pint: Pairwise integration of heterogeneous functional genomics data

Olli-Pekka Huovilainen and Leo Lahti

March 29, 2010

1 Introduction

Paired genomic observations of related biological aspects in the same patients are increasingly available, including measurements of gene- and miRNA expression, gene copy number, and methylation status. Study of the dependencies can reveal functional mechanisms and interactions not seen in the individual data sets. For example, integration of gene expression and copy number has been shown to reveal cancer-associated chromosomal regions and associated genes with potential diagnostic, prognostic and clinical impact [4].

The *pint* package provides tools for the discovery and analysis of statistical dependencies between co-occurring data sources. The currently implemented methods are based on the probabilistic canonical correlation analysis (pCCA) framework [2] and its extensions [1, 3, 4]. The package provides tools to guide dependency modeling through Bayesian priors [4]. The models assume approximately Gaussian distributed observations. Probabilistic formulation deals rigorously with uncertainty associated with small sample sizes common in biomedical studies.

We demonstrate how to integrate gene or micro-RNA expression with DNA copy number (aCGH) measurements to discover functionally active mutations. The models capture the strongest shared signal in paired observations, and indicates affected genes and patients. The dependency modeling framework is potentially applicable also to other types of biomedical data, including methylation, SNPs, alternative splicing and transcription factor binding, or in other application fields.

2 Examples

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)
> 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.

2.2 Discovering functionally active copy number changes

Screen the genome to discover chromosomal regions with statistical dependencies between the measurements (gene expression and copy number in our example).

```
> model17qpSimCCA <- screen.chromosome(geneExp, geneCopyNum, windowSize = 10,
+     chr = 17, arm = "q", method = "pSimCCA")
```

This example uses the default method (pSimCCA; [4]) to screen over chromosome arm 17q for dependencies. The dependency is measured within a chromosomal region ('window') around each gene, defined with a fixed window size of 10 closest genes. A fixed window size quarantees the comparability of the dependency scores between windows.

2.3 Visualizing the results

Chromosomal regions with the strongest dependencies can be seen from a dependency plot:

```
> plot(model17qpSimCCA, showTop = 10)
```

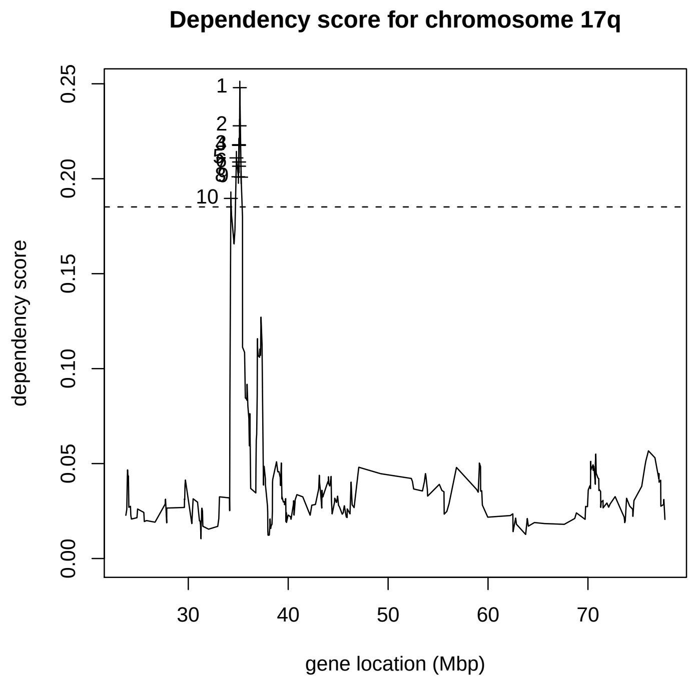

Figure 1: The dependency plot reveals chromosomal regions with high dependency between gene expression and copy number.

The highest dependency is in the area between 30Mbp and 40Mbp. This is a known gastric cancer associated region (see e.g. [4]). The top-5 genes with the highest dependency in their chromosomal neighborghood can be retrieved with:

```
> findHighestGenes(model17qpSimCCA, 5)

[1] "ENSG00000141738" "ENSG00000141736" "ENSG00000131748" "ENSG00000173991"
[5] "ENSG00000125686"
```

We can also investigate the contribution of individual patients or probes on the overall dependency (Fig. **??**- **??**). The model parameters W and z are easily retrieved from the models (see [4] for description of the model parameters). In 1-dimensional case their interpretation is straightforward; z indicates the strength of shared signal in each sample (patient), and W describes how the signal is captured by each probe. Wz describes how the shared signal is manifested in each data set. 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.

```
> model <- findHighestModels(model17qpSimCCA, 1)[[1]]
> plot(model, geneExp, geneCopyNum)
```

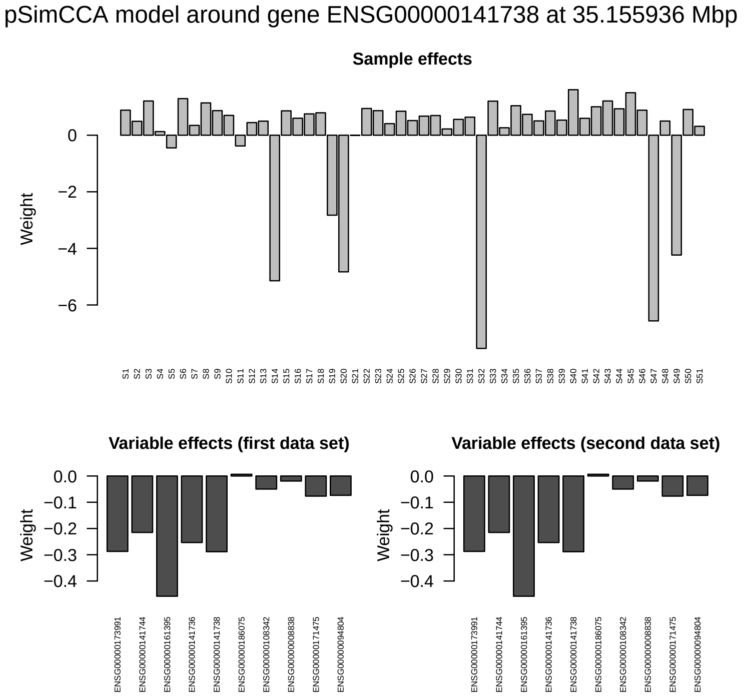

Figure 2: Samples and variable contribution to the dependencies around the gene with the highest dependency score between gene expression and copy number measurements in its chromosomal neighborghood. The visualization highlights affected patients and the associated genes.

3 Implemented dependency models

The package provides alternative probabilistic dependency measures (set the 'method' variable in screen.chromosome). These include:

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

These correspond to different assumptions regarding the structure of the data set-specific covariances. The dimensionality of the shared latent variable Z can also be set by the user. The SimCCA method guides the dependency search with priors on W.

For example, use probabilistic CCA with 1-dimensional Z:

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

Acknowledgements

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

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 Hollm??, 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.