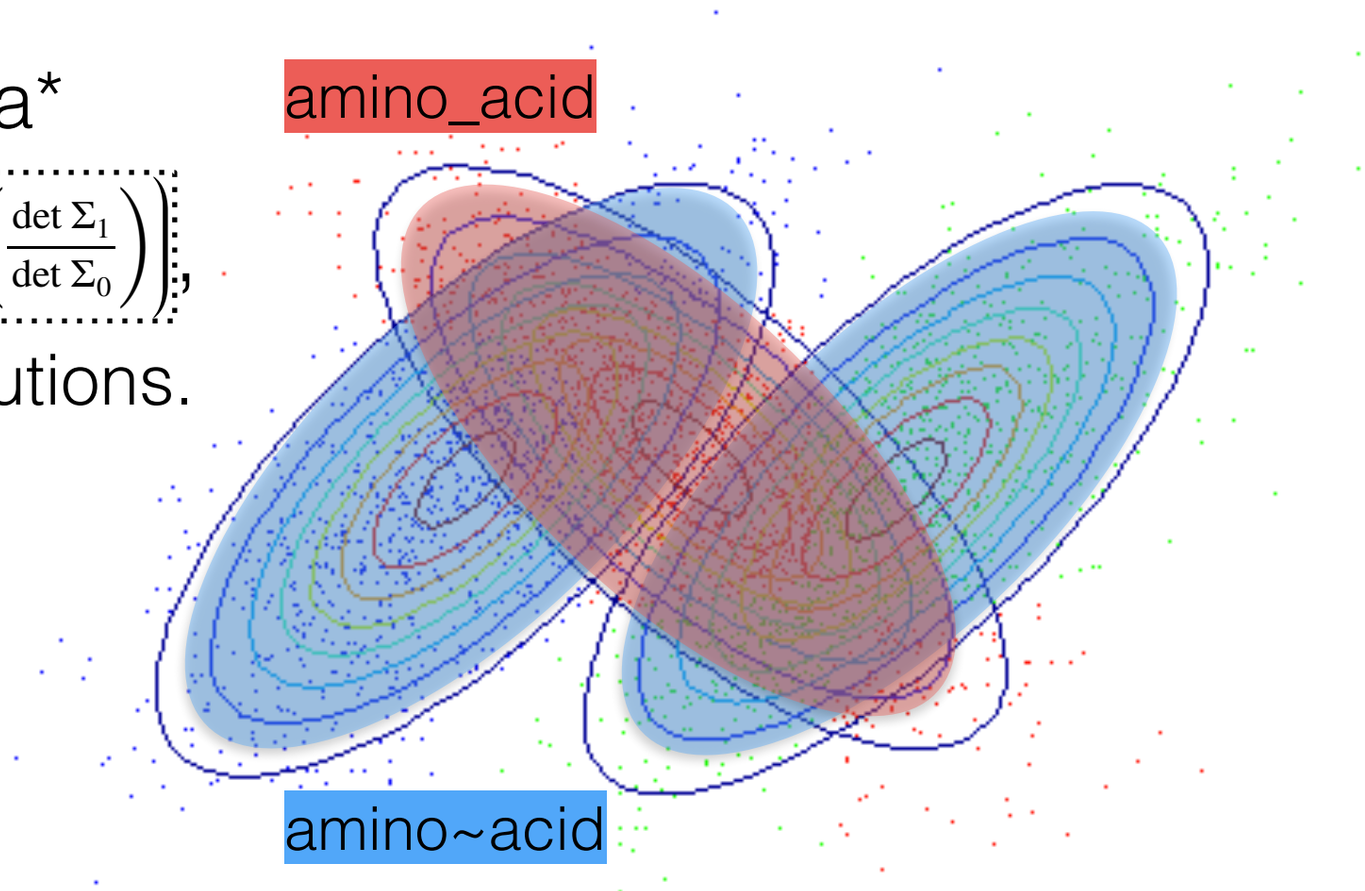
* - https://en.wikipedia.org/wiki/Kullback–Leibler_divergence

Closed KL Divergence

Closed KL divergence formula*

$$D_{KL}(\mathcal{N}_0 \,||\, \mathcal{N}_1) = \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k + \ln\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right),$$

where k is dimensionality of distributions.



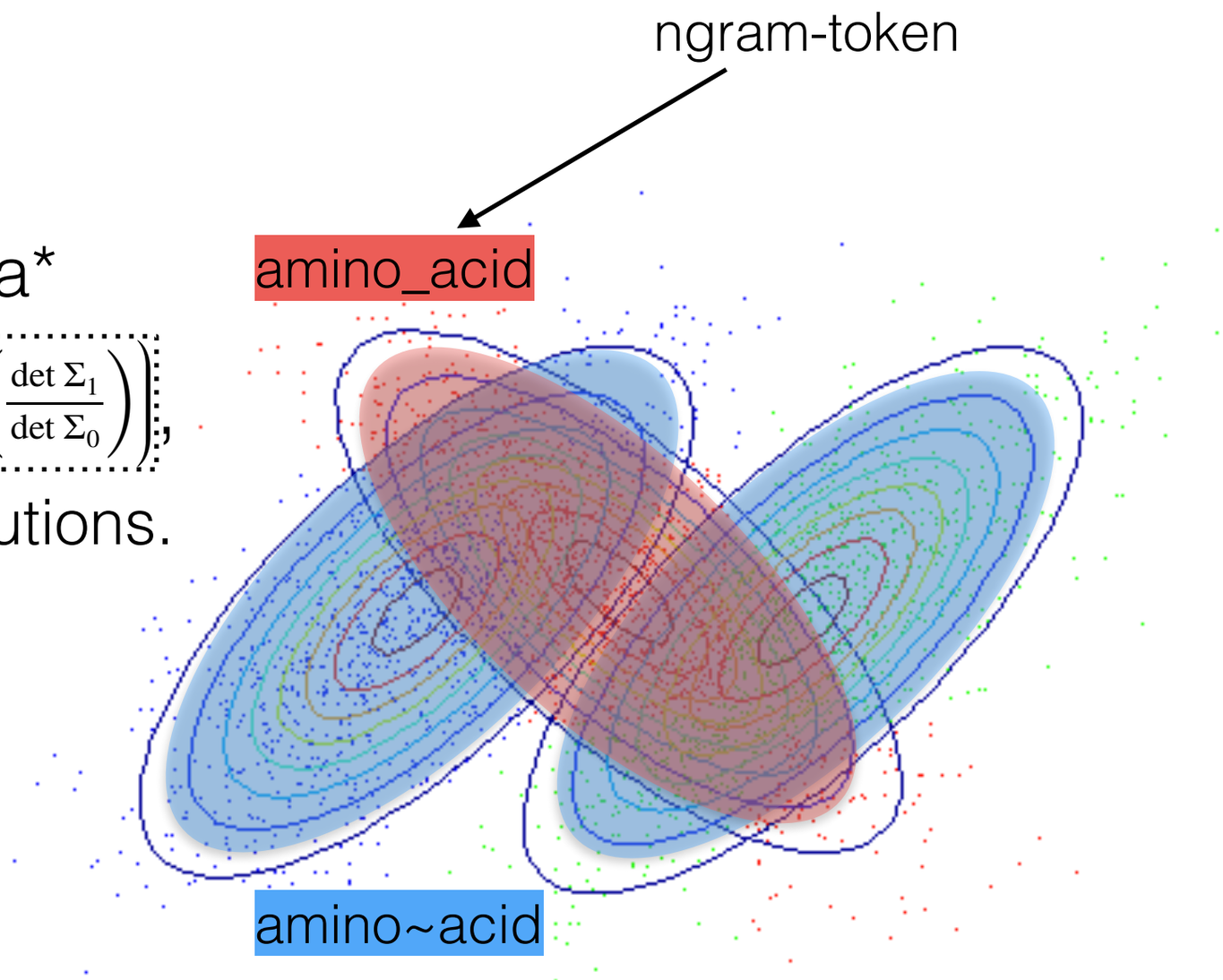
* - https://en.wikipedia.org/wiki/Kullback–Leibler_divergence

Closed KL Divergence

Closed KL divergence formula*

$$D_{KL}(\mathcal{N}_0 \,||\, \mathcal{N}_1) = \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k + \ln\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right),$$

where k is dimensionality of distributions.



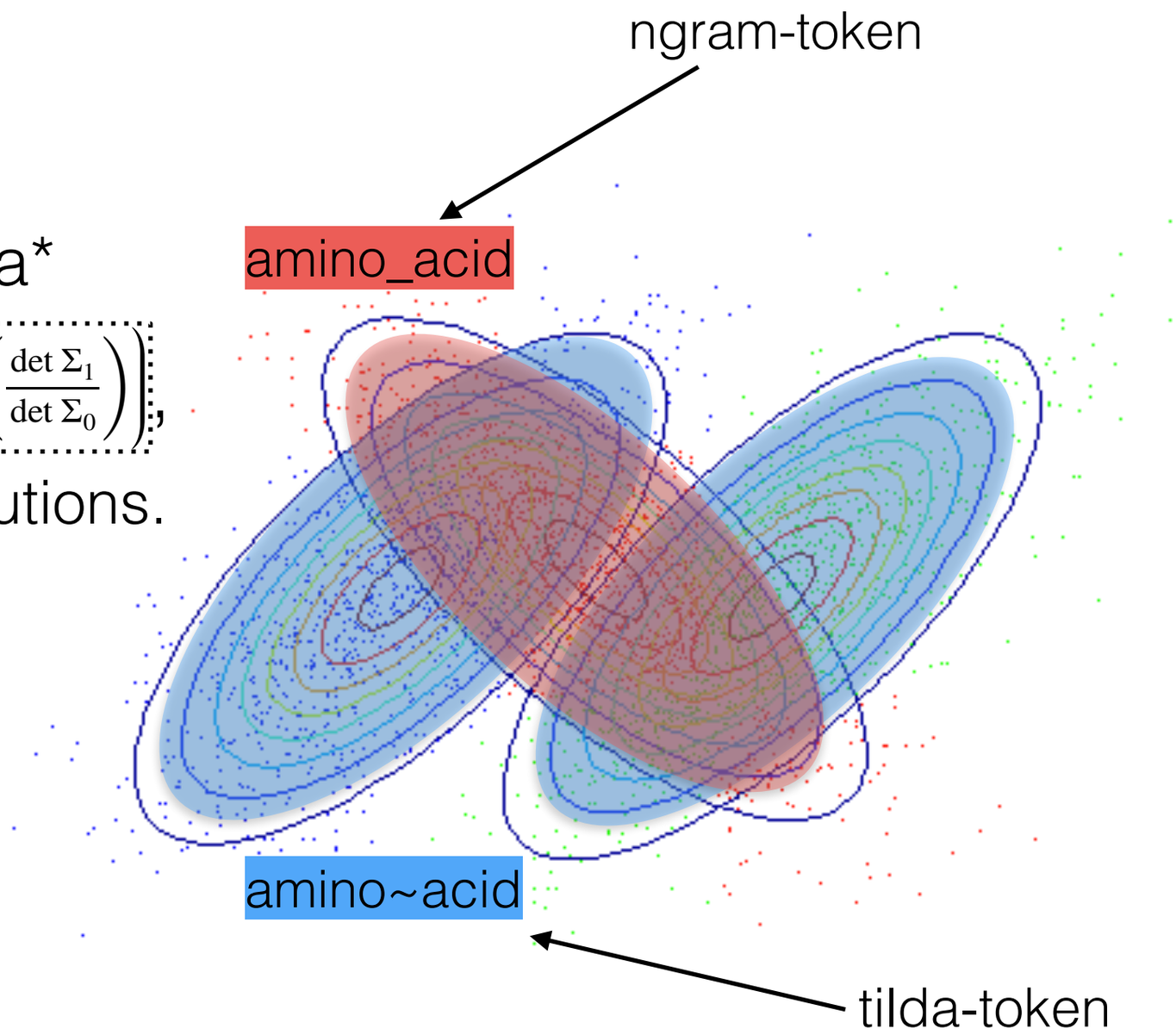
* - https://en.wikipedia.org/wiki/Kullback–Leibler_divergence

Closed KL Divergence

Closed KL divergence formula*

$$D_{KL}(\mathcal{N}_0 \,||\, \mathcal{N}_1) = \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k + \ln\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right),$$

where k is dimensionality of distributions.



* - https://en.wikipedia.org/wiki/Kullback–Leibler_divergence

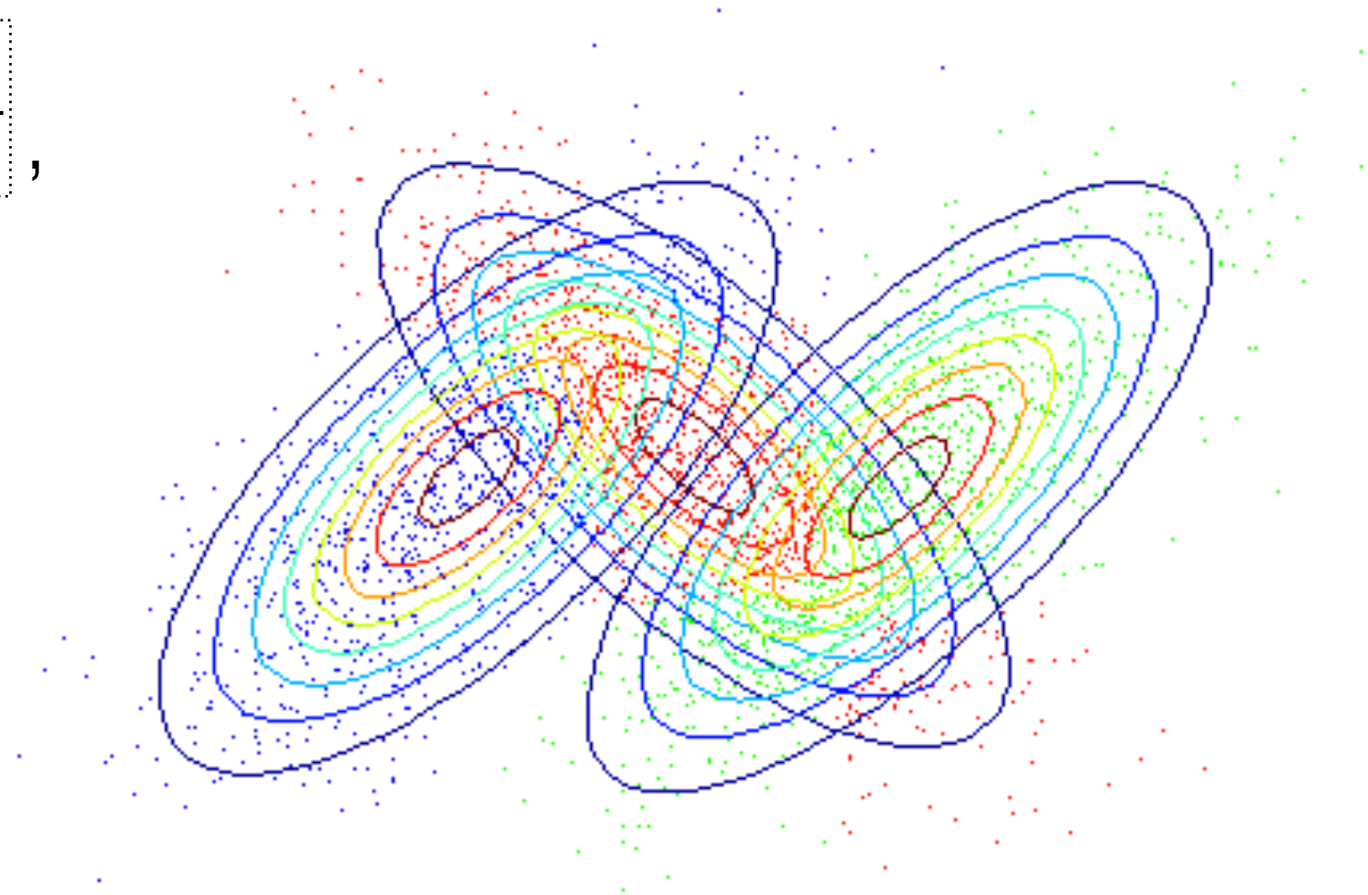
Variational KL Divergence

Variational KL divergence formula*

$$D_{variational}(f\,||\,g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(f_a||f_{a'})}}{\sum_b \omega_b e^{-D(f_a||g_b)}},$$

where
$$\left.\begin{aligned} f(x) &= \sum_a \pi_a \mathcal{N}(x;\mu_a;\Sigma_a), \\ g(x) &= \sum_b \omega_b \mathcal{N}(x;\mu_b;\Sigma_b) \end{aligned}\right\}$$

and
$$\left.\begin{aligned} f_a(x) &= \mathcal{N}(x;\mu_a;\Sigma_a), \\ g_b(x) &= \mathcal{N}(x;\mu_b;\Sigma_b). \end{aligned}\right\}$$



* - J. R. Hershey and P. A. Olsen, Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models, 2007.

Variational KL Divergence

Variational KL divergence formula*

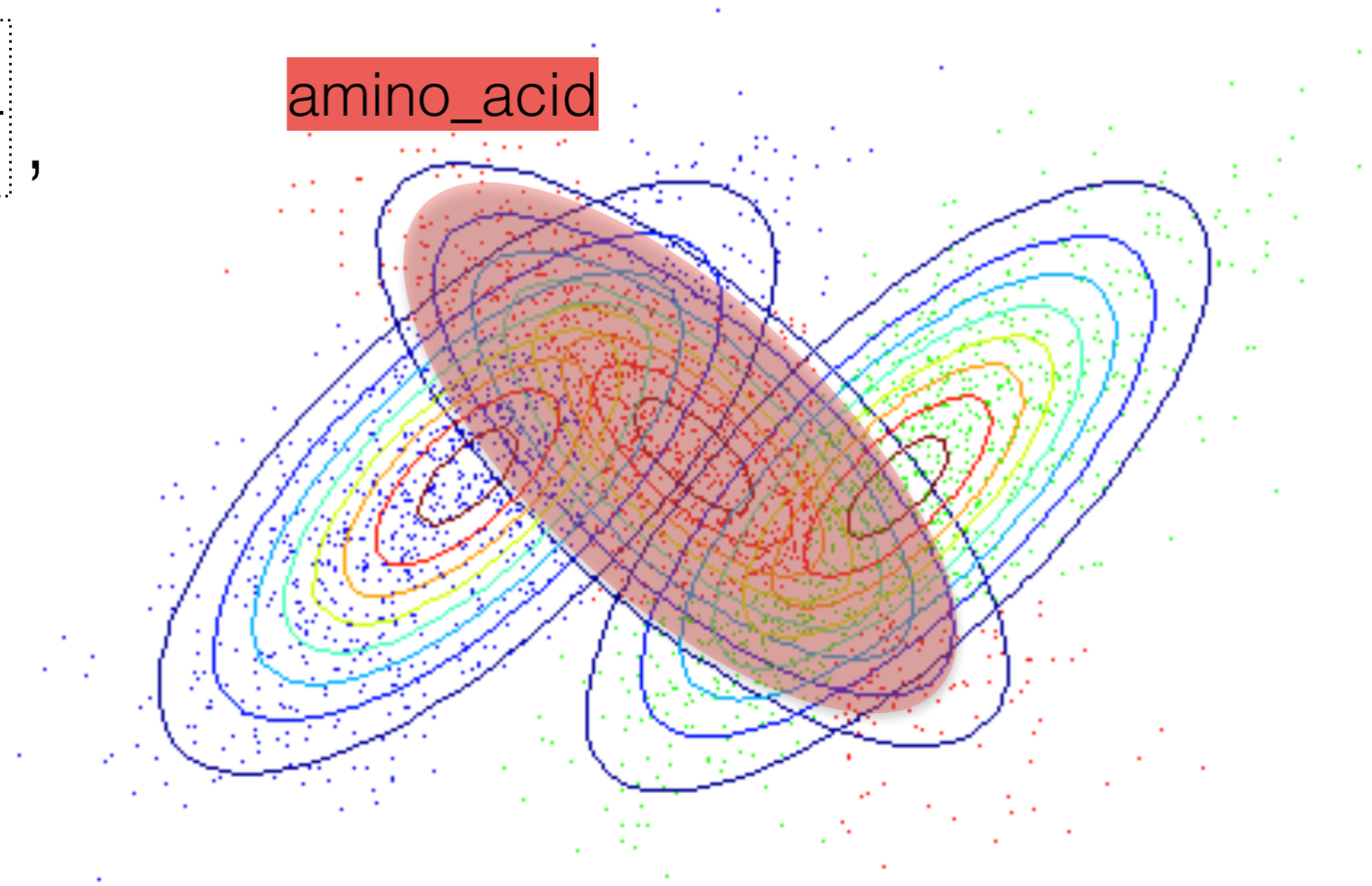
$$D_{variational}(f\,||\,g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(f_a||f_{a'})}}{\sum_b \omega_b e^{-D(f_a||g_b)}},$$

where

$$\left.\begin{aligned} f(x) &= \sum_a \pi_a \mathcal{N}(x;\mu_a;\Sigma_a), \\ g(x) &= \sum_b \omega_b \mathcal{N}(x;\mu_b;\Sigma_b) \end{aligned}\right\}$$

and

$$\left.\begin{aligned} f_a(x) &= \mathcal{N}(x;\mu_a;\Sigma_a), \\ g_b(x) &= \mathcal{N}(x;\mu_b;\Sigma_b). \end{aligned}\right\}$$



* - J. R. Hershey and P. A. Olsen, Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models, 2007.

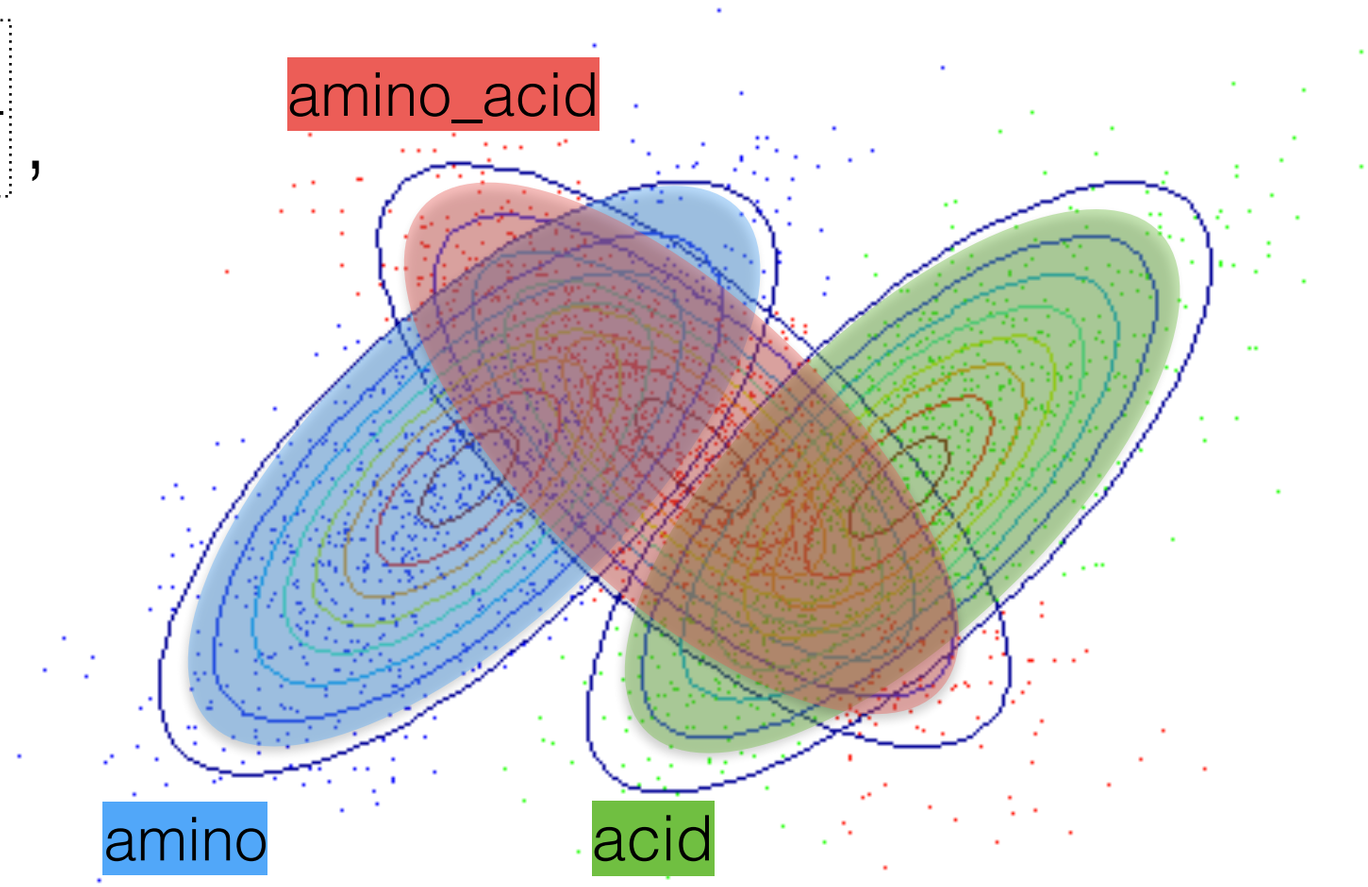
Variational KL Divergence

Variational KL divergence formula*

$$D_{variational}(f\,||\,g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(f_a||f_{a'})}}{\sum_b \omega_b e^{-D(f_a||g_b)}},$$

where
$$f(x) = \sum_a \pi_a \mathcal{N}(x;\mu_a;\Sigma_a),$$
$$g(x) = \sum_b \omega_b \mathcal{N}(x;\mu_b;\Sigma_b)$$

and
$$f_a(x) = \mathcal{N}(x;\mu_a;\Sigma_a),$$
$$g_b(x) = \mathcal{N}(x;\mu_b;\Sigma_b).$$



* - J. R. Hershey and P. A. Olsen, Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models, 2007.

Variational KL Divergence

Variational KL divergence formula*

$$D_{variational}(f\,||\,g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(f_a||f_{a'})}}{\sum_b \omega_b e^{-D(f_a||g_b)}},$$

where

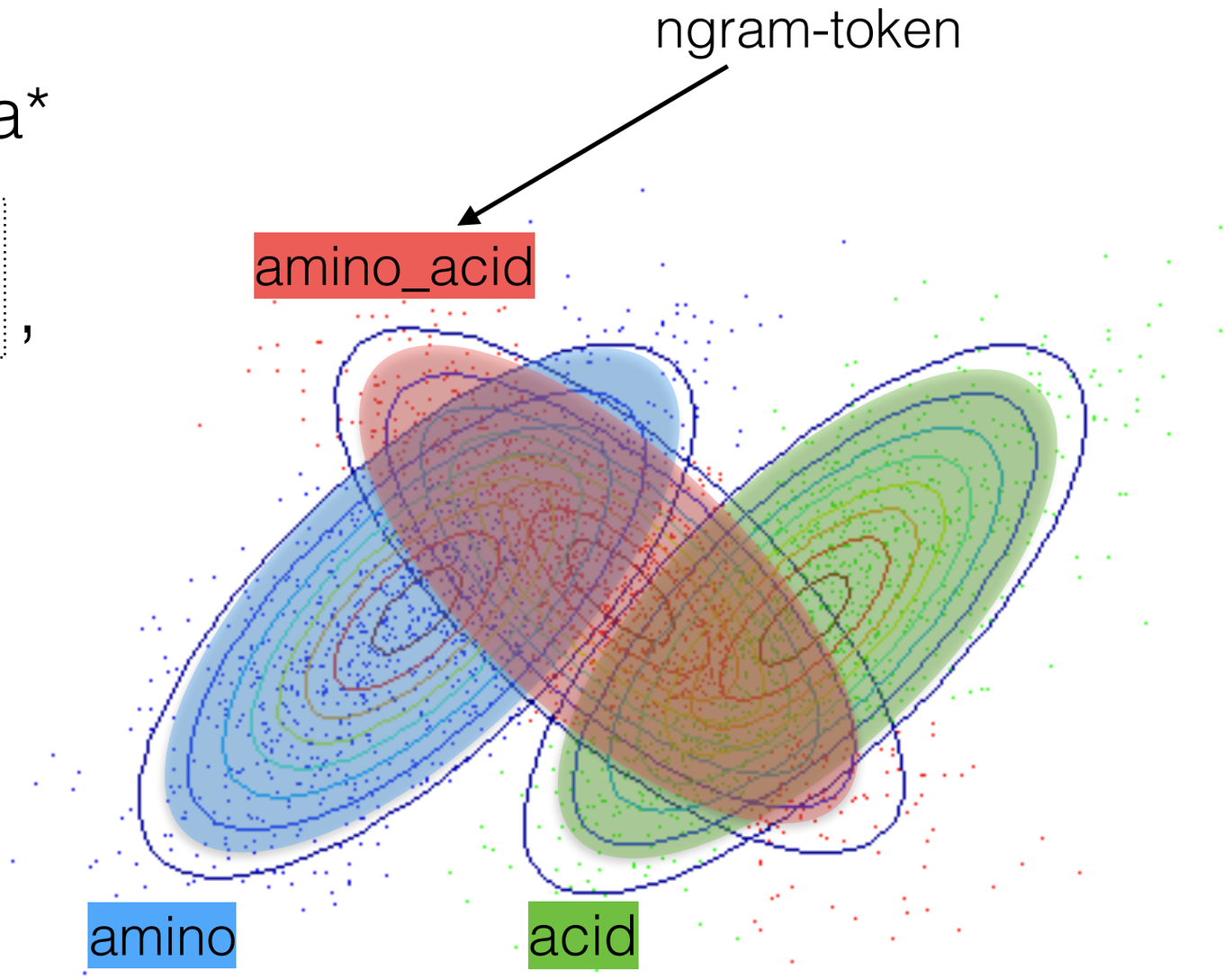
$$f(x) = \sum_a \pi_a \mathcal{N}(x;\mu_a;\Sigma_a),$$

$$g(x) = \sum_b \omega_b \mathcal{N}(x;\mu_b;\Sigma_b)$$

and

$$f_a(x) = \mathcal{N}(x;\mu_a;\Sigma_a),$$

$$g_b(x) = \mathcal{N}(x;\mu_b;\Sigma_b).$$



* - J. R. Hershey and P. A. Olsen, Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models, 2007.

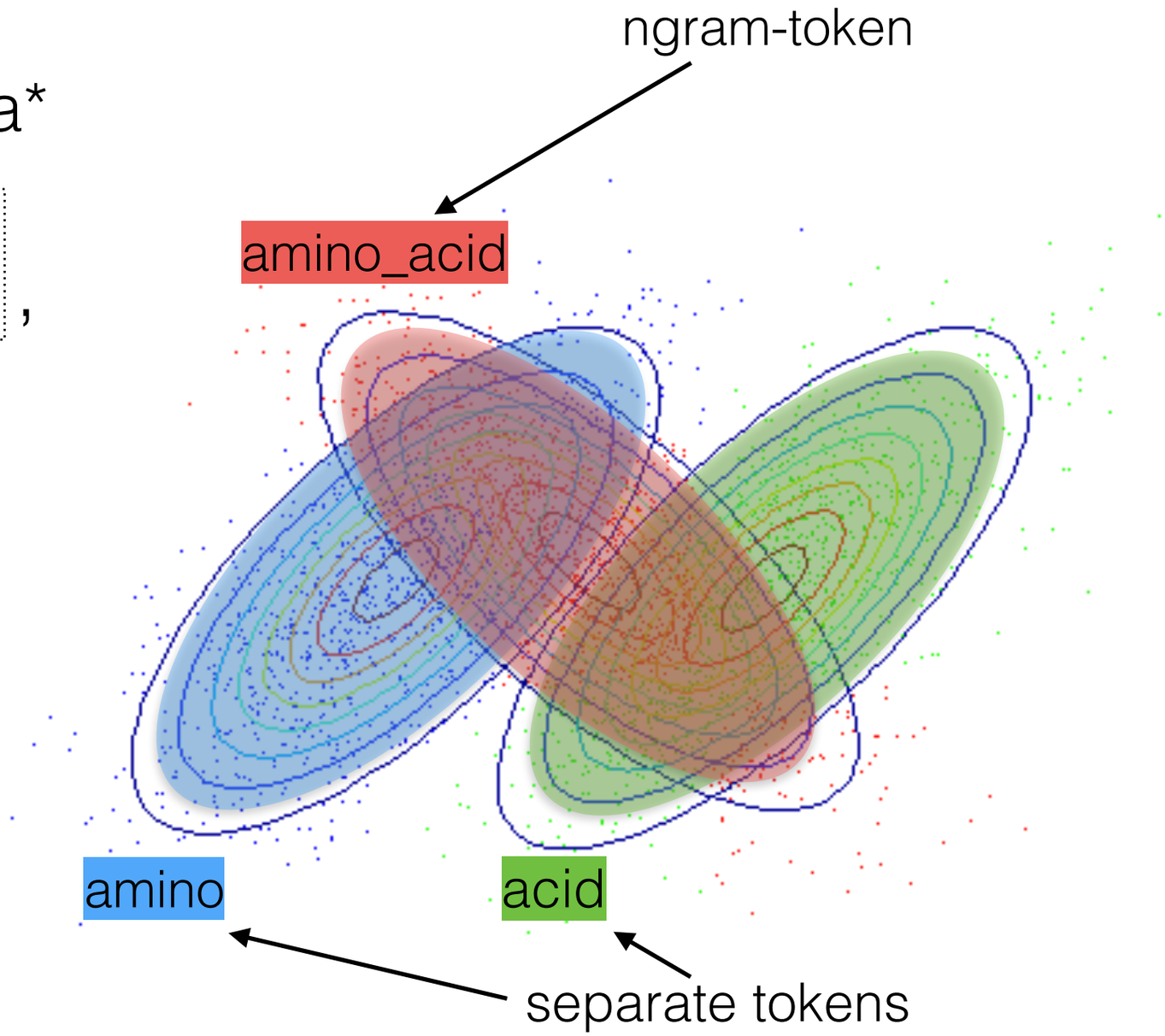
Variational KL Divergence

Variational KL divergence formula*

$$D_{variational}(f||g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(f_a||f_{a'})}}{\sum_b \omega_b e^{-D(f_a||g_b)}},$$

where
$$f(x) = \sum_a \pi_a \mathcal{N}(x; \mu_a; \Sigma_a),$$
$$g(x) = \sum_b \omega_b \mathcal{N}(x; \mu_b; \Sigma_b)$$

and
$$f_a(x) = \mathcal{N}(x; \mu_a; \Sigma_a),$$
$$g_b(x) = \mathcal{N}(x; \mu_b; \Sigma_b).$$



* - J. R. Hershey and P. A. Olsen, Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models, 2007.

Results

	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	voltage gated	case series	present study	fetal cells	tumorigenic clones	db db
2	resonance energy	therapy exposure	breast cancer	induced transition	glun1 glun2b	w w
3	direct suppression	prenatal diagnosis	wild type	gene expression	ophthalmology journals	substantia nigra
4	tumor samples	mid log	results suggest	suggested guidelines	egfp nachralpha3	tumorigenic clones
5	prenatal diagnosis	tumorigenic clones	gene expression	complete follow	numbered tag	intestinal tract
6	serine threonine	l lactis	cell lines	coagulation function	substantia nigra	numbered tag
7	n acetylcysteine	resonance energy	long term	related infections	volar oblique	alfalfa peroxidase
8	antioxidant potential	agarose gel	cell cycle	clinical studies	elastin matrix	qbeta replicase
9	negative affectivity	test whether	risk factors	first data	nursing home	hydroxymethylglutaryl coenzyme
10	therapy exposure	tumor samples	polymerase chain	pathway conduction	regenerative medicine	transdermal buprenorphine
11	ovx zol	williams syndrome	growth factor	treated seeds	hydroxymethylglutaryl coenzyme	lysinibacillus sp
12	light chain	foot protein	chain reaction	gene increase	lysinibacillus sp	oral contraceptives
13	pedot go	isolates type	significant difference	similar trends	hypocotyl elongation	tick borne
14	rs10468017 variant	neuron areas	cell death	higher levels	intestinal tract	distant metastases
15	tumorigenic clones	mosaic virus	cancer cells	investigated lesions	carbonic anhydrase	lxxll motif
16	mirror neuron	eec syndrome	mg kg	bone marrow	mirror neuron	glun1 glun2b
17	neuron areas	voltage gated	cell line	cu b	charnley classification	bm examination
18	elevated umbilical	nci n78	significantly higher	molecular computing	voluntary interruptions	xl cgd
19	proteomic analysis	distant metastasis	important role	zokor species	moogoo udder	clinicaltrials gov
20	cultured fibroblasts	clinical importance	cell surface	affected organs	try ends	multivariate logistic

Results

	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	voltage gated	case series	present study	fetal cells	tumorigenic clones	db db
2	resonance energy	therapy exposure	breast cancer	induced transition	glun1 glun2b	w w
3	direct suppression	prenatal diagnosis	wild type	gene expression	ophthalmology journals	substantia nigra
4	tumor samples	mid log	results suggest	suggested guidelines	egfp nachralpha3	tumorigenic clones
5	prenatal diagnosis	tumorigenic clones	gene expression	complete follow	numbered tag	intestinal tract
6	serine threonine	l lactis			substantia nigra	numbered tag
7	n acetylcysteine	resonance energy			volar oblique	alfalfa peroxidase
8	antioxidant potential	agarose gel			elastin matrix	qbeta replicase
9	negative affectivity	test whether			nursing home	hydroxymethylglutaryl coenzyme
10	therapy exposure	tumor samples			regenerative medicine	transdermal buprenorphine
11	ovx zol	williams syndrome	growth factor	treated seeds	hydroxymethylglutaryl coenzyme	lysinibacillus sp
12	light chain	foot protein	chain reaction	gene increase	lysinibacillus sp	oral contraceptives
13	pedot go	isolates type	significant difference	similar trends	hypocotyl elongation	tick borne
14	rs10468017 variant	neuron areas	cell death	higher levels	intestinal tract	distant metastases
15	tumorigenic clones	mosaic virus	cancer cells	investigated lesions	carbonic anhydrase	lxxll motif
16	mirror neuron	eec syndrome	mg kg	bone marrow	mirror neuron	glun1 glun2b
17	neuron areas	voltage gated	cell line	cu b	charnley classification	bm examination
18	elevated umbilical	nci n78	significantly higher	molecular computing	voluntary interruptions	xl cgd
19	proteomic analysis	distant metastasis	important role	zokor species	moogoo udder	clinicaltrials gov
20	cultured fibroblasts	clinical importance	cell surface	affected organs	try ends	multivariate logistic

TF-IDF and TextRank

Results

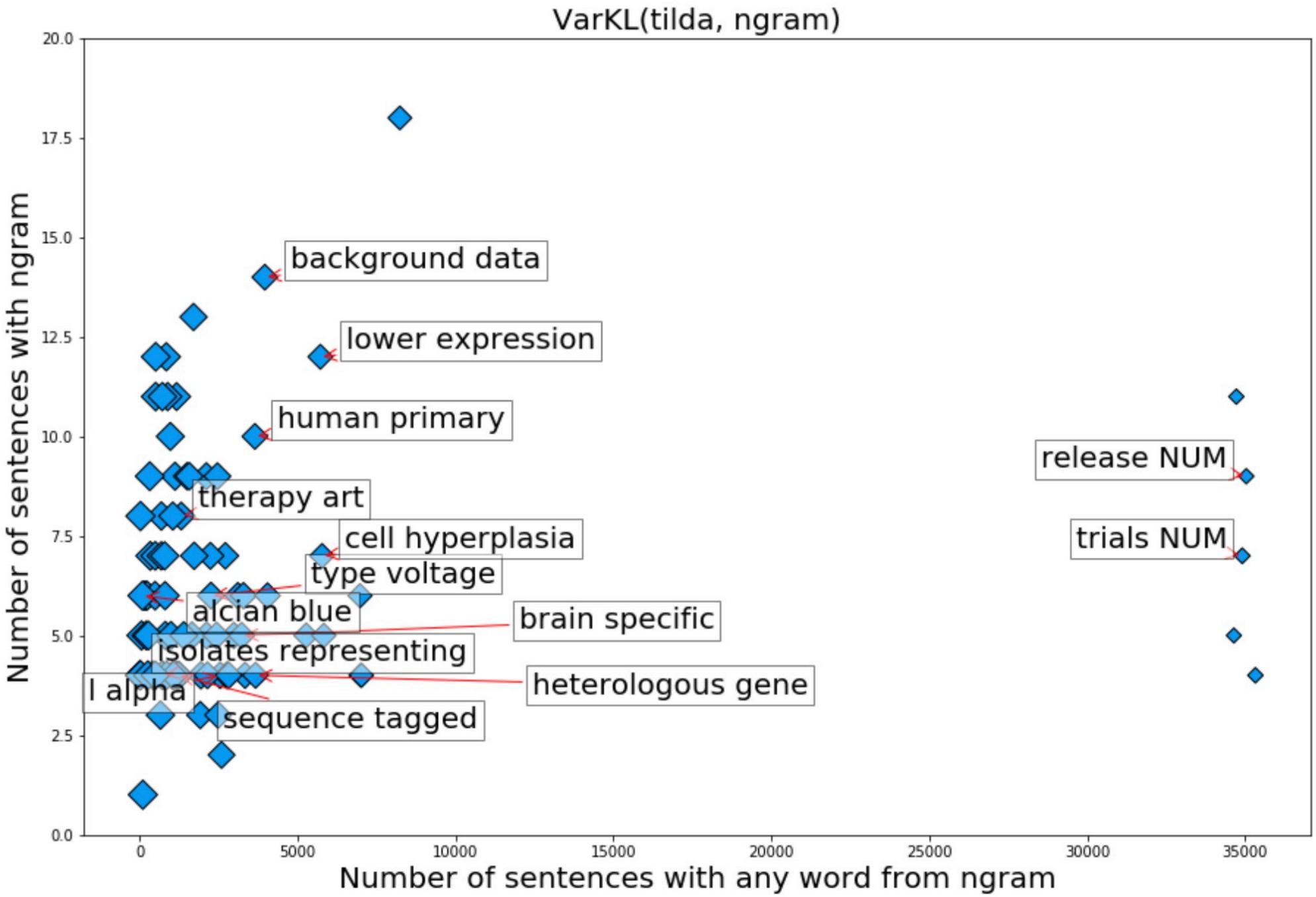
	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	voltage gated	case series	present study	fetal cells	tumorigenic clones	db db
2	resonance energy	therapy exposure	breast cancer	induced transition	glun1 glun2b	w w
3	direct suppression	prenatal diagnosis	wild type	gene expression	ophthalmology journals	substantia nigra
4	tumor samples	mid log	results suggest	suggested guidelines	egfp nachralpha3	tumorigenic clones
5	prenatal diagnosis	tumorigenic clones	gene expression	complete follow	numbered tag	intestinal tract
6						
7						
8						
9						yme
10						e
11	uvx 20r	williams syndrome	growth factor	treated seeds	hydroxymethylglutaryl coenzyme	lysinibacillus sp
12	light chain	foot protein	chain reaction	gene increase	lysinibacillus sp	oral contraceptives
13	pedot go	isolates type	significant difference	similar trends	hypocotyl elongation	tick borne
14	rs10468017 variant	neuron areas	cell death	higher levels	intestinal tract	distant metastases
15	tumorigenic clones	mosaic virus	cancer cells	investigated lesions	carbonic anhydrase	lxxll motif
16	mirror neuron	eec syndrome	mg kg	bone marrow	mirror neuron	glun1 glun2b
17	neuron areas	voltage gated	cell line	cu b	charnley classification	bm examination
18	elevated umbilical	nci n78	significantly higher	molecular computing	voluntary interruptions	xl cgd
19	proteomic analysis	distant metastasis	important role	zokor species	moogoo udder	clinicaltrials gov
20	cultured fibroblasts	clinical importance	cell surface	affected organs	try ends	multivariate logistic

Closed KL

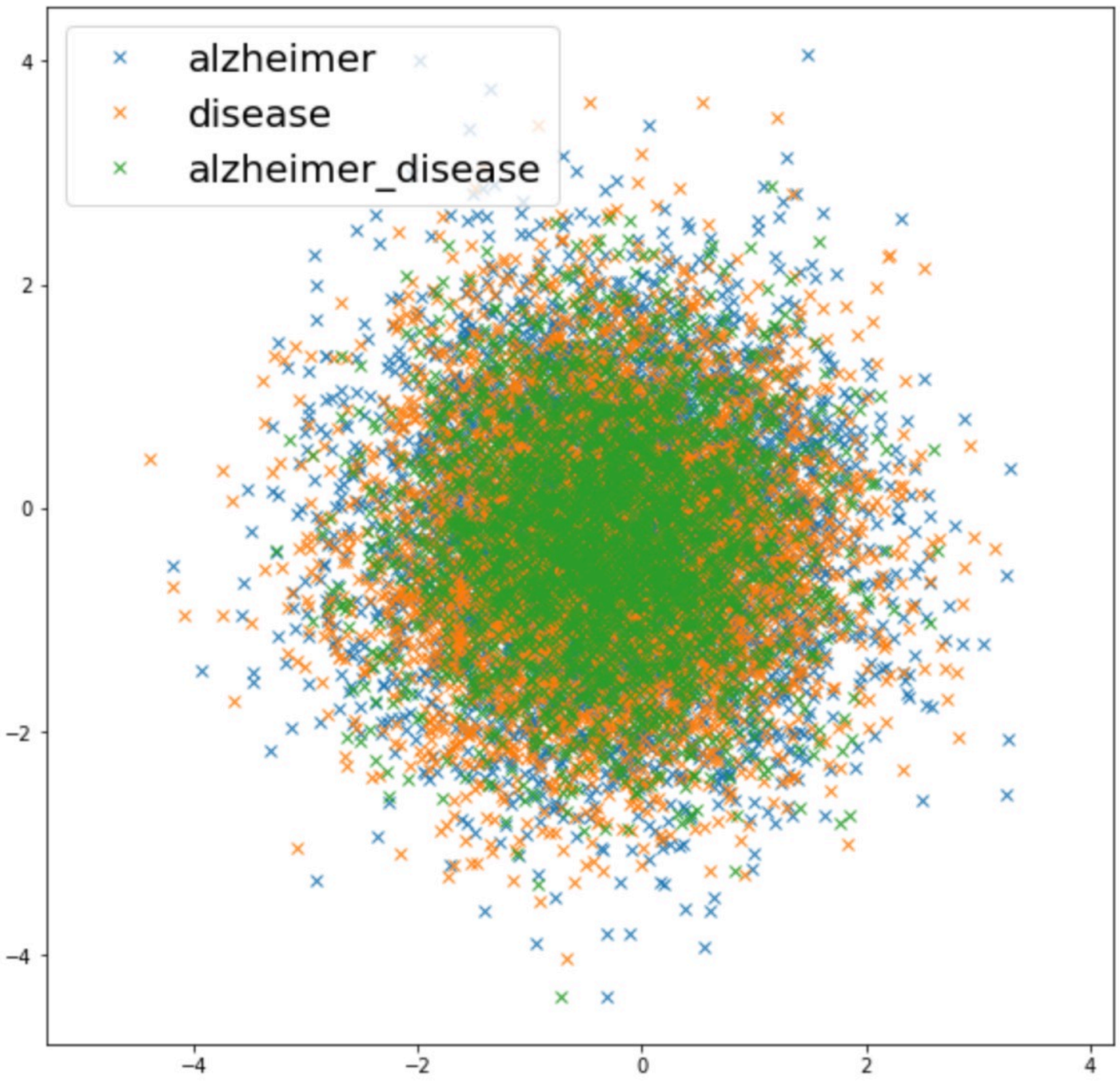
TF-IDF and TextRank

Variational KL

Results


VarKL(tilda, ngram)
Number of sentences with any word from ngram
Number of sentences with ngram
background data
lower expression
human primary
therapy art
cell hyperplasia
type voltage
alcian blue
brain specific
isolates representing
heterologous gene
l alpha
sequence tagged
release NUM
trials NUM
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
5000
10000
15000
20000
25000
30000
35000

Variational KL



Summary

- TF-IDF
- PageRank (TextRank)
- New methods (Closed KL / Variational KL)