Applying LSTM to text generation

J. Leroux, N. Laliberte, S. Frederic Boileau

MILA

Université de Montréal

Introduction

- A conditional independence graph is a concise representation of pairwise conditional independence among many variables.
- We present a general framework for estimating pairwise conditional independence relationships among variables.

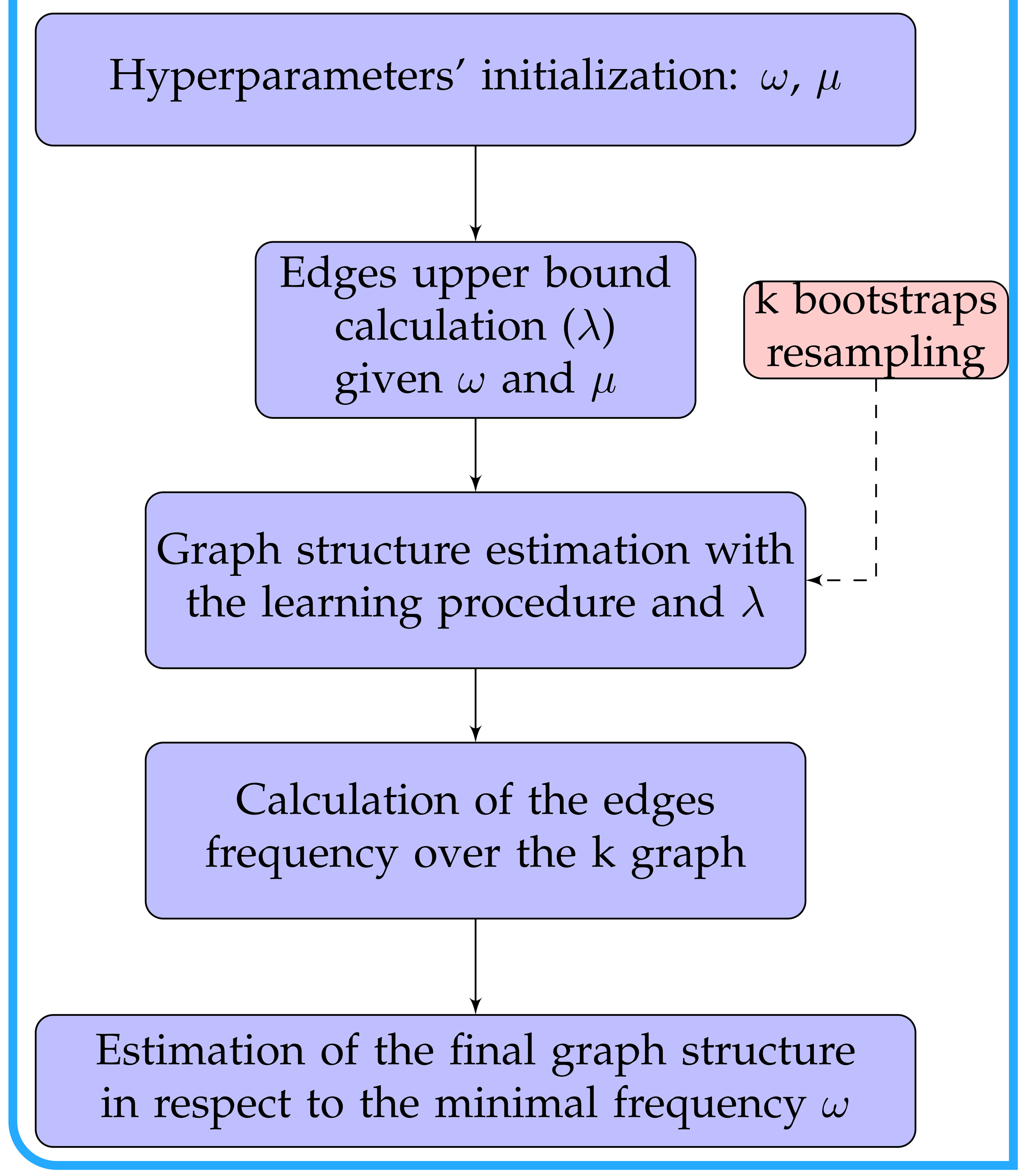
Stability Selection Algorithm

- Number of edges is a tuning parameter in any graphical model estimator.
- No obvious number constitutes a good choice.
- Stability Selection helps choosing the parameter (λ) with respect to a bound on the expected number of false positives μ :

$$\lambda = \lfloor \sqrt{(2\omega - 1) \cdot \mu \cdot p \cdot (p-1)/2} \rfloor$$

where ω is a threshold of the minimum relative frequency of edges and p is the number of variables.

Pseudo code



Algorithms

Graphical Random Forest (GRaFo)

Edges ranking scheme based on Random Forests' permutation importance measure.

Importance measure :

Predictor's relevance based on error difference between a regular Random Forest fit and a Random Forest fit within which one predictor's values are permuted.

Main steps :

1. Perform RF regression of a single variable on the remaining $p-1$ variables.
2. Calculate predictors' importance measure.
3. Repeat steps 1 and 2 for every p variables.
4. Define the (i, j) pair as i the outcome and j the predictor.
5. Best λ ranking pairs are added as edges.

Graphical Lasso (GLasso)

Estimation of the regularized precision matrix $\mathbf{\Lambda}$ under multivariate normal hypothesis:

$$\hat{\mathbf{\Lambda}}_{\mathrm{GL}} = \mathrm{argmax}_{\mathbf{\Lambda}} \left\{ \log |\mathbf{\Lambda}| - \mathbf{tr}(S\mathbf{\Lambda}) - \rho ||\mathbf{\Lambda}||_1^1 \right\}$$

where S is the empirical covariance matrix and ρ the regularization term.

Important points :

- The L_1 regularization impose the sparsity of the precision matrix.
- Null covariance is equivalent to independence under normal distribution.
- Graphical Lasso is a good solution for estimate sparse undirected graphical models.

References

Meinshausen, N., Bhlmann, P., 2010. Stability selection. J Roy Stat Soc B 72, 417—473.

Results

Evaluation of known graph structures

- Analysis of the ROC curve (Sensitivity $\sim$ Specificity) :

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad \text{Specificity} = \frac{TN}{TN+FP}$$

- Selection criterion : graph structure maximizing sensitivity and specificity sum (symbolized by a black diamond on the graphics).

Scenario 1 : Ising Model

- Simulate pairwise dependencies between a set of binary variables.

Scenario 2 : Gaussian Networks

- Simulate continuous variables following multivariate Gaussian distribution with sparse covariance matrix.

Conclusion

- Stability selection algorithm stabilizes the graphical model estimation on artificial datasets.
- Graphical Lasso outperforms GRaFo algorithm.

Blablabl

- **Comorbidity** : Any distinct additional clinical entity that has existed during the clinical course of a patient.
- **CLSA database** : 50,000 patients and 58 comorbidity factors.
- **Objective** : Profile the burden of comorbidity based on conditional dependences.

Conclusion

- Highlight mechanisms of disease purely based on observational data.
- Focus on few interesting associations which can then be specically tested in model organisms and clinical trials.