Outline

- Cognitive component analysis: A definition
- Cognitive component analysis: A motivation
- Machine learning tools (ICA, sparse representations)
- Examples
  - Phonemes as cognitive components
  - Communities in social networks
  - Example: Grouping of mixed media data
- Theory: Aspects of generalizability in unsupervised learning
- Conclusion and outlook
Cognitive Component Analysis

A broad definition of cognition
- “The act or process of knowing - Cognition includes every mental process that may be described as an experience of knowing (including perceiving, recognizing, conceiving and reasoning) as distinguished from an experience of feeling and willing.”
  -Brittanica Online (2005)

Cognitive component analysis (COCA)
- The process of unsupervised grouping of data so that the resulting group structure is well-aligned with grouping based on human cognitive activity:

Cognitive compatibility as ”Micro Turing” tests....
Ecological modeling

Important for engineering proxies for human information processing...

Cf. efficient coding of “context-to-action” mapping
Cognitive component analysis (COCA)

- The process of unsupervised grouping of data so that the ensuing group structure is well-aligned with that resulting from human cognitive activity.

The “object” is a basic notion in cognitive psychology; E.g. “number of objects in short time memory”.

- Definition: “…there is a strong focus on objects as perceptual signaling units. A pragmatic definition of an object is a "signal source that maintains a minimum of independent behaviors in a given environment".
- Can cognitive component analysis be a step towards an general purpose definition of an “object”?

Theoretical issues: we are interested in the relation between supervised and unsupervised learning. Related to the discussion of the utility of unlabeled examples in supervised learning.
Many generalizations are possible – which ones will make sense to a human?
Cognitive modeling: Mental models

- Human cognition is often to act on weak signals, i.e., lack of information or poor signal to noise conditions.

- Mental models can be more or less well-aligned with actual physics/ecology, c.f. Friston et al.’s Predictive coding model

Qualia, sensory data

- Qualitative data often mapped with MDS multidimensional scaling: low-dimensional, neighbor preserving Euclidean representation.

- Austen Clark in Sensory Qualities (1993):
  - "The number of dimensions of the MDS space corresponds to the number of independent ways in which stimuli in that modality can be sensed to resemble or differ, but the dimensions per se have no meaning."
Gärdenfors’ “Conceptual Spaces”

- Cognitive models:
  - Symbolic, associative/connectionist, geometrical

- Human cognition ~ similarity judgments ~ Gestalt theory ~ geometrical proximity

- How to identify conceptual spaces, i.e., geometrical representations?
  (Gärdenfors, 2000)
Kemp-Tenenbaum – Discovery of structural form
(PNAS, 2008)

Human mind has only access to relatively low complexity modeling tools.
Cognitive Component Analysis: Why independence?

- Cognitive component analysis (COCA)
  - The process of unsupervised grouping of cognitive events so that the resulting group structure is well-aligned with manual grouping

- The **object** is a basic notion in cognitive psychology;
  - E.g. “modeling number of objects in short time memory”.
  - A pragmatic definition of **an object** is “a signal source that maintains a minimum of independent behavior in a given environment”.
  - Thus, independent component analysis could attain a key role in understanding cognition? (Hypothesis presented at AKRR, Helsinki 2005)

- Theoretical issues: we are interested in the relation between supervised and unsupervised learning. **How compatible are the hidden representations of supervised and unsupervised models?** Related to the discussion of the utility of unlabeled examples in supervised learning.
Cognitive compatibility

Unsupervised Learning

Hidden variable

\[ p(s | x, w_u) \propto p(x | s, w_u) p(s | w_u) \]

"Cognitive event": Data, sound, image, behavior

Supervised learning

\[ p(y \mid x, w_s) \]

"Cognitive" label, i.e. provided by a human observer

How well do these learned representations match: \( s = y \)?
When can COCA be expected to work?

- If “statistical structure” in the relevant feature space is well aligned with the label structure, we expect high **cognitive compatibility**.

- Unsupervised-then-supervised learning can explain “learning from single example”.

- Benign case, malign case, worst case....

Labels:
- “A”
- “B”
How will COCA help computers understand media content?

- Understand = simulate cognitive processing in humans
- Metadata estimation from media (sound/images/video/deep web objects)
- Basic signal processing tools are known (perceptual models...)
- ISO/MPEG standardization
Vector space representation

- Abstract representation - can be used for all digital media
- A “cognitive event” is represented as a point in a high-dimensional “feature space” – document similarity ~ spatial proximity in a given metric

- Text: Term/keyword histogram, N-grams
- Image: Color histogram, texture measures
- Video: Object coordinates (tracking), active appearance models
- Sound: Spectral coefficients, mel cepstral coefficients, gamma tone filters

*Contexts can be identified by their feature associations* ( = Latent semantics )
The independent context hypothesis (ICA)

- Challenge: Presence of multiple agentscontexts
- Need to "blindly" separate source signals = learn contexts
- Machine learning come to rescue!

\[ x(feature, time) = \sum_k A(feature, k) s(k, time) \]
Linear mixing generative model ICA - “Synthesis”
simplistic model incorporating sparsity and independence

Normal sources

Dense sources

Sparse sources

Space-time matrix
Component’s “where”
Vector of “what”

\[ x(loc, time) = \sum_k A(loc, k) s(k, time) \]
Protocol for comparing supervised and unsupervised learning

- Use the “unsupervised-then-supervised” scheme to implement a classifier:
  - Train the unsupervised scheme, e.g., ICA
  - Freeze the ICA representation (A matrix)
  - Train a simple (e.g., Naïve Bayes) classifier using the features obtained in unsupervised learning

- Compare with supervised classifier
  - Error rates of the two systems
  - Compare posterior probabilities

Research question: Can this simple linear model based on independence account for the pattern of human errors?
Phoneme classification

**Nasal vs oral:** “Esprit project ROARS” (Alinat et al., 1993)

Binary classification

Error rates: 0.23 (sup.), 0.22 (unsup.)

Bitrates: 0.48 (sup.), 0.39 (unsup.)
Cognitive components of speech

- Basic representation: Mel weighted cepstral coefficients (MFCCs)
- Modeling at different time scales 20 msec – 1000 msec
- Phonemes
- Gender
- Speaker identity

Figure 3: The latent space is formed by the two first principal components of data consisting of four separate utterances representing the sounds ‘s’, ‘o’, ‘t’, ‘a’. The structure clearly shows the sparse component mixture, with ‘rays’ emanating from the origin (0,0). The ray embraced in a rectangle contains a mixture of ‘s’ and ‘t’ features, a cognitive component associated with the vowel /o/ sound.
**Error rate comparison**
For the given time scales and thresholds, data locate around $y = x$, and the correlation coefficient $\rho = 0.67, p < 1.38 \times 10^{-9}$.

**Sample-to-sample correlation**
- Three groups: vowels eh, ow; fricatives s, z, f, v; and stops k, g, p, t.
- 25-d MFCCs; EBS to keep 99% energy; PCA reduces dimension to 6.
- Two models had a similar pattern of making correct predictions and mistakes, and the percentage of matching between supervised and unsupervised learning was 91%.
Longer time scales

Time integrated (1000ms) MFCC’s: text independent speaker recognition....

Feng & Hansen (CIMCA, 2005)
Error rate correlations super/unsupervised learning for different cognitive time scale events (phoneme, gender, height, speaker identity)

Fig. 4. Figure shows test error rates of both supervised and unsupervised learning on four topics: phonemes, gender, height, and identity. Solid lines indicate \( y = x \) in the coordinate systems. All data located along this line, meaning high correlation between supervised and unsupervised learning.
Digital media: Vector space representation

- Abstract representation - can be used for all digital media
- Document is represented as a point in a high-dimensional "feature space" – document similarity ~ spatial proximity in a given metric

- Text: Term/keyword histogram, N-grams
- Image: Color histogram, texture measures
- Video: Object coordinates (tracking), active appearance models
- Sound: Spectral coefficients, cepstral coefficients, gamma tone filters

*Contexts can be identified by their feature associations (= Latent semantics)*

Latent semantics

- Document features are correlated, the pattern of correlation reflects "associations".
- Associations are context specific
- Word sets are activated in concert in a given context
  ape ~ zoo, zoo ~ elephant => ape ~ elephant

*Latent semantic analysis: Contexts can be identified by term co-variance patterns (PCA)*

Linear mixture of independent agents in term-document scatterplots.

Linear mixture of independent contexts observed in short time features (mel-cepstrum) in a music database.
Social networks: Linear mixtures of independent communities?

"Movie actor network"
- A collaborative small world network
  128,000 movies
  380,000 actors

Genre patterns in expert’s opinion on similar music artists

(AMG400, Courtesy D. Ellis)
Independent contexts in document data bases

- \( x(j,t) \) is the occurrence of the \( j \)'th word in the \( t \)'th document.

- \( s(k,t) \) quantifies how much the \( k \)'th context is expressed in \( t \)'th document.

- \( A(j,k) \) quantifies the typical importance of the \( j \)'th word in the \( k \)'th context

PCA vs ICA document scatterplots
Independent contexts in dynamic text:

Chat room analysis

We logged a day's chat in a CNN "news cafe".

The database involves 120 users chatting during an 8 hour period.

```
<html>
<body>
<p>few weeks.</p>
<p><b>MieZ</b> hey seagate</p>
<p><b>Recycle</b> denise: he deserved it for stealing os code in his early days</p>
<p><b>Zeno</b> ok Sharonelle</p>
<p><b>deny</b> LOL @ Recycle</p>
<p><b>heartattackagain</b> Ed Shore...lol...We might have an operating system that doesn't crash every thirty minutes....lololol......</p>
<p><b>EdShore</b> Shoooby, I don't believe you, I've been doing this sine PET, TRS-80, and PIRATES! Don't tell me you've been CHATTING! PROVE IT!</p>
<p><b>Zeno</b> Recycle LOL ethical and criminal laws are different for the business world</p>
<p><b>Seagate</b> Recycle, thats what the technology business is all about.</p>
<p><b>tribe</b> I heard a local radio talk show host saying last night that he has noticed everytime this Etan issue slows down, something happens to either the family in Miami or in Cuba to put it right back in the headlines. He mentioned the cousin's hospitalization as just the latest saga</p>
<p><b>Diogenes</b> If Bill Gates was in Silicon Valley never a word would you have ever heard.</p>
<p><b>Zeno</b> EdC you may have been doing sbe but I have been doing cosine.</p>
<p><b>shooby</b> EdShore: Compuserve since, heck, 76?</p>
<p><b>Zeno</b> I mean EdShore</p>
<p><b>Recycle</b> rumor has it that he was even dumpster diving at school for soda</p>
</body>
</html>
```
ICA by dynamic decorrelation (Molgedey-Schuster, Kolenda et al. 2001)

The Bayes factor - $P(M|D)$ - of each model is estimated in the BIC approximation.

Source autocorrelations
Chat room analysis

Keywords from prototype histograms (A-columns)

Topic 1: Chat, join, pm, cnn, board,...
Topic 2: Gun show
Topic 3: Susan, Smith, mother, children, kid, life
Topic 4: People, census, elian, state, clinton,...

“Topic 5” is a rest group
Example: Independent contexts in multi-media

- Organizing webpages in categories
- Labels obtained from Yahoo’s directory
- Features: Text, color, and texture subsets of MPEG image features

For each text/image 1 .. N sample

\[ X^T = \begin{bmatrix} \text{Word histogram} & \text{Gabor Texture histogram} & \text{HSV Color histogram} \end{bmatrix} \]

3591 192 768

Feature / document matrix

Performance of the system trained by associating unsupervised independent components with labels – generalization based on Yahoo categories

<table>
<thead>
<tr>
<th>Modality</th>
<th>Classification Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>23.0%</td>
</tr>
<tr>
<td>Texture</td>
<td>18.0%</td>
</tr>
<tr>
<td>Texture/Color</td>
<td>11.5%</td>
</tr>
<tr>
<td>Text</td>
<td>5.7%</td>
</tr>
<tr>
<td>Combined (texture/color/text)</td>
<td>2.8%</td>
</tr>
</tbody>
</table>

Fig. 3. Scatterplots of the text and image multimedia data, projected to a two-dimensional subspace found by PCA. Grey value of points corresponds to the three classes considered, see Fig. 4. The ray-like structure strongly suggest an ICA interpretation, however, the relevance of this representation can only be determined by a subsequent inspection of the recovered source signals. As we will see in section 4.6, it turns out that there is an interesting alignment of the source signals and a manual labeling of the multimedia documents.
**Performance**

Combined error rate: 2.8%
Single best error rate: 5.7%

<table>
<thead>
<tr>
<th>Image</th>
<th>Label</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1$</td>
<td>Sports</td>
<td>position college weight born lbs height guard</td>
</tr>
<tr>
<td>$I_2$</td>
<td>Aviation</td>
<td>na air convair wing</td>
</tr>
<tr>
<td>$I_3$</td>
<td>Paintball</td>
<td>check darkside force gog strike odt</td>
</tr>
</tbody>
</table>
"Brede" -- tools for a neuroimaging search engine

- Exponential growth in publications/images/data
- Distributed (www) heterogeneous databases
- Multimedia facilities needed
- Tools:
  - Novelty detection (Nielsen&Hansen, HBM 2002)
  - Finding similar volumes (Nielsen&Hansen, AIM 2003)
  - Brede toolbox

Primary co-worker: Finn Årup Nielsen
...consult his homepage for a neuroinformatics informed search engine
Analytic search example: “The posterior cingulate cortex”

- cyto-architecturally well defined brain region (Vogt et al, 2001)

- no “consensus” about its function: Several functions are reported including involvement in emotion, pain, memory etc
Materials

- 245 abstracts of functional imaging studies that responded to PubMed query (March 13th, 2003):

  ("posterior cingulate" OR "posterior cingulum" OR "retrosplenial" OR "retrosplenium") AND
  ("magnetic resonance imaging" OR "positron emission tomography")

We find independent components of abstracts and locate them using the Brede database: Two components are rich in terms related to "memory" and "emotion". Localization may actually be slightly different pointing to a regional specialization.
Linear independent components (Factor) models seem to be relevant to many cognitive ecologies

Let us take a look under the hood...

- Can we understand swift human learning?

- ...a closer look at generalizability of unsupervised learning in factor models
Factor models

- Represent a datamatrix by a low-dimensional approximation

\[ X(t, l) = \sum_{k=1}^{K} A(l, k) S(k, t) \]
Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of \( m = 2,429 \) facial images, each consisting of \( n = 19 \times 19 \) pixels, and constituting an \( n \times m \) matrix \( V \). All three find approximate factorizations of the form \( V = WH \), but with three different types of constraints on \( W \) and \( H \), as described more fully in the main text and methods. As shown in the 7 \( \times \) 7 montages, each method has learned a set of \( r = 49 \) basis images. Positive values are illustrated with black pixels and negative values with red pixels. A particular instance of a face, shown at top right, is approximately represented by a linear superposition of basis images. The coefficients of the linear superposition are shown next to each montage, in a 7 \( \times \) 7 grid, and the resulting superpositions are shown on the other side of the equality sign. Unlike VQ and PCA, NMF learns to represent faces with a set of basis images resembling parts of faces.

Learning the parts of objects by non-negative matrix factorization

Daniel D. Lee* & H. Sebastian Seung**†

* Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA
† Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
Generative model for hidden variables

\[ x = As + \varepsilon, \quad \varepsilon \sim N(0, \Sigma) \]

\[
p(x \mid A, \theta) = \int p(x \mid A, s, \Sigma)p(s \mid \theta)ds
\]

\[
p(x \mid A, s, \Sigma) = \sqrt{2\pi\Sigma}^{-1/2} e^{-\frac{1}{2}(x-As)^T\Sigma^{-1}(x-As)}
\]

Source distribution:
- PCA: ... normal
- ICA: ... other
- IFA: ... Gauss. Mixture

PCA: \[ \Sigma = \sigma^2 \cdot 1 \]

FA: \[ \Sigma = D \]

S known: GLM
(1-A^{-1}) sparse: SEM
S,A positive: NMF

Højen-Sørensen, Winther, Hansen, Neural Comp (2002), Neurocomputing (2002)
Modeling the generalizability of FA

- **Rich physics literature on “retarded” learning**
- **Universality**
  - Generalization for a “single symmetry breaking direction” is a function of ratio of $N/D$ and signal to noise $S$
  - For subspace models-- a bit more complicated -- depends on the component SNR’s and eigenvalue separation
  - For a single direction, the mean squared overlap $R^2 = \langle (u^T_1 u_0)^2 \rangle$ is computed for $N,D \to \infty$

$$R^2 = \begin{cases} 
(\alpha S^2 - 1) / S(1 + \alpha S) & \alpha > 1 / S^2 \\
0 & \alpha \leq 1 / S^2 
\end{cases}$$

$$\alpha = N / D \quad S = 1 / \sigma^2 \quad N_c = D / S^2$$


$N_c = (0.0001, 0.2, 2, 9, 27, 64, 128, 234, 400, 625)$

$\sigma = (0.01, 0.06, 0.12, 0.17, 0.23, 0.28, 0.34, 0.39, 0.45, 0.5)$
Generalization in unsupervised learning

- Looking for universality by simulation
  - learning two clusters in white noise.

- Train $K=2$ component factor models.

- Measure overlap between line of sight and plane spanned by the two factors.

Experiment

Variable: $N$, $D$
Fixed: SNR
Restoring the generalizability of FA

- Now what happens if you are on the slope of generalization, i.e., N/D is just beyond the transition to retarded learning?

- The estimated projection is offset, hence, future projections will be too small!

- ...problem if discriminant is optimized for unbalanced classes in the training data!
Heuristic: Leave-one-out re-scaling of SVD test projections

Kjems, Hansen, Strother: “Generalizable SVD for Ill-posed data sets” NIPS (2001)
Re-scaling the component variances

Possible to compute the new scales by leave-one-out doing N SVD’s of size N << D

Compute \( \mathbf{U}_0 \mathbf{\Lambda}_0 \mathbf{V}_0^T = \text{svd}(X) \) and \( \mathbf{Q}_0 = \begin{bmatrix} q_j \end{bmatrix} = \mathbf{\Lambda}_0 \mathbf{V}_0^T \)

foreach \( j = 1...N \)

\[
\bar{q}_{-j} = \frac{1}{N-1} \sum_{j' \neq j} q_{j'}
\]

Compute \( \mathbf{B}_{-j} \mathbf{\Lambda}_{-j} \mathbf{V}_{-j}^T = \text{svd}(\mathbf{Q}_{-j} - \bar{Q}_{-j}) \)

\( z_j = \mathbf{B}_{-j} \mathbf{B}_{-j}^T (q_j - \bar{q}_{-j}) \)

\( \hat{\lambda}_i^2 = \frac{1}{N-1} \sum_j z_{ij}^2 \)
What about other factorizations?

- NMF can be adjusted similarly using histogram equalization

- Note the re-scaling problem is almost trivial if the classifier is Naïve Bayes-like (threshold adaption) taking the independent components as features
- More complex classifiers may need more coordination
Conclusions

- Evidence that phonemes, gender, identity are independent components ‘objects’ in the (time stacked) MFCC representation

- Evidence that human categorization is based on sparse independent components in social networks, text, digital media

- Conjecture: that objects in digital media can be identified as independent components: The brain uses old tricks from perception to solve complex “modern” problems.
Outlook

- Compatibility in social networks
  - Are links and features compatible?

- The independent context hypothesis:
  - Are the Gestalt laws simply results of ICA representations?
  - Next step to understand dynamics
Acknowledgments

• Danish Research Councils
• EU Commission
• NIH Human Brain Project grant (P20 MH57180)
References

