DeCovarT, a multidimensional probalistic model for the deconvolution of heterogeneous transcriptomic samples

Bastien Chassagnol, Grégory Nuel, Etienne Becht

DeCovarT: A Probabilistic and Multidimensional Framework for Cellular Deconvolution in Heterogeneous Biological Samples

Bastien Chassagnol[1,2,*], Grégory Nuel[2], Etienne Becht[1]

1 Institut De Recherches Internationales Servier (IRIS), FRANCE
2 LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), Sorbonne Université, 4, place Jussieu, 75252 PARIS, FRANCE

*** bastien_chassagnol@laposte.net**

Abstract

Although bulk transcriptomic analyses have greatly contributed to a better understanding of complex diseases, their sensibility is hampered by the highly heterogeneous cellular compositions of biological samples. To address this limitation, computational deconvolution methods have been designed to automatically estimate the frequencies of the cellular components that make up tissues, typically using reference samples of physically purified populations. However, they perform badly at differentiating closely related cell populations.

We hypothesised that the integration of the covariance matrices of the reference samples could improve the performance of deconvolution algorithms. We therefore developed a new tool, DeCovarT, that integrates the structure of individual cellular transcriptomic network to reconstruct the bulk profile. Specifically, we inferred the ratios of the mixture components by a standard maximum likelihood estimation (MLE) method, using the Levenberg-Marquardt algorithm to recover the maximum from the parametric convolutional distribution of our model. We then consider a reparametrisation of the log-likelihood to explicitly incorporate the simplex constraint on the ratios. Preliminary numerical simulations suggest that this new algorithm outperforms previously published methods, particularly when individual cellular transcriptomic profiles strongly overlap.

1 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 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. For instance, the gold-standard deconvolution algorithm *CIBERSORT* [New15] applies nu-support vector regression (ν-SVR) to recover the minimal subset of the most informative genes in the purified signature matrix. However, this machine learning approach assumes that the transcriptomic expressions are independent.

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. The generative probabilistic model of our algorithm, *DeCovarT* (Deconvolution using the Transcriptomic Covariance), implements this integrated approach.

2 Model

First, we introduce the following notations:

- $(\boldsymbol{y} = (y_{gi}) \in \mathbb{R}_{+}^{G \times N}$ is the global bulk transcriptomic expression, measured in N individuals.
- $\boldsymbol{X} = (x_{gj}) \in \mathcal{M}_{\mathbb{R}^{G \times J}}$ the signature matrix of the mean expression of G genes in J purified cell populations.
- $\boldsymbol{p} = (p_{ji}) \in]0,1[^{J \times N}$ the unknown relative proportions of cell populations in N samples

As in most traditional deconvolution models, we assume that the total bulk expression can be reconstructed by summing the individual contributions of each cell population weighted by its frequency, as stated explicitly in the following linear matricial relationship (Equation (1)):

$$\boldsymbol{y} = \boldsymbol{X} \times \boldsymbol{p} \tag{1}$$

In addition, we consider unit simplex constraint on the cellular ratios, $\boldsymbol{p}$ (Equation (2)):

$$\begin{cases} \sum_{j=1}^{J} p_j = 1 \\ \forall j \in \widetilde{J} \quad p_j \geq 0 \end{cases} \tag{2}$$

2.1 Standard linear deconvolution model

However, in real conditions with technical and environmental variability, strict linearity of the deconvolution does not usually hold. Thus, an additional error term is usually considered, and without further assumption on the distribution of this error term, the usual approach to retrieve the best of parameters is by minimising the squared error term between the mixture expressions predicted by the linear model and the actual observed response. This optimisation task is achieved through the ordinary least squares (OLS) approach (Equation (3)),

$$\hat{\boldsymbol{p}}_i^{\mathrm{OLS}} \equiv \underset{\boldsymbol{p}_i}{\arg\min} ||\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i||^2 = \underset{\boldsymbol{p}_i}{\arg\min} ||\boldsymbol{X}\boldsymbol{p}_i - \boldsymbol{y}_i||^2 = \sum_{g=1}^{G} \left(y_{gi} - \sum_{j=1}^{J} x_{gj} p_{ji} \right) \tag{3}$$

If we additionally assume that the stochastic error term follows a *homoscedastic* zero-centred Gaussian distribution and that the value of the observed covariates (here, the purified expression profiles) is determined (see the corresponding graphical representation in Figure 1a and the set of equations describing it Equation (4)),

$$y_{gi} = \sum_{j=1}^{J} x_{gj} p_{ji} + \epsilon_i, \quad y_{gi} \sim \mathcal{N}\left(\sum_{j=1}^{J} x_{gj} p_{ji}, \sigma_i^2\right), \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2) \tag{4}$$

then, the MLE is equal to the OLS, which, in this framework, is given explicitly by Equation (5):

$$\hat{\boldsymbol{p}}_i^{\text{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}_i \tag{5}$$

and is known under the the Gauss-Markov theorem.

2.2 Motivation of using a probabilistic convolution framework

In contrast to standard linear regression models, we relax in the DeCovarT modelling framework the *exogeneity* assumption, by considering the set of covariates $\boldsymbol{X}$ as random variables rather than fixed measures, in a process close to the approach of DSection algorithm and DeMixt algorithms. However, to our knowledge, we are the first to weaken the independence assumption between observations by explicitly considering a multivariate distribution and integrating 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, given by Equation (6):

$$\text{Det}(2\pi\boldsymbol{\Sigma}_j)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_{.j})\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_{.j})^\top\right) \tag{6}$$

and parametrised by:

- $\boldsymbol{\mu}_{.j}$, the mean purified transcriptomic expression of cell population j
- $\boldsymbol{\Sigma}_j$, the covariance matrix of each cell population, constrained to be *positive-definite* (see Appendix A.1). Precisely, we retrieve it from inferring its inverse, known as the precision matrix, through the gLasso [Maz11] algorithm. We define $\boldsymbol{\Theta}_j \equiv \boldsymbol{\Sigma}_j^{-1}$ the corresponding *precision matrix*, whose inputs, after normalisation, store the partial correlation between two genes, conditioned on all the others. Notably, pairwise gene interactions whose corresponding off-diagonal terms in the precision matrix are null are considered statistically spurious, and discarded.

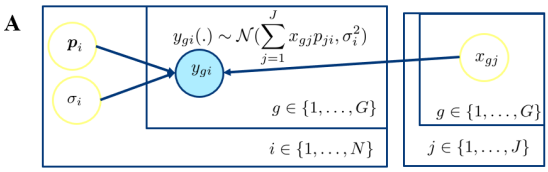
To derive the log-likelihood of our model, first we *plugged-in* the mean and covariance parameters $\zeta_j = \left(\boldsymbol{\mu}_{.j}, \boldsymbol{\Sigma}_j\right)$ estimated for each cell population in the previous step. Then, setting $\boldsymbol{\zeta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = (\boldsymbol{\mu}_{.j})_{j\in\widetilde{J}} \in \mathcal{M}_{G\times J}, \quad \boldsymbol{\Sigma} \in \mathcal{M}_{G\times G}$ the known parameters and $\boldsymbol{p}$ the unknown cellular ratios, we show that the conditional distribution of the observed bulk mixture, conditioned on the individual purified expression profiles and their ratios in the sample, $\boldsymbol{y}|(\boldsymbol{\zeta}, \boldsymbol{p})$, is the convolution of pairwise independent multivariate Gaussian distributions. Using the *affine invariance* property of Gaussian distributions, we can show that this convolution is also a multivariate Gaussian distribution, given by Equation (7).

$$\boldsymbol{y}|(\boldsymbol{\zeta}, \boldsymbol{p}) \sim \mathcal{N}_G(\boldsymbol{\mu}\boldsymbol{p}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{\mu} = (\boldsymbol{\mu}_{.j})_{j\in\widetilde{J}}, \quad \boldsymbol{p} = (p_1, \ldots, p_J) \text{ and } \boldsymbol{\Sigma} = \sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j \tag{7}$$

The DAG associated to this modelling framework is shown in Figure 1b).

In the next section, we provide an explicit formula of the log-likelihood of our probabilistic framework, its gradient and hessian, which in turn can be used to retrieve the MLE of our distribution.

(a) Standard linear model representation.



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

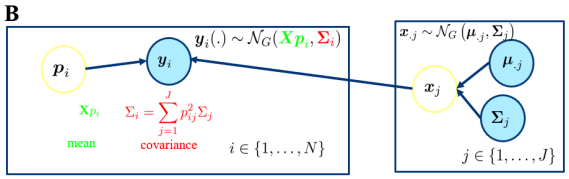


Figure 1. We use the standard graphical convention of graphical models, as depicted in RevBayes webpage. For identifiability reasons, we conjecture that all variability proceeds from the stochastic nature of the covariates.

2.3 Derivation of the log-likelihood

From Equation (7), the conditional log-likelihood is readily computed and given by Equation (8):

$$\ell_{\boldsymbol{y}|\boldsymbol{\zeta}}(\boldsymbol{p}) = C + \log\left(\operatorname{Det}\left(\sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j\right)^{-1}\right) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{p}\boldsymbol{\mu})^\top \left(\sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j\right)^{-1} (\boldsymbol{y} - \boldsymbol{p}\boldsymbol{\mu}) \tag{8}$$

2.4 First and second-order derivation of the unconstrained DeCovarT log-likelihood function

The stationary points of a function and notably maxima, are given by the roots (the values at which the function crosses the x-axis) of its gradient, in our context, the vector: $\nabla\ell : \mathbb{R}^J \to \mathbb{R}^J$ evaluated at point $\nabla\ell(\boldsymbol{p}) :]0,1[^J \to \mathbb{R}^J$. Since the computation is the same for any cell ratio p_j, we give an explicit formula for only one of them (Equation (9)):

$$\begin{aligned}
\frac{\partial \ell_{\boldsymbol{y}|\boldsymbol{\zeta}}(\boldsymbol{p})}{\partial p_j} &= \tfrac{\partial \log(\operatorname{Det}(\boldsymbol{\Theta}))}{\partial p_j} - \tfrac{1}{2}\left[\tfrac{\partial(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top}{\partial p_j}\boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p}) + (\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \tfrac{\partial \boldsymbol{\Theta}}{\partial p_j}(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p}) + (\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta} \tfrac{\partial(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})}{\partial p_j}\right] \\
&= -\operatorname{Tr}\left(\boldsymbol{\Theta}\tfrac{\partial \boldsymbol{\Sigma}}{\partial p_j}\right) - \tfrac{1}{2}\left[-\boldsymbol{\mu}_{.j}^\top \boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p}) - (\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \Theta \tfrac{\partial \boldsymbol{\Sigma}}{\partial p_j}\Theta(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p}) - (\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\mu}_{.j}\right] \\
&= -2p_j \operatorname{Tr}\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_j\right) + (\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\mu}_{.j} + p_j(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})
\end{aligned} \tag{9}$$

Since the solution to $\nabla\left(\ell_{\boldsymbol{y}|\boldsymbol{\zeta}}(\boldsymbol{p})\right) = 0$ is not closed, we had to approximate the MLE using iterated numerical optimisation methods. Some of them, such as the Levenberg–Marquardt algorithm, require a second-order approximation of the function, which needs the computation of the Hessian matrix. Deriving once more Equation (9) yields the Hessian matrix, $\mathbf{H} \in \mathcal{M}_{J\times J}$ is given by:

$$\begin{aligned}
\mathbf{H}_{i,i} = \frac{\partial^2 \ell}{\partial^2 p_i} &= -2\operatorname{Tr}\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_i\right) + 4p_i^2 \operatorname{Tr}\left(\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_i\right)^2\right) - 2p_i(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - \boldsymbol{\mu}_{.i}^\top \boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - \\
&\quad 2p_i(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - (\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\left(4p_i^2\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_i\right)\boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p}), \quad i \in \widetilde{J} \\
\mathbf{H}_{i,j} = \frac{\partial^2 \ell}{\partial p_i \partial p_j} &= 4p_j p_i \operatorname{Tr}\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}\boldsymbol{\Sigma}_i\right) - 2p_i(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{.j} - \boldsymbol{\mu}_{.i}^\top \boldsymbol{\Theta}\boldsymbol{\mu}_{.j} - \\
&\quad 2p_j(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - 4p_i p_j(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu}\boldsymbol{p}), \quad (i,j) \in \widetilde{J}^2, i \neq j
\end{aligned} \tag{10}$$

in which the coloured sections pair one by one with the corresponding coloured sections of the gradient, given in Equation (9).

Matrix calculus can largely ease the derivation of complex algebraic expressions, thus we remind in Appendices A.1 and A.2 relevant matrix properties and derivations. The numerical consistency of these derivatives was asserted with the **numDeriv** package, using the more stable Richardson's extrapolation ([For81]).

However, the explicit formulas for the gradient and the hessian matrix of the log-likelihood function, given in Equation (9) and Equation (10) respectively, do not take into account the simplex constraint assigned to the ratios. While some optimisation methods use heuristic methods to solve this problem, we consider alternatively a reparametrised version of the problem, detailed comprehensively in Appendix A.3.

3 Simulations

3.1 Simulation of a convolution of multivariate Gaussian mixtures

To assert numerically the relevance of accounting the correlation between expressed transcripts, we designed a simple toy example with two genes and two cell proportions. Hence, using the simplex constraint (Equation (2)), we only have to estimate one free unconstrained parameter, θ_1, and then uses the mapping function, defined in Appendix A.3 to recover the ratios in their original space.

We simulated the bulk mixture, $\boldsymbol{y} \in \mathcal{M}_{G\times N}$, for a set of artificial samples $N = 500$, with the following generative model:

- We have tested two levels of cellular ratios, one with equi-balanced proportions ($\boldsymbol{p} = (p_1, p_2 = 1 - p_1) = (\frac{1}{2}, \frac{1}{2})$ and one with highly unbalanced cell populations: $\boldsymbol{p} = (0.95, 0.05)$.
- Then, each purified transcriptomic profile is drawn from a multivariate Gaussian distribution. We compared two scenarios, playing on the mean distance of centroids, respectively $\mu_{.1} = (20, 22), \mu_{.2} = (22, 20)$ and $\mu_{.2} = (20, 40), \mu_{.2} = (40, 20)$) and building the covariance matrix, $\boldsymbol{\Sigma} \in \mathcal{M}_{2\times 2}$ by assuming equal individual variances for each gene (the diagonal terms of the covariance matrix, $\text{Diag}(\boldsymbol{\Sigma_1}) = \text{Diag}(\boldsymbol{\Sigma_1}) = \boldsymbol{I}_2$) but varying the pairwise correlation between gene 1 and gene 2, $\mathbb{C}\text{ov}\left[x_{1,2}\right]$, on the following set of values: $\{-0.8, -0.6, \ldots, 0.8\}$ for each of the cell population.
- As stated in Equation (1), we assume that the bulk mixture, $\mathbf{y}_{.i}$ could be directly reconstructed by summing up the individual cellular contributions weighted by their abundance, without additional noise.

3.2 Iterated optimisation

The extremum, and by extension the MLE, is a root of the gradient of the log-likelihood. However, in our generative framework, the inverse function cancelling the gradient of Equation Equation (8) is non-closed. Instead, iterated numerical optimisation algorithms that consider first or second-order approximations of the function to optimise are used to approximate the roots.

The *Levenberg-Marquardt (LM)* algorithm ([Lev44]) 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. Far from the endpoint, a second-order descent is favoured for its faster convergence pace, while the steepest approach is privileged close to the extremum since it allows careful refinement of the step size. Specially, we used the LM implementation of R package `marqLevAlg` to infer estimates of the cellular ratios from the bootstrap simulations ([Phi+21]). It notably includes an additional convergence criteria, the relative distance to the maximum (RDM), that sets apart extrema from spurious saddle points.

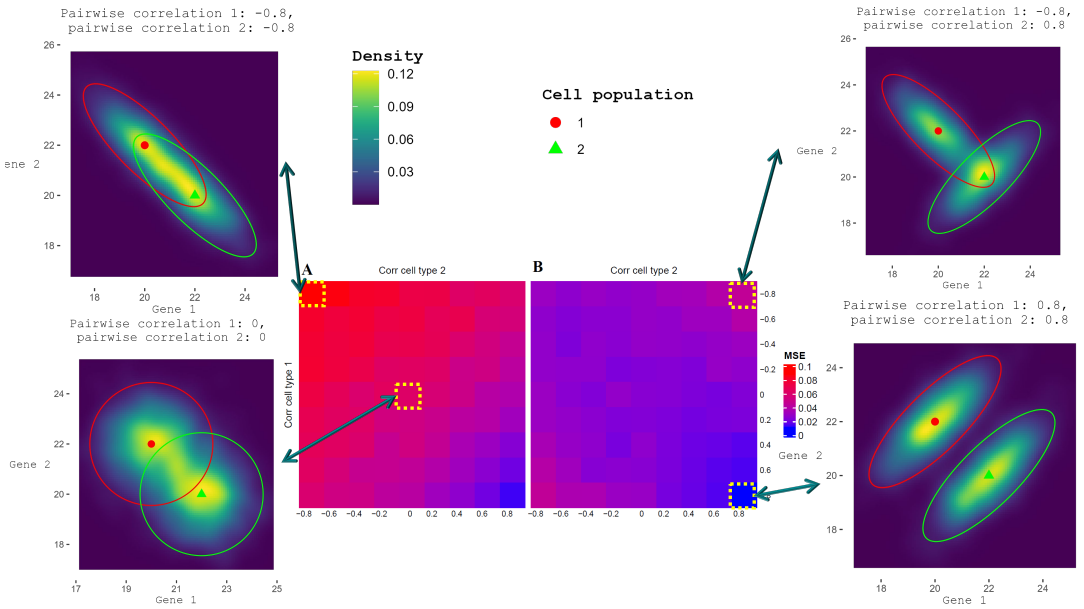


Figure 2. We used the package **ComplexHeatmap** to display the mean square error (MSE) of the estimated cell ratios, comparing the NNLS output, as implemented in the DeconRNASeq algorithm ([Gon13]), in Panel **A**, with our newly implemented DeCovarT algorithm, in Panel **B**. The lower the MSE, the least noisy and biased the estimates. In addition, we added the two-dimensional density plot for the intermediate scenario, for which each population is parameterised by a diagonal covariance matrix, and the most extreme scenarios (those with the highest correlation between genes). The ellipsoids represent for each cell population the 95% confidence region and the red spherical icon and the green triangular icon represent respectively the centroids (average expression of gene 1 and gene 2) of cell population 1 and cell population 2.

3.3 Results

We compared the performance of DeCovarT algorithm with the DeconRNASeq algorithm ([Gon13]).

Even with a limited toy example including two cell populations characterised only by two genes, we observe that the overlap was a good proxy of the quality of the estimation: the less the overlap between the two cell distributions, the better the quality of the estimation Figure 2.

The package used to generate the simulations and infer ratios from virtual or real biological mixtures with the DeCovarT algorithm is implemented on my personal Github account DeCovarT.

4 Perspectives

The new deconvolution algorithm that we implemented, DeCovarT, is the first one based on a multivariate generative model while enforcing explicitly the simplex constraint. Hence, it provides a strong basis to further derive statistical tests to assert whether the proportion of a given cell population differs significantly between two distinct biological conditions.

Extend the Simulation Framework To evaluate the biological and statistical interest of `DeCovarT`, we need to expand the simulation framework, by encompassing a larger number of cell types, genes, and testing the sensitivity of the model by voluntarily including noise in the benchmark evaluation.

The next phase of our evaluation involves real-world experiments, encompassing both blood and solid tumoral samples. To that end, we could start from the Kassandra benchmark, by [Zai22]. This large-scale project evaluates the performance of five established gold-standard and signature-based deconvolution algorithms, including EPIC [Rac17], CIBERSORT [New15], CIBERSORTx [New+19], quanTIseq [Fin19], and ABIS [Mon+19]. The evaluation involves deconvolving six publicly available datasets annotated with both flow cytometry and bulk RNA-seq expression.

Enhanced Inference and Integration of Co-Expression Networks All the popular differential gene expression analyses, such as *limma + voom* ([Rit+15]), *EdgeR* ([RMS10]) and *Deseq2* ([Var+16]), tend to overlook gene-gene interactions, comparing independently the expression between two conditions for each gene. The usual univariate approach of DGEAs additionally underlies the need of adjusting $p-$values, as numerous genes are examined simultaneously, and without accounting for interactions between them, the probability of observing false positives increases.

To account for correlations among observations, two consecutive papers, by [CL23] and [CCB22], present an innovative Bayesian framework which models proteomic expression across diverse biological conditions as multivariate Gaussian distributions. Insightful discussions with the main author, Marie Chion, suggest a straightforward extension of the method to transcriptomic expression, given the close relationship between two kinds of omics, both depicting counts.

While the methodology was originally designed to delineate differentially expressed genes between two conditions, the method can be readily extended to incorporate a *one-vs-all strategy*. This extension allows for the identification of markers specific to a particular cell population in comparison to all others. Furthermore, the generative model aligns closely with our deconvolution framework, leveraging the same distributions to describe cellular omic profiles. Alternatively, differential network approaches, such as INDEED, by [Zuo16], implement heuristic and dual-optimisation approaches, finding the sweet spot between maximising the mean differences between purified expression profiles and differentiating the neighbourhood network structure.

The gLasso algorithm used to derive the precision matrix associated to each purified cell profile is subjected to *parameter shrinkage*, like any penalty regularisation approach. Notably, in our setting, shrinkage tends to systematically underestimate the non-zero partial correlations of the precision matrix.

To mitigate this issue, one approach is to incorporate the *support* (indicating non-null inputs), derived from the gLasso output, into a conventional Maximum Likelihood Estimation (MLE) framework. The general concept is to utilising the true "zeros" to impose topological constraints on the final Gaussian Graphical Model (GGM). However, it's important to note that unless the undirected Markov network obtained from the gLasso output is a *chordal graph*, there is usually no straightforward mapping between the two topological spaces.

Finally, the inclusion of prior biological knowledge, such as the strength of relationships between transcription factors, retrieved from Protein-Protein Interaction (PPI) networks, can help reduce the exponential space of undirected graphs to explore.

Enhanced Inference and Integration of Co-Expression Networks All the methods outlined in Section 4 yield a subset of genes that distinguish a particular cell population from all others. However, when we combine these gene subsets, we often end up with a non-scalable signature matrix, presenting strong multicollinearity resulting from the redundancy between the gene markers identified.

To further refine the final set of genes able to delineating any cell population included in the signature matrix, *AutoGeneS*, by [Ali21], introduces a greedy genetic approach coupled with a dual optimisation approach [1]. Precisely, the loss function involves minimising inter-population correlation while simultaneously maximising the distance of the centroids.

We propose instead of this dual optimisation approach the minimisation of the *global overlap* between the concatenated distributions of the cellular profiles. Indeed, this metric not only captures in a single criterion the combined influence of mean inter-cluster distance and differential network structure in delineating purified cellular expression profiles, but supplies a straightforward score easy to interpret. The *overlap* metric precisely measures the shared probability mass and the degree of concurrence in probability densities. In simpler terms, it quantifies the global probability of incorrectly assigning an expression profile to the wrong cell subtype when utilising a maximum a posteriori approach, with the knowledge of each cellular profile's individual parameters.

[1] In standard approaches that rely on linear regression, the condition number serves as the gold-standard metric for assessing the level of precision of the linear model achievable with the design matrix

Joint Estimation of purified Expression Profiles and Cellular Ratios The generative model underlying the `DeCovarT` framework (Figure 1b) assumes that both the ratios and the purified cellular expression profiles are unobserved and need to be inferred from our model. However, we derived explicit formulas for the Gradient (eq. (9)) and Hessian (eq. (10)) of the associated log-likelihood function as if the purified expression profiles had been observed, by heuristically replacing the unknown and sample-specific purified expression profiles $\boldsymbol{X}_{.i}$ with their averaged counterparts $\boldsymbol{\mu}$. However, jointly optimising the cellular ratios and the purified expression profiles results in a non identifiable problem exhibiting an infinite number of solutions, without strong prior assumptions or regularisation of the unknown parameters to estimate. Finally, it's a highly intractable analytical task, and it is quite likely that no explicit form of the Gradient, nor the Hessian, could be derived.

We detail in Appendix B a Gibbs sampler to approximate the target distribution, here the joint value of the purified profiles and the cellular ratios. In addition, MCMC sampling allows for straightforward incorporation of prior knowledge, and streamlines the derivation of Maximum a Posteriori (MAP) estimates and *credible intervals*.

Precisely, by coupling Gibbs and Metropolis Hasting samplers, we ensure at each iteration that the estimated parameters adhered to the "balance condition", an essential property guaranteeing the convergence of MCMC chains to a stationary distribution identifiable to the target distribution.

References

[Lev44] Levenberg, Kenneth. "A method for the solution of certain non-linear problems in least squares". In: *Quarterly of Applied Mathematics* (1944). ISSN: 0033-569X, 1552-4485. DOI: 10.1090/qam/10666. URL: https://www.ams.org/qam/1944-02-02/S0033-569X-1944-10666-0/.

[For81] Bengt Fornberg. "Numerical Differentiation of Analytic Functions". In: *ACM Trans. Math. Softw.* (1981). DOI: 10.1145/355972.355979. URL: https://doi.org/10.1145/355972.355979.

[RMS10] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data". In: *Bioinformatics* (Jan. 1, 2010). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616. URL: https://doi.org/10.1093/bioinformatics/btp616.

[Maz11] Mazumder, Rahul and Hastie, Trevor. "The Graphical Lasso: New Insights and Alternatives". In: *Electronic Journal of Statistics* (2011). DOI: 10.1214/12-EJS740.

[Gon13] Gong, Ting and Szustakowski, Joseph D. "DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq Data". In: *Bioinformatics (Oxford, England)* (2013). DOI: 10.1093/bioinformatics/btt090.

[New15] Newman, Aaron and Liu, Chih and others. "Robust Enumeration of Cell Subsets from Tissue Expression Profiles". In: *Nature methods* (2015). DOI: 10.1038/nmeth.3337.

[Rit+15] Matthew E. Ritchie et al. "Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies". In: *Nucleic Acids Research* (Apr. 20, 2015). ISSN: 0305-1048. DOI: 10.1093/nar/gkv007. URL: https://doi.org/10.1093/nar/gkv007.

[Var+16] Hugo Varet et al. "SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data". In: *PLOS ONE* (June 9, 2016). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0157022. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157022.

[Zuo16] Zuo, Yiming and Cui, Yi and others. "INDEED: Integrated Differential Expression and Differential Network Analysis of Omic Data for Biomarker Discovery". In: *Methods (San Diego, Calif.)* (2016). DOI: 10.1016/j.ymeth.2016.08.015.

[Rac17] Racle, Julien and de Jonge, Kaat and others. "Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data". In: *eLife* (2017). Ed. by Alfonso Valencia. DOI: 10.7554/eLife.26476.

[Fin19] Finotello, Francesca and Mayer, Clemens and others. "Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-seq Data". In: *Genome Medicine* (2019). DOI: 10.1186/s13073-019-0638-6.

[Mon+19] Gianni Monaco et al. "RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types". In: *Cell Reports* (2019). DOI: 10.1016/j.celrep.2019.01.041.

[New+19] Aaron M. Newman et al. "Determining cell type abundance and expression from bulk tissues with digital cytometry". In: *Nature Biotechnology* (2019). DOI: 10.1038/s41587-019-0114-2.

[Ali21] Aliee, Hananeh and Theis, Fabian J. "AutoGeneS: Automatic Gene Selection Using Multi-Objective Optimization for RNA-seq Deconvolution". In: *Cell Systems* (2021). DOI: 10.1016/j.cels.2021.05.006.

[Phi+21] Viviane Philipps et al. "Robust and Efficient Optimization Using a Marquardt-Levenberg Algorithm with R Package marqLevAlg". In: *The R Journal* (2021). ISSN: 2073-4859. DOI: 10.32614/RJ-2021-089. URL: http://arxiv.org/abs/2009.03840.

[CCB22] Marie Chion, Christine Carapito, and Frédéric Bertrand. "Accounting for Multiple Imputation-Induced Variability for Differential Analysis in Mass Spectrometry-Based Label-Free Quantitative Proteomics". In: *PLOS Computational Biology* (Aug. 29, 2022). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010420. URL: http://arxiv.org/abs/2108.07086.

[Zai22] Zaitsev, Aleksandr and Chelushkin, Maksim and others. "Precise Reconstruction of the TME Using Bulk RNA-seq and a Machine Learning Algorithm Trained on Artificial Transcriptomes". In: *Cancer Cell* (2022). DOI: 10.1016/j.ccell.2022.07.006.

[CL23] Marie Chion and Arthur Leroy. "A Bayesian Framework for Multivariate Differential Analysis Accounting for Missing Data". In: (July 18, 2023). DOI: 10.48550/arXiv.2307.08975. arXiv: 2307.08975 [stat]. URL: http://arxiv.org/abs/2307.08975 (visited on 02/01/2024). preprint.