Bayesian Probabilistic Tensor Decompositions with Spike and Slab Sparsity Prior

Chuqiao Ren[a]

[a]*Columbia University*

Abstract

This report will present a graphical model of CANDECOMP/PARAFAC decomposition of a tensor with Normal-Inverse-Wishart prior, Normal-Gammar prior and Spike-and-Slab sparsity prior. The posterior probability will be evaluated by Gibbs Sampler, and the detailed update rules for each prior will be given.

Keywords: Tensor,Tensor Decomposition, Bayesian, Probability, Spike and Slab, Gaussian distribution, Wishart distribution, Gibbs Sampler.

1. Model Overview

Factor analysis has been extensively used in dimension reduction. The dimension of the gene data we are currently analyzing is relatively high; hence, it is reasonable to implement a latent factor model to extract the pattern from this data. Because of the nature of the data, we can construct a third-order tensor with column fiber, row fiber and tube fiber as individual, gene and tissue, respectively. Let $\mathcal{R} \in \mathbb{R}^{M\times N\times P}$ be a third order tensor, whose three dimensions correspond to individual, gene and tissue slices with sizes M, N and P. The ultimate goal is to decompose this third-order tensor $\mathcal{R}$ to three factor matrices, $\mathbf{V} \in \mathbb{R}^{N\times K}$, $\mathbf{U} \in \mathbb{R}^{M\times K}$ and $\mathbf{T} \in \mathbb{R}^{P\times K}$, where K is the dimension of the latent factors. We write the basic factorization as

$$\mathcal{R} = \sum_{k=1}^{K} U_{k:} \circ V_{k:} \circ T_{k:}, \tag{1}$$

where $\circ$ denotes the vector outer product. We can also express **equation (1)** entry-wise

$$R^p_{mn} \approx < V_n, U_m, T_p > \equiv \sum_{k=1}^{K} U_{mk} V_{nk} T_{pk}, \tag{2}$$

where we assume that the overall distribution is $\mathcal{N}(< V_n, U_m, T_p >, \alpha^{-1})$. This is a instance of CANDECOMP/PARAFAC decomposition [1]. Furthermore, we assume that $\mathbf{U}$ and $\mathbf{T}$ follow

[☆]A final research report to Dr. Istik

Email address: cr2826@columbia.edu (Chuqiao Ren)

the Normal-Inverse-Wishart distribution and $\mathbf{V}$ follows the Normal-Gamma distribution. In order to enforce the sparsity over the gene data, we will add the Spike and Slab prior on $\mathbf{v_n} = [v_{n1}, \ldots, v_{nK}]$, with a K-dimensional binary indicator vector $\mathbf{z}_n$ and its conjugate prior π_k. **Figure (1)** shows the graphical representation of our latent factor model.

According to the overall Graphical Model in **Figure (1)**, we can write the joint posterior distribution as

$$p(\mathbf{U}, \mathbf{V}, \mathbf{T}, \mathbf{z}, \boldsymbol{\pi}, \Theta_U, \Theta_V, \Theta_T \mid \boldsymbol{\mathcal{R}}) \propto$$
$$p(\mathbf{U} \mid \boldsymbol{\mathcal{R}}, \mathbf{V}, \mathbf{T}, \Theta_U) p(\mathbf{V} \mid \boldsymbol{\mathcal{R}}, \mathbf{U}, \mathbf{T}, \mathbf{z}, \Theta_V) p(\mathbf{T} \mid \boldsymbol{\mathcal{R}}, \mathbf{U}, \mathbf{V}, \Theta_T)$$
$$p(\Theta_U) p(\Theta_V) p(\Theta_T) p(\mathbf{z} \mid \boldsymbol{\mathcal{R}}, \boldsymbol{\pi}, \mathbf{U}, \mathbf{V}, \mathbf{T}) p(\boldsymbol{\pi}) p(\alpha)$$

where $\Theta_U \equiv \{\mu_U, \Lambda_U\}$, $\Theta_T \equiv \{\mu_T, \Lambda_T\}$ and $\Theta_V \equiv \{\mu_N, \lambda_N\}$. In more details, we have

- $p(\alpha) = Gamma(\alpha \mid \alpha, \beta) \propto \alpha^{\alpha-1} e^{-\alpha\beta}$
- $p(\mathbf{U} \mid \boldsymbol{\mathcal{R}}, \mathbf{V}, \mathbf{T}\Theta_U) = \prod_m^M p(U_m \mid \boldsymbol{\mathcal{R}}, \mathbf{V}, \mathbf{T}, \alpha, \Theta_U)$ and
  $p(\mathbf{T} \mid \boldsymbol{\mathcal{R}}, \mathbf{U}, \mathbf{V}, \Theta_T) = \prod_p^P p(T_p \mid \boldsymbol{\mathcal{R}}, \mathbf{U}, \mathbf{V}, \alpha, \Theta_T)$ as Normal Distribution[1].
- $p(\mathbf{V} \mid \boldsymbol{\mathcal{R}}, \mathbf{U}, \mathbf{T}, \mathbf{z}, \Theta_V)$ as Normal Distribution with Spike-and-Slab prior.
- $p(\Theta_U)$, $p(\Theta_T)$ as Normal-Wishart distribution (NW) and $p(\Theta_V)$ as Normal-Gamma distribution (NG).
- $p(\mathbf{z} \mid \boldsymbol{\mathcal{R}}, \boldsymbol{\pi}, \mathbf{U}, \mathbf{V}, \mathbf{T})$ and $p(\boldsymbol{\pi})$ as Bernoulli-Beta conjugate prior

2. Sampling

Because we constructed the graphical model (shown in **Figure (1)**) using distributions from the exponential family for most of the part, and the parent-child relationships preserve conjugacy in our model, we are able to use Gibbs sampling to sample most of the probability terms in the posterior distribution [2]. The following sections will briefly introduce how we update each variable in Gibbs sampling for all the priors we used in this model.

2.1. Spike and Slab Prior

Spike-and-Slab is a class of sparse priors that can be used based on a discrete mixture of point mass at zero (spike) and any other distribution (slab). It is first proposed by Mitchell and Beauchamp (1988) and Ishwaran and Rao (2005) [3]. Suppose the latent variable v_{nk}

[1]Note: Product of normal distributions is also a normal distribution

from $\mathbf{v}_n = \{v_{n1}, \ldots, v_{nK}\}$ is encoded by considering independent prior distributions given

$$p(\mathbf{v_n}|\mathbf{z_n}) = \prod_k p(v_{nk}|z_{nk})$$
$$p(v_{nk} \mid z_{nk}) = (1 - z_{nk})\delta_0(v_{nk}) + z_{nk}\pi(v_{nk})$$

where $\mathbf{z}_n = \{v_{n1}, \ldots, z_{nK}\}$ is a binary vector that controls the independent prior distribution, δ_0 is the delta function at zero (corresponding to the spike) and $\pi(v_{nk})$ is a fixed Gaussian distribution. Furthermore, we use a hierarchical specification with Bernoulli-Beta conjugate priors [4]:

$$p(\mathbf{z}_n \mid \boldsymbol{\pi}) = \prod_k \mathcal{B}(z_{nk} \mid \pi_k) = \prod_k \pi_k^{z_{nk}}(1 - \pi_k)^{(1-z_{nk})}$$
$$p(\pi_k \mid \alpha, \gamma) \sim \beta(\pi_k \mid \alpha, \gamma) = \frac{1}{\mathcal{B}(\alpha, \gamma)}\pi_k^{\alpha-1}(1 - \pi_k)^{\gamma-1}$$

where

$$\mathcal{B}(\alpha, \gamma) = \frac{\Gamma(\alpha + \gamma)}{(\Gamma(\alpha)\Gamma(\gamma))}.$$

In summary, we can represent the Spike-and-Slab prior as

$$p(\mathbf{v}_n \mid \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_k \mathcal{N}(v_{nk} \mid z_{nk}\mu_k, z_{nk}\sigma_k^2)$$

where μ_k is the mean of Gaussian and $\boldsymbol{\Sigma}$ (diagonal covariance) has elements σ_k^2 for slab component.

2.2. Bernoulli-Beta Conjugate Prior

As we mentioned in **Section 2.1**, we used conjugate Beta-Bernoulli priors to model $\mathbf{z}_n$. As a review, consider the following definitions of $\mathbf{z}_n$ and its prior π_k for each z_{nk}:

$$p(\mathbf{z}_n \mid \boldsymbol{\pi}) = \prod_k \mathcal{B}(z_{nk} \mid \pi_k) = \prod_k \pi_k^{z_{nk}}(1 - \pi_k)^{(1-z_{nk})}$$
$$p(\pi_k \mid \alpha, \gamma) \sim \beta(\pi_k \mid \alpha, \gamma) = \frac{1}{\mathcal{B}(\alpha, \gamma)}\pi_k^{\alpha-1}(1 - \pi_k)^{\gamma-1}$$

The posterior distribution for π_k is a Beta distribution and can be described as:

$$\begin{aligned} p(\pi_k) &\propto p(\mathbf{z}_k \mid \pi_k)p(\pi_k) \\ &= \big(\prod_n \pi_k^{z_{nk}}(1 - \pi_k)^{(1-z_{nk})} \cdot p(\pi_k)\big) \\ &= \pi_k^{\sum_n(z_{nk})}(1 - \pi_k)^{N-\sum_n(z_{nk})} \cdot p(\pi_k) \\ &= \pi_k^{\alpha+\sum_n(z_{nk})-1}(1 - \pi_k)^{N-\sum_n(z_{nk})+\gamma-1} \end{aligned}$$

Hence

$$p(\pi_k) \sim \beta(\pi_k \mid \bar{\alpha}, \bar{\gamma}),$$

and for each step, we can update $\bar{\alpha}$ and $\bar{\gamma}$ by

$$\bar{\alpha} = \alpha + \sum_n z_{nk}$$
$$\bar{\gamma} = N - \sum_n z_{nk} + \gamma$$

2.3. Gamma Prior

In **Section 1**, we introduced the overall normal distribution as $\mathcal{N}(< V_n, U_m, T_p >, \alpha^{-1})$. Since the mean μ is known as $< V_n, U_m, T_p >$ but the precision α is unknown, we can use Gamma prior to model this normal distribution [5]. As a review, the prior probability can be represented as

$$p(\alpha) = Gamma(\alpha \mid \alpha, \beta) \propto \alpha^{\alpha-1} e^{-\alpha\beta}.$$

Therefore, the conditional probability $p(\alpha \mid \mathcal{R})$ follows the Gamma distribution and can be expressed as

$$p(\alpha \mid \mathcal{R}) = Gamma(\alpha \mid \alpha_n, \beta_n).$$

So we are able to update α_n and β_n as follows [5]:

$$\alpha_n = \alpha + (N + M + K)/2$$
$$\beta_n = \beta + \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{p=1}^{P} (R_{mn}^p - \mu)^2$$

2.4. Normal-Gamma Prior

In **Section 1**, we specified that $p(\Theta_V)$ is a Normal-Gamma prior. From the graphical model (**Figure (1)**) we know that for each v_{nk} instance, we have normal prior with unknown μ_k and λ_k. The univariate conjugate pair of normal distribution with unknown μ and λ is Gamma distribution[5]. As a review, in our model, we have the following definition:

- $\mathbf{v}_n$ is a 1-dimensional vector $(v_n 1, \ldots, v_n K)$
- $\mu \mid \sigma \sim \mathcal{N}(\mu_0, \sigma/\kappa_0)$
  $\lambda \sim Gamma(\alpha_0, \beta_0)$

The conditional joint probability of μ, λ given $\mathbf{v}_n$ is

$$p(\mu, \lambda \mid \mathbf{v}_n) = NG(\mu, \lambda \mid \mu_k, \kappa_k, \alpha_k, \beta_k),$$

where $NG(\cdot)$ denotes the Normal-Gamma distribution which can be defined as

$$\begin{aligned} NG(\mu, \lambda \mid \mu_0, \kappa_0, \alpha_0, \beta_0) &:= \mathcal{N}(\mu \mid \mu_0, (\kappa_0 \lambda)^{-1}) Gamma(\lambda \mid \alpha_0, \beta_0) \\ &= \frac{1}{Z} \lambda^{\alpha_0 - \frac{1}{2}} \exp\left(-\frac{\lambda}{2}(\kappa_0(\mu - \mu_0)^2 + 2\beta_0)\right) \end{aligned}$$

where

$$Z = \frac{\Gamma(\alpha_0)}{\beta_0^{\alpha_0}} (\frac{2\pi}{\kappa_0})^{\frac{1}{2}}.$$

Therefore, we can update $\mu_k, \kappa_k, \alpha_k, \beta_k$ as following [5]:

$$\begin{aligned} \mu_k &= \frac{\kappa_0 \mu_0 + K \bar{v_n}}{\kappa_0 + K} \\ \kappa_k &= \kappa_0 + K \\ \alpha_k &= \alpha_0 + \frac{K}{2} \\ \beta_k &= \beta_0 + \frac{1}{2} \sum_k^K (v_{nk} - \bar{v_n})^2 + \frac{\kappa_0 k (\bar{v_n} - \mu_0)^2}{2(\kappa_0 + K)} \\ \bar{v_n} &= \frac{1}{K} \sum_k^K v_{nk} \end{aligned}$$

2.5. Normal-Wishart Prior

In **Section 1**, we specified that $p(\Theta_U)$ and $p(\Theta_T)$ is a Normal-Wishart prior. Due to the symmetry in our model for U and T, here we only discuss $p(\Theta_U)$, and $p(\Theta_T)$ will be the same. Normal-Wishart prior is the multivariate analogous of the Normal-Gamma prior [5]. From the graphical model (**Figure (1)**) we know that for each k-dimensional vector $\mathbf{U}_m$, we have normal prior with unknown $\mu_{\mathbf{U}}$ and $\mathbf{\Lambda_U}$. The multivariate conjugate pair of normal distribution with unknown μ and $\mathbf{\Lambda}$ is Wishart distribution[5]. As a review, in our model, we have the following definition:

- $\mathbf{U}$ is a k-dimensional vector $(U_1, \ldots, U_m)$
- $\boldsymbol{\mu} \mid \Sigma \sim \mathcal{N}(\mu_0, \Sigma/\kappa_0)$
  $\Lambda \sim W_{i_{\nu_0}}(T_n)$

The conditional joint probability of $\boldsymbol{\mu}, \Lambda$ given $\mathbf{U}$ is

$$p(\boldsymbol{\mu}, \Lambda \mid \mathbf{U}) = \mathcal{N}(\boldsymbol{\mu} \mid \mu_m, (\kappa_m \Lambda)^{-1}) W_{i_{\nu_m}}(\Lambda \mid T_m).$$

The prior $p(\boldsymbol{\mu}, \Lambda)$ can be defined as

$$p(\boldsymbol{\mu}, \Lambda) = NW_i(\boldsymbol{\mu}, \Lambda \mid \mu_0, \kappa, \nu, T) = \mathcal{N}(\boldsymbol{\mu} \mid \mu_0, (\kappa_m \Lambda)^{-1}) W_{i_\nu}(\Lambda \mid T)$$
$$\frac{1}{Z}|\Lambda|^{\frac{1}{2}} \exp\Big(-\frac{\kappa}{2}(\boldsymbol{\mu} - \mu_0)^T \Lambda (\boldsymbol{\mu} - \mu_0)\Big) |\Lambda|^{(\kappa - d - 1)/2} \exp\big(-\frac{1}{2}\mathbf{tr}(T^{-1}\Lambda)\big)$$

where

$$Z = (\frac{\kappa}{2\pi})^{d/2} |T|^{\kappa/2} 2^{d\kappa/2} \Gamma_d(\kappa/2).$$

Therefore, we can update $\mu_m, T_m, \kappa_m, \nu_m$ as following [5]:

$$\begin{aligned}
\mu_m &= \frac{\kappa\mu_0 + m\bar{U}}{\kappa + m} \\
T_m &= T + S + \frac{\kappa m}{\kappa + m}(\mu_0 - \bar{U})(\mu_0 - \bar{U})^T \\
\nu_m &= \nu + m \\
\kappa_m &= \kappa + m \\
S &= \sum_m^M (U_m - \bar{U})(U_m - \bar{U})^T
\end{aligned}$$

2.6. Decomposed Latent Matrices

Now we are ready to use the conclusion we derived from **Section (2.1), Section (2.2), Section (2.3), Section (2.4)** and **Section (2.5)** to derive the sampling of latent matrices. Due to the symmetry in our graphical model of $\mathbf{U}$ and $\mathbf{T}$, we omit the discussion of $\mathbf{T}$ for brevity here.

2.6.1. Sampling Latent Matrix $\mathbf{U}$

From the graphical model we can easily derive the conditional probability of $\mathbf{U}$ as

$$P(\mathbf{U} \mid \mathbf{R}, \mathbf{V}, \mathbf{T}, \Theta_U) = \prod_{m=1}^{M} p(U_m \mid \mathbf{R}, \mathbf{V}, \mathbf{T}, \Theta_U).$$

We assume that U_m has Gaussian distribution; hence, we have

$$p(U_m \mid \mathbf{R}, \mathbf{V}, \mathbf{T}, \Theta_U) \sim \mathcal{N}(U_m \mid \mu_m, (\Lambda)^{-1}).$$

We are able to update μ_m and Λ_m as follows [6]:

$$\mu_m := \mathbf{\Sigma}_m(\Sigma_U \mu_U + \alpha \sum_{p=1}^{P} \sum_{n=1}^{N} I_{mn}^p R_{mn}^p V_n T_p)$$

$$\Lambda_m = \Lambda_U + \alpha \sum_{p=1}^{P} \sum_{n=1}^{N} I_{mn}^p R_{mn}^p V_n T_p (V_n T_p)^T)$$

where I_{mn}^p is one if R_{mn}^p is available and zero otherwise.

2.6.2. Sampling Latent Matrix $\mathbf{V}$

In **Section (2.1)** we discussed the Spike-and-Slab prior. As a review, we have the following definitions

$$p(\mathbf{v_n}|\mathbf{z_n}) = \prod_k p(v_{nk}|z_{nk}) \tag{3}$$

$$p(v_{nk} \mid z_{nk}) = (1 - z_{nk})\delta_0(v_{nk}) + z_{nk}\pi(v_{nk}) \tag{4}$$

From **Equation (4)** above, we can easily conclude that if $z_{nk} = 1$, $\mathbf{v_k}$ will have a Gaussian distribution. Furthermore, since $v_{n_i k} \perp\!\!\!\perp v_{n_j k}$ if $i \neq j$, we are able to pick all the entries of $\mathbf{v}$ as $z_{nk} = 1$, and these entries will have a Gaussian distribution. More formally, let us define $k' = \{k \mid z_{nk} = 1\}$ with cardinality $|k'| = L$. Then $\boldsymbol{\mu}_{k'} = [\mu_1, \ldots, \mu_L]$ is a vector of μ_k where $z_{nk} = 1$ and $\mathbf{\Lambda}_{k'}$ is a covariance matrix with only diagonal entries consisted of $\lambda_{k'}$. Furthermore, let U'_m, T'_p be the subset of U_m, T_p corresponding to $z_{nk} = 1$ in $\mathbf{v}_n$, respectively. The conditional probability of $\mathbf{V}$ can be defined as

$$P(\mathbf{V} \mid \mathbf{R}, \mathbf{U}, \mathbf{T}, \Theta_U) = \prod_{k'=1}^{L} P(V_{k'} \mid \mathbf{R}, \mathbf{U}, \mathbf{T}, \Theta_U).$$

Each $V_{k'}$ has Gaussian distribution; hence, we have

$$P(V_{k'} \mid \mathbf{R}, \mathbf{U}, \mathbf{T}, \Theta_U) \sim \mathcal{N}(V_{k'} \mid \mu_{k'}, (\Lambda)^{-1}).$$

Now we are able to update $\mu_{k'}$ and $\Lambda_{k'}$ as follows:

$$\mu_{k'} := \mathbf{\Sigma}_{k'}(\Sigma_V \mu_V + \alpha \sum_{p=1}^{P} \sum_{m=1}^{M} I_{mk'}^p R_{mk'}^p U'_m T'_p)$$

$$\Lambda_{k'} = \Lambda_V + \alpha \sum_{p=1}^{P} \sum_{m=1}^{M} I_{mk'}^p R_{mk'}^p (U'_m T'_p)(U'_m T'_p)^T$$

3. Discussion and Future Work

The only term left in the overall joint posterior distribution is

$$p(z_{nk} \mid \mathbf{X}, \boldsymbol{\pi}, \mathbf{V}_{\neg nk}, \mathbf{U}, \mathbf{T}).$$

When $z_{nk} = 0$, we have

$$\begin{aligned} & p(z_{nk} = 0 \mid \mathbf{X}, \boldsymbol{\pi}, \mathbf{V}_{\neg nk}) \\ & \propto \int p(z_{nk} = 0, v_{nk} = 0, \mathbf{X} \mid \boldsymbol{\pi}, \mathbf{V}_{\neg nk}, \mathbf{U}, \mathbf{T}) dv_{nk} \\ & = (1 - \pi_k) p(\boldsymbol{\mathcal{R}} \mid \mathbf{V}_{\neg nk}, \mathbf{U}, \mathbf{T}, v_{nk} = 0) \end{aligned}$$

and this is easy to compute. However, if $z_{nk} = 1$, we have

$$\begin{aligned} & p(z_{nk} = 1 \mid \mathbf{X}, \boldsymbol{\pi}, \mathbf{V}_{\neg nk}) \\ & \propto \int p(z_{nk} = 1, v_{nk}, \mathbf{X} \mid \boldsymbol{\pi}, \mathbf{V}_{\neg nk}, \mathbf{U}, \mathbf{T}) dv_{nk} \\ & = \pi_k \int p(\boldsymbol{\mathcal{R}} \mid \mathbf{V}_{\neg nk}, \mathbf{U}, \mathbf{T}, v_{nk}) \mathcal{N}(v_{nk} \mid \mu_k, \sigma_k^2) dv_{nk} \end{aligned}$$

and the integral is not tractable[4]. Hence, for the case of intractable integral, we need to use Laplace approximation. Due to the limitation of time and the priority of work, although we have the idea of how to evaluate the Laplace approximation, we didn't proceed to evaluate it.

The code of this work is available on github. It has a Python version, which is mainly developed by Shuo Yang, and a cpp version, which is mainly developed by myself. Due to the limitation of time, we are only able to implement the Gaussian distribution (NW and NG) without Spike-and-Slab prior on $\mathbf{V}$. In the future, we hope to complete the implementation of the sparsity prior.

4. Code

One may find a Python implementation (by Shuo Yang) and a cpp implementation of the Gaussian distribution (NW and NG) without Spike-and-Slab prior on github (`https://github.com/ComputationalBiology-CS-CU/prob_tensor_decomp.git`).

5. Acknowledgment

This work cannot be done without the generous help from Shuo Yang. Also, I would like to thank Dr. Itsik Pe'er for providing me this precious opportunity to work in his lab.

References

[1] T. G. Kolda, B. W. Bader, S. N. Laboratories, Tensor Decompositions and 51 (3) (2008) 455–500.

[2] C. M. Bishop, Pattern Recognition and Machine Learning, Information Science and Statistics, Springer, 2006.

[3] H. Ishwaran, J. S. Rao, Spike and slab variable selection: Frequentist and bayesian strategies, Annals of Statistics 33 (2) (2005) 730–773. arXiv:0505633, doi:10.1214/009053604000001147.

[4] S. Mohamed, Generalised Bayesian matrix factorisation models (February) (2011) 1–140.

[5] K. P. Murphy, Conjugate bayesian analysis of the gaussian distribution, preprint.

[6] L. Xiong, X. Chen, T.-k. Huang, J. Schneider, J. G. Carbonell, Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization, Proceedings of the SIAM International Conference on Data Mining (2010) 211—-222.

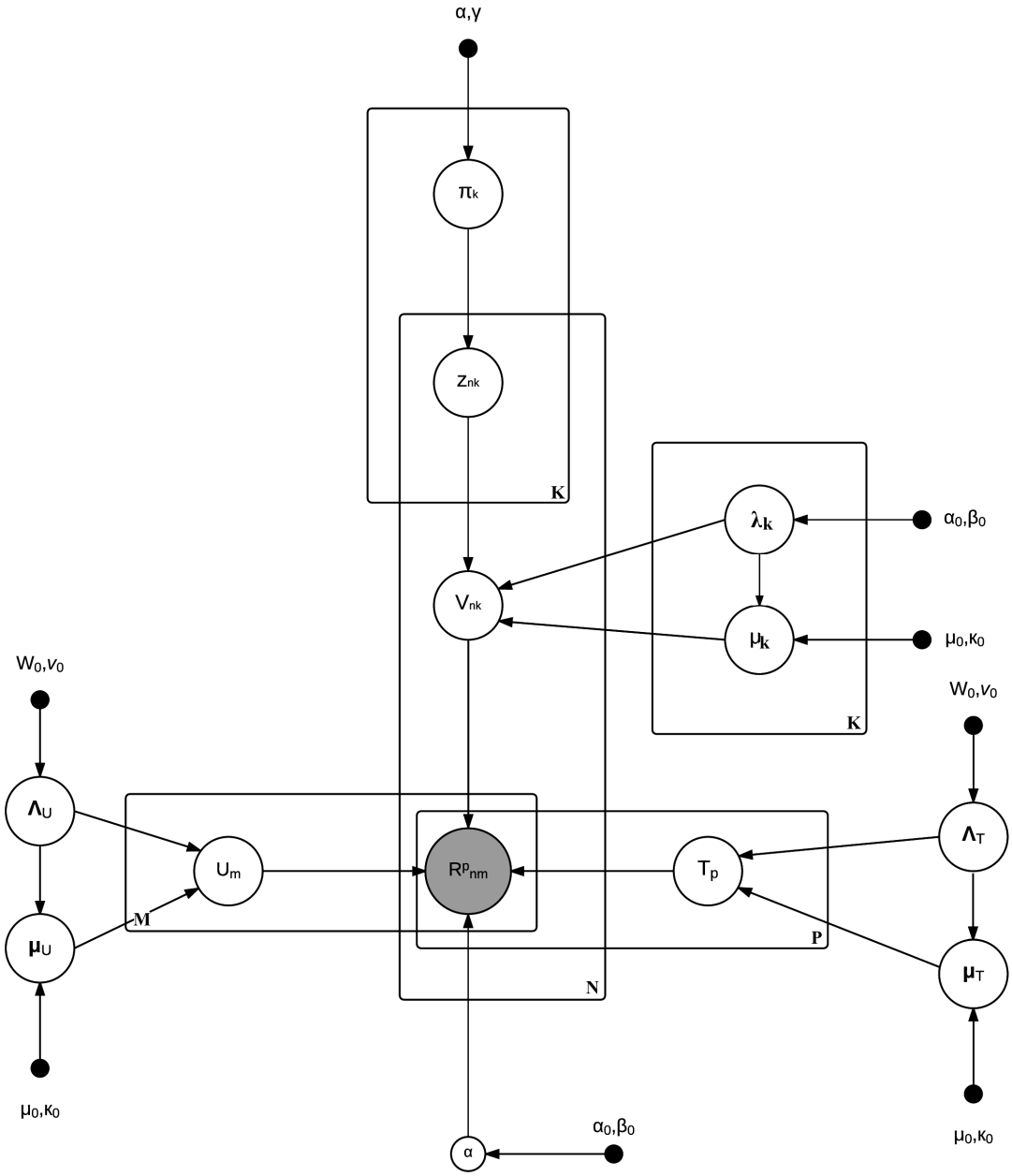

Figure 1: The overall graphical model of tensor decompsition.