Fast EM Inference for Bayesian Factor Analysis with Indian Buffet Process Prior

Inferenza Veloce con EM per l'Analisi Fattoriale Bayesiana mediante Indian Buffet Process

Veronika Ročková and Edward I. George

Abstract We consider a nonparametric Bayesian approach to parsimonious latent factor modeling of high-dimensional covariance matrices. Central to our approach is the Indian Buffet Process (IBP) prior, which allows the number of factors to be unbounded while remaining finite with probability one. The combinatorial complexity of the IBP makes posterior simulation very challenging. As an alternative, we present a fast EM algorithm for MAP learning in Bayesian factor analysis with spike-and-slab priors, exploiting the stick-breaking representation of IBP.

Abstract *Discutiamo un approccio Bayesiano non parametrico alla modellazione parsimoniosa di fattori latenti per matrici di covarianza di grandi dimensioni. L'Indian Buffet Process (IBP), usato come prior, permette un numero illimitato di fattori pur conservando probabilita' unitaria. La complessita' combinatoriale dell'IBP rende la simulazione della posterior onerosa. Qui proponiamo un algoritmo EM veloce per addestramento MAP in analisi fattoriale Bayesiana con prior a spike-and-slab, usando la stick-breaking dell'IBP.*

Key words: EM, Indian Buffet Process, Spike and Slab

1 Bayesian Factor Analysis Revisited

We will be focusing on scenarios when n observations are collected on G related responses and factor analysis is deployed to find regularities in the data, that can be attributed to a set of hidden causes (factors). The data matrix of n standardized G dimensional responses will be denoted by $Y = [y_1, \ldots, y_n]'$. Assuming a fixed number K of latent factors, the generic factor model assumes

$$Y_i \mid \omega_i, B, \Sigma \overset{\text{ind}}{\sim} \mathrm{N}_G\left(B\omega_i, \Sigma\right), \quad 1 \leq i \leq n, \tag{1}$$

Veronika Ročková
University of Pennsylvania, Philadelphia (PA), e-mail: vrockova@wharton.upenn.edu

where ω_i is a $(K \times 1)$ vector of unobserved latent factors arising from $\mathrm{N}_K(0, \sigma_\omega^2 \mathrm{I}_K)$, $\Sigma = \mathsf{diag}\{\sigma_j^2\}_{j=1}^G$ is a diagonal matrix of unknown positive scalars and B is the $(G \times K)$ matrix of factor loadings. Margining out the latent data induces dependence among the responses by yielding $f(y_i \mid B, \Sigma) = \mathrm{N}_G(0, BB' + \Sigma)$. Model (1) is invariant under factor rotation and therefore non-identifiable. This indeterminacy will be less of an issue if the primary goal of the analysis is prediction or estimation of the covariance matrix. Finding a unique factor rotation, that is also interpretable, is commonly achieved by introducing structural identifiability constraints coupled with some form of shrinkage on the factor loadings. However, these identifiability constraints can be concerning, because they render the variables no longer exchangeable. In Section 2, we will instead consider a nonparametric prior which guarantees the exchangeability, while only inducing a soft identifiability constraint against permutational invariance.

1.1 Spike and Slab Priors

The Bayesian approach to sparse factor modeling uses priors on the elements in $B = \{\beta_{jk}\}_{j,k=1}^{G,K}$ that induce either zeroes or values close to zero with high-probability. We take the latter approach, exploiting the continuous relaxation of a point-mass mixture prior (as in Rockova and George (2014)). Denoting $B = [\beta_1, \ldots, \beta_G]'$, each β_j is assigned a conjugate Gaussian mixture prior

$$\pi(\beta_j) \sim \mathrm{N}_K(0, \sigma_j^2 D_j)$$

where $D_j = \mathsf{diag}\{(1-\gamma_{jk})v_0 + \gamma_{jk} v_1\}_{k=1}^K$ and $v_0 << v_1$ are set to be small and large to distinguish the ignorable β_{jk} values from those that describe a functional relationship between the jth response and the kth factor. Here $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jK})'$ denotes a vector of binary allocation indicators that characterize which features are possessed by jth response. For σ_j^2 we assume independent Inverse Gamma priors $\mathrm{IG}(\eta/2, \eta\lambda/2)$.

1.2 Finite Latent Allocation Model

To define the joint prior distribution on the finite allocation matrix $\Gamma = (\gamma_1, \ldots, \gamma_G)'$, we use a partially exchangeable beta-bernoulli prior with intermediate occurrence probabilities $\theta = (\theta_1, \ldots, \theta_K)'$, one for each column in B. Written hierarchically, the prior is

$$\begin{aligned} \pi(\gamma_{jk} | \theta_k) &\overset{\text{ind}}{\sim} \mathrm{Bernoulli}(\theta_k), \\ \pi(\theta_k) &\overset{\text{ind}}{\sim} \mathscr{B}\left(\frac{\alpha}{K}, 1\right). \end{aligned} \tag{2}$$

It may be inappropriate to constrain the number of factors a priori. To this end, Grifiths and Ghahramani (2007) proposed a stochastic process on feature allocations. Named the Indian Buffet Process (IBP), employing this process as a prior enables inference about the effective factor cardinality.

2 Going Infinite with Indian Buffet Process

The metaphor of the Indian Buffet Process describes customers tasting dishes in an infinite buffet. Let γ_{jk} be a binary indicator characterizing whether the jth customer tastes dish k. Event $\gamma_{jk} = 1$ occurs with probability m_k/i, where m_k is the number of customers that have previously taken a serving from dish k. Having sampled from dishes that were taken by the preceding customers, customer j then proceeds to take a number of new dishes by drawing from a Poisson(α/j) distribution. Modulo a reordering of the dishes (features), the Indian buffet process generates an exchangeable distribution over infinite binary matrices. The effective dimension (the number of sampled dishes) however remains finite with probability one.

The Indian Buffet Process (IBP) is formally defined as a limiting distribution over equivalence classes of finite allocation matrices, obtained by integrating out θ_k in (2) and by taking the limit $K \to \infty$. Proceeding marginally over θ_k leads directly to a Gibbs sampler. Instead, we will use the stick-breaking representation of Teh et al. (2008) and proceed conditionally on a particular ordering of θ_k's.

2.1 Stick Breaking Construction

Stick-breaking refers to the process of recursively breaking off random fractions of the unit interval. Let $\nu_1, \nu_2, \dots$ be an infinite sequence of random variables, where each ν_j represents the jth breaking fraction. After k steps of recursive breaking, the length of the stick that we have left is $\prod_{j<k} \nu_j$. Unlike the stick-breaking construction for the Dirichlet process prior, here we recurse on the length of the remaining piece rather than the discarded piece. Teh et al. (2008) showed an interesting connection between this reversed stick breaking representation and the Indian Buffet Process.

Let $\theta_{(1)} > \theta_{(2)} > ... > \theta_{(K)}$ be a decreasing ordering of $\theta = (\theta_1, \dots, \theta_K)'$, where each $\theta_k \sim \mathscr{B}(\alpha, 1)$. Teh et al. (2008) showed that in the limit as $K \to \infty$, the $\theta_{(k)}$'s obey the following stick-breaking law,

$$\theta_{(k)} = \prod_{l=1}^{k} \nu_l, \quad \text{where} \quad \nu_l \overset{\text{iid}}{\sim} \mathscr{B}(\alpha, 1). \tag{3}$$

The implicit ordering induces a soft identifiability constraint against permutational invariance. Assuming that $\theta_{(k)} = 0$ for $k > K^\star$, (3) provides a truncated approximation to the IBP, which is at the heart of our EM approach. We will also assume $\beta_{jk} = 0$ for $k > K^\star$.

3 EM for Factor Analysis with the IBP

The goal of our unsupervised learning algorithm is to find the allocation matrix most likely (a posteriori) to have generated the data. We proceed indirectly in terms of the truncated loading matrix B (of dimension $(G \times K^\star)$) and treat the allocation matrix Γ (of dimension $(G \times K^\star)$) as additional latent data. If the hidden variables were known, the loading matrix could be easily estimated. Similarly, both Γ and Ω cannot be inferred unless the loading matrix is known. This chicken-and-egg problem can be can be solved by iterating between two steps: (E-step) computing the expectation of the latent data given the current loadings and (M-step) finding the most likely loadings given the expected missing data. These two steps form a basis of our Expectation-Maximization (EM) algorithm.

The EM algorithm locates posterior modes of $\pi(B, \Sigma, \nu \,|\, Y)$ iteratively by maximizing the expected augmented log posterior:

$$Q(B, \nu, \Sigma) = \mathsf{E}_{\Gamma, \Omega | \cdot} \left[\log \pi \left(\underbrace{B, \Sigma, \nu}_{\text{unknown parameters}} \;,\; \overbrace{\Gamma, \Omega}^{\text{missing data}} \;\Big|\; \underbrace{Y}_{\text{observed data}} \right) \right], \tag{4}$$

where $\mathsf{E}_{\Gamma, \Omega | \cdot}(\cdot)$ denotes the conditional expectation given the observed data and current parameter estimates at the mth iteration, and $\nu = (\nu_1, \ldots, \nu_{K^\star})'$ is the vector of breaking fractions. The allocation matrix Γ is then obtained by probabilistic thresholding of the loading matrix (Section 4). The objective function can as well be parametrized in terms of θ's using the stick-breaking relationship $\nu_k = \theta_{(k)} / \theta_{(k-1)}$. We note in passing that such parametrization is beneficial for a feasible M-step.

3.1 The E-step

The E-step requires computing the expected log complete posterior (4). This simplifies to computing the conditional mean and covariance of the latent data in Ω and the conditional allocation probabilities in Γ according to the closed form expressions:

$$\langle \omega_i \rangle = M^{(m)} B^{(m)\prime} \Sigma^{(m)-1} \left(y_i - A^{(m)} x_i \right) \quad \text{with} \quad M^{(m)} = \left(B^{(m)\prime} \Sigma^{(m)-1} B^{(m)} + \mathrm{I}_{K^\star} \right)^{-1},$$

$$\langle \gamma_{jk} \rangle = \frac{\mathrm{N}(\beta_{jk}^{(m)}; 0, \sigma_j^{(m)2} v_1) \theta_k^{(m)}}{\mathrm{N}(\beta_{jk}^{(m)}; 0, \sigma_j^{(m)2} v_1) \theta_k^{(m)} + \mathrm{N}(\beta_{jk}^{(m)}; 0, \sigma_j^{(m)2} v_0)(1 - \theta_k^{(m)})}. \tag{5}$$

3.2 The M-step

Once the latent data have been updated, the M-step consists of maximizing (4) with respect to the unknown parameters. Denote by y^j the jth column in Y, $\langle \Omega \rangle = [\langle \omega_1 \rangle, \ldots, \langle \omega_n \rangle]'$ and $D_j^\star = M^{(m)} + \mathrm{diag}\{\langle \gamma_{jk} \rangle / v_1 + (1 - \langle \gamma_{jk} \rangle) / v_0\}_{k=1}^{K^\star}$. The updates at the mth iteration are

$$\beta_j^{(m)} = \left(\langle\Omega\rangle'\langle\Omega\rangle + D_j^\star\right)^{-1}\langle\Omega\rangle' y^j,$$
$$\sigma_j^{(m)2} = \frac{||y^j - \langle\Omega\rangle\beta_j^{(m)}||^2 + ||D_j^{\star 1/2}\beta_j^{(m)}||^2 + \nu\lambda}{n + K^\star + \nu}.$$

Treating the ν's as parameters to be estimated, the $\nu^{(m)}$ update would require maximizing

$$Q(\nu) = \sum_{i=1}^{G}\sum_{k=1}^{K^\star}\left\{\langle\gamma_{ik}\rangle\sum_{j=1}^{k}\log\nu_j + (1-\langle\gamma_{ik}\rangle)\log\left(1-\prod_{j=1}^{k}\nu_j\right)\right\} + (\alpha-1)\sum_{k=1}^{K^\star}\log(\nu_k), \tag{6}$$

which is a rather complex nonlinear function in ν. Instead, we perform the optimization in terms of θ's. Plugging $\nu_k = \theta_{(k)}/\theta_{(k-1)}$ into (6), the new objective becomes

$$Q^\star(\theta) = \sum_{i=1}^{G}\sum_{k=1}^{K^\star}\left[\langle\gamma_{ik}\rangle\log\theta_k + (1-\langle\gamma_{ik}\rangle)\log(1-\theta_k)\right] + (\alpha-1)\log\theta_K, \tag{7}$$

whose maximum $\theta^{(m)}$ can be found by solving a linear program subject to a series of constraints

$$\theta_k - \theta_{k-1} \le 0, k = 2,\ldots,K^\star,$$
$$0 \le \theta_k \le 1, k = 1,\ldots,K^\star.$$

Had we assumed the finite beta-bernoulli prior from Section 1.2, the updates of the occurrence probabilities would simply become $\theta_k^{(m)} = \frac{\sum_{j=1}^{G}\langle\gamma_{jk}\rangle + a - 1}{a+b+G-2}$. Although the ordering constraint provides more stability in estimation, it does not guarantee rotational invariance of the factor model. Realizations of IBP which approximately correspond to identifiable loading matrices can be obtained by considering small α or through intermediate sparsification step described in the next section

4 EM Factor Structure Detection using Sparsification

We illustrate how well our EM algorithm fares in recovering the effective factor cardinality and the underlying sparse pattern of covariance with a simple example. We consider a simulated experiment with $n = 100$ observations from (1) with $G = 200$ responses and $K = 5$ generative factors using a structured loading matrix (Figure 1 left). Responses were generated assuming $\sigma_j^2 = 1$ and nonzero $\beta_{jk} = 4$ and then centered. We consider the order of the truncated stick breaking approximation $K^\star = 20, \alpha = 1, v_0 = 0.005$ and $v_1 = 1\,000$.

In order to obtain interpretable orientations of the loading matrix, we perform an additional sparsification step at every iteration by thresholding out loadings β_{jk}, for which $\langle\gamma_{jk}\rangle < 0.5$. Such a step stabilizes estimation of the truly relevant associations, while effectively removing intermediate noise coefficients.

Initializing the algorithm at $\beta_{jk}^{(0)} = 1$ and performing the sparsification at every iteration, we obtain the estimated pattern of zeroes in the loading matrix depicted in

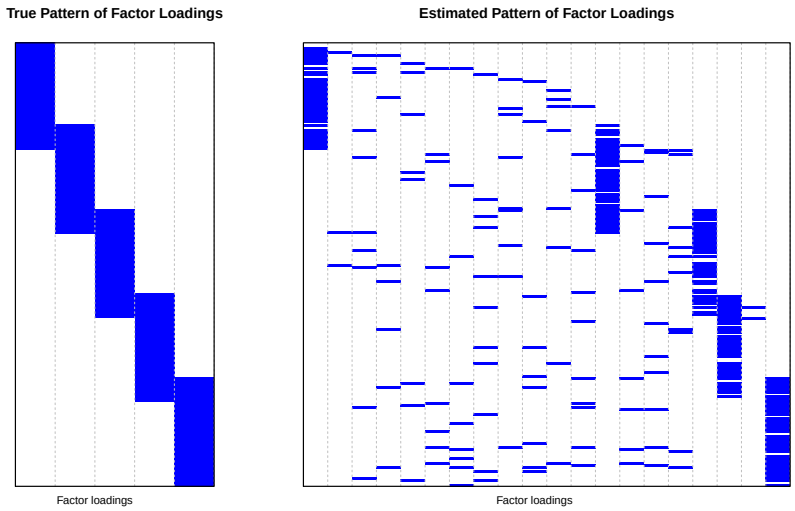

Fig. 1 True and estimated pattern of zeroes in the loading matrix presented in the left-ordered-form. Nonzero values marked in blue.

Figure 1 (right). The estimated allocation of zeroes strongly suggests $\widehat{K} = 5$ as the effective factor cardinality. The distribution pattern of the nonzero entries indicates a grouping structure, which approximately recovers the $K = 5$ clusters that actually generated the data.

5 Discussion

We have presented a fast EM approach to Bayesian factor analysis using spike-and-slab priors coupled with an Indian Buffet Process prior on the allocation matrix. A sparsification variant of our EM algorithm enhances the efficient recovery of the actual factor cardinality and exposes the hidden grouping pattern among the responses.

References

[1] Rockova, V., George, E.: EMVS: The Bayesian Approach to Bayesian Variable Selection. Journal of American Statistical Association (to appear) (2014)
[2] Griffiths, T., Ghahramani, Z.: Infinite Latent Feature Models and the Indian Buffet Process. In Advances in Neural Information Processing Systems **18** (2004), MIT Press, Cambridge, MA
[3] Teh, Y., Görür, Ghahramani, Z.: In Proceedings of the International Conference on Artificial Intelligence and Statistics (2006)