pint:
Pairwise integration of functional genomics data

Olli-Pekka Huovilainen*and Leo Lahti
Department of Information and Computer Science,
Aalto University School of Science and Technology, Finland

January 5, 2011

1 Introduction

Multiple types of genomic observations from the same patients are increasingly available in biomedical studies, including measurements of gene- and micro-RNA expression levels, DNA copy number, and methylation status. By investigating the dependencies between the various functional layers of the genome it is possible to discover mechanisms and interactions that are not seen in the individual data sets. For instance, integration of gene expression and DNA copy number can reveal cancer-associated chromosomal regions and associated genes with potential diagnostic, prognostic and clinical impact (4).

We demonstrate how to integrate gene or micro-RNA expression with DNA copy number (aCGH) measurements to discover functionally active chromosomal aberrations. The models capture the shared signal in paired observations, indicating the affected genes and patients. The methods are potentially applicable also to other types of biomedical data, including epigenetic modifications, SNPs, alternative splicing and transcription factor binding, or in other application fields.

The methods are based on a particular latent variable model, probabilistic canonical correlation analysis (2) and its extensions (1; 3; 4). Probabilistic formulation deals rigorously with uncertainty associated with small sample sizes common in biomedical studies. Additional tools to guide dependency modeling through Bayesian priors are provided (4).

2 Examples

This Section shows how to apply the methods for dependency detection in functional genomics studies. The general dependency modeling framework and tools are described in Section ˜3.

*ohuovila@gmail.com

2.1 Example data

Use of the package is demonstrated with an example data set containing paired observations of gene expression and copy number from a set of gastric cancer patients (5).

Load the package and example data:

```
> require(pint)

pint Copyright (C) 2008-2011 Olli-Pekka Huovilainen and Leo Lahti.

This program comes with ABSOLUTELY NO WARRANTY.

This is free software, and you are welcome to redistribute it under the FreeBSD license.

See the licensing terms for details.

> data(chromosome17)
```

Each example data set ($geneExp$ and $geneCopyNum$) consists of a list with two items: $data$ and $info$. The probes in gene expression and gene copy number are assumed to be paired. $data$ is a data matrix with gene expression or gene copy number data. Genes are in rows and samples in columns and rows and columns should be named. $info$ is a data frame with additional information about genes. It has three elements: loc, chr and arm. loc indicates the genomic location of the probes in base pairs (numeric); chr and arm are factors indicating the chromosome and chromosomal arm of the probe.

2.1.1 Modeling assumptions

Note that the currently implemented dependency models assume approximately Gaussian distributed observations. With microarray data sets, this is typically obtained by presenting the data in the log_2 domain; this is the default in many microarray preprocessing methods.

2.2 Discovering functionally active copy number changes

Chromosomal regions with frequent copy number alterations and associated gene expression changes are potential candidates for cancer genes. Such regions are assumed to have high dependency between gene copy number and expression levels. To detect these regions, we measure the dependency between expression and copy number for each region and pick the regions showing the highest dependency. In practice, a sliding window over the genome is used.

The following example screens chromosome arm 17q for dependendent regions.

```
> models <- screen.cgh.mrna(geneExp, geneCopyNum, windowSize = 10,
+     chr = 17, arm = "q")
```

The dependency is measured separately for each gene within a chromosomal region ('window') around the gene. A fixed dimensionality (window size) is necessary to ensure comparability of the dependency scores between windows.

The scale of the chromosomal regions can be tuned by changing the window size ('windowSize'). The default dependency modeling method is a constrained version of probabilistic CCA; (4). See help(screen.cgh.mrna) for further options.

2.3 Application to micro-RNA or epigenetic measurements

Micro-RNA and epigenetic measurements are increasingly available in biomedical studies, accompanying observations of DNA copy number changes and mRNA expression levels. Assuming that appropriate preprocessing is performed, and these data types are provided as matrices with matched samples and chromosomal location information for the measurement probes, the current functions provide tools to discover chromosomally local dependencies between any of these data sources.

2.4 Visualizing the results

A dependency plot reveals chromosomal regions with the strongest dependency between gene expression and copy number:

```
> plot(models, showTop = 10)
```

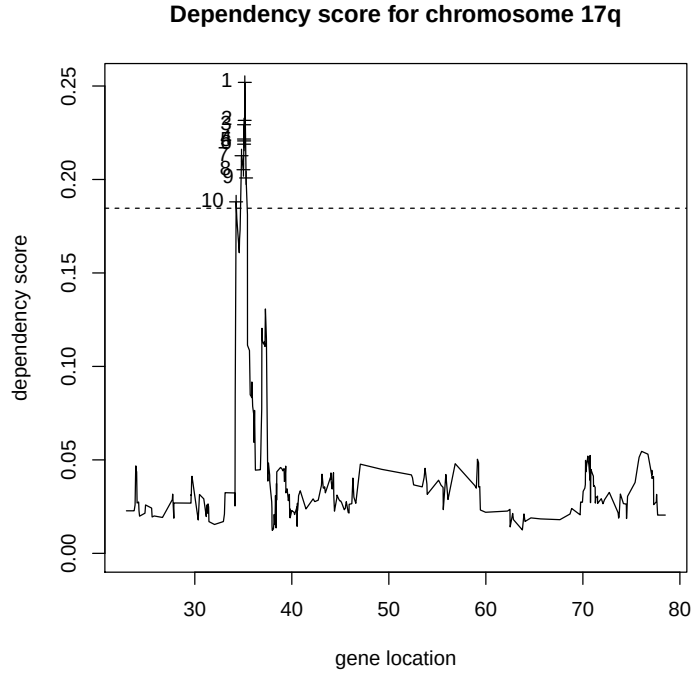

Figure 1: The dependency plot reveals chromosomal regions with the strongest dependency between gene expression and copy number.

Here the highest dependency is between 30-40Mbp which is a known gastric

cancer-associated region. Note that the display shows the location in megabase-pairs while location is provided in basepairs. The top-5 genes with the highest dependency in their chromosomal neighborghood can be retrieved with:

```
> topGenes(models, 5)
```

```
[1] "ENSG00000141738" "ENSG00000141736" "ENSG00000173991" "ENSG00000131748"
[5] "ENSG00000161395"
```

It is possible investigate the contribution of individual patients or probes on the overall dependency (Fig. ˜2).

This is based on the model parameters W and the latent variable z that are easily retrieved from the learned dependency model (see Section˜3 for description of the model parameters). In 1-dimensional case the interpretation is straightforward: z indicates the strength of shared signal in each sample (patient), and W describes how this signal is captured by each gene expression or copy number probe. With multi-dimensional W and z, the variable- and sample effects are approximated (for visualization purposes) by the loadings and projection scores corresponding of the first principal component of Wz which describes the shared signal in each data set.

2.5 Additional parameters

The dimensionality of the shared latent variable Z and the dependency modeling method can also be set by the user. For example, use probabilistic CCA with 1-dimensional Z:

```
> model17qpCCA <- screen.cgh.mrna(geneExp, geneCopyNum, windowSize = 10,
+     chr = 17, arm = "q", method = "pCCA", params = list("zDimension = 1"))
```

```
> model <- topModels(models, 1)[[1]]
> plot(model, geneExp, geneCopyNum)
```

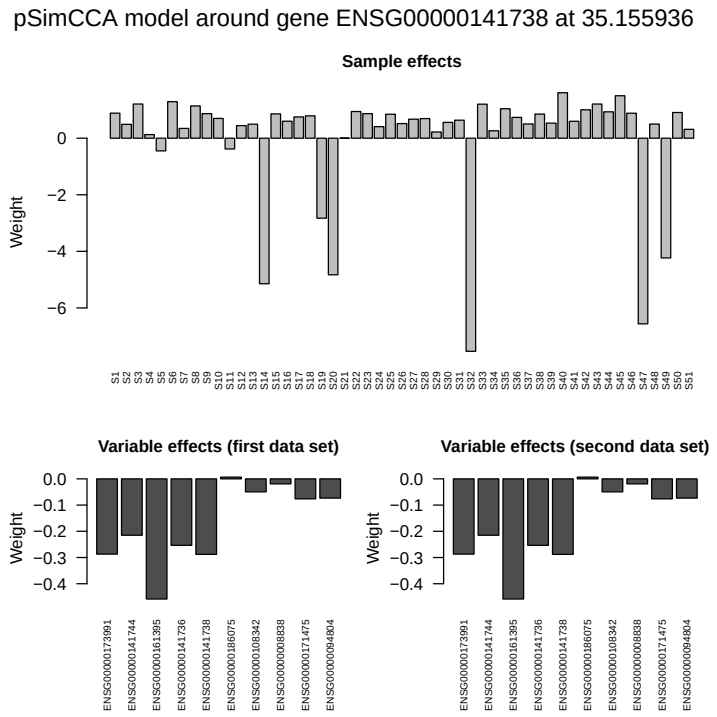

Figure 2: Samples and variable contribution to the dependencies around the gene with the highest dependency score between gene expression and copy number measurements in the chromosomal region. The visualization highlights the affected patients and genes.

3 Dependency modeling framework

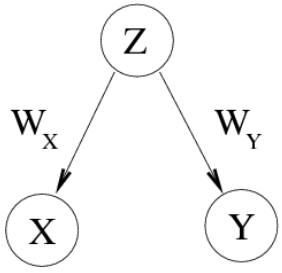

Figure 3: Graphical description of the shared latent variable model showing generation of data sets X and Y from latent shared variable $\mathbf{z}$ through W_x and W_y

Modeling of dependencies is based on the probabilistic canonical correlation analysis framework (2) and its extensions (1; 4). This is a latent variable model that assumes that the two data sets, X and Y can be decomposed in *shared* and *data set-specific* components (Figure˜3). Our task is to discover these components, given modeling assumptions.

The shared signal is modeled with a shared latent variable $\mathbf{z}$. Intuitively, this measures the strength of the shared signal in each patient. While the variation is shared, it can have different manifestation in each data set. This is described by $W_x z$ and $W_y z$ where W_x, W_y indicate how the shared signal is observed in the individual data sets. Assuming a Gaussian model for the shared latent variable and data set-specific effects, this leads to the following model:

$$X \sim \mathcal{N}(W_x \mathbf{z}, \Psi_x) \tag{1}$$

$$Y \sim \mathcal{N}(W_y \mathbf{z}, \Psi_y) \tag{2}$$

The latent variable $\mathbf{z}$ is assumed to follow standard multivariate normal distribution, i.e. $\mathbf{z}˜N(0, I)$. The data set-specific effects are described by the covariance matrices Ψ_x and Ψ_y. The model parameters are estimated with an expectation-maximization (EM) algorithm (see 'fit.dependency.model'). After fitting the model parameters W, Ψ, a maximum-likelihood estimate of $\mathbf{z}$ can be calculated (see 'z.expectation').

3.1 Special cases

Particular models are obtained as special cases of the above modeling framework. This leads to a set of alternative models for dependency detection, including

- probabilistic PCA (pPCA)
- probabilistic factor analysis (pFA)
- probabilistic CCA (pCCA)
- similarity-constrained probabilistic CCA (pSimCCA)

These correspond to different assumptions regarding the structure of the data set-specific effects and types of dependency. While PCA and factor analysis are typically used for analysing individual data sets, they are special cases of the described framework and can therefore also be used to model dependencies between data sets. For discussion of the differences between these models, see (2; 4).

3.1.1 Probabilistic PCA

Probabilistic PCA (pPCA) assumes an isotropic model for the data set-specific effects, with identical covariance matrices:

$$\Psi_x = \Psi_y = \sigma I. \tag{3}$$

This model is called pPCA since it is identical to concatenating X, Y, and fitting ordinary probabilistic PCA on the concatenated data set.

3.1.2 Probabilistic factor analysis

Probabilistic factor analysis (pFA) assumes a diagonal model for Ψ_x, Ψ_y. Note that in general, $\Psi_x \neq \Psi_y$. The package also implements a special case with isotropic but not necessarily identical (as in pPCA) covariance matrices.

This model is called pFA since it is identical to concatenating X, Y, and fitting ordinary probabilistic factor analysis on the concatenated data set.

3.1.3 Probabilistic CCA

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

3.1.4 Probabilistic SimCCA

We also provide toos to guide dependency modeling through Bayesian priors (4). Similarity-constrained probabilistic CCA (pSimCCA) imposes a prior on the relation between W_x and W_y. This can be used to guide modeling to focus on certain types of dependencies, and to avoid overfitting.

The relationship between W_x and W_y is described with $W_y = TW_x$. A prior on T can be used to focus the modeling on certain types of dependencies. We use matrix normal prior distribution:

$$P(T) = N_m(H, \sigma_T^2 I, \sigma_T^2 I) \tag{4}$$

By default, $H = I$ and $\sigma_T^2 = 0$, which results in identical manifestation of the shared signal in the two data sets: $W_y = W_x$. This model is denoted pSimCCA in the package. However, 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, which leads to 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.

3.2 Quantifying dependency

Dependency between the data sets X, Y is measured by the ratio of shared vs. data set-specific signal (see '?dependency.score'):

$$\frac{Tr(WW^T)}{Tr(\Psi)} \tag{5}$$

3.3 Functions for dependency modeling

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

Acknowledgements

We would like to thank prof. Sakari Knuutila (University of Helsinki) for providing the example data set.

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

[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.

[5] Samuel Myllykangas, Siina Junnila, Arto Kokkola, Reija Autio, Ilari Scheinin, Tuula Kiviluoto, Marja-Liisa Karjalainen-Lindsberg, Jaakko Hollmén, Sakari Knuutila, Pauli Puolakkainen, Outi Monni Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes *International Journal of Cancer*, 123(4):817-25, 2008.