Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

A Bayesian Factorised Covariance Model for Image Analysis*

Jun Li and Dacheng Tao
Centre for Quantum Computation & Intelligent Systems
University of Technology, Sydney, Australia
{Jun.Li, Dacheng.Tao}@uts.edu.au

Abstract

This paper presents a specialised Bayesian model for analysing the covariance of data that are observed in the form of matrices, which is particularly suitable for images. Compared to existing general-purpose covariance learning techniques, we exploit the fact that the variables are organised as an array with two sets of ordered indexes, which induces innate relationship between the variables. Specifically, we adopt a factorised structure for the covariance matrix. The covariance of two variables is represented by the product of the covariance of the two corresponding rows and that of the two columns. The factors, i.e. the row-wise and column-wise covariance matrices are estimated by Bayesian inference with sparse priors.

Empirical study has been conducted on image analysis. The model first learns correlations between the rows and columns in an image plane. Then the correlations between individual pixels can be inferred by their locations. This scheme utilises the structural information of an image, and benefits the analysis when the data are damaged or insufficient.

1 Introduction

Statistical learning is about inducing trend from data. Given observations of a set of attributes, trend discovery means to learn the relationship between those interested attributes. General relationship is an elusive notion, which may involve attribute groups of any size and can manifest themselves in arbitrary patterns in the observed data. Nevertheless, one can learn rich information about the population of the data from the knowledge about how attribute pairs vary in a coordinated manner. That is, given instances of a set of variables, a useful analysis is to estimate the covariance matrix of the variables.

Learning the covariance is theoretically straightforward but practically difficult. The sample covariance matrix asymptotically approximates the population covariance. But when the number of variables, p, is large compared with the sample size N, the estimation is unreliable [Marcenko and Pastur, 1967]. This is hardly surprising. A non-degenerated population of p variables has a covariance matrix lying on a Stiefel manifold characterised by $p(p-1)/2$ parameters. In the context of modern high-dimensional data analysis, even for a relatively moderate p, the N required for the naive covariance estimator to work properly could be prohibitively large. Therefore, given limited observations, it is necessary to exploit side information about the data population to obtain a reliable estimate of covariance.

In this paper, we are concerned with a special setting for the covariance estimating problem, where the variables are organised as a matrix. In particular, we propose a Bayesian model, where the covariance between two variables factorises into two parts: (i) the covariance between the two corresponding rows of the matrix and (ii) that of the two columns. The row- and column-wise covariance matrices are considered as random variables and learned by Bayesian inference. Therefore, we employ the prior knowledge about the matrix structure of the variables as the side information regularising the learning, and use the name of *Bayesian factorised covariance model* (BFCM). The inference in BFCM is more manageable than directly learning the original covariance matrix, which is of much higher dimension than either of the two factor covariance matrices. The Bayesian treatment, i.e. prior distribution of the factor covariance, provides additional regularisation.

BFCM is particularly suited to images. Generally, direct estimate of pixel-to-pixel covariance is difficult, because practical images often contain a large number of pixels. The hypothesis of BFCM states that the correlation between two pixels is induced by their respective positions in the image, i.e.

$$\begin{aligned}&\mathrm{Cov}\left[\mathrm{Pixel}(i_1, j_1), \mathrm{Pixel}(i_2, j_2)\right]\\ =&\mathrm{Cov}\left[\mathrm{Row}(i_1), \mathrm{Row}(i_2)\right] \times \mathrm{Cov}\left[\mathrm{Col}(j_1), \mathrm{Col}(j_2)\right].\end{aligned}$$

The hypothesis is sensible, because a category of images is often characterised by the relations between the rows and those between the columns in those images.

More generally, the proposed model benefits covariance estimation for matrix data in two aspects. First, the factorisation structure significantly reduces the degree of freedom in the covariance estimator and improves its scalability. Second, the factorisation structure makes the covariance estimator more robust. For example, if the information source is faulty, e.g. due to a damaged sensor, some variables can be missing in *all* observations, it would be infeasible to estimate element-wise

*This work was supported by the Australian Research Council under Discovery Project ARC DP-120103730.

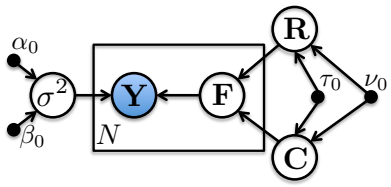

Figure 1: Graphical illustration of BFCM.

covariance, but relations between rows and columns can still be inferred.

2 Bayesian factorised covariance model

This section specifies the definition and inference of BFCM. Frequently used denotations are listed as follows. The symbol $\otimes$ represents the Kronecker product of two matrices, operator $\text{vec}\,(\cdot)$ transforms a matrix to a column vector by stacking the columns of the matrix, $\text{vec}\,(\mathbf{A}) = [a_{11}\dots a_{m1},\dots,a_{1n}\dots a_{mn}]^T$, and function $\text{etr}[\mathbf{A}] := \exp\{\text{tr}(\mathbf{A})\}$, where $\text{tr}(\cdot)$ is matrix trace. The formulations of standard probability density functions follow those in [Gelman *et al.*, 2003].

2.1 Model Definition

BFCM focuses on the relations between the entries in a random matrix. Fig. 1 shows a graphical illustration of the probabilistic structure of BFCM, and the Bayesian model is defined as follows. The data are N samples of random matrices. We consider an observed matrix $\mathbf{Y}^{(n)} \in \mathbb{R}^{n_r\times n_c}$ to be consisting of an underlying pattern $\mathbf{F}^{(n)} \in \mathbb{R}^{n_r\times n_c}$ and independent Gaussian noises

$$p(y_{ij}|f_{ij},\sigma^2) = \mathcal{N}(y_{ij}|f_{ij},\sigma^2), \tag{1}$$

for $i = 1\dots n_r$ and $j = 1\dots n_c$, where σ^2 represents the variance of the noise. Note that the sample index $^{(n)}$ will be omitted when there is no ambiguity. Then the covariance between the elements of $\mathbf{F}$ is of interest. According to BFCM, the covariance has a factorised form

$$\text{cov}\left[f_{i_1j_2}, f_{i_2j_2}\right] = \text{cov}\left[\Omega^r_{i1},\Omega^r_{i2}\right] \times \text{cov}\left[\Omega^c_{j1},\Omega^c_{j2}\right]. \tag{2}$$

According to the discussion in Sec. 1, Ω^r_i and Ω^c_j denote the indexing sets of the rows and the columns. Note that these indexes only represent conceptual objects as place holders in $\text{cov}\,[\cdot,\cdot]$ function to give rise to their respective covariance matrices. The objects themselves are not of interest, and will have no significance in the following development of BFCM. Let the row- and column-covariance matrices be $\mathbf{R}$ and $\mathbf{C}$, respectively:

$$\mathbf{R}:\ R_{i_1i_2} = \text{cov}\left[\Omega^r_{i_1},\Omega^r_{i_2}\right],$$
$$\mathbf{C}:\ C_{j_1j_2} = \text{cov}\left[\Omega^c_{j_1},\Omega^c_{j_2}\right].$$

Then the factorisation (2) states that the element-wise covariance of $\mathbf{F}$, $\text{cov}\left[\text{vec}\,(\mathbf{F}),\text{vec}\,(\mathbf{F})\right]$, is the Kronecker product of $\mathbf{R}$ and $\mathbf{C}$. Given $\mathbf{R}$ and $\mathbf{C}$, the prior of $\mathbf{F}$ can be specified by the following Gaussian distribution

$$p(\text{vec}\,(\mathbf{F})\,|\mathbf{R},\mathbf{C}) = \mathcal{N}(\text{vec}\,(\mathbf{F})\,|0,\mathbf{C}\otimes\mathbf{R}). \tag{3}$$

An equivalent form of (3) is a matrix variate normal distribution [Gupta and Nagar, 2002],

$$p(\mathbf{F}|\mathbf{R},\mathbf{C}) = (2\pi)^{-n_cn_r/2} \times\det(\mathbf{R})^{-n_c/2}\det(\mathbf{C})^{-n_r/2}\text{etr}[-\frac{1}{2}\mathbf{R}^{-1}\mathbf{F}\mathbf{C}^{-1}\mathbf{F}^T], \tag{4}$$

where $\det(\cdot)$ denotes matrix determinant. BFCM explicitly treats $\mathbf{R}$ and $\mathbf{C}$ as latent random variables, and defines priors and performs inferences on them. Technically, the model is formulated in terms of $\mathbf{R}^{-1}$ and $\mathbf{C}^{-1}$. The prior on $\mathbf{R}^{-1}$ is defined as

$$\begin{aligned} &p(\mathbf{R}^{-1}|\tau_0,\nu_0,\rho) \\ \propto&\mathcal{W}(\mathbf{R}^{-1}|\tau_0\mathbf{I}_{n_r},\nu_0)\cdot\mathcal{L}(\mathbf{R}^{-1}|\rho) \\ \propto&\det(\mathbf{R})^{-(\nu_0-n_r-1)/2}\text{etr}\left[-\frac{\tau_0}{2}\mathbf{R}^{-1}\right]\exp\left[-\rho\|\mathbf{R}^{-1}\|_1\right], \end{aligned} \tag{5}$$

where $\mathcal{W}(.)$ and $\mathcal{L}(.)$ represent the Wishart and Laplace density functions and $\|\cdot\|_1$ is the 1-norm of a matrix. With respect to $\mathbf{R}^{-1}$, the prior has the conjugate form to (4). Furthermore, the Laplace term encourages $\mathbf{R}^{-1}$ to be sparse. The prior belief of sparsity on $\mathbf{R}^{-1}$ is sensible and helping inference. An entry in $\mathbf{R}^{-1}$ corresponds to *direct* connections between two rows, which is a strong relation: if $[\mathbf{R}^{-1}]_{i_1,i_2} \neq 0$, then knowing the i_1-th row provides unique information about the i_2-th row, which is *not* provided by knowning all the remaining rows, and *vice versa*. The sparse prior encourages the model to focus on significant relations, which reduces the risk of over-fitting and improves efficiency of evaluating functions like (4). The distribution of $\mathbf{C}$ is defined in the same form as that of $\mathbf{R}$.

We employ the conjugate prior for the noise variance σ^{-2}, which is a gamma distribution

$$p(\sigma^{-2}|\alpha_0,\beta_0) = \mathcal{G}(\sigma^{-2}|\alpha_0,\beta_0), \tag{6}$$

where $\mathcal{G}(.)$ stands for the gamma density function. Therefore the Bayesian model structure in Fig. 1 is completed by the joint density

$$\begin{aligned} &p(\mathbf{F}^{(1\dots N)},\mathbf{R},\mathbf{C},\sigma^2,\mathbf{Y}^{(1\dots N)}) \\ =\ &p(\mathbf{R})p(\mathbf{C})p(\sigma^2)\prod_{n=1}^{N}\left[p(\mathbf{Y}^{(n)}|\mathbf{F}^{(n)},\sigma^2)p(\mathbf{F}^{(n)}|\mathbf{R},\mathbf{C})\right], \end{aligned} \tag{7}$$

where we use concise denotations for the priors: $p(\mathbf{R}), p(\mathbf{C})$ are formulated by (5) and $p(\sigma^2)$ by (6).

2.2 Inference

In this subsection, we are concerned with the posterior distributions of the latent variables given the data. The posterior is determined by the joint probability. Substituting (4), (5), (6)

and (1) in 7, and taking the logarithmic form, we have

$$
\begin{aligned}
L = &-\frac{1}{2}\sum_n \mathrm{tr}\left[\mathbf{R}^{-1}\mathbf{F}^{(n)}\mathbf{C}^{-1}\mathbf{F}^{(n)T}\right] \\
&+\sum_n\left\{-n_c n_r \log\sigma - \frac{\sigma^{-2}}{2}\left\|\mathbf{F}^{(n)}-\mathbf{Y}^{(n)}\right\|_F^2\right\} \\
&-\frac{\nu_0+Nn_c}{2}\log\det(\mathbf{R}) - \frac{\tau_0}{2}\mathrm{tr}[\mathbf{R}^{-1}] - \rho\|\mathbf{R}^{-1}\|_1 \\
&-\frac{\nu_0+Nn_r}{2}\log\det(\mathbf{C}) - \frac{\tau_0}{2}\mathrm{tr}[\mathbf{C}^{-1}] - \rho\|\mathbf{C}^{-1}\|_1 \\
&+(\alpha_0-1)\log\sigma^{-2} - \beta_0\sigma^{-2} + \mathrm{Const.}
\end{aligned} \tag{8}
$$

The joint posterior w.r.t. all the latent variables is not computationally tractable. Thus the inference is performed in an alternating manner, where we compute the posterior of each variable in turn by conditioning (8) on the remaining variables.

Posterior distributions

Posterior of $\mathbf{F}^{(n)}$ In (8), $\mathbf{F}$ (omitting the sample index) appears in the product of two Gaussian density functions $p(\mathbf{F}|\mathbf{R},\mathbf{C})p(\mathbf{Y}|\mathbf{F})$, which is specified by (1) and (3), respectively. Thus the posterior of $\mathbf{F}$ given $\mathbf{R},\mathbf{C}$ and $\mathbf{Y}$ is a Gaussian

$$
\begin{aligned}
p(\mathrm{vec}\,(\mathbf{F})\,|\mathbf{R},\mathbf{C},\mathbf{Y}) =& \mathcal{N}(\mu_F,\mathbf{\Sigma}_F), && (9)\\
\mu_F =& \mathbf{\Sigma}_F \times \mathrm{vec}\,(\mathbf{Y})\,/\sigma^2 && (10)\\
\mathbf{\Sigma}_F^{-1} =& (\mathbf{C}\otimes\mathbf{R})^{-1} + \sigma^{-2}\mathbf{I}. && (11)
\end{aligned}
$$

Posterior of $\mathbf{R}$ and $\mathbf{C}$ Since the prior (5) is derived from the conjugate family of (4), the posterior of $\mathbf{R}$ has the same functional form as the prior. Considering the relevant terms in (8) gives the posterior as

$$
\begin{aligned}
&\log\ p(\mathbf{R}|\mathbf{F}^{(1\ldots N)},\mathbf{C}) \\
=&-\frac{1}{2}\mathrm{tr}(\mathbf{R}^{-1}\tilde{\mathbf{S}}) - \frac{\nu_0+Nn_c}{2}\log|\mathbf{R}| - \rho\|\mathbf{R}\|_1 + \mathrm{Const},
\end{aligned} \tag{12}
$$

where $\tilde{\mathbf{S}} = \tau_0\mathbf{I} + \sum_n \mathbf{F}^{(n)}\mathbf{C}^{-1}\mathbf{F}^{(n)T}$. The posterior of $\mathbf{C}$ can be derived similarly.

Posterior of σ^{-2} The gamma prior of σ^{-2} is conjugate to the likelihood (1), thus the posterior is also a gamma distribution:

$$
\begin{aligned}
&p(\sigma^{-2}|\mathbf{F},\mathbf{Y}) \\
=&\mathcal{G}(\alpha_0 + \frac{n_c n_r N}{2}, \beta_0 + \frac{1}{2}\sum_n \|\mathbf{F}^{(n)}-\mathbf{Y}^{(n)}\|_F^2),
\end{aligned} \tag{13}
$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Computation procedure

We compute the *maximum a posteriori* (MAP) solution to the posterior distributions of the latent variables. We choose MAP not only because there are off-the-shelf optimisation algorithms providing convenient and fast computations, but also because the posterior modes of $\mathbf{R}^{-1}$ and $\mathbf{C}^{-1}$ are sparse, which improves model robustness and interpretability. We implement MAP as an alternating optimisation of (8) in terms of $\mathbf{F}$, σ^{-2}, $\mathbf{R}^{-1}$ and $\mathbf{C}^{-1}$. The computation of $\mathbf{F}$ and that of $\mathbf{R}^{-1}$ and $\mathbf{C}^{-1}$ need special attention.

Compute posterior mode of $\mathbf{F}$ The posterior of $\mathbf{F}$ is a standard Gaussian distribution given all the other variables. The corresponding mode is the mean parameter of the density function (9). However, $\mathbf{F}$ is a multivariate distribution of $n_r \times n_c$ variables, and the naive computation of the posterior mode following (10) and (11) involves an expensive matrix inversion. Thus we compute efficiently [Yan *et al.*, 2011] by using SVD of the factor covariance $\mathbf{R}$ and $\mathbf{C}$. Let

$$
\begin{aligned}
\mathbf{R} &= \mathbf{U}_R\mathbf{S}_R\mathbf{U}_R^T \\
\text{and } \mathbf{C} &= \mathbf{U}_C\mathbf{S}_C\mathbf{U}_C^T.
\end{aligned}
$$

Then $\mathbf{\Sigma}_F$ can be obtained by reformulating (11)

$$
\mathbf{\Sigma}_F = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T,
$$

where $\mathbf{U} = \mathbf{U}_C \otimes \mathbf{U}_R$, $\mathbf{S} = \mathbf{S}_C \otimes \mathbf{S}_R$, and the diagonal matrix $\mathbf{\Lambda}$ is given by

$$
\Lambda_{ii} = \frac{1}{S_{ii}^{-1} + \sigma^{-2}},\ i = 1,\ldots,n_r \times n_c.
$$

Comparing (10), we can compute μ_F efficiently as

$$
\mu_F = \sigma^{-2}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \times \mathrm{vec}\,(\mathbf{Y})\,.
$$

By the relation of Kronecker product and matrix vectorisation, we have

$$
\begin{aligned}
\mathbf{U}^T \times \mathrm{vec}\,(\mathbf{Y}) &= (\mathbf{U}_C^T \otimes \mathbf{U}_R^T)\mathrm{vec}\,(\mathbf{Y}) \\
&= \mathrm{vec}\left(\mathbf{U}_R^T\mathbf{Y}\mathbf{U}_C\right),
\end{aligned} \tag{14}
$$

and $\sigma^{-2}\mathbf{\Lambda} \times \mathrm{vec}\left(\mathbf{U}_R^T\mathbf{Y}\mathbf{U}_C\right)$ can be arranged as a element-wise product as

$$
\begin{aligned}
\sigma^{-2}\mathbf{\Lambda}\cdot\mathrm{vec}\left(\mathbf{U}_R^T\mathbf{Y}\mathbf{U}_C\right) &= \mathrm{vec}\left(\mathbf{\Psi}\odot\left(\mathbf{U}_R^T\mathbf{Y}\mathbf{U}_C\right)\right) \\
\mathbf{\Psi} &= \frac{\boldsymbol{\lambda}_r\boldsymbol{\lambda}_c^T}{\sigma^2 + \boldsymbol{\lambda}_r\boldsymbol{\lambda}_c^T},
\end{aligned}
$$

where $\mathbf{\Psi}$ is a $n_r \times n_c$ matrix of the diagonal of $\mathbf{\Lambda}$ and $\boldsymbol{\lambda}_r$ and $\boldsymbol{\lambda}_c$ are column vectors of the diagonal of $\mathbf{S}_R$ and $\mathbf{S}_C$, respectively. The operator $\odot$ represents element-wise product. Employing the same trick as (14) completes the computation of the posterior mean μ_F

$$
\begin{aligned}
\mu_F &= \mathbf{U}\cdot\mathrm{vec}\left(\mathbf{\Psi}\odot\left(\mathbf{U}_R^T\mathbf{Y}\mathbf{U}_C\right)\right) \\
&= \mathrm{vec}\left(\mathbf{U}_R\left[\mathbf{\Psi}\odot\left(\mathbf{U}_R^T\mathbf{Y}\mathbf{U}_C\right)\right]\mathbf{U}_C^T\right).
\end{aligned}
$$

Compute posterior model of $\mathbf{R}$ and $\mathbf{C}$ The posterior of $\mathbf{R}$ is given by (12). We consider the following change of variables in the equation, letting $\mathbf{Q} = \mathbf{R}^{-1}$, $\tilde{\mathbf{S}}' = \tilde{\mathbf{S}}/(\nu_0 +$

Nn_c) and $\rho' = 2\rho/(\nu_0 + Nn_c)$. Then maximising (12) w.r.t. $\mathbf{R}^{-1}$ is equivalent to

$$\arg\max_{\mathbf{Q}} - \mathrm{tr}(\mathbf{Q}\tilde{\mathbf{S}}') + \log\det(\mathbf{Q}) - \rho'\|\mathbf{R}\|_1.$$

The maximisation can be solved efficiently by employing the graph Lasso algorithm [Friedman *et al.*, 2008].

In theory, the time complexity for each iteration is $\mathcal{O}(\hat{n}^3)$, where $\hat{n} = \max(n_r, n_c, N)$. In practice, the computation is dominated by the computations of $\mathbf{R}^{-1}$ and $\mathbf{C}^{-1}$, for which the worst scenario time complexities are $\mathcal{O}(n_r^3)$ and $\mathcal{O}(n_c^3)$ [Friedman *et al.*, 2008].

3 Related Work

Factorised covariance model is related to several existing lines of research in machine learning and statistics. The covariance matrix plays an essential role in latent factor models [Jolliffe, 2002; Tao *et al.*, 2009], which analyse the covariance to design a small number of new variables (latent factors), so that varying the latent factors reproduces the joint heterogeneity of the observed variables. BFCM can be specialised as a covariance estimation step that precedes a traditional latent factor analysis. In particular, we can stack each matrix into a long column vector so all the data are transformed into a single matrix $\mathbf{Y}_D$ of $[n_v \times N]$, where $n_v = n_r \times n_c$. BFCM is applied to $\mathbf{Y}_D$: we solve for $\mathbf{R}$ and fix $\mathbf{C} = \mathbf{I}$ so that the N samples are independent. Then the learned $\mathbf{R}$ summarises the correlations between all observed variables without considering the matrix structure, and a compact parameterisation of $\mathbf{R}$ leads to the traditional latent factors, e.g. a reduced rank SVD of $\mathbf{R}$ represents a Bayesian implementation of probabilistic PCA [Tipping and Bishop, 1999; Li and Tao, 2013].

There are several efforts addressing the small sample size (SSS) problem related to covariance estimation for high-dimensional data. For images, [Shashua and Levin, 2001] pointed that for compression, a model should account for not only redundancies between images, but also the spatial redundancies (row/column-correlations) within individual images, and they represented a set of images as a 3-D array. The model relieves the SSS problem, since it treats rows and columns, rather than entire images as units of analysis – it is possible to compute a valid model from a single image. In [Shashua and Levin, 2001], the computation is based on addition of rank-1 arrays, while in [Vasilescu and Terzopoulos, 2002; Tao *et al.*, 2007b; 2007a], it has been shown that high-order SVD [Tucker, 1966] provides a natural and effective computational tool for multi-dimensional arrays of image data.

Alternatively, reliable covariance estimators have been proposed by regularising the problem based on sparsity assumptions. For example, banded structure and thresholds have been proposed in [Bickel and Levina, 2008b] and [Bickel and Levina, 2008a], respectively. Regularisation of pairwise correlations has been proposed in our previous work [Li and Tao, 2012]. Learning a sparse *inverse* covariance has also been studied [Rothman *et al.*, 2008], which corresponds to trimmed connections between the variables from the viewpoint of graphical models. Based on the efficient gradient descent algorithm for lasso [Friedman *et al.*, 2007], a fast solution to the optimisation problem has been developed in [Friedman *et al.*, 2008], which is employed for the inference of BFCM, as well as many other regularised models [Guan *et al.*, 2012]. In addition to sparse inverse, BFCM exploits the matrix form of the data, and assumes the convariance matrix is the Kronecker product of two covariance matrices: one of the rows and the other of the columns.

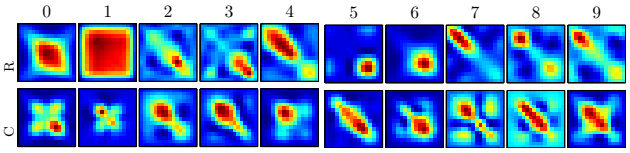

Figure 2: Row/column covariance of hand-written digits.

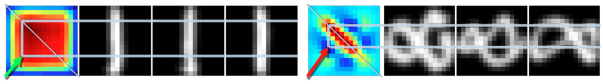
Figure 3: Some revealed relations in the images

Factorisation-models have also been employed in relational data analysis, where the observed array consists of records of the interactions among multiple groups of entities [Nickel *et al.*, 2011]. In [Yu *et al.*, 2006], a relational data model has been developed by defining multiple Gaussian processes on the participating entity groups. As a result, the joint covariance is given by the product of the covariance of those constituent Gaussian processes. The model can be alternatively motivated by considering the observations as the result of applying a (possibly infinite) multilinear transformation to a random core matrix of independent unit Gaussian entries [Yan *et al.*, 2011]. Extensions to tensor data has also been considered in [Chu *et al.*, 2009] and [Xu *et al.*, 2012]. Most relational data models treat $\mathbf{R}$ and $\mathbf{C}$ as kernel matrices to be fixed in learning, in contrast, BFCM explicitly employs a Bayesian model for the factor covariance matrices. We impose priors and perform inference on $(\mathbf{R}, \mathbf{R}^{-1})$ and $(\mathbf{C}, \mathbf{C}^{-1})$ directly. Probabilistic treatment is also employed by [Yu *et al.*, 2006], where an inverse Wishart density is imposed on Gaussian process kernels. However, the kernels in that model operate separately on two sets of factors generating the relational matrix, which differs from the direct covariance learning in BFCM.

4 Experiments

Discover image structure We first show the model is effective in discovering knowledge about an image class from a relatively small number of samples. BFCM is used to analyse hand-written digits from USPS dataset [Hull, 1994]. A BFCM is learned for each digit by using 10 randomly chosen images.

Fig. 2 shows the learned row- and column-wise covariance matrices. A brief investigation of the covariance matrices suggests that they are meaningful and representing structural information of the corresponding image classes. For example, Fig. 3 illustrates the interpretation of two entries of the covariance: one from the row-covariance for digit "1" and the other from the column-covariance for digit "8". The row-covariance of "1" suggests that most rows of a "1"-image are

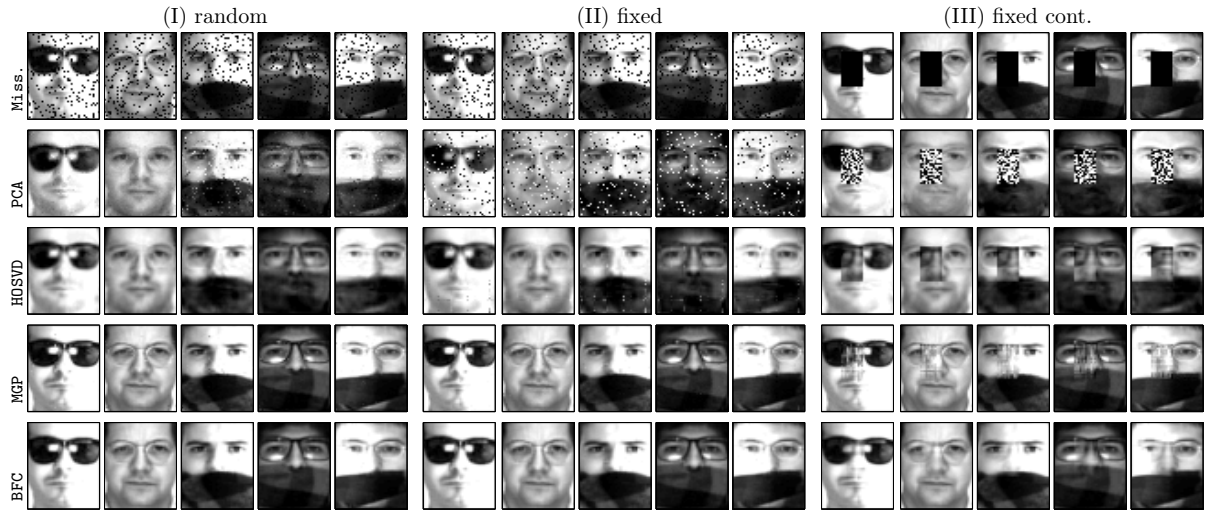

Figure 4: Predict missing pixels in face images. **(I)** missing at random positions in each image, **(II)** missing at fix positions in all images, **(III)** missing at a fixed and contiguous rectangle in all images. The top row shows example images with missing pixels. The 2nd to the 5th rows shows predicted images by different probabilistic models: PCA, HOSVD, Matrix Gaussian process (MGP) and BFCM. Note: for prediction task (II) and (III), the 2D-array structure of data must be accounted for, because the missing pixels are unseen in *all* images.

highly correlated, which is consistent with how "1" is written and is shown by the example images in Fig. 3. On the other hand, the marked entry of the column-covariance of "8" indicates that the corresponding pair of columns are negatively correlated. This is validated by the (transposed) example images , where the pen stokes of "8" cover different parts on the two columns. It is also noticeable that the symmetrical structures of the images manifest themselves in the anti-diagonal entries of the covariance matrices. Moreover, the images of "5/6" has similar row-covariance, which contrasts to that of "7/9".

In this experiment, the pixels are normalised so the average variance is 1. Diffuse prior parameters are used: $\alpha_0 = 1$, $\beta_0 = 0.01$, $\tau_0 = 1$, $\nu_0 = 1$, and we let $\rho = 0$ to disable sparsity, because the demonstration focuses on the factorised covariance structure for images.

Predict missing pixels Secondly, we learn probabilistic models from facial images to predict missing pixels. The test uses 50 facial images from the AR dataset[1]. Essentially, prediction is to relate the unknowns to observations. One strategy is to exploit the connections between the samples. For example, the problem of overcoming occlusions in facial images has been cast as an ℓ_1 regularised regression task in [Wright *et al.*, 2009]. The regression formulation is effective when a fully observed reference database is available. If all observed samples may contain missing values, a probabilistic model is necessary to scale the analysis of connections from the samples down to the constituent variables. E.g., probabilistic PCA (PPCA) [Tipping and Bishop, 1999] can learn the correlations between individual pixels, so that missing pixels can be inferred from observed ones.

Examples of applying this scheme are shown in Group (I) in Fig. 4. In the top row, the five images contain 10% randomly missing pixels. The second row shows the images given by PPCA. The estimated missing pixels are roughly consistent with their ground-truth values according to the figure. However, there are visible artefacts, because the observed samples (50) is insufficient compared to the number of variables ($50 \times 40 = 2,000$) for learning PPCA reliably.

[1] www2.ece.ohio-state.edu/~aleix/ARdatabase.html

	Rand.	Fix	Fix Cont.
PCA	0.1826	5.799	5.803
HOSVD	0.066	0.152	1.756
MGP	0.028	0.026	0.488
BFCM	0.016	0.017	0.306

Table 1: Mean square error of predicted missing pixels. The average pixel variance is 1.0.

To address the small sample size problem, we represent the images as a multi-dimensional array by using high-order SVD (HOSVD) [Vasilescu and Terzopoulos, 2002]. The prediction of missing values is an iterative procedure: (i) performing HOSVD on complete data and (ii) completing the data by using the prediction of HOSVD. The procedure starts with randomly initialised missing pixels. The third row shows the results by HOSVD, which supresses over-fitting and improves the prediction.

We also test the matrix Gaussian process (MGP) [Yu *et al.*, 2006; Yan *et al.*, 2011] and BFCM, which also produce better result than PCA. Similar to HOSVD, MGP and BFCM avoid over-fitting by accounting for the array-structure of images and dealing with a smaller learning problem. On the other hand, MGP and BFCM consider the correlations in both the rows and the columns, and thus give more precise predictions. In Table 1, the left column compares the MSE of the missing value predictions, where the average variance of the pixels is normalised to 1.

Besides the over-fitting issue, the strategy of focusing relations between rows or columns instead of those between

	YALE	ORL	AR
RAW	66.7 ± 4.4	94.6 ± 2.6	43.0 ± 1.5
PCA[2]	68.3 ± 3.6	91.1 ± 3.4	38.6 ± 1.1
HOSVD	42.1 ± 1.1	90.4 ± 1.7	44.2 ± 2.7
MGP	69.0 ± 6.5	91.1 ± 1.9	42.1 ± 1.2
BFCM	75.3 ± 4.2	94.2 ± 2.9	50.7 ± 4.5

Table 2: Classification rates on different representations.

pixels helps the learning by expanding the applicability. A noticeable prerequisite of learning the pixel-wise covariance is that *all* pixels must be observed at least once. This is appropriate if the missing is at random, e.g. when data are lost due a poor communication channel. However, the assumption fails for an interesting scenario of image processing: when the images are acquired by a compromised *sensor*, which results missing pixels at fixed positions in *all* images.

Group (II) in Fig. 4 displays examples of predicting missing pixels at fixed positions. As above, the second row shows predicted images by PPCA. The model stops working, because it is infeasible to explicitly learn the covariance between pixel pairs, if one of them has never been observed. On the other hand, HOSVD, MGP and BFCM are not affected by fixing missing positions. The second column of Table 1 lists the MSEs. The predictions by PCA are invalid, and those by HOSVD, MGP and BFCM are similar to the previous test.

We also test the model by letting the missing happen in a fixed contiguous rectangle. Although the missing rectangle of 20×12 contains a similar number of missing pixels as in the previous tests, the task is more challenging. An image pixel is most likely to be related to its neighbours, and if a contiguous area has never been observed, it is difficult to obtain useful knowledge about the inner pixels.

Group (III) shows the example images and the prediction results. HOSVD gives unreliable predictions in this test. In principle, HOSVD considers the joint correlations between columns and rows. However, in implementation, when computing the correlations in one mode, e.g. those of rows, all rows in the images are treated as independent samples. The computation becomes unreliable when significant number of rows contain missing values at the same positions. MGP and BFCM suffer less from the contiguous missing pattern. In particular, BFCM explicitly learns the covariance and is better at discovering and exploiting the symmetrical structure between the columns in facial images. The visual assessment of the results is supported by the MSEs in the third column of Table 1.

In the experiment, we determine the model parameters by the performance in the first test (missing at random), including the rank of PCA and HOSVD, the kernel width and the latent dimension in MGP, and τ_0 in BFCM (ν_0, α_0 and β_0 are set as previously).

Learn representation by a single example Thirdly, we test model-based representation for the facial images. In particular, we challenge the model by using only a single image for training. This setting of sample scarcity seems to be somewhat extreme. However, it shows the strength of a factorised structure. One image contains many columns and rows. Thus it is possible to learn both the row and column covariance from a single image, which is an ill-defined problem for learning pixel-wise covariance. Once a BFC model is learned, a projection can be derived by following the rule of maximum variance, which resembles standard PCA. Specifically, the projection is given by performing SVD on the estimated covariance. After $\mathbf{R}$ and $\mathbf{C}$ are estimated, the SVD and the projection is obtained by (14). A similar projection can be derived for MGP by using the estimated kernels of the row and column Gaussian processes. We also include PCA[2] and HOSVD in this experiment. For classification, the simple nearest neighbour rule (NN) is used. The test is done on three face datasets (AR, YALE[3] and ORL[4]) and the results are listed in Table 2. As a baseline, we also include results of applying NN on the raw data. The parameters of the algorithms in this experiment is determined by cross-validation.

The results shows that BFCM learns effective covariance from a single observation and leads to new representation. For discrimination, the new representation is superior to the raw data for the AR and YALE datasets, which contain relatively large appearance variations. On the other hand, the other models do not provide representations that are significantly more discriminative than the raw features. Hence direct regularisation of the covariance enables BFCM to learn a useful facial image model from a single example.

5 Conclusion

A Bayesian covariance estimator for matrix data has been developed. A factorised structure is adopted. The structure is motivated by meaningful relations in the row- and column-indexes of the matrix, and it helps reduce the covariance learning problem to a manageable scale for practical matrix data sizes.

Different from existing relational data models using factorised covariance, this work explicitly deals with the factor covariance matrices, treats them as random variables in a Bayesian model, imposes meaningful and sparsity-inducing priors and takes advantage of a fast optimisation algorithm for inference. In particular, the priors are defined w.r.t. the inverse of the factor covariance, consisting of a Wishart and a Laplace term. The components in the prior jointly encourage independency, which further regularises the learning problem and makes the model applicable for scarce observations.

Experiments demonstrate that the model is capable of learning useful relations for image understanding. The knowledge can be applied to infer missing data and to extract informative representation for subsequent classification tasks. Noticeably, in the former task, the model recovered a pixel even if it is missing in all images, and its relation to the remaining pixels is not directly available; in the latter task, the model learned effective representations from a single image.

[2] PCA does not work on a single training image, and is computed using up to 100 images from 5 subjects excluded from test set.

[3] cvc.yale.edu/projects/yalefaces/yalefaces.html

[4] www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

References

[Bickel and Levina, 2008a] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Stat.*, 36(6):2577–2604, 2008.

[Bickel and Levina, 2008b] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Stat.*, 36(1):199–227, 2008.

[Chu *et al.*, 2009] W. Chu, S. Clara, and Z. Ghahramani. Probabilistic models for incomplete multi-dimensional arrays. In *Proc. AI & Stat. (AISTATS)*, 2009.

[Friedman *et al.*, 2007] J. Friedman, T. Hastie, H. Hoefling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, pages 302–332, 2007.

[Friedman *et al.*, 2008] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–41, 2008.

[Gelman *et al.*, 2003] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2003.

[Guan *et al.*, 2012] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. NeNMF: An optimal gradient method for nonnegative matrix factorization. *IEEE Trans. Signal Proc.*, 60(6):2882–2898, 2012.

[Gupta and Nagar, 2002] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman & Hall/CRC, 2002.

[Hull, 1994] J. J. Hull. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intel.*, 16(5):550–554, 1994.

[Jolliffe, 2002] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2002.

[Li and Tao, 2012] Jun Li and Dacheng Tao. On preserving original variables in bayesian PCA with application to image analysis. *IEEE Trans. Image Proc.*, 21(12):4830–4843, 2012.

[Li and Tao, 2013] Jun Li and Dacheng Tao. Simple exponential family PCA. *IEEE Trans. Neural Networks and Learning Systems*, 2013. doi:10.1109/TNNLS.2012.2234134.

[Marcenko and Pastur, 1967] V. A. Marcenko and L. A. Pastur. Distributions of eigenvalues of some sets of random matrices. *Math. USSR-Sb*, pages 507–536, 1967.

[Nickel *et al.*, 2011] M. Nickel, V. Tresp, and H. Kriegel. A three-way model for collective learning on multi-relational data. In *Proc. Int. Conf. Mach. Learning (ICML)*, 2011.

[Rothman *et al.*, 2008] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic J. Stat.*, 2:494–515, 2008.

[Shashua and Levin, 2001] A. Shashua and Anat Levin. Linear image coding for regression and classification using the tensor-rank principle. In *Conf. Computer Vision and Pattern Recog. (CVPR)*, 2001.

[Tao *et al.*, 2007a] Dacheng Tao, Xuelong Li, Xindong Wu, Weiming Hu, and Stephen J. Maybank. Supervised tensor learning. *Knowl. Inf. Syst.*, 13(1):1–42, 2007.

[Tao *et al.*, 2007b] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J. Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1700–1715, 2007.

[Tao *et al.*, 2009] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J. Maybank. Geometric mean for subspace selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):260–274, 2009.

[Tipping and Bishop, 1999] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J. Royal Statist. Soc. – B*, 61(3):611–622, 1999.

[Tucker, 1966] L.R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

[Vasilescu and Terzopoulos, 2002] M.A.O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: TensorFaces. In *European Conf. Computer Vision (ECCV)*, 2002.

[Wright *et al.*, 2009] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intel.*, 31(2):1–17, 2009.

[Xu *et al.*, 2012] Z. Xu, F. Yan, and Y. Qi. Infinite Tucker decomposition: Nonparametric Bayesian models for multiway data analysis. In *Proc. Int. Conf. Mach. Learning (ICML)*, 2012.

[Yan *et al.*, 2011] F. Yan, Z. Xu, and Y. Qi. Sparse matrix-variate Gaussian process blockmodels for network modeling. In *Proc. Uncertainty Artif. Intel. (UAI)*, 2011.

[Yu *et al.*, 2006] K. Yu, W. Chu, S. Yu, and V. Tresp. Stochastic relational models for discriminative link prediction. In *Adv. in Neural Inf. Process. Syst. (NIPS)*, 2006.