Random Mixed Field Model for Mixed-Attribute Data Restoration

Qiang Li[†,⋆] and **Wei Bian**[†] and **Richard Yi Da Xu**[♯] and **Jane You**[⋆] and **Dacheng Tao**[†]
[†] Centre for Quantum Computation and Intelligent Systems, FEIT, University of Technology Sydney
[♯] School of Computing and Communications, FEIT, University of Technology Sydney
[⋆] Department of Computing, The Hong Kong Polytechnic University
leetsiang.cloud@gmail.com
{wei.bian, yida.xu, dacheng.tao}@uts.edu.au
csyjia@comp.polyu.edu.hk

Abstract

Noisy and incomplete data restoration is a critical preprocessing step in developing effective learning algorithms, which targets to reduce the effect of noise and missing values in data. By utilizing attribute correlations and/or instance similarities, various techniques have been developed for data denoising and imputation tasks. However, current existing data restoration methods are either specifically designed for a particular task, or incapable of dealing with mixed-attribute data. In this paper, we develop a new probabilistic model to provide a general and principled method for restoring mixed-attribute data. The main contributions of this study are twofold: a) a unified generative model, utilizing a generic random mixed field (RMF) prior, is designed to exploit mixed-attribute correlations; and b) a structured mean-field variational approach is proposed to solve the challenging inference problem of simultaneous denoising and imputation. We evaluate our method by classification experiments on both synthetic data and real benchmark datasets. Experiments demonstrate, our approach can effectively improve the classification accuracy of noisy and incomplete data by comparing with other data restoration methods.

Introduction

Real world data usually contain noise and missing values, which could severely degrade the performance of learning algorithms (Maimon and Rokach 2010; Chen et al. 2015; Wang and Oates 2015). The task of data restoration is to reduce the effect of noise and missing values, and plays a critical preprocessing step in developing effective learning algorithms. Attribute-level noise and missing values are two of the major concerns in data restoration. In the literature, attribute-level noise refers to undesirable incorrect measurements in some specific attribute of all instances. Different from the characteristics of noise, missing values are unavailable measurements. In practice, there are various reasons leading to noise and missing values, such as incaution or unwillingness in manual data entry process, equipment failure and high acquisition cost.

Data denoising targets to estimate the true value from noisy measurements based on certain assumptions. Unlike popular image denoising research (Buades, Coll, and Morel 2005b; 2005a), the research on attribute-level data denoising has long been limited. One possible reason is that images usually have strong local smoothness which can benefit denoising, but such local smoothness rarely exists in general data, such as survey reports. To achieve good local smoothness, (Li et al. 2002) proposed to use cluster labels to rearrange all the instances in Ecoli dataset (Lichman 2013). Then a wavelet shrinkage method is employed to filter certain attribute across instances. Although such "ad hoc" smoothness could make sense, it is intrinsically unfounded and prone to random instability. Different from deterministic approaches, probabilistic models provide a more rational approach for handling noisy data. The key assumption is the observed corrupted data are generated by adding random noise to latent noise-free data. Through such generative models, one can exploit informative priors over latent variables so as to encode attribute correlations.

Data imputation aims at providing good estimates of missing values. Deterministic approaches resort to modify classical regressors/classifiers to impute missing attributes. For example, K nearest neighbors imputation (KNNI) (Troyanskaya et al. 2001) imputes missing values with the mean/mode of K nearest neighbors for continuous/discrete attributes, as well as other techniques including local least squares imputation (LLSI) (Kim, Golub, and Park 2005), support vector machine imputation (SVMI) (Honghai et al. 2005), multiple kernel learning imputation (MKLI) (Zhu et al. 2011), and random forest imputation (Stekhoven and Bühlmann 2012). On the other hand, probabilistic latent variable models are employed to find the most probable imputation. For example, (Ghahramani and Jordan 1994) addressed the imputation problem by learning mixture models from an incomplete dataset. (Schneider 2001) designed a regularized expectation-maximization imputation (REMI) method by modelling the latent variables with multivariate Gaussian. Singular value decomposition imputation (SVDI) (Troyanskaya et al. 2001) combined principle component (PC) regression and EM estimation to estimate missing values. Some researchers also developed Bayesian principle component analysis based imputation (Oba et al. 2003), which jointly conducts PC regression, Bayesian estimation and EM learning.

Despite their effectiveness in exploiting attribute correlations and/or instance similarities, existing methods have two

main limitations: (1) they are specifically designed for a particular task, either denoising or imputation; (2) most of them are incapable of dealing with mixed-attribute data directly, and a prerequisite conversion step can inevitably cause information loss. In this paper, we formulate the mixed-attribute data restoration problem with a random mixed field (RMF) model. Moreover, to solve the resulting challenging inference problem, we derive a structured variational approach based on the mean field assumption. By exploiting mixed-attribute correlations, the proposed framework is capable of mixed-attribute data denoising and imputation at the same time.

Related Works

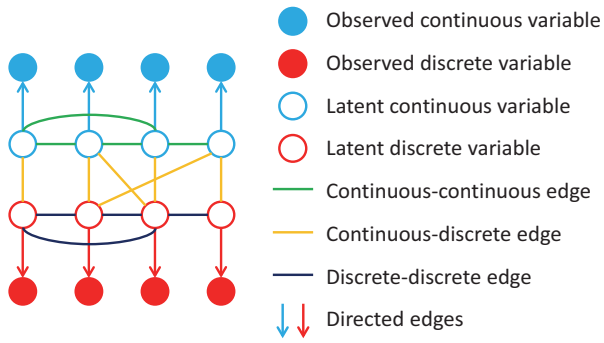
Recently, mixed graphical models have attracted increasing attentions (Lee and Hastie 2013; Cheng, Levina, and Zhu 2013; Yang et al. 2014) to meet the need for heterogeneous multivariate data modelling and analysis (Wang et al. 2015; Lian et al. 2015; Ding, Ming, and Fu 2014). In general, mixed graphical models extend classical graphical models by letting nodes to emerge from different types of random variables, such as continuous, discrete and count.

In the literature, mixed graphical models were first proposed in (Lauritzen and Wermuth 1989) to model mixed continuous and discrete variables. In this seminal work, the multinomial and conditional Gaussian distributions are used to represent the joint heterogeneous multivariate distribution. However, the number of model parameters scales exponentially with the number of discrete variables. To reduce the number of model parameters, (Lee and Hastie 2013) considered only pairwise interactions and fixed precision matrix for continuous variables. (Cheng, Levina, and Zhu 2013) further explored triple interactions between two discrete and one continuous variable. (Yang et al. 2014) considered mixed graphical models via a unified exponential family distribution to handle mixed continuous, discrete and count variables.

Though an RMF model belongs to general mixed graphical models, we propose to investigate the inference and parameter learning aspects of RMF model which is indeed complementary to the latest structure learning research. Specifically, 1) a structured mean field approach is derived to solve the inference problem of RMF model; 2) a variational expectation maximization algorithm is implemented to estimate the noise parameters given a fixed RMF prior.

Random Mixed Field Model

An RMF model is usually constructed by a hidden network playing the prior part and a corresponding set of observed nodes playing the likelihood part. See Figure 1 for a general example of RMF model. Note that, RMF model can be regarded as a specification of general mixed graphical models. In the following, we first describe the general framework of RMF model, and then give derivations of the inference algorithm. Parameter learning and data restoration algorithms will also be discussed. To simplify discussion, we will consider a fully-connected, pairwise, and continuous-discrete mixed graph in the next part of the paper.



(a) Mixed-net (b) Symbol explanations

Figure 1: An example of random mixed field model. (a) The hidden network is a "mixed-net" consisting of both continuous and discrete nodes. (b) explains all the four types of nodes and five types of edges.

Representation

Given a general mixed pairwise graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we have the vertex set $\mathcal{V} = \mathcal{V}_u \cup \mathcal{V}_v \cup \mathcal{V}_x \cup \mathcal{V}_y$ representing latent continuous/discrete, observed continuous/discrete variables, and the edge set $\mathcal{E} = \mathcal{E}_{uu} \cup \mathcal{E}_{vv} \cup \mathcal{E}_{uv} \cup \mathcal{E}_{ux} \cup \mathcal{E}_{vy}$ denoting the union of continuous-continuous, discrete-discrete, continuous-discrete connections and emissions. Consider the mixed-net example in Figure 1(a) which consists of four types of nodes and five types of edges. In detail, $\mathcal{V}_u$ and $\mathcal{V}_v$ are represented by cyan and red circles, $\mathcal{V}_x$ and $\mathcal{V}_y$ are denoted by cyan and red filled circles. On the other hand, $\mathcal{E}_{uu}$, $\mathcal{E}_{vv}$ and $\mathcal{E}_{uv}$ correspond to green, purple and yellow line segments; $\mathcal{E}_{ux}$ and $\mathcal{E}_{vy}$ correspond to cyan and red directed line segments.

An RMF model defines a joint distribution over the latent and observed variables according to some specific graphical configuration. In general, the joint distribution can be factorized into the prior and likelihood parts as below,

$$p(u, v, x, y|\Theta) = p(u, v|\Theta_p)p(x, y|u, v; \Theta_n), \qquad (1)$$

where $\Theta = \Theta_p \cup \Theta_n$ represents the union of prior and noise parameters.

The prior distribution is defined over latent variables via a Gaussian-Potts mixed potential,

$$p(u, v|\Theta_p) \propto \exp\left(\sum_{s=1}^{m}\sum_{t=1}^{m} -\frac{1}{2}\beta_{st}u_s u_t + \sum_{s=1}^{n}\alpha_s u_s + \sum_{s=1}^{m}\sum_{j=1}^{n}\rho_{sj}(v_j)u_s + \sum_{j=1}^{n}\sum_{k=1}^{n}\phi_{jk}(v_j, v_k)\right), \qquad (2)$$

where $\Theta_p = [\{\beta_{st}\}, \{\alpha_s\}, \{\rho_{sj}\}, \{\phi_{jk}\}]$ denotes the prior parameters. In particular, β_{st}, α_s, ρ_{sj} and ϕ_{jk} parameterizes continuous-continuous edge potential, continuous node potential, continuous-discrete edge potential, and discrete-discrete edge potential, respectively. Upon this mixed Gaussian-Potts prior distribution, we can also obtain node-wise conditional distributions for each variable. Specifically,

the conditional distribution of a continuous variable u_s given all its neighboring variables is a Gaussian distribution with a linear regression model for the mean and β_{ss}^{-1} being the unknown variance,

$$p(u_s|u_{\setminus s}, v) = \frac{\sqrt{\beta_{ss}}}{\sqrt{2\pi}} \exp(\zeta), \tag{3}$$

$$\zeta = \frac{-\beta_{ss}}{2} \left(u_s - \frac{\left(\alpha_s + \sum_j \rho_{sj}(v_j) - \sum_{t \neq s} \beta_{st} u_t\right)}{\beta_{ss}} \right)^2.$$

Note that, the backslash operator $\setminus$ is used to exclude variable s in defining the set of neighboring variables. The conditional distribution of a discrete variable v_j given its neighbors is a multinomial distribution with L_j states,

$$p(v_j|v_{\setminus j}, u) = \frac{\exp(\xi_{v_j})}{\sum_{l=1}^{L_j} \exp(\xi_l)}, \tag{4}$$

$$\xi_l = \left(\sum_s \rho_{sj}(l) u_s + \phi_{jj}(l,l) + \sum_{k \neq j} \phi_{jk}(l, v_k) \right).$$

The likelihood is defined based on the assumption that all observed variables are independent to each other conditioned on the latent variables,

$$p(x, y|u, v; \Theta_n) = \prod_{s=1}^{m} p(x_s|u_s) \prod_{j=1}^{n} p(y_j|v_j), \tag{5}$$

where $\Theta_n = [\{\sigma_s\}, \{\varphi_j\}]$ denotes the noise parameters of Gaussian and multinomial distributions. In other words, the continuous emission corresponds to additive white Gaussian noise (AWGN), and the discrete emission represents random flipping noise (RFN). Consequently, the distribution of x_s conditioned on u_s is modelled as a Gaussian with the noise parameter σ_s,

$$p(x_s|u_s) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{1}{2\sigma_s^2} (x_s - u_s)^2 \right). \tag{6}$$

And the distribution of y_j given v_j is modelled as a multinomial distribution parameterized by noise parameter φ_j,

$$p(y_j|v_j) = \frac{\exp(\varphi_j(y_j, v_j))}{\sum_{l=1}^{L_j} \exp(\varphi_j(l, v_j))}. \tag{7}$$

Structured Mean Field

With an RMF model, data restoration can be achieved by the inference over posterior distribution $p(u, v|x, y; \Theta)$. Since the calculation of the likelihood $p(x, y; \Theta)$ is intractable, we seek to approximate inference approaches. Specifically, we use the variational approach, which is considered to be more efficient than sampling methods. Based on the mean field assumption, the optimal variational approximation of $p(u, v|x, y; \Theta)$ is given by

$$q^*(u, v) = \arg \min_{\substack{q(u,v)= \\ q(u)q(v)}} \mathrm{KL}[q(u, v) \| p(u, v|x, y; \Theta)]. \tag{8}$$

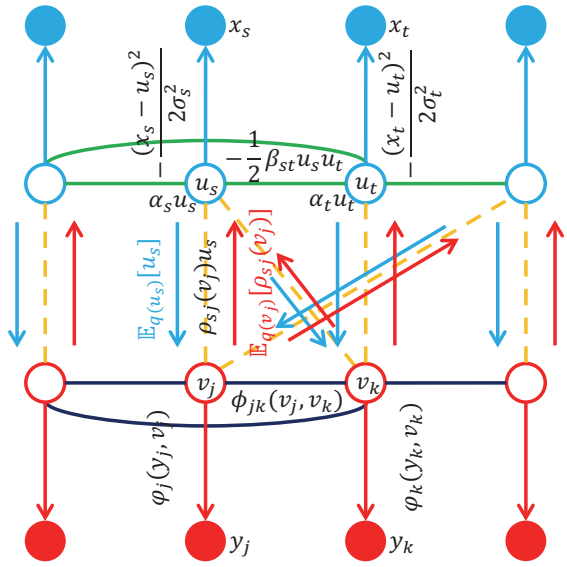


Figure 2: The proposed structured mean field approximation can be regarded as cutting off those mixed-type edges and absorbing the interactions in the form of expected sufficient statistics, i.e., $\mathbb{E}_{q(u_s)}[u_s]$ and $\mathbb{E}_{q(v_j)}[\rho_{sj}(v_j)]$, respectively. Such a posterior approximation will result in two separate subgraphs, which are much easier to handle. In addition, it is required to alternately update each of the two subgraphs' joint distributions until convergence.

The minimization of the Kullback-Leibler divergence in (8) can be achieved by maximizing a lower bound,

$$\mathcal{L}(q) = \mathbb{E}_{q(u)q(v)} \left[\ln \frac{p(x, y, u, v)}{q(u)q(v)} \right] \tag{9}$$

of the log evidence $\ln p(x, y) = \ln \sum_v \int_u p(x, y, u, v)$ w.r.t. $q(u)$ and $q(v)$, respectively. Accordingly, the update formula for $q(u)$ and $q(v)$ are given by (Bishop 2006),

$$q(u) \leftarrow \frac{1}{Z_u} \exp \mathbb{E}_{q(v)}[\ln p(u, v, x, y)] \tag{10}$$

$$q(v) \leftarrow \frac{1}{Z_v} \exp \mathbb{E}_{q(u)}[\ln p(u, v, x, y)], \tag{11}$$

where $\mathbb{E}_p[f]$ calculates the expectation of function f w.r.t. distribution p, and Z_u and Z_v are the normalization terms.

To solve Eqn. (10) for updating $q(u)$, we evaluate the expectation w.r.t. $q(v)$,

$$\begin{aligned}
\mathbb{E}_{q(v)}[\ln p(u, v, x, y)] &\equiv -\frac{1}{2} \sum_{s=1}^{m} \sum_{t=1}^{m} \beta_{st} u_s u_t \\
&+ \sum_{s=1}^{m} \alpha_s u_s + \sum_{s=1}^{m} \sum_{j=1}^{n} \mathbb{E}_{q(v_j)}[\rho_{sj}(v_j)] u_s \\
&+ \sum_{s=1}^{m} \left[-\frac{(x_s - u_s)^2}{2\sigma_s^2} \right] + \sum_{j=1}^{n} \mathbb{E}_{q(v_j)}[\varphi_j(y_j, v_j)] \\
&\equiv -\frac{1}{2} u^T \hat{B} u + \hat{\gamma}(v, x)^T u,
\end{aligned} \tag{12}$$

where the notation $\equiv$ denotes the two terms on the left and right hand sides are equivalent up to a constant, and $B = \{\beta_{st}\}$, $\hat{B} = B + \text{diag}\{\frac{1}{\sigma_s^2}\}$, $\{\hat{\gamma}(v,x)\}_s = \alpha_s + \sum_j \mathbb{E}_{q(v_j)}[\rho_{sj}(v_j)] + \frac{x_s}{\sigma_s^2}$. Fortunately, $q(u)$ follows a multivariate Gaussian distribution, $q(u) = \mathcal{N}(u|\hat{B}^{-1}\hat{\gamma}(v,x), \hat{B}^{-1})$. Notice that, for the Gaussian-Potts model defined in Eqn. (2), we do not need to calculate the inverse of the updated precision matrix $\hat{B}$. The reason is that the continuous-discrete edge potentials are only absorbed into the first-order term of $q(u)$. In addition, the noise term $\text{diag}\{\frac{1}{\sigma_s^2}\}$ is added to the diagonal of B thus does not affect the original graphical connections defined in B. Consequently, algorithms such as Gauss elimination and GaBP (Bickson 2008) can be employed to efficiently infer the mean $\hat{B}^{-1}\hat{\gamma}(v,x)$ when B is sparse.

Regarding Eqn. (11), we have the expectation w.r.t. $q(u)$,

$$\begin{aligned}
&\mathbb{E}_{q(u)}[\ln p(u,v,x,y)] \\
&\equiv \sum_{j=1}^{n}\sum_{k=1}^{n}\phi_{jk}(v_j,v_k) + \sum_{j=1}^{n}\sum_{s=1}^{m}\rho_{sj}(v_j)\mathbb{E}_{q(u_s)}[u_s] \\
&+ \sum_{s=1}^{m}\mathbb{E}_{q(u_s)}\left[-\frac{(x_s-u_s)^2}{2\sigma_s^2}\right] + \sum_{j=1}^{n}\varphi_j(y_j,v_j) \\
&\equiv \sum_{j=1}^{n}\sum_{k=1}^{n}\phi_{jk}(v_j,v_k) + \sum_{j=1}^{n}\hat{\varphi}_j(y_j,v_j,u), \qquad (13)
\end{aligned}$$

where $\hat{\varphi}_j(y_j,v_j,u) = \sum_s \rho_{sj}(v_j)\mathbb{E}_{q(u_s)}[u_s] + \varphi_j(y_j,v_j)$. In other words, $q(v)$ follows a pairwise discrete MRF, $q(v) \propto \exp\left\{\sum_j\sum_k \phi_{jk}(v_j,v_k) + \sum_j \hat{\varphi}_j(y_j,v_j,u)\right\}$. Note that, for the Gaussian-Potts model defined in Eqn. (2), those interaction and emission terms $\{\hat{\varphi}_j\}$ do not affect the original graphical connection defined in $\{\phi_{jk}\}$. Thus, we can use the loopy belief propagation (Murphy, Weiss, and Jordan 1999) algorithm to solve the pairwise discrete MRF inference problem.

In the above derivations, the mixed-type edge terms appeared in both $q(u)$ and $q(v)$ but in different forms. Figure 2 illustrates the process of formulating two interacting subgraphs via $\mathbb{E}_{q(u_s)}[u_s]$ and $\mathbb{E}_{q(v_j)}[\rho_{sj}(v_j)]$ when the hidden network is a mixed-net. The alternative updating between $q(u)$ and $q(v)$ is performed until convergence to a stationary point. The stationary point corresponds to two completely independent subgraphs that jointly approximate the whole mixed graph.

Parameter Estimation

For the data restoration task considered in this paper, we follow the setting of a clean training dataset and a corrupted testing dataset (Saar-Tsechansky and Provost 2007). Regarding "clean", we mean the samples are noise-free and complete and "corrupted" means the samples contain noise and missing values. According to this setting, the RMF prior parameters Θ_p can be learned from the clean training dataset. Fortunately, several third-party learning techniques, such as, variants of graphical LASSO (Friedman, Hastie, and Tibshirani 2008), ℓ_1 regularized pseudo-likelihood (Besag 1974; Lee and Hastie 2013) and ℓ_1 regularized node-wise regression (Yang et al. 2012), can be utilized to learn this generic prior. [1]

When restoring the corrupted testing dataset, domain knowledge can be employed to yield a good estimate of the noise parameters Θ_n. If unfortunately this method fails, a variational EM algorithm can be adopted to estimate noise parameters given all testing data and the generic RMF prior. In general, given N i.i.d. observation samples $X = \{x^{(i)}\}$ and $Y = \{y^{(i)}\}$, the variational EM algorithm iterates between variational inference (E-step) and parameter estimation (M-step). Since the RMF prior parameters are fixed after learning from the training dataset, we can only iteratively infer $q(u,v)$ and estimate Θ_n on the testing dataset until convergence. The corresponding noise parameter estimation in M-step is achieved by taking derivatives of the objective function $\mathcal{Q}(\Theta)$ w.r.t. σ_s^2 and $\varphi_j(a,b)$ respectively,

$$\begin{aligned}
\sigma_s^2 \leftarrow &\frac{1}{N}\sum_i (x_s^{(i)})^2 - \frac{2}{N}\sum_i x_s^{(i)}\mathbb{E}_{q(u_s^{(i)})}[u_s^{(i)}] \\
&+ \frac{1}{N}\sum_i \mathbb{E}_{q(u_s^{(i)})}[(u_s^{(i)})^2]; \qquad (14)
\end{aligned}$$

$$\frac{\exp(\varphi_j(a,b))}{\sum_l \exp(\varphi_j(l,b))} \leftarrow \frac{\sum_i \mathbb{I}(y_j^{(i)}=a)q(v_j^{(i)}=b)}{\sum_i q(v_j^{(i)}=b)}. \qquad (15)$$

Note that we have made explicit the testing sample index i for clarity.

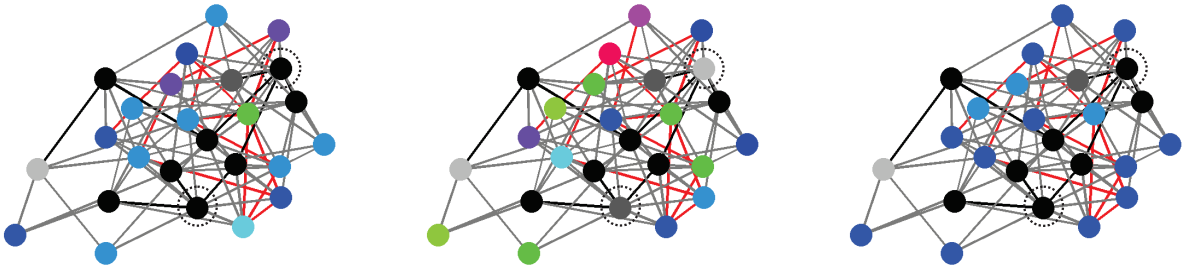
So far, it seems that our derivation only considers noise. However, it is very straightforward to modify the proposed framework to handle missing values. Consider some of the observed variables of sample i are missing, say $x_m^{(i)}$ and $y_m^{(i)}$, we can simply delete those $p(x_m^{(i)}|u_m^{(i)})$ and $p(y_m^{(i)}|v_m^{(i)})$ terms, and keep all the other terms totally unchanged. Then the proposed variational inference procedure is modified accordingly. It is worth mentioning that our framework can resort to the generic RMF prior even when heavy missingness occurs. Thus, the simple deletion strategy is also applicable when the missing values become prevalent.

Evaluation on Synthetic Data

We design a simulation study to show that mixed-attribute correlations can effectively help reduce noise effects and improve classification performance. Consider a mixed-net graph consisting of 15 continuous and 10 discrete nodes with correlation parameters defined as below,

$$\begin{aligned}
&\alpha_s = 1, \beta_{ss} = 1, \forall s \in \mathcal{V}_u, \beta_{st} = 4, \forall st \in \mathcal{E}_{uu}; \\
&\rho_{sj} = [3\ 2\ 1], \forall sj \in \mathcal{E}_{uv}; \\
&\phi_{jj} = 0, \forall j \in \mathcal{V}_v, \phi_{jk} = \begin{bmatrix} 1.5 & 0.5 & 0.5 \\ 0.5 & 1.5 & 0.5 \\ 0.5 & 0.5 & 1.5 \end{bmatrix}, \forall jk \in \mathcal{E}_{vv}.
\end{aligned}$$

[1] Although these techniques are originally designed for structure learning, the resulting sparsified parameters will not only indicate the graphical structure, but also provide a good parameterization of the generic prior.



(a) Ground truth (b) Noisy observation (c) Denoised result

Figure 3: The mixed-net graph used in our simulation contains 15 continuous (HSV-colored) and 10 discrete (grey-colored) nodes. The nodes are colored according to attribute values of a representative example. The three types of edges (continuous-continuous in red, discrete-discrete in black and continuous-discrete in light grey) are randomly chosen from all possible edges.

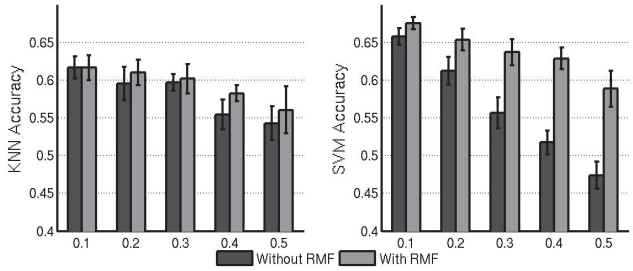


Figure 4: KNN (left plot) and SVM (right plot) classification accuracies of noisy (black) and denoised (light grey) data under different levels of random noise (the noise strength τ ranges from 0.1 to 0.5). Each bar represents the mean and standard deviation of 10 independent experiments.

In addition, we formulate two classes by adding two small but different random numbers ($\delta_1, \delta_2 \in [-0.5, 0.5]$) to all elements of ρ_{sj}. According to this setting, we generate 750 random examples for each class. Then we split all the examples into training and testing sets with a ratio of $2:1$. The training set is utilized to train RMF model and KNN, SVM classifiers. And the testing set is used to generate noisy testing sets by injecting different levels of AWGN to continuous attributes and RFN to discrete attributes. In addition, the corruption strength is defined at five different percentages, i.e., $\tau = 0.1, 0.2, 0.3, 0.4, 0.5$. For continuous attributes, the noise standard deviations are $\sigma_s = \tau\breve{\sigma}_s$, $s = 1, 2, \ldots, m$, with $\breve{\sigma}_s$ being the signal standard deviations. For discrete attributes, the flipping probabilities are formulated as $p(y_j \neq a | v_j = a) = \tau$, $j = 1, 2, \ldots, n$.

Figure 3 illustrates the synthetic mixed-net graph structure and a representative example. We observe that the colors of these denoised continuous nodes are much closer to ground truth than noisy observation. In addition, the observed wrong state values of the discrete variables (in dotted circles) are also corrected after applying our inference algorithm. Besides the qualitative result, we also conduct quantitative classification experiment and summarize the results in Figure 4. According to the error bars, RMF improves the performance of classification significantly.

Evaluation on Real Data

In this section, we present experimental results on four real-world mixed-attribute datasets from the UCI machine learning repository (Lichman 2013), which are "Adult", "Credit", "Statlog-Australian" and "Statlog-German". The "Adult" dataset has already been split into train/test in approximately $2/3$, $1/3$ proportions. As for the "Credit", "Statlog-Australian" and "Statlog-German" datasets, we simply select the first $2/3$ proportion of all the instances as the training set and the remaining as the testing set.

Furthermore, to specifically consider the effect of all comparison methods on handling noise/missingness at testing stage, the experimental setting is clean training data versus corrupted testing data. The same methodology has also been widely employed in the literature, for example (Saar-Tsechansky and Provost 2007; Dekel, Shamir, and Xiao 2010; Maaten et al. 2013). Consequently, all models and classifiers are built using clean training data and applied to handle corrupted testing data.

Except where no corruption is applied, each reported result is the average classification accuracy over 10 independent experiments in which random noise and missingness are injected into the testing data. More importantly, all comparison methods are carried out on the same random noisy or incomplete testing data.

Data Denoising

For data denoising task, we employ the same noisy data generation strategy used in previous simulation study. More specifically, five different levels of noise strength ($\tau = 0.1, 0.2, 0.3, 0.4, 0.5$) are applied to all the four UCI datasets. Table 1 presents the classification accuracies of standard classifiers, before and after applying RMF denoising. As expected, the classification accuracy decreases as the noise strength increases compared to noise-free data classification. On the other hand, for most cases, the classification accuracies are effectively improved after RMF denoising. In addition, SVM classifier is more sensitive to noise than KNN classifier as the performance drops faster. In fact, SVM makes predictions using pre-trained fixed hyperplane weights while KNN is a lazy learner which can make adjustments for new instances.

Table 1: Classification Accuracies with/without Data Denoising.

τ	0	0.1	0.2	0.3	0.4	0.5	0	0.1	0.2	0.3	0.4	0.5
Method	Adult						Credit					
KNN	0.8248	**0.8174**	0.8063	0.7944	0.7821	0.7694	0.8227	**0.7727**	0.7191	0.6605	0.6105	0.5273
RMF+KNN		**0.8174**	**0.8083**	**0.7967**	**0.7865**	**0.7748**		0.7700	**0.7386**	**0.6936**	**0.6764**	**0.6455**
SVM	0.8467	0.8356	0.8243	0.8084	0.7951	0.7817	0.8636	**0.7895**	0.7200	0.6455	0.5900	0.4968
RMF+SVM		**0.8413**	**0.8317**	**0.8186**	**0.8053**	**0.7920**		**0.7895**	**0.7382**	**0.6859**	**0.6664**	**0.6336**
Method	Statlog-Australian						Statlog-German					
KNN	0.8783	0.8052	0.7274	0.6617	0.6148	0.5526	0.7417	**0.7189**	0.6895	0.6700	0.6703	0.6474
RMF+KNN		**0.8091**	**0.7613**	**0.7396**	**0.7074**	**0.6635**		0.7147	**0.6970**	**0.6880**	**0.6757**	**0.6655**
SVM	0.8478	0.7791	0.6948	0.6422	0.5752	0.4965	0.7688	0.7486	0.7204	0.6955	0.6778	0.6580
RMF+SVM		**0.7974**	**0.7661**	**0.7361**	**0.7078**	**0.6657**		**0.7523**	**0.7411**	**0.7210**	**0.7270**	**0.7096**

Table 2: Classification Accuracies with Noisy Data Imputation.

ρ	0	0.1	0.2	0.3	0.4	0.5	0	0.1	0.2	0.3	0.4	0.5
Method	Adult						Credit					
KNNI+KNN	0.8063	0.8018	0.7814	0.7543	0.7075	0.3577	0.7191	0.7114	0.6859	0.6377	0.5673	0.4645
REMI+KNN	0.8063	**0.8043**	**0.7918**	**0.7683**	0.7333	0.6929	0.7191	0.7136	0.7045	0.6705	**0.6441**	0.6005
RMFI+KNN	**0.8083**	0.7997	0.7794	0.7631	**0.7553**	**0.7543**	**0.7386**	**0.7245**	**0.7041**	**0.6732**	0.6355	**0.6150**
KNNI+SVM	0.8243	0.8144	0.7944	0.7703	0.7236	0.4947	0.7200	0.7095	0.6764	0.6300	0.5477	0.4555
REMI+SVM	0.8243	0.8190	**0.8088**	**0.7889**	0.7574	0.7032	0.7200	0.7150	0.6995	0.6668	0.6291	0.5327
RMFI+SVM	**0.8317**	**0.8222**	0.7972	0.7743	**0.7589**	**0.7547**	**0.7382**	**0.7205**	**0.7091**	**0.6795**	**0.6377**	**0.6155**
Method	Statlog-Australian						Statlog-German					
KNNI+KNN	0.7274	0.7261	0.6952	0.6470	0.6000	0.5696	0.6895	0.6952	0.6913	0.6550	0.6057	0.5637
REMI+KNN	0.7274	0.7317	**0.7222**	0.6917	**0.6596**	**0.6000**	0.6895	**0.7030**	0.6874	**0.6784**	**0.6604**	0.5817
RMFI+KNN	**0.7613**	**0.7635**	0.7217	**0.6926**	0.6470	0.5857	**0.6970**	0.6913	**0.6946**	0.6712	0.6465	**0.6141**
KNNI+SVM	0.6948	0.7022	0.6800	0.6474	0.5896	0.4917	0.7204	0.7282	0.7015	0.6895	0.5844	0.4147
REMI+SVM	0.6948	0.7070	0.7017	**0.7013**	**0.6600**	0.5770	0.7204	**0.7327**	0.6994	0.6976	0.6838	0.6793
RMFI+SVM	**0.7661**	**0.7700**	**0.7317**	0.7009	0.6561	**0.5857**	**0.7411**	0.7312	**0.7144**	**0.7129**	**0.7072**	**0.7081**

Noisy Data Imputation

We further evaluate RMF's capability on the task of data imputation under noise. A little different from previous setting, the corrupted testing data are generated by first injecting noise ($\tau = 0.2$), then adding different levels of missingness. Missing completely at random (MCAR) strategy is employed to randomly annihilate a percentage ($\rho = 0.1, 0.2, 0.3, 0.4, 0.5$) of continuous and discrete attributes of each instance in the testing data. Table 2 compares classification accuracies of standard classifiers utilizing KNN imputation (KNNI), regularized-EM imputation (REMI) and the proposed RMF imputation (RMFI) techniques. According to the experimental results, RMFI obtained better performance with other imputation methods. Note that the proposed RMFI framework is capable of reducing noise effect during imputation, thus RMFI is very suitable for the noisy data imputation task.

Note that, KNNI imputes missing values with the mean/mode of K nearest neighbors for continuous/discrete attributes, while REMI is to impute missing values via a regularized expectation-maximization iterative procedure. In all our experiments, we employ the KNNI implementation, "knnimpute.m", from Matlab's Bioinformatics toolbox. For KNNI settings, we choose $K = 3$ and use weighted Euclidean distance measure, which is also suggested by the authors (Troyanskaya et al. 2001). The REMI source code is available at the author's homepage http://www.clidyn.ethz.ch/imputation/, and the default setting is used. Before applying the KNNI and REMI methods, we first transform those nominal attributes into dummy variables. Then the imputed testing data are post-processed to satisfy the constraint that the dummy vector of each nominal attribute should contain exactly one numerical value "1". It is worth mentioning that we have also tried other methodologies, such as specially impute nominal attributes with mode, but obtained no better results than the above one.

Conclusions

Data restoration is common and critical for real-world data analysis practice. Although major problems, e.g, data denoising and imputation, have been widely studied in the literature, there still lacks a principled approach that is able to dress the generic data restoration problem. The proposed RMF model reduces this gap, by providing a principled approach to jointly handle data denoising and imputation within the probabilistic graphical model scope. An efficient inference algorithm for the RMF model was derived based on a structured variational approach. Empirical evaluations confirmed the effectiveness of RMF and showed its competitiveness by comparing with other data restoration methods.

Acknowledgement

This research is supported by the Chancellor's Postdoctoral Research Fellowship of University of Technology Sydney, and Australian Research Council Projects (No: DP-140102164 and FT-130101457).

References

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* 192–236.

Bickson, D. 2008. Gaussian belief propagation: Theory and aplication. *arXiv preprint arXiv:0811.2518*.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. Springer.

Buades, A.; Coll, B.; and Morel, J.-M. 2005a. A non-local algorithm for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 60–65. IEEE.

Buades, A.; Coll, B.; and Morel, J.-M. 2005b. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation* 4(2):490–530.

Chen, X. C.; Faghmous, J. H.; Khandelwal, A.; and Kumar, V. 2015. Clustering dynamic spatio-temporal patterns in the presence of noise and missing data. In *International Joint Conference on Artificial Intelligence*, 2575–2581.

Cheng, J.; Levina, E.; and Zhu, J. 2013. High-dimensional mixed graphical models. *arXiv preprint arXiv:1304.2810*.

Dekel, O.; Shamir, O.; and Xiao, L. 2010. Learning to classify with missing and corrupted features. *Machine learning* 81(2):149–178.

Ding, Z.; Ming, S.; and Fu, Y. 2014. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI Conference on Artificial Intelligence*, 1192–1198.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441.

Ghahramani, Z., and Jordan, M. I. 1994. Supervised learning from incomplete data via an em approach. In *Advances in Neural Information Processing Systems*, 120–127.

Honghai, F.; Guoshun, C.; Cheng, Y.; Bingru, Y.; and Yumei, C. 2005. A svm regression based approach to filling in missing values. In *Knowledge-Based Intelligent Information and Engineering Systems*, 581–587. Springer.

Kim, H.; Golub, G. H.; and Park, H. 2005. Missing value estimation for dna microarray gene expression data: local least squares imputation. *Bioinformatics* 21(2):187–198.

Lauritzen, S. L., and Wermuth, N. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics* 31–57.

Lee, J., and Hastie, T. 2013. Structure learning of mixed graphical models. In *International Conference on Artificial Intelligence and Statistics*, 388–396.

Li, Q.; Li, T.; Zhu, S.; and Kambhamettu, C. 2002. Improving medical/biological data classification performance by wavelet preprocessing. In *IEEE International Conference on Data Mining*, 657–660. IEEE.

Lian, W.; Rai, P.; Salazar, E.; and Carin, L. 2015. Integrating features and similarities: Flexible models for heterogeneous multiview data. In *AAAI Conference on Artificial Intelligence*, 2757–2763.

Lichman, M. 2013. UCI machine learning repository.

Maaten, L.; Chen, M.; Tyree, S.; and Weinberger, K. Q. 2013. Learning with marginalized corrupted features. In *International Conference on Machine Learning*, 410–418.

Maimon, O., and Rokach, L. 2010. Data mining and knowledge discovery handbook.

Murphy, K. P.; Weiss, Y.; and Jordan, M. I. 1999. Loopy belief propagation for approximate inference: An empirical study. In *The Conference on Uncertainty in Artificial Intelligence*, 467–475. Morgan Kaufmann Publishers Inc.

Oba, S.; Sato, M.-a.; Takemasa, I.; Monden, M.; Matsubara, K.-i.; and Ishii, S. 2003. A bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19(16):2088–2096.

Saar-Tsechansky, M., and Provost, F. 2007. Handling missing values when applying classification models. *The Journal of Machine Learning Research* 8:1625–1657.

Schneider, T. 2001. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* 14(5):853–871.

Stekhoven, D. J., and Bühlmann, P. 2012. Missforest-nonparametric missing value imputation for mixed-type data. *Bioinformatics* 28(1):112–118.

Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; and Altman, R. B. 2001. Missing value estimation methods for dna microarrays. *Bioinformatics* 17(6):520–525.

Wang, Z., and Oates, T. 2015. Imaging time-series to improve classification and imputation. In *International Joint Conference on Artificial Intelligence*, 3939–3945.

Wang, C.; Chi, C.-H.; Zhou, W.; and Wong, R. 2015. Coupled interdependent attribute analysis on mixed data. In *AAAI Conference on Artificial Intelligence*, 1861–1867.

Yang, E.; Allen, G.; Liu, Z.; and Ravikumar, P. K. 2012. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, 1358–1366.

Yang, E.; Baker, Y.; Ravikumar, P.; Allen, G.; and Liu, Z. 2014. Mixed graphical models via exponential families. In *International Conference on Artificial Intelligence and Statistics*, 1042–1050.

Zhu, X.; Zhang, S.; Jin, Z.; Zhang, Z.; and Xu, Z. 2011. Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering* 23(1):110–121.