Disentangling Emotional Information from Speech Signals via Representation Learning

Fabio Massimo Zennaro
Supervisor: Dr Ke Chen
Co-supervisor: Dr Xiaojun Zeng
Advisor: Prof Stephen Furber

University of Manchester
End of Second Ph.D. Year

June 26th, 2015

Outline of the presentation

- *Introduction* - presenting the field of research and our project;
- *Progress of the Research* - explaining the work and the results obtained so far;
- *Plan for the Research* - laying out our plan for the next year's work.

Background

Affective Computing - research on the development of emotional-aware computers.

Emotional Speech - one of the main channels through which emotions are expressed.

Emotional Information Extraction from Speech - problem of extracting emotional information from acoustic over-informative signal.

Our vision

Emotional Information Extraction

Emotional Information Disentanglement: generating representations containing *all and only* emotional information.

Semantic Representation Learning: generating representation *homomorphic* with human understanding and meaning.

Disentangled and *semantic* representations usable across a wide array of emotional-related tasks.

Our contribution

Emotional Information Disentanglement

Machine Learning: developing algorithms for information disentanglement.

Affective Computing: showing the effectiveness of our solution in a real-world scenario.

Strands of Research

Approaches to Information Disentanglement

Feature Distribution Learning: study of learning in the feature distribution space. [2]

Information Theoretic Learning: study of learning guided by information theory. [4]

DSF: Disentangling Sparse Filtering

ITLR: Information Theoretic Representation Learning

Feature Distribution Learning

Feature Distribution Learning is an approach to *unsupervised learning* focusing on learning the distribution of data in the *feature space* instead of the *data space*.

Sparse Filtering (SF) is a prototypical algorithm for feature distribution learning based on the learning of a *sparse distribution*.

SF was shown to be a good algorithm with respect to performance, number of hyperparameters and computational cost.

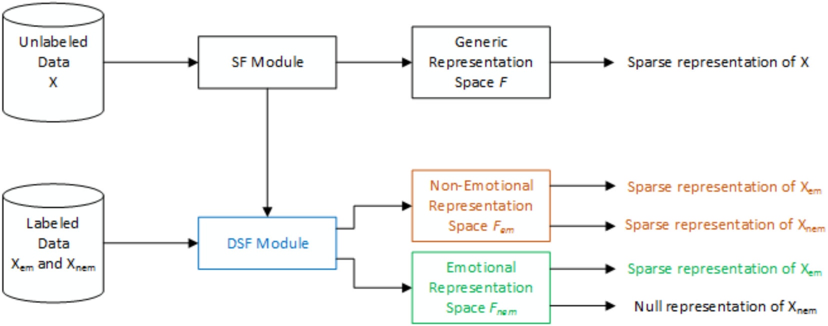
Disentangling Sparse Filtering (1)

Scenario: detecting the presence of emotion in speech in real-time, relying on vast amounts of unlabelled recorded data.

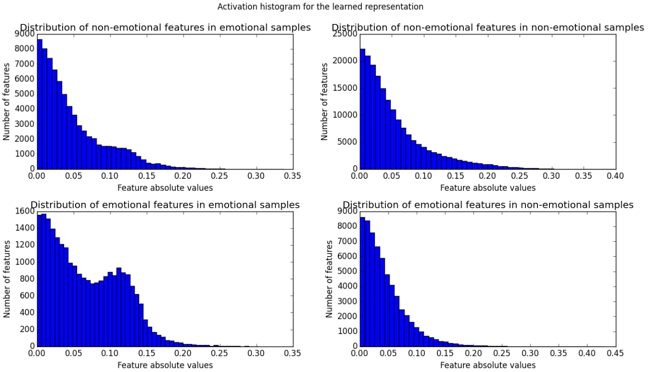
Starting from SF, we worked on feature distribution learning and we:

1. Extended SF to *online settings*;
2. Extended SF to *semi-supervised settings*;
3. Developed new algorithms for learning *disentangled sparse representations* (DSF_D) or *orthogonal sparse representations* (DSF_{AD}) of emotional speech.

Disentangling Sparse Filtering (2)

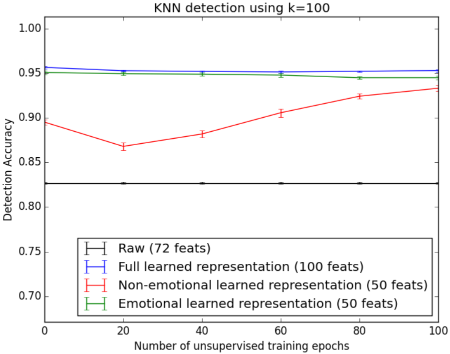


Preliminary Results (1)



Activation: we learned a markedly different distribution of emotional information over the emotional samples compared to non-emotional samples.

Preliminary Results (2)



Detection Accuracy: the learned emotional representation allows us to achieve high accuracy in emotion detection.

Future Work (1)

Finalization of Work on DSF

- Evaluation of DSF using different datasets;
- Evaluation of DSF on different emotional tasks;
- Comparison of DSF against other methods presented in the literature.

Outcome: journal article (*IEEE TAC* or *IEEE NNLS*) or conference paper (*ICML* or *ICLR*)

Future Work (2)

Improving Emotional Information Disentanglement

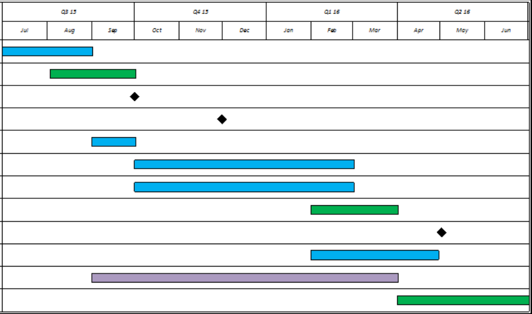
DSF+ITLR: using information theoretic learning for disentangling feature distribution learning.

Deep DSF: stacking together DSF learning modules.

Outcome: journal article (*IEEE TAC* or *IEEE NNLS*) or conference paper (*NIPS*)

Gantt Chart

ID	Task Name	Start	Finish
1	Finalization of the work on DSF	01/07/2015	01/09/2015
2	Collection of the result of DSF for publication	03/08/2015	01/10/2015
3	Deadline: ICML	01/10/2015	01/10/2015
4	Deadline: ICLR	01/12/2015	01/12/2015
5	Feasibility study for further work on emotional information disentanglement	01/09/2015	01/10/2015
6	Development of disentangling algorithms based on DSF and ITLR	01/10/2015	01/03/2016
7	Development of deep disentangling algorithms	01/10/2015	01/03/2016
8	Collection of the result for publication	01/02/2016	01/04/2016
9	Deadline: NIPS	02/05/2016	02/05/2016
10	Implementation of a case study scenario	01/02/2016	29/04/2016
11	Formalization of the definition of disentanglement	01/09/2015	01/04/2016
12	Writing up dissertation	01/04/2016	01/07/2016



Thank you!

Thank you for listening!

References I

[1] Geoffrey E. Hinton and Ruslan R. Salakhutdinov.
Reducing the dimensionality of data with neural networks.
Science, 28:504–507, 2006.

[2] Jiquan Ngiam, Zhenghao Chen, Sonia A Bhaskar, Pang W Koh, and Andrew Y Ng.
Sparse filtering.
In *Advances in Neural Information Processing Systems*, pages 1125–1133, 2011.

[3] Robert Plutchik.
Emotion: A Psychoevolutionary Synthesis.
Harper and Row, 1980.

References II

[4] Jose C Principe.
Information theoretic learning: Rényi's entropy and kernel perspectives.
Springer, 2010.

[5] James A. Russell.
A circumplex model of affect.
Journal of Personality and Social Psychology, 39:1161–1178, 1980.

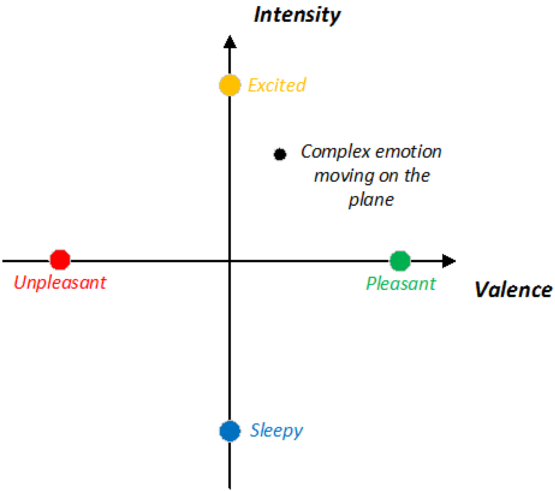
[6] James A. Russell.
Core affect and the psychological construction of emotion.
Psychological Review, 110:145–172, 2003.

Applications of Emotional-Aware Computing

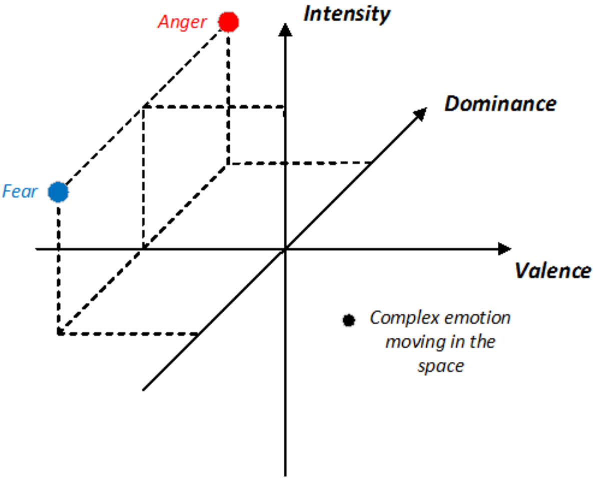
Several *applications* may take advantage of computers able to deal with emotions, such as:

- Diagnostic Systems,
- On-Line Learning Environments,
- Artificial Agents for Social Assistance,
- Customer Satisfaction Systems,
- Mood-Driven Applications,
- Virtual Games.

Continuous Theories of Emotion (1)



Continuous Theories of Emotion (2)



Discrete Theories of Emotion (1)

Ekman (1969) "Big Six"	Ekman (1999)	Lazarus (1999)	Buck (1999)	Lewis and Havilland (1993)	Banse and Scherer (1996)	Cowie (1999)
Anger	Anger	Anger	Anger	Anger / Hostility	Rage / Hot Anger	Angry
					Irritation / Cold Anger	
Fear	Fear	Fright	Fear	Fear	Fear / Terror	Afraid
Sadness	Sadness / Distress	Sadness	Sadness	Sadness	Sadness / Dejection	Sad
					Grief / Desperation	
		Anxiety	Anxiety	Anxiety	Worry / Anxiety	Worried
Happiness	Sensory pleasure	Happiness	Happiness	Happiness	Happiness	Happy
					Elation / Joy	
	Amusement			Humour		Amused
	Satisfaction					Pleased
	Contentment					Content
			Interested			Interested
			Curious			

Discrete Theories of Emotion (2)

Ekman (1969) "Big Six"	**Ekman (1999)**	**Lazarus (1999)**	**Buck (1999)**	**Lewis and Havilland (1993)**	**Banse and Scherer (1996)**	**Cowie (1999)**
Surprise			Surprised			
	Excitement					Excited
			Bored		Boredom / Indifference	Bored
						Relaxed
			Burn out			
Disgust	Disgust	Disgust	Disgust	Disgust	Disgust	
	Contempt		Scorn			
	Pride	Pride	Pride	Pride		
			Arrogance			
		Jealousy	Jealousy			
		Envy	Envy			
	Shame	Shame	Shame	Shame	Shame / Guilt	
	Guilt	Guilt	Guilt	Guilt		
	Embarassment			Embarassment		
						Disappointed
	Relief	Relief				
		Hope				
						Confident
		Gratitude				
		Love		Love		Loving
						Affectionate
		Compassion	Pity			
			Moral rapture			
			Moral indignation			
		Aesthetic				

Unified Theory of Emotion

We suggest a unified theory of emotion, which is rooted in the models proposed by Russell (*two-dimensional circumplex* [5]), Whissel and Plutchik (*emotion wheel* [3]).

This unified model is built around the concepts [6] of:

- *Core Affects*, that is, neurophysiological states experienced as a continuous feeling described by hedonic and arousal values;
- *Emotional Episodes*, that is, discrete events during which the core affect undergoes a sensible change.

Emotional Speech

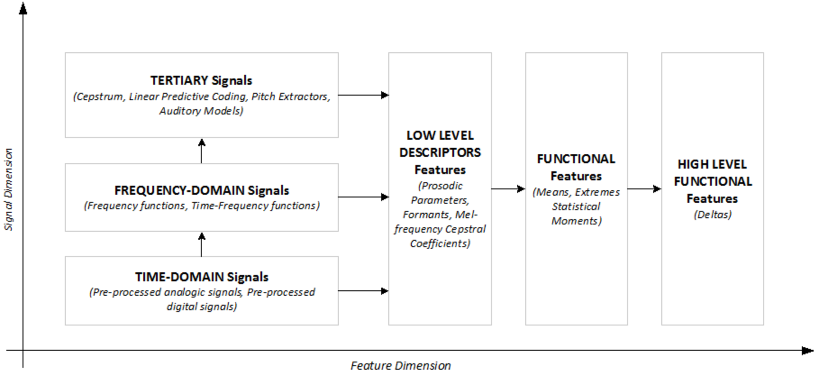
Speech is an *overinformative signal* containing many elements of information, such as:

- *Linguistic information*, related to the meaning of the uttered sounds;
- *Paralinguistic information*, related to the inner state of the speaker;
- *Extralinguistic information*, related to the cultural traits of the speaker.

Emotional Datasets

Corpus	*Year*	*Rec.*	*Lang*	*Speakers*	*Audio/Video*	*Emotions*	*#Sam*
Berlin Emotional Database	1997	acted (studio)	Ger	5F, 5M	A	Discrete theory with 7 basic emotions (anger, boredom, disgust, fear, joy, neutral, sadness)	700+100 sen-tences
DES (Danish Emotional Speech)	1996	acted (studio)	Dan	2F, 2M	A	Discrete theory with 5 basic emotions (anger, happiness, neutral, sadness, surprise)	260+81 utter-ances
MAV (Montreal Affective Voices)	2008	acted (studio)	Fre	15F, 15M	A	Discrete theory with 9 basic emotions (anger, disgust, fear, pain, happiness, neutral, pleasure, sadness, surprise) and continuous theory with 3 dimensions (valence, arousal, intensity)	90 bursts
VAM (Vera Am Mittag Corpus)	2008	natural	Ger	36F, 11M	AV	Continuous theory with 3 dimensions (valence, intensity, dominance)	1018 utter-ances
eNTERFACE	2004	induced	Eng	8F, 34M	AV	Discrete theory with 7 basic emotions (anger, disgust, fear, happiness, neutral, sadness, surprise)	1166 se-quences
TIMIT	1993	acted (studio)	Eng	192F, 438M	A	Not emotional	6300 sen-tences

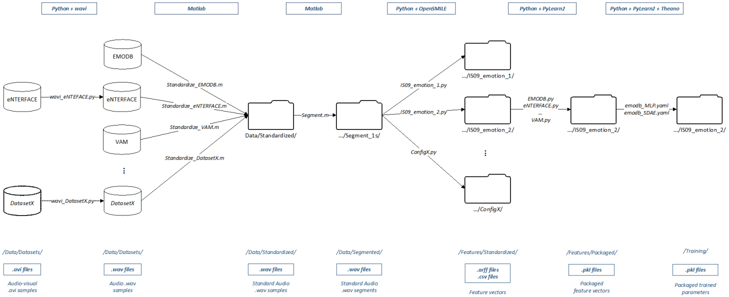
Representations



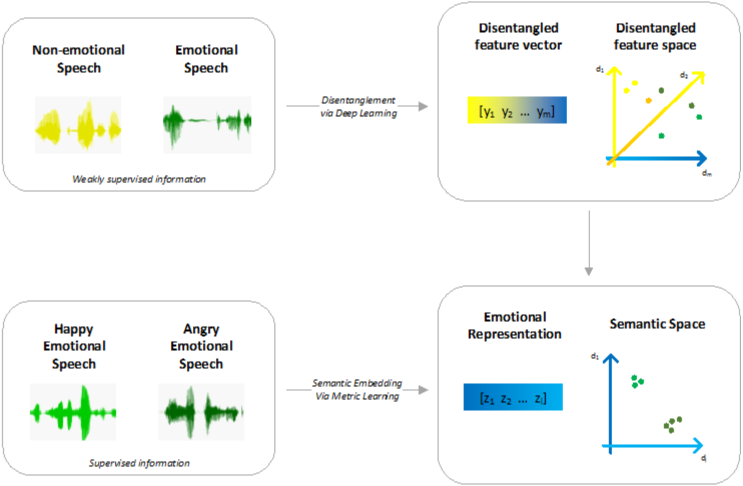
Low-Level Descriptors

Family of Features	Types of Features	Examples of Features
Prosodic	Fundamental Frequency	F_0, characterising points, contours
	Intensity	Energy, characterising points, root mean energy
	Time	Duration, voice and unvoiced segments ratio, zero-crossing rate
	Voice Quality	Band-energies
Spectral	Formants	Formants
	Spectral Shape	Band-energies, roll-off, centroid, flux, spectral balance
Tertiary	Cepstral	Cepstral Coefficients, MFCC
	LPC	LPC Coefficients, PLPC
	Other Tertiary	Gammatone Frequency Cepstral Coefficient (GFCC) and Power Normalized Coefficient (PNCC)
Voice Source	Voice Source	Jitter, shimmer, microprosody, NHR, HRN
Wavelets	Wavelets	Band-energies, Teager energy, modulation spectograms, RASTA, Gabor features, cortical features
Harmonic	Harmonic	Filtered sub-bands amplitude, correlogram
Zipf	Zipf	Entropy of inverse Zipf of frequency coding

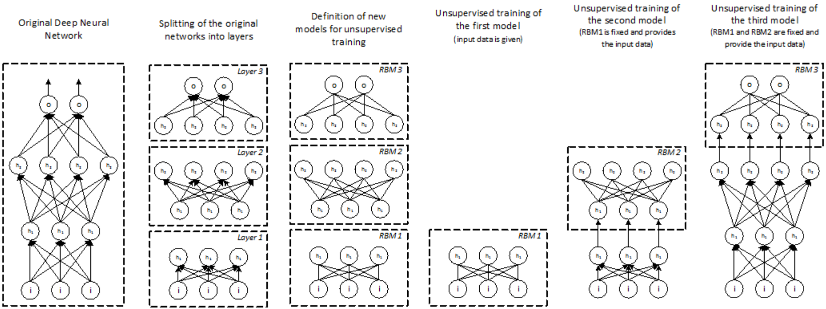
Pre-processing Pipeline



Disentangled Representations and Semantic Representations



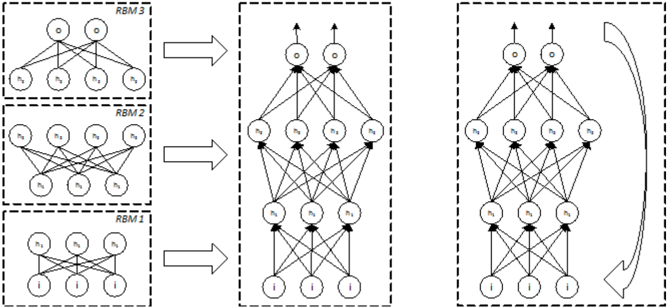
Deep Learning - Unsupervised Greedy Training



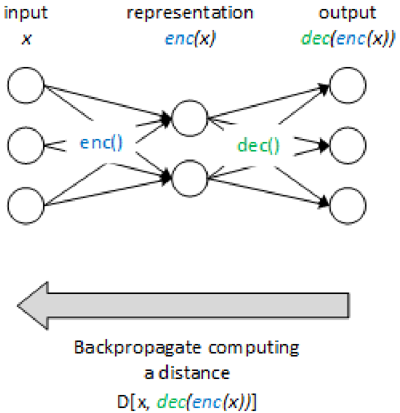
Deep Learning - Supervised Fine Tuning

Initialization of the original network using the values computed while training the unsupervised models

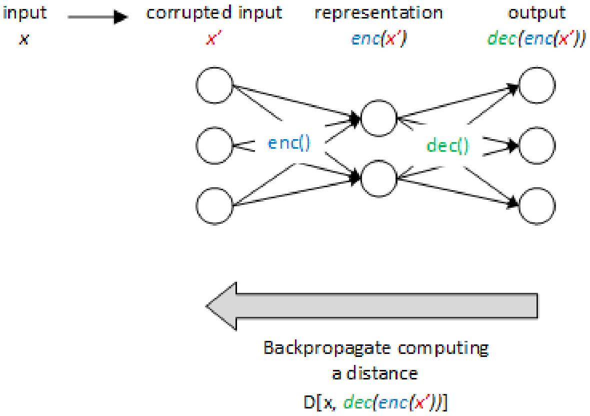
Supervised fine-tuning of the original network using a gradient-based back-propagation algorithm



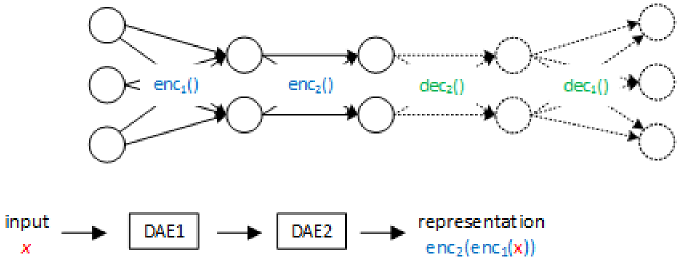
Autoencoders



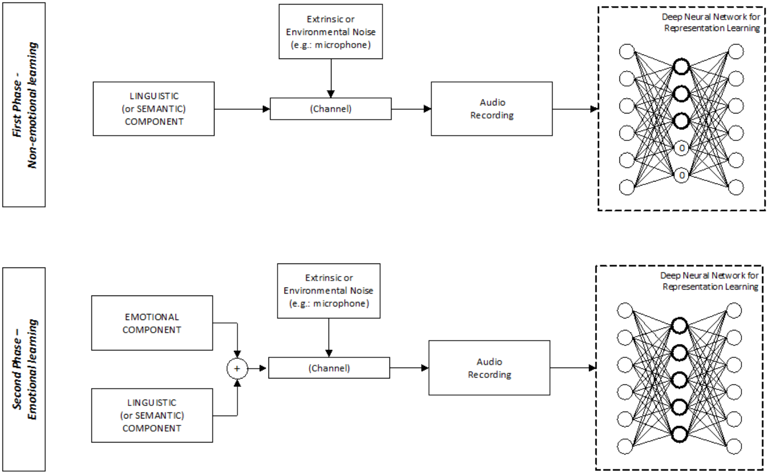
Denoising Autoencoders



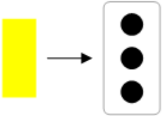
Stacked Denoising Autoencoders



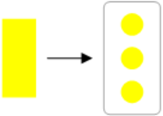
Contrastive Gradual Representation Learning - Idea



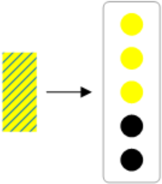
Contrastive Gradual Representation Learning - Algorithm (1)



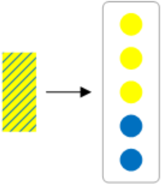
1. Present non-emotional samples to a DAE and train it.



2. The DAE learns to model non-emotional components.

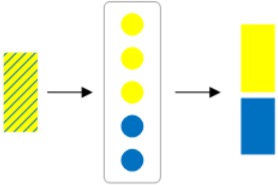


3. Present emotional samples to the DAE and train it.

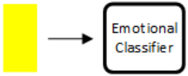


4. The DAE learns to model non-emotional and emotional components.

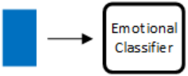
Contrastive Gradual Representation Learning - Algorithm (2)



5. Process new emotinal sample through the DAE and get two representations.



6. Perform emotional classification using the non-emotional representation.



7. Perform emotional classification using the emotional representation.

Information Theoretic Learning (1)

Several machine learning methods works through the optimization of a *loss function*.

Often, this loss function is defined over the *tacit assumption* that the error of the learned mapping function is *Gaussian*.

This leads to the definition of learning through the minimization of the second-order moment of the error (*MSE*)

Information Theoretic Learning (2)

Information theoretic learning drops the hypothesis of Gaussianity of the error and optimize information-theoretic estimators of the error.

For example, minimizing the entropy of the error (*MEE*), we can achieve the maximum transfer of information between the data and the model.

Information Theoretic Learning (3)

A core concept in information theoretic learning is the *quadratic information potential* estimator $I\hat{P}_2$.

- It is used as an *abstract descriptor* of probability distribution;
- It is defined starting from *Renyi's entropy*;
- It is more computationally friendly than Shannon's entropy;
- It is used to derive other *quadratic* theoretic information measures (distances and mutual informations).

Information Theoretic Learning (4)

Shannon's entropy:

$$H_S(X) = -\int p_X \ln p_X$$

Renyi's entropy:

$$H_\alpha(X) = \frac{1}{1-\alpha} \ln \int p_X^\alpha$$

Renyi's quadratic entropy:

$$H_2(X) = -\ln \int p_X^2$$

Information Theoretic Learning (5)

Estimated Renyi's quadratic entropy:

$$\hat{H}_2(X) = -\ln\left[\frac{1}{N^2}\sum\sum G_{\sigma\sqrt{2}}(x_i - x_j)\right]$$

Quadratic Information Potential Estimator:

$$\hat{IP}_2 = \frac{1}{N^2}\sum\sum G_{\sigma\sqrt{2}}(x_i - x_j)$$

Information Theoretic Representation Learning

Disentanglement may be learned through the maximization of the distance between the distribution of emotional and non-emotional representations.

Minimal Mutual Information (mMI) tries to learn independent distributions for emotional and non emotional samples by minimizing the *quadratic Euclidean distance* in the distribution space of the joint and the marginals.

Data Distribution Learning

Data Distribution Learning is the traditional approach to unsupervised learning in which, given data $\mathcal{D}$, we try to model the distribution of the process that generated $\mathcal{D}$.

Several mainstream algorithms: *Boltzmann machines*, *autoencoders*, *indipendent component analysis* [2].

Implicit assumption: learning the *true structure of the data* (i.e.: the statistical description of the process generating the data) will automatically provide a *useful* representation.

Feature Distribution Learning

Feature Distribution Learning is an innovative approach to unsupervised learning in which, given data $\mathcal{D}$, we try to model the distribution of the representation $\mathcal{R}$ in order to maximize its usefulness.

SF being the first algorithm of this kind [2].

Implicit assumption: some forms of representation are better than others and they will automatically provide a *useful* representation.

Sparsity

A *sparse* distribution, that is a distribution where most of the values are zero.

- *Practical reason:* sparse representation proved successful in many machine learning task (e.g.: *sparse deep belief networks* [**?**] or *k-sparse autoencoders* [**?**]);
- *Analogical reason:* biological systems implements sparse distributed representations (e.g.: modelling V1 cortex coding [**?**]);
- *Formal reason:* sparse distribution has low entropy ([**?**])

Sparse Filtering

SF achieve sparsity enforcing three properties:

1. *Population Sparsity:* each sample has few non-zero values;
2. *Lifetime Sparsity:* each feature has few non-zero values;
3. *High Dispersal:* activity on each row should be constant.

Given a dataset[1]:

$$\underbrace{\text{raw features}\left\{\begin{bmatrix} .3 & .4 & .3 & \cdots & .7 \\ .2 & .7 & .3 & \cdots & .3 \\ .3 & .8 & .5 & \cdots & .6 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ .2 & .1 & .8 & \cdots & .4 \end{bmatrix}\right.}_{\text{samples}} \xrightarrow{SF} \underbrace{\left.\begin{bmatrix} 0 & 0 & 0 & \cdots & .7 \\ 0 & 0 & 0 & \cdots & .6 \\ 0 & .7 & 0 & \cdots & 0 \\ 0 & .8 & 0 & \cdots & 0 \\ .9 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & .8 & \cdots & 0 \end{bmatrix}\right\}\text{SF features}}_{\text{samples}}$$

[1] notice the slightly unusual convention of having features along the rows and samples along the columns

SF Algorithm

Minimize the following *loss function*

$$\mathrm{argmin}_W \left\| \left\| \| f(WX) \|_{L2,row} \right\|_{L2,column} \right\|_{L1}$$

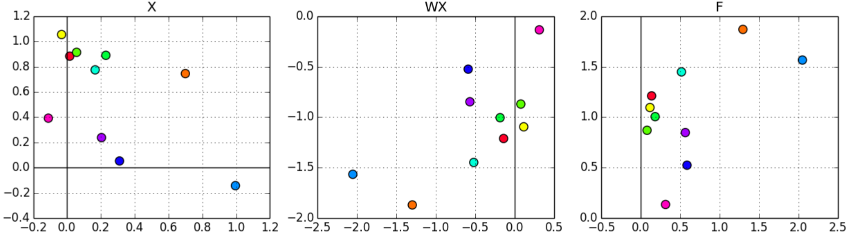
through *gradient descent*.

This ugly formula can be decomposed into four intuitive steps.

SF Algorithm - Step 1

Non-linear processing:

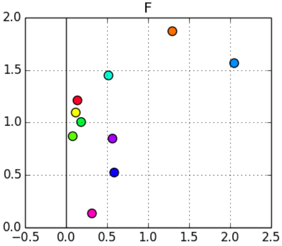
$$F = f(WX) = |WX|$$



SF Algorithm - Step 2

Normalization along the rows (features):

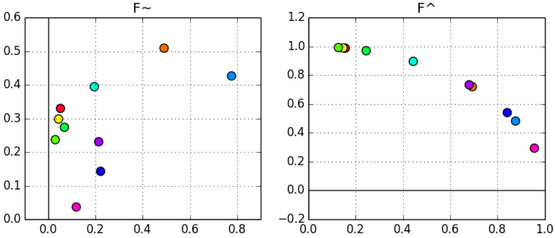
$$\tilde{F} = \frac{F}{\|F\|_{L2,row}}$$



SF Algorithm - Step 3

Normalization along the columns (samples):

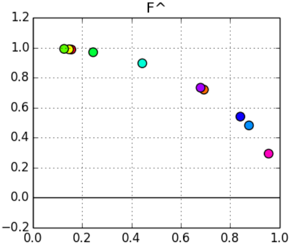
$$\hat{F} = \frac{\tilde{F}}{\left\|\tilde{F}\right\|_{L2,column}}$$

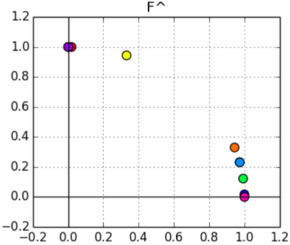


SF Algorithm - Step 4

Minimization of L1 norm:

$$\left\|\hat{F}\right\|_{L1}$$





Online Scenario

Scenario: test samples come in real-time and must be processed independently and efficiently.

Solution:

1. Learn from training data offline;
2. Estimate SF L_1 and L_2 parameters offline;
3. Process test data online normalizing it using the estimated parameters.

Questions:

- How unbiased are the estimates of the parameters?

Semi-Supervised Scenario

Scenario: training data is made up by a small set of labelled data and a large set of unlabelled data.

Solution:

1. Learn sparsity running SF on the unlabelled training set;
2. Save the learned weights;
3. Learn disentanglement running DSF on the labelled training set.

Questions:

- When do we stop the unsupervised learning?
- How to balances sparsity and disentanglement?

(Emotional) Disentangling Sparse Filtering

DSF achieve disentangling sparsity enforcing:

1. *Sparsity:* as in SF;
2. *Disentanglement:* non-emotional samples are represented in a lower dimensional space than emotional samples.

$$\textit{raw features}\left\{\left[\underbrace{\begin{matrix} .3 & .4 & .3\\ .2 & .7 & .3\\ .3 & .8 & .5\\ \cdots & \cdots & \cdots\\ .2 & .1 & .8 \end{matrix}}_{\substack{\textit{emo}\\ \textit{samples}}} \quad \begin{matrix} \cdots\\ \cdots\\ \cdots\\ \cdots\\ \cdots \end{matrix} \quad \underbrace{\begin{matrix} .7 & .3\\ .3 & .2\\ .6 & .9\\ \cdots & \cdots\\ .4 & .1 \end{matrix}}_{\substack{\textit{nem}\\ \textit{samples}}} \right]\right. \xrightarrow{DSF} \left[\begin{array}{c|c} A & B\\ \hline C & D \end{array}\right]$$

$$\text{where: } \underbrace{\begin{matrix} [A]\\ [C] \end{matrix}}_{\substack{\textit{emo}\\ \textit{sample}}} \qquad \underbrace{\begin{matrix} [B]\}\\ [D]\} \end{matrix}}_{\substack{\textit{nem}\\ \textit{samples}}} \begin{matrix} \textit{nem features}\\ \textit{emo features} \end{matrix}$$

DSF Algorithm

Loss functions for DSF:

$$\mathcal{L}_{DSF_D} = \left\| \left\| \left\| \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right\|_{L2,row} \right\|_{L2,column} \right\|_{L1} + \lambda_D \left\| D \right\|_{L1}$$

$$\mathcal{L}_{DSF_{AD}} = \left\| \left\| \left\| \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right\|_{L2,row} \right\|_{L2,column} \right\|_{L1} + \lambda_D \left\| D \right\|_{L1} + \lambda_A \left\| A \right\|_{L1}$$