DeCovarT, a R package for a robust deconvolution of cell mixture in transcriptomic samples using a multivariate Gaussian generative framework

Bastien CHASSAGNOL* Yufei Luo† Gregory Nuel‡ Etienne Becht§

Transcriptomic analyses have contributed greatly to a better understanding of the biological processes involved in the evolution of complex and versatile diseases. However, bulk transcriptomic analyses ignore the heterogenous contribution of diverse cell populations to samples heterogeneity. Thus, computational deconvolution methods have been developed to analyse the cellular composition of tissues. However, the performance of these algorithms is limited in distinguishing between cell populations with very similar expression profiles, and we hypothesised that integrating the covariance between genes could enhance the performance of deconvolution algorithms for closely related cell populations. We therefore developed a new deconvolution algorithm, DeCovarT, which takes into account the transcriptomic network structure of each cell population. To do so, we represented the set of transcriptomic interactions as a multivariate Gaussian distribution, assuming a sparse network structure deduced from the precision matrix returned by the gLasso algorithm. Next, we reconstruct the overall mixing profile by a generative model, in which we show, under reasonable assumptions, that the law describing the overall expression profile conditional on the cell ratios and purified expression profiles also follows a multivariate Gaussian distribution. The maximum likelihood estimate (MLE) of the associated function, i.e. the cell ratios optimising the probability of observing the observed transcriptomic distribution, is estimated in our paper by first reparametrising the log-likelihood function into an unconstrained version and then optimising it by consecutive iterations of the Levenberg-Marquardt algorithm. This allows us to obtain an estimator that respects the simplex constraint and to derive the corresponding asymptotic confidence bands. In addition to the introduction of a new statistical modelling paradigm, we plan in our presentation to briefly review the standard optimisation methods implemented in R with their specific features and main restrictions. Notably, we benchmarked them on a toy example that highlights strong behavioural differences in the context of constrained optimisation.

Mots-clefs : cellular deconvolution – gLasso – generative model – bulk RNA Sequencing – Levenberg-Marquard – constrained optimisation

Introduction

The analysis of the bulk transcriptome provided new insights on the mechanisms underlying disease development. However, such methods ignore the intrinsic cellular heterogeneity of complex biological samples, by averaging measurements over several distinct cell populations. Failure to account for changes of the cell composition is likely to result in a loss of *specificity* (genes mistakenly identified as differentially expressed, while they only reflect an increase in the cell population naturally producing them) and *sensibility* (genes

*LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), bastien.chassagnol@upmc.fr
†Les Laboratoires Servier, IDRS, yufei.luo@servier.com
‡LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), Gregory.Nuel@math.cnrs.fr
§Les Laboratoires Servier, IRIS, etienne.becht@servier.com

expressed by minor cell populations are amenable being masked by highly variable expression from major cell populations).

Accordingly, a range of computational methods have been developed to estimate cellular fractions, but they perform poorly in discriminating cell types displaying high phenotypic proximity. Indeed, most of them assume that purified cell expression profiles are fixed observations, omitting the variability and intrinsically interconnected structure of the transcriptome. In contrast to these approaches, we hypothesised that integrating the pairwise covariance of the genes into the reference transcriptome profiles could enhance the performance of transcriptomic deconvolution methods. We therefore introduce *DeCovarT* (Deconvolution using the Transcriptomic Covariance), a new holistic probabilistic approach.

Objectives

Rationale of the new generative model

As in most traditional deconvolution models, we assume that the overall measured gene expression can be reconstructed by summing the individual contributions of each cell population weighted by its frequency. Formally, let $\boldsymbol{X} = (x_{gj}) \in \mathcal{M}_{\mathbb{R}^{G\times J}}$ the signature matrix representing the purified transcriptomic profiles of J cell populations and $\boldsymbol{p} = (p_{ji}) \in]0,1[^{J\times N}$ the unknown relative proportions of cell populations in N samples, then the linear relation relating the bulk expression ($\boldsymbol{y} = (y_{gi}) \in \mathbb{R}_+^{G\times N}$ to the individual cell expression profiles is given by the matrix product: $\boldsymbol{y} = \boldsymbol{X} \times \boldsymbol{p}$.

However, in real conditions with technical and environmental variability, the strict linearity of the deconvolution does not strictly hold. Thus, an additional error term is usually added, assumed to follow a *homoscedastic* zero-centred Gaussian distribution and with pairwise independent response measures while the exogenous variables (here, the purified expression profiles) are supposed determined: this set of conditions is referred to as the Gaussian-Markow assumptions. In that configuration, the MLE (maximum likelihood estimate) that bast describes this standard linear model is equal to the ordinary least squares (OLS) estimate (subfigure 1a).

In contrast to this canonical approach, in DeCovarT, we relax the *exogeneity* property by treating exogenous variables $\boldsymbol{X}$ as random variables rather than determined measures, in a process close to the approach of the DSection algorithm [1]. However, to our knowledge, we are the first to weaken the independence assumption between observations by explicitly incorporating the intrinsic covariance structure of the transcriptome of each purified cell population. To do so, we conjecture that the G-dimensional vector $\boldsymbol{x}_j$ characterising the transcriptomic expression of each cell population follows a multivariate Gaussian distribution: $\boldsymbol{x}_j \sim \mathcal{N}_G(\boldsymbol{\mu}_{.j}, \boldsymbol{\Sigma}_j)$, with $\boldsymbol{\mu}_{.j}$ the mean purified transcriptomic expression and $\boldsymbol{\Sigma}_j$ the covariance matrix, that we constrain to be positive-definite and of full rank and that is inferred using the output of the gLasso algorithm [2] (subfigure 1b).

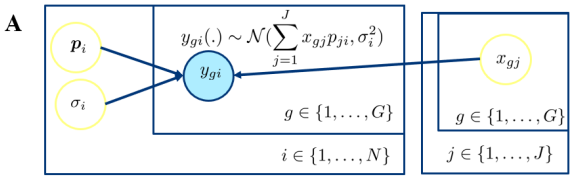
Derivation of the log-likelihood

First, we *plugged-in* the mean and covariance parameters $\zeta_j = \left(\boldsymbol{\mu}_{.j}, \boldsymbol{\Sigma}_j\right)$ inferred in the previous step. Then, by letting $\boldsymbol{\zeta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\quad \boldsymbol{\mu} = (\boldsymbol{\mu}_{.j})_{j\in\widetilde{J}} \in \mathcal{M}_{G\times J}$, $\quad \boldsymbol{\Sigma} \in \mathcal{M}_{G\times G}$ the known parameters and $\boldsymbol{p}$ the unknown cellular ratios, the conditional distribution $\boldsymbol{y}|(\boldsymbol{\zeta}, \boldsymbol{p})$ is the convolution of pairwise independent multivariate Gaussian distributions, which is also a multivariate Gaussian distribution 1, deduced from the *affine invariant* property of Gaussian distributions.

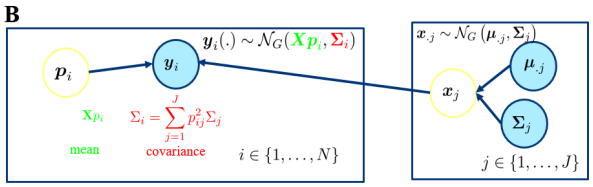
$$\boldsymbol{y}|(\boldsymbol{\zeta}, \boldsymbol{p}) \sim \mathcal{N}_G(\boldsymbol{\mu}\boldsymbol{p}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{\mu} = (\boldsymbol{\mu}_{.j})_{j\in\widetilde{J}}, \quad \boldsymbol{p} = (p_1, \ldots, p_J) \text{ and } \boldsymbol{\Sigma} = \sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j \tag{1}$$

From Equation 1, we readily compute the associated conditional log-likelihood (Equation 2):

$$\ell_{\boldsymbol{y}|\boldsymbol{\zeta}}(\boldsymbol{p}) = C + \log\left(\text{Det}\left(\sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j\right)^{-1}\right) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{p}\boldsymbol{\mu})^\top \left(\sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j\right)^{-1} (\boldsymbol{y} - \boldsymbol{p}\boldsymbol{\mu}) \tag{2}$$



(a) Standard linear model representation.



(b) The generative model used for the DeCovart framework.

Figure 1: We use the standard graphical convention of graphical models, as depicted in RevBayes webpage. For identifiability reasons, we conjecture that all variability arises from the stochastic nature of the covariates.

Iterated optimisation

The MLE is traditionally retrieved from the roots of the gradient of the log-likelihood. However, in our generative framework, cancelling the gradient of Equation @ref(eq:loglikelihood-multivariate-gaussian) reveals a non-closed form. Instead, iterated numerical optimisation algorithms can be used to proxy the roots, most of them considering first or second-order approximations of the function to optimise.

The *Levenberg-Marquardt algorithm* bridges the gap between between the steepest descent method (first-order) and the Newton-Raphson method (second-order) by inflating the diagonal terms of the Hessian matrix. Away from the endpoint, a second-order descent is favoured for its faster convergence pace, while the steepest approach is privileged close to the extremum, as it allows careful refinement of the step size. We use the implementation available in the **marqLevAlg** package. In particular, it introduces a stringent convergence criteria, the relative distance to the maximum (RDM), which sets apart extrema from spurious saddle points [3].

We provide additional theoretical results, such as analytical formulas for the Gradient and the Hessian in their constrained and unconstrained versions as well as simulation outputs in the vignette of the DeCovarT Github webpage.

References

[1] T. Erkkilä, S. Lehmusvaara, P. Ruusuvuori, T. Visakorpi, I. Shmulevich, and H. Lähdesmäki, "Probabilistic analysis of gene expression measurements from heterogeneous tissues," *Bioinformatics*, vol. 26, no. 20, pp. 2571–2577, Oct. 2010, doi: 10.1093/bioinformatics/btq406.

[2] R. Mazumder and T. Hastie, "The Graphical Lasso: New Insights and Alternatives," *Electronic Journal of Statistics*, vol. 6, Nov. 2011, doi: 10.1214/12-EJS740.

[3] M. Prague, D. Commenges, J. Guedj, J. Drylewicz, and R. Thiébaut, "NIMROD: A program for inference via a normal approximation of the posterior in models with random effects based on ordinary differential equations," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 2, pp. 447–458, Aug. 2013, doi: 10.1016/j.cmpb.2013.04.014.