Dependency modelling toolbox

Leo Lahti*, Olli-Pekka Huovilainen, Abhishek Tripathi,
Ilkka Huopaniemi, and Tommi Suvitaival
Department of Information and Computer Science,
Aalto University School of Science and Technology, Finland

January 13, 2011

1 Introduction

Investigation of dependencies between multiple data sources allows the discovery of regularities and interactions that are not seen in individual data sets. The importance of such methods is increasing with the availability and size of co-occurring data sets in computational biology, open data initiatives, and in other domains. Practical, open access implementations of general-purpose algorithms will help to realize the full potential of these information sources.

This package provides general-purpose tools for the discovery and analysis of statistical dependencies between co-occurring data sources. The implementations are based on well-established models such as probabilistic canonical correlation analysis and multi-task learning [1, 2, 3, 4, 5]. Probabilistic framework deals rigorously with the uncertainties associated with small sample sizes, and allows incorporation of prior information in the analysis through Bayesian priors [4]. The applicability of the models has been demonstrated in previous case studies [3, 4, 5]. This is a development version. Your feedback and contributions are welcome. See the project page at R-Forge[1], or contact project authors.

1.1 Installation

Install dmt from within R using command
'install.packages("dmt", repos="http://R-Forge.R-project.org")'

2 Functionality

- regularized dependency detection [2, 4]
- dependency-based dimensionality reduction [5]
- multi-way modeling of co-occurrence data[2]; [3]. Currently available as example source code only.

*leo.lahti@iki.fi
[1] http://dmt.r-forge.r-project.org/
[2] http://www.cis.hut.fi/projects/mi/software/multiWayCCA/

Below is a brief summary of the functionality and installation instructions.

2.1 Documentation

The package implements the dependency modeling framework explained below (see function 'fit.dependency.model'), and provides wrappers for the special cases of the model.

Currently only online-documentation for the package is available for dependency-based dimensionality reduction. See
http://www.cis.hut.fi/projects/mi/software/drCCA/dochtml/00Index.html

3 Probabilistic dependency modeling framework

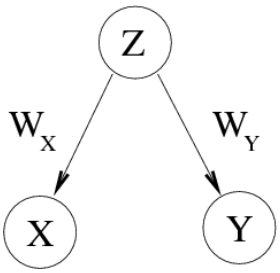

Figure 1: Graphical description of the shared latent variable model showing generation of data sets X and Y from latent shared variable $\mathbf{z}$ through W_x and W_y

The package [4] implements the probabilistic dependency modeling framework presented in [2] and extensions [1, 6, 4]. The latent variable model assumes that the two data sets, X and Y can be decomposed in *shared* and *data set-specific* components (Figure˜1). Our tools help to discover these components, given modeling assumptions.

The shared signal is modeled with a latent variable $\mathbf{z}$. Intuitively, this measures the strength of the shared signal in each patient. Shared signal can have different manifestation in each data set, described by $W_x z$ and $W_y z$ where W_x, W_y. Assuming a standard Gaussian model for the shared latent variable $\mathbf{z} \sim N(0, I)$ and data set-specific effects, this leads to the following model:

$$\begin{aligned} X &\sim W_x \mathbf{z} + \varepsilon_x \\ Y &\sim W_y \mathbf{z} + \varepsilon_y \\ \varepsilon_\cdot &\sim \mathcal{N}(0, \Psi_\cdot) \\ \mathbf{z} &\sim \mathcal{N}(0, I) \end{aligned} \tag{1}$$

The data set-specific effects are modelled by the covariance matrices Ψ_x, Ψ_y. Model parameters are estimated with an EM algorithm.

3.1 Special cases

Special cases of the model include probabilistic versions of canonical correlation analysis, factor analysis, and principal component analysis, and regularized

versions of them.

Probabilistic CCA (pCCA) assumes full covariance matrices Ψ_x, Ψ_y. This gives the most detailed model for the data set specific effects. The connection of this latent variable model and the traditional canonical correlation analysis has been established in [2].

Probabilistic factor analysis (pFA) is obtained with a diagonal covariances Ψ_x, Ψ_y. In addition, a special case is implemented where each covariance matrix $\Psi_.$ is isotropic but they are not necessarily identical (as would be the case in pPCA). This model is identical to concatenating X, Y, and fitting ordinary probabilistic factor analysis on the concatenated data set. The structure of the covariances is simpler than in pCCA. This regularizes the solution and can potentially reduce overfitting in some applications.

Probabilistic PCA (pPCA) is obtained with identical isotropic covariances for the data set-specific effects: $\Psi_x = \Psi_y = \sigma I$. This model is identical to concatenating X, Y, and fitting ordinary probabilistic PCA on the concatenated data set.

3.2 Regularized dependency modelling

We provide toos to guide dependency modeling through Bayesian priors [4]. Prior on the relation between W_x and W_y can be used to guide modeling to focus on certain types of dependencies, and to avoid overfitting. The relationship is described through $W_y = TW_x$. We use matrix normal prior distribution: $P(T) = N_m(H, \sigma_T^2 I, \sigma_T^2 I)$. By default, $H = I$ and $\sigma_T^2 = 0$, giving $W_y = W_x$. This model is denoted pSimCCA in the package. The prior can be loosened by tuning $sigma_T^2$. With $sigma_T^2 \to \infty$, estimation of W_x and W_y become independent, yielding ordinary probabilistic CCA. It is also possible to tune the mean matrix H. This would set a particular relationship between the manifestations of the shared component in each data set, and $sigma_T^2$ is again be used to tune the strength of such prior.

4 Dependency-based dimensionality reduction

The drCCA [5] method can be used for dependency-based dimensionality reduction that retains the variation shared between the original data sources, while reducing data set-specific effects. The approach utilizes generalized canonical correlation analysis to perform a linear projection on the collection of data sets. Linearity makes it fast on large data sets. The package includes regularization and tools to select the final dimensionality of the combined data set automatically.

4.1 Applications

For applications in functional genomics, see [3, 4, 5], and the associated pint[3] BioConductor package. This can be installed from within R with source('http://bioconductor.org/biocLite.R') biocLite('pint')

[3] http://bioconductor.org/packages/release/bioc/html/pint.html

5 Multi-way multi-view models (multiWayCCA)

multiWayCCA[4] provides tools for multi-way, multi-source modeling. This is particularly usefule for simultaneous multi-way (anova-type) modelling of multiple related data sources. For details, see the original paper [3].

5.1 Installing & documentation of multiWayCCA

Download the source[5]. Then uncompress the folder; readme.txt in the uncompressed folder contains instructions for running the analysis. For documentation and examples, see the readme.txt file included in the package.

5.2 Licensing terms

The dmt package is licensed under GNU GPL 2 open software license.

Acknowledgements

The project is a joint effort by several people: Leo Lahti, Tommi Suvitaival, Olli-Pekka Huovilainen, Ilkka Huopaniemi, Abhishek Tripathi, Arto Klami, and Samuel Kaski from the Statistical Machine Learning and Bioinformatics group at the Department of Information and Computer Science, Aalto University School of Science and Technology, Finland. Abhishek Tripathi is with Department of Computer Science, University of Helsinki, Finland. The authors belong also to Helsinki Institute for Information Technology HIIT and Adaptive Informatics Research Centre AIRC.

References

[1] Cédric Archambeau, Nicolas Delannay, and Michel Verleysen. Robust probabilistic projections. In W.W. Cohen and A.˜Moore, editors, *Proceedings of the 23rd International conference on machine learning*, pages 33–40. ACM, 2006.

[2] Francis˜R. Bach and Michael˜I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[3] Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešic, and Samuel Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 2010. (ISMB 2010, to appear).

[4] Leo Lahti, Samuel Myllykangas, Sakari Knuutila, and Samuel Kaski. Dependency detection with similarity constraints. In *Proc. MLSP'09 IEEE International Workshop on Machine Learning for Signal Processing*, 2009.

[4] http://www.cis.hut.fi/projects/mi/software/multiWayCCA/

[5] http://www.cis.hut.fi/projects/mi/software/multiWayCCA/multiWayCCA-package-100326.zip

[5] Arto˜Klami Abhishek˜Tripathi and Samuel Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9(111), 2008.

[6] Arto Klami and Samuel Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72(1-3):39–46, 2008.