DNA Methylation Network Estimation with Sparse Latent Gaussian Graphical Model

Bernard Ng[1], **Sina Jafarzadeh**[1], **Daniel Cole**[2], **Anna Goldenberg**[2], **Sara Mostafavi**[1]
[1]Centre for Molecular Medicine and Therapeutics, UBC, Canada
[2]SickKids Hospital, University of Toronto, Canada
bernardyng@gmail.com

Abstract

Inferring molecular interaction networks from genomics data is important for advancing our understanding of biological processes. Whereas considerable research effort has been placed on inferring such networks from gene expression data, network estimation from DNA methylation data has received very little attention due to the substantially higher dimensionality and complications with result interpretation for non-genic regions. To combat these challenges, we propose here an approach based on sparse latent Gaussian graphical model (SLGGM). The core idea is to perform network estimation on q latent variables as opposed to d CpG sites, with $q<<d$. To impose a correspondence between the latent variables and genes, we use the distance between CpG sites and transcription starting sites of the genes to generate a prior on the CpG sites' latent class membership. We evaluate this approach on synthetic data, and show on real data that the gene network estimated from DNA methylation data significantly explains gene expression patterns in unseen datasets.

1 Introduction

Uncovering networks of interacting genes provides insights into the biological mechanisms that give rise to phenotypes. Under this network perspective, genes correspond to nodes with their interactions modeled via weighted edges. Over the past decade, gene interaction networks have primarily been constructed from gene expression data due to their large abundance [1]. In contrast, less effort has been placed on using DNA methylation data, despite their relatively more robust (less dynamic) nature and increasing availability for large cohorts. Most studies that do use methylation data estimate networks by directly correlating all CpG site pairs, with a focus on module detection [2-6]. However, the typical small sample-to-variable ratio limits the accuracy of the resulting networks [7]. Also, interpreting methylation networks is more difficult, since less is known about the functional role and gene targets of non-coding regulatory regions. Some studies employ canonical correlation analysis (CCA) to combine gene-proximal CpG sites in reducing dimensionality and enabling gene level interpretation [8; 9], but CCA cannot distinguish direct interactions from indirect interactions. To address the above challenges, we propose here an approach based on sparse latent Gaussian graphical model (SLGGM) [10]. The idea is to estimate a network between q latent variables as opposed to d CpG sites, and tie the latent variables to genes via a prior on the CpG-to-gene mapping. This way, the scale of the network estimation problem is greatly reduced and gene level interpretation is facilitated. Also, SLGGM inherently

estimates partial correlation, which helps isolate direct interactions [11]. We assess this approach on synthetic data, and further show that the gene network constructed from a large-scale methylation dataset (ROSMAP [12]) significantly explains gene expression patterns of unseen datasets from various related tissues (GTEx [13]).

2 Methods

2.1 Sparse Latent Gaussian Graphical Model

Given a $n \times d$ DNA methylation data matrix, $\mathbf{X}$, where n is the number of samples and d is the number of CpG sites, our goal is to learn a $q \times q$ sparse inverse covariance matrix, $\mathbf{K}$, where $q<d$ and $\mathbf{K}_{ij} \neq 0$ indicates that genes i and j are conditionally associated with each other, i.e. only direct interactions are captured. This problem can be formulated as finding a $\mathbf{K}$ that maximizes the following distribution [10]:

$$P(\mathbf{X},\mathbf{Z},\mathbf{L},\boldsymbol{\Sigma},\mathbf{K}) = P(\mathbf{X} \mid \mathbf{Z},\mathbf{L},\boldsymbol{\Sigma})P(\mathbf{L} \mid \mathbf{K})P(\mathbf{K})p(\mathbf{Z}), \tag{1}$$

$$P(\mathbf{X} \mid \mathbf{Z},\mathbf{L},\boldsymbol{\Sigma})P(\mathbf{Z}) = \prod_{i=1}^{d}\prod_{j=1}^{q} \boldsymbol{\pi}_j^{\mathbf{Z}_{ij}} N(\mathbf{X}_i \mid \mathbf{L}_j, \sigma_j^2 \mathbf{I}_{n\times n})^{\mathbf{Z}_{ij}}, \tag{2}$$

$$P(\mathbf{L} \mid \mathbf{K})P(\mathbf{K}) = N(\mathbf{L} \mid \mathbf{0}_{n\times q}, \mathbf{K}^{-1})L(\mathbf{K}). \tag{3}$$

Each column of $\mathbf{X}$, denoted as $\mathbf{X}_i$, is assumed to follow a mixture of Gaussians with means $\mathbf{L} = (\mathbf{L}_1,\ldots,\mathbf{L}_q)$ and variances $\boldsymbol{\Sigma} = \{\sigma_1^2\mathbf{I}_{n\times n},\ldots,\sigma_q^2\mathbf{I}_{n\times n}\}$. The $n \times q$ latent variable $\mathbf{L}$ is assumed to follow a centered multivariate Gaussian with an inverse covariance $\mathbf{K}$ having a Laplace prior $L(\mathbf{K})$ to promote sparse $\mathbf{K}$ estimation, since $n<<q$ typically. $\mathbf{Z}$ is a $d \times q$ binary matrix, where $\mathbf{Z}_{ij}=1$ indicates that $\mathbf{X}_i$ belongs to latent class j. Each $\mathbf{X}_i$ is assumed to belong to only one latent class as required for standard Gaussian mixture models. Optimal $\mathbf{L}$, $\boldsymbol{\Sigma}$, and $\mathbf{K}$ can be found with coordinate ascent [10], i.e. alternate between finding $\mathbf{L}$, $\boldsymbol{\Sigma}$, and $\mathbf{K}$ that maximizes (1). The update procedures for $\mathbf{L}$, $\boldsymbol{\Sigma}$, and $\mathbf{K}$ are described in Sections 2.2, 2.3, and 2.4, respectively.

To deploy SLGGM for learning a gene network from methylation data, we impose a distance-based prior by setting $\boldsymbol{\gamma}_{ij} = E(\mathbf{Z}_{ij})$ to $f(1/\mathbf{D}_{ij})$, where $\mathbf{D}_{ij}$ is the base-pair distance between CpG site i and the transcription starting site (TSS) of gene j, and $f(\cdot)$ normalizes $1/\mathbf{D}_{ij}$ such that $\Sigma_j\, \boldsymbol{\gamma}_{ij}=1$. The assumption is that CpG sites closer to a given gene are more likely to affect that gene, as often observed [14]. In this work, we fix $\boldsymbol{\gamma}_{ij}$ to retain the correspondence between the genes and the latent variable $\mathbf{L}$, and leave updating $\boldsymbol{\gamma}_{ij}$ to refine the CpG-to-gene mapping for future work. Also, we note that in contrast to methylation values per sample, which tend to be bimodal, the distribution of methylation values at each CpG site is by and large unimodal and approximately Gaussian in the real data, hence meeting the assumptions of SLGGM.

2.2 L Update

By taking the log of (1), retaining terms involving $\mathbf{L}$, and differentiating w.r.t. $\mathbf{L}_j$, one can show that the optimal $\mathbf{L}_j$ at iteration k+1 is given by:

$$\mathbf{L}_j^{k+1} = \left(\mathbf{X}\boldsymbol{\gamma}_j - \sigma_j^{2k}\left(\mathbf{L}_{\tilde{j}}^k \mathbf{K}_{\tilde{j}j}^k\right)\right) \Big/ \left(n_j + \sigma_j^{2k}\mathbf{K}_{jj}^k\right), \tag{4}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_q)$ is a $d \times q$ matrix containing the distance-based prior, $n_j = \Sigma_i\, \boldsymbol{\gamma}_{ij}$, and $\tilde{j} = \{1,\ldots,q\} \setminus j$, i.e. $\mathbf{L}_{\tilde{j}}^k$ denotes all columns of $\mathbf{L}^k$ except the j^{th} and $\mathbf{K}_{\tilde{j}j}^k$ denotes the j^{th} column of $\mathbf{K}^k$ excluding the j^{th} element. We initialize $\mathbf{L}$ as $\mathbf{X}\boldsymbol{\gamma}$.

2.3 σ_j^2 Update

Similarly, differentiating the log of (1) w.r.t. σ_j^2, the optimal σ_j^2 is given by:

$$\sigma_j^{2^{k+1}} = 1/n \sum_{m=1}^{n} \sum_{l=1}^{d} (\mathbf{X}_{ml} - \mathbf{L}_{mj}^{k})^2 \boldsymbol{\gamma}_{lj} / n_j \,, \tag{5}$$

where the noise variance of $\mathbf{X}$ is assumed to be the same across the n samples for each latent class j. σ_j^2 is initialized to 1.

2.4 K update

Finding $\mathbf{K}$ that maximizes (1) is equivalent to solving the following problem [10]:

$$\min_{\mathbf{K}>0} \quad -ln|\mathbf{K}| + tr(\mathbf{L}^T\mathbf{L}\mathbf{K}) + \lambda \|\mathbf{K}\|_1 \,, \tag{6}$$

where the first two terms correspond to $-\log(N(\mathbf{L}|\mathbf{0}_{n\times q},\mathbf{K}^{-1}))$ with constants removed, and the last term is the l_1 norm of $\mathbf{K}_{ij}$, $i\neq j$, corresponding to $-\log(L(\mathbf{K}))$, which promotes a sparse $\mathbf{K}$ estimate. For solving (6), we employ QUIC [15] and BigQUIC [16], depending on q. λ is set to $\alpha\lambda_{\max}$ for different α's, where $\lambda_{\max}=max_{ij}\,|\mathbf{L}^T\mathbf{L}|$, $i\neq j$.

3 Materials

Synthetic data. We created 60 synthetic datasets for three q/n ratios: 1=100/100, 5=500/100, and 0.5=500/1000, with 20 datasets for each q/n ratio. For each dataset, we first generated a random $q\times q$ sparse inverse covariance matrix, $\mathbf{K}$, with 10% of the elements being non-zero. We then drew a random $n\times q$ matrix, $\mathbf{L}$, from $N(\mathbf{0}_{n\times q}, \mathbf{K}^{-1})$. A $n\times 1$ vector, $\mathbf{X}_i$, was then drawn from $N(\mathbf{L}_j,\sigma_j^2\mathbf{I}_{n\times n})$ for i=1 to d, where the choice of j was based on $\mathbf{Z}_{ij}$=1. The CpG-to-gene mapping, $\mathbf{Z}$, was established by assigning each CpG site to its closest gene based on the real data. On average, d≈3,000 and 15,000 CpG sites for q=100 and 500 random genes. The membership prior, $\boldsymbol{\gamma}_{ij}$, was set to $f(1/\mathbf{D}_{ij})$, where $\mathbf{D}_{ij}$ is the distance between CpG site i and TSS of gene j, and $f(\cdot)$ ensures $\Sigma_j\,\boldsymbol{\gamma}_{ij}$=1. Since $\mathbf{Z}$ is supposedly unknown, we included $\mathbf{D}_{ij}$ of all CpG-gene pairs that are within 1Mb from each other in computing $\boldsymbol{\gamma}$ to emulate uncertainty in the CpG-to-gene mapping. These smaller scale problems permit rigorous testing of SLGGM within a practical amount of computation time.

Real data. DNA methylation data (450K Illumina array) were generated from 702 post-mortem samples of the dorsolateral prefrontal cortex as part of the ROSMAP study [12] (available on Synapse). In addition to standard quality control, age at death, sex, batch, post mortem interval, and top 10 principal components (PC) from the DNA methylation data were regressed out. For method evaluation, we used gene expression data (13,484 genes) from 508 individuals of the ROSMAP study [17]. Standard quality control was performed, and age at death, sex, batch, post-mortem interval, RNA integrity, sequencing depth, genotyping PCs, and top 10 PCs from expression data were removed. Only genes and CpG sites within 1Mb from each other were retained, resulting in 416,452 CpG sites and 13,004 genes. We also downloaded gene networks constructed from expression data of 13 brain tissues as well as expression data of 35 peripheral tissues from the GTEx portal [13].

4 Results and Discussion

Synthetic data. Average precision and recall estimated by applying SLGGM (with QUIC), and comparing the sparse edge patterns of the ground truth and estimated networks are plotted in Fig. 1. For comparison, we applied SGGM (with QUIC) to the ground truth $\mathbf{L}$. The performance of SLGGM applied to $\mathbf{X}$ is similar to SGGM applied to $\mathbf{L}$ for the q/n ratios tested, with SLGGM's performance being well within the standard error of SGGM (not plotted to avoid clutter). If we weight the ground truth edges by $|\mathbf{K}_{ij}|$ in the precision and recall estimation, both SLGGM and SGGM achieved a precision of ~1, indicating that the stronger edges were reliably extracted.

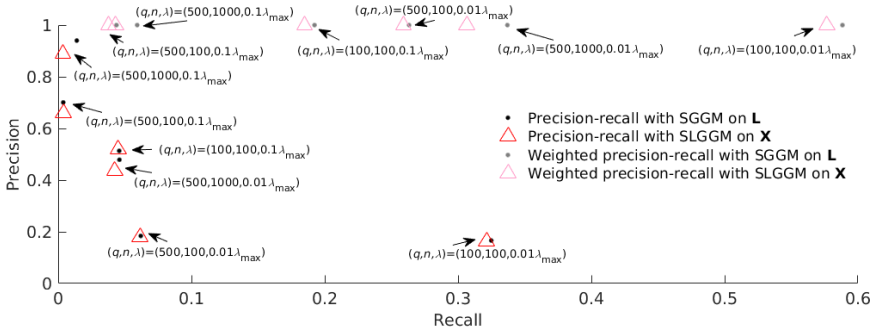

Fig. 1: Synthetic data results. For various q/n ratios, applying SLGGM on $\mathbf{X}$ ($n\times d$) resulted in similar performance as applying SGGM on the ground truth $\mathbf{L}$ ($n\times q$), where $d>>q$.

Real data. Since no ground truth network is available for real data, we used gene networks estimated from ROSMAP and GTEx expression data for assessing SLGGM's performance. With the ROSMAP gene expression data, we applied SGGM (BigQUIC, $\lambda=0.5\lambda_{\max}$) and stability selection [18] on 100 random subsamples ($0.8n$ subjects each). Network edges with selection frequency >0.5, i.e. estimated to be non-zero for >50% of the subsamples, were assumed reliable. With the 13 GTEx brain tissue-based networks, we assumed edges that were non-zero for >50% of the networks were reliable. As a baseline, we compared the reliable network edges between ROSMAP and GTEx. With ROSMAP as the reference, the estimated precision and recall are 0.023 and 0.061. If we weight the reliable network edges by the selection frequency, the weighted precision and recall are 0.917 and 0.062, demonstrating that the more reliable edges tend to be extracted. Comparing the network edges extracted by SLGGM (BigQUIC, $\lambda=0.2\lambda_{\max}$) and the reliable edges in the ROSMAP and GTEx expression networks, the precision and recall are 0.001 and 0.059 for ROSMAP and 0.001 and 0.023 for GTEx. The low precision is likely due to the small sample size. Nonetheless, the weighted precisions of SLGGM are 0.920 for ROSMAP and 0.306 for GTEx, indicating that the more reliable network edges are indeed extracted. We note that assessing only the sparse edge patterns neglect the information encoded by the edge weights. To capture this information, we applied kernel machines [19] with the ROSMAP and GTEx expression data as response and $\mathbf{K}^{-1}$ estimated with SLGGM as the kernel. Specifically, we averaged the expression values across subjects for each gene but without subject mean removal to retain the genome-wide expression pattern. This pattern of average expression values was used as the response. The association between the ROSMAP expression pattern and $\mathbf{K}^{-1}$ is statistically significant (p=0 to machine precision), demonstrating that $\mathbf{K}^{-1}$ well explains the gene expression pattern. Also, over the 35 tissues in the GTEx data, stronger associations are observed for brain tissues on average, Fig. 2.

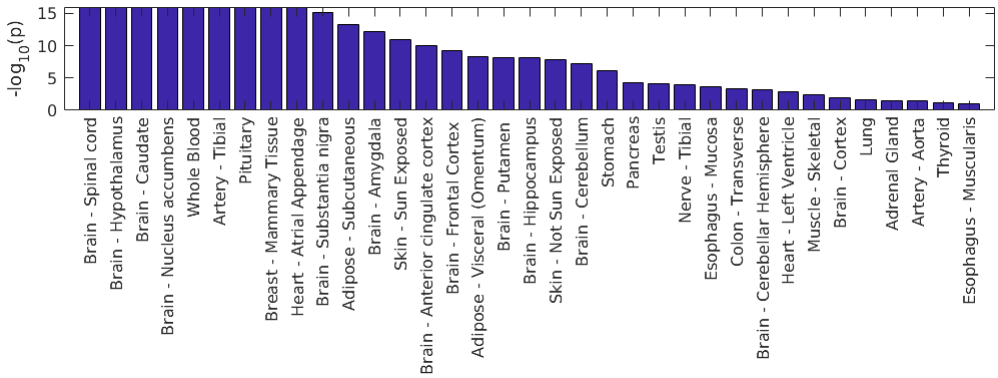

Fig. 2: Real data results. The $\mathbf{K}^{-1}$ estimated by SLGGM well explains the expression pattern of brain tissues. Note p-values that are 0 (to machine precision) were set to 10^{-16} for clearer visualization.

References

[1] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., & di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol Syst Biol, 3*, 78.

[2] Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., …, Ophoff, R. A. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol, 13*(10), R97.

[3] Ma, X., Liu, Z., Zhang, Z., Huang, X., & Tang, W. (2017). Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data. *BMC Bioinformatics, 18*(1), 72.

[4] Didier, G., Brun, C., & Baudot, A. (2015). Identifying communities from multiplex biological networks. *PeerJ, 3*, e1525.

[5] Cantini, L., Medico, E., Fortunato, S., & Caselle, M. (2015). Detection of gene communities in multi-networks reveals cancer drivers. *Sci Rep, 5*, 17386.

[6] Jiao, Y., Widschwendter, M., & Teschendorff, A. E. (2014). A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics, 30*(16), 2360-2366.

[7] Guo, W., Calixto, C. P. G., Tzioutziou, N., Lin, P., Waugh, R., Brown, J. W. S., & Zhang, R. (2017). Evaluation and improvement of the regulatory inference for large co-expression networks with limited sample size. *BMC Syst Biol, 11*(1), 62.

[8] Bartlett, T. E., Olhede, S. C., & Zaikin, A. (2014). A DNA methylation network interaction measure, and detection of network oncomarkers. *PLoS One, 9*(1), e84573.

[9] Zhou, Y., Liu, Y., Li, K., Zhang, R., Qiu, F., Zhao, N., & Xu, Y. (2015). ICan: an integrated co-alteration network to identify ovarian cancer-related genes. *PLoS One, 10*(3), e0116095.

[10] Celik, S., Logsdon, B., & Lee, S.-I. (2014). Efficient dimensionality reduction for high-dimensional network estimation. *Proc. Int. Conf. Mach. Learn.*, pp. 1953-1961, 2014.

[11] Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika, 94*(1), 19-35.

[12] De Jager, P. L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L. C., Yu, L., …, McCabe, C. (2014). Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature Neuroscience, 17*(9), 1156-1163.

[13] GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. (2015). *Science, 348*(6235), 648-660.

[14] Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., & Blanchette, M. (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol, 15*(2), R37.

[15] Hsieh, C. J., Sustik, M. A., Ravikumar, P., & Dhillon, I. S. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in Neural Information Processing Systems*, pp. 2330-2338.

[16] Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., & Poldrack, R. (2013). BIG & QUIC: Sparse inverse covariance estimation for a million variables. *Proc. Advances in Neural Information Processing Systems*, pp. 3165-3173.

[17] Lim, A. S., Klein, H. U., Yu, L., Chibnik, L. B., Ali, S., Xu, J., …, De Jager, P. L. (2017). Diurnal and seasonal molecular rhythms in human neocortex and their relation to Alzheimer's disease. *Nat Commun, 8*, 14931.

[18] Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(4), 417-473.

[19] Liu, D., Lin, X., & Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics, 63*(4), 1079-1088.