Probabilistic Matrix Addition

Amrudin Agovic AAGOVIC@CS.UMN.EDU
Arindam Banerjee BANERJEE@CS.UMN.EDU
Dept of Computer Science & Engg, University of Minnesota, Twin Cities

Snigdhansu Chatterjee CHATTERJEE@STAT.UMN.EDU
School of Statistics, Univeristy of Minnesota, Twin Cities

Abstract

We introduce Probabilistic Matrix Addition (PMA) for modeling real-valued data matrices by simultaneously capturing covariance structure among rows and among columns. PMA additively combines two latent matrices drawn from two Gaussian Processes respectively over rows and columns. The resulting joint distribution over the observed matrix does not factorize over entries, rows, or columns, and can thus capture intricate dependencies in the matrix. Exact inference in PMA is possible, but involves inversion of large matrices, and can be computationally prohibitive. Efficient approximate inference is possible due to the sparse dependency structure among latent variables. We propose two families of approximate inference algorithms for PMA based on Gibbs sampling and MAP inference. We demonstrate the effectiveness of PMA for missing value prediction and multi-label classification problems.

1. Introduction

We introduce a novel approach for modeling data matrices which can simultaneously capture (nonlinear) covariance structures among rows as well as columns. For a $n \times m$ matrix X, existing approaches which consider both covariance structures can be broadly divided into two categories: Gaussian Process approaches which suitably modify a given kernel to incorporate relational information and subsequently draw outputs (rows) i.i.d. from a single GP (Silva et al., 2007; Xu et al., 2009; Higdon, 2002); and Linear Models of Corregionalization (LMC) (Wackernagel, 2003; Gelfand & Banerjee, 2010), a widely used family of models from Geostatistics, which effectively flattens out the data matrix into a long vector, and uses suitable covariance structures over the vectorized form. As a result, the entries, rows, or columns of the matrix are not independent as covariances between all entries are modeled. However, the lack of (conditional) independence can lead to serious scalability problems for inference in LMCs.

In this paper, we introduce the Probabilistic Matrix Addition (PMA) model which simultaneously considers two (nonlinear) kernels $\mathcal{K}_1$ and $\mathcal{K}_2$ corresponding to the rows and the columns of the matrix respectively. The kernels are utilized in two Gaussian Processes (GPs), from which we draw two latent matrices with independence along rows and along columns respectively. The latent matrices from the two GPs are added to obtain the final matrix, yielding a nonparametric generative model for real-valued data matrices of any size. As GPs define priors over functions $f(x)$, PMA can be viewed as a simple but non-trivial way to define priors over functions $f(x, y)$ which when instantiated lead to finite sized matrices.

Similar to LMCs, the joint distribution of PMA over the matrix entries does not factorize over entries, rows, or columns, and thus can capture intricate dependencies among the entries. Unlike LMC, PMA does not assume stationarity. It exhibits a conditional independence structure over the latent variables, which allows for fast approximate inference algorithms. We present two methods for approximate inference in PMA, respectively based on Gibbs sampling and MAP inference. The Gibbs sampler is efficient since it takes full advantage of the conditional independence structure, and precision matrices over the conditioning variables can be computed using a suitable application of the Sherman-Morrison formula. The MAP inference is obtained by solving a Sylvester equation (Golub et al., 1979; Wachspress, 1988) where both row and column

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

covariances play a role in determining a latent variable matrix. For parameter estimation and missing values prediction, the inference methods are used in a suitable alternating update framework.

We illustrate the effectiveness of PMA on two tasks: matrix missing value prediction, where the goal is to infer multiple missing values in a given data matrix; and multi-label classification, where the goal is to predict an entire new row given a matrix which may also have missing values. For matrix missing value prediction, we compare PMA to a single GP capturing covariances either across rows or columns, Probabilistic Matrix Factorization (PMF) (Salakhutdinov & Mnih, 2007) and LMC (Gelfand & Banerjee, 2010). PMA clearly outperforms a single GP, and is competitive or better than PMF and LMC. For multi-label classification, we compare PMA to three baselines including state-of-the art approaches designed specifically for multi-label classification. Across all evaluation measures and datasets, PMA consistently outperforms the other methods.

The rest of the paper is organized as follows. In Section 2, we introduce PMA, discuss its properties, and contrast it with LMCs. We consider the missing value prediction problem in Section 3, propose two inference approaches for PMA, and present empirical evaluation of the ideas. In Section 4, we discuss how new rows (or columns) can be predicted using PMA, and present empirical evaluation on the multi-label prediction problem. We briefly discuss related work in Section 5 and conclude in Section 6.

2. The Model

The Probabilistic Matrix Addition (PMA) model defines distributions over real valued matrices. Let X be a $n \times m$ matrix. We start by outlining a generative model for any such matrix for arbitrary n and m. Consider two Gaussian processes $G_1 \equiv GP(0, \mathcal{K}_1)$ with covariance function $\mathcal{K}_1$ corresponding to rows and $G_2 \equiv GP(0, \mathcal{K}_2)$ with covariance function $\mathcal{K}_2$ corresponding to columns. For n rows, we get the following distribution over any column $f \in \mathbb{R}^n$ from G_1:

$$p(f|G_1) = \frac{1}{(2\pi)^{n/2}|\mathcal{K}_1|^{1/2}} \exp\left(-\frac{1}{2} f^T \mathcal{K}_1^{-1} f\right) \ . \quad (1)$$

Since the matrix will have m columns, we sample $f_1, \ldots, f_m \in \mathbb{R}^n$ independently following the above distribution. The samples form the following $n \times m$ matrix F:

$$F = \begin{bmatrix} f_1 & \cdots & f_m \end{bmatrix} = \begin{bmatrix} f_1(1) & \cdots & f_m(1) \\ \vdots & \ddots & \vdots \\ f_1(n) & \cdots & f_m(n) \end{bmatrix} \ . \quad (2)$$

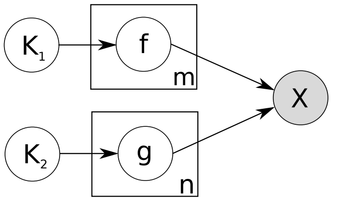


Figure 1. Graphical model for PMA: X is generated as the sum of F, sampled by column, and G, and sampled by row.

For m columns, we get the following distribution over any row $g \in \mathbb{R}^m$ from G_2:

$$p(g|G_2) = \frac{1}{(2\pi)^{m/2}|\mathcal{K}_2|^{1/2}} \exp\left(-\frac{1}{2} g^T \mathcal{K}_2^{-1} g\right) \ . \quad (3)$$

Since the matrix will have n rows, we sample $g_1, \ldots, g_n \in \mathbb{R}^m$ independently following the above distribution. The samples form the following $n \times m$ matrix G:

$$G = \begin{bmatrix} g_1^T \\ \vdots \\ g_n^T \end{bmatrix} = \begin{bmatrix} g_1(1) & \cdots & g_1(m) \\ \vdots & \ddots & \vdots \\ g_n(1) & \cdots & g_n(m) \end{bmatrix} \ . \quad (4)$$

Given the two random matrices F and G, we generate the $n \times m$ random matrix X as

$$X = F + G \ . \quad (5)$$

In particular, each entry of X is (Figure 1)

$$x_{ij} = f_j(i) + g_i(j) \ . \quad (6)$$

While the generative process for X is simple, it leads to intricate dependencies between its entries, in particular capturing (nonlinear) covariance structures along rows as well as columns.

2.1. Joint and Conditional Distributions

Joint Distribution: First, we consider the joint distribution of the components of the entire matrix $X = [x_{ij}] \in \mathbb{R}^{m \times n}$. Since $f_j(i) \sim N(0, \mathcal{K}_{1,(i,i)})$ and $g_i(j) \sim N(0, \mathcal{K}_{2,(j,j)})$, the marginal distribution of x_{ij} is a univariate Gaussian: $x_{ij} \sim N(0, \mathcal{K}_{1,(i,i)} + \mathcal{K}_{2,(j,j)})$. To compute the joint covariance, first note that $E[f_j(i) f_j(\ell)] = K_1(i, \ell)$, $E[f_j(i) f_k(\ell)] = 0$, $E[g_i(j) g_i(k)] = K_2(j,k)$, and $E[g_i(j) g_\ell(k)] = 0$. As a result, $E[x_{ij}^2] = K_1(i,i) + K_2(j,j)$, $E[x_{ij} x_{ik}] = K_2(j,k)$, $E[x_{ij} x_{\ell j}] = K_1(i, \ell)$, and $E[x_{ij} x_{\ell k}] = 0$. Putting everything together, if $vec(X)^T = [X_{(:,1)}^T, \ldots, X_{(:,m)}^T]$ denotes the vectorized version of X, then the joint distribution of $vec(X) \in \mathbb{R}^{mn}$ is a multivariate Gaussian, i.e., $vec(X) \sim N(0, \Sigma_{vec(X)})$ where

$$\Sigma_{vec(X)} = (\mathbb{I}_m \otimes \mathcal{K}_1) + (\mathcal{K}_2 \otimes \mathbb{I}_n) = \mathcal{K}_1 \oplus \mathcal{K}_2 . \quad (7)$$

where $\oplus$ denotes the Kronecker sum (Laub, 2005).

Conditional Distributions: We now consider the conditional distribution of each f_{ij}, g_{ij}, and x_{ij} given the rest of the latent matrices, i.e., $F_{(-i,-j)}$ and $G_{(-i,-j)}$. Given $F_{(-i,-j)}$, $f_j(i)$ only depends on $f_j(-i)$, the other elements of f_j. Further, $f_j(i)$ is conditionally independent of $G_{(-i,-j)}$ given $F_{(-i,-j)}$. To see this, note that in the PMA graphical model there are two types of paths connecting $f_j(i)$ to elements in $G_{(-i,-j)}$: paths going through the collider x_{ij} and paths going through elements in $f_j(-i)$. The conditional d-separation between $f_j(i)$ and $G_{(-i,-j)}$ given $F_{(-i,-j)}$ stems from the fact that there is no conditioning on the collider x_{ij} for paths of the first type and conditioning on the non-colliders (elements of $f_j(-i)$) for paths of the second type. By a similar argument, $g_i(j)$ depends only on $g_i(-j)$, the other elements of g_i, and is conditionally independent of $F_{(-i,-j)}$ given $G(-i,-j)$. As a result, we have

$$\begin{aligned} f_j(i)|(F_{(-i,-j)}, G_{(-i,-j)}) &\sim N(m_j^f(i), s_j^f(i)) \ , \\ g_i(j)|(F_{(-i,-j)}, G_{(-i,-j)}) &\sim N(m_i^g(j), s_i^g(j)) \ , \end{aligned} \tag{8}$$

where

$$\begin{aligned} m_j^f(i) &= \mathcal{K}_{1,(i,-i)}\mathcal{K}_{1,(-i,-i)}^{-1} f_j(-i) \ , \\ s_j^f(i) &= \mathcal{K}_{1,(i,i)} - \mathcal{K}_{1,(i,-i)}\mathcal{K}_{1,(-i,-i)}^{-1}\mathcal{K}_{1,(-i,i)} \ , \\ m_i^g(j) &= \mathcal{K}_{2,(j,-j)}\mathcal{K}_{2,(-j,-j)}^{-1} g_i(-j) \ , \\ s_i^g(j) &= \mathcal{K}_{2,(j,j)} - \mathcal{K}_{2,(j,-j)}\mathcal{K}_{2,(-j,-j)}^{-1}\mathcal{K}_{2,(-j,j)} \ . \end{aligned} \tag{9}$$

Since $x_{ij} = f_j(i) + g_i(j)$, we have

$$x_{ij}|(F_{(-i,-j)}, G_{(-i,-j)}) \sim N(m_{ij}, s_{ij}) \ , \tag{10}$$

where

$$m_{ij} = m_j^f(i) + m_i^g(j) \ , \ \ s_{ij} = s_j^f(i) + s_i^g(j) \ . \tag{11}$$

2.2. Relationship with LMCs

Linear Models of Corregionalization (LMCs) are a broad family of related models widely studied in Geostatistics (Wackernagel, 2003; Gelfand & Banerjee, 2010). We compare and contrast the proposed PMA with LMCs. The simplest form of LMC, also known as the separable model or intrinsic specification (Mardia & Goodall, 1993; Gelfand & Banerjee, 2010), works with vectors $X(s_j) \in \mathbb{R}^m$ at locations $s_j, j = 1, \ldots, n$. The objective is to capture associations within a given location and across locations. Following common notation from Geostatistics (Gelfand & Banerjee, 2010), let $X(s) = Aw(s)$, be a process where $A \in \mathbb{R}^{m \times m}$ is a full rank matrix and $w_j(s) \sim N(0,1)$ are i.i.d. processes with stationary correlation function $\rho(s-s') = corr(w_j(s), w_j(s'))$ not depending on j. Let $T = AA^T \in \mathbb{R}^{m\times m}$ denote the local covariance matrix. The cross covariance $\Sigma_{X(s),X(s')}$ can then be expressed as $\Sigma_{X(s),X(s')} = C(s-s') = \rho(s-s')T$. Thus, by flattening out X as $vec(X) \in \mathbb{R}^{mn}$, the joint distribution of $vec(X) \sim N(0, \Sigma_{vec(X)})$ where $\Sigma_{vec(X)} = R \otimes T$, $R_{ss'} = \rho(s-s')$, and $\otimes$ denotes the Kronecker product. More general versions of LMC can be obtained by abandoning the i.i.d. assumption on $w_j(s)$ or by considering a nested covariance structure (Gelfand & Banerjee, 2010): $C(s-s') = \sum_u \rho_u(s-s')T^{(u)}$. Since the component processes are zero mean, the intrinsic formulation of LMC (Gelfand & Banerjee, 2010) only requires the specification of the second moment of the differences in measurements, given by $\Sigma_{X(s)-X(s')} = \Psi(s-s') = C(0) - C(s-s') = T - \rho(s-s')T = \gamma(s-s')T$. The function $\gamma(s-s') = \rho(0) - \rho(s-s')$, where $\rho(0) = 1$, is referred to as a variogram. Learning and inference in LMCs are typically performed by assuming a parametric form for the variogram (Zhang, 2007; Wackernagel, 2003). Several recent publications in machine learning (Bonilla et al., 2008; Teh & Seeger, 2005) can be seen as special cases of LMCs.

The proposed PMA is related to LMCs as is evident from the structure of the joint covariance matrices. However, there are important differences between the two models, including modeling assumptions as well as efficiency of inference algorithms. We briefly discuss these aspects below. First, LMCs are stationary models where the covariance depends on $(s - s')$, whereas PMA does not make such an assumption. Further, generally LMCs do not have an explicit latent variable based generative model in their specification. In particular, the statistical dependency structure of the elements of X tends to be complete. As a result, inference in LMCs typically involve one or both of the following possible issues: (i) Inverting large covariance matrices, say $\mathbb{R}^{(mn-p)\times(mn-p)}$ matrices for p missing entries, which is computationally prohibitive, (ii) Assuming a parametric form of the variogram which greatly restricts modeling flexibility (Gelfand & Banerjee, 2010). In contrast, PMA has a latent variable based model specification and the statistical dependency structure in PMA is significantly sparse. The sparsity can be exploited to develop efficient approximate inference algorithms (see Section 3). Since the joint distribution is Gaussian, exact inference can be done in PMA but has the same computational issues as in LMCs.

3. Predicting Missing Values

For missing value prediction, we are given a partially observed data matrix X. The goal is to infer the missing values based on the structure of the known observations. In this section we outline two approaches for missing value prediction, respectively based on Gibbs sampling and MAP inference. We conclude the section with an experimental evaluation of PMA for missing value prediction.

3.1. Gibbs Sampling

Let $\tilde{X}$ be a full matrix, where the missing values have been initialized to random values. For given gram matrices (K_1, K_2), the sampler updates the latent matrices and the missing entries in $\tilde{X}$. Since $X = F + G$, it is sufficient to sample only F or only G—we choose to sample G. If K_1 and/or K_2 is unknown, we alternate between sampling (G, X) and estimating K_1 and/or K_2.

Sampling G: Given K_1, K_2, and a full data matrix $\tilde{X}$, using Bayes rule we have

$$p(g_i(j)|G_{(-i,-j)}, \tilde{X}, K_1, K_2) \propto$$
$$p(g_i(j)|G_{(-i,-j)}, \tilde{X}_{(-i,-j)}, K_1, K_2) p(\tilde{x}_{ij}|G, \tilde{X}_{(-i,-j)}, K_1, K_2)$$
$$= p(g_i(j)|G_{(-i,-j)}, K_2) p(\tilde{x}_{ij} - g_i(j)|F_{(-i,-j)}, K_1) \ ,$$

due to conditional independence and the fact that $F_{(-i,-j)} = \tilde{X}_{(-i,-j)} - G_{(-i,-j)}$. Note that the individual distributions are univariate Gaussians as in (8) and (9). Since the product of two Gaussians is also a Gaussian, we have

$$p(g_i(j)|G_{-i,-j}, X, K_1, K_2) \propto N(g_i(j)|\mu_{ij}, \sigma_{ij}^2) \qquad (12)$$

where

$$\mu_{ij} = \frac{m_i^g(j) s_j^f(i) + m_j^f(i) s_i^g(j)}{s_j^f(i) + s_i^g(j)} \ , \ \sigma_{ij}^2 = \frac{s_j^f(i) s_i^g(j)}{s_j^f(i) + s_i^g(j)} \ , \qquad (13)$$

with m^g, m^f, s^g, s^f are from (9).

The sampler involves several matrix inverses, viz $K_{1,(-i,-i)}^{-1}$ and $K_{2,(-j,-j)}^{-1}$, but these can be computed efficiently from K_1^{-1} and K_2^{-1}. For computations involving K_1, instead of computing n inverses of $(n-1) \times (n-1)$ sub-matrices of K_1 (see (9)), we can obtain each such inverse from rank-2 modifications to K_1. Assuming that K_1^{-1} has been computed, consider the problem of computing $K_{1,(-1,-1)}^{-1} = K_{1,(2:n,2:n)}^{-1}$. According to the Sherman-Morrison formula, we have $(K_1 + uv^t)^{-1} = K_1^{-1} - (K_1^{-1} uv^t K_1^{-1})/(1 + v^t K_1^{-1} u)$, where $1 + v^t K_1^{-1} u \neq 0$ and $u, v \in \mathbb{R}^n$. We construct rank-2 updates to zero out entries $K_{1,(2:n,1)}$ and $K_{1,(1,2:n)}$. This can be accomplished in two steps, first we obtain $A = K_1 + u_1 v_1^T$ where $u_{1(1)} = 0$, $u_{1(2:n)} = -K_{1(2:n,1)}$, $v_{1(1)} = 1$, $v_{1(2:n)} = 0$. Then we obtain $B = A + u_2 v_2^T$ where $u_{2(1)} = 1$, $u_{2(2:n)} = 0$, $v_{2(1)} = 0$, $v_{2(2:n)} = -K_{1(1,2:n)}$. Applying the Sherman-Morrison formula twice we compute $B^{-1} = (K_1 + u_1 v_1^T + u_2 v_2^T)^{-1}$. From basic properties of block matrices it follows: $K_{1,(-1,-1)}^{-1} = K_{1,(2:\,n,2:\,n)}^{-1} = B_{(2:n,2:n)}^{-1}$. We follow a similar computation for all the n submatrices $K_{1,(-i,-i)}$. Similarly, we can efficiently compute the m inverses of $(m-1) \times (m-1)$ submatrices $K_{2,(-j,-j)}$ of K_2.

Sampling $\tilde{X}$: Missing values in X are sampled by extending the sampler and treating the missing $\tilde{x}_{ij}$ as latent variables. In particular, we sample $\tilde{x}_{ij}$ conditioned on $\tilde{X}_{(-i,-j)}$ and one of F and G. Conditioning on F, we have

$$p(\tilde{x}_{ij}|\tilde{X}_{(-i,-j)}, F, K_1, K_2) = N(x_{ij}|\bar{x}_{ij}, \zeta_{ij}) \qquad (14)$$

where

$$\bar{x}_{ij} = f_j(i) + K_{2,(i,-i)} K_{2,(-i,-i)}^{-1} (\tilde{x}_{-i,j} - f_j(-i)) \ ,$$
$$\zeta_{ij} = K_{2,(i,i)} - K_{2,(i,-i)} K_{2,(-i,-i)}^{-1} K_{2,(-i,i)} \ . \qquad (15)$$

Parameter Estimation: If K_1 and K_2 are unknown, we initialize $\hat{K}_1 \succ 0, \hat{K}_2 \succ 0$, and alternate between sampling $(G, \tilde{X})$ and estimating $(\hat{K}_1, \hat{K}_2)$. We have already outlined how to sample G and $\tilde{X}$. Let $F = \tilde{X} - G$. Then, we have $\hat{K}_1 = \frac{1}{m} \sum_{i=1}^m f_j f_j^T$ and $\hat{K}_2 = \frac{1}{n} \sum_{i=1}^n g_i g_i^T$.

3.2. MAP Inference

As before, we start with a full matrix $\tilde{X}$, where the missing values have been filled at random. For given gram matrices (K_1, K_2), we alternate between estimating F (or G) and $\tilde{X}$.

Estimating F: Given $\tilde{X}, K_1$, and K_2 the joint log-likelihood over (X, F) is:

$$\log p(\tilde{X}, F|K_1, K_2) = \log p(F|K_1) + \sum_{i=1}^n \log p(\tilde{x}_{i:}|f_{:}(i), K_2) \ .$$

For a given $\tilde{X}$, the MAP F can be obtained by maximizing the joint log-likelihood, or equivalently minimizing

$$\sum_{i=1}^n (\tilde{x}_i - F^T e_i^n)^T K_2^{-1} (\tilde{x}_i - F^T e_i^n) + \sum_{j=1}^m e_j^{mT} F^T K_1^{-1} F e_j^m \ ,$$

where $e_i^n \in \mathbb{R}^n, e_j^m \in \mathbb{R}^m$ are vectors of all zeros with the i^{th} and j^{th} position set to one respectively. A direct calculation shows that the solution has to satisfy the following Sylvester equation

$$FK_2 + K_1F = K_1\tilde{X} \ . \quad (16)$$

A solution to the Sylvester equation exists if and only if no eigenvalue of K_1 is equal to the negative of an eigenvalue of K_2 (Golub et al., 1979). Since both K_1 and K_2 are positive definite, the condition is satisfied, and the solution can be obtained by standard methods (Golub et al., 1979; Wachspress, 1988).

Estimating $\tilde{X}$: We iteratively update the originally missing entries $\tilde{x}_{ij}$ based on the mode of the distribution $p(\tilde{x}_{ij}|F, \tilde{X}_{(-i,-j)}, K_1, K_2)$, given by

$$\tilde{x}_{ij}^{new} = f_j(i) + \mathcal{K}_{2,(j,-j)}\mathcal{K}_{2,(-j,-j)}^{-1}(\tilde{x}_{-i,j} - f_j(-i)) \ . \quad (17)$$

Note that the expression is similar to (15), where we sample from the corresponding distribution.

Parameter Estimation: We initialize $\hat{K}_1 \succ 0$, $\hat{K}_2 \succ 0$, and the missing values of X randomly. Then, we alternate between updating $(\tilde{X}, F)$, and estimating $\hat{K}_1$, $\hat{K}_2$. We have already discussed updates for $(\tilde{X}, F)$. Let $G = \tilde{X} - F$. Then $\hat{K}_1 = \frac{1}{m}\sum_{j=1}^{m} f_j f_j^T$ and $\hat{K}_2 = \frac{1}{n}\sum_{i=1}^{n} g_i g_i^T$.

3.3. Experimental Evaluation

We report results from two sets of experiments for missing value prediction. In the first set, we compare PMA to Gaussian Process regression (GPR) on simulated datasets. In the second set, we compare PMA to other algorithms, including GPR, Probabilistic Matrix Factorization (PMF) (Salakhutdinov & Mnih, 2007), and intrinsic LMC (I-LMC) on benchmark datasets.

Datasets and Evaluation: For the first set, we use PMA to generate artificial datasets, of size 50 × 20 and 50 × 50. Evaluation is done using mean square error of the predicted values. For the second set, we use two multi-label classification datasets—Emotions (Trohidis et al., 2008) and Scene (Boutell et al., 2004). In multi-label classification, for n points and m classes, class memberships are represented as $n \times m$ binary matrix B. We consider a truncated log-odds matrix X, with $x_{ij} = c$ if $b_{ij} = 1$, and $x_{ij} = -c$ if $b_{ij} = 0$. For the experiments, certain entries x_{ij} are assumed to be missing. Evaluation is done using class membership prediction accuracy based on $\text{sign}(\hat{x}_{ij})$.

Methodology: All experiments were conducted using five-fold cross validation. For the first set, both K_1 and K_2 are assumed to be unknown, and are estimated from data. For the second set, K_1 is instantiated using the RBF kernel function $\mathcal{K}_1$ based on feature vectors of the data objects, and K_2 is estimated from data.

Algorithms: For the first set, we compare PMA to GPR. GPR-D1 treats rows as data points, while GPR-D2 treats columns as data points. We utilize one implementation of PMA based on MAP inferece (PMA-MAP) and one based on Gibbs sampling (PMA-GIBBS). For the second set, we compare five different algorithms: GPR, PMA-MAP, PMA-GIBBS, PMA-EXACT, PMF, and I-LMC. PMA-EXACT performs exact prediction by flattening out the matrix X into a vector, assuming a covariance matrix of the form $\mathcal{K}_1 \oplus \mathcal{K}_2$. I-LMC corresponds to intrinsic LMC in the prediction step, whereby we utilize a covariance matrix of the form $\mathcal{K}_1 \otimes \mathcal{K}_2$. For both PMA-EXACT and I-LMC we use a provided kernel $\mathcal{K}_1$ and K_2 is estimated from data. The purpose of comparing the latter two algorithms is to see whether PMA suffers by assuming a sparser dependency structure.

Performance: The results for the first set involving simulated data is in Figure 2. PMA performs better than both GPR-D1 and GPR-D2, suggesting that there is a clear benefit in modeling both $\mathcal{K}_1$ and $\mathcal{K}_2$. While not all matrices will necessarily have relevant correlation structure across both dimensions, when this is the case, PMA appears to do better. PMA-GIBBS and PMA-MAP perform similarly, with PMA-GIBBS appearing slightly better.

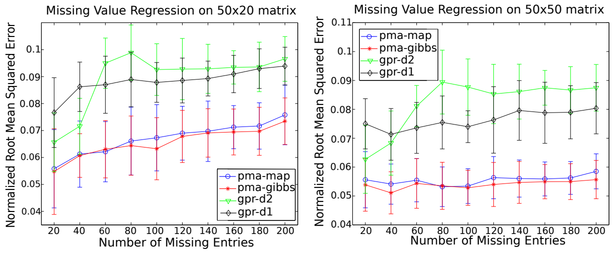
The results for the second set involving benchmark datasets is in Table 1. We make the following observations: (i) PMA clearly outperforms a GPR, illustrating the value in modeling correlations across both rows and columns; (ii) The differences in performance between PMA-GIBBS, PMA-EXACT and I-LMC are negligible. The results indicate that PMA does not suffer by assuming a sparser dependency structure. Further, PMA-GIBBS is fairly accurate when compared to PMA-EXACT, which can be computationally prohibitive for large datasets; (iii) PMA-GIBBS appears to perform slightly but consistently better than PMA-MAP; (iv) PMA is competitive compared to PMF. On Emotions, PMF appears slightly better. On Scene, PMA is significantly better, suggesting a clear advantage for certain datasets.

4. Predicting New Rows

We consider the problem of predicting a new row in the data matrix X assuming that $\mathcal{K}_1$ is a known kernel function. The motivation comes from multi-label classification, where a new row translates to all labels for a data point not encountered before. The methods developed can also be applied to predict new columns assuming $\mathcal{K}_2$ is a known kernel function. As before, we outline two inference approaches based on Gibbs sam-

Table 1. Error rates for recovering missing labels obtained using five-fold cross validation on the Emotions and Scene data sets. Performance of GPR, PMA-MAP, PMA-GIBBS, PMA-EXACT, I-LMC and PMF is evaluated while an increasing percentage of labels are missing in the training data. Missing Labels are randomly selected. The error rates reflect the percentage of missing labels incorrectly recovered.

	10%	15%	20%	25%	30%
Emotions					
GPR	32.3 ± 5.9	33.1 ± 4.7	32.6 ± 5.0	34.6 ± 2.3	35.3 ± 2.3
PMA-MAP	23.9 ± 6.5	25.3 ± 3.8	26.9 ± 4.4	29.7 ± 4.8	30.8 ± 4.7
PMA-GIBBS	23.3 ± 5.3	$\mathbf{24.8 \pm 3.2}$	$\mathbf{25.1 \pm 3.8}$	$\mathbf{27.2 \pm 3.9}$	28.0 ± 4.0
PMA-EXACT	$\mathbf{19.7 \pm 4.9}$	$\mathbf{23.6 \pm 6.9}$	25.8 ± 4.0	$\mathbf{27.3 \pm 5.4}$	$\mathbf{27.9 \pm 4.1}$
I-LMC	$\mathbf{20.3 \pm 4.6}$	25.1 ± 5.9	$\mathbf{25.7 \pm 3.7}$	27.6 ± 4.5	$\mathbf{27.8 \pm 3.8}$
PMF	$\mathbf{21.8 \pm 5.0}$	$\mathbf{22.6 \pm 2.4}$	$\mathbf{24.6 \pm 3.0}$	$\mathbf{26.3 \pm 1.6}$	$\mathbf{26.0 \pm 3.7}$
Scene					
GPR	14.7 ± 1.7	34.5 ± 8.0	17.2 ± 2.1	17.4 ± 1.7	18.0 ± 2.1
PMA-MAP	11.9 ± 1.0	13.6 ± 2.5	13.8 ± 2.7	13.9 ± 3.2	14.8 ± 1.5
PMA-GIBBS	$\mathbf{10.3 \pm 1.4}$	$\mathbf{10.9 \pm 2.6}$	$\mathbf{11.1 \pm 1.8}$	$\mathbf{11.3 \pm 2.1}$	$\mathbf{12.3 \pm 1.2}$
PMA-EXACT	$\mathbf{10.4 \pm 1.0}$	$\mathbf{11.0 \pm 1.0}$	$\mathbf{11.6 \pm 1.8}$	$\mathbf{11.9 \pm 1.2}$	$\mathbf{12.5 \pm 2.3}$
I-LMC	10.4 ± 1.0	$\mathbf{10.9 \pm 1.0}$	$\mathbf{11.8 \pm 1.7}$	$\mathbf{11.8 \pm 1.2}$	$\mathbf{12.9 \pm 2.6}$
PMF	$\mathbf{9.2 \pm 2.2}$	13.8 ± 3.0	16.1 ± 3.4	18.5 ± 2.8	20.1 ± 3.0



(a) 50x20 Matrix

(b) 50x50 Matrix

Figure 2. Five fold cross validation on two artificially created data sets. PMA clearly benefits from modeling both covariance structures. The GPR both across rows (GPR-D1) and columns (GPR-D2) is weaker. Gibbs sampling appears to do slightly better compared to MAP, when it comes to inference in PMA.

pling and MAP inference respectively. We evaluate PMA for new row prediction in the task of multi-label classification.

We first focus on initializing the new row $x_{(n+1):}$ of X. Since F and G do not have values for this new row, one needs to get suitable extensions for F and G. Since F has dependencies along columns, for each column j, we obtain the MAP estimate $f_j(n+1)$ using GP regression and $\mathcal{K}_1$ yielding the extended matrix $\tilde{F} \in \mathbb{R}^{(n+1)\times m}$. Since G does not have dependencies along columns, we sample a new row $g_{n+1}^T \sim GP(0, \mathcal{K}_2)$ yielding the extended $\tilde{G} \in \mathbb{R}^{(n+1)\times m}$.

For Gibbs Sampling, we obtain the initial extended matrix as $\tilde{X} = \tilde{F} + \tilde{G}$. Then we proceed as in Section 3.1, while treating the entire last row of $\tilde{X}$ as latent, in addition to $\tilde{G}$ and any other missing entries in $\tilde{X}$. For MAP inference, since g_{n+1}^T is zero mean, the $(n+1)^{st}$ row of $\tilde{F}$ serves directly as an estimate for the new row $x_{(n+1):}$. In either setting, if K_2 is unknown, we alternate between sampling/estimating $\tilde{X}$ and estimating K_2.

4.1. Experimental Evaluation

We compare the performance of PMA (PMA-GIBBS) to existing state-of-the-art methods for multi-label classification on a number of benchmark data sets. We use the Scene (Boutell et al., 2004) and Emotions (Trohidis et al., 2008) datasets for evaluation. As before, we use truncated log-odds during learning, and the sign of the predicted score for evaluation.

Algorithms and Methodology: We evaluate PMA-GIBBS against three multi-label classification algorithms. For PMA-GIBBS, we assume that $\mathcal{K}_1$ is an RBF Kernel over the points, where its parameters are estimated using cross validation, and K_2 is unknown and estimated from the data. The inputs into $\mathcal{K}_1$ are given by feature vectors of the data points. As a baseline, we consider one-vs-rest SVM as a multi-label classifier, which we refer to as MLSVM. We also consider two state-of-the-art approaches for multi-label learning: Multi Label K-nearest Neighbors (MLKNN) (Zhang, 2007), a method which applies the k-nearest neighbor idea to the multi-label setting; and Instance Based Learning by Logistic Regression (IBLR) (Cheng & Hüllermeier, 2009), where features are first transformed to incorporate label information from local neighborhoods prior to applying logistic regression. In all multi-label experiments, we utilize an RBF Kernel in PMA, where the parameter σ is chosen by cross-validation.

We also consider the setting where the training set has partial labels, i.e., missing entries in X. While PMA can utilize partial labels, the other algorithms cannot. Hence, we construct a reduced training set discarding points with partial labels. The corresponding models are called PMA-GIBBS-D, MLKNN-D, IBLRML-D and MLSVM-D.

We evaluated multi-label classification performance using three well known multi-label evaluation measures: one error, precision, and ranking loss (Cheng & Hüllermeier, 2009) by running five-fold cross-validation. Low values for one error and ranking loss are preferred, while high values for precision are desirable.

Performance: We evaluate performance on both multi-label datasets by considering an increasing number of labeled points. As seen in Figure 3, PMA outperforms the other three methods on both data sets and for all three performance measures, and the improvements are significant. Further, all three methods designed specifically for multi-label classification outperform MLSVMs on all datasets and evaluation measures.

We also tested the prediction performance when missing labels are present in the training data. As seen in Table 2, PMA-GIBBS significantly outperforms the other models due to its ability to leverage partially labeled data. Among algorithms which discard points with partial labels, PMA-GIBBS-D outperforms the others.

Table 2. Five fold cross validation on the Scene data set with 25% of label entries missing. PMA-GIBBS utilizes all available data while training. PMA-GIBBS-D, MLKNN-D,IBLRML-D and MLSVM-D discard data points with missing label entries in the training stage.

	OneError	AvePrec	Coverage
PMA-GIBBS	$\mathbf{29.7 \pm 4.2}$	$\mathbf{82.3 \pm 2.7}$	$\mathbf{10.6 \pm 2.3}$
PMA-GIBBS-D	51.1 ± 7.5	67.5 ± 4.8	22.4 ± 3.7
MLKNN-D	70.5 ± 2.0	46.3 ± 4.3	23.7 ± 8.1
IBLRML-D	61.9 ± 8.9	36.9 ± 3.9	54.8 ± 3.9
MLSVM-D	87.9 ± 2.0	40.9 ± 1.6	83.1 ± 1.4

5. Related Work

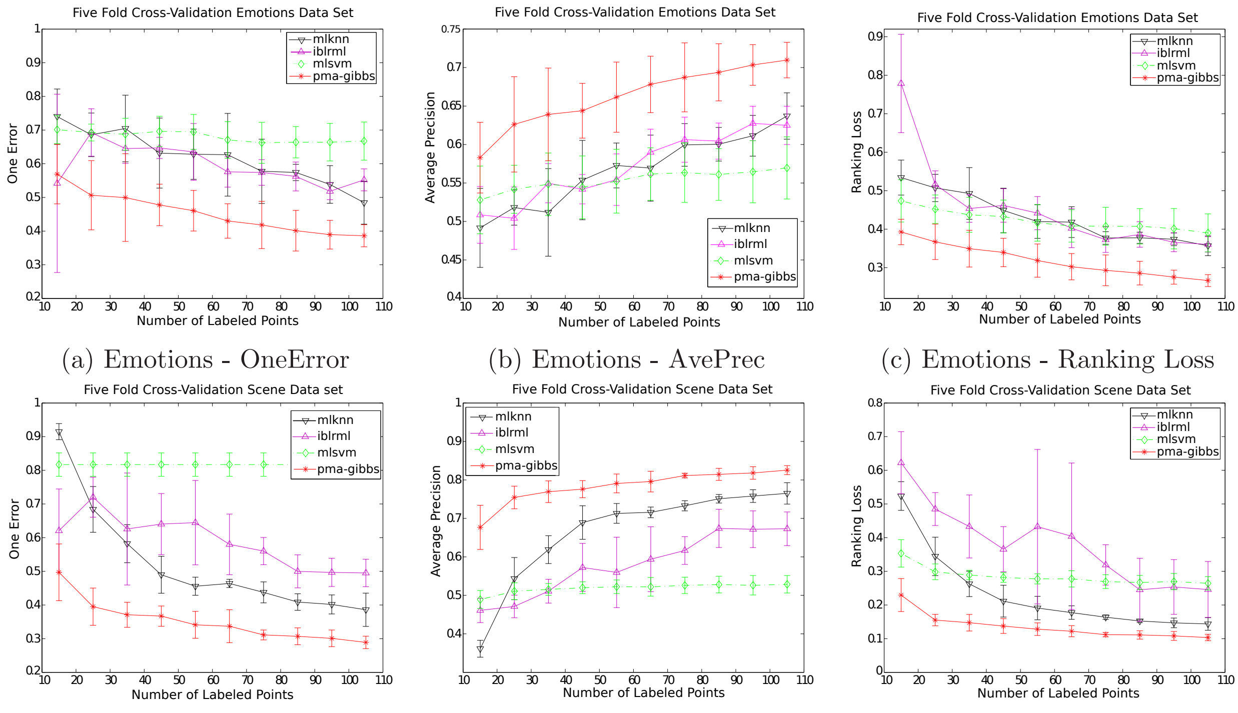
In this section, we briefly review related work in extending Gaussian Processes to capture correlated outputs, and related work in multi-label classification. A popular approach to capture correlations among outputs of a GP is to utilize Convolution Processes (CPs) (Boyle & Frean, 2005; Higdon, 2002). In CPs, each output is represented as the convolution of a smoothing kernel and a latent function, whereby the outputs are assumed to be drawn i.i.d. from a Gaussian Process. PMA does not make the i.i.d. assumption. (Chu et al., 2007) introduce a model capable of incorporating relational side information in the form of a graph, resulting in correlated outputs of a GP. Unlike PMA, this approach does not model correlations both across rows and columns explicitly. (Silva et al., 2007) propose a model which assumes latent functions to be the sum of two random variables, one of which contains relational side information. Unlike PMA, the resulting model can be represented by a single GP with a modified kernel, from which points/rows are drawn i.i.d. In (Xu et al., 2009) an approach is proposed that combines ideas from (Chu et al., 2007) and (Silva et al., 2007). Latent variables are assumed to be a sum of multiple random variables which encode relational information, whereby the aggregate latent variables are representable as outputs of a single GP (see discussion after (11) in (Xu et al., 2009)). Further, in (Xu et al., 2009), links are explicitly modeled.

The literature on multi-label classification has methods which attempt to capture correlation among labels. In (Zhang & Zhou, 2007), label statistics from neighborhoods are used to build a Bayesian classifier. In (Cheng & Hüllermeier, 2009), features are constructed based on label information from neighborhoods and subsequently used in logistic regression.

6. Conclusions

We have introduced a novel model for matrix data analysis capable of capturing correlations among rows and columns simultaneously. PMA has sparse statistical dependency structures yielding fast approximate inference algorithms. We have presented preliminary experiments demonstrating the advantage of PMA over single GPs for matrix analysis, as well as its ability to handle missing data. The ability of PMA to capture correlations along rows and columns simultaneously appears especially beneficial in domains such as multi-label classification. Our empirical evaluation shows that PMA can significantly outperform some of the existing multi-label classification algorithms. Further, PMA can readily be extended to higher order structures such as tensors which we plan to investigate in future work.

Acknowledgements: We warmly thank Daniel Boley for several helpful discussions. This research was supported by NSF grants IIS-0916750, IIS-0812183, IIS-1029711, NetSE-1017647, and NSF CAREER award IIS-0953274.



(a) Emotions - OneError

(b) Emotions - AvePrec

(c) Emotions - Ranking Loss

(d) Scene - OneError

(e) Scene - AvePrec

(f) Scene - Ranking Loss

Figure 3. Five fold cross validation on the Emotions and Scene data sets using three evaluation measures. PMA consistently outperforms the other methods on all datasets according to all evaluation measures.

References

Bonilla, E., Chai, K.M., and Williams, C. Multi-task Gaussian process prediction. In *NIPS*, 2008.

Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

Boyle, P. and Frean, M. Dependent Gaussian processes. In *NIPS*, 2005.

Cheng, W and Hüllermeier, E. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2-3):211–225, 2009.

Chu, W., Sindhwani, V., Ghahramani, Z., and Keerthi, S. Sathiya. Relational learning with Gaussian processes. In *NIPS*, 2007.

Gelfand, A. E. and Banerjee, S. Multivariate spatial process models. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (eds.), *Handbook of Spatial Statistics*. CRC Press, 2010.

Golub, G., Nash, S., and Van Loan, C. A Hessenberg-Schur method for the problem AX + XB= C. *IEEE Transactions on Automatic Control*, 24(6), 1979.

Higdon, D. M. Space and space-time modelling using process convolutions. In *Quantitative methods for current environmental issues*, pp. 37–56. Springer-Verlag, 2002.

Laub, A. J. *Matrix Analysis for Scientists and Engineers*. SIAM, 2005.

Mardia, K. V. and Goodall, C. R. Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*, pp. 347–386. Elsevier Science Publishers B.V., 1993.

Salakhutdinov, R. and Mnih, A. Probabilistic matrix factorization. In *NIPS*, 2007.

Silva, R., Chu, W., and Ghahramani, Z. Hidden common cause relations in relational learning. In *NIPS*, 2007.

Teh, Y. W. and Seeger, M. Semiparametric latent factor models. In *AISTATS*, 2005.

Trohidis, K., Tsoumakas, G., Kalliris, G., and Vlahavas, I. Multilabel classification of music into emotions. In *ISMIR*, 2008.

Wachspress, E. L. Iterative solution of the Lyapunov matrix equation. *Applied Mathematics Letters*, 1(1):87–90, 1988.

Wackernagel, H. *Multivariate Geostatistics: An Introduction With Applications*. Springer-Verlag Berlin, 2003.

Xu, Z., Kersting, K., and Tresp, V. Multi-relational learning with gaussian processes. In *IJCAI*, 2009.

Zhang, H. Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics*, pp. 125–139, 2007.

Zhang, M. and Zhou, Z. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40 (7):2038–2048, 2007.