35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Identifying candidate disease genes using a trace norm constrained bipartite raking model

Cheng H. Lee, *Student Member, IEEE*, Oluwasanmi Koyejo, and Joydeep Ghosh, *Fellow, IEEE*

***Abstract*— Computational prediction of genes that play roles in human diseases remains an important but challenging task. In this work, we formulate candidate gene prediction as a bipartite ranking problem combining a task-wise ordered observation model with a latent multitask regression function using the matrix-variate Gaussian process (MV-GP). We then use a trace-norm constrained variational inference approach to obtain the bipartite ranking model variables and the parameters of the underlying multitask regression model. We use this model to predict candidate genes from two gene-disease association data sets and show that our model outperforms current state-of-the-art methods. Finally, we demonstrate the practical utility of our method by successfully recovering well characterized gene-disease associations hidden in our training data.**

I. Introduction

Identifying genes that play roles in human diseases is critical for developing new therapeutic approaches and improving care for afflicted patients. However, despite significant progress, the genetic architectures of many human diseases have not been fully elucidated. This is due in part to the large number of human genes and the high costs of experimentally verifying the role of even a single candidate gene in the etiology of a disease. Various computational approaches that produce prioritized lists of candidate genes for further experimental evaluation have been proposed as means for reducing the search space; a fairly comprehensive review of these methods has recently been published [1].

A key challenge in developing computational methods for predicting candidate disease genes is that the observed responses are generally positive gene-disease associations, and the states of the unobserved responses remain unknown; i.e., few, if any, published reports show that a gene is definitely *not* associated with a disease. These types of problems are known as *single class* or *positive unlabeled* (PU) learning tasks [2]. One solution to this problem, taken by algorithms like ProDiGe [3], is to develop a model that maximizes the classification accuracy between the class of known (positive) examples and the class of unknown (or unlabeled) examples [4]. The collaborative filtering literature has also addressed single class problems using low-rank matrix factorization models [5]. Other recent work has approached the single class problem as a bipartite ranking problem and developed a model that ranks positive examples ahead of unknown examples, based on the notion that such a model would also rank unobserved positive associations ahead of unobserved negative associations [2].

Here, we describe a novel method of predicting candidate disease genes that treats the problem as a bipartite ranking task. We develop a model that combines a latent multitask regression function with task-wise ordered observation variables. We employ a non-parametric matrix-variate Gaussian process (MV-GP) prior for the multitask regression and propose a novel trace constrained variational inference approach that imposes useful low rank structure on the multitask regression. Treating the problem as a bipartite ranking task fulfills the scientific need for an accurately ranked list of potential candidate genes for a given disease [3], [6], while low rank structure induces significant correlations among predictions for different diseases, which matches empirical observations that similar diseases often have similar genetic architectures [7], [8].

This paper is organized as follows. We first describe two curated gene-disease data sets used to predict novel gene-disease associations; we also describe a gene-gene interaction network and a disease relationship graph used by our model to capture similarities among genes and diseases, respectively, and to improve the quality of our predictions. We then briefly introduce our generative model and the variational inference approach we use to train it. We use several information retrieval metrics to evaluate our model's performance, and with such measures, our method significantly outperforms ProDiGe [3], which to the best of our knowledge, represents the state-of-the-art in this field. Finally, we discuss some predictions made by our model to show its utility in producing leads for experimental validation and suggest areas where our model might be improved.

II. Materials and Methods

A. Data Sets

We train and evaluate our models using two sets of gene-disease association data curated from the literature. The first, which we call the **OMIM data set**, is based on the Online Mendelian Inheritance in Man (OMIM) database and is representative of the candidate gene prediction task for monogenic or near monogenic diseases, i.e., diseases caused by only one or at most a few genes. The data matrix contains a total of $M = 3{,}210$ diseases, $N = 13{,}614$ genes, and $T = 3{,}636$ known associations (data density of 0.0083%). We note that the extreme sparsity of this data set makes the prediction problem extremely difficult.

The second dataset, which we call the **Medline data set**, is a much larger data set and is representative of predicting

C.H. Lee is with the Department of Biomedical Engineering, and O. Koyejo and J. Ghosh are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, 78712. E-mail: {chlee@, sanmi.k@, ghosh@ece}.utexas.edu

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

candidate genes for both monogenic as well as polygenic diseases, i.e., diseases caused by the interactions of tens or even hundreds of genes. The set of genes in this data set is defined using the NCBI ENTREZ Gene database [9], and the set of diseases is defined using the "Disease" branch of the NIH Medical Subject Headings (MeSH) ontology [10]. We extract co-citations of these genes and diseases from the PubMed/Medline database [11] to identify positive gene-disease associations. This resulting data set contains a total of of M = 4,496 diseases, N = 21,243 genes, and T = 250,190 known associations (data density of 0.26%).

We use information about biological interactions among genes and known relationships among diseases to improve the accuracy of our model, since similar diseases very often have similar genetic causes. We derive our **gene networks** from the HumanNet database [12], a genome-wide functional network of human genes constructed using multiple lines of evidence.For both the OMIM and Medline data sets, our gene-gene interaction network contains a total of 433,224 links. Our **disease network** is derived from the term hierarchy established in the 2011 release of the MeSH ontology. The disease network for the Medline data set contains a total of 13,922 links. However, because we do not have a direct mapping of OMIM diseases to MeSH terms, we do not use a disease network for the OMIM data set.

B. Model

We treat the problem of predicting candidate disease gene as a bipartite ranking task. Ranking refers to the task of learning an ordering for a set of items, in this case, the association between a disease and a set of genes. In the bipartite ranking setting, the items are drawn from two sets, known as the positive set and the negative set, and the task is to learn an ordering that places the positive items ahead of the negative items [13], [14]. Because our observed only contain known (i.e., positive) associations, we follow the method used by ProDiGe [3] and randomly sample the gene-disease association matrices to generate "negative" observations.

The data thus consist of a small set of known gene-disease associations and a large set of unknowns. Let $\mathsf{M} \ni m$ denote the set of diseases, and $\mathsf{N} \ni n$ denote the set of genes. These are collected into a set $\mathsf{T} \ni m, n$. Let $y_{m,n} \in [+1, -1]$ be the label for the m^{th} disease gene and the n^{th} gene pair, and let $\mathsf{T}_m = \{n \,|\, (m,n) \in \mathsf{T}\}$ be the set of labeled genes for the m^{th} disease. The set of known gene links for disease m is given by $\mathsf{D}_m^+ = \{(n \in \mathsf{T}_m \,|\, y_{m,n} = +1\}$, and the set of sampled unknowns is denoted by the set $\mathsf{D}_m^- = \{n \in \mathsf{T}_n \,|\, y_{m,n} = -1\}$. The vector of labels for the m^{th} disease is given by $\boldsymbol{y}_m \in [-1, +1]^{T_m}$.

We solve this ranking problem using task-wise ordered observation variables and a latent multitask regression function with a non-parametric matrix-variate Gaussian process (MV-GP) prior. Shown schematically in Fig. 1, our proposed generative model for $\boldsymbol{y}_m$ is

$$p(\boldsymbol{y}_m|\boldsymbol{r}_m) \propto \prod_{l \in \mathsf{D}_m^+} \prod_{l' \in \mathsf{D}_m^-} \mathbb{I}_{\left[r_{m,l} \geq r_{m,l'}\right]}. \qquad (1)$$

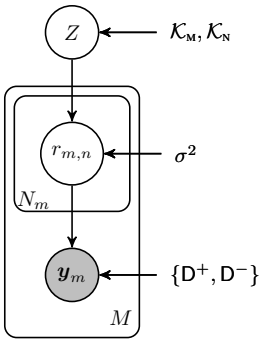

Fig. 1: Plate model of generative bipartite ranking with the latent matrix-variate Gaussian process.

where $\mathbb{I}_{[\cdot]}$ is the indicator function ($\mathbb{I}_{[b]} = 1$ if $b = 1$ and $\mathbb{I}_{[b]} = 0$ otherwise). Note that $p(\boldsymbol{y}_m|\boldsymbol{r}_m)$ is nonzero if and only if $\boldsymbol{r}_m$ satisfies the ordering defined by $\{\mathsf{D}_m^+, \mathsf{D}_m^-\}$. It follows that any vector $\boldsymbol{r}_m$ s.t. $p(\boldsymbol{y}_m|\boldsymbol{r}_m)$ is nonzero also maximizes the AUC performance metric (described in the "*Model evaluation*" section below).

The variables $\boldsymbol{r}_m$ are generated from a Gaussian distribution $r_{m,n} \sim \mathcal{N}\left(z_{m,n}, \sigma^2\right)$ with mean $z_{m,n}$ and variance σ^2. We couple the diseases by jointly generating the latent mean variables from a zero mean matrix-variate Gaussian process $Z \sim \mathcal{GP}\left(0, \mathcal{K}_M, \mathcal{K}_N\right)$ with disease covariance $\mathcal{K}_M$ and gene covariance $\mathcal{K}_N$. These covariances are computed from the disease network and the gene network, respectively [15].

We utilize a novel trace-constrained variational inference to train the bipartite ranking model variables and the underlying multitask regression model. We assume a fully factored representation with $q(\boldsymbol{r}, Z) = \mathbb{I}_{[\boldsymbol{r}=\boldsymbol{r}^*]}q(Z)$, where $\boldsymbol{r} = \{\boldsymbol{r}_m\}$. The variational lower bound of the log likelihood is thus

$$\ln p(\boldsymbol{y}|\mathcal{D}) \geq \ln p(\boldsymbol{y}|\boldsymbol{r}) + \mathbb{E}_{\mathbf{Z}}[\ln p(\boldsymbol{r}, \mathbf{Z})] - \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z})],$$

where $\boldsymbol{y} = \{\boldsymbol{y}_m\}$ and $\boldsymbol{r} = \{\boldsymbol{r}_m\}$. To enforce the low rank constraint, we restrict our search to the space for $q(Z)$ of Gaussian processes $q(Z) = \mathcal{GP}\left(\psi, S\right)$ subject to a trace norm constraint $\|\psi\|_{\mathcal{K},*} \leq C$, where C is a user defined constant. Such a trace norm constraint is known to encourage low rank of matrix valued variables [16], [17].

The combined inference is solved by alternating optimization for $q(Z)$ and $\boldsymbol{r}^*$. The proposed model is trained to estimate bipartite ranking scores for each task and the underlying multitask latent regression distribution. Item rankings are predicted by sorting the expected noise-free scores of the trained model $\mathbb{E}[z_{m,n}|\mathcal{D}] = \psi(m,n)$.

C. Model evaluation

Because of time and monetary constraints, only a small set of the top-ranked predicted genes are viable candidates for experimental validation. Hence, we evaluate our model using the following metrics, which focus on behavior at the top of a ranked list [18]:

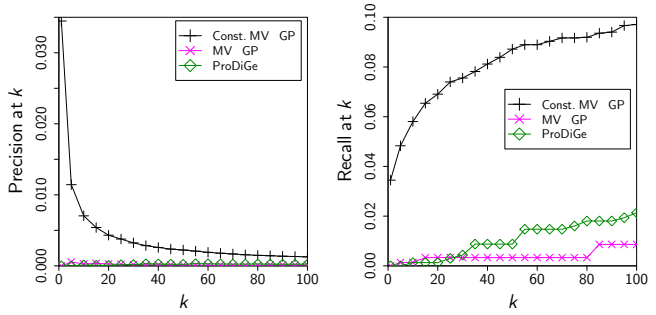

Fig. 2: Average precision (left) and recall (right) for various models using the OMIM data set. "MV-GP" refers to the full-rank model without the trace norm constraint, and "Const. MV-GP" refers to the low-rank model obtained by using the trace norm constraint.

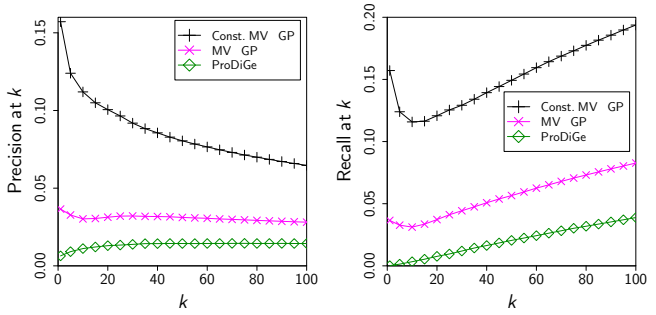

Fig. 3: Average precision (left) and recall (right) for various models using the Medline data set. "MV-GP" refers to the full-rank model without the trace norm constraint, and "Const. MV-GP" refers to the low-rank model obtained by using the trace norm constraint.

TABLE I: AUC and MAP means (std. dev.) for various models using the OMIM data set

	Const. MV-GP	MV-GP	ProDiGe
AUC	0.654 (0.029)	**0.686 (0.016)**	0.524 (0.018)
MAP100	**0.041 (0.002)**	0.001 (0.001)	0.001 (0.000)

TABLE II: AUC and MAP means (std. dev.) for various models using the Medline data set

	Const. MV-GP	MV-GP	ProDiGe
AUC	**0.793 (0.002)**	0.687 (0.002)	0.716 (0.001)
MAP100	**0.042 (0.003)**	0.009 (0.001)	0.003 (0.000)

1) Area under the receiver operating characteristic (ROC) curve (AUC): measures the overall ranking performance of the model.
2) The precision at $k \in \{1, 2, \ldots, 100\}$: measures what fraction of the top k predicted genes are known to be associated with a given disease.
3) The recall at $k \in \{1, 2, \ldots, 100\}$: measures what fraction of the known associated genes are retrieved within the top k predicted genes
4) Mean average precision at $k = 100$ (MAP100): computed as the mean (over all diseases) of the average precision at $k = 100$. The average precision for a single disease is given as:

$$\mathrm{AP}_{@k} = \frac{\sum_{l=1}^{k} \mathbb{I}_{[\bar{\boldsymbol{g}}_l=1]} \mathrm{P}_{@l}}{\min(G_m, k)}$$

In addition to these metrics, we examine cases where well known gene-disease associations were removed from the training set but retained in the testing set to see if our method is able to correctly place such associations at the top of the candidate gene list.

We generate five training sets by randomly sampling a small fraction of the known associations (positives) in each of the OMIM and Medline data sets. The remaining known associations for each set are then used as a testing set for model evaluation. The average of each metric over these training sets is used evaluate model performance.

III. Results

We train two variants of our model: the first a full-rank MV-GP model without the trace norm constraint, and the second a low-rank MV-GP model obtained using the trace norm constraint. We compare the performance of our models to ProDiGe, which ranks candidate disease genes using multitask support vector machines trained with kernels derived from the gene and disease networks. We use ProDiGe as our "gold standard" benchmark because it consistently outperforms other approaches, including distance-based learning methods like Endeavour [19], [20] and label propagation methods like PRINCE [21], and this paper's short format precludes a more extensive comparison of the various methods described in [1].

Results from both the OMIM (Fig. 2 and Table I) and Medline (Fig. 3 and Table II) show that the trace-norm constrained model performs significantly better than both the full-rank MV-GP and ProDiGe, which supports the effectiveness of trace regularization and low-rank approximations for this type of single class problem. The usefulness of the correlations among diseases induced by the disease network and the low-rank structure is evident in the Medline data set results where the training data consisted of an average of fewer than 3 known associations out of a possible 13,614 per disease. The much poorer performance of all models on the OMIM data set is attributable to various factors, including the lack of a disease network, the greater sparsity of the data, and the challenge of trying to predict only a few (usually < 5) positive candidate genes per disease.

Further, as Table III shows, our model is able to correctly identify several extensively studied and experimentally validated gene-disease associations that were missing in various training sets but present in the corresponding testing sets. Moreover, our method consistently ranks these associations highly, which is of practical importance as researchers pursing candidate disease genes for experimental validation are unlikely to look beyond the top several predictions.

TABLE III: Examples of known disease genes in the Medline data set correctly identified in testing by the trace norm constrained MV-GP model

Disease	Gene	Rank	Description
Alzheimer's	APOE	16	Lipoprotein component [22]
Asthma	CD14	10	Immune receptor [23]
High cholesterol	APOE	3	Lipoprotein component [24]
High cholesterol	APOB	14	Lipoprotein component [25]
Prostate cancer	CRP	2	Inflammation marker [26]
Prostate cancer	VEGFA	9	Vascular growth factor [27]

IV. Conclusion

This paper presents the candidate gene prediction as a bipartite ranking problem that can be molded by combining a trace norm constrained matrix-variate Gaussian process (MV-GP) with per-task (i.e., per-disease) ordered observation variables. We showed that the trace norm constraint leads to a low rank model that captures the similar genetic underpinnings of similar diseases. We applied this model to gene-disease association data sets derived from the OMIM and PubMed/Medline databases and demonstrated that our model is a significant improvement over the reasonably strong ProDiGe baseline model. Finally, we highlighted the capability of our model in identifying viable candidate genes for further experimental validation.

We plan to expand this work by exploring other regularization methods for the basic MV-GP model. We also intend to improve on the quality of our candidate gene predictions by refining the data sets and models to handle potentially unreliable reports of gene-disease associations and by incorporating other sources of gene-gene interaction and disease similarity data. Finally, we will also explore approaches that avoid "popularity effects", where by frequently studied genes and genes in network hubs are given unusually high rankings.

Acknowledgements

This work was supported by NSF grant IIS 1016614 and by the Schlumberger Centennial Chair in Engineering (University of Texas at Austin). We thank Sreangsu Acharyya for helpful discussions on bipartite ranking. We also thank U. Martin Blom and Edward Marcotte for helpful discussions and providing the OMIM data set.

References

[1] R. M. Piro and F. Di Cunto, "Computational approaches to disease-gene prediction: rationale, classification and successes," *The FEBS journal*, vol. 279, pp. 678–696, 2012.

[2] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 213–220.

[3] F. Mordelet and J.-P. Vert, "Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples." *BMC Bioinformatics*, vol. 12, p. 389, 2011.

[4] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *UAI*, 2009, pp. 452–461.

[5] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang, "One-class collaborative filtering," in *ICDM*, 2008, pp. 502–511.

[6] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Comput Biol*, vol. 6, no. 1, p. e1000641, 01 2010.

[7] P. G. Sun, L. Gao, and S. Han, "Prediction of human disease-related gene clusters by clustering analysis," *International Journal of Biological Sciences*, vol. 7, no. 1, pp. 61–73, 2011.

[8] N. Natarajan, U. M. Blom, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, "Predicting gene-disease associations using multiple species data," Department of Computer Science, University of Texas at Austin, Tech. Rep. TR-11-37, October 2011.

[9] D. R. Maglott, J. Ostell, K. D. Pruitt, and T. A. Tatusova, "Entrez gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 39, no. Database-Issue, pp. 52–57, 2011.

[10] N. L. of Medicine, "Medical Subject Headings," http://www.nlm.nih.gov/mesh/.

[11] ——, "PubMed," http://www.ncbi.nlm.nih.gov/pubmed/.

[12] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome Research*, vol. 21, no. 7, pp. 1109–1121, May 2011.

[13] C. Cortes and M. Mohri, "Auc optimization vs. error rate minimization," in *Advances in Neural Information Processing Systems*. MIT Press, 2004.

[14] W. Kotlowski, K. Dembczynski, and E. Huellermeier, "Bipartite ranking through minimization of univariate loss," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ser. ICML '11, L. Getoor and T. Scheffer, Eds. New York, NY, USA: ACM, June 2011, pp. 1113–1120.

[15] A. J. Smola and I. Kondor, "Kernels and regularization on graphs," in *Proceedings of the Annual Conference on Computational Learning Theory*, ser. Lecture Notes in Computer Science, B. Schölkopf and M. Warmuth, Eds. Springer, 2003.

[16] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert, "A new approach to collaborative filtering: Operator estimation with spectral regularization," *Journal of Machine Learning Research*, vol. 10, pp. 803–826, 2009.

[17] M. Dudík, Z. Harchaoui, and J. Malick, "Lifted coordinate descent for learning with trace-norm regularization." *AISTATS*, vol. 22, pp. 327–336, 2012.

[18] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. ACM, Jun. 2006, pp. 233–240.

[19] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, "Gene prioritization through genomic data fusion." *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, May 2006.

[20] L.-C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts, and Y. Moreau, "ENDEAVOUR update: a web resource for gene prioritization in multiple species." *Nucleic Acids Research*, vol. 36, no. Web Server issue, pp. W377–84, Jul. 2008.

[21] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Computational Biology*, vol. 6, no. 1, p. e1000641, 2010.

[22] L. Bertram and R. E. Tanzi, "The genetics of Alzheimer's disease," *Progress in Molecular Biology and Translational Science*, vol. 107, pp. 79–100, 2012.

[23] R. Tesse, R. C. Pandey, and M. Kabesch, "Genetic variations in toll-like receptor pathway genes influence asthma and atopy," *Allergy*, vol. 66, pp. 307–316, 2011.

[24] K. Greenow, N. J. Pearce, and D. P. Ramji, "The key role of apolipoprotein E in atherosclerosis," *Journal of Molecular Medicine*, vol. 83, pp. 329–342, 2005.

[25] M. A. Austin, C. M. Hutter, R. L. Zimmern, and S. E. Humphries, "Genetic causes of monogenic heterozygous familial hypercholesterolemia: a HuGE prevalence review," *American Journal of Epidemiology*, vol. 160, pp. 407–420, 2004.

[26] K. Saito and K. Kihara, "C-reactive protein as a biomarker for urological cancers," *Nature Reviews Urology*, vol. 8, pp. 659–666, 2011.

[27] M. T. Schweizer and M. A. Carducci, "From bevacizumab to tasquinimod: angiogenesis as a therapeutic target in prostate cancer," *Cancer Journal*, vol. 19, pp. 99–106, 2013.