Non-parametric Bayesian Hierarchical Factor Modeling and Regression

Piyush Rai[1] and Hal Daumé III
School of Computing, University of Utah
Salt Lake City - Utah 84112
{piyush,hal}@cs.utah.edu

Introduction

We address the problem of sparse Bayesian factor regression from high-dimensional gene-expression data where the number and inter-relationship of factors is not known apriori. We take a non-parametric Bayesian approach based on a variant of the *Indian Buffet Process* [1]. This leads to an interpretable model for gene-pathway relationships, a simple inference procedure, and allows us to consider more complex models for factor modeling than is allowed by model-selection based approaches [2]. Our motivation is that non-parametric approaches yield models that let us look at structurally rich problems in a coherent manner. In particular, the non-parametric approach allows us to directly consider variable selection and hierarchical factors in a unified model. Variable selection models the fact that most genes are not involved in any pathway for a given dataset, and leads to computational benefits. The non-parametric nature of our model allows us to impose a hierarchy over the discovered factors which helps in explaining the correlations within sets of factors having similar functions. Finally, we couple the regression task within the factor modeling framework itself, instead of considering them as separate tasks [3].

Bayesian Factor Regression Models

In gene-expression studies, we model the relationships between high-dimensional ($P \gg 1000$) gene-expression data and a small number ($k \approx 10$) of underlying *subpathways* (latent factors). The gene versus factor relationship is inherently sparse: each gene affects and is affected by a very small number of latent factors. We use Bayesian factor-regression: couple a factor analysis with regression modeling. For the factor analysis aspect, we have gene expression data having N samples with P variables (genes) each ($P \gg N$) in a matrix $\mathbf{X}$ of size $P \times N$. We model the data as: $\mathbf{X} = \mathbf{AF} + \mathbf{\Psi}$, where $\mathbf{A}$ is a *sparse* $P \times k$ factor-loading matrix (capturing the relationship between genes and the latent factors), $\mathbf{F}$ is a $k \times N$ latent factor matrix relating factors to samples, and $\mathbf{\Psi}$ is idiosyncratic noise. The predictive modeling for regression can be done in terms of the latent factors: $\mathbf{y} = \theta'\mathbf{F} + \varepsilon$, where θ is a $k \times 1$ vector of regression parameters and ε is noise. The latent factor regression yields considerable statistical and computational saving since the dimensionality (k) is significantly smaller than that of the original data (P). The predictive modeling can be coupled with the factor analysis framework leading to a unified sparse regression model for the predictors $\mathbf{X}$ and the responses variables $\mathbf{y}$. It is also straightforward to extend the model for multivariate response variables [2].

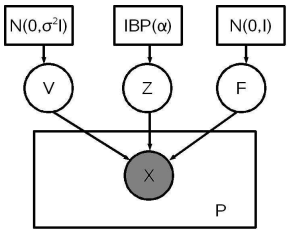
A common approach to the factor analysis problem is to place a *sparse* prior on $\mathbf{A}$ [3]. However, typically, the assumption is that the number of factors (k) is known in advance, which is often not the case. The existing approaches to model selection are based on reversible jump MCMC methods or evolutionary stochastic model search based methods which may take too long to converge [2].

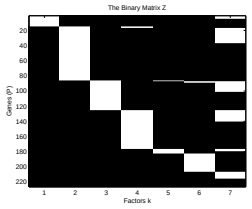
Non-parametric Bayesian Factor Regression

We take a different approach to sparse factor regression which is based on the *Indian Buffet Process* (IBP) [1]. Although IBP has been applied to non-parametric factor analysis in the past [1], the standard IBP formulation places IBP prior on the factor matrix ($\mathbf{F}$) associating *samples* (i.e., set of data points) with factors. However, this assumption is inappropriate in the gene-expression context: we are interested in associations between genes (i.e., variables) and factors. Thus, it is more appropriate to place the IBP prior on the factor loading matrix ($\mathbf{A}$) instead. In the IBP culinary analogy, the customers are *genes* which are selecting dishes (factors). Note, however, that since $\mathbf{A}$ and $\mathbf{F}$ are related with each other via the number of factors k, modeling $\mathbf{A}$ non-parametrically allows our model to also have an unbounded number of factors, essentially leading to an alternative non-parametric latent factor model.

[1]Category: Learning Algorithms/Graphical Models. Preference: Oral

$\mathbf{X} = (\mathbf{Z} \odot \mathbf{V})\mathbf{F} + \mathbf{\Psi}$
$\mathbf{Z} \sim IBP(\alpha); \alpha \sim Gamm(a, b)$
$\mathbf{V} \sim Nor(0, \sigma_v); \sigma_v \sim Gamm(e, f)$
$\mathbf{F} \sim Nor(0, \mathbf{I})$
$\mathbf{\Psi} \sim diag(\Psi_1, .., \Psi_P)$





In the above formulation, the factor loading matrix $\mathbf{A}$ is a convolution of a sparse binary matrix ($\mathbf{Z}$) and Gaussian vectors ($\mathbf{V}$), similar to the standard approaches taken in sparse modeling. $\mathbf{Z}$ is drawn from an IBP prior. IBP follows a sequential generative process and we use the standard Gibbs sampler for inference. However, for non-conjugate models, it may also be beneficial to use the stick-breaking construction [4] of IBP, which additionally addresses the problem of slow-mixing. The rightmost figure above shows the result of sampling for a run of 500 iterations for a gene-expression dataset having 25 samples of 226 genes each. The number of factors discovered is 7 which is close to the ground truth. Our approach gave a mean squared reconstruction error of about 0.32 on this data as compared to BFRM [3] for which it is about 0.36 and it shows that our approach indeed does quite well. As in [2], combining predictive modeling to the model is again straightforward. This is done by simply extending the model by prepending the response variables ($\mathbf{y_i}$) to the gene-expression vectors ($\mathbf{x_i}$). The MCMC analysis in this case can be extended by treating $\mathbf{y_i}$s as missing variables to be imputed.

Variable Selection and Hierarchical Factor Modeling

Typical gene-expression datasets are of the order of several thousand or tens of thousands of genes. However, in most datasets, many genes will not be associated with any pathway (factor). In the standard formulation, these are accounted for only by the idiosyncratic noise term, which is not an appropriate model. We propose a variable selection prior in form of a sparse P-dimensional vector $\mathbf{T}$ (each entry of $\mathbf{T}$ corresponds to a row in $\mathbf{Z}$), placing a beta prior ($Beta(1, a)$) over $\mathbf{T}$. Following the IBP culinary analogy, this corresponds to a customer entering and immediately deciding not to eat, before looking at any dishes. Typically, we set a to be a constant (10 or 100) or give it a prior and sample over a. The Gibbs sampler samples for a particular row in $\mathbf{Z}$ only if the corresponding entry in $\mathbf{T}$ is sampled as 1, otherwise the entire row of $\mathbf{Z}$ is set as zero. Due to the large size of these datasets, sampling the $\mathbf{Z}$ matrix (and the associated $\mathbf{V}$ matrix) can be quite expensive. Variable selection can yield considerable computation savings.

Another limitation of the standard factor analysis models comes from the fact that they assume a priori independent latent factors. This fails to capture the correlations that may exists among gene pathways. Such correlations are relevant, especially in gene-expression contexts, since factors with similar tasks tend to regulate related genes. In such contexts, the assumption that gene pathways (the factors) are independent is no longer true. Some pathways are involved in transcription, some in synthesis, some in signaling, etc. In light of this, we wish to model a hierarchical latent structure on factors. To do so, we employ the coalescent as a hierarchical prior [5]. The "leaves" of the coalescent tree are precisely the rows in our factor matrix. The coalescent is an attractive choice because it is also non-parametric with a well-defined predictive density, allowing for simple extensions to our Gibbs sampling procedure. Our initial results have been encouraging and we expect to report compelling results at the workshop.

References

[1] Z. Ghahramani, T.L. Griffiths, and P. Sollich. Bayesian nonparametric latent feature models. In *Bayesian Statistics 8. Oxford University Press*, 2007.

[2] C. Carvalho, J. Lucas, Q. Wang, J. Chang, J. Nevins, and M. West. High-dimensional sparse factor modelling - applications in gene expression genomics. In *Journal of the American Statistical Association*, 2008.

[3] M. West. Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics 7*, 2003.

[4] Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11, 2007.

[5] Y. W. Teh, H. Daumé III, and D. M. Roy. Bayesian agglomerative clustering with coalescents. In *Advances in Neural Information Processing Systems*, volume 20, 2008.