Robust deconvolution of transcriptomic samples using the gene covariance structure

Bastien CHASSAGNOL[1,2,3], Pierre-Henri WUILLEMIN[1], Gregory NUEL[2] and Etienne BECHT[3]
[1] LIP6 (Laboratoire d'Informatique Paris 6), 4 Place Jussieu, 75005, Paris, FRANCE
[2] LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), 4 Place Jussieu, 75005, Paris, FRANCE
[3] Les Laboratoires Servier, 50 Rue Carnot, 92150, Suresnes, FRANCE

Corresponding author: bastien.chassagnol@upmc.fr

Transcriptomic analyses have increasingly contributed to our understanding of the intricate biological processes involved in the emergence of auto-immune diseases or tumour-promoting environments. However, classical bulk analyses ignore the intrinsic complexity of biological samples, by averaging measurements over multiple distinct cell populations. It is therefore unclear whether a change in the gene expression between samples results from a variation of the cell type proportions, from an environmental signal or a mutation [1].

To remove this ambiguity, deconvolution algorithms can estimate the proportions of cell populations from a bulk transcriptome using the reference transcriptome of purified cell populations. Traditionally, most approaches, including the gold standard CIBERSORT algorithm [2], retrieve the cell proportions of a mixture assuming the linear assumption that each gene expression is the sum of each cell population's contribution weighted by their corresponding relative frequency in the sample.

However, none of these methods account for the transcriptomic covariance structure and address the crucial problem of co-expression between distinct genes. The first goal of our project aims at studying the impact of correlation structures in the quality of the estimation performed by canonical deconvolution algorithms that assume *iid* distributions between the genes and use a fixed averaged expression profile for each cell type. The transcriptomic pathways were learnt from publicly purified cell data only, hypothesising that the network structure was sparse. Direct connections between the genes are represented for each population by non-zeros entries, learnt by plugging in the *MLE* covariance estimate, with zeros inputs shrunk by the gLasso algorithm [3,4].

Then, we develop a new deconvolution method that model each purified cellular expression profile as a multivariate Gaussian distribution [5], whose covariance parameter is the *plugged-in* estimate learnt beforehand to reconstitute the bulk profile. Next, we will optimise the estimation of the cellular expression profiles, by determining the MLE optimising the associated convolution of density functions of purified multivariate Gaussian transcriptomic profiles. Finally, we will compare our method to standard deconvolution algorithms, showing its interest to supply estimates more faithful to the biological reality.

References

[1] Shai S. Shen-Orr and Renaud Gaujoux. Computational deconvolution: Extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology*, 5, 2013.

[2] Aaron Newman, Chih Liu, Michael Green, Andrew Gentles, Weiguo Feng, Yue Xu, Chuong Hoang, Maximilian Diehn, and Ash Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12, 2015.

[3] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 3:432–441, 2008.

[4] Joachim Dahl, Vwani Roychowdhury, and Lieven Vandenberghe. Maximum likelihood estimation of gaussian graphical models: Numerical implementation and topology selection. *Journal of Multivariate Analysis*, 2005.

[5] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.