Author: Kpogbezan, G.B.
Title: Prior information and variational Bayes in high dimensional statistical network inference
Issue Date: 2018-12-10

Introduction

The rapid evolution of data acquisition technologies in the last 25 years has enabled a massive production of high-dimensional and highly complex datasets in many scientific domains, including genomics, finance and statistical pattern recognition, to name but a few. In genomics, high-throughput platforms such as microarrays provide measurements of many thousands of molecular aspects (e.g. gene expression) of the cell. While between 20,000 and 25,000 genes of a single patient are easily characterized simultaneously, the number n of patients runs typically in the tens or hundreds. This typically gives rise to data characterized by 'large p, small n'. The analysis of such high-dimensional data ($n \ll p$) is very challenging as the traditional statistical methods become useless. For instance, the sample covariance matrix becomes rank deficient and can not be inverted. Our contribution in this thesis consists of incorporating prior knowledge in the analysis of these data, which we model using mainly graphical models.

0.1 Graphical models

A graphical model is a way to 'marry' probability theory with graph theory. A graph $\mathcal{G}$, as used in this thesis, is a pair $(\mathcal{I}, \mathcal{E})$, where $\mathcal{I}$ is a set of indices (or vertices) and the set of edges $\mathcal{E}$ is a subset of the set $\mathcal{I} \times \mathcal{I}$ of ordered pairs of distinct vertices . An edge between vertices r and s is *undirected* if both (r, s) and (s, r) are in $\mathcal{E}$, whereas an edge $(r, s) \in \mathcal{E}$ whose opposite $(s, r) \notin \mathcal{E}$ is called *directed.* In the diagram of $\mathcal{G}$ an *undirected* edge is usually represented by a line between the corresponding vertices whereas a *directed* edge is represented by an arrow. A graph is called undirected if it possesses only undirected edges, and it is called directed if all edges are directed. Let $Y = (Y_1, Y_2, \cdots, Y_p)$ denote a random field with index set $\mathcal{I} = \{1, 2, \cdots, p\}$ taking values in probability spaces $\mathcal{Y}_i$, $i \in \mathcal{I}$ and $\mathcal{Y} = \times_{i \in \mathcal{I}} \mathcal{Y}_i$ being the product space. Furthermore, let $\mathcal{D}$ denote the set of all probability distributions on $\mathcal{Y}$. A graphical model consists of a graph $\mathcal{G} = (\mathcal{I}, \mathcal{E})$ and a set of properties (called *Markov properties*) that together determine a sub-family of probability distributions in $\mathcal{D}$. According to both the type of the allowed graphs $\mathcal{G}$ and the set of properties we distinguish between several graphical models - e.g. Markov networks.

Markov networks (or Markov Random Fields) arise when only undirected graphs over the p vertices are allowed and the family of distributions in $\mathcal{D}$ consists of probability distributions on $\mathcal{Y}$ obeying the *local Markov property*. The latter states that: conditional on its adjacent variables, any variable is independent of all the remaining variables. If furthermore, the distribution admits a strictly positive density with respect to some product measure μ on $\mathcal{Y}$ then, the local Markov property is equivalent to the *pairwise Markov property* [108]: any two non-adjacent variables are conditionally independent given all other variables. In the latter case the pairwise Markov property is in turn equivalent to the *global Markov property*. An undirected graph satisfiying the local Markov property is also referred to as a Conditional Independence Graph (CIG). Conditional independence graphs are of prime interest in this thesis.

Other type of graphical models are Bayesian networks which are based on directed acyclic graphs (DAG) [78, 79] and independence chain graphs based on chain graphs [4, 78]. Chain graphs contain both directed and undirected edges.

Gaussian Graphical Models.

A Gaussian Graphical Model (GGM) assumes data are drawn from a multivariate normal distribution:

$$Y \sim \mathrm{N}(0, \Omega_p^{-1}) \tag{1}$$

where Y is a p-dimensional random vector comprising the p random variables $Y_1, \ldots, Y_p$ corresponding to the nodes of $\mathcal{I}$ and Ω_p^{-1} is a non-singular $(p \times p)$-dimensional covariance matrix. The matrix Ω_p is referred to as the *precision matrix*. For a GGM the edge set $\mathcal{E}$ of the underlying conditional independence graph corresponds to the nonzero elements of Ω_p [78]. Hence, reconstructing the conditional independence graph is equivalent to determining the non-zero elements of this matrix.

Both frequentist and Bayesian approaches are used in the literature to estimate the conditional independence graph. Standard frequentist approaches rely on penalized likelihood estimation. The augmented penalty to the likelihood aims at resolving the high-dimensionality issue of the data. The commonly employed lasso and ridge penalties amount to adding the ℓ_1- and ℓ_2-norm, respectively, of the precision matrix to the likelihood [43, 139]. Both penalties shrink the elements of the precision matrix towards zero. The lasso penalty may shrink these to exactly zero, thus performing variable selection. The ridge penalization requires a post-hoc step to sparsify its precision matrix estimate. The usual Bayesian approach is to put a prior π on the structure of the conditional independence graph $\mathcal{G}$ and given $\mathcal{G}$ a prior $p(\Omega_p|\mathcal{G})$ on the precision matrix [33, 50, 68]. The joint density is given by

$$p(\mathcal{G}, \Omega_p, \mathbf{Y}) = \pi(\mathcal{G})p(\Omega_p|\mathcal{G})p(\mathbf{Y}|\mathcal{G}, \Omega_p) \qquad \text{where} \qquad \mathbf{Y} = (Y^1, \cdots, Y^n)$$

and a joint structural and quantitative learning is performed by computing the posterior $p(\mathcal{G}, \Omega_p|\mathbf{Y}) \propto p(\mathcal{G}, \Omega_p, \mathbf{Y})$. Except in very small problems, the space of graphs to consider is typically restricted to - e.g. decomposable graphs, forests, or trees.

In a multivariate Gaussian distribution, all conditional distributions are Gaussian linear regressions. Hence, to Gaussian model determination (with non-decomposable graphs), [33] propose estimating these conditional regressions from data using (Bayesian) sparse regression techniques (often called Simultaneous Equations Models).

Simultaneous Equations Models.

Simultaneous Equations Models (SEMs) are a framework for modeling and coding path diagrams. We will use the *very basic* SEMs consisting in modeling the full conditional distribution of each univariate random variable Y_i, $i \in \mathcal{I}$ and thus resulting in a system of regressions

$$Y_i = \sum_{t \neq i} \beta_{i,t} Y_t + \epsilon_i, \quad \epsilon_i \perp\!\!\!\perp \{Y_t; t \neq i\}, \quad i \in \mathcal{I}. \tag{2}$$

SEMs are flexible tools and computationally very attractive. They account for experimental or biological covariates in the regressions and are appropriate for many types of data distribution [3, 25, 115]. They allow the integration of multiple data sets and at the same time are scalable to large datasets in their computational complexity. Moreover, there is an equivalence between GGM and SEMs, namely, the regression coefficients $\beta_i = (\beta_{i,t} : t \neq i)$ can be expressed in the precision matrix of Y as [97]

$$\beta_{i,t} = -\frac{(\Omega_p)_{it}}{(\Omega_p)ii},$$

in which case the residuals in (2) when regressing a single coordinate Y_i of a multivariate Gaussian vector linearly on the other coordinates Y_t, for $t \neq i$, are Gaussian. That means, the (non)zero entries in the ith row vector of the precision matrix Ω_p

correspond to the (non)zero coordinates of β_i. Consequently, the problem of identifying the Gaussian graphical model can be cast as a variable selection problem in the p regression models (2). This approach of recasting the estimation of the (support of the) precision matrix as a collection of regression problems was first suggested by [33] and latter introduced by [97], who employed Lasso regression [43, 130] to estimate the parameters. Other variable selection methods can be employed as well [73].

In this thesis, we introduce a Bayesian approach of the SEMs. In Chapter 1 we develop a Bayesian formulation of the SEMs with Gaussian, ridge-type priors on the regression coefficients. In Chapter 2, we extent the latter model to incorporate prior knowledge on the conditional independence graph. A disadvantage of the Gaussian priors employed in these papers is that they are not able to selectively shrink parameters, but shrink them jointly towards zero (although prior information used in Chapter 2 alleviates this by making this dependent on prior group). Chapter 3 proposes a general framework for analysing large-scale data sets with complex dependence structures using a collection of linear regression models corresponding to p characteristics (e.g. genes). The *horseshoe* prior [19, 20] has been introduced in order to better model the sparsity of the explanatory variables, thus being able to selectively shrink parameters towards zero. Reconstruction of conditional independence graphs by incorporating prior information is a special case of the proposed framework in Chapter 3.

0.2 Prior information

High-dimensional modeling is important in many scientific areas but is also a challenging task. In genomics, the identification of gene regulatory networks is crucial for understanding gene function, and hence important for both treatment and prediction of diseases. This challenge of analysing data consisting of few replicate measurements against large number of covariates "$n \ll p$" can be alleviated by incorporating external (or "prior") information in the analysis. In gene regulatory networks reconstruction, prior knowledge on the topology on the to-be-reconstructed network is readily available. For instance, the current beliefs on interactions among genes is condensed in repositories like KEGG and Reactome. The Bayesian framework provides a natural architecture to incorporate and accommodate such prior information. It may be believed that such priors can affect the integrity of the current study results and can even lead to conclusions that are driven not by the data but by a prior resulting from some non-relevant previous studies. However, the incorporation in a soft manner, so that it informs the analysis if correct, but can be overruled if completely incompatible with the data, helps overcoming this situation.

Many works have already been devoted to incorporating prior knowledge into network reconstruction. These sudies include [64, 65, 87, 102, 127, 149] for the incorporation of many types of different prior knowledge, including literature-based knowledge in Bayesian network learning and dynamic Bayesian network learning. However, none of these proposed methods explicitly estimate the agreement of the prior knowledge with the data at hand.

Our approach in this thesis is based on prior modelling of the regression parameters of the SEMs in a soft manner using respectively the Gaussian, ridge-type prior in Chapter 2 with a prior on the regularization parameter that depends on external information, and the horseshoe prior in Chapter 3 with a prior on the sparsity index that also depends on external information. Multiple sources of information are incorporated simultaneously. The proposed scheme attaches a latent variable to each source of information independently across sources. These latent variables enter the prior distributions of the coordinates of β_i, which marginally given the latent variable are scale mixtures of the normal distribution. Our soft borrowing of prior information is based on the estimation of these prior hyperparameters by an appealing empirical Bayes procedure (called *global empirical Bayes*).

In Chapter 4, we investigate how gene regulatory networks (GRNs) can be reconstructed from combining observational and time-course gene expression (cell line) data. We present strategies to borrow information respectively in a soft and hard manner from either study type in reconstructing both the CIG-based gene regulatory network and the *human* independence (or time-series) chain graph. The hard borrowing of prior information here means that the prior information is hard-wired in our analysis, because we intend to steer the results for reasons of interpretation or because we have a strong belief in the prior information.

0.3 Variational Bayes approximation

In Bayesian statistics, a prior is assigned to the parameter of interest. The prior belief is subsequently updated by means of current data and inference is based on the posterior distribution. Traditional Bayesian computation methods rely on Markov Chain Monte Carlo (MCMC). However, modern datasets (e.g. gene expression data) are extremely high-dimensional and the use of MCMC is often a computational bottleneck due to high-dimensional integral computations. Approximate Bayesian methods have emerged in recent years as fast alternatives methods to MCMC to overcome these shortcomings. Among the proposed methods *variational Bayes approximations* seem very promising.

Variational approximations are a set of deterministic methods used to make approximate inference for parameters in complex statistical models. The name *variational approximations* originates from the mathematical topic known as *variational calculus*. The latter is concerned with the problem of optimizing a functional over a class of functions. The problem becomes usually feasible when the domain of the functional is restricted to some sub-class of functions. The variational Bayes approximation to a distribution is the closest element q^* in a given target set $\mathcal{Q}$ of distributions, usually with "distance" measured by Kullback-Leibler divergence [141]. The set $\mathcal{Q}$ is chosen as a compromise between computational tractability and accuracy of approximation. If θ denote the parameter of interest in a generic Bayesian model and $\mathbf{Y}$ the observed data, the Kullback-Leibler divergence is defined as

$$KL\Big(q||p(\cdot|\,\mathbf{Y})\Big) = \mathbf{E}_q \log \frac{q(\theta)}{p(\theta|\,\mathbf{Y})} = \log p(\mathbf{Y}) - \mathbf{E}_q \log \frac{p(\mathbf{Y},\theta)}{q(\theta)}, \tag{3}$$

where $\theta \mapsto p(\theta|\,\mathbf{Y})$ is the posterior density, the expectation is taken with respect to θ having the density $q \in \mathcal{Q}$, and $(y,\theta) \mapsto p(y,\theta) = p(y|\,\theta)\,\pi(\theta)$ and $y \mapsto p(y) = \int p(y,\theta)\,d\theta$ are the joint density of $(\mathbf{Y},\theta)$ and the marginal density of $\mathbf{Y}$, respectively, in the model with prior density π on θ. Minimization of (3) is equivalent to the maximization of the expression on the far right hand side of (3) which is usually referred to as "the evidence lower bound", or "elbo". By the non-negativity of the Kullback-Leibler divergence it holds

$$\log p(\mathbf{Y}) \geq \mathbf{E}_q \log \frac{p(\mathbf{Y},\theta)}{q(\theta)} =: \mathrm{elbo}(q;\mathbf{Y}). \tag{4}$$

Early applications involved standard distributions such as Gaussian, Dirichlet, Laplace and extreme value models [5–7, 96, 142]. In the present thesis we use nonparametric approximations, restricted only by the assumption that the various parameters are (block) independent. (This may be referred to as *mean-field* variational Bayes, although this term appears to be used more often for independence of all univariate marginals, whereas we use block independence.) The restriction of $\mathcal{Q}$ to a subclass of product densities gives rise to explicit solutions for each product component in terms of the others, leading to iterative scheme for obtaining the solutions. Precisely, the assumption $q(\theta) = \prod_{i=1}^{M} q_i(\theta_i)$ yields

$$\begin{aligned}
\mathrm{elbo}(q;\mathbf{Y}) &= \int q(\theta)\log\frac{p(\mathbf{Y},\theta)}{q(\theta)}d\theta \\
&= \int\prod_{i=1}^{M} q_i(\theta_i)\Big[\log p(\mathbf{Y},\theta) - \sum_{i=1}^{M}\log q_i(\theta_i)\Big]d\theta_1\cdots d\theta_M \\
&= \int q_1(\theta_1)\Big[\int \log p(\mathbf{Y},\theta)\prod_{i=2}^{M} q_i(\theta_i)d\theta_2\cdots d\theta_M\Big]d\theta_1 \\
&\quad - \int q_1(\theta_1)\log q_1(\theta_1)d\theta_1 + \text{terms not involving } q_1
\end{aligned}$$

Define

$$G_1(\theta_1) = \int \log p(\mathbf{Y},\theta)\prod_{i=2}^{M} q_i(\theta_i)d\theta_2\cdots d\theta_M$$

Then,

$$\begin{aligned}
\mathrm{elbo}(q;Y) &= \int q_1(\theta_1)\log\Big(\frac{\exp(G_1(\theta_1))}{q_1(\theta_1)}\Big)d\theta_1 + \text{terms not involving } q_1 \\
&= \int q_1(\theta_1)\log\left(\frac{\exp(G_1(\theta_1))/\int\exp(G_1(\theta_1))d\theta_1}{q_1(\theta_1)}\right)d\theta_1 + \text{terms not involving } q_1 \\
&= -KL\left(q_1||\frac{\exp(G_1(\theta_1))}{\int\exp(G_1(\theta_1))d\theta_1}\right) + \text{terms not involving } q_1.
\end{aligned}$$

Hence by the non-negativity of the Kullback-Leibler divergence, the optimal q_1^* satisfies

$$\begin{aligned}
q_1^*(\theta_1) &= \frac{\exp(G_1(\theta_1))}{\int\exp(G_1(\theta_1))d\theta_1} \\
&\propto \exp\Big[\int\log p(\mathbf{Y},\theta)\prod_{i=2}^{M} q_i(\theta_i)d\theta_2\cdots d\theta_M\Big] \\
&= \exp\Big[E_{q_{-1}}\log p(\mathbf{Y},\theta)\Big]
\end{aligned}$$

where $E_{q_{-1}}$ indicates the expectation over $(\theta_2,\cdots,\theta_M)$ with respect to $q_2\times\cdots\times q_M$. Unfortunately, this expression depends on $q_2,\cdots,q_M$. However, analog expressions for $q_2^*,\cdots,q_M^*$ can be derived, and it is hoped that repeatedly updating a density q_i^* using the current values of $q_1^*,\cdots,q_{i-1}^*,q_{i+1}^*\cdots,q_M^*$ will in the limit yield the maximizer of (4).

Variational Bayes typically produces accurate approximations to posterior means, but have been observed to underestimate posterior spread [12, 18, 48, 94, 131, 143, 145, 151], even for the marginal distributions. We find that in our setting the approximations agree reasonably well to MCMC approximations of the marginals, although the latter take much longer to compute.

High-dimensional Bayesian regressions

In high-dimensional linear regression, a regularization is required to guarantee the existence and accuracy of estimates. This is done in the Bayesian case by introducing a latent variable in the parameter vector θ_i, and the priors on the regression coefficients β_i are referred to as *regularization priors*. *Scale mixtures* of normal distributions are a well-known class of regularization priors giving rise to different priors for different choices of the mixing densities. In Chapter 1 and 2 we used an inverse-gamma mixing density which results in a ridge-type prior for the regression coefficients, whereas in Chapter 3 we employ a half-Cauchy mixing density. The latter is known as *horseshoe prior* [19, 20]. We fix the hyperparameters to the same values across regressions, thus allowing their estimation by our *global empirical Bayes* procedure. The classical empirical Bayes procedure estimates prior hyperparameters by maximizing the marginal likelihood of the data. Our *global empirical Bayes* procedure maximizes a sum of marginal likelihoods which is enabled by our global-local type prior for modeling multiple related high-dimensional and complex datasets. The procedure has been shown to be very efficient, specially in very high-dimensional settings [135]. The global empirical Bayes enables the borrowing of information across regressions.

0.4 Outline of this thesis

The thesis consists of four chapters organized as follows.

Chapter 1: *Gene network reconstruction using global-local shrinkage priors*

This chapter introduces a new global-local shrinkage ridge-type prior for undirected networks reconstruction based on SEMs with posterior edge selection. The proposed approach is computationally fast and outperforms known competitors such as the *graphical lasso*.

Chapter 2: *An empirical Bayes approach to network recovery using external knowledge*

Chapter 2 extends Chapter 1 to include prior information in reconstructing undirected networks. The incorporation of the prior knowledge is done in a soft manner allowing the data at hand to overrule the prior information if not relevant. Furthermore, the proposed method is able to explicitly estimate the agreement of the prior knowledge with the data at hand which is a novelty in incorporating prior information in network inference.

Chapter 3: *Incorporating prior information and borrowing information in high-dimensional sparse regression using the horseshoe and variational Bayes*

Chapter 3 introduces a framework for simultaneously analysing multiple related high-dimensional and complex datasets. Such analyses include gene regulatory network reconstruction, genetic association studies (e.g. eQTL mapping) and data integration in genomics, to name but a few. To enable the analysis for small n relative to large p, we introduce the *horseshoe* prior which allows for sparsity; a desired property for the analysis of such data. We illustrate the approach by two applications, namely: to the reconstruction of gene regulatory networks and to eQTL mapping.

Chapter 4: *Borrow network information between observational and time-course studies: explorations*

This chapter explores several approaches to reconstruct gene regulatory networks from combining observational (*in vivo*) and time-course cell line (*in vitro*) gene expression data. The dynamics of the human cell are assumed to obey a first-order vector autoregression VAR(1) model and it is investigated how the underlying model parameters can be efficiently learned using the two types of datasets. We saw in an application to real data that reconstruction of the conditional independence graph by borrowing information from the cell line data improved significantly. Moreover, our newly proposed strategies to learn the VAR(1) model parameters are able to indicate preserved transcriptional dynamics between the *in vitro* and *in vivo* environments.