Notes on Finding Joint Latent Space of Networks and Traits

1 Brief Motivation and Summary

It is important to link a population of networks A_i (or multi-dimensional tensors in general, such as images) to their corresponding traits y_i over subject $i = 1, \ldots, n$. Especially there is a large interest in using A_i to predict y_i. Besides the direct tensor regression treatment, an empirical alternative would be a two-step approach: first obtain a low dimensional representation x_i for each A_i, then use x_i as predictors to predict y_i. However, this hurestics commonly does not work – because the finding of the low-dimension vector is unsupervised, there is no guarantee that it contains good signal about predicting y_i.

This motivates us to consider a supervised approach. Consider both A_i and y_i are realizations over a low-dimensional latent space, with x_i as the coordinate. Then x_i can be treated as the low-rank core tensor in a probabilistic Tucker decomposition of A_i, and as latent coordinate in a Gaussian process with y_i as the outcome. Despite of nonlinearity we introduced, as shown in the following, this joint modeling problem is well defined as a joint decomposition problem of two tensors, with a special factor structure in Gaussian process. The posterior sampling can be efficiently carried out using Hamiltonian Monte Carlo.

2 Matrix Factorization Form for Gaussian Process Latent Variable Model

Let y_{ij} be the continuous outcome for the ith subject and jth outcome, $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The Gaussian process latent variable model assumes that for each p-dimensional vector $y_i = [y_{i1}, y_{i2}, \ldots, y_{ip}]$, there is a latent coordinate $x_i \in \mathbb{R}^d$, such that y_i is a p-dimensional Gaussian process realized on these coordindate. Their goal is commonly to treat the

coordinate as nonlinear low dimensional representation for the data, when $d \ll p$.

Assuming the different outcomes for $j = 1, \ldots, p$ are independent, Gaussian process stipulates that, for any two y_{ij} and $y_{i'j}$ from two different subjects on the jth outcome, they follow a bivariate normal distribution with mean 0 and covariance $Cov(y_{ij}, y_{i'j}) = \phi_j K_j(||x_i - x_i'||) + \sigma^2 \delta_{i,i'}$, where $K_j(.)$ is a kernel that maps $\mathbb{R}^+ \to [0, 1]$ and ϕ_j is the scale parameter and σ_j^2 is the variance of independent measurement error.

Gaussian process is computationally challenging, however, with a spectral approximation over r frequencies, each vector of $y_{\cdot j}$ can be represented as:

$$\begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{nj} \end{bmatrix} = \begin{bmatrix} c_1(x_1) & s_1(x_1) & \ldots & \ldots & c_r(x_1) & s_r(x_1) \\ c_1(x_2) & s_1(x_2) & \ldots & \ldots & c_r(x_2) & s_r(x_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ c_1(x_n) & s_1(x_n) & \ldots & \ldots & c_r(x_n) & s_r(x_n) \end{bmatrix} \begin{bmatrix} \alpha_{1j} \\ \beta_{1j} \\ \vdots \\ \vdots \\ \alpha_{rj} \\ \beta_{rj} \end{bmatrix} + \begin{bmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \vdots \\ \epsilon_{nj} \end{bmatrix} \quad (1)$$

where $c_h(x_i)$ and $s_h(x_i)$ are the spectral bases over $h = 1, \ldots, r$ that depend on x_i, specifically, $c_h(x_i) = \cos(x_i' w_h)$ and $s_h(x_j) = \sin(x_i' w_h)$; w_h is uniformly distributed in $(0, \pi)^d$; $\alpha_{hj} \sim N(0, g(w_h))$ and $\beta_{hj} \sim N(0, g(w_h))$ with $g(w_h)$ as the spectral density corresponding to $\phi_j K_j(||x||)$, all $\alpha_{\cdot}$ and $\beta_{\cdot}$ are independent; $\epsilon_{\cdot j} \overset{iid}{\sim} N(0, \sigma_j^2)$. It can be verified that marginalizing out all α and β will yield an discrete Fourier transform of the spectral density, which is a good approximation for the Gaussian process covariance.

However, we do not marginalize out α and β, but treat them as parameters instead. This representation gives a **matrix factorization** form for Gaussian process. Combining all $y_{\cdot j}$ into a matrix $Y = \begin{bmatrix} y_{\cdot 1} & y_{\cdot 2} & \ldots & y_{\cdot p} \end{bmatrix} \in \mathbb{R}^{n \times p}$, we take the following factorization:

$$Y = C(X)\theta + E$$

where $C(X) \in \mathbb{R}^{n \times 2r}$ is the same matrix for spectral basis in (1) and $\theta \in \mathbb{R}^{2r \times p}$ is the matrix containing $\alpha_{\cdot j}$ and $\beta_{\cdot j}$ as defined above for its jth column; $E \in \mathbb{R}^{n \times p}$ is the measurement error.

3 Joint Model of Tucker Decomposition and Gaussian Process

We focus on Tucker decomposition as the low rank representation for A_i. For simplicity, we assume $A_i \in \{0,1\}^{m\times m}$ is a binary and symmetric adjacency matrix, and can be modeled as follows:

$$\begin{aligned} A_i &\sim \text{Bernoulli}(\frac{1}{1+\exp(-\psi_i)}) \\ \psi_i &= Z + U\tilde{x}_i U' \end{aligned} \tag{2}$$

where "$\sim$ Bernoulli" denotes the lower triagular elements of A_i are distributed elementwise as Bernoulli; $Z \in \mathbb{R}^{m\times m}$ is shared among all i's to induce low rank structure; $U \in \mathbb{R}^{m\times \tilde{d}}$. Most importantly, $\tilde{x}_i$ is a $\tilde{d}\times\tilde{d}$ symmetric matrix with the lower triangular part (including the diagonal) representing the latent coordinates. We denote $x_i = LowerTri(\tilde{x}_i)$.

Commonly, unsupervised Tucker is subject to rotation problem. This can be reduced by a supervised approach for searching a particular orientation best for Gaussian process modeling. Combining decomposition likelihood $L(A_i; x_i)$ with the Gaussian process factorization likelihood $L(y_i; x_i)$, leading to a model with joint density:

$$\pi_0(\theta)\prod_i L(A_i;x_i)L(y_i;x_i)\pi_0(x_i)$$

where we omit the other parameters $\theta = \{U, \alpha, \beta\}$ in the likelihood; $\pi_0(.)$ are the prior distribution.

As the dimension of x_i is undertermined. We further use a population-level shrinkage prior on $\pi_0(x_i)$, by letting $x_{il} \sim N(0, \tau_l^2)$ across all i, where τ_l^2 follows a Dirichlet-Laplace distribution. In the Tucker decompostion, this helps in the parsimony control of the core parameters. In the Gaussian process part, this allows us to use an simple isotropic correlation for the kernel, such as $\exp(-\frac{\sum_l (x_{il}-x_{i'l})^2}{\rho})$, since shrinking the sub-coorinates $x_{.l}$ toward a common 0 reduces the pairwise distance in the subspace, avoiding the effort of doing variable selection with more complicated kernel.