MTTTS16 Learning from Multiple Sources

5 ECTS credits

Autumn 2019, University of Tampere
Lecturer: Jaakko Peltonen

Lecture 10:
Learning Sample Correspondence

On this lecture:

- How to use multi-view setups that require paired samples, when the pairing is not known

Canonical Correlation Analysis when data pairing across views is unknown

Canonical correlation analysis without known pairs

- **Matching samples between two data sets** is a task that appears in several kinds of applications:
  - For example, in bioinformatics, gene activities are measured with microarrays of different types, manufacturers etc. Each array contains several “probes” for short DNA sequences (activity is measured by how much genetic material binds to each probe). Sometimes it can be unclear which probes match each other between different microarrays.
  - Cellular activity such as metabolic activity (presence of metabolites) or activity of genes can be measured for different species, and sometimes we need to find corresponding metabolites / genes across species.
  - When analyzing content of webpages or text articles we may need to match photos with their textual descriptions, which may be somewhere in the page/article.
  - In translated documents (e.g. EU laws), which sentence from language A matches which in language B? Order and number of sentences may depend on language.

Following the approach from Tripathi, A., Klami, A., Virpioja S. **Bilingual sentence matching using kernel CCA.** In proceedings of MLSP 2010, International Conference on Machine Learning for Signal Processing. Images from that paper.

Canonical correlation analysis without known pairs

- One way to solve the matching is to follow this principle: “The correct matching results in the **highest statistical dependency** between the two collections.”
- Statistical dependency could be measured e.g. by **mutual information**, or by simpler measures such as **correlation** which notices some (but not all) dependencies.
- Even with a perfect matching of samples, not everything between the two data sets (two views) may be correlated: it is enough to find at least some **subspace** of both data sets that becomes highly correlated when the matching is successful
- Canonical correlation analysis (CCA) could be used for this purpose!

- This shows that CCA could be useful for matching. What about the other way around, can matching be useful for CCA?

Following the approach from Tripathi, A., Klami, A., Virpioja S. **Bilingual sentence matching using kernel CCA.** In proceedings of MLSP 2010, International Conference on Machine Learning for Signal Processing. Images from that paper.

Canonical correlation analysis without known pairs

- Canonical correlation analysis (CCA) tries to find a feature subspace (linear projection) from each of two views so that the projections into the subspace are as correlated as possible.
- CCA requires that samples arrive in pairs (x1,x2) where each sample has a known feature vector in both views. Essentially this means each sample has a large number of features which are divided into two feature sets.
- In some applications, there are two views, but samples do not come with known features in both views, only for one view or the other.
- Sometimes we can assume that each sample (having a feature in view 1) has some pair in the other view (view 2) but we just don't know it.
- If we have a large collection of samples from both views, can we try to find a matching pair for each sample from one view to the other? If we can, then we can run CCA as before!

Following the approach from Tripathi, A., Klami, A., Virpioja S. **Bilingual sentence matching using kernel CCA.** In proceedings of MLSP 2010, International Conference on Machine Learning for Signal Processing. Images from that paper.

Canonical correlation analysis without known pairs

• Matching samples, and CCA, are useful for each other - they make each other possible.

• Sometimes external annotation is available to help the task; sometimes "vague" information like priors about the matching is available; sometimes we can only use the data sets themselves.

• In general, given two data sets $\mathbf{X} \in \mathbb{R}^{N \times D_x}$ and $\mathbf{Y} \in \mathbb{R}^{M \times D_y}$, we want to find a **permutation** of the samples, **p**, so that the i:th sample in **X** is matched with sample $\mathbf{p}(i)$ in **Y**.

• We try to find lower-dimensional mappings $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{y})$ to maximize the dependency

$$\max_{\mathbf{p},\mathbf{f},\mathbf{g}} \quad \mathrm{Dep}\left(\mathbf{f}(\mathbf{X}), \mathbf{g}(\mathbf{Y}(\mathbf{p}))\right)$$

with respect to the permutation **p** and the mappings **f** and **g**, where $\mathrm{Dep}(\cdot,\cdot)$ is some measure of dependency, and $\mathbf{Y}(\mathbf{p}) \in \mathbb{R}^{N \times D_y}$ denotes **Y** with rows (samples) permuted according to **p**

Following the approach from Abhishek Tripathi, Arto Klami, and Samuel Kaski. **Using dependencies to pair samples for multi-view learning**. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1561–1564. Images from that paper.

Canonical correlation analysis without known pairs

- The dependency measure Dep and parameterization of the mappings **f** and **g** can be chosen freely.
- We use Pearson correlation for Dep and linear projections for **f** and **g**: $\mathbf{f}(\mathbf{x}) = \mathbf{x}\mathbf{W}_x$, $\mathbf{g}(\mathbf{y}) = \mathbf{y}\mathbf{W}_y$
- Then the optimization problem becomes

$$\max_{\mathbf{p},\mathbf{W}_x,\mathbf{W}_y} \operatorname{corr}\left(\mathbf{X}\mathbf{W}_x, \mathbf{Y}(\mathbf{p})\mathbf{W}_y\right)$$

- This can be solved iteratively. First, keep the projections fixed, and maximize with respect to the permutation: with a finite sample set this becomes

$$\max_{\mathbf{p}} \frac{\mathbf{W}_x^T \mathbf{X}^T \mathbf{Y}(\mathbf{p}) \mathbf{W}_y}{\|\mathbf{X}\mathbf{W}_x\| \|\mathbf{Y}(\mathbf{p})\mathbf{W}_y\|}$$

- The numerator of the above can be written in terms of distances between projected samples. The denominator is constant if M = N (same number of samples in both sets) and can be approximated as constant even if M is slightly larger than N.

Following the approach from Abhishek Tripathi, Arto Klami, and Samuel Kaski. **Using dependencies to pair samples for multi-view learning**. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1561–1564. Images from that paper.

Canonical correlation analysis without known pairs

- Then the optimization of the permutation becomes

$$\min_{\mathbf{p}} \sum_{i=1}^{N} \|\mathbf{x}_i \mathbf{W}_x - \mathbf{y}_{\mathbf{p}(i)} \mathbf{W}_y\|^2$$

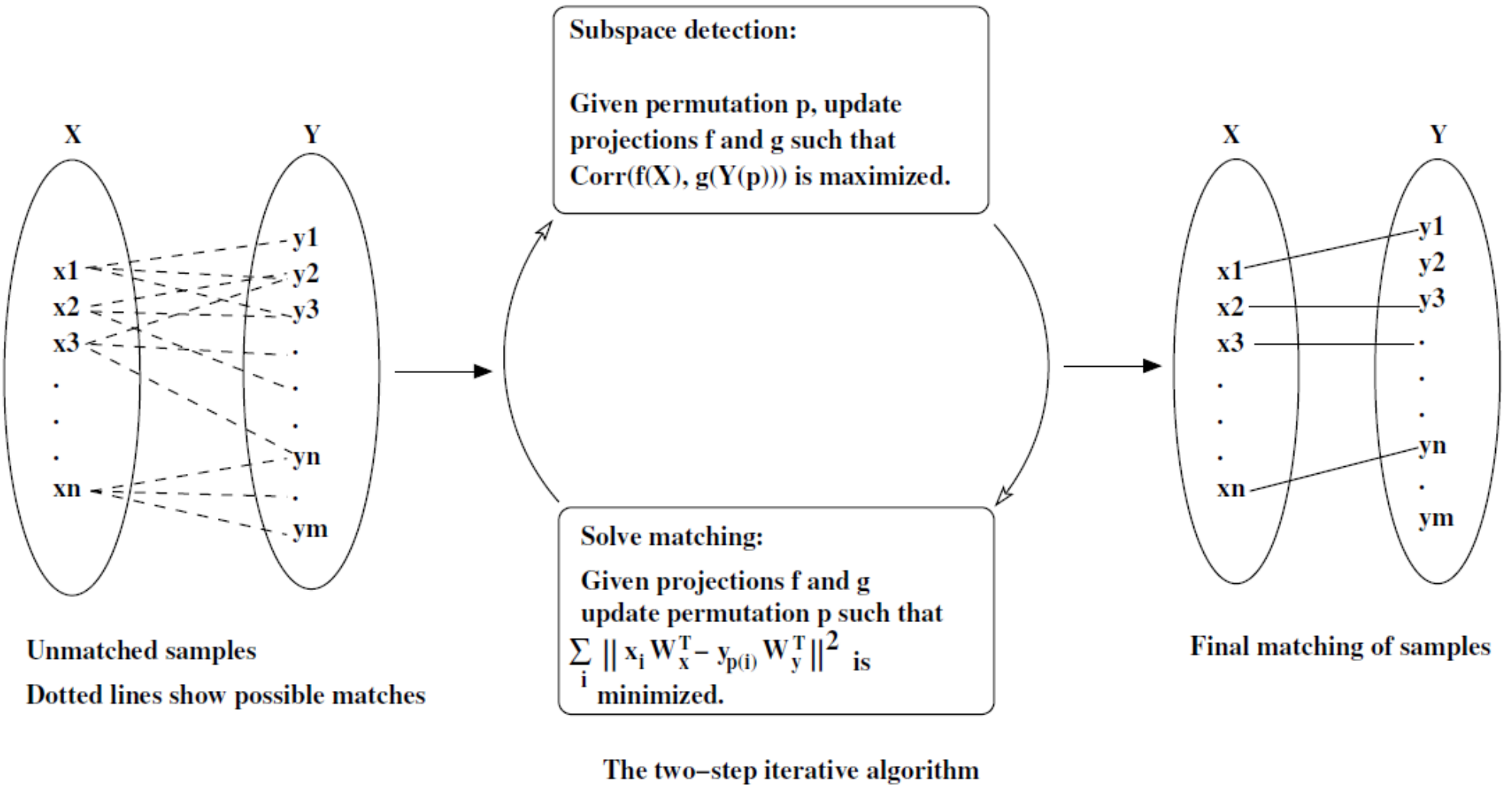
which tries, for each sample $\mathbf{x}_i$, to find the sample $\mathbf{y}_{\mathbf{p}(i)}$ whose projection $\mathbf{y}_{\mathbf{p}(i)} \mathbf{W}_y$ is as close as possible to the projection $\mathbf{x}_i \mathbf{W}_x$

- If each sample **y** can be used once, the above is an **assignment problem** which can be solved exactly with e.g. the Hungarian algorithm. (http://en.wikipedia.org/wiki/Hungarian_algorithm)
- Given a fixed pairing, the projections can then be solved by canonical correlation analysis (CCA) as on lecture 1.
- These two steps can be combined into an **iteration**: 1. start with a random pairing, 2. find optimal projections for that pairing by CCA. 3. find a new pairing (solve the assignment problem). Repeat steps 2. and 3. until convergence.

Following the approach from Abhishek Tripathi, Arto Klami, and Samuel Kaski. **Using dependencies to pair samples for multi-view learning**. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1561–1564. Images from that paper.

Canonical correlation analysis without known pairs

- Illustration of the basic idea of the method:



Following the approach from Tripathi, A., Klami, A., Virpioja S. **Bilingual sentence matching using kernel CCA.** In proceedings of MLSP 2010, International Conference on Machine Learning for Signal Processing. Images from that paper.

Canonical correlation analysis without known pairs

- One detail is: when we extract more than one CCA component in the CCA step, we have to decide how important each component is in the matching step: we must decide the weight (or scale) of each component, when computing the Euclidean distance between samples projected to the CCA components.
- CCA does not fix scales of the features (correlation is the same regardless of the overall scale of the projection).
- One option is to ignore this and use a uniform scale.
- A better option is to weight each component by the corresponding correlation value.

- Another detail: if we know prior information about which pairs are possible, we can take that into account when finding the matching: disallow matches p(i) where p(i) is known to be an impossible match for i.

Following the approach from Abhishek Tripathi, Arto Klami, and Samuel Kaski. **Using dependencies to pair samples for multi-view learning**. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1561–1564. Images from that paper.

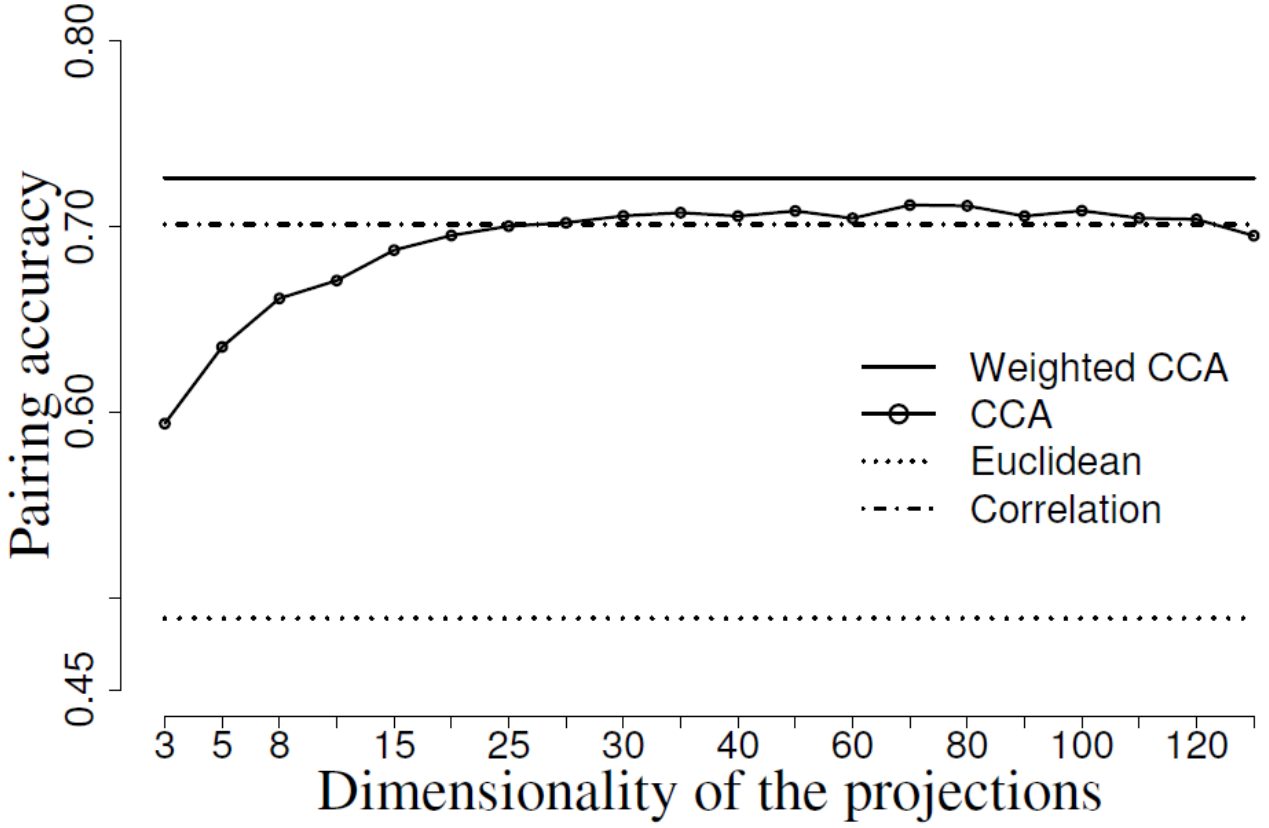
Canonical correlation analysis without known pairs

- **Experiment:** let's try to find matches between probes of two different microarray types (probe sets), Affymetrix HGU-95 and Affymetrix HGU-133.
- Each probe from HGU-95 is a sample **x**, and each probe from HGU-133 is a sample **y**.
- For each HGU-95 probe (sample) **x**, and for each HGU-133 probe (sample) **y**, the features of the sample are activity of the probe across a set of tissues from patients with acute lymphoblastic leukemia (ALL).
- For these two microarray types, the company Affymetrix provides a comparison sheet providing “true pairs” between HGU-95 and HGU-133. We can use this to evaluate the accuracy of the matches we find.

Following the approach from Abhishek Tripathi, Arto Klami, and Samuel Kaski. **Using dependencies to pair samples for multi-view learning**. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1561–1564. Images from that paper.

Canonical correlation analysis without known pairs

- If we pair samples randomly, we get accuracy 0.31.
- With the iteration of CCA and matching, we can do better.
- We will try different amounts of extracted CCA components for use in pairing.



Following the approach from Abhishek Tripathi, Arto Klami, and Samuel Kaski. **Using dependencies to pair samples for multi-view learning**. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1561–1564. Images from that paper.

Canonical correlation analysis without known pairs

- Variants: if we want to get more general dependencies than simple canonical correlations, we can replace the CCA step with some other mapping from original features of both data sets to a lower-dimensional space where the pairs have high dependency.
- For example, instead of CCA (lecture 1) we can use kernel CCA (lecture 4), which is computed based on kernel matrices $\mathbf{K_X}(i,j) = \mathbf{k_x}(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{K_Y}(i,j) = \mathbf{k_y}(\mathbf{y}_i, \mathbf{y}_j)$
- as on lecture 4, kernel CCA solves

$$\underset{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i \in \mathbb{R}^N}{\operatorname{argmax}} \; \boldsymbol{\alpha}_i \mathbf{K_X} \mathbf{K_{Y(p)}} \boldsymbol{\beta}_i^T$$

$$\text{subject to} \;\; \boldsymbol{\beta}_i \mathbf{K_{Y(p)}} \mathbf{K_{Y(p)}} \boldsymbol{\beta}_i^T = 1, \;\; \boldsymbol{\alpha}_i \mathbf{K_X} \mathbf{K_X} \boldsymbol{\alpha}_i^T = 1$$

- Then the matching step can be rewritten as

$$\min_{\mathbf{p}} \sum_{i=1}^{N} \| \mathbf{K}_{\mathbf{X}}^i \boldsymbol{\mathcal{A}} - \mathbf{K}_{\mathbf{Y(p)}}^i \boldsymbol{\mathcal{B}} \|^2$$

- To avoid overlearning, kernel matrices can be regularized as $\mathbf{K_X} + \gamma \mathbf{I}$ for some multiplier gamma.

Following the approach from Tripathi, A., Klami, A., Virpioja S. **Bilingual sentence matching using kernel CCA.** In proceedings of MLSP 2010, International Conference on Machine Learning for Signal Processing. Images from that paper.

Canonical correlation analysis without known pairs

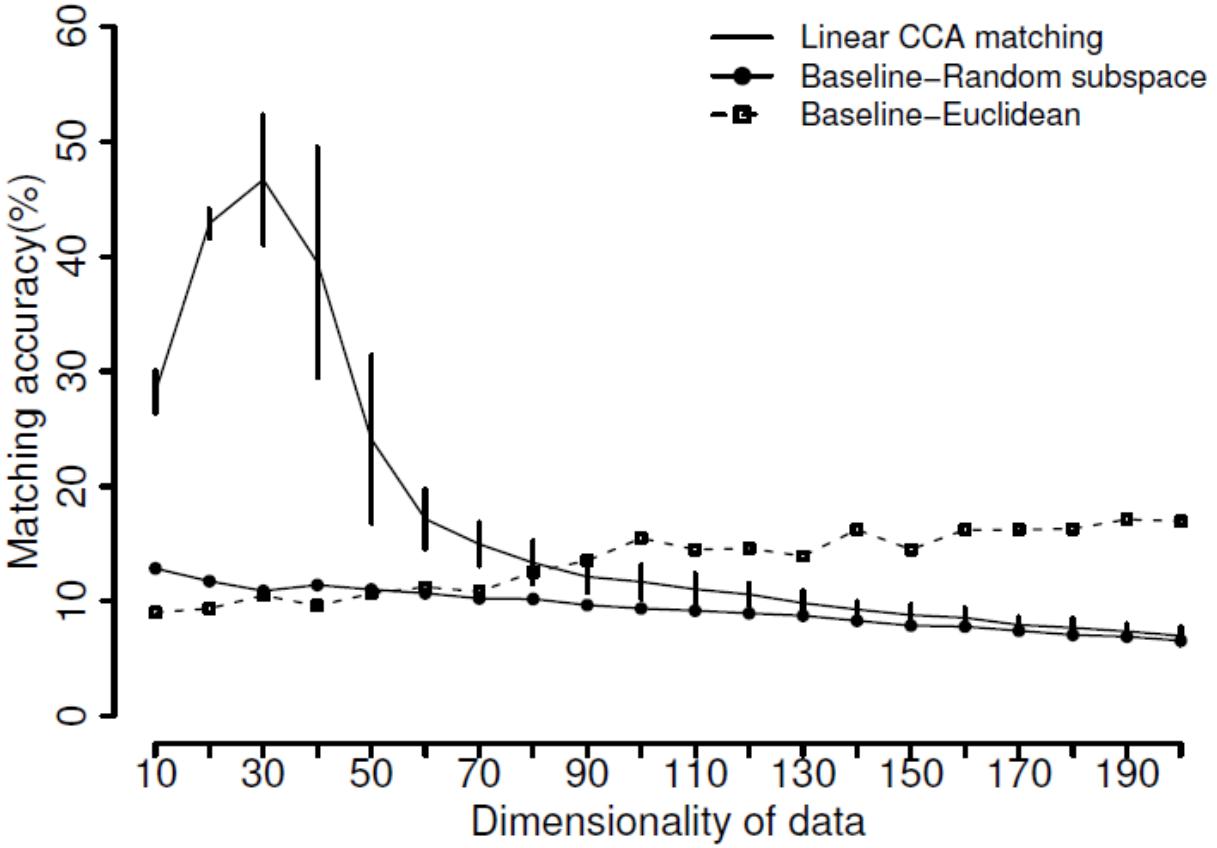
- **Experiment:** let's try to find matches between part of the "Europarl" corpus: proceedings of the European Parliament from September 2003 in Finnish and English, yielding 21358 sentences over 8 days.
- The sentences contain 496044 tokens (words etc.) in English and 382866 in Finnish; we reduce the high-dimensional spaces by principal component analysis (PCA) to 200 for each data set:
- The result is a 21358x200 matrix for Finnish and a 21358x200 matrix for English, with unknown pairing.
- We train on 1266 sentences from the first day, and test on the rest.
- We try using different amounts of the PCA components.
- The Europarl corpus includes its own sentence alignment tool, we use its results as the "true matches" for performance evaluation.

Following the approach from Abhishek Tripathi, Arto Klami, and Samuel Kaski. **Using dependencies to pair samples for multi-view learning**. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1561–1564. Images from that paper.

Canonical correlation analysis without known pairs

• Result: first try linear CCA. Comparisons: match in a random subspace, or match without finding a subspace using all given PCA features.

(technically, prior information about possible matches was included as a penalty for matching very far-off sentences like matching 1st English sentence to 10th Finnish sentence. Details omitted.)



• Next try kernel CCA: improves further.

Linear matching	$58.4 \pm 2.4\%$
Kernel matching	$61.1 \pm 2.2\%$

Following the approach from Abhishek Tripathi, Arto Klami, and Samuel Kaski. **Using dependencies to pair samples for multi-view learning**. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1561–1564. Images from that paper.

References

- Abhishek Tripathi, Arto Klami, and Samuel Kaski. **Using dependencies to pair samples for multi-view learning**. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1561–1564.
- Tripathi, A., Klami, A., Virpioja S. **Bilingual sentence matching using kernel CCA.** In proceedings of MLSP 2010, International Conference on Machine Learning for Signal Processing.