Tradict

Surojit Biswas

Contents

Tradict attempts to learn a small set of statistically representative *marker* genes from a collection of training transcriptomes. In prospective scenarios, where only these marker genes are measured, Tradict uses these sparse measurements to reconstruct the expression of biological pathways or the entire transcriptome as needed.

The rest of this document is organized as follows: In Section 1 we define key mathematical concepts and terms needed for what follows. In Section 2 we detail Tradict's underlying generative model. In Section 3, we then discuss model inference as it relates to encoding the transcriptome and decoding a sparse sampling of it.

1 Preliminaries

For a matrix A, $A_{:i}$ and $A_{i:}$ index the i^{th} column and row, respectively. For a set of indices, q, we use $-q$ to refer to all indices not specified by q.

We define the continuous relaxation of the Poisson distribution (hereafter, Continuous-Poisson) to have the following density function:

$$f(x|\lambda) = C_\lambda \frac{e^{-\lambda}\lambda^x}{\Gamma(x+1)}$$

where C_λ is a normalization constant.

Our collection n of training transcriptomes is represented by $x \in \mathbb{R}^{n \times g}$, where x_{ij} denotes the measured transcripts per million (TPM) of gene j in sample i.

2 Model

Tradict uses a Continuous-Poisson Multivariate Normal (CP-MVN) hierarchical model to model transcript abundances, measured as transcripts per million (TPM). Specifically, given a set of representative markers, Tradict first attempts to hierarchically model the expression of biological pathways given these marker abundances. It then models the expression of all genes in the transcriptome given the expression of these pathways.

We assume that for the TPM measured for gene j, t_j, there is an associated, unmeasured latent abundance z_j. This latent abundance (up to a increasing-monotonic transformation) represents the gene's true abundance, of which the measured TPM is a noisy realization. Let $s \in \mathbb{R}^G$ be a vector of *scores* (discussed below) representing the expression of a given set of G pathways, and let m be a set of indices for a given panel of representative markers.

We define the CP-MVN model as follows:

$$\begin{aligned} z_m &\sim \mathcal{N}\left(\mu^{(m)}, \Sigma^{(m)}\right) && \text{[Layer 4]} \\ s|z_m &\sim \mathcal{N}(\mu_{s|z_m}, \Sigma_{s|z_m}) && \text{[Layer 3]} \\ z_{-m}|s &\sim \mathcal{N}(\mu_{z_{-m}|s}, \Sigma_{z_{-m}|s}) && \text{[Layer 2]} \\ t_j &\sim \text{Continuous-Poisson}(\exp(z_j)) && \text{[Layer 1]} \end{aligned}$$

Here, $\mu^{(m)}$ and $\Sigma^{(m)}$ refer to mean vector and covariance matrix of z_m. Given these the conditional mean and covariance of the pathway scores and the latent abundances of all non-marker genes can be obtained through gaussian conditioning. Specifically, for two normally distributed variables a and b the conditional mean of b given a is given by $\mu_{b|a} = \mu_b + (a - \mu_a)\Sigma_a^{-1}\sigma_{ab}$ and $\Sigma_{b|a} = \Sigma_b - \sigma_{ab}^T\Sigma_a^{-1}\sigma_{ab}$, where σ_{ab} is the cross-covariance between a and b, and Σ_a and Σ_b are the covariance matrices of a and b, respectively.

3 Inference

During decoding, Tradict attempts to infer pathway expression scores and the transcriptome using only TPM measurements of the representative markers. We first note that due self-conjugacy of the Normal distribution Layers 2 and 3 in the CP-MVN model above can be collapsed. Therefore, the model may be partitioned into two separate models for pathway score prediction and transcriptome prediction. For pathway scores we have,

$$\begin{aligned} z_m &\sim \mathcal{N}\left(\mu^{(m)}, \Sigma^{(m)}\right) \\ s|z_m &\sim \mathcal{N}(\mu_{s|z_m}, \Sigma_{s|z_m}) \end{aligned}$$

For gene expression we have,

$$\begin{aligned} z_m &\sim \mathcal{N}(\mu^{(m)}, \Sigma^{(m)}) \\ z_{-m}|z_m &\sim \mathcal{N}(\mu_{z_{-m}|z_m}, \Sigma_{z_{-m}|z_m}) \\ t_j &\sim \text{Continuous-Poisson}(\exp(z_j)) \end{aligned}$$

Notice that predictions of both pathway scores and the remaining transcriptome depend on the latent abundances of the representative markers (not the measured abundances). Therefore, Tradict must first learn these latent abundances using the observed TPM measurements, t_m, and estimates of $\mu^{(m)}$ and $\Sigma^{(m)}$ it has learned from the training data. Once the marker latent abundances are known, Tradict can calculate expected marker scores and TPMs of non-marker genes.

3.1 Latent abundance inference for markers

Specifically, suppose Tradict has estimates of $\mu^{(m)}$ and $\Sigma^{(m)}$ given by $\hat{\mu}^{(m)}$ and $\hat{\Sigma}^{(m)}$. Given these and the measured marker TPMs, we can calculate a maximum *a posteriori* (MAP) estimate of z_m.

The posterior distribution over z_m is given by

$$\begin{aligned}
p(z_m|t_m, \hat{\mu}^{(m)}, \hat{\Sigma}^{(m)}) &= \frac{p(t_m|z_m, \hat{\mu}^{(m)}, \hat{\Sigma}^{(m)})p(z_m|\hat{\mu}^{(m)}, \hat{\Sigma}^{(m)})}{\int_k p(t_m|k, \hat{\mu}^{(m)}, \hat{\Sigma}^{(m)})p(k|\hat{\mu}^{(m)}, \hat{\Sigma}^{(m)})\mathrm{d}k} \\
&\propto \prod_{i=1}^{n} p(t_{im}|z_{im}, \hat{\mu}^{(m)}, \hat{\Sigma}^{(m)})p(z_{im}|\hat{\mu}^{(m)}, \hat{\Sigma}^{(m)}) \\
&= \prod_{i=1}^{n} \left[\prod_{j=1}^{|m|} C_{\exp(z_{ij})} \exp(z_{ij})^{t_{ij}} e^{-\exp(z_{ij})}/\Gamma(t_{ij}+1) \right] \\
&\times \frac{1}{\sqrt{2\pi|\hat{\Sigma}^{(m)}|}^{|m|}} \exp\left(-\frac{1}{2}(z_{i:} - \hat{\mu}^{(m)})\mathrm{inv}\left(\hat{\Sigma}^{(m)}\right)(z_{i:} - \hat{\mu}^{(m)})^T \right)
\end{aligned}$$

where for notational clarity we have used $\mathrm{inv}(\cdot)$ to represent matrix inverse.

Given z is a matrix parameter, this may be difficult to solve directly. However, note that given z_{ij}, t_{ij} is conditionally independent of t_{i-j}. Additionally, given z_{i-j}, z_{ij} is normally distributed with mean and covariance

$$\begin{aligned}
a_{ij} &= \mu_j^{(m)} + \left(z_{i,-j} - \mu_{-j}^{(m)} \right) \mathrm{inv}\left(\Sigma_{-j,-j}^{(m)} \right) \Sigma_{-j,j}^{(m)} \\
\sigma_j &= \Sigma_{j,j}^{(m)} - \Sigma_{j,-j}^{(m)} \mathrm{inv}\left(\Sigma_{-j,-j}^{(m)} \right) \Sigma_{-j,j}^{(m)}
\end{aligned}$$

respectively. Taken together, this suggests an iterative conditional modes algorithm in which we maximize the posterior one column of z at a time, while conditioning on all others.

Let $\hat{z}$ denote our current estimate of z. Our objective is given by,

$$\begin{aligned}
\hat{z}_{ij} &= \underset{z_{ij}|z_{i,-j}}{\mathrm{argmax}} \log p(z_{ij}|t_{ij}, \hat{z}_{i,-j}, \hat{\mu}^{(m)}, \hat{\Sigma}^{(m)}) \\
&= \underset{z_{ij}|z_{i,-j}}{\mathrm{argmax}} \log p(t_{ij}|z_{ij}, \hat{z}_{i,-j}, \hat{\mu}^{(m)}, \hat{\Sigma}^{(m)})p(z_{ij}|\hat{z}_{i,-j}, \hat{\mu}^{(m)}, \hat{\Sigma}^{(m)}) \\
&= \underset{z_{ij}|z_{i,-j}}{\mathrm{argmax}} \log p(t_{ij}|z_{ij})p(z_{ij}|\hat{z}_{i,-j}, \hat{\mu}^{(m)}, \hat{\Sigma}^{(m)}) \\
&= \underset{z_{ij}|z_{i,-j}}{\mathrm{argmax}} \log \left[\exp(z_{ij})^{t_{ij}} e^{-\exp(z_{ij})} \exp\left(-\frac{1}{2\sigma_j}(z_{ij} - a_{ij})^2 \right) \right] \\
&= \underset{z_{ij}|z_{i,-j}}{\mathrm{argmax}} \, t_{ij} z_{ij} - \exp(z_{ij}) - \frac{1}{2\sigma_j}(z_{ij} - a_{ij})^2
\end{aligned}$$

Differentiating we get,

$$\begin{aligned}
&\frac{\partial}{\partial z_{ij}} t_{ij} z_{ij} - \exp(z_{ij}) - \frac{1}{2\sigma_j}(z_{ij} - a_{ij})^2 \\
&\qquad = t_{ij} - \exp(z_{ij}) - \frac{1}{\sigma_j}(z_{ij} - a_{ij})
\end{aligned}$$

Because z_{ij} appears as a linear and exponential term, Tradict cannot solve this gradient analytically. It therefore utilizes Newton-Raphson optimization. For this it also requires the Hessian, which is given by,

$$\begin{aligned}
&\frac{\partial}{\partial z_{ij}} t_{ij} - \exp(z_{ij}) - \frac{1}{\sigma_j}(z_{ij} - a_{ij}) \\
&\qquad = -\exp(z_{ij}) - \frac{1}{\sigma_j} < 0
\end{aligned}$$

Notice the Hessian is always negative-definite, which implies each update has a single, unique optimum.

In practice, the Newton-Raphson updates can be performed in vectorized fashion iteratively for each column of z. We generally find that this optimization takes 5-15 iterations (full passes over all columns of z) and less than a minute to converge.

3.2 Prediction of pathway scores and TPM abundance of remaining genes

Given the learned latent marker abundances, Tradict can calculate expected pathway scores and the TPM abundances of the remaining non-marker genes in the transcriptome. For pathway scores we have,

$$\mathbb{E}[s|z_m] = \mu_{s|z_m} = \hat{\mu}_s + \left(z_m - \hat{\mu}^{(m)}\right) \text{inv}\left(\hat{\Sigma}^{(m)}\right) \hat{\sigma}_{z_m,s}.$$

Here, $\hat{\mu}_s$ and $\hat{\sigma}_{z_m,s}$ represent estimates of the unconditional mean of s and the cross-covariance matrix between z_m and s. These are learned from the training data.

For the remaining, non-marker genes in the transcriptome we have,

$$\begin{aligned}
\mathbb{E}[t_{ij}|z_{im}] &= \int_{-\infty}^{\infty} \mathbb{E}[t_{ij}|z_{ij}] p(z_{ij}|z_{im}) \mathrm{d}z_{ij} \\
&= \int_{-\infty}^{\infty} \exp(z_{ij}) \mathcal{N}(z_{ij}|\mu_{z_{ij}|z_{im}}, \Sigma_{z_{ij}|z_{im}}) \mathrm{d}z_{ij} \\
&= \mathbb{E}_{\mathcal{N}}[\exp(z_{ij})|z_{im}]
\end{aligned}$$

Recall the Moment Generating Function of a Normal random variable X with mean μ and variance σ^2 is given by $M(t) = \mathbb{E}[\exp(tX)] = \exp(\mu t + \sigma^2 t^2/2)$. Therefore we have,

$$\mathbb{E}[t_{ij}|z_{im}] = \mathbb{E}_{\mathcal{N}}[\exp(z_{ij})|z_{im}] = M(1) = \exp\left(\mu_{z_{ij}|z_{im}} + \frac{1}{2}\Sigma_{z_{ij}|z_{im}}\right)$$

where,

$$\begin{aligned}
\mu_{z_{ij}|z_{im}} &= \hat{\mu}_j + \left(z_{im} - \hat{\mu}^{(m)}\right) \text{inv}\left(\hat{\Sigma}^{(m)}\right) \hat{\sigma}_{z_m,z_j} \\
\Sigma_{z_{ij}|z_{im}} &= \hat{\sigma}_{jj} - \hat{\sigma}^T_{z_m,z_j} \text{inv}\left(\hat{\Sigma}^{(m)}\right) \hat{\sigma}_{z_m,z_j}
\end{aligned}$$

Here, $\hat{\mu}_j$ and $\hat{\sigma}_{z_m,z_j}$ represent estimates of the unconditional mean of z_j and the cross-covariance matrix between z_m and z_j. These are learned from the training data.

4 Properties of the CP-MVN hierarchical model

Mean-Variance and overdispersion. Noise correction/buffering. Skewness Useful properties? Noise correction of markers.