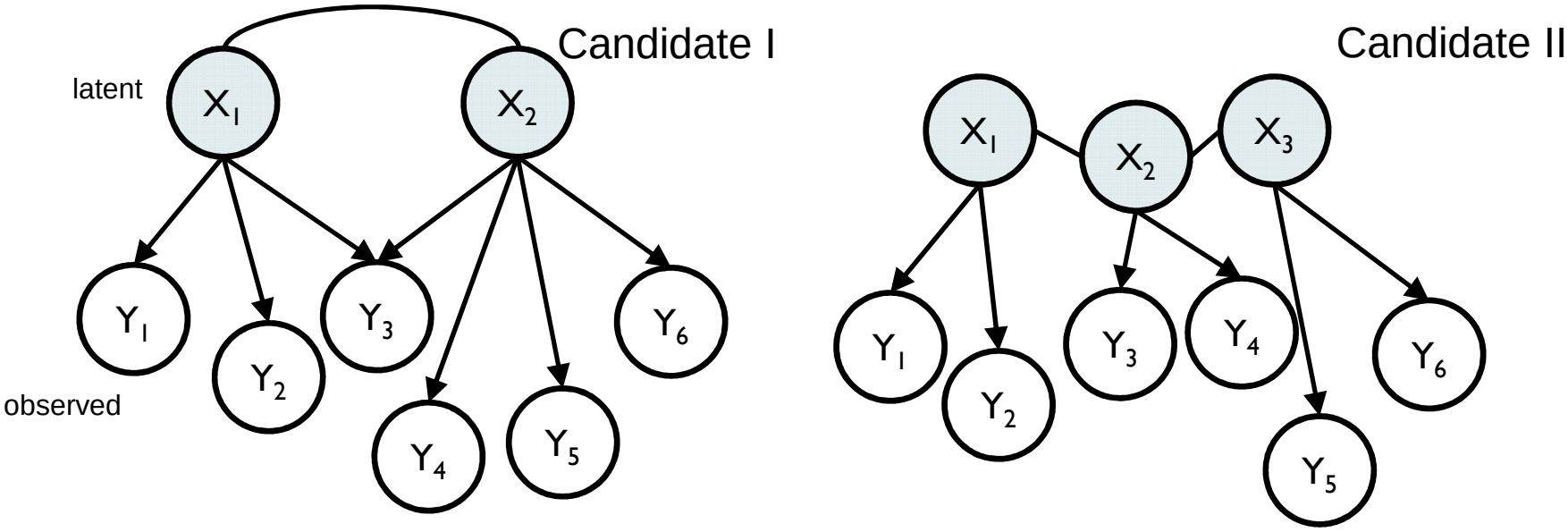
Latent Composite Likelihood Learning for the Structured Canonical Correlation Model

Ricardo Silva
University College London
ricardo@stats.ucl.ac.uk

UAI 2012 – Avalon, CA

Learning Latent Structure



- Difficulty on computing scores or tests
- Identifiability: theoretical issues and implications to optimization

Leveraging Domain Structure

- Exploiting "main" factors

YOUR JOB AND ORGANISATION

7. To what extent do you agree or disagree with the following statements about your immediate manager?

My immediate manager...

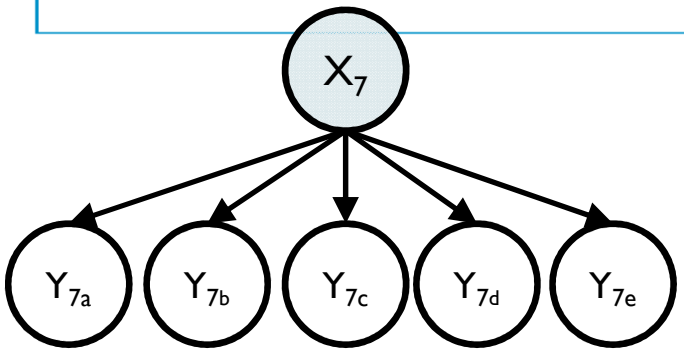
a. ...encourages those who work for her/him to work as a team.

b. ...can be counted on to help me with a difficult task at work.

c. ...gives me clear feedback on my work.

d. ...asks for my opinion before making decisions that affect my work.

e. ...is supportive in a personal crisis.

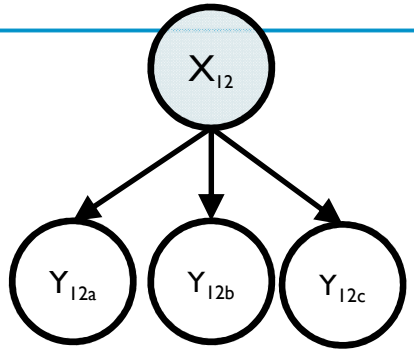


12. To what extent do you agree or disagree with the following statements?

a. I often think about leaving this Trust.

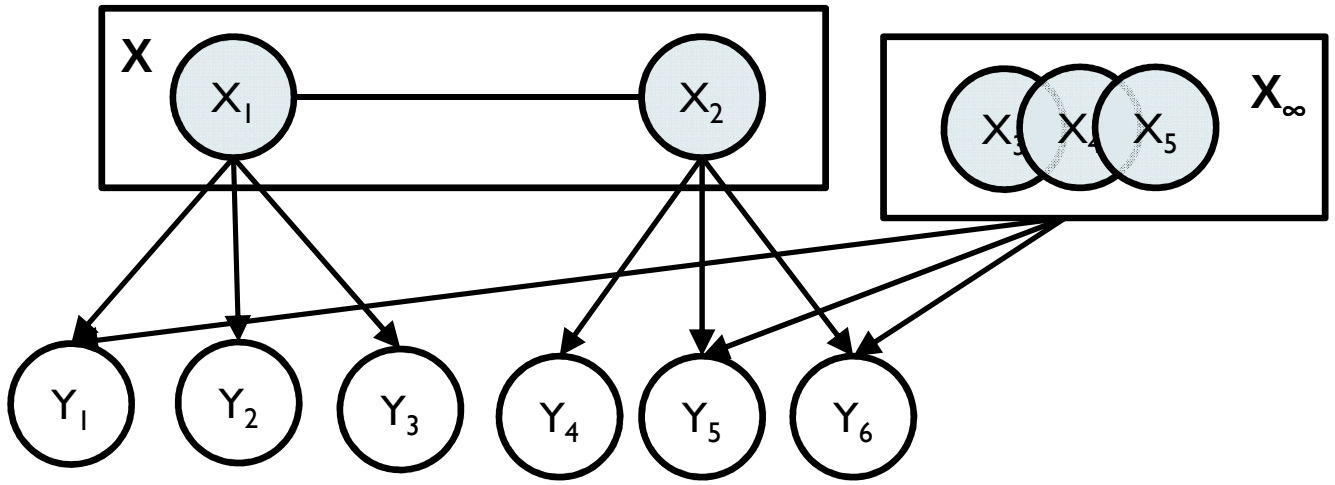
b. I will probably look for a job at a new organisation in the next 12 months.

c. As soon as I can find another job, I will leave this Trust.



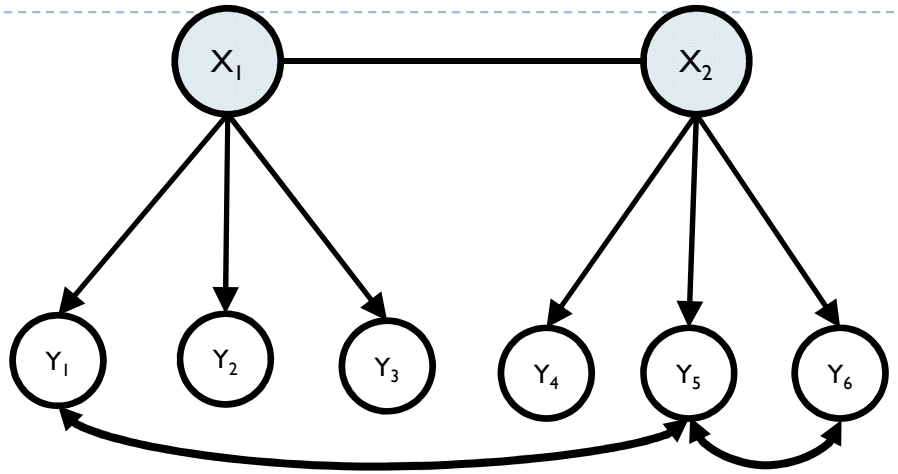
(NHS Staff Survey, 2009)

The "Structured Canonical Correlation" Structural Space



- Set of pre-specified latent variables **X**, observations **Y**
  - Each *Y* in **Y** has a pre-specified single parent in **X**
- Set of unknown latent variables $\mathbf{X}_\infty \perp\!\!\!\perp \mathbf{X}$
  - Each *Y* in **Y** can have potentially infinite parents in $\mathbf{X}_\infty$
- "Canonical correlation" in the sense of modeling dependencies within a partition of observed variables

The "Structured Canonical Correlation": Learning Task



- Assume a partition structure of Y according to $\mathbf{X}$ is known
- Define the *mixed graph projection* of a graph over ($\mathbf{X}$, $\mathbf{Y}$) by a bi-directed edge $Y_i \leftrightarrow Y_j$ if they share a common ancestor in $\mathbf{X}_\infty$
- Practical assumption: bi-directed substructure is sparse
- Goal: learn bi-directed structure (and parameters) so that one can estimate functionals of P($\mathbf{X} \mid \mathbf{Y}$)

Parametric Formulation

- $\mathbf{X} \sim N(0, \Sigma)$, Σ positive definite
  - Ignore possibility of causal/sparse structure in $\mathbf{X}$ for simplicity
- For a fixed graph *G*, parametrize the *conditional cumulative distribution function (CDF)* of $\mathbf{Y}$ given $\mathbf{X}$ according to bi-directed structure:
- $F(\mathbf{y} \mid \mathbf{x}) \equiv P(\mathbf{Y} \leq \mathbf{y} \mid \mathbf{X} = \mathbf{x}) \equiv \prod P_i(\mathbf{Y_i} \leq y_i \mid \mathbf{X}_{[i]} = \mathbf{x}_{[i]})$
  - Each set $\mathbf{Y}_i$ forms a bi-directed clique in *G*, $\mathbf{X}_{[i]}$ being the corresponding parents in $\mathbf{X}$ of the set $\mathbf{Y}_i$
  - In this paper we assume each *Y* is binary for simplicity

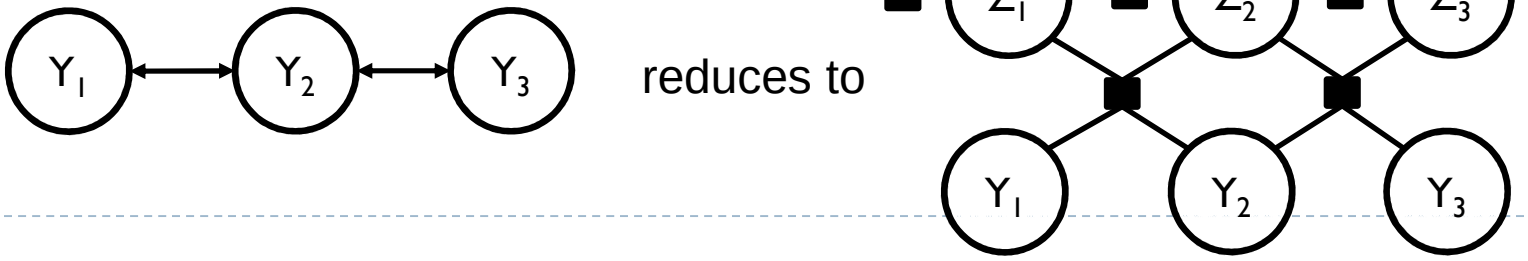
(Further details: Silva et al. 2011, Huang and Frey, 2011)

Parametric Formulation

- In order to calculate the likelihood function, one should convert from the (conditional) CDF to the probability mass function (PMF)
  - $P(\mathbf{y}, \mathbf{x}) = \{\Delta F(\mathbf{y} \mid \mathbf{x})\} P(\mathbf{x})$
  - Where $\Delta F(\mathbf{y} \mid \mathbf{x})$ represents a difference operator. For *p*-dimensional binary (unconditional) $F(\mathbf{y})$ this boils down to

$$\mathcal{P}(\mathbf{Y} = \mathbf{y}) = \sum_{z_1=0}^{1} \cdots \sum_{z_p=0}^{1} (-1)^{\sum_{i=1}^{p} z_i} F(\mathbf{y} - \mathbf{z})$$

- Message passing formulation – Example:



Learning with Marginal Likelihoods

- For X_j parent of Y_i in $\mathbf{X}$:

$$\mathcal{P}(Y_i = 0 \mid X_j = x_j) \equiv \Phi(0; \beta_{i1} x_j + \beta_{i0}, 1)$$

- Let

$$\mathcal{D} = \{\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(N)}\}$$

- Marginal likelihood:

$$\mathcal{P}(\mathcal{D} \mid \mathcal{G}_m, \{\beta_{i1}, \beta_{i0}\}, \Sigma) =$$

$$\int \mathcal{P}(\mathcal{D}, \mathbf{X}^{1:N}, \theta \mid \mathcal{G}_m, \beta, \Sigma)\, d\mathbf{X}^{1:N}\, d\theta$$

- Pick graph G_m that maximizes the marginal likelihood (maximizing also with respect to Σ and β), where θ parameterizes local conditional CDFs $F_i(\mathbf{y_i} \mid \mathbf{x_{[i]}})$

Computational Considerations

- Intractable, of course
  - Including possible large tree-width of bi-directed component
- First option: marginal bivariate composite likelihood

$$CL(\theta; \mathcal{D}) = \prod_{k \in K} \mathcal{L}_k(\theta; \mathcal{D})^{w_k}$$

$$PCL(\mathcal{G}_m, \beta, \Sigma) \equiv \mathcal{F}_{\mathcal{G}_m}^{(\beta,\Sigma)} + \log \pi(\mathcal{G}_m),$$

$$\mathcal{F}_{\mathcal{G}_m}^{(\beta,\Sigma)} \equiv \sum_{i<j} \log \mathcal{P}(\mathbf{Y}_i^{1:N}, \mathbf{Y}_j^{1:N} \mid \mathcal{G}_m, \beta, \Sigma)$$

Integrates θ_{ij} and $\boldsymbol{X}^{1:N}$ with a crude quadrature method

Algorithm 1 Pairwise Structured CCA Learning

1: **procedure** LEARNSTRUCTUREDCCA-I($\mathcal{S}$,$\mathcal{D}$)
2: $\{\mathbf{Y}, \mathbf{X}_{\mathcal{S}}, \mathcal{G}_m\} \leftarrow$ GETDAG($\mathcal{S}$)
3: $\{\beta, \Sigma\} \leftarrow$ INITPARAMETERS($\mathcal{G}_m, \mathcal{D}$)
4: **repeat**
5: $\{\beta, \Sigma\} \leftarrow \arg\max_{(\Omega_\beta, \Omega_\Sigma)} PCL(\mathcal{G}_m, \beta, \Sigma)$
6: $\mathcal{G}_m \leftarrow \arg\max_{(\mathcal{G}_m^{+/-})} PCL(\mathcal{G}_m, \beta, \Sigma)$
7: **until** $\mathcal{G}_m$ has not changed.
8: **return** $\mathcal{G}_m$
9: **end procedure**

$G_m^{+/-}$ is the space of graphs that differ from G_m by at most one bi-directed edge

Beyond Pairwise Models

- Wanted: to include terms that account for more than pairwise interactions
  - Gets expensive really fast
- An indirect compromise:
  - Still only pairwise terms just like PCL
  - However, integrate θ_{ij} not over the prior, but over some posterior that depends on more than on $\mathbf{Y}_i^{1:N}$, $\mathbf{Y}_j^{1:N}$:
    - Key idea: collect evidence from $p(\theta_{ij} \mid \mathbf{Y}_{\mathbf{S}}^{1:N})$, $\{i, j\} \subset \mathbf{S}$, plug it into the expected log of marginal likelihood $\mathcal{P}(\mathbf{Y}_i^{1:N}, \mathbf{Y}_j^{1:N} \mid \mathcal{G}_m, \beta, \Sigma)$. This corresponds to bounding each term of the log-composite likelihood score with different distributions for θ_{ij}:

$$\sum_{i<j} \int q_{ij}(\theta_{ij}) \log \frac{\mathcal{P}_{ij}(\mathbf{Y}_i^{1:N}, \mathbf{Y}_j^{1:N}, \theta_{ij} \mid \mathcal{G}_m, \beta, \Sigma)}{q_{ij}(\theta_{ij})} d\,\theta_{ij}$$

Beyond Pairwise Models

- New score function

$$\mathcal{Q}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})} = \sum_{m<n} \sum_{Y_i \in \mathcal{S}_m} \sum_{Y_j \in \mathcal{S}_n} \int q_{mn}(\theta_{ij}) \log \mathcal{P}(\mathbf{Y}_i^{1:N}, \mathbf{Y}_j^{1:N} \mid \mathcal{G}_m, \beta, \Sigma, \theta_{ij}) d\,\theta_{ij} + \frac{1}{|\mathcal{S}|-1} \sum_{m=1}^{|\mathcal{S}|} \sum_{n \neq m} \sum_{\{Y_i, Y_j\} \subset \mathcal{S}_m} \int q_{mn}(\theta_{ij}) \log \mathcal{P}(\mathbf{Y}_i^{1:N}, \mathbf{Y}_j^{1:N} \mid \mathcal{G}_m, \beta, \Sigma, \theta_{ij}) d\,\theta_{ij}$$

  - S_k: observed children of X_k in $\mathbf{X}$
  - Notice: multiple copies of likelihood for θ_{ij} when Y_i and Y_j have the same latent parent
- Use this function to optimize parameters $\{\beta, \Sigma\}$
  - (but not necessarily structure)

Algorithm 2

Algorithm 2 Modified Pairwise Structured CCA Learning

procedure LEARNSTRUCTUREDCCA-II($\mathcal{S}$,$\mathcal{D}$)
  $\{\mathbf{Y}, \mathbf{X}_{\mathcal{S}}, \mathcal{G}_m\} \leftarrow$ GETDAG($\mathcal{S}$)
  $\{\beta, \Sigma\} \leftarrow$ INITPARAMETERS($\mathcal{G}_m, \mathcal{D}$)
  repeat
    for $1 \leq m < n \leq |\mathcal{S}|$ **do**
      $q_{mn}(\cdot) \leftarrow \mathcal{P}(\Theta_{mn} \mid \mathbf{Y}_{mn}^{1:N}, \beta, \Sigma, \mathcal{G}_m)$
    end for
    $\{\beta, \Sigma\} \leftarrow \arg\max_{(\Omega_\beta, \Omega_\Sigma)} Q_{\mathcal{G}_m}^{(\beta, \Sigma, \{q_{mn}(\cdot)\})}$
    $\mathcal{G}_m \leftarrow \arg\max_{(\mathcal{G}_m^{+/-})} PCL(\mathcal{G}_m, \beta, \Sigma)$
  until $\mathcal{G}_m$ has not changed.
  return $\mathcal{G}_m$
end procedure

- q_{mn} comes from conditioning on all variables that share a parent with Y_i and Y_j
  - Laplace approximation
- In practice, we use PCL when optimizing structure
  - EM issues with discrete optimization: model without edge has an advantage, sometimes bad saddlepoint

Experiments: Synthetic Data

- 20 networks of 4 latent variables with 4 children per latent variable
  - Average number of bi-directed edges: ~18
- Evaluation criteria:
  - Mean-squared error of estimate of slope β for each observed variable
  - Edge omission error (false negatives)
  - Edge commission error (false positives)
- Comparison against "single-shot" learning
  - Fit model without bi-directed edges, add edge $Y_i \leftrightarrow Y_j$ if implied pairwise distribution $P(Y_i, Y_j)$ doesn't fit the data
    - Essentially a single iteration of Algorithm 1

Experiments: Synthetic Data

- Quantify results by taking the difference between number of times Algorithm 2 does better than Algorithm 1 and 0 ("single-shot" learning)

	1000		5000		10000	
Slope	I	0	I	0	I	0
number	13	6	17	15	15	13
p-value	0.22	0.25	*	*	*	0.06
Omission	I	0	I	0	I	0
number	11	18	6	14	6	9
p-value	0.17	*	0.82	*	0.62	0.22
Commision	I	0	I	0	I	0
number	5	2	15	16	16	18
p-value	0.28	*	*	*	*	*

  - The number of times where the difference is positive with the corresponding p-values for a Wilcoxon signed rank test (stars indicate numbers less than 0.05)

Experiments: NHS Data

- Fit model with 9 factors and 50 variables on the NHS data, using questionnaire as the partition structure
  - 100,000 points in training set, about 40 edges discovered
- Evaluation:
  - Test contribution of bi-directed edge dependencies to P(**X** | **Y**): compare against model without bi-directed edges
  - Comparison by predictive ability: find embedding for each $\mathbf{X}^{(d)}$ given $\mathbf{Y}^{(d)}$ by maximizing

$$\sum_{ij} \log \mathcal{P}(Y_i^{(d)}, Y_j^{(d)}, X_i^{(d)}, X_j^{(d)} \mid \beta, \Sigma, \mathcal{G}_m)$$

  - Test on independent 50,000 points by evaluating how well we can predict other 11 answers based on latent representation using logistic regression

Experiments: NHS Data

- MCCA: mixed graph structured canonical correlation model
- SCCA: null model (without bi-directed edges)
- Table contains AUC scores for each of the 11 binary prediction problems using estimated **X** as covariates:

	MCCA	SCCA		MCCA	SCCA
Q1	0.71	0.71	Q7	0.80	0.79
Q2	0.75	0.75	Q8	0.82	0.81
Q3	0.86	0.82	Q9	0.86	0.83
Q4	0.90	0.82	Q10	0.69	0.69
Q5	0.79	0.80	Q11	0.78	0.75
Q6	0.73	0.72			

Conclusion

- Marginal composite likelihood and mixed graph models are a good match
  - Still requires some choices of approximations for posteriors over parameters, and numerical methods for integration
- Future work:
  - Theoretical properties of the alternative marginal composite likelihood estimator
  - Identifiability issues
  - Reduction on the number of evaluations of q_{mn}
  - Non-binary data
    - Which families could avoid multiple passes over data?