pint: probabilistic data integration for functional genomics

Olli-Pekka Huovilainen[1]*and Leo Lahti[1,2]
(1) Dpt. Information and Computer Science, Aalto University, Finland
(2) Dpt. Veterinary Bioscience, University of Helsinki, Finland

April 16, 2015

1 Introduction

Multiple genomic observations from the same samples are increasingly available in biomedical studies, including measurements of gene- and micro-RNA expression levels, DNA copy number, and methylation status. By investigating dependencies between different functional layers of the genome it is possible to discover mechanisms and interactions that are not seen in the individual measurement sources. 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 [6].

This package implements probabilistic models for integrative analysis of mRNA expression levels with DNA copy number (aCGH) measurements to discover functionally active chromosomal alterations. The algorithms can be used to discover functionally altered chromosomal regions and to visualize the affected genes and samples. The algorithms can be applied 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 latent variable models including probabilistic canonical correlation analysis [2] and related extensions [1, 4, 6], implemented in the *dmt* package in CRAN [5, 3]. Probabilistic formulation deals rigorously with uncertainty associated with small sample sizes common in biomedical studies and provides tools to guide dependency modeling through Bayesian priors [6].

1.0.1 Dependencies

The CRAN packages *dmt* and *mvtnorm* are required for installation.

*ohuovila@gmail.com

2 Examples

This Section shows how to apply the methods for dependency detection in functional genomics. For further details on the dependency modeling framework, see the dependency modeling package *dmt* in CRAN[1].

2.1 Example data

Our example data set contains matched observations of gene expression and copy number from a set of gastric cancer patients [7]. Load the package and example data with:

```
> library(pint)
> data(chromosome17)
```

The example data contains (*geneExp* and *geneCopyNum*) objects. These lists contain two elements:

- *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 and the probes and samples are matched between the two data sets.
- *info* data frame with additional information about the genes in the *data* object; in particular, *loc* indicates the genomic location of each probe in base pairs; *chr* and *arm* indicate the chromosome and chromosomal arm of the probe.

The models assume *approximately* Gaussian distributed observations. With microarray data sets, this is typically obtained by presenting the data in the log_2 domain, which is the default in many microarray preprocessing methods.

2.2 Discovering functionally active copy number changes

Chromosomal regions that have simultaneous copy number alterations and gene expression changes will reveal potential cancer gene candidates. To detect these regions, we measure the dependency between expression and copy number for each region and pick the regions showing the highest dependency as such regions have high dependency between the two data sources. A sliding window over the genome is used to quantify dependency within each region. Here we show a brief example on chromosome arm 17q:

```
> 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; [6]. See help(screen.cgh.mrna) for further options.

[1] http://dmt.r-forge.r-project.org/

2.3 Application in other genomic data integration tasks

Other genomic data sources such as micro-RNA or epigenetic measurements are increasingly available in biomedical studies, accompanying observations of DNA copy number changes and mRNA expression levels [8]. Given matched probes and samples, the current functions can be used to screen for dependency between any pair of genomic (or other) data sources.

3 Summarization and interpretation

3.1 Visualization

Dependency plots will reveals chromosomal regions with the strongest dependency between gene expression and copy number changes:

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

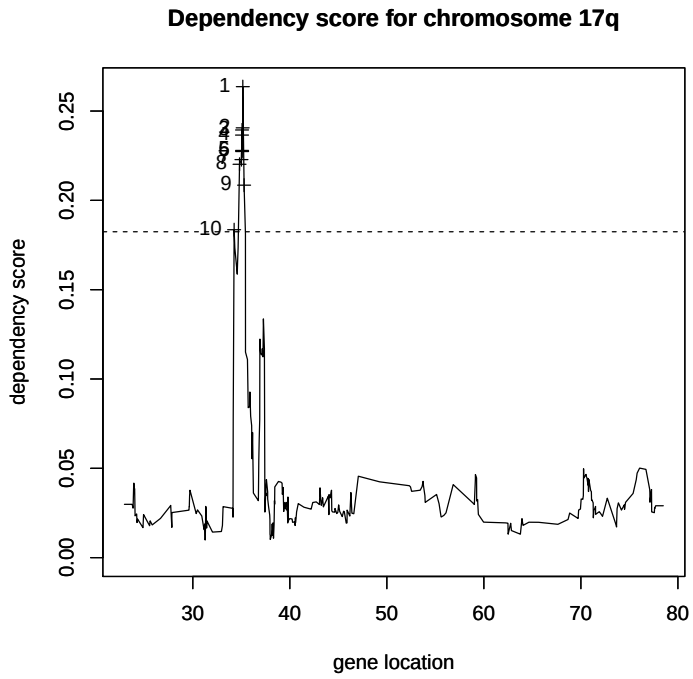

Figure 1: The dependency plot reveals chromosomal regions with the strongest dependency between gene expression and copy number.

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] "ENSG00000141744"
```

It is also possible investigate the contribution of individual patients or probes on the overall dependency based on the model parameters W and the latent variable $\mathbf{z}$ that are easily retrieved from the learned dependency model (Fig. 2). In 1-dimensional case the interpretation is straightforward: $\mathbf{z}$ will indicate the shared signal strength in each sample and W describes how the shared signal is reflected in each data source. With multi-dimensional W and $\mathbf{z}$, the variable- and sample effects are approximated (for visualization purposes) by the loadings and projection scores corresponding of the first principal component of $W\mathbf{z}$ is used to summarize the shared signal in each data set.

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

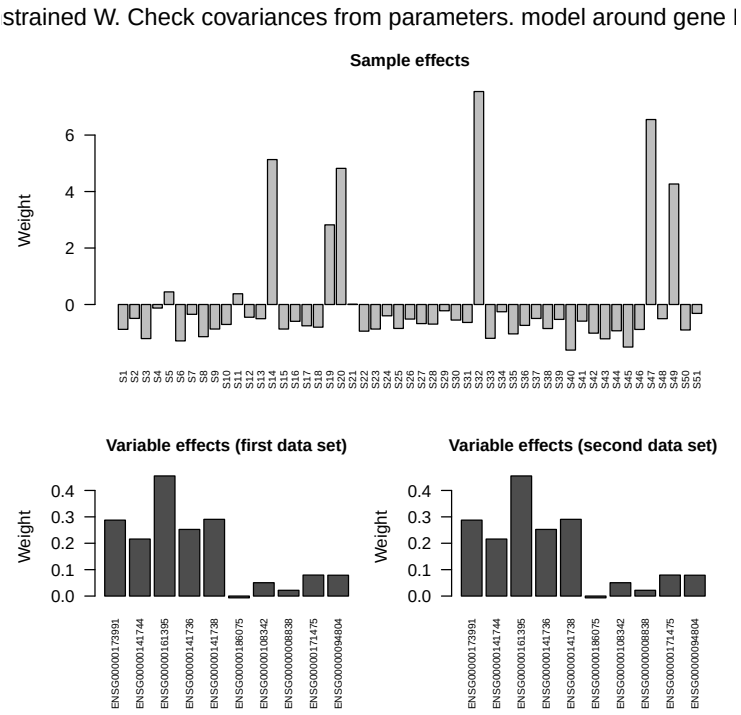

Figure 2: Samples and variable contribution to the dependencies around the gene with the highest dependency score between gene expression and copy number measurements in the chromosomal region. The visualization highlights the affected patients and genes.

3.2 Summarization

Other useful functions for summarizing and investigating the results include:

- *join.top.regions*: merge overlapping models that exceed the threshold; gives a list of distinct, continuous regions detected by the models.
- *summarize.region.parameters*: provides a summary of sample and probe effects over partially overlapping models

4 The dependency modeling framework

Detailed description of the model parameters and available dependency detection methods is provided with the *dmt* package in CRAN [5]. The models are based on probabilistic canonical correlation analysis and related extensions [2, 6]. In summary, the shared signal between two (multivariate) observations X, Y is modeled with a shared latent variable $\mathbf{z}$. This can have different manifestation in each data set, which is described by linear transformations W_x and W_y. Standard multivariate normal distribution for the shared latent variable and data set-specific effects gives the following model:

$$\begin{aligned} X &\sim W_x \mathbf{z} + \varepsilon_x \\ Y &\sim W_y \mathbf{z} + \varepsilon_y \end{aligned} \tag{1}$$

The data set-specific effects are modeled with multivariate Gaussians $\varepsilon_\cdot \sim \mathcal{N}(0, \Psi_\cdot)$ with covariances Ψ_x, Ψ_y, respectively. Dependency between the data sets X, Y is quantified by the ratio of shared vs. data set-specific signal (see '?dependency.score'), calculated as

$$\frac{Tr(WW^T)}{Tr(\Psi)} \tag{2}$$

5 Details

- *Licensing terms:* the package is licensed under FreeBSD open software license
- *Citing pint:* Please cite [6]

This document was written using:

```
> sessionInfo()

R version 3.2.0 (2015-04-16)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: Ubuntu 14.04.2 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] pint_1.18.0   dmt_0.8.20    MASS_7.3-40   Matrix_1.2-0  mvtnorm_1.0-2

loaded via a namespace (and not attached):
[1] tools_3.2.0     grid_3.2.0      lattice_0.20-31
```


Acknowledgements

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

References

[1] C. Archambeau, N. Delannay, and M. Verleysen. Robust probabilistic projections. In W. Cohen and A. Moore, editors, *Proceedings of the 23rd International conference on machine learning*, volume 148, pages 33–40, Pittsburgh, Pennsylvania, 2006. ACM.

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

[3] O.-P. Huovilainen. Screening of functional copy number changes with dependency models. Master's thesis, Aalto University School of Science and Technology, Department of Information and Computer Science, 2010.

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

[5] L. Lahti et al. Dependency modeling toolkit. ICML workshop, June 2010. Implementation provided in the package *dmt* at CRAN.

[6] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski. Dependency detection with similarity constraints. In *Proceedings MLSP'09 IEEE International Workshop on Machine Learning for Signal Processing XIX*, pages 89–94, Piscataway, NJ, September 2-4 2009. IEEE Signal Processing Society.

[7] S. Myllykangas, S. Junnila, A. Kokkola, R. Autio, I. Scheinin, T. Kiviluoto, M.-L. Karjalainen-Lindsberg, J. Hollmén, S. Knuutila, P. Puolakkainen, and O. Monni. Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes. *International Journal of Cancer*, 123(4):817–825, 2008.

[8] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008.