Bayesian matrix factorisation for learning latent pathway inter-dependencies and identifying responsive pathways

Naruemon Pratanwanich and Pietro Lio'

Computer Laboratory, University of Cambridge, UK

Abstract

Background In a pathway genes are grouped based on their properties such as their functionality and their physical interactions, and will represent a layer of complexity above the single genes. Thus, identifying pathways responsive to drug treatments or disease conditions has been considered a further step beyond predicting individual differentially expressed genes. Since a gene can regulate others in different pathways via a series of actions, it implies that pathways themselves are correlated. Although it is of great interest in the identification of responsive pathways, the computational methods for learning pathway dependency structure in a large scale is limited. To the best of our knowledge, our work is the first effort to reconstruct a graph demonstrating the pairwise relationships among pathways concurrently with responsive pathway identification from gene expression data.

Methodology We have developed a Bayesian matrix factorization of gene expression data by treating annotated pathways as latent variables. Importantly, we propose a Gaussian distribution with respect to an undirected graph called Gaussian Markov Random Field (GMRF) on one of the factors in the matrix factorisation. With GMRF, the pairwise dependencies among pathways can be captured simultaneously with the pathway identification. More formally, we have performed the analysis of the differential gene expression data of G genes in D condtions. Our assumption is that the differential gene expression data arise from the effects of perturbed genes existing in the pathways responsive to each condition *e.g.* a drug or a disease. This concept can be implemented by matrix factorisation, where the observed differential gene expression data matrix $\mathbf{X} \in \mathbb{R}^{G \times D}$ is decomposed into two matrices: $\mathbf{X} \sim \mathbf{BS}$ [6]. The first matrix $\mathbf{B} \in \mathbb{R}^{G \times P}$ denotes the strength of gene membership in each pathway, called a gene-pathway matrix. The second matrix $\mathbf{S} \in \mathbb{R}^{P \times D}$ corresponds to the degree that each pathway responsive to each condition, called a pathway-condition matrix. It is noted that P is the number of pathways shared by matrix $\mathbf{B}$ and matrix $\mathbf{S}$, since pathways are regarded as the latent factors underlying both genes and conditions. Similar to the FacPad [6], we used the prior knowledge of gene-pathway associations matrix $\mathbf{K} \in \{0,1\}^{G \times P}$ from Kyoto Encyclopedia of Genes and Genomes (KEGG) [3] to force the sparsity pattern of matrix $\mathbf{B}$. Unlike the FacPad [6], we developed a GMRF model with the aim of capturing pathway dependencies. A Gaussian distribution was imposed on matrix $\mathbf{S}$ with a precision (inverse covariance) matrix $\mathbf{\Phi} \in \mathbb{R}^{P \times P}$. As a result, we can draw the undirected links among pathways corresponding to the non-zero off-diagonal elements of matrix $\mathbf{\Phi}$. Meanwhile, we are still able to identify responsive pathways from matrix $\mathbf{S}$. Figure 1 shows our methodology in a schematic view.

Results In simulation studies, our model can identify responsive pathways (matrix $\mathbf{S}$) with high accuracy as well as reconstruct any arbitrary pathway-pathway structures (matrix $\mathbf{\Phi}$) with high precision and recall. Furthermore, we applied our model to analyse the gene expression data of drug treatments from the Connectivity Map project (build 02) [5]. We find that our model performs favourably in the case of multiple enriched pathways, supporting our assumption of pathway correlations. This capability is beneficial to the analysis of disease gene expression data in the future as the number of associated pathways is increasing in proportion to disease complexity, comorbidity [1], pregression time [2]. Moreover, the overall performance of our approach surpass the existing methods such as the well-known Gene Set Enrichement Analysis (GSEA) mainly because of the GMRF model. Concurrently, our model can infer pathway inter-interactions, some of which were previously reported in literature such as the well-studied association of FcεRI signaling pathway and NF-κB signaling pathway [4]. Other unprecedented interactions may suggest new hypotheses. In addition, our model can capture the dependencies among pathways although they have no overlapping genes. This result is in accordance with the fact that pathways can interact at any other molecular level apart from genes.

Conclusion The novel connectivity network among pathways can assist the discovery of linkages between two perturbed states such as comorbiditiy and drug-disease mechanisms, bringing on an important step towards personlised medicine.

Keywords: Bayesian inference, matrix factorisation, Gaussian Markov Random Field, biological pathway, gene expression data

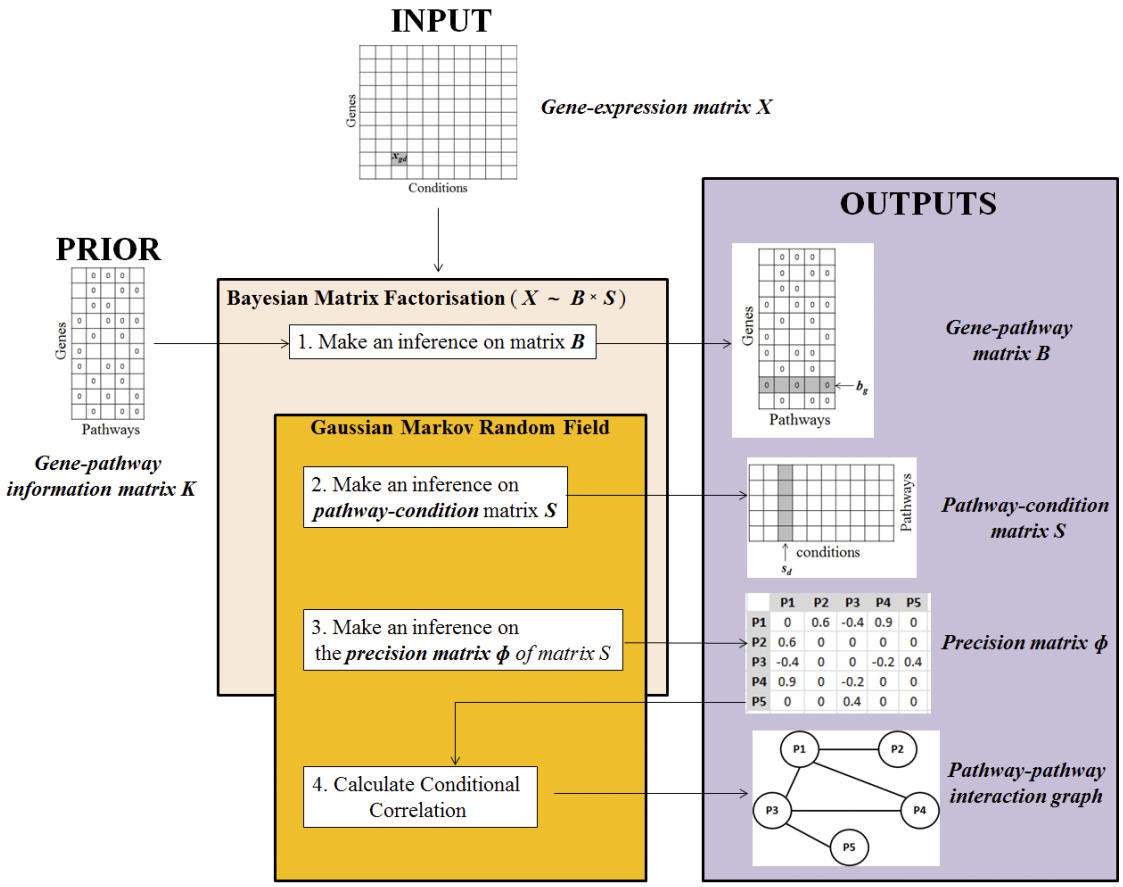

Figure 1: **Method overview.** Matrix $\mathbf{X}$ and Matrix $\mathbf{K}$ are the inputs of our models. The matrix $\mathbf{X}$ is then decomposed into matrix $\mathbf{B}$ and matrix $\mathbf{S}$: $\mathbf{X} \sim \mathbf{BS}$. In the first step, the matrix $\mathbf{K}$ is used as a prior to guide the sparsity pattern for the inference of matrix $\mathbf{B}$. The next two steps are to infer the matrix $\mathbf{S}$ and its precision matrix $\mathbf{\Phi}$ subject to the GMRF model. Finally, given the matrix $\mathbf{\Phi}$, we can calculate the conditional correlation of every pathway pair for the construction of the pathway interaction network.

References

[1] Enrico Capobianco and Pietro Lio. Comorbidity: a multidimensional approach. *Trends in molecular medicine*, 2013.

[2] Leroy Hood. Systems biology and p4 medicine: past, present, and future. *Rambam Maimonides medical journal*, 4(2):e0012, 2013.

[3] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, (Database issue):D109–14, 2012.

[4] Stefanie Klemm and Jürgen Ruland. Inflammatory signal transduction from the fcεri to nf-κb. *Immunobiology*, 211(10):815–820, 2006.

[5] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A Armstrong, Stephen J Haggarty, Paul A Clemons, Ru Wei, Steven A Carr, Eric S Lander, and Todd R Golub. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–35, 2006.

[6] Haisu Ma and Hongyu Zhao. FacPad: Bayesian sparse factor modeling for the inference of pathways responsive to drug treatment. *Bioinformatics*, 28(20):2662–70, 2012.