



# Sparse conditional copula models for structured output regression



Minyoung Kim

Department of Electronics & IT Media Engineering, Seoul National University of Science & Technology, Seoul 139-743, South Korea

## ARTICLE INFO

### Article history:

Received 17 March 2015  
 Received in revised form  
 1 February 2016  
 Accepted 19 March 2016  
 Available online 4 April 2016

### Keywords:

Multiple output regression  
 Sparse inverse covariance estimation  
 Copula models  
 Gaussian random fields

## ABSTRACT

We deal with the multiple output regression task where the central theme is to capture the sparse output correlation among the output variables. Sparse inverse covariance learning of linear Gaussian conditional models has been recently studied, shown to achieve superb prediction performance. However, it can fail when the underlying true input–output process is non-Gaussian and/or non-linear. We introduce a novel sparse conditional copula model to represent the joint density of the output variables. By incorporating a Gaussian copula function, yet modeling univariate marginal densities by (non-Gaussian) mixtures of experts, we achieve high flexibility in representation that admits non-linear and non-Gaussian densities. We then propose a sparse learning method for this copula-based model that effectively imposes sparsity in the conditional dependency among output variables. The learning optimization is efficient as it can be decomposed into gradient-descent marginal density estimation and the sparse inverse covariance learning for the copula function. Improved performance of the proposed approach is demonstrated on several interesting image/vision tasks with high dimensions.

© 2016 Elsevier Ltd. All rights reserved.

## 1. Introduction

We consider the multiple output (or vector output) regression problem where the goal is to predict multiple response variables from input covariates. Contrary to the conventional single (scalar) output regression in statistics and machine learning fields, the central aspect is to capture the statistical correlation among the output variables efficiently. In this sense, the simple strategy of applying single output regression estimation independently for each of the output variables is considered to be suboptimal since it completely ignores the output correlation in learning a predictive model. In the machine learning literature, there has been considerable recent research work on estimating more accurate multiple output regressors beyond the independent treatment [1–3].

In the practitioner's point of view, effective output correlation modeling in regression is highly beneficial for several application problems. For instance, for the task of denoising hand-written character images, the output vector comprises of pixel intensities of a character image, in which neighboring dimensions tend to have similar values (as either foreground character pixels or background), clearly exhibiting strong conditional dependency patterns among the output variables. Also in computer vision, in the motion estimation problem predicting motions for next few frames, an instance of typical forecasting problems, as the image features at two consecutive frames typically conform to the

motion smoothness constraint, the data contain high statistical correlation among the output variables.

However, for situations where the input/output data are high dimensional, one has to trade off the faithful output correlation modeling against model complexity. One sensible approach is to confine the model parameter space with sparseness constraints, which motivates the recently emerging *sparse inverse covariance estimation* [4–9]. Under the Gaussian random field models (either joint or conditional densities), these approaches attempt to impose the sparseness of the inverse covariance matrix based on the following fact: a Gaussian density factorizes into pairwise potentials on output variables (say,  $y_i$  and  $y_j$ ), with the corresponding coefficient equal (up to constant) to the  $(i,j)$  component of the inverse covariance matrix. Therefore, having many zero entries in the inverse covariance leads to a loose statistical dependency structure.

While the sparse inverse covariance estimation has achieved superb prediction performance, the main weak point is that most approaches are built on the strong assumption of linear Gaussianity of output given input. Although the assumption leads to convex optimization, the linear Gaussian model family is less flexible and considerably restricted in representational capacity. In the sequel, when the true data generating process is far from linear Gaussian, it may suffer from inaccurate prediction.

In this paper, we propose a novel sparse copula-based density model that admits much richer non-Gaussian and non-linear density models. The copula model is studied considerably in

E-mail address: [mikim21@gmail.com](mailto:mikim21@gmail.com)

statistics community [10–15], while its application to machine learning is relatively rare. A main difficulty in multivariate density modeling is to choose a proper function family beyond the multivariate Gaussian, and the copula alleviates it by decoupling the task into two subtasks: (i) build marginal distributions for individual variables which is easy due to univariate modeling, and (ii) model the inter-correlation among the variables by the so-called *copula* function. The copula function can be any function that only satisfies certain uniform marginal constraints.

One may argue that defining the copula function is as difficult as choosing a multivariate density. However, the most popular strategy is to adopt the Gaussian copula function. It is known that as far as the marginal densities are not linear Gaussians, the resulting density is non-Gaussian. Moreover, by choosing sensible (non-Gaussian) marginal densities, the model can be endowed with high representational power beyond Gaussian, hence providing a great deal of flexibility in modeling a non-linear and non-Gaussian multivariate distribution.

Indeed the proposed model employs, as the univariate marginal densities, the mixture of sigmoid-weighted Gaussians. Also known as the mixture of experts [16,17], it divides the input space into different regimes, each delineated by multiple hyperplanes, and in each regime, the output variable conforms to a mixture of Gaussians, a rich density family that can approximate any density with high precision.

In addition to the enriched representational capacity with flexible density modeling, the intriguing property of the proposed model is that the conditional independence among the output variables can be directly imposed by a sparse inverse covariance in the copula function. Imposing sparseness of the inverse covariance in the copula function in learning a density model is the central idea in our approach, which is beneficial for capturing the most salient correlation in the output variables for high-dimensional data scenarios. Moreover we derive that the sparse inverse copula covariance learning with parameters of marginal densities fixed, becomes an instance of convex optimization.<sup>1</sup>

Furthermore, the learning problem can be framed exactly into the existing sparse inverse covariance learning for linear Gaussian models by replacing the empirical second-order moments with the statistics based on the marginal copula-linked features. In the computational perspective, we can benefit from the existing fast-convergent sparse inverse covariance learning algorithms (e.g., [9]) without any modification. In a nutshell, our copula-based sparse density model is endowed with rich representational power, and at the same time, avoids overfitting by loose statistical dependency modeling through sparse inverse covariance in the copula function. This promising aspect is indeed demonstrated on several interesting real-world multiple output regression tasks.

The paper is organized as follows: after describing the formal problem setup and notation in Section 1.1, we briefly discuss some recent related work on sparse inverse covariance learning for Gaussian models in Section 2. Then in Section 3, our conditional copula model is introduced, for which we propose the sparse learning algorithm based on block coordinate descent. In Section 4 the experimental results on several regression tasks, contrasting the proposed model against existing approaches are provided.

<sup>1</sup> More precisely speaking, as derived in Section 3.3, we have relaxed the constraints of the diagonal entries of the correlation matrix being all 1. Although the relaxation may result in improper density in the sequel, however, it is advantageous to make the problem remain convex (when the expert gating parameters are fixed). In practice, the solution obtained from the relaxed problem gives superior prediction performance, which justifies the modification of the problem.

## 1.1. Notations and problem setup

The *multiple output regression* is considered where the main goal is to predict the response vector  $\mathbf{y} \in \mathbb{R}^p$  from the input feature vector (or the covariates)  $\mathbf{x} \in \mathbb{R}^d$ . One straightforward approach can be applying standard scalar-output regression, estimating a regression model for each output dimension independently. However, it can lead to a suboptimal solution in that the underlying inter-correlation among the output (random) variables is ignored. Properly capturing the statistical dependency of the response variables is the key to yield better prediction models. We are particularly interested in situations where the input/output dimensions ( $d$  and  $p$ ) are relatively large compared to the training sample size. This implies that one can have potentially overfitted models unless certain regularization or constraints are properly imposed on the model space.

For notations, we use boldfaced symbols for vectors and matrices, and plain letters for scalar values. Also,  $y_i$ , for  $i = 1, \dots, p$ , indicates the  $i$ -th entry of  $\mathbf{y}$ , while  $\mathbf{y}_C$  for a set  $C \subseteq \{1, \dots, p\}$  stands for the set of variables in  $\mathbf{y}$  whose indices belong to  $C$ . We denote by  $\mathcal{N}(\mathbf{z}; \mathbf{m}, \mathbf{V})$  the multivariate Gaussian density with mean  $\mathbf{m}$  and covariance  $\mathbf{V}$ . For a matrix  $\mathbf{A}$ ,  $A_{ij}$  indicates the  $(i,j)$ -entry of  $\mathbf{A}$ . The determinant and trace of a positive definite matrix  $\mathbf{\Sigma}$  (i.e.,  $\mathbf{\Sigma} \succ 0$ ) are denoted by  $\det(\mathbf{\Sigma})$  and  $\text{Tr}(\mathbf{\Sigma})$ , respectively.

## 2. Related work on sparse inverse covariance learning

In this section we briefly review some related work on sparse inverse covariance learning for Gaussian random field models and related models. Beginning with general sparse Gaussian density estimation (Section 2.1), we describe its extension to conditional linear Gaussian models for multiple output prediction tasks (Section 2.2).

### 2.1. Sparse inverse covariance density estimation

For the task of  $k$ -variate Gaussian density estimation  $P(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{V})$ ,  $\mathbf{z} \in \mathbb{R}^k$ , with given data  $\{\mathbf{z}^i\}_{i=1}^n$  and the zero-mean data assumption without loss of generality, the standard maximum-likelihood estimate (MLE) admits the closed form solution,  $\mathbf{V} = \mathbf{S}$  where  $\mathbf{S}$  is the sample covariance matrix:

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}^i \mathbf{z}^{i\top}. \quad (1)$$

Despite its asymptotic consistency, MLE becomes quickly unreliable when the data are high-dimensional ( $k$  large) and the number of data samples is limited ( $n$  small).

To address the issue, one needs to exploit certain prior knowledge about the domains of interest, among which imposing statistical independency among the random variables in  $\mathbf{z}$  is a reasonable assumption to be made. In essence, imposing many entries of the inverse covariance to be 0, has an effect of reducing the number of parameters to be estimated, which consequently yields a sensible and reliable model. More importantly, in the Gaussian model, the sparse correlation directly translates to sparse inverse covariance.

Formally, letting  $\mathbf{\Omega} = \mathbf{V}^{-1}$ , this can be easily seen from the density representation,

$$P(\mathbf{z}) \propto \exp\left(-\frac{1}{2} \mathbf{z}^\top \mathbf{\Omega} \mathbf{z}\right) = \exp\left(-\frac{1}{2} \sum_{ij} z_i z_j \Omega_{ij}\right), \quad (2)$$

which can be seen as the undirected graphical model, more specifically, Gaussian Markov random field (GMRF) [18]. By the Hammersley–Clifford theorem [19] or the local/global Markov

Download English Version:

<https://daneshyari.com/en/article/531814>

Download Persian Version:

<https://daneshyari.com/article/531814>

[Daneshyari.com](https://daneshyari.com)