Mach Learn (2014) 97:103–127
DOI 10.1007/s10994-014-5444-1

A constrained matrix-variate Gaussian process for transposable data

Oluwasanmi Koyejo · Cheng Lee · Joydeep Ghosh

Received: 5 March 2013 / Accepted: 23 April 2014 / Published online: 10 June 2014

Abstract Transposable data represents interactions among two sets of entities, and are typically represented as a matrix containing the known interaction values. Additional side information may consist of feature vectors specific to entities corresponding to the rows and/or columns of such a matrix. Further information may also be available in the form of interactions or hierarchies among entities along the same mode (axis). We propose a novel approach for modeling transposable data with missing interactions given additional side information. The interactions are modeled as noisy observations from a latent noise free matrix generated from a matrix-variate Gaussian process. The construction of row and column covariances using side information provides a flexible mechanism for specifying a-priori knowledge of the row and column correlations in the data. Further, the use of such a prior combined with the side information enables predictions for new rows and columns not observed in the training data. In this work, we combine the matrix-variate Gaussian process model with low rank constraints. The constrained Gaussian process approach is applied to the prediction of hidden associations between genes and diseases using a small set of observed associations as well as prior covariances induced by gene-gene interaction networks and disease ontologies. The proposed approach is also applied to recommender systems data which involves predicting the item ratings of users using known associations as well as prior covariances induced by social networks. We present experimental results that highlight the performance

Editors: Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný.

O. Koyejo (✉)
Imaging Research Center, University of Texas at Austin, Austin, TX, USA
e-mail: sanmi.k@utexas.edu

C. Lee
Department of Biomedical Engineering, University of Texas at Austin, Austin, TX, USA
e-mail: chlee@utexas.edu

J. Ghosh
Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, USA
e-mail: ghosh@ece.utexas.edu

of constrained matrix-variate Gaussian process as compared to state of the art approaches in each domain.

Keywords Constrained Bayesian inference · Gaussian process · Transposable data · Nuclear norm · Low rank

1 Introduction

Transposable data describes relationships between pairs of entities. Such data can be organized as a matrix, with one set of entities as the rows, the other set of entities as the columns. In such datasets, both the rows and column of the matrix are of interest. Transposable data matrices are often sparse, and of primary interest is the prediction of unobserved matrix entries representing unknown interactions. In the machine learning community, the modeling of transposable data is often encountered as multitask learning (Stegle et al. 2011). In addition to the matrix, transposable datasets often include features describing each row entity and each column entity, or graphs describing relationships between the rows and the columns. These features and graphs can be useful for improving in-matrix prediction performance and for extending model predictions outside of the observed matrix, thus alleviating the *cold-start* problem. In this work, we combine the matrix-variate Gaussian process model with low rank constraints for the predictive modeling of transposable data.

In recent years, the matrix variate Gaussian distribution (MV-G) has emerged as a popular model for transposable data (Allen and Tibshirani 2010, 2012) as it compactly decomposes correlations between the matrix entries into correlations between the rows, and correlations between the columns. Although the MV-G has been shown to be effective for modeling matrix data with missing entries, model predictions do not extend to rows and columns that are unobserved in the training data. One approach to remedy this deficiency is to replace the MV-G with the nonparametric matrix-variate Gaussian process (MV-GP) (Stegle et al. 2011). This is achieved by replacing the row and column covariance matrices of the MV-G with parameterized row and column covariance *functions*. Thus, the resulting model can provide predictions for new rows and columns given features. The MV-GP may also be described as an extension of the scalar valued Gaussian process (GP) (Rasmussen and Williams 2005), a popular model for scalar functions, to vector valued responses. The MV-GP has been applied to link analysis, transfer learning, collaborative prediction and other multitask learning problems (Yu and Chu 2008; Bonilla et al. 2008; Yan et al. 2011). Despite its wide use for transposable data and multitask learning, the MV-GP does not capture low rank structure.

Rank constraints have become ubiquitous in matrix prediction tasks (Yu et al. 2007; Zhu et al. 2009; Koyejo and Ghosh 2011; Zhou et al. 2012; Koyejo and Ghosh 2013a). The low rank assumption implies that matrix-valued parameters of interest can be decomposed as the inner product of low dimensional factors. This reduces the degrees of freedom in the matrix model and can improve the parsimony of the results. Recent theoretical (Candès and Recht 2009) and empirical (Koren et al. 2009) results have provided additional motivation for the low rank approach. The low rank assumption is also motivated by computational concerns. Consider the computational requirements of a full matrix regression model such as a Gaussian process regression (Rasmussen and Williams 2005). Here, the memory requirements scale quadratically with data size, and naïve inference via using a matrix inverse scales cubically with data size (Álvarez et al. 2012). In contrast, training low rank models can scale linearly with the data size and quadratically with the underlying matrix rank (using the factor representation).

Further, efficient optimization methods have been proposed (Koren et al. 2009; Dudik et al. 2012).

We propose a novel constrained Bayesian inference approach that combines the flexibility and extensibility of the matrix-variate Gaussian process with the parsimony and empirical performance of low rank models. Constrained Bayesian inference (Koyejo and Ghosh 2013a) is a principled approach for enforcing expectation constraints on the Bayesian inference procedure. It is a useful approach for probabilistic inference when the problem of interest requires constraints that are difficult to capture using standard prior distributions alone. Examples include linear inequality constraints (Gelfand et al. 1992) and margin constraints (Zhu et al. 2012). To enforce these restrictions, constrained Bayesian inference represents the Bayesian inference procedure as a constrained relative entropy minimization problem. The resulting optimization problem can often be reduced to constrained parameter estimation and solved using standard optimization theoretic techniques.

The main contributions of this paper are as follows:

- We propose a novel approach for capturing the low rank characteristics of transposable data by combining the matrix-variate Gaussian process prior with constrained Bayesian inference subject to nuclear norm constraints.
- We show that (i) the distribution that solves the constrained Bayesian inference problem is a Gaussian process, (ii) its inference can be reduced to the estimation of a finite set of parameters, and (iii) the resulting optimization problem is strongly convex in these parameters.
- We evaluate the proposed model empirically and show that it performs as well as (or better than) the state of the art domain specific models for disease-gene association prediction with gene network and disease ontology side information and recommender systems with social network side information.

We begin by discussing relevant background on the matrix-variate Gaussian process and nuclear norm constraints for matrix-variate functions in Sect. 2. We introduce the concept of constrained inference in Sect. 2.4 and apply it to the matrix-variate Gaussian process to compute a low rank prediction (Sect. 4). We present the empirical performance of the proposed model compared to state of the art domain specific models for transposable data in the disease-gene association domain (Sect. 5.1) and the recommender systems domain (Sect. 5.2). Finally, we conclude in Sect. 6.

2 Background

This section describes the problem statement (Sect. 2.2), and the main building blocks of our approach (i) the matrix-variate Gaussian process (Sect. 2.3) and (ii) constrained Bayesian inference (Sect.2.4).

2.1 Preliminaries

We denote vectors by bold lower case e.g. $\mathbf{x}$ and matrices by bold upper case e.g. $\mathbf{X}$. Let $\mathbf{I}_D$ represent the $D \times D$ identity matrix. Given a matrix $\mathbf{A} \in \mathbb{R}^{P \times Q}$, $\text{vec}(\mathbf{A}) \in \mathbb{R}^{PQ}$ is the vector obtained by concatenating columns of $\mathbf{A}$. Given matrices $\mathbf{A} \in \mathbb{R}^{P \times Q}$ and $\mathbf{B} \in \mathbb{R}^{P' \times Q'}$, the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is denoted as $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{PP' \times QQ'}$. A useful property is the *Kronecker identity*: $\text{vec}(\mathbf{AXB}) = (\mathbf{B}^\top \otimes \mathbf{A})\text{vec}(\mathbf{X})$, where $\mathbf{X} \in \mathbb{R}^{Q \times P'}$ and $\mathbf{B}^\top$ represents the transpose of $\mathbf{B}$.

Let $\mathrm{E}[\cdot]$ be the *expectation operator* with $\mathrm{E}_p[f(z)] = \int_z p(z) f(z) dz$. The *Kullback-Leibler (KL) divergence* between densities $q(z)$ and $p(z)$ is given by:

$$\mathrm{KL}(q(z) \| p(z)) = \mathrm{E}_q \left[\log q(z) - \log p(z) \right].$$

Let $\mathbf{x} \in \mathbb{R}^P$ be drawn from a *multivariate Gaussian* distribution. The density is given as:

$$\mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left[(\mathbf{x}-\mathbf{m})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right]\right)}{(2\pi)^{P/2}|\boldsymbol{\Sigma}|^{P/2}},$$

where $\mathbf{m} \in \mathbb{R}^P$ is the mean vector and $\boldsymbol{\Sigma} \in \mathbb{R}^{P \times P}$ is the covariance matrix. $|\cdot|$ denotes the matrix determinant and $\mathrm{tr}(\cdot)$ denotes the matrix trace.

2.2 Transposable data notation and problem statement

Let $\mathbb{M} \ni m$ be the index set of rows and $\mathbb{N} \ni n$ be the index set of columns. The index set of observed matrix entries is represented by $\mathsf{L} = \{(m, n)\} \subset \mathbb{M} \times \mathbb{N}$ with every $l = (m, n) \in \mathsf{L}$. We define the subset of observed rows as the set $\mathsf{M} = \{m \,|\, (m, n) \in \mathsf{L}\} \subset \mathbb{M}$ with size $|\mathsf{M}| = M$, and the subset of observed columns as the set $\mathsf{N} = \{n \,|\, (m, n) \in \mathsf{L}\} \subset \mathbb{N}$ with size $|\mathsf{N}| = N$ so $L = |\mathsf{L}| \leq M \times N$. Let each entry in the matrix be represented by y_l. The observed subset of the transposable matrix is represented by $\mathbf{y} = \left[y_{l_1} \ \ldots \ y_{l_L}\right]^\top$. Our goal is to estimate a predictive model for any unobserved entries $\{y_{l'} \,|\, l' \notin \mathsf{L}\}$ including entries not observed within the bounds of the training matrix i.e. $\{y_{l'} \,|\, l' \notin \mathsf{M} \times \mathsf{N}\}$.

2.3 Matrix-variate Gaussian process for transposable data

The matrix-variate Gaussian process is a doubly indexed stochastic process $\{Z_{m,n}\}_{m \in \mathbb{M}, n \in \mathbb{N}}$ where finitely indexed samples follow a multivariate Gaussian distribution. As with the scalar Gaussian process (Rasmussen and Williams 2005), the MV-GP is completely specified by its mean and covariance functions. We use the notation $\mathcal{MGP}(\phi, \mathscr{C}_\mathrm{N}, \mathscr{C}_\mathrm{M})$ to denote the MV-GP with mean function $\phi : \mathbb{M} \times \mathbb{N} \mapsto \mathbb{R}$, row covariance function $\mathscr{C}_\mathrm{M} : \mathbb{M} \times \mathbb{M} \mapsto \mathbb{R}$ and the column covariance function $\mathscr{C}_\mathrm{N} : \mathbb{N} \times \mathbb{N} \mapsto \mathbb{R}$. The covariance function of the prior MV-GP has a Kronecker product structure (Álvarez et al. 2012). This form assumes that the prior covariance between matrix entries can be decomposed as the product of the row and column covariances. The joint covariance function of the MV-GP decomposes into product form as $\mathscr{C}\left((m, n), (m', n')\right) = \mathscr{C}_\mathrm{M}(m, m')\mathscr{C}_\mathrm{N}(n, n')$, or equivalently, $\mathscr{C} = \mathscr{C}_\mathrm{N} \otimes \mathscr{C}_\mathrm{M}$. We use the notation $\mathcal{GP}(\psi, \mathscr{C})$ to denote the scalar valued Gaussian process with mean function $\psi : \mathsf{L} \mapsto \mathbb{R}$ and covariance function $\mathscr{C} : \mathsf{L} \times \mathsf{L} \mapsto \mathbb{R}$.

Let $Z \sim \mathcal{MGP}(\phi, \mathscr{C}_\mathrm{M}, \mathscr{C}_\mathrm{N})$, and define the matrix $\mathbf{Z} \in \mathbb{R}^{M \times N}$ with entries $z_{m,n} = Z(m, n)$ for $m, n \in \mathsf{M} \times \mathsf{N}$, $\mathrm{vec}(\mathbf{Z})$ is a distributed as a multivariate Gaussian with mean $\mathrm{vec}(\boldsymbol{\Phi})$ and covariance matrix $\mathbf{C}_\mathrm{N} \otimes \mathbf{C}_\mathrm{M}$, i.e., $\mathrm{vec}(\mathbf{Z}) \sim \mathcal{N}(\mathrm{vec}(\boldsymbol{\Phi}), \mathbf{C}_\mathrm{N} \otimes \mathbf{C}_\mathrm{M})$, where $\phi_{m,n} = \phi(m, n)$, $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ is the mean matrix, $\mathbf{C}_\mathrm{M} \in \mathbb{R}^{M \times M}$ is the row covariance matrix and $\mathbf{C}_\mathrm{N} \in \mathbb{R}^{N \times N}$ is the column covariance matrix. This definition extends to finite subsets $\mathsf{L} \subset \mathbb{M} \times \mathbb{N}$ that are not complete matrices. For any subset L, the vector $\mathbf{z} = \left[z_{l_1} \ \ldots \ z_{l_L}\right]$ is distributed as $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\Phi}_\mathsf{L}, \mathbf{C})$ where the vector $\boldsymbol{\Phi}_\mathsf{L} = [\phi(1) \ldots \phi(L)] \in \mathbb{R}^L$ are arranged from the entries of the mean matrix corresponding to the set $l \in \mathsf{L}$, and $\mathbf{C}$ is the covariance matrix evaluated only on pairs $l, l' \in \mathsf{L} \times \mathsf{L}$.

The MV-GP is a popular prior distribution for transposable matrix data. Here we combine it with a Gaussian observation noise model as follows (see Fig. 1):

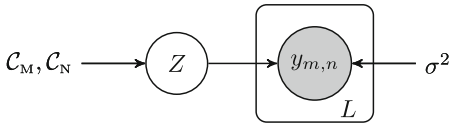


Fig. 1 Plate diagram of the hierarchical matrix-variate Gaussian process model with i.i.d Gaussian observation noise. $Z(m, n)$ is the hidden noise-free matrix entry

1. Draw the function Z from a zero mean MV-GP as $Z \sim \mathcal{MGP}(0, \mathcal{C}_{\mathrm{M}}, \mathcal{C}_{\mathrm{N}})$.
2. Draw observed response independently as $y_{m,n} \sim \mathcal{N}\left(z_{m,n}, \sigma^2\right)$ given $z_{m,n} = Z(m, n)$.

The hidden matrix $\mathbf{Z} \in \mathbb{R}^{M \times N}$ with entries $z_{m,n} = Z(m, n)$ may be interpreted as the latent noise-free matrix. The inference task is to estimate the posterior distribution $Z|\mathcal{D}$, where $\mathcal{D} = \{\mathbf{y}, \mathsf{L}\}$. It follows that the posterior distribution is a Gaussian process (Rasmussen and Williams 2005) given by $Z|\mathcal{D} \sim \mathcal{GP}(\phi, \Sigma)$, with mean and covariance functions:

$$\phi(m, n) = \mathbf{C}_{\mathsf{L}}(m, n)[\mathbf{C} + \sigma^2\mathbf{I}]^{-1}\mathbf{y} \tag{1a}$$

$$\Sigma\left((m, n), (m', n')\right) = \mathcal{C}((m, n), (m', n')) - \mathbf{C}_{\mathsf{L}}(m, n)[\mathbf{C} + \sigma^2\mathbf{I}]^{-1}\mathbf{C}_{\mathsf{L}}(m, n)^{\top}. \tag{1b}$$

The covariance function $\mathbf{C}_{\mathsf{L}}(m, n)$ corresponds to the sampled covariance matrix between the index (m, n) and all training data indexes $(m', n') \in \mathsf{L}$, $\mathbf{C}$ is the covariance matrix between all pairs (m, n), $(m', n') \in \mathsf{L} \times \mathsf{L}$, and $\mathbf{I}$ is the $L \times L$ identity matrix. The closed form follows directly from the definition of a MV-GP as a scalar GP (Rasmussen and Williams 2005) with appropriately vectorized variables. The computational complexity of applying the GP model scales with the number of observed samples L. Storage of the covariance matrix requires $\mathcal{O}(L^2)$ memory, and the naïve inference requires $\mathcal{O}(L^3)$ computation.

2.4 Constrained Bayesian inference

Probabilistic inference involves estimating the distribution of latent variables given new information such as observed data and constraints. This is often achieved via Bayes rule. Given the prior distribution of the latent variables, Bayes rule is a simple formula for computing the latent variable distribution conditioned on the observed data. However, Bayes rule may be inadequate when the constraints one seeks to impose on a latent variable distribution are computationally intractable to enforce by careful selection of the prior distribution alone. An alternative approach is to enforce these constraints as part of the inference procedure. While this can be achieved via rejection sampling and related techniques (Gelfand et al. 1992), such methods are computationally intractable for high dimensional variables as a large proportion of the samples will be discarded. Constrained Bayesian inference via variational optimization is a useful alternative in such cases. Constrained Bayesian inference converts the probabilistic inference into an optimization problem, thus allowing the application of standard optimization techniques.

Let z represent the latent variables and y represent the observations. Bayes rule can be used to compute the posterior density $p(z|y)$ as:

$$p(z|y) = \frac{p(y|z)p(z)}{p(y)}$$

where the conditional density $p(y|z)$ is known as the likelihood, $p(z)$ is the prior density and $p(y)$ is the evidence. An alternative approach was proposed by Zellner (1988), who showed

that the Bayesian posterior can be computed as the solution of the variational optimization problem:

$$p(z|y) = \arg\min_{q \in \mathscr{P}} \mathrm{KL}(q(z) \| p(z)) - \mathrm{E}_q \left[\log p(y|z) \right]. \tag{2}$$

where $\mathscr{P} = \{q \mid \int_z q(z)dz = 1\}$.

Constrained Bayesian inference (Koyejo and Ghosh 2013a) can be used to enforce additional structure on the posterior distribution. It involves enforcing additional constraints on the variational optimization posed in (2). This paper will focus on expectation constraints applied to *feature functions* of the latent variables. Given a vector of feature functions $\boldsymbol{\gamma}(z)$ and a constraint set C, let $\mathscr{R}_{\mathsf{C}} = \{q \in \mathscr{P} \mid \mathrm{E}_q \left[\boldsymbol{\gamma}(z) \right] \in \mathsf{C}\}$ represent the set of densities that satisfy the constraint $\mathrm{E}_q \left[\boldsymbol{\gamma}(z) \right] \in \mathsf{C}$. Constrained Bayesian inference requires solving one of the following equivalent variational optimization problems (Ganchev and Ja 2010; Zhu et al. 2012; Koyejo and Ghosh 2013a):

$$q_*(z) = \arg\min_{q \in \mathscr{R}_{\mathsf{C}}} \mathrm{KL}(q(z) \| p(z)) - \mathrm{E}_q \left[\log p(y|z) \right]. \tag{3a}$$

$$q_*(z) = \arg\min_{q \in \mathscr{R}_{\mathsf{C}}} \mathrm{KL}(q(z) \| p(z|y)). \tag{3b}$$

Thus, the solution is an information projection of the Bayesian posterior distribution onto the constraint set $\mathscr{C}$. Following Zellner, we call q_* the *postdata* density to distinguish it from the unconstrained Bayesian posterior density. Further discussion of constrained Bayesian inference is provided in Appendix 7.

3 Related work

Constrained Bayesian inference is a special case of constrained relative entropy minimization where some of the constraints are generated from observed data (Koyejo and Ghosh 2013). Constrained relative entropy minimization and constrained entropy maximization have been studied in several application domains including natural language processing (Berger et al. 1996) and ecology (Dudík et al. 2007). Applications in the machine learning literature include maximum entropy discrimination (MED) (Jaakkola et al. 1999), and other models inspired by MED have been proposed for combining nonparametric topic models with large margin constraints for document classification (Zhu et al. 2009) and multitask classification (Zhu et al. 2011). Constrained relative entropy models have also been applied to collaborative filtering (Xu et al. 2012) and link prediction (Zhu et al. 2012) Other work using nonparametric priors (Zhu et al. 2009, 2011) has resulted in intractable inference, requiring the application of variational approximations with tractable assumptions made for the independence structure and parametric families of the solution. Our work appears to be the first that uses nonparametric prior distributions without requiring such simplifying assumptions. In addition, we consider constraints on the *function space* of the Gaussian process, which generalize the evaluation based constraints proposed in prior work i.e. constraints on the entire mean function as opposed to constraints on the mean of a set of matrix entries.

Factor models such as principal component analysis (PCA) (Bishop 2006) and its variants are popular methods for extracting information from matrix data. The standard PCA model can be extended to handle missing data using a Bayesian approach (Bishop 2006) that marginalizes over the missing data. The Gaussian process latent variable model (GP-LVM) (Lawrence and Hyvärinen 2005) was proposed to extend PCA to model non-linear relation-

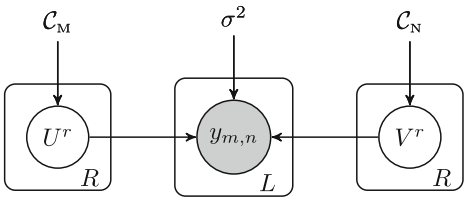


Fig. 2 Hierarchical low rank factor Gaussian process

ships by replacing the covariance matrix with a non-linear kernel. This kernel approach has been applied to non-linear matrix factorization (Lawrence and Urtasun 2009). The GP-LVM integrates out one of the factors and estimates the other. The rank of the factor model must be pre-specified in such models, and is often fixed via expensive cross-validation. Implementations of Kernel PCA typically capture prior correlations over the rows *or* the columns, but not both[1]. Our proposed model is designed capture prior correlations simultaneously over the rows and columns via the matrix-variate Gaussian process prior. Further, the nuclear norm provides an avenue for automatic (implicit) rank selection.

The most common common approach for low rank matrix data modeling in the Gaussian process literature is the hierarchical low rank factor model. In particular, the hierarchical low rank factor Gaussian process (factor GP) has been proposed to capture low rank structure (Yu et al. 2007; Zhu et al. 2009; Zhou et al. 2012). We discuss this approach in some detail as it is used as our main baseline. Here, Gaussian processes are used as the priors for the low dimensional factors. With a fixed model rank R, the generative model for the factor GP is as follows (see Fig. 2):

1. For each $r \in \{1 \dots R\}$, draw row functions: $U^r \sim \mathcal{GP}(0, \mathcal{C}_{\mathrm{M}})$. Let $\mathbf{u}_m \in \mathbb{R}^R$ with entries $u^r_m = U^r(m)$.
2. For each $r \in \{1 \dots R\}$, draw column functions: $V^r \sim \mathcal{GP}(0, \mathcal{C}_{\mathrm{N}})$. Let $\mathbf{v}_n \in \mathbb{R}^R$ with $v^r_n = V^r(n)$.
3. Draw each matrix entry independently: $y_{m,n} \sim \mathcal{N}\left(\mathbf{u}_m^\top \mathbf{v}_n, \sigma^2\right) \ \forall (m,n) \in \mathsf{L}$.

where $\mathbf{u}_m$ is the m^{th} row of $\mathbf{U} = [\mathbf{u}^1 \dots \mathbf{u}^R] \in \mathbb{R}^{M \times R}$, and $\mathbf{v}_n$ is the n^{th} row of $\mathbf{V} = [\mathbf{v}^1 \dots \mathbf{v}^R] \in \mathbb{R}^{N \times R}$. The maximum-a-posteriori (MAP) estimates of $\mathbf{U}$ and $\mathbf{V}$ can be computed as the solution of the following optimization problem:

$$\arg\min_{\mathbf{U},\mathbf{V}} \frac{1}{\sigma^2} \sum_{(m,n)\in L} (y_{m,n} - \mathbf{u}_m^\top \mathbf{v}_n)^2 + \mathrm{tr}\left(\mathbf{U}^\top \mathbf{C}_{\mathrm{M}}^{-1} \mathbf{U}\right) + \mathrm{tr}\left(\mathbf{V}^\top \mathbf{C}_{\mathrm{N}}^{-1} \mathbf{V}\right) \tag{4}$$

where $\mathrm{tr}(\mathbf{X})$ is the trace of the matrix $\mathbf{X}$. Statistically, the factor GP may be interpreted as the sum of rank-one factor matrices. Hence the law of large numbers can be used to show that the distribution of Z converges to $\mathcal{GP}(0, \mathcal{C}_{\mathrm{N}} \otimes \mathcal{C}_{\mathrm{M}})$ as the rank $R \to \infty$ (Yu et al. 2007).

Despite its success, the factor Gaussian process approach has some deficiencies when applied for probabilistic inference. First, posterior distributions of interest are generally intractable. Specifically, neither the joint posterior distribution of $\{\mathbf{U}, \mathbf{V}\}$ nor the distribution of $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ is Gaussian, and their posterior distributions are quite challenging to characterize. As a result, the posterior mean is challenging to compute without sampling and practitioners often apply the MAP approach. Second order statistics such as the posterior covariance are also computationally intractable. Instead, various approximate inference techniques have been applied. A Laplace approximation was proposed by (Yu et al. 2007)

[1] The choice to capture either row or column covariances in PCA and GPLVM is not fundamental to these models i.e. it is primarily a modeling choice.

and (Zhu et al. 2009) utilized sampling techniques. Further, in most cases, the rank must be fixed a-priori. More recently, Bayesian models for matrix factorization that include a nonparametric prior for the number of latent factors have been proposed based on the Indian buffet process (Zhu 2012; Xu et al. 2012) and multiplicative gamma process (Zhang and Carin 2012). Inference with these models is generally intractable, and requires approximations or sampling, which may result in slow or inaccurate inference for large datasets. Further, many of these approaches have focused on in-matrix prediction, and have not been applied to out-of-matrix predictions.

Other related literature include Li and Yeung (2009), where the authors proposed a regularized matrix factorization model exploiting relation information. The proposed model is identical to the Gaussian process factor model[2] (Zhou et al. 2012) with an appropriate choice of kernel. Li et al. (2009b) proposed an approach for learning a kernel based on network links that can then be applied to predictive modeling tasks. Li et al. (2009a) proposed a Bayesian probabilistic PCA model for full matrix prediction exploiting relational data information by constructing a covariance matrix that accounted for the relational data. An alternative approach focusing on learning additive Gaussian process kernels was proposed by (Xu et al. 2009), and an approach for nonparametric relational data modeling using co-clustering (instead of matrix factorization) was proposed by Xu et al. (2006). Several works have focused on the matrix prediction task alone without the use of side information. For example, Sutskever et al. (2009) utilized the clustering of factors to model the latent relationships as an alternative to designing covariance matrices.

4 Proposed approach: the nuclear norm constrained MV-GP

We propose nuclear norm constrained Bayesian inference for modeling low rank transposable data as an alternative to the low rank factor approach. The proposed approach constrains the model by directly regularizing the rank of the expected prediction via a constraint on its nuclear norm. Optimization with the rank constraint is computationally intractable, and the popular factor representation results in a nonconvex optimization problem that is susceptible to local minima (Dudik et al. 2012). The nuclear norm constraint has been proposed as a tractable surrogate regularization for the low rank constraint, which is in turn motivated by parsimony of the low rank representation, and the superior empirical performance of low rank models in many application domains. The nuclear norm of a matrix variate function is given by the sum of its singular values (Abernethy et al. 2009), and is the tightest convex hull of its rank. Under certain conditions, it can be shown that nuclear norm regularization recovers the true low rank matrix (Pong et al. 2010). Further details on the nuclear norm of matrix functions are provided in Appendix 8.

With no loss of generality, we assume a set of rows M and a set of columns N of interest so $\mathsf{L} \subset \mathsf{M} \times \mathsf{N}$. Let $\mathbf{Z} \in \mathbb{R}^{M \times N}$ be the matrix of hidden variables, with $\mathbf{z} = \mathrm{vec}(\mathbf{Z}) \in \mathbb{R}^{M \times N}$. Given any finite index set of observations at indices $l \in \mathsf{L}$, the finite dimensional prior distribution is a Gaussian distribution given by $\mathcal{N}(0, \mathbf{C})$ where $\mathbf{C} \in \mathbb{R}^{MN \times MN}$. We seek a postdata density $q(Z|\mathscr{D})$ that optimizes (3a) subject to the constraint $\left|\left|\left| \mathrm{E}_q [\, Z \,] \right|\right|\right|_1 \leq \eta$ where $|||\cdot|||_1$ is the nuclear norm. For any finite index set, the unconstrained Bayesian posterior distribution is Gaussian (Sect. 2.3). Following the steps of Sect. 2.4 (see also Appendix 7), it is straightforward to show that since the feature function $\gamma(Z) = Z$ is linear, the constrained Bayes solution must also take a Gaussian form. All that remains is to solve for the mean and

[2] See experiments (Sect. 5) for further discussion.

covariance. We may apply either the prior form (3a) or the equivalent posterior form (3b) for constrained inference. We discuss both approaches for illustrative purposes.

Let the Bayesian posterior be given by $\mathcal{N}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ as described in (1) where $\boldsymbol{\phi} = \text{vec}(\boldsymbol{\Phi}) \in \mathbb{R}^{M\times N}$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{MN\times MN}$. Let the postdata distribution be given by $\mathcal{N}(\boldsymbol{\psi}, \mathbf{S})$, where $\boldsymbol{\psi} = \text{vec}(\boldsymbol{\Psi}) \in \mathbb{R}^{M\times N}$, and $\mathbf{S} \in \mathbb{R}^{MN\times MN}$. Using the posterior form (3b), the postdata distribution is found by minimizing the KL divergence between the Gaussian distribution $\mathcal{N}(\boldsymbol{\psi}, \mathbf{S})$ and the Bayesian posterior distribution $\mathcal{N}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$. This is given by:

$$\min_{\boldsymbol{\psi},\mathbf{S}} \text{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) + (\boldsymbol{\phi}-\boldsymbol{\psi})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}-\boldsymbol{\psi}) - \log|\mathbf{S}| + \log|\boldsymbol{\Sigma}| \;\text{ s.t. } \left|\left|\left|\text{E}_q[Z]\right|\right|\right|_1 \le \eta$$

where $\boldsymbol{\psi} = \text{vec}(\boldsymbol{\Psi})$. The optimization decouples between the mean term $\boldsymbol{\psi}$ and the covariance term $\mathbf{S}$ as:

$$\min_{\boldsymbol{\psi}} (\boldsymbol{\phi}-\boldsymbol{\psi})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}-\boldsymbol{\psi}) \;\text{ s.t. } \left|\left|\left|\text{E}_q[Z]\right|\right|\right|_1 \le \eta \tag{5a}$$

$$\min_{\mathbf{S}} \text{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) - \log|\mathbf{S}| + \log|\boldsymbol{\Sigma}| \tag{5b}$$

The minimum in terms of the covariance is achieved for $\mathbf{S} = \boldsymbol{\Sigma}$ and the mean optimization is given by the solution of a constrained quadratic optimization.

Direct optimization of (5a) requires the computation, storage and inversion of the covariance matrix $\boldsymbol{\Sigma}$. This may become computationally infeasible for high dimensional data. In such situations, estimation of the postdata mean using the prior form (3a) is a more computationally feasible approach. The result is the optimization problem:

$$\mathfrak{L}(\boldsymbol{\Psi}, \mathbf{S}) = \min_{\boldsymbol{\psi},\mathbf{S}}\Big[\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z})] - \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{y},\mathbf{Z})] \text{ s.t. } \left|\left|\left|\text{E}_q[Z]\right|\right|\right|_1 \le \eta\Big]. \tag{6}$$

Let $\mathbf{P} \in \mathbb{R}^{L\times MN}$ be a selection matrix such that $\mathbf{S}_L = \mathbf{PSP}^\top$ is the postdata covariance matrix of the subset of observed entries $l \in \mathsf{L}$, and $\mathbf{C}_L = \mathbf{PCP}^\top$ is the prior covariance of the corresponding subset of entries. Evaluating expectations, the cost function (6) results in the following inference cost function (omitting terms independent of $\boldsymbol{\psi}$ and $\mathbf{S}$):

$$\mathfrak{L}(\boldsymbol{\Psi}, \mathbf{S}) = \min_{\{\boldsymbol{\psi} \mid \left|\left|\left|\text{E}_q[Z]\right|\right|\right|_1 \le \eta\},\mathbf{S}} \begin{bmatrix} \frac{1}{2\sigma^2}\sum_{m,n\in\mathsf{L}}(y_{m,n}-\psi_{m,n})^2 + \frac{1}{2}\boldsymbol{\psi}^\top\mathbf{C}^{-1}\boldsymbol{\psi} \\ -\ln|\mathbf{S}| + \frac{1}{2\sigma^2}\text{tr}(\mathbf{S}_L) + \frac{1}{2}\text{tr}\left(\mathbf{C}^{-1}\mathbf{S}\right) \end{bmatrix}.$$

First, we compute gradients with respect to $\mathbf{S}$. After setting the gradients to zero, we compute:

$$\mathbf{S}_* = \left(\mathbf{C}^{-1} + \frac{1}{\sigma^2}\mathbf{P}^\top\mathbf{P}\right)^{-1} = \mathbf{C} - \mathbf{CP}^\top\left(\mathbf{C}_L + \frac{1}{\sigma^2}\mathbf{I}_L\right)^{-1}\mathbf{PC} \tag{7}$$

The second equality is a consequence of the matrix inversion lemma. We note that this is the exact same result as was found by using the posterior approach (5b). Next, collecting the terms involving the mean results in the optimization problem:

$$\boldsymbol{\psi}_* = \arg\min_{\boldsymbol{\psi}} \frac{1}{2\sigma^2}\sum_{m,n\in\mathsf{L}}(y_{m,n}-\psi_{m,n})^2 + \frac{1}{2}\boldsymbol{\psi}^\top\mathbf{C}^{-1}\boldsymbol{\psi} \;\text{ s.t. } \left|\left|\left|\text{E}_q[Z]\right|\right|\right|_1 \le \eta \tag{8}$$

This is a convex regularized least squares problem with a convex constraint set. Hence, (8) is convex, and $\boldsymbol{\psi}_*$ is unique. Using the Kronecker identity, we can re-write the cost function in

parameter matrix form. We can also replace the nuclear norm constraint with the equivalent regularizer weighed by λ. This leads to the equivalent optimization problem:

$$\boldsymbol{\Psi}_* = \arg\min_{\boldsymbol{\Psi}} \frac{1}{2\sigma^2} \sum_{m,n \in \mathsf{L}} (y_{m,n} - \psi_{m,n})^2 + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Psi}^\top \mathbf{C}_\mathrm{M}^{-1} \boldsymbol{\Psi} \mathbf{C}_\mathrm{N}^{-1}\right) + \lambda \left|\left|\left| \mathrm{E}_q\left[Z \right] \right|\right|\right|_1. \quad (9)$$

The final step is to define the term $\left|\left|\left| \mathrm{E}_q\left[Z \right] \right|\right|\right|_1$. We note that since the prior distribution is a Gaussian process, a valid postdata distribution must extend to arbitrary index sets. Hence the postdata mean is a matrix-variate function. The parametric representation of the postdata mean can be defined using the posterior distribution of the Gaussian process outlined by Csató (2002) and applying the representation theorem (18). Thus, we recover the parametric form of the mean function as $\boldsymbol{\Psi} = \mathbf{C}_\mathrm{M} \boldsymbol{A} \mathbf{C}_\mathrm{N}$ where $\boldsymbol{A} \in \mathbb{R}^{M \times N}$. We may now solve for $\boldsymbol{A}$ directly:

$$\boldsymbol{A}_* = \arg\min_{\boldsymbol{A}} \frac{1}{2\sigma^2} \sum_{m,n \in \mathsf{L}} \left(y_{m,n} - (\mathbf{C}_\mathrm{M} \boldsymbol{A} \mathbf{C}_\mathrm{N})_{m,n}\right)^2 + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}^\top \mathbf{C}_\mathrm{M} \boldsymbol{A} \mathbf{C}_\mathrm{N}\right) + \lambda |||\psi_{\boldsymbol{A}}|||_{1\text{-}\mathscr{H}_\mathscr{C}}. \quad (10)$$

where $\psi_{\boldsymbol{A}}$ is the mean function corresponding to the parameter $\boldsymbol{A}$ (see (18)), and $|||\cdot|||_{1\text{-}\mathscr{H}_\mathscr{C}}$ represents the nuclear norm in the Hilbert space $\mathscr{H}_\mathscr{C}$ (defined in Appendix 8). We also note that the optimization problem (10) is strongly convex.

We now seek to extend the solution from the finite observed index set to the nonparametric domain. Our approach will rely on Kolmogorov's Extension theorem (Bauer 1996) which provides a mechanism for describing infinite dimensional random processes via their finite dimensional marginals (Orbanz and Teh 2010). We will apply the theorem to extend the solution estimated by (6) using a finite index to a corresponding nonparametric Gaussian process. This will be achieved by showing that the solution can be extended to an arbitrary index set with a consistent functional form for the mean and the covariance.

Theorem 1 *The postdata distribution* $\mathscr{N}(\boldsymbol{\psi}, \mathbf{S})$ *is a finite dimensional representation of the Gaussian process* $\mathscr{GP}(\psi, S)$ *sampled at indices* L *where the mean function* ψ *is given by* (10) *and the covariance function* S *is given by* (1b)*.*

Sketch of proof: The requirements of Kolmogorov's extension theorem can be reduced to a proof that for a fixed training set $\mathscr{D}$, the postdata distribution of the superset $(\mathsf{M} \times \mathsf{N}) \cup (m', n')$ has a consistent function representation[3]. The mean and covariance of the postdata density are decoupled in the optimization and the postdata covariance function can be computed in closed form. Thus, for the covariance, this follows trivially from the functional form of (1b). The functional form of the mean follows from the finite representation (18) that solves the optimization problem (10). Note that the solution does not change with the addition of indices $l' = (m', n') \notin \mathsf{L}$ without corresponding observations $y_{l'}$. Uniqueness of the solution follows from the strong convexity of (10). We refer the reader to the dissertation (Koyejo 2013) for further details. □

4.1 Alternative representation of the nuclear norm constrained inference

The mean function optimization (10) may also be represented in terms of matrix parameters that are amenable to direct optimization. With the index set fixed, compute a basis. $\mathbf{G}_\mathrm{M} \in$

[3] See (Rasmussen and Williams 2005, Section 2.2) for an analogous proof applied to Gaussian process regression.

$\mathbb{R}^{M\times D_M}$ and $\mathbf{G}_N \in \mathbb{R}^{N\times D_N}$ such that $\mathbf{C}_M = \mathbf{G}_M\mathbf{G}_M^\top$ and $\mathbf{C}_N = \mathbf{G}_N\mathbf{G}_N^\top$. The mean function can be re-parameterized as $\psi(m,n) = \mathbf{G}_M(m)\mathbf{B}\mathbf{G}_N(n)^\top$, where $\mathbf{B} \in \mathbb{R}^{D_M\times D_N}$. The nuclear norm of ψ can now be computed directly as the nuclear norm of the parameter matrix (Abernethy et al. 2009, Theorem 3). The resulting optimization problem is:

$$\mathbf{B}_* = \arg\min_{\mathbf{B}} \frac{1}{2\sigma^2}\sum_{m,n\in\mathsf{L}}\left(y_{m,n} - (\mathbf{G}_M\mathbf{B}\mathbf{G}_N^\top)_{m,n}\right)^2 + \frac{1}{2}|||\mathbf{B}|||_2^2 + \lambda|||\mathbf{B}|||_1. \tag{11}$$

where $\mathbf{B}$ is the estimated parameter matrix, and $|||\cdot|||_2^2$ and $|||\cdot|||_1$ represent the matrix squared Frobenius norm and the matrix nuclear norm respectively. In this form, the mean function can be estimated directly using standard solvers for large scale nuclear norm constrained optimization (e.g. Dudik et al. 2012; Laue 2012).

To improve scalability, large scale nuclear norm regularized solvers generally represent the parameter matrix in low rank form, avoiding storage of the full matrix. Further, the rank of the parameter matrix is automatically estimated during the optimization. We provide a short summary of the approaches in Dudik et al. (2012) and Laue (2012). Interested readers are referred to the relevant papers for further details. The parameter matrix can be estimated starting from a rank one solution, then the rank is increased until additional factors do not improve the cost any further. The first step consists of determining a good descent direction, and the second step consists of optimizing the factors given the initial direction. In the first step, a descent direction is determined by computing the singular vectors associated with the maximum singular value of the sparse gradient matrix.This step does not need to be accurate and is usually achieved using a few iterations of the power method. The factor optimization in the second step is analogous to the standard matrix factorization optimization, so the large scale nuclear norm solvers mainly differ from standard matrix factorization in the determination of an initial descent direction (matrix factorization is generally randomly initialized), and in the automatic determination of the number of required factors i.e. the rank. Thus the computational requirements of large scale nuclear norm regularized regression are comparable to standard matrix factorization methods.

5 Experiments

We completed experiments with transposable datasets from the disease-gene association domain and the recommender system domain. **Prior covariances:** All the datasets studied consist of transposable data matrices with corresponding row and/or column graphs. We experimented with the identity prior covariance $\mathbf{C} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, and the diffusion prior covariance (Smola and Kondor 2003) given as $\mathbf{C} = \exp(-a\mathbf{L}) + b\mathbf{I}$, where $\mathbf{L}$ is the normalized graph Laplacian matrix. Let $\mathbf{A}$ be the adjacency matrix for the graph and $\mathbf{D}$ be a diagonal matrix with entries $\mathbf{D}_{i,i} = (\mathbf{A}\mathbf{1})_i$. The normalized Laplacian matrix is computed as $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$. We set the $a = b = 1$. No further optimization was performed, and more detailed experimental validation of covariance parameter selection is left for future work.

Models: We present results for the proposed constrained MV-GP approach (**Con. MV-GP**), and the special cases using only the nuclear norm (**Trace GP**)[4] and using only the Hilbert norm (**MV-GP**) i.e. the standard MV-GP regression. To the best of our knowledge, the special case of Trace GP is a novel contribution. As baselines, we implemented kernelized

[4] The nuclear norm is also known as the trace norm.

probabilistic matrix factorization (**KPMF**) (Zhou et al. 2012) and probabilistic matrix factorization (**PMF**) (Mnih and Salakhutdinov 2007) using rank 5 and rank 20 factors. PMF is identical to KPMF using an identity covariance. KPMF has been shown to outperform PMF and and other baseline models in various domains. We note that the rank constraint ensures that all of the proposed models except for MV-GP can be used for in-matrix predictions even with the identity prior covariance. Out-of-matrix predictions require the use of other covariance matrices.

We implemented Con. MV-GP using the representation outlined in Sect. 4.1. The Cholesky decomposition of the covariance matrices was used as the basis representation. The model hyperparameter λ was selected using 5 values logarithmically spaced between 10^{-3} and 10^{3} and the noise hyperparameter was selected σ^2 using 20 values logarithmically spaced between 10^{-3} and 10^{3} for all the models. We experimented with learning the data noise variance term σ^2, but found the results worse than using parameter selection. In particular, the estimated noise variance often approached zero - indicating overfitting. A possible solution we plan to explore is to introduce a prior distribution for σ^2 (see e.g. Bayesian linear regression in Bishop (2006, Chapter 3.3) that may help to regularize the noise term away from zero.

The standard MV-GP is often implemented as a scalar GP with the row and column prior covariance matrices multiplied as shown in (1). We found this "direct" approach computationally intractable as the memory requirements scale quadratically with the size of the observed transposable data matrix. Instead, we implemented the MV-GP in matrix form as a special case of (11) with $\lambda = 0$. This allowed us to scale the model to the larger datasets at the expense of more computation. The nuclear norm regularized optimization in (11) was solved using the large scale approach of Laue (2012). All numerical optimization was implemented using the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm.

Experiment design and cross validation: We performed two kinds of experiments. In the rest of this discussion, "rows" will refer to either the disease (disease-gene prediction) or the user (recommender system). The *known rows* experiment was designed to evaluate the performance of the model for entries selected randomly over the observed values in the matrix. In contrast, the *new rows* experiment was designed to evaluate the generalization ability of the model for new rows not observed in the training set. We partitioned each dataset into five-fold crossvalidation sets. The model was trained on 4 of the 5 sets and tested on the held out set. The results presented are the averaged fivefold cross validation performance. For the "known row" experiments, the cross validation sets were randomly selected over the matrix. For the *new row* experiments, the cross validation was performed row-wise, i.e., we selected training set row and test set rows. Note that the identity prior covariance cannot be used for new row prediction, but due to the low rank constraint, it can be used for known row prediction.

5.1 Disease-gene prediction

Genes are segments of DNA that determine specific characteristics; over 20,000 genes have been identified in humans, which interact to regulate various functions in the body. Researchers have identified thousands of diseases, including various cancers and respiratory diseases such as asthma (NCBI 1998), caused by mutations in these genes. Genetic association studies (McCarthy et al. 2008) are the standard approach for discovering disease-causing genes. However, these studies are often tedious and expensive to conduct. Hence, computational methods that can reduce the search space by predicting the list of candidate genes associated with a given disease are of significant scientific interest. The disease gene prediction task has been the subject of a significant amount of study in recent years (Vanunu

et al. 2010; Li and Toh 2010; Mordelet and Vert 2011; Singh-Blom et al. 2013). The task is challenging because all the observed responses correspond to known associations, and there are no reliable negative examples. Disease gene association shares the binary matrix representation of the *one class* (also known as implicit feedback) matrix prediction studied in the collaborative filtering literature (Pan et al. 2008; Hu et al. 2008).

Additional baseline: In addition to the matrix factorization baseline models, we compared our proposed approach to ProDiGe (Mordelet and Vert 2011); a start-of-the-art approach that has been shown to be superior to previous top-performing approaches, including distance-based learning methods like Endeavour (Aerts et al. 2006) and label propagation methods like PRINCE (Vanunu et al. 2010). ProDiGe estimates the prioritization function using a multitask support vector machine (SVM) trained with the gene prior covariance and disease prior covariance as kernels. Of the models implemented, ProDiGe is most similar to the MV-GP. In fact, MV-GP and ProDiGe mainly differ in their loss functions (squared loss and hinge loss respectively). The SVM regularization parameter for ProDiGe was selected from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. We note that also PMF represents the matrix factorization baseline often applied to similar implicit feedback datasets in the recommendation system literature (Pan et al. 2008).

Sampling "negative" entries: Following Mordelet and Vert (2011), we sampled the unknown entries as "negative" observations randomly over the disease-gene association matrix. We sampled 5 different negatively labeled item sets. All models were trained with the positive set combined with one of the negative labeled sets. The model scores were computed by averaging the scores over the 5 trained models. All models were trained using the same samples.

Metrics: Experimental validation of disease-gene associations in a laboratory can be time consuming and costly, so only a small set of the top ranked predictions are of practical interest. Hence, we focus on metrics that capture the ranking behavior of the model at the top of the ranked list. All the ranking metrics were computed on the test set after removing *all* genes that had been observed in the training set. We computed precision ($\mathrm{P}_{@k}$) and recall ($\mathrm{R}_{@k}$) where $k = 1, \ldots, 20$. Let $\mathbf{g}_l$ denote the labels of gene l as sorted by the predicted scores of the trained regression model, and let $G_m = \sum_l \mathbf{1}_{[\mathbf{g}_l=1]}$ be the total number of relevant genes for disease m in the test data after removing relevant genes observed in the training data. The precision at k computes the fraction of relevant genes retrieved out off all retrieved genes at position k. The recall at k computes the fraction of relevant genes retrieved out of all relevant genes that can be retrieved with a list of length k. These are computed as:

$$\mathrm{P}_{@k} = \frac{\sum_{l=1}^{k} \mathbf{1}_{[\mathbf{g}_l=1]}}{k}, \quad \mathrm{R}_{@k} = \frac{\sum_{l=1}^{k} \mathbf{1}_{[\mathbf{g}_l=1]}}{G_m}.$$

All metrics were computed per disease and then averaged over all the diseases in the test set. Model selection was computed separately per metric. Higher values reflect better performance for the $\mathrm{P}_{@k}$ and $\mathrm{R}_{@k}$ metrics and their maximum value is 1.0.

Datasets: We trained and evaluated 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

Table 1 OMIM disease-gene dataset

Model	$P_{@20}$	$R_{@20}$
MV-GP (D)	0.000 (0.000)	0.003 (0.002)
Con. MV-GP (D)	0.008 (0.001)	0.146 (0.031)
Con. MV-GP (I)	**0.010 (0.001)**	**0.175 (0.025)**
Trace GP (D)	0.006 (0.001)	0.117 (0.021)
Trace GP (I)	0.009 (0.001)	0.157 (0.023)
KPMF-5 (D)	**0.010 (0.001)**	0.167 (0.028)
PMF-5 (I)	0.002 (0.000)	0.034 (0.004)
KPMF-20 (D)	0.009 (0.002)	0.161 (0.040)
PMF-20 (I)	0.002 (0.000)	0.039 (0.008)
ProDiGe	0.000 (0.000)	0.001 (0.003)

Avg. (std.) $P_{@20}$ and $R_{@20}$ performance. (I) Identity prior covariance (D) Diffusion prior covariance. Low precision values in OMIM data are due to the high class imbalance of the test data (average of 1.2 genes per disease) Bold values indicate best performance

and is representative of predicting 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 (Maglott et al. 2011), and the set of diseases is defined using the "Disease" branch of the NIH Medical Subject Headings (MeSH) ontology (National Library of Medicine 2012). We extracted co-citations of these genes and diseases from the PubMed/Medline database (National Library of Medicine 2012) 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 0.36 %).

Information about biological interactions among genes and known relationships among diseases were used to improve the accuracy of our model, since similar diseases very often have similar genetic causes. We derive **gene networks** from the HumanNet database (Lee et al. 2011), a genome-wide functional network of human genes constructed using multiple lines of evidence, including gene co-expression, protein-protein interaction data, and networks from other species. 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. As a result, we are unable to test our model's ability to produce predictions for "new" diseases, i.e., diseases with no associated genes in the training set.

The OMIM dataset contains an average of 1.2 test genes (positive items) per disease, and the model is required to rank more than $13,000$ genes per disease. Hence, the gene prediction task is particularly challenging. This difficulty is reflected in the low precision values observed in Table 1 and Fig. 3a. Despite this extreme sparsity, we found that the proposed approaches (Con. MV-GP and Trace GP) performed as well or better than the matrix factorization baselines (KMPF, PMF), and significantly outperformed the domain specific baseline (ProdiGe). In fact, both full rank models (MV-GP and ProDiGe) performed poorly, suggesting the importance of the low rank / nuclear norm constraint. The results in Fig. 3a, b further highlight the performance of the proposed models at the very top of the list.

We were unable to run ProDiGe on the Medline dataset due computational issues. In particular, the implementation of ProDiGe requires the full kernel matrix as an input. The memory required to store the full kernel is quadratic in the transposable data size. We did not pursue an alternative implementation with reduced memory requirements as experiments

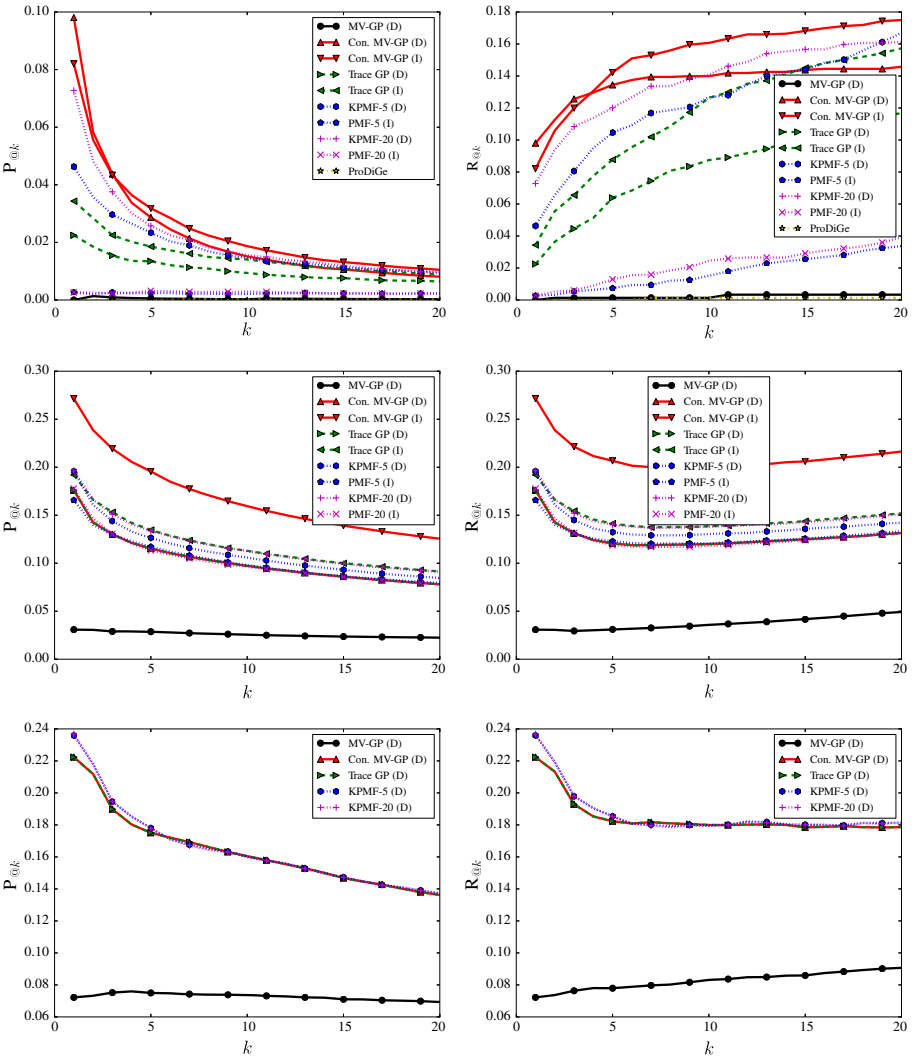


Fig. 3 Disease-gene prediction. Precision (*left*) and Recall (*right*) @$k = 1, 2, \ldots, 20$. (I): Identity prior covariance, (D): Diffusion prior covariance. Low precision values in OMIM are due to the high class imbalance in the test data (avg of 1.2 genes per disease). The identity prior covariance does not generalize to new diseases. Constrained MV-GP out-performs ProDiGe (domain specific baseline), KPMF and PMF. ProDiGe was unable to scale to the full curated dataset (see text) **a** OMIM precision@k **b** OMIM recall@k **c** Medline (known diseases) precision@k **d** Medline (known diseases) recall@k **e** Medline (new diseases) precision@k **f** Medline (new diseases) recall@k

with OMIM and initial experiments with subsampled data indicated inferior performance. The Medline dataset contained an average of 59.2 positive items per disease. Correspondingly, the tested models achieved a higher precision than in the OMIM dataset. Our experimental results (Table 2) show that the proposed models (Con. MV-GP, Trace GP) significantly outperformed the matrix factorization baselines (PMF, KMPF) on the known diseases, and performed as least as well as KMPF on the new diseases. The results in Fig. 3c, d show that the proposed

Table 2 Medline disease-gene dataset

Model	Known diseases		New Diseases	
	$P_{@20}$	$R_{@20}$	$P_{@20}$	$R_{@20}$
MV-GP (D)	0.022 (0.000)	0.049 (0.002)	0.069 (0.020)	0.091 (0.022)
Con. MV-GP (D)	0.078 (0.001)	0.131 (0.004)	**0.137 (0.029)**	**0.181 (0.026)**
Con. MV-GP (I)	**0.126 (0.001)**	**0.216 (0.002)**	–	–
Trace GP (D)	0.078 (0.001)	0.131 (0.004)	**0.137 (0.029)**	**0.181 (0.026)**
Trace GP (I)	0.091 (0.001)	0.152 (0.004)	–	–
KPMF-5 (D)	0.085 (0.001)	0.142 (0.004)	0.136 (0.032)	0.179 (0.032)
PMF-5 (I)	0.079 (0.002)	0.133 (0.003)	–	–
KPMF-20 (D)	0.091 (0.001)	0.151 (0.004)	0.136 (0.032)	0.179 (0.032)
PMF-20 (I)	0.078 (0.001)	0.131 (0.002)	–	–

Avg. (std.) $P_{@20}$ and $R_{@20}$ performance. (I): Identity prior covariance (D): Diffusion prior covariance. The dataset contains an average of 59.2 test genes per disease. The identity prior covariance does not generalize to new diseases. ProDiGe was unable to scale to the full curated dataset
Bold values indicate best performance

models outperform the baselines for known diseases prediction at all levels of precision and recall we measured. The results for new disease prediction in Fig. 3e, f show similar performance for both approaches on the new diseases.

In summary, the presented results suggest that the low rank constraint is useful for describing the structure of disease-gene association. We also found that in all the datasets, the constrained Bayesian models (Con. MV-GP and Trace GP) performed the same or better than the Bayesian factor models (KPMF and PMF) and the unconstrained Bayesian model (MV-GP). This shows the utility of the constrained Bayesian inference approach as compared to the Bayesian factor model approach. Constrained MV-GP with the identity kernel was the best single performing method, matching results in the literature suggesting that the network information is not always helpful for in-matrix predictions (Koyejo and Ghosh 2011; Zhou et al. 2012), though it remains essential for generalization beyond the training matrix. Future work will include further examination of these issues.

5.2 Recommender systems

The goal of a recommender system is to suggest items to users based on past feedback and other user and item information. Recommender systems may also be used for targeted advertising and other personalized services. The low rank matrix factorization approach has proven to be a popular and effective model for the recommender systems data (Mnih and Salakhutdinov 2007; Koren et al. 2009; Koyejo and Ghosh 2011). Several authors (Yu et al. 2007; Zhu et al. 2009; Zhou et al. 2012) have studied the factor GP approach for recommender systems, and have shown that prior covariances extracted from the social network can improve the prediction accuracy and may be used to provide predictions with no training ratings (Koyejo and Ghosh 2011; Zhou et al. 2012). Kernelized probabilistic matrix factorization (KPMF) is of particular interest, as it has been shown to outperform PMF (Mnih and Salakhutdinov 2007) and SoRec (Ma et al. 2008), strong baseline methods for predicting user item preferences with social network side information.

Metrics The model performance was measured using a combination of regression and ranking metrics. Recommender systems are typically most concerned with presenting the few

items that the user is very likely to be interested in, and accurately predicting the score of the other items is less important. Several authors (Steck 2010; Steck and Zemel 2010) have shown that measuring the recall ($R_{@k}$) of the top relevant items compared to all available items can provide an unbiased estimate of the predicted ranking. As suggested by (Steck 2010) we measure the ability of the model to predict relevant items (ratings greater than 4) ahead of other entries (both missing and observed entries with rating less than or equal to 4) using recall at 20 ($R_{@20}$). Recall per user was computed on the test set after removing *all* items that had been observed in the training set, and averaged over all users. For regression, we used the root mean square error (RMSE) metric (Koren et al. 2009) given by $\sqrt{\frac{1}{L}\sum_{l=1}^{L}(y_l - \hat{y}_l)^2}$ where $\hat{y}_l$ is the prediction for index l. Lower values reflect better performance for the RMSE.

Datasets: We trained and evaluated our models using two publicly available recommender systems datasets with social network side information - Flixster and Epinions datasets. **Flixster**[5] is a website where users share film reviews and ratings. The users can also signify social connections. We utilized the dataset described by (Jamali and Ester 2010) which contains a ratings matrix and the social network. We selected the $M = 5,000$ users with the most friends in the network and $N = 5,000$ movies with the most ratings. This resulted in a matrix with $L = 33,182$ (density $= 0.001\,\%$) ratings and $211,702$ undirected user social connections. The identity prior covariance was used for the movies. Ratings in Flixster take one of 10 values in the set $\{0.5, 1, 1.5, \ldots, 5.0\}$. **Epinions**[6] is an item review site where users can also specify directed association by signifying a trust link. We utilized the extended Epinions dataset (Massa and Avesani 2006) and converted all the directed trust links into undirected links. We selected the $M = 5,000$ users with the most trust links in the network and $N = 5,000$ movies with the most ratings. This resulted in a matrix with $L = 187,163$ (density $0.007\,\%$) ratings and $550,298$ user social connections. The identity prior covariance was used for the items. Ratings in the Epinions dataset take one of five values in the set $\{1.0, 2.0, \ldots, 5.0\}$.

We present five fold cross validation performance for in matrix and new user predictions on both Flixster and Epinions datasets. We found that the model that selected using RMSE as the validation metric did not always perform best in terms of recall (and vice versa). This matches the results by other researchers (Steck 2010; Steck and Zemel 2010). Hence, we performed cross validation separately for RMSE and $R_{@20}$. The results on the Flixster dataset are shown in Table 3. For known users, we found that the tested models performed similarly in terms of RMSE, but the proposed models (Con MV-GP, Trace GP) significantly outperformed the matrix factorization baselines (KPMF, PMF) in terms of recall. These results are further highlighted in the $R_{@k}$ performance as shown in Fig. 4a. The results were often equivalent for new user predictions Fig. 4b. Thus our experimental results suggest that the proposed models are more accurate in terms of ranking while retaining competitive regression performance.

The RMSE and $R_{@20}$ performance on the Epinions dataset is shown in Table 4. Our results here mirror the results on the Flixster dataset. Our experiments show similar RMSE performance for all models, and a significant gain in performance in terms of $R_{@20}$ for the proposed constrained approach for known users. A similar trend is also highlighted in Fig. 4c. Con. MV-GP, Trace GP and KPMF perform similarly when tested on new users as shown in Fig. 4d with a slight performance improvement for Con. MV-GP. Comparing the Bayesian MV-GP to its constrained variant clearly shows the utility of the nuclear norm constraint in both recommender systems datasets. In all, our results suggest that the nuclear norm constrained MV-GP is effective for regression *and* for ranking in recommender systems.

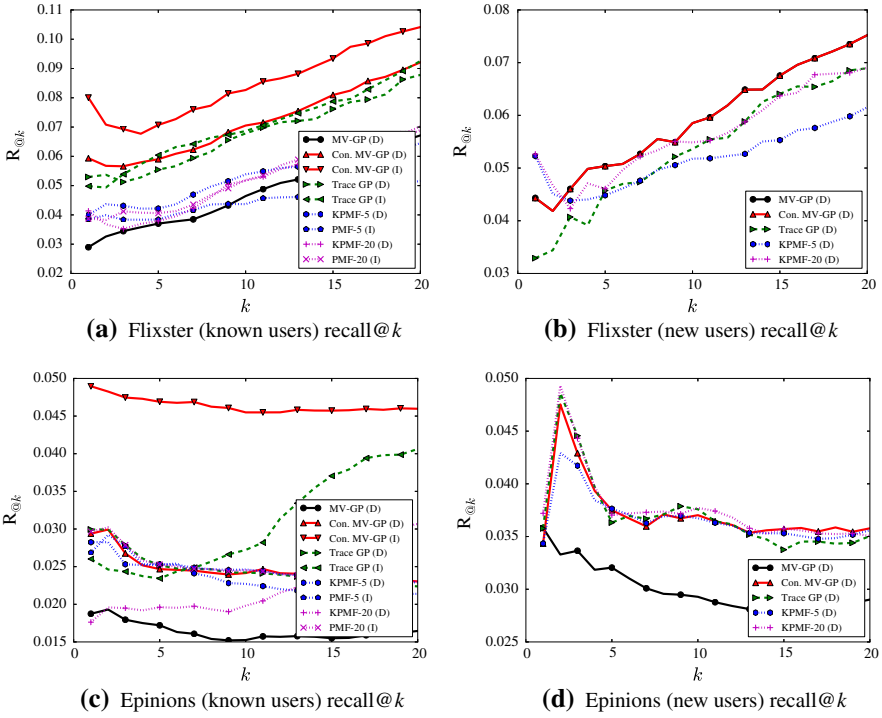
[5] www.flixster.com

[6] www.epinions.com

Table 3 Flixster dataset

Model	Known users		New users	
	RMSE	$R_{@20}$	RMSE	$R_{@20}$
MV-GP (D)	1.066 (0.006)	0.067 (0.008)	1.066 (0.088)	**0.075 (0.017)**
Con. MV-GP (D)	0.989 (0.002)	0.092 (0.012)	1.066 (0.088)	**0.075 (0.017)**
Con. MV-GP (I)	**0.982 (0.001)**	**0.104 (0.004)**	–	–
Trace GP (D)	0.989 (0.002)	0.088 (0.008)	1.066 (0.088)	0.069 (0.015)
Trace GP (I)	**0.982 (0.001)**	0.093 (0.003)	–	–
KPMF-5 (D)	0.993 (0.003)	0.064 (0.012)	1.066 (0.088)	0.062 (0.014)
PMF-5 (I)	0.995 (0.003)	0.052 (0.006)	–	–
KPMF-20 (D)	0.986 (0.001)	0.069 (0.007)	1.066 (0.088)	0.069 (0.015)
PMF-20 (I)	0.989 (0.002)	0.070 (0.003)	–	–

Avg. (std.) RMSE and $R_{@20}$ performance comparison. Smaller RMSE indicates better performance, Larger $R_{@20}$ indicates better performance. (I): Identity prior covariance, (D): Diffusion prior covariance
Bold values indicate best performance



(a) Flixster (known users) recall@k **(b)** Flixster (new users) recall@k

(c) Epinions (known users) recall@k **(d)** Epinions (new users) recall@k

Fig. 4 Performance results on recommender systems datasets. Recall $@k = 1, 2, \ldots, 20$ for known users (*left*) and new users (*right*). The prior covariances and constraints have the largest effect for very sparse data. (I): Identity prior covariance, (D): Diffusion prior covariance. Con. MV-GP outperforms KPMF (Zhou et al. 2012), which has been shown to outperform PMF (Mnih and Salakhutdinov 2007) and SoRec (Ma et al. 2008)

Table 4 Epinions dataset

Model	Known users		New users	
	RMSE	$R_{@20}$	RMSE	$R_{@20}$
MV-GP (D)	0.323 (0.007)	0.016 (0.000)	0.329 (0.020)	0.029 (0.002)
Con. MV-GP (D)	0.273 (0.005)	0.023 (0.001)	0.307 (0.022)	**0.036 (0.009)**
Con. MV-GP (I)	0.274 (0.006)	**0.046 (0.002)**	–	–
Trace GP (D)	0.273 (0.005)	0.022 (0.001)	0.307 (0.022)	0.035 (0.009)
Trace GP (I)	0.274 (0.006)	0.041 (0.003)	–	–
KPMF-5 (D)	0.274 (0.004)	0.021 (0.002)	**0.305 (0.022)**	**0.036 (0.009)**
PMF-5 (I)	**0.272 (0.004)**	0.023 (0.001)	–	–
KPMF-20 (D)	0.275 (0.005)	0.031 (0.003)	0.306 (0.022)	0.035 (0.007)
PMF-20 (I)	0.273 (0.005)	0.023 (0.001)	–	–

Avg. (std.) RMSE and $R_{@20}$ performance comparison. Smaller RMSE indicates better performance, Larger $R_{@20}$ indicates better performance. (I): Identity prior covariance, (D): Diffusion prior covariance
Bold values indicate best performance

6 Conclusion

This paper introduces a novel approach for the predictive modeling of low rank transposable data with the matrix-variate Gaussian process. The low rank is achieved using a nuclear norm constrained inference; recovering a mean function of low rank. We showed that inference for the Gaussian process with the nuclear norm constraint is convex. The proposed approach was applied to the disease-gene association task and to the recommender system task. The proposed model was effective for regression and for ranking with highly imbalanced data, and performed at least as well as (and often significantly better than) state of the art domain specific baseline models.

Recent work (Yu et al. 2013) characterizing necessary and sufficient conditions for the existence of a representer theorem points to the potential scope of the constrained inference approach combined with nonparametric processes. Thus, we plan to explore other constraint sets in addition to the nuclear norm constraint explored here. We are also interested in exploring covariance constraints as outlined for Gaussian distributions in Koyejo and Ghosh (2013a) applied to nonparametric processes. We are interested in applications of nonparametric constrained Bayesian inference to more complicated models beyond Gaussian distributions. Finally, we intend to explore the biological implications of these constrained disease gene association results in collaboration with domain experts.

Acknowledgments Authors acknowledge support from NSF grant IIS 1016614. We also thank U. Martin Blom and Edward Marcotte for providing the OMIM data set. The authors thank the anonymous reviewers for insightful comments that helped to improve this manuscript.

7 Constrained Bayesian inference

Altun and Smola (2006) studied the constrained inference approach when the constraint set is a norm ball $\mathscr{C} = \{\mathbf{c} \mid \|\mathbf{c}-\mathbf{b}\| \leq \epsilon\}$. They showed that one can apply the Fenchel duality theory to solve (3a) subject to such norm constraints and Zhu et al. (2012) extended their approach

to more general convex constraint sets. More recently Koyejo and Ghosh (2013a) showed that the constrained Bayesian inference problem satisfied a representer theorem in terms of exponential family distributions under weak conditions. The discussion in this section follows the approach of Koyejo and Ghosh (2013a).

Let $\mathscr{X}$ be a Banach space and $\mathscr{X}^*$ be its dual space. The Legendre-Fenchel transformation (convex conjugate) of a function $f : \mathscr{X} \mapsto [-\infty, +\infty]$ is $f^* : \mathscr{X}^* \mapsto [-\infty, +\infty]$ where $f^*(x^*) = \sup_{x \in \mathscr{X}} \{x^\top x^* - f(x)\}$. Further details on Fenchel duality may be found in (Borwein and Zhu 2005).

Let $g(\cdot)$ denote a regularization function defined to match the properties of the constraint set C. For instance, we may define $g(\cdot)$ as an indicator function of set membership in C or a soft penalty on the set membership. The following theorem characterizes the solution of (3b) when C is convex.

Theorem 2 *(Zhu et al. (2012)) Let g be a convex function and denote its Legendre-Fenchel conjugate by g^*,*

$$\min_{q \in \mathscr{P}} \mathrm{KL}(q(z) \| p(z)) + g(\mathrm{E}_q \left[\boldsymbol{\gamma}(z) \right]) \tag{12}$$

$$= \max_{\boldsymbol{\kappa}} \; -\log \int_z p(z) \exp(\boldsymbol{\kappa}^\top \boldsymbol{\gamma}(z)) dz - g^*(-\boldsymbol{\kappa}) \tag{13}$$

and the unique solution is given by $q_*(z) = p(z) \exp((\kappa_*)^\top \boldsymbol{\gamma}(z) - \Lambda_{\boldsymbol{\kappa}_*})$ where $\boldsymbol{\kappa}_*$ is the solution of the finite dimensional dual optimization (13) and $\Lambda_{\boldsymbol{\kappa}_*}$ ensures normalization.

Solving the resulting dual optimization (13) is often challenging. An alternative primal approach is to separate the problem into two parts. First, define the parametric form of the optimizing postdata density, then directly optimize over that parametric family. Unlike the dual approach, the proposed primal approach does not require convexity of the constraint set. However, both approaches require that a solution exists i.e. the set of densities that satisfy (3b) is not empty. For completeness, we present the details of the solution.

Denote the constraint set subject to equality constraints as $\mathscr{E}_{\mathbf{c}} = \{q \in \mathscr{P} \mid \mathrm{E}_q \left[\boldsymbol{\gamma}(z) \right] = \mathbf{c}\}$. The constrained Bayes optimization problem can be written as:

$$\min_{\mathbf{c} \in \mathsf{C}} \left[\min_{q \in \mathscr{E}_{\mathbf{c}}} \mathrm{KL}(q(z) \| p(z|y)) \right], \tag{14}$$

which requires the solution of an inner optimization:

$$q_{\mathbf{c}}(z) = \arg\min_{q \in \mathscr{E}_{\mathbf{c}}} \mathrm{KL}(q(z) \| p(z|y)). \tag{15}$$

Let $\mathsf{A} \subset \mathsf{C}$ be the set of points where the minimizer of (15) is achievable. We can associate a density function $q_{\mathbf{c}}(z)$ with every element $\mathbf{c} \in \mathsf{A}$. The *feasible set* is characterized by the set of densities $\mathscr{S} = \{q_{\mathbf{c}}(z) \mid \mathbf{c} \in \mathsf{A}\}$. The following proposition is a direct consequence of Theorem 2 and is stated without proof.

Proposition 1 (Koyejo and Ghosh 2013a) *For any $\mathbf{c} \in \mathsf{A}$, the unique minimizer of (15) is given by: $q_{\mathbf{c}}(z) = p(z|y) \exp(\boldsymbol{\kappa}_{\mathbf{c}}^\top \boldsymbol{\gamma}(z) - \Lambda_{\boldsymbol{\kappa}_{\mathbf{c}}})$ where $\boldsymbol{\kappa}_{\mathbf{c}}$ is the solution of the finite dimensional dual optimization (13) with the constraint set $\mathscr{C}' = \{\mathrm{E}_q \left[\boldsymbol{\gamma}(z) \right] = \mathbf{c}\}$ and $\Lambda_{\boldsymbol{\kappa}_{\mathbf{c}}}$ ensures normalization.*

Theorem 3 (Koyejo and Ghosh 2013a) *Let* $\mathscr{S} = \{q_{\mathbf{c}} \,|\, \mathbf{c} \in \mathsf{A}\}$ *denote the feasible set of* (15). *The postdata density given by the minimizer of* (3b) *is the solution of:*

$$q_*(z) = \arg\min_{q \in \mathscr{S}} \mathrm{KL}(q(z) \| p(z|y))$$

and the solution is given by $q_*(z) = q_{\mathbf{a}}(z)$ *for the optimal* $\mathbf{a} \in \mathsf{A}$ *with* $q_*(z) = p(z|y) \exp((\boldsymbol{\kappa}_{\mathbf{a}})^\top \boldsymbol{\kappa}(z) - \Lambda_{\boldsymbol{\kappa}_{\mathbf{a}}})$ *where* $\boldsymbol{\kappa}_{\mathbf{a}}$ *is the solution of the finite dimensional dual optimization* (13) *with the constraint set* $\mathscr{C}' = \{\mathrm{E}_q \left[\boldsymbol{\gamma}(z) \right] = \mathbf{a}\}$ *and* $\Lambda_{\boldsymbol{\kappa}_{\mathbf{a}}}$ *ensures normalization. The solution is unique if* A *is convex.*

The key insight from Proposition 1 is that the solution of (15) fully specifies the parametric form of the density. In other words, all the members of the set $\mathscr{S} = \{q_{\mathbf{c}} \,|\, \mathbf{c} \in \mathsf{A}\}$ have the same parametric form with $q_{\mathbf{c}} = f_{\boldsymbol{\theta}_{\mathbf{c}}}(z)$ is determined by the choice of $\mathbf{c}$. Note that all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $\boldsymbol{\Theta}$ is the constraint set of the parametric distribution family specified by f. The existence of this parameterized family follows from Theorem 3.

Corollary 1 *The postdata density given by the minimizer of* (3b) *is given by* $q_*(z) = f_{\boldsymbol{\theta}_*}(z)$ *where* $\boldsymbol{\theta}_*$ *is the solution of:*

$$\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\mathrm{KL}(f_{\boldsymbol{\theta}}(z) \| p(z|y)) \; s.t. \; \mathrm{E}_{f_{\boldsymbol{\theta}}} \left[\boldsymbol{\gamma}(z) \right] \in \mathscr{C} \right].$$

The expectation $\mathrm{E}_{f_{\boldsymbol{\theta}}} \left[\boldsymbol{\gamma}(z) \right]$ will be a fixed function of $\boldsymbol{\theta}$ depending on the specific parametric family. Hence Corollary 1 becomes a finite dimensional constrained optimization over $\boldsymbol{\theta}$. Corollary 1 suggests the following recipe for constrained Bayesian inference. First, Proposition 1 is applied to specify the parametric form of q_*, then Corollary 1 is applied to convert the variational problem into a finite dimensional parametric optimization.

8 Spectral norms of compact operators

Let $\mathscr{H}_{\mathscr{C}_{\mathrm{M}}}$ denote the Hilbert space of functions induced by the row prior covariance $\mathscr{C}_{\mathrm{M}}$. Similarly, let $\mathscr{H}_{\mathscr{C}_{\mathrm{N}}}$ denote the Hilbert space of functions induced by the column prior covariance $\mathscr{C}_{\mathrm{N}}$. Let $\mathbf{x} \in \mathscr{H}_{\mathscr{C}_{\mathrm{M}}}$ and $\mathbf{y} \in \mathscr{H}_{\mathscr{C}_{\mathrm{N}}}$ define (possibly infinite dimensional) feature vectors. The mean function the MV-GP is defined by a linear map $W : \mathscr{H}_{\mathscr{C}_{\mathrm{M}}} \mapsto \mathscr{H}_{\mathscr{C}_{\mathrm{N}}}$. This is the bilinear form on $\mathscr{H}_{\mathscr{C}} = \mathscr{H}_{\mathscr{C}_{\mathrm{M}}} \times \mathscr{H}_{\mathscr{C}_{\mathrm{N}}}$ given by $\psi(m, n) = \langle \mathbf{x}_m, W\mathbf{y}_n \rangle_{\mathscr{H}_{\mathscr{C}_{\mathrm{M}}}}$.

Let $\mathscr{B}$ denote the set of compact bilinear operators mapping $\mathscr{H}_{\mathscr{C}_{\mathrm{M}}} \mapsto \mathscr{H}_{\mathscr{C}_{\mathrm{N}}}$. A compact operator $W \in \mathscr{B}$ admits a spectral decomposition (Abernethy et al. 2009) with singular values given by $\{\xi_i(W)\}$. **The nuclear norm** is given by the L1 norm on the spectrum of W:

$$|||\psi|||_{1\text{-}\mathscr{H}_{\mathscr{C}}} = \sum_{i=1}^{D} \xi_i(W) \tag{16}$$

Another common regularizer is the induced **Hilbert norm** given by the L2 norm on the spectrum of W:

$$|||\psi|||^2_{2\text{-}\mathscr{H}_{\mathscr{C}}} = \sum_{i=1}^{D} \xi_i^2(W) \tag{17}$$

Further details may be found in (Berlinet and Thomas-Agnan 2004)

Let $L(\psi, \mathbf{y}, \mathsf{L})$ represent the loss function for a finite set of training data points $\mathsf{L} \in \mathsf{M} \times \mathsf{N}$ and $Q(\psi)$ be a spectral regularizer. We define the regularized risk functional:

$$L(\psi, \mathbf{y}, \mathsf{L}) + \lambda Q(\psi)$$

where $\lambda \geq 0$ is the regularization constant. A representer theorem exists, i.e., the function ψ that optimizes the regularized risk can be represented as a finite weighted sum of the prior covariance functions evaluated on training data (Abernethy et al. 2009). Hence, the optimizing function can be computed as:

$$\begin{aligned} \psi(m, n) &= \sum_{m' \in \mathsf{M}} \sum_{n' \in \mathsf{N}} \alpha_{m',n'} \mathscr{C}_{\mathrm{M}}(m, m') \mathscr{C}_{\mathrm{N}}(n, n') \\ &= \mathbf{C}_{\mathrm{M}}(m) \boldsymbol{A} \mathbf{C}_{\mathrm{N}}(n) \end{aligned} \tag{18}$$

where $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ is a parameter matrix, $\mathbf{C}_{\mathrm{M}}(m)$ is the prior covariance matrix evaluated between m and $m' \in \mathsf{M}$, i.e., the m^{th} row of $\mathbf{C}_{\mathrm{M}}$, and $\mathbf{C}_{\mathrm{N}}(n)$ is the prior covariance matrix evaluated between n and all $n' \in \mathsf{N}$.

9 Parameter estimation for the nuclear norm regularized MV-GP

Like other Bayesian modeling approaches, the constrained Bayesian inference procedure provides a mechanism for parameter estimation. This is achieved by optimizing the cost function[7] (3a) with respect to the parameters. The parameters of interest include the noise variance and the parameters of the prior row and column covariance functions. The optimization for the noise variance parameter is given by:

$$\min_{\sigma^2} L \log \sigma^2 + \frac{1}{\sigma^2} \left[\sum_{m,n \in \mathsf{L}} (y_{m,n} - \psi_{m,n})^2 + \mathrm{tr}(\mathbf{S}_L) \right]$$

This can be solved in closed form. The solution is given by:

$$\sigma^2 = \frac{1}{L} \left[\sum_{m,n \in \mathsf{L}} (y_{m,n} - \psi_{m,n})^2 + \mathrm{tr}(\mathbf{S}_L) \right] \tag{19}$$

Similarly, we can solve for the parameters that define the prior covariance functions. Suppose the row covariance and column covariance have parametric forms $\mathbf{C}_{\mathrm{M}}(\rho)$ and $\mathbf{C}_{\mathrm{N}}(\tau)$ respectively. Let $\mathbf{C}(\rho, \tau) = \mathbf{C}_{\mathrm{N}}(\tau) \otimes \mathbf{C}_{\mathrm{M}}(\rho)$ represent the joint prior covariance. We can select the covariance parameters by optimizing (3a) as:

$$J(\rho, \tau) = \min_{\rho, \tau} \ \frac{1}{2} \log |\mathbf{C}(\rho, \tau)| + \frac{1}{2} \boldsymbol{\psi}^{\top} \mathbf{C}(\rho, \tau)^{-1} \boldsymbol{\psi} + \frac{1}{2} \mathrm{tr} \Big(\mathbf{C}(\rho, \tau)^{-1} \mathbf{S} \Big)$$

The gradient with respect to ρ is given by:

$$\begin{aligned} \frac{\partial J(\rho, \tau)}{\partial \rho} = \frac{N}{2} \mathrm{tr} \left(\mathbf{C}_{\mathrm{M}}(\rho)^{-1} \frac{\partial \mathbf{C}_{\mathrm{M}}(\rho)}{\partial \rho} \right) - \frac{1}{2} \boldsymbol{\psi}^{\top} \mathbf{C}(\rho, \tau)^{-1} \frac{d \mathbf{C}(\rho, \tau)}{d \rho} \mathbf{C}(\rho, \tau)^{-1} \boldsymbol{\psi} \\ - \frac{1}{2} \mathrm{tr} \left(\mathbf{C}(\rho, \tau)^{-1} \frac{d \mathbf{C}(\rho, \tau)}{d \rho} \mathbf{C}(\rho, \tau)^{-1} \mathbf{S} \right) \end{aligned}$$

[7] Note that the model evidence term must be added back in order to use the posterior form of the constrained Bayesian inference cost function (3b).

where $\frac{\partial \mathbf{C}(\rho,\tau)}{\partial \rho} = \mathbf{C}_N \otimes \frac{\partial \mathbf{C}_M(\rho)}{\partial \rho}$, and $\frac{\partial \mathbf{C}_M(\rho)}{\partial \rho}$ is the element-wise gradient. This can be simplified further by collecting terms, and similar gradients can be computed with respect to τ. See (Rasmussen and Williams 2005, Chapter 5) for more details on the closely related approach of Gaussian process covariance parameter selection by marginal likelihood optimization. We note that the prior covariance hyperparameters may be computationally challenging to optimize in practice as the proposed updates require the storage and computation of large covariance matrices.

References

Abernethy, J., Bach, F., Evgeniou, T., & Vert, J. P. (2009). A new approach to collaborative filtering: Operator estimation with spectral regularization. *JMLR: The Journal of Machine Learning Research*, *10*, 803–826.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., et al. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, *24*(5), 537–544.

Allen, G. I., & Tibshirani, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, *4*(2), 764–790.

Allen, G. I., & Tibshirani, R. (2012). Inference with transposable data: Modelling the effects of row and column correlations. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), *74*(4), 721–743.

Altun, Y., & Smola, A. J. (2006). Unifying divergence minimization and statistical inference via convex duality. In: COLT.

Álvarez, M. A., Rosasco, L., & Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, *4*(3), 195–266.

Bauer, H. (1996). Probability Theory. De Gruyter Studies in Mathematics Series: De Gruyter.

Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Comput Linguist*, *22*(1), 39–71.

Berlinet, A., & Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Boston, Dordrecht, London: Kluwer Academic Publishers.

Bishop, C. M. (2006). *Pattern recognition and machine learning (information science and statistics)*. Secaucus, NJ, USA: Springer.

Bonilla, E., Chai, K. M., & Williams, C. (2008).Multi-task gaussian process prediction. In: NIPS ,20, 153–160.

Borwein, J., & Zhu, Q. (2005). *Techniques of variational analysis, CMS books in mathematics*. Berlin: Springer.

Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, *9*(6), 717–772.

Csató, L. (2002). Gaussian processes: Iterative sparse approximations. PhD thesis, Aston University.

Dudík, M., Phillips, S. J., & Schapire, R. E. (2007). Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, *8*, 1217–1260.

Dudik, M., Harchaoui, Z., Malick, J., et al. (2012). Lifted coordinate descent for learning with trace-norm regularization. In: AISTATS-proceedings of the fifteenth international conference on artificial intelligence and statistics-2012, Vol. 22.

Ganchev, K., & Ja, Graça. (2010). Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, *11*, 2001–2049.

Gelfand, A. E., Smith, A. F. M., & Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523–532.

Hu, Y., Koren, Y., Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In: Data Mining, 2008. ICDM'08. Eighth IEEE international conference on, IEEE, pp. 263–272.

Jaakkola, T., Meila, M., Jebara, T. (1999). Maximum entropy discrimination. In: NIPS, MIT Press.

Jamali, M., & Ester, M. (2010). A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on recommender systems, ACM, pp. 135–142.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*, 30–37.

Koyejo, O. (2013). Constrained relative entropy minimization with applications to multitask learning. PhD thesis, The University of Texas at Austin.

Koyejo, O., & Ghosh, J. (2011). A kernel-based approach to exploiting interaction-networks in heterogeneous information sources for improved recommender systems. In: Proceedings of the 2nd international workshop on information heterogeneity and fusion in recommender systems, ACM, pp. 9–16.
Koyejo, O., & Ghosh, J. (2013). Constrained Bayesian inference for low rank multitask learning. In: Proceedings of the 29th conference on Uncertainty in artificial intelligence (UAI).
Koyejo, O., & Ghosh, J. (2013). A representation approach for relative entropy minimization with expectation constraints. In: ICML workshop on divergences and divergence learning (WDDL).
Laue, S. (2012). A hybrid algorithm for convex semidefinite optimization. In: Proceedings of the 29th international conference on machine learning (ICML-12), pp. 177–184.
Lawrence, N., & Hyvärinen, A. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, *6*, 1783–1816.
Lawrence, N. D., & Urtasun, R. (2009). Non-linear matrix factorization with gaussian processes. In: Proceedings of the 26th annual international conference on machine learning, ACM, pp. 601–608.
Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., & Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, *21*(7), 1109–1121.
Li, L., & Toh, K. C. (2010). An inexact interior point method for l 1-regularized sparse covariance selection. *Mathematical Programming Computation*, *2*(3–4), 291–315.
Li, W. J., & Yeung, D. Y. (2009). Relation regularized matrix factorization. In: Proceedings of the 21st international joint conference on artificial intelligence, IJCAI'09, pp. 1126–1131.
Li, W. J., Yeung, D. Y., & Zhang, Z. (2009). Probabilistic relational PCA. In: Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 1123–1131).
Li, W. J., Zhang, Z., Yeung D. Y. (2009). Latent Wishart processes for relational kernel learning. In: D. A. V. Dyk & M. Welling (Eds.), AISTATS, pp. 336–343.
Ma, H., Yang, H., Lyu, M. R., King, I. (2008). Sorec: Social recommendation using probabilistic matrix factorization. In: Proceeding of the 17th ACM conference on Information and knowledge management, ACM, New York, NY, USA, CIKM '08, pp. 931–940.
Maglott, D. R., Ostell, J., Pruitt, K. D., & Tatusova, T. A. (2011). Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research*, *39*(Database–Issue), 52–57.
Massa, P., & Avesani, P. (2006). Trust-aware bootstrapping of recommender systems. In: ECAI 2006 workshop on recommender systems, pp. 29–33.
McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., et al. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369.
Mnih, A., & Salakhutdinov, R. (2007). Probabilistic matrix factorization. In: J. C. Platt, D. Koller, Y. Singer & S. T. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 1257–1264).
Mordelet, F., & Vert, J. P. (2011). Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, *12*, 389.
National Library of Medicine. (2012) Medical subject headings. http://www.nlm.nih.gov/mesh/. Retrieved from March 2012.
National Library of Medicine. (2012). PubMed. http://www.ncbi.nlm.nih.gov/pubmed/. Retrieved from March 2012.
NCBI. (1998). Genes and disease. Online, URL http://www.ncbi.nlm.nih.gov/books/NBK22183/. Retrieved from January 10, 2011.
Orbanz, P., & Teh, Y. W. (2010). Bayesian nonparametric models. In: C. Sammut & G. I. Webb (Eds.),*Encyclopedia of machine learning*. Berlin: Springer.
Pan, R., Zhou, Y., Cao, B., Liu, N. N., Lukose, R., Scholz, M., Yang, Q. (2008). One-class collaborative filtering. In: Data mining, 2008. ICDM'08. eighth IEEE international conference on, IEEE, pp. 502–511.
Pong, T. K., Tseng, P., Ji, S., & Ye, J. (2010). Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, *20*(6), 3465–3489.
Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning (adaptive computation and machine learning series)*. Cambridge, MA: The MIT Press.
Singh-Blom, U. M., Natarajan, N., Tewari, A., Woods, J. O., Dhillon, I. S., & Marcotte, E. M. (2013). Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PloS One*, *8*(5), e58,977.
Smola, A. J., & Kondor, R. (2003). Kernels and regularization on graphs. In: B. Schölkopf & M. K. Warmuth (Eds.), *Learning theory and kernel machines* (pp. 144–158). Berlin: Springer.
Steck, H. (2010). Training and testing of recommender systems on data missing not at random. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 713–722.

Steck, H., & Zemel, R. S. (2010). A generalized probabilistic framework and its variants for training top-k recommender systems. In: PRSAT.

Stegle, O., Lippert, C., Mooij, J. M., Lawrence, N. D., Borgwardt, K. M. (2011). Efficient inference in matrix-variate gaussian models with iid observation noise. In: *Advances in neural information processing systems* (pp 630–638).

Sutskever, I., Tenenbaum, J. B., Salakhutdinov, R. (2009). Modelling relational data using bayesian clustered tensor factorization. In: *Advances in neural information processing systems* (pp 1821–1828).

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., & Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, *6*(1), e1000641.

Xu, M., Zhu, J., & Zhang, B. (2012). Nonparametric max-margin matrix factorization for collaborative prediction. *Advances in Neural Information Processing Systems*, *25*, 64–72.

Xu, Z., Tresp, V., Yu, K., Kriegel, H. P. (2006). Learning infinite hidden relational models. Uncertainity in, Artificial Intelligence (UAI2006).

Xu, Z., Kersting, K., & Tresp, V. (2009). Multi-relational learning with gaussian processes. In: Proceedings of the 21st international joint conference on artificial intelligence, IJCAI'09, pp. 1309–1314.

Yan, F., Xu, Z., Qi, Y. A. (2011). Sparse matrix-variate gaussian process blockmodels for network modeling. In: UAI.

Yu, K., & Chu, W. (2008). Gaussian process models for link analysis and transfer learning. In: NIPS, pp 1657–1664.

Yu, K., Chu, W., Yu, S., Tresp, V., & Xu, Z. (2007). Stochastic relational models for discriminative link prediction. *Advances in neural information processing systems 19* (pp. 1553–1560). Cambridge, MA: MIT Press.

Yu, Y., Cheng, H., Schuurmans, D., Szepesvri, C. (2013). Characterizing the representer theorem. In: ICML.

Zellner, A. (1988). Optimal information processing and bayes's theorem. *The American Statistician*, *42*(4), 278–280.

Zhang, X., & Carin, L. (2012). Joint modeling of a matrix with associated text via latent binary features. *Advances in Neural Information Processing Systems*, *25*, 1565–1573.

Zhou, T., Shan, H., Banerjee, A., Sapiro, G. (2012). Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In: SDM, pp 403–414.

Zhu, J. (2012). Max-margin nonparametric latent feature models for link prediction. In: Proceedings of the 29th international conference on machine learning (ICML-12), pp 719–726.

Zhu, J., Ahmed, A., Xing, E. P. (2009). Medlda: Maximum margin supervised topic models for regression and classification. In: Proceedings of the 26th annual international conference on machine learning, ACM, pp 1257–1264.

Zhu, J., Chen, N., Xing, E. P. (2011). Infinite latent SVM for classification and multi-task learning. In: J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), Advances in neural information processing systems (Vol. 24, pp 1620–1628).

Zhu, J., Chen, N., Xing, E. P. (2012). Bayesian inference with posterior regularization and infinite latent support vector machines. CoRR abs/1210.1766.