Discussion of "Learning Scale Free Networks by Reweighted ℓ_1 regularization"

Deepak Agarwal
Yahoo! Research

The paper provides a new regularization method for covariance selection [1], an important problem in several applications like genetics, network analysis, image analysis, and many others. Stated formally, the problem is as follows — given n iid observations $X_1, \cdots, X_n$ from a p dimensional multivariate normal with mean $\mathbf{0}$ and concentration (inverse covariance) matrix Ω, obtain an estimate of Ω with desirable statistical properties. The maximum likelihood estimator obtained by inverting the sample covariance matrix is noisy and unreliable in high dimensional problems, hence some form of regularization is called for. The entries in Ω can also be interpreted as edge weights in a graph where nodes represent co-ordinates in the observation vector X, and edge weights represent partial associations among nodes. It is often reasonable to assume this graph is sparse in real applications, hence regularization based on ℓ_1 penalty on the elements of Ω that simultaneously ensures sparsity and positive definiteness is attractive. The paper extends this line of research by modifying existing ℓ_1 regularization techniques to explicitly encourage long-range dependencies (power law) in the graph. More specifically, it modifies the convex optimization problem in equation (3) into a non-convex optimization problem that promotes power law degree distribution in the dependency graph as described in equation (4) through a sum-log ℓ_1 penalty. The resulting optimization can be solved through a minorize-maximize(MM) algorithm as described in section 4. The E-step computes the regularization coefficients and the M-step solves a standard Graphical Lasso problem. Only a few iterations are required as illustrated empirically.

The new penalty is a welcome addition to existing literature on ℓ_1 based covariance selection. Indeed, a pure ℓ_1 penalty may fail to capture the hub like structure in the dependency graph that is often the case in several applications. The fact that estimation only involves running a few iterations of re-weighted ℓ_1 regularizers makes the method computationally feasible and simple to apply in practice. Throughout, the authors provide useful insights into the methods. Thanks to some thoughtful comments on an earlier draft by a reviewer of this paper, comparison to related methods are now comprehensive both qualitatively and quantitatively. Despite a large body of work on covariance selection and ℓ_1 regularization, I believe the ideas discussed in this paper will promote future work in covariance selection. The paper is written well and provides a good introduction to someone who is new to this area. The discussion and relationship to existing literature is thorough.

There are several technical issues that have not been fully explored in the paper, it maybe be useful to address them elsewhere. One major issue that is not clear is the choice of parameters ϵ_i. The authors do provide some intuitive explanation and recommendations, but in the experimental section the choice seems quite arbitrary. Robustness to the choice of these parameters is an important issue, does the local minima change drastically with small perturbations in ϵ_is? Another issue not discussed is the sensitivity of solutions to initialization. Finally, scalability is another concern for high dimensional problems. Although Glasso scales to a few thousand nodes, techniques like psuedo-likelihood can handle larger graphs, some experiments on large problems would be useful to understand the potential of the method in large scale applications. The only real-world experiment is on a bioinformatics application where the validation is weak (no accuracy numbers for instance), better experiments that clearly illustrates the benefits of the approach in terms of predictive accuracy will further strengthen the research. This may involve looking for dataset outside bioinformatics where predictive accuracy is the method of choice to evaluate methods. Finally, the method assumes continuous multivariate data from a multivariate normal, generalization to other forms of multivariate data (e.g. binary, counts) would be interesting to consider.

I will end this discussion by pointing out a few other

recent methods that aim to capture long-range dependence in covariance selection but using different techniques. The first method is based on deep belief networks (DBN) described in [3]. Assuming detrended data, the concentration matrix is modeled as linear combination of outer-products of rank one matrices. Mathematically, concentration matrix of X_k given a latent vector $\boldsymbol{h}_k$ is $\sum_{l=1}^{F} h_{k,l} C_l C_l' + I$, where C_l are p dimensional vectors. The latent vectors associated with each observation are assumed to be binary. This induces a mixture distribution that is shown to capture long-range dependencies remarkably well in applications. Computation is however more intensive than the method described in this paper. Nevertheless, comparison on datasets used to fit DBN would be interesting. Another work that is more closely related is to assume a decomposition of Ω as a sum of low-rank and sparse matrices [2]. The low-rank component captures short-range dependencies and the long-range dependencies missed out are captured by the second component that is assumed to be sparse. The resulting optimization problem is convex, the authors provide an efficient algorithm to perform the optimization. A careful comparison to these methods both qualitatively and quantitatively looks exciting.

References

[1] A. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[2] H.Xu, C.Caramanis, and S.Sanghavi. Robust pca via outlier pursuit. In *NIPS*, 2010.

[3] M.Ranzato, V.Mnih, and G.Hinton. Generating more realistic images using gated mrf's. In *NIPS*, 2010.