DeCovarT, a generative model for the deconvolution of heterogeneous transcriptomic samples

17/06/2023

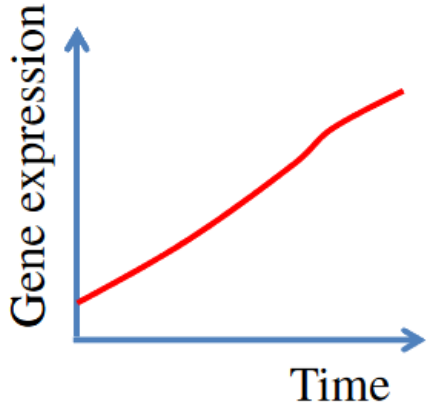
Bastien CHASSAGNOL
Gregory NUEL, Etienne BECHT, Yufei LUO

SERVIER
moved by you

Confusing biological noise

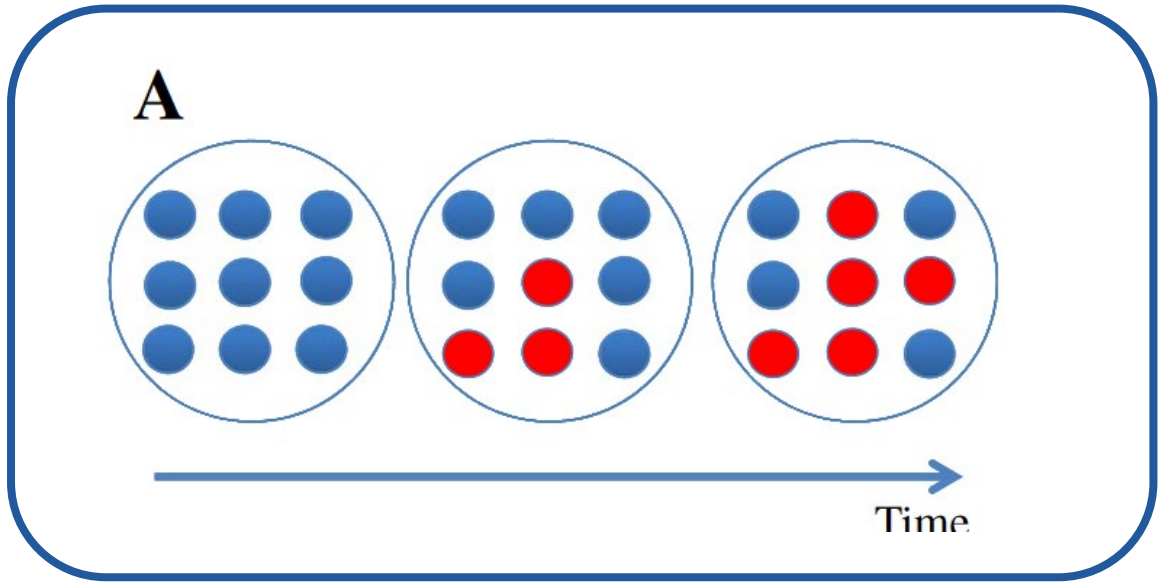
resting cell population 1

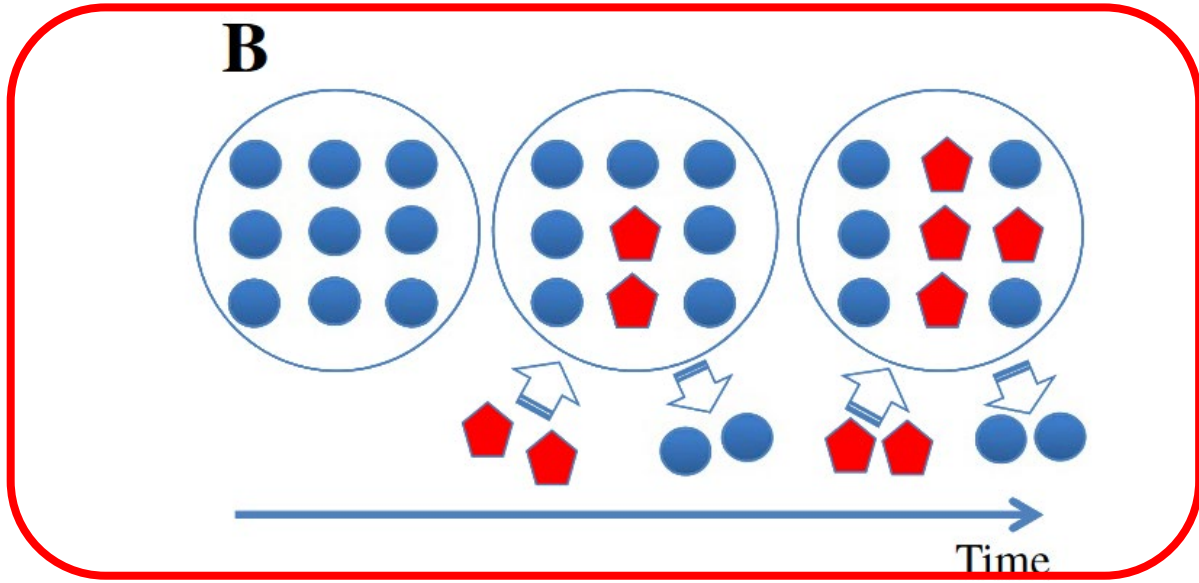
activated cell population 1



Shoemaker et al. 2012c

activated cell population 2

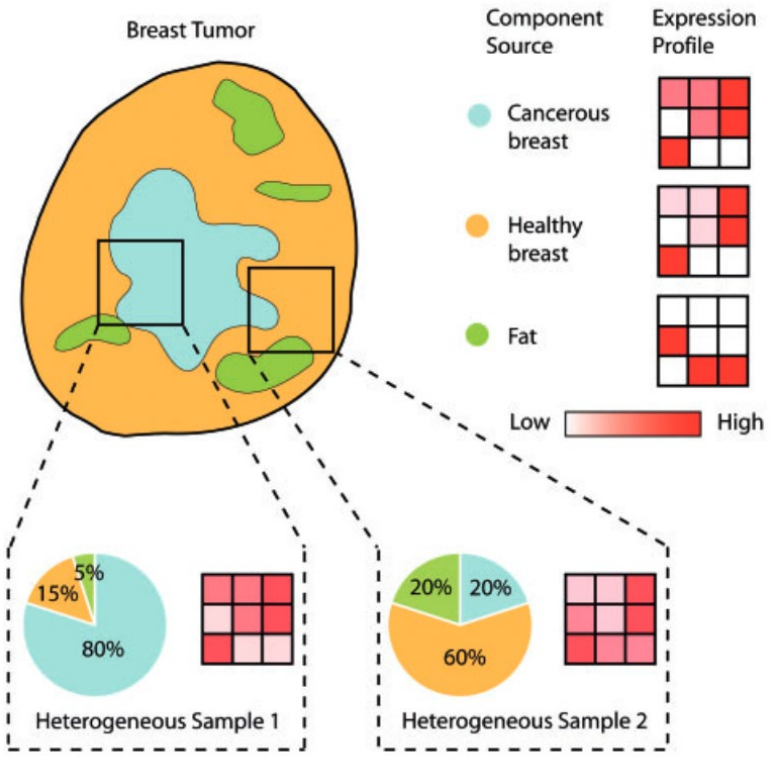




Scenario A: increase of the gene expression is generated by an **activation** of cell population 1

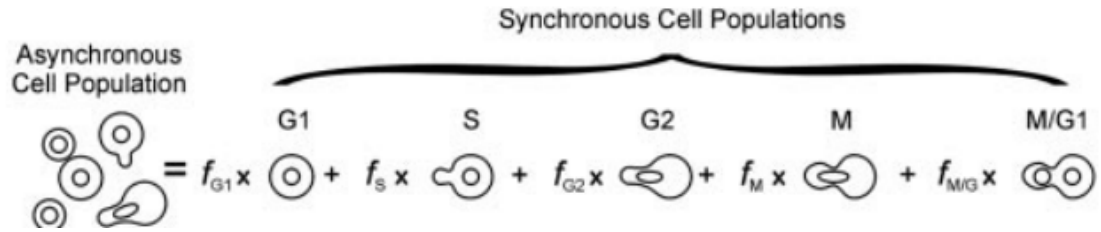
Scenario B: the gene expression increases due to the **arrival** of a **new** cell population 2

Heterogenity of tissues



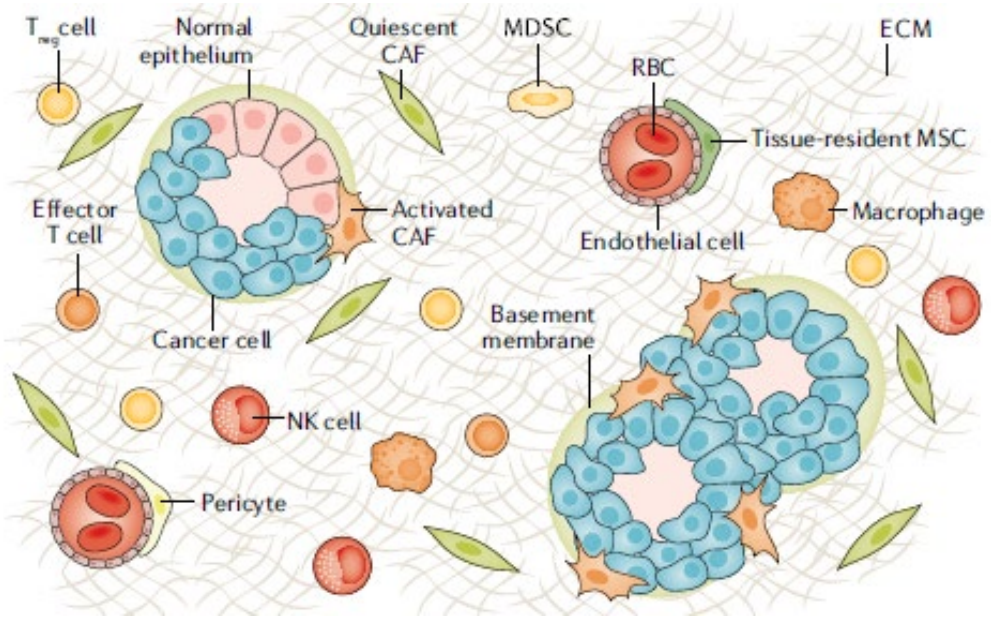
Mixture of tissues

Quon and Morris, 2009



Mixture of cell phases

Lu et al, 2003



Mixture of cell populations

Finotello and Trajanoski 2018

Deconvolution inputs

$$\begin{pmatrix} x_{1,1} & \cdots & x_{1,J} \\ \vdots & \ddots & \vdots \\ x_{G,1} & \cdots & x_{G,J} \end{pmatrix} \times \begin{pmatrix} p_{1,1} & \cdots & p_{1,N} \\ \vdots & \ddots & \vdots \\ p_{J,1} & \cdots & p_{J,N} \end{pmatrix} = \begin{pmatrix} y_{1,1} & \cdots & y_{1,N} \\ \vdots & \ddots & \vdots \\ y_{G,1} & \cdots & y_{G,N} \end{pmatrix}$$

X stores purified cellular expression profiles

p is the individual vector of cell ratios

Y stores the resulting bulk expression values

$$\begin{pmatrix} x_{G_1,1} & \cdots & 0 \\ 0 & x_{G_2,2} & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_{G_k,k} \end{pmatrix}$$

Marker-based

Linear box constraints

$$\boldsymbol{X} p_i = \boldsymbol{y}_i$$

$$\begin{cases} \sum_{j=1}^{J} p_{ji} = 1 \\ \forall j \in \{1, \ldots, J\}, \quad p_{ji} \geq 0 \end{cases}$$

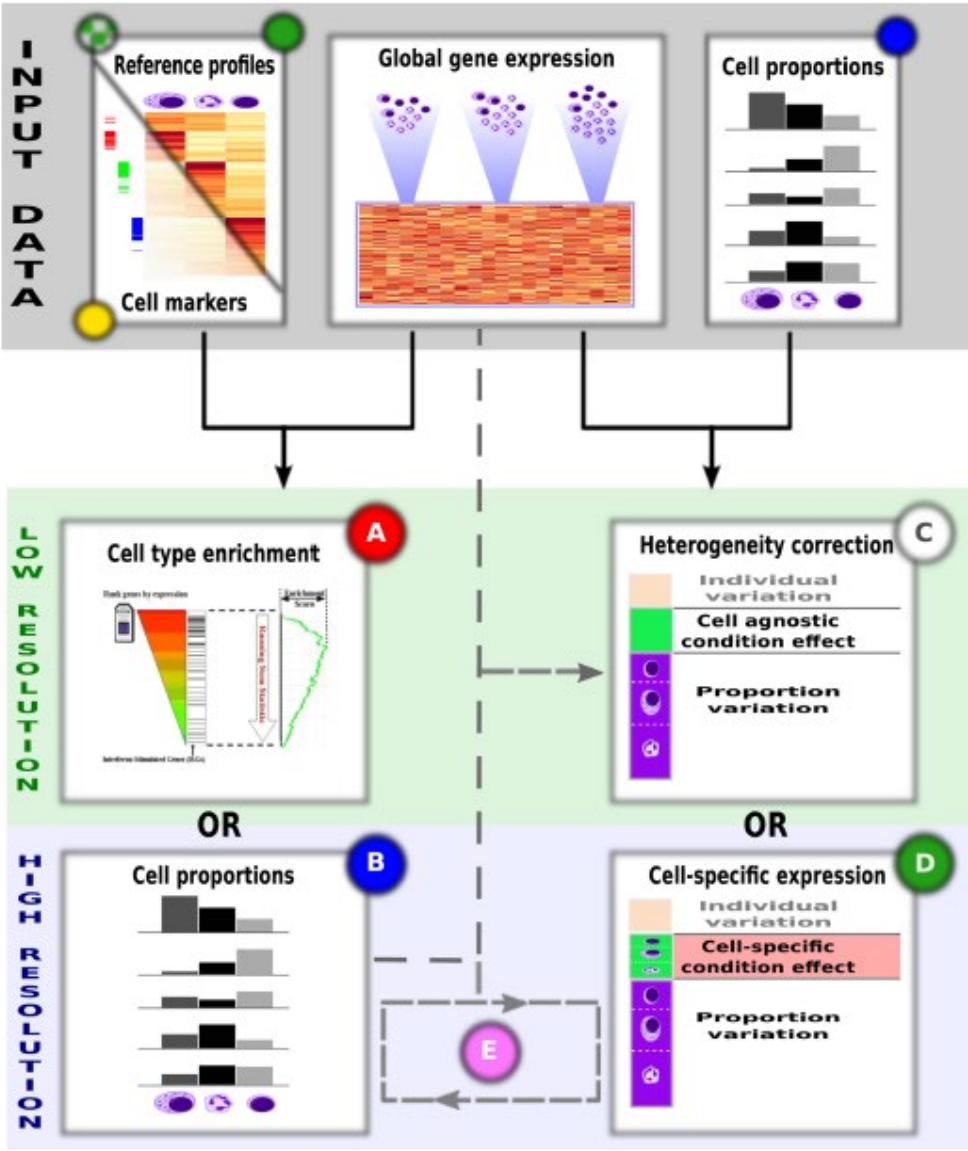
Deconvolution ecosystem

Partial deconvolution

- Estimate the ratios *p* for all individuals with the purified cell signature **X** and bulk mixture **y.**
- Try to infer cell specific expression profiles **X** based on **p** and **y.**

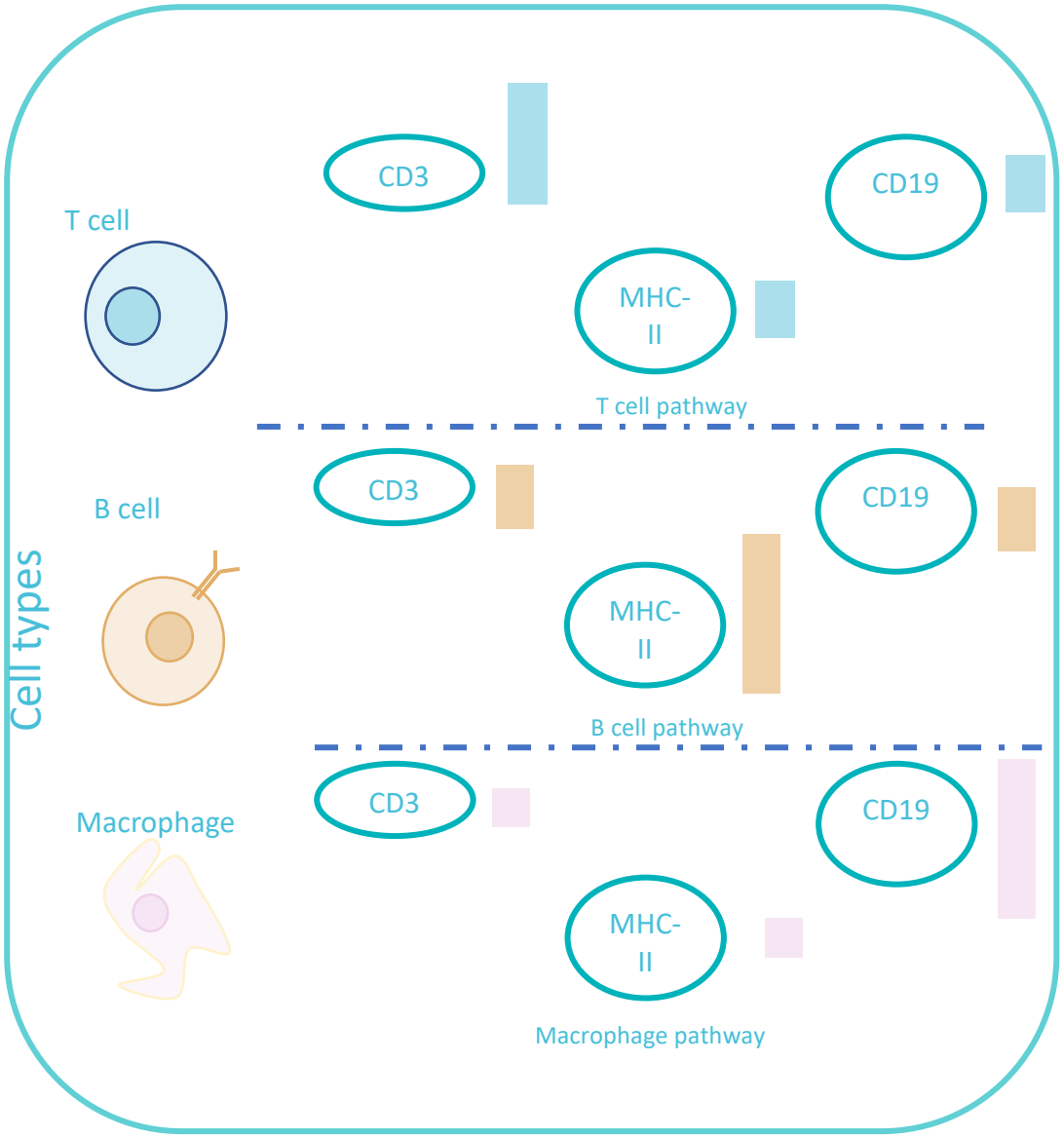
Complete deconvolution

- Try to infer alternatively both **p** and **X** (unsupervised, reference-free methods). Undetermined problem without prior.

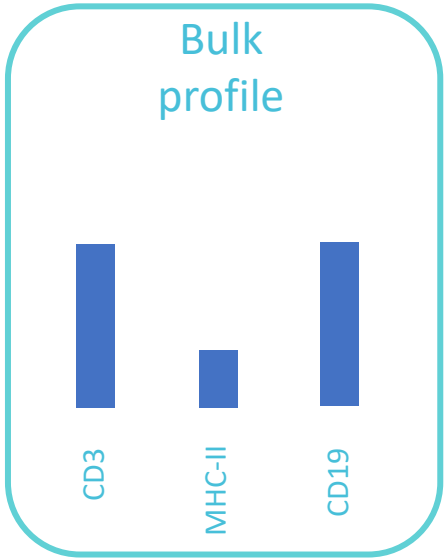


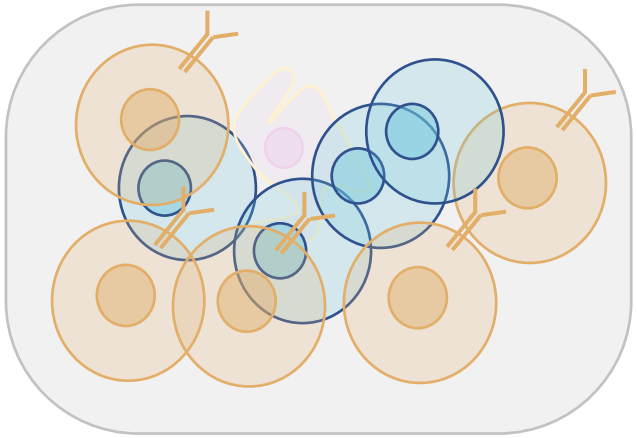
Shen-Orr et al, 2013

Standard deconvolution framework

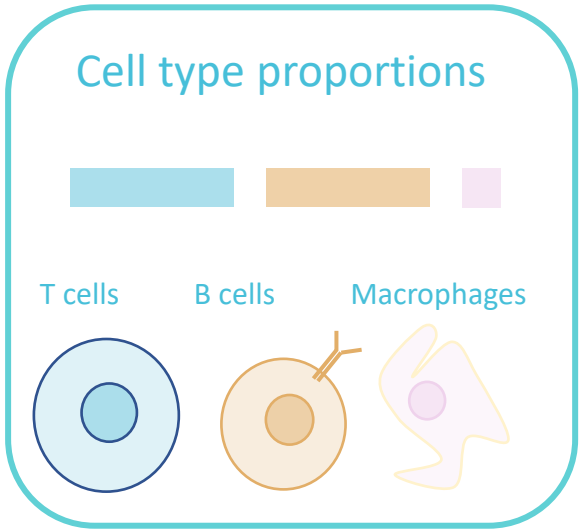


Purified cellular profiles

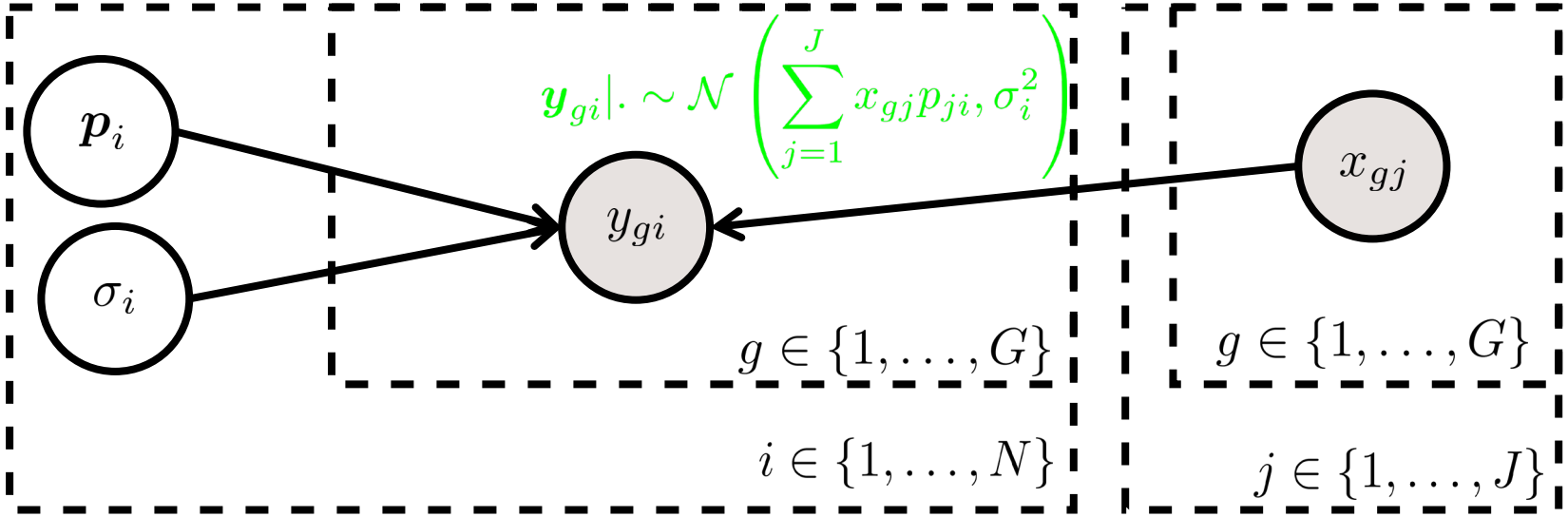




Standard deconvolution



Graphical model of linear regression



Nodes

Constant

Deterministic variable

Stochastic variable

Observations

Parameters

Estimated parameters

$\theta = (\boldsymbol{p}, \sigma)$

Distribution probabilities

Likelihood Laws

$f(\mathcal{D}|\theta)$

DeCovarT framework

Cell types

T cell

CD3

CD19

MHC-II

T cell pathway

B cell

CD3

CD19

MHC-II

B cell pathway

Macrophage

CD3

CD19

MHC-II

Macrophage pathway

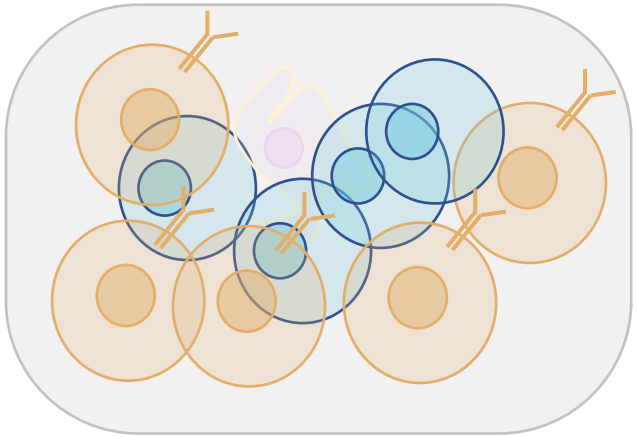
Purified cellular profiles

Bulk profile

CD3

MHC-II

CD19



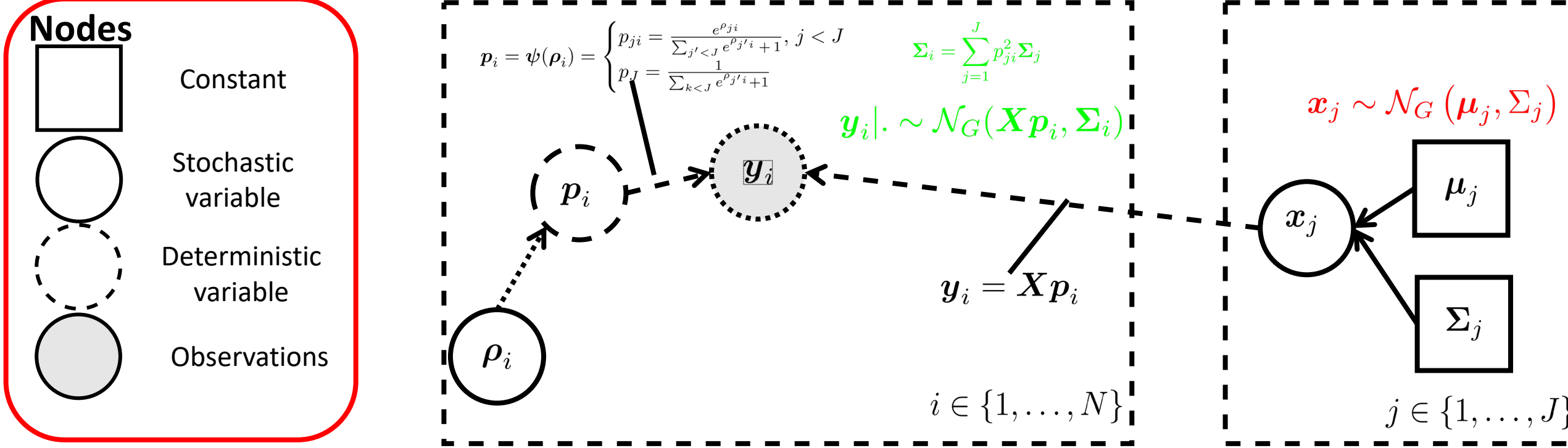
DeCovarT

Cell type proportions

T cells

B cells

Macrophages



Graphical model of DeCoVarT

Parameters

Prior parameters
$\zeta = (\mu, \Sigma)$

Estimated parameters
$\theta = (p, X)$

Distribution probabilities

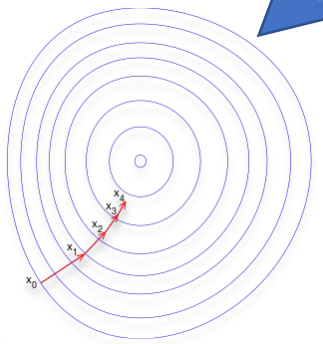
Prior laws
$f(\theta|\xi)$

Likelihood Laws
$f(\mathcal{D}|\theta)$

Optimisation algorithms in R

$$\ell_{\mathbf{y}|\mathbf{X},\Sigma}(\mathbf{p}) = C + \log\left(\det\left(\sum_{j=1}^{J} p_j^2 \Sigma_j\right)^{-1}\right) - \frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{p})^\top \left(\sum_{j=1}^{J} p_j^2 \Sigma_j\right)^{-1} (\mathbf{y}-\mathbf{X}\mathbf{p})$$

Log-likelihood associated to the model = log-likelihood of a multivariate Gaussian distribution



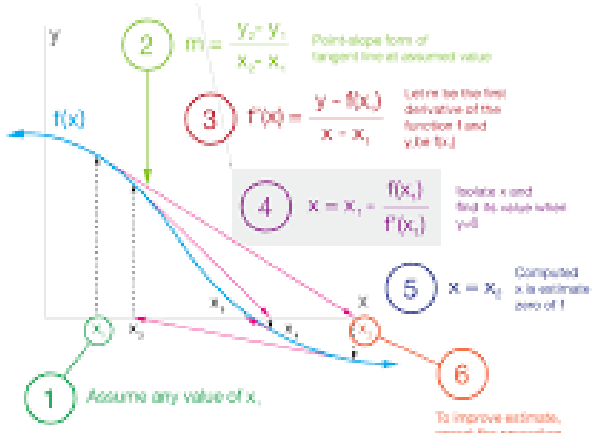
Steepest descent method

$$\theta^{(q+1)} = \theta^{(q)} + \gamma \nabla \ell\left(\theta^{(q)}\right)$$

Updated **Marquardt-Levenberg,** in package *marqLevAlg*

BP Hejblum, 2021, R Journal

$$\theta^{(q+1)} = \theta^{(q)} + \gamma_q \frac{\nabla \ell\left(\theta^{(q)}\right)}{\widetilde{\mathbf{H}}_\ell\left(\theta^{(q)}\right)}$$



Newton Raphson method

$$\theta^{(q+1)} = \theta^{(q)} + \frac{\nabla \ell\left(\theta^{(q)}\right)}{\mathbf{H}_\ell\left(\theta^{(q)}\right)}$$

#RR2023 Rencontres R AVIGNON 2023

LPSM

SERVIER