

# Gaussian Conditional Random Fields for Classification

Andrija Petrović<sup>a\*</sup>    Mladen Nikolić<sup>b</sup>    Miloš Jovanović<sup>a</sup>    Boris Delibašić<sup>a</sup>

<sup>a</sup> Faculty of Organizational Sciences, University of Belgrade, Center for Business Decision Making, Jove Ilica 154, 11000 Belgrade, Serbia

<sup>b</sup> Faculty of Mathematics, University of Belgrade, Department of Computer Science, Studentski trg 16, 11000 Belgrade, Serbia

## Abstract

Gaussian conditional random fields (GCRF) are a well-known used structured model for continuous outputs that uses multiple unstructured predictors to form its features and at the same time exploits dependence structure among outputs, which is provided by a similarity measure. In this paper, a Gaussian conditional random fields model for structured binary classification (GCRFBC) is proposed. The model is applicable to classification problems with undirected graphs, intractable for standard classification CRFs. The model representation of GCRFBC is extended by latent variables which yield some appealing properties. Thanks to the GCRF latent structure, the model becomes tractable, efficient and open to improvements previously applied to GCRF regression models. In addition, the model allows for reduction of noise, that might appear if structures were defined directly between discrete outputs. Additionally, two different forms of the algorithm are presented: GCRFBCb (GCRGBC - Bayesian) and GCRFBCnb (GCRFBC - non Bayesian). The extended method of local variational approximation of sigmoid function is used for solving empirical Bayes in Bayesian GCRFBCb variant, whereas MAP value of latent variables is the basis for learning and inference in the GCRFBCnb variant. The inference in GCRFBCb is solved by Newton-Cotes formulas for one-dimensional integration. Both models are evaluated on synthetic data. It was shown that both models achieve better prediction performance than unstructured predictors. Furthermore, computational and memory complexity is evaluated. Advantages and disadvantages of the proposed GCRFBCb and GCRFBCnb are discussed in detail.

Keywords: Structured classification, Gaussian conditional random fields, Empirical Bayes, Local variational approximation

\* Corresponding author: email: aapetrovic@mas.bg.ac.rs, tel.: +381 62 295 278

---

## 1 Introduction

Increased quantity and variety of sources of data with correlated outputs, so called structured data, created an opportunity for exploiting additional information between dependent outputs to achieve better prediction performance. An extensive review on topic of binary and multi-label classification with structured output is provided in [Su, 2015]. The structured classifiers were compared in terms of accuracy and speed.

One of the most successful probabilistic models for structured output classification problems are conditional random fields (CRF) [Sutton and McCallum, 2006]. CRFs were successfully applied on a variety of different structured tasks, such as: low-resource named entity recognition [Cotterell and Duh, 2017], image segmentation [Zhang et al., 2015], chord recognition [Masada and Bunescu, 2017] and word segmentation [Zia et al., 2018]. The main advantages of CRFs lies in their discriminatory nature, resulting in the relaxation of independence assumptions and the label bias problem that are present in many graphical models. Additionally, availability of exact gradient evaluation and probability information made CRFs widely used in different applications. Aside of many advantages, CRFs have also many drawbacks. Gradient computation and partition function evaluation can be computationally costly, especially for large number of feature functions. That is the reason why CRFs can be very computationally expensive during inference and learning, and consequently slow. Moreover, the CRFs with complex structure usually do not support decode-based learning [Sun and Ma, 2018]. Sometimes even the gradient computation is impossible or exact inference is intractable due to complicated partition function.

In order to solve these problems, a wide range of different algorithms have been developed and adapted for various task. The mixtures of CRFs capable to model data that come from multiple different sources or domains is presented in [Kim, 2017]. The method is related to the well known hidden-unit CRF (HUCRF) [Maaten et al., 2011]. The conditional likelihood and expectation minimization (EM) procedure for learning have been derived there. The mixtures of CRF models were implemented on several real-world applications resulting in prediction improvement. Recently, the model based on unification of deep learning and CRF was developed by [Chen et al., 2016]. The deep CRF model showed better performance compared to either shallow CRFs or deep learning methods on their own. Similarly, the combination of CRFs and deep convolutional neural networks was evaluated on an example of environmental microorganisms labeling [Kosov et al., 2018]. The spatial relations among outputs were taken in consideration and experimental results have shown satisfactory results.

Structured models for regression based on CRFs have recently been a focus of many researchers. One of the popular methods for structured regression – Gaussian conditional random fields (GCRF) – has the form of multivariate Gaussian distribution. The main assumption of the model is that the relations between outputs are presented in quadratic form. The multivariate Gaussian distribution representation of a CRF has many advantages, like convex loss function and, consequently, efficient inference and learning.

The GCRF model was first implemented for the task of low-level computer vision [Tappen et al., 2007]. Since than, various different adaptations and approximations of GCRF were proposed [Radosavljevic et al., 2014]. The parameter space for the GCRF model is extended to facilitate joint modelling of positive and negative influences [Glass et al., 2016]. In addition, the model is extended by bias term into link weight and solved as a part of convex optimization. Semi-supervised model marginalized Gaussian conditional random fields (MGCRF) for dealing with missing variables were proposed by [Stojanovic et al., 2015]. The benefits of the model were proved on partially observed data and showed better prediction performance then alternative semi-supervised structured models.

In this paper, a new model of Gaussian conditional random fields for binary classification is

proposed (GCRFBC). The model assumes that discrete outputs  $y_i$  are conditionally independent for given continuous latent variables  $z_i$  which follow a distribution modeled by a GCRF. That way, relations between discrete outputs are not expressed directly. Two different inference and learning approaches are proposed in this paper. The first one is based on evaluating empirical Bayes by marginalizing latent variables (GCRFBCb), whereas MAP value of latent variables is the basis for learning and inference in the second model (GCRFBCnb). The presented models are tested on synthetic data. This is a discrete output problem, so it is not possible to use standard GCRFs for regression.

In section 2 the related work is reviewed and the GCRF model is briefly presented. The details of the proposed models along with the inference and learning are described in section 3. Experimental results on synthetic data and real-world applications are shown in section 4. Final conclusions are given in section 5.

## 2 Related Work and Background Material

GCRF is a discriminative graph-based regression model [Radosavljevic et al., 2010]. Nodes of the graph are variables  $\mathbf{y} = (y_1, y_2, \dots, y_N)$ , which need to be predicted given a set of features  $\mathbf{x}$ . The attributes  $\mathbf{x} = (x_1, x_2, \dots, x_N)$  interact with each node  $y_i$  independently of one another, while the relations between outputs are expressed by pairwise interaction function. In order to learn parameters of the model, a training set of vectors of attributes  $\mathbf{x}$  and real-valued response variables  $\mathbf{y}$  are provided. The generalized form of the conditional distribution  $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$  is:

$$P(\mathbf{y}|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \exp \left( \sum_{i=1}^N A(\boldsymbol{\alpha}, y_i, \mathbf{x}_i) + \sum_{i \neq j} I(\boldsymbol{\beta}, y_i, y_j) \right) \quad (1)$$

Two different feature functions are used: association potential  $A(\boldsymbol{\alpha}, y_i, \mathbf{x})$  to model relations between outputs  $y_i$  and corresponding input vector  $\mathbf{x}_i$  and interaction potential  $I(\boldsymbol{\beta}, y_i, y_j)$  to model pairwise relations between nodes. Vectors  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are parameters of the association potential  $A$  and the interaction potential  $I$ , whereas  $Z$  is partition function. The association potential is defined as:

$$A(\boldsymbol{\alpha}, y_i, \mathbf{x}_i) = - \sum_{k=1}^K \alpha_k (y_i - R_k(\mathbf{x}_i))^2 \quad (2)$$

where  $R_k(\mathbf{x}_i)$  represents unstructured predictor of  $y_i$  for each node in the graph. This unstructured predictor can be any regression model that gives prediction of output  $y_i$  for given attributes  $\mathbf{x}_i$ .  $K$  is the total number of unstructured predictors. The interaction potential functions are defined as:

$$I(\boldsymbol{\beta}, y_i, y_j) = - \sum_{l=1}^L \sum_{k=1}^K \beta_l S_{ij}^l (y_i - y_j)^2 \quad (3)$$

where  $S_{ij}^l$  is value that express similarity between nodes  $i$  and  $j$  in graph  $l$ .  $L$  is the total numbers of graphs (similarity functions). Graphs can express any kind of relations between nodes e.g., spatial and temporal correlations between outputs. One of the main advantages of GCRF is the ability to express different relations between outputs by variety of graphs. Moreover, the GCRF is able to learn which graph is significant for outputs prediction.

The quadratic form of interaction and association potential enables conditional distribution  $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$  to be expressed as multivariate Gaussian distribution. The canonical form of GCRF is [Radosavljevic et al., 2010]:

$$P(\mathbf{y}|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right) \quad (4)$$

where  $|\cdot|$  denotes determinant. Precision matrix  $\Sigma^{-1} = 2Q$  and distribution mean  $\boldsymbol{\mu} = \Sigma\mathbf{b}$  is defined as, respectively:

$$Q = \begin{cases} \sum_{k=1}^K \alpha_k + \sum_{h=1}^N \sum_{l=1}^L \beta_l S_{ih}^l, & \text{if } i = j \\ -\sum_{l=1}^L \beta_l S_{ij}^l, & \text{if } i \neq j \end{cases} \quad (5)$$

$$b_i = 2 \left( \sum_{k=1}^K \alpha_k R_k(\mathbf{x}_i) \right) \quad (6)$$

Due to concavity of multivariate Gaussian distribution the inference task  $\underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$  is straightforward. The maximum posterior estimate of  $\mathbf{y}$  is the distribution expectation  $\boldsymbol{\mu}$ . The objective of the learning task is to optimize parameters  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  by maximizing conditional log likelihood  $\underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\operatorname{argmax}} \sum_{\mathbf{y}} \log P(\mathbf{y}|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ . One way to ensure positive definiteness of the covariance matrix of GCRF is to require diagonal dominance [Strang et al., 1993]. This can be ensured by imposing constraints that all elements of  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  be greater than 0 [Radosavljevic et al., 2010].

Large number of different studies connected with graph based methods for regression can be found in the literature [Fox, 2015]. A comprehensive review of continuous conditional random fields (CCRF) was provided in [Radosavljevic et al., 2010]. The sparse conditional random fields obtained by  $l_1$  regularization are first proposed and evaluated by [Wytock and Kolter, 2013]. Additionally, [Frot et al., 2018] presented GCRF with the latent variable decomposition and derived convergence bounds for the estimator that is well behaved in high dimensional regime.

One of the adaptations of GCRF on discrete output was briefly discussed in [Radosavljevic, 2011], as a part of future work directions that should be considered. Namely, the model should assume existence of latent continuous variables that follows distribution modeled by GCRF, whereas the distribution of discrete outputs  $\mathbf{y}$  is in multivariate normal distribution form with diagonal covariance matrix. The parameters of the models are obtained by EM algorithm. Moreover, since  $\mathbf{y}$  and  $\mathbf{z}$  are unknown, the inference is performed by first calculating marginal expectation for  $\mathbf{z}$ . Discrete values of outputs are found as average values of  $\mathbf{z}$  over positive and negative examples. The models developed in this paper are motivated by the preceding discussion.

### 3 Methodology

One way of adapting GCRF to classification problem is by approximating discrete outputs by suitably defining continuous outputs. Namely, GCRF can provide dependence structure over continuous variables which can be passed through sigmoid function. That way relationship between regression GCRF and classification GCRF is similar to the relationship between linear and logistic regression, but with dependent variables. Aside from allowing us to define a classification variant of GCRF, this may result in additional appealing properties:

- The model is applicable to classification problems with undirected graphs, intractable for standard classification CRFs. Thanks to the GCRF latent structure, the model becomes tractable, efficient and open to improvements previously applied to GCRF regression models.
- Defining correlations directly between discrete outputs may introduce unnecessary noise to the model [Tan et al., 2010]. This problem can be solved by defining structured relations on a latent continuous variable space.

- In case that unstructured predictors are unreliable, which is signaled by their large variance (diagonal elements in the covariance matrix), it is simple to marginalize over latent variable space and obtain better results.

It is assumed that  $y_i$  are discrete binary outputs and  $z_i$  are continuous latent variables assigned to each  $y_i$ . In addition, each output  $y_i$  is conditionally independent of the others given  $z_i$ . The illustration of dependencies expressed by GCRFBC model is presented in Fig. 1.

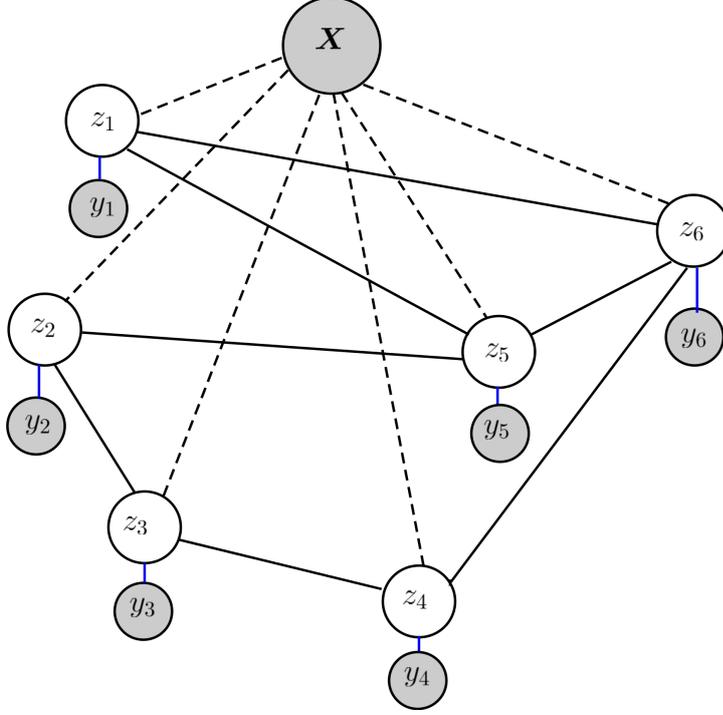


Figure 1: Graphical representation of dependencies expressed by GCRFBC model

The conditional probability distribution  $P(y_i|z_i)$  is defined as Bernoulli distribution:

$$P(y_i|z_i) = \text{Ber}(y_i|\sigma(z_i)) = \sigma(z_i)^{y_i} (1 - \sigma(z_i))^{1-y_i} \quad (7)$$

where  $\sigma(\cdot)$  is sigmoid function. Due to conditional independence assumption, the joint distribution of outputs  $y_i$  can be expressed as:

$$P(y_1, y_2, \dots, y_N | \mathbf{z}) = \prod_{i=1}^N \sigma(z_i)^{y_i} (1 - \sigma(z_i))^{1-y_i} \quad (8)$$

Furthermore, the conditional distribution  $P(\mathbf{z}|\mathbf{x})$  is the same as in the classical GCRF model and has canonical form defined by multivariate Gaussian distribution. Hence, joint distribution of continuous latent variables  $\mathbf{z}$  and outputs  $\mathbf{y}$  is:

$$P(\mathbf{y}, \mathbf{z} | \mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^N \sigma(z_i)^{y_i} (1 - \sigma(z_i))^{1-y_i} \cdot \frac{1}{(2\pi)^{N/2} |\boldsymbol{\Sigma}(\mathbf{x})|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}(\mathbf{x}))^T \boldsymbol{\Sigma}(\mathbf{x})^{-1} (\mathbf{z} - \boldsymbol{\mu}(\mathbf{x}))\right) \quad (9)$$

where  $\boldsymbol{\theta} = (\alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_L)$ .

Equation 9 presents the general form of mathematical model representation that will be further discussed in this paper.

We consider two ways of inference and learning in GCRFBC model:

- GCRFBCb - with conditional probability distribution  $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ , in which variables  $\mathbf{z}$  are marginalized over, and
- GCRFBCnb - with conditional probability distribution  $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \mu_{\mathbf{z}})$ , in which variables  $\mathbf{z}$  are substituted by their expectations.

### 3.1 Inference

#### 3.1.1 Inference in GCRFBCb Model

Prediction of discrete outputs  $\mathbf{y}$  for given features  $\mathbf{x}$  and parameters  $\boldsymbol{\theta}$  is analytically intractable due to integration of the joint distribution  $P(\mathbf{y}, \mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$  with respect to latent variables. However, due to conditional independence between nodes, it is possible to obtain  $P(y_i = 1|\mathbf{x}, \boldsymbol{\theta})$ .

$$P(y_i|\mathbf{x}, \boldsymbol{\theta}) = \int_{\mathbf{z}} P(y_i|\mathbf{z})P(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})d\mathbf{z} \quad (10)$$

$$P(y_i = 1|\mathbf{x}, \boldsymbol{\theta}) = \int_{\mathbf{z}} \sigma(z_i)P(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})d\mathbf{z} \quad (11)$$

As a result of independence properties of the distribution, it holds  $P(y_i = 1|\mathbf{z}) = P(y_i = 1|z_i)$ , and it is possible to marginalize  $P(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$  with respect to latent variables  $\mathbf{z}' = (z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_N)$ :

$$P(y_i = 1|\mathbf{x}, \boldsymbol{\theta}) = \int_{z_i} \sigma(z_i) \left( \int_{\mathbf{z}'} P(\mathbf{z}', z_i|\mathbf{x}, \boldsymbol{\theta})d\mathbf{z}' \right) dz_i \quad (12)$$

where  $\int_{\mathbf{z}'} P(\mathbf{z}', z_i|\mathbf{x}, \boldsymbol{\theta})d\mathbf{z}'$  is normal distribution with mean  $\mu = \mu_i$  and variance  $\sigma_i^2 = \Sigma_{ii}$ . Therefore, it holds:

$$P(y_i = 1|\mathbf{x}, \boldsymbol{\theta}) = \int_{-\infty}^{+\infty} \sigma(z_i)\mathcal{N}(z_i|\mu_i, \sigma_i^2)dz_i \quad (13)$$

The evaluation of  $P(y_i = 0|\mathbf{x}, \boldsymbol{\theta})$  is straightforward and is expressed as:

$$P(y_i = 0|\mathbf{x}, \boldsymbol{\theta}) = 1 - P(y_i = 1|\mathbf{x}, \boldsymbol{\theta}) \quad (14)$$

The one-dimensional integral is still analytically intractable, but can be effectively evaluated by one-dimensional numerical integration. Additionally, the surface of the function expressed by the product of the univariate normal distribution and sigmoid function is mostly concentrated closely around the mean, except in cases in which variance of normal distribution is high. The plot of function  $\sigma(z)N(z|\mu_{ii}, \sigma^2)$  with respect to the variance of normal distribution is illustrated by Fig. 2. Therefore, the limits of the integral  $(-\infty, +\infty)$  can be reasonably approximated by the interval  $(\mu - 10\sigma_i, \mu + 10\sigma_i)$ . This approximation improves integration precision, especially in case that Newton-Cotes formulas are used for numerical integration [Davis and Rabinowitz, 2007]. The proposed inference approach can be effectively used in case of huge number of nodes, due to low computational cost of one-dimensional numerical integration.

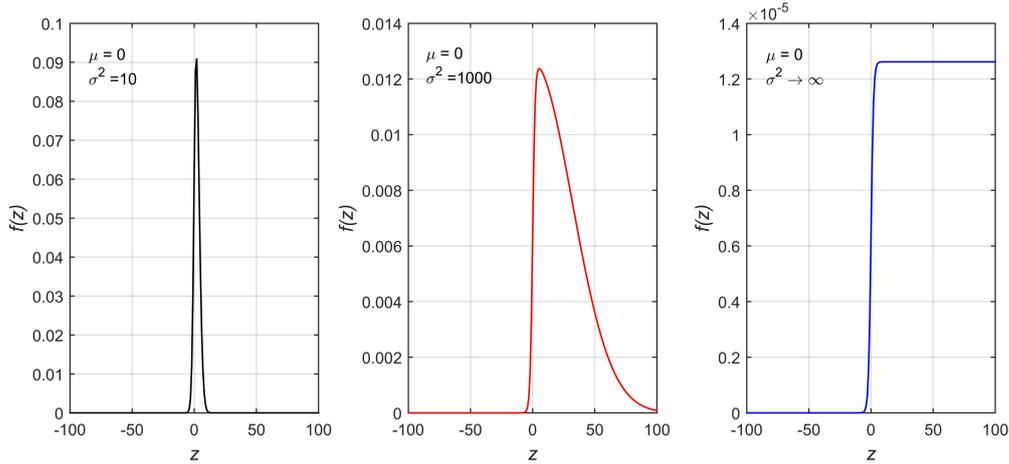


Figure 2: Shapes of function  $\sigma(z)N(z_i|\mu_{ii},\sigma^2)$  for three different choices of variance of the normal distribution.

### 3.1.2 Inference in GCRFBCnb Model

The inference procedure in GCRFBCnb is much simpler, because marginalization with respect to latent variables is not performed. To predict  $\mathbf{y}$ , it is necessary to evaluate posterior maximum of latent variable  $\mathbf{z}_{\max} = \underset{\mathbf{z}}{\operatorname{argmax}} P(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$ , which is straightforward due to normal form of GCRF. Therefore, it holds  $\mathbf{z}_{\max} = \boldsymbol{\mu}_{z,i}$ . The conditional distribution  $P(y_i = 1|\mathbf{x}, \boldsymbol{\mu}_{z,i}, \boldsymbol{\theta})$  can be expressed as:

$$P(y_i = 1|\mathbf{x}, \boldsymbol{\mu}_{z,i}, \boldsymbol{\theta}) = \sigma(\mu_{z,i}) = \frac{1}{1 + \exp(-\mu_{z,i})} \quad (15)$$

where  $\mu_{z,i}$  is expectation of latent variable  $z_i$ .

## 3.2 Learning

### 3.2.1 Learning in GCRFBCb Model

In comparison with inference, learning procedure is more complicated. Evaluation of the conditional log likelihood is intractable, since latent variables cannot be analytically marginalized. The conditional log likelihood is expressed as:

$$\begin{aligned} \mathcal{L}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) &= \log \left( \int_{\mathbf{Z}} P(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta}, \mathbf{X}) d\mathbf{Z} \right) = \sum_{j=1}^M \log \left( \int_{z_j} P(\mathbf{y}_j, z_j|\boldsymbol{\theta}, \mathbf{x}) dz_j \right) \\ &= \sum_{j=1}^M \mathcal{L}_j(\mathbf{y}_j|\mathbf{x}, \boldsymbol{\theta}) \end{aligned} \quad (16)$$

$$\mathcal{L}_j(\mathbf{y}_j|\mathbf{x}, \boldsymbol{\theta}) = \log \int_{z_j} \prod_{i=1}^N \sigma(z_{ji})^{y_{ji}} (1 - \sigma(z_{ji}))^{1-y_{ji}} \frac{\exp(-\frac{1}{2}(\mathbf{z}_j - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{z}_j - \boldsymbol{\mu}_j))}{(2\pi)^{N/2} |\boldsymbol{\Sigma}_j|^{1/2}} dz_j \quad (17)$$

where  $\mathbf{Y} \in \mathbb{R}^{M \times N}$  is complete dataset of outputs,  $\mathbf{X} \in \mathbb{R}^{M \times N \times A}$  is complete dataset of features,  $M$  is the total number of instances and  $A$  is the total number of features. Please note that each instance is structured, so while different instances are independent of each other, variables within one instance are dependent.

One way to approximate integral in conditional log likelihood is by local variational approximation. [Jaakkola and Jordan, 2000] derived lower bound for sigmoid function, which can be expressed as:

$$\sigma(x) \geq \sigma(\xi) \exp\{(x - \xi)/2 - \lambda(\xi)(x^2 - \xi^2)\} \quad (18)$$

where  $\lambda(\xi) = -\frac{1}{2\xi} \cdot \left[\sigma(\xi) - \frac{1}{2}\right]$  and  $\xi$  is a variational parameter. The Eq. 18 is called  $\xi$  transformation of sigmoid function and it yields maximum value when  $\xi = x$ . The sigmoid function with lower bound is shown in Fig. 3.

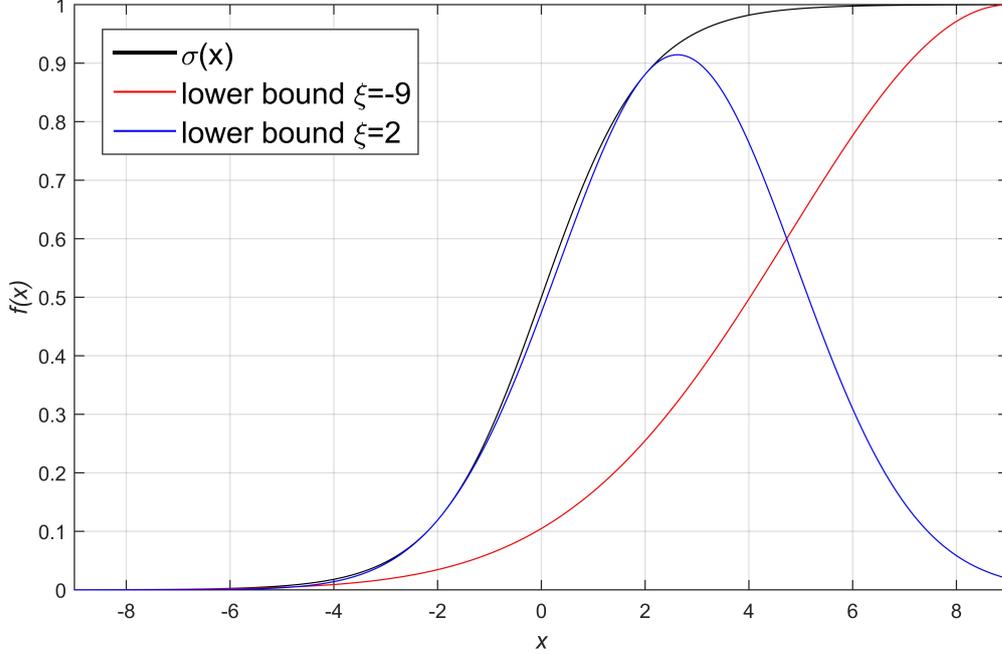


Figure 3: The sigmoid function with its lower bound

This approximation can be applied to the model defined by Eq. 17, but the variational approximation has to be further extended because of the product of sigmoid functions, such that:

$$P(\mathbf{y}_j, \mathbf{z}_j | \boldsymbol{\theta}, \mathbf{x}) = P(\mathbf{y}_j | \mathbf{z}_j) P(\mathbf{z}_j | \mathbf{x}, \boldsymbol{\theta}) \geq \underline{P}(\mathbf{y}_j, \mathbf{z}_j | \boldsymbol{\theta}, \mathbf{x}, \boldsymbol{\xi}_j) \quad (19)$$

$$\underline{P}(\mathbf{y}_j, \mathbf{z}_j | \boldsymbol{\theta}, \mathbf{x}, \boldsymbol{\xi}_j) = \prod_{i=1}^N \sigma(\xi_{ji}) \exp\left(z_{ji} y_{ji} - \frac{z_{ji} + \xi_{ji}}{2} - \lambda(\xi_{ji})(z_{ji}^2 - \xi_{ji}^2)\right) \cdot \frac{1}{(2\pi)^{N/2} |\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}_j - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{z}_j - \boldsymbol{\mu}_j)\right) \quad (20)$$

The Eq. 20 can be arranged in the form suitable for integration. The lower bound of conditional log likelihood  $\underline{\mathcal{L}}(\mathbf{y}_j | \boldsymbol{\theta}, \mathbf{x}, \boldsymbol{\xi}_j)$  is defined as:

$$\underline{\mathcal{L}}(\mathbf{y}_j | \mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{\xi}_j) = \log \underline{P}(\mathbf{y}_j | \mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{\xi}_j) = \sum_{i=1}^N \left( \log \sigma(\xi_{ji}) - \frac{\xi_{ji}}{2} + \lambda(\xi_{ji}) \xi_{ji}^2 \right) - \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j + \frac{1}{2} \mathbf{m}_j^T \mathbf{S}_j^{-1} \mathbf{m}_j + \frac{1}{2} \log |\mathbf{S}_j| \quad (21)$$

where:

$$\mathbf{S}_j^{-1} = \boldsymbol{\Sigma}_j^{-1} + 2\boldsymbol{\Lambda}_j \quad (22)$$

$$\mathbf{m}_j = \Sigma_j \left( (\mathbf{y}_j - \frac{1}{2}\mathbf{I}) + \Sigma_j^{-1} \boldsymbol{\mu}_j \right) \quad (23)$$

$$\Lambda_j = \begin{bmatrix} \lambda(\xi_{j1}) & 0 & 0 & \dots & 0 \\ 0 & \lambda(\xi_{j2}) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda(\xi_{jN}) \end{bmatrix} \quad (24)$$

GCRFBCb uses the derivative of conditional log likelihood in order to find the optimal values for parameters  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$  and matrix of variational parameters  $\boldsymbol{\xi} \in \mathbb{R}^{M \times N}$  by gradient ascent method. In order to ensure positive definiteness of normal distribution involved, it is sufficient to constrain parameters  $\boldsymbol{\alpha} > 0$  and  $\boldsymbol{\beta} > 0$ . The partial derivative of conditional log likelihood  $\frac{\partial \underline{\mathcal{L}}_j(\mathbf{y}_j | \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\xi}_j)}{\partial \alpha_k}$  is computed as:

$$\begin{aligned} \frac{\partial \underline{\mathcal{L}}_j(\mathbf{y}_j | \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\xi}_j)}{\partial \alpha_k} &= -\frac{1}{2} \text{Tr} \left( S_j \frac{\partial S_j^{-1}}{\partial \alpha_k} \right) + \frac{\partial \mathbf{m}_j^T}{\partial \alpha_k} S_j^{-1} \mathbf{m}_j + \frac{1}{2} \mathbf{m}_j^T \frac{\partial S_j^{-1}}{\partial \alpha_k} \mathbf{m}_j \\ &\quad - \frac{\boldsymbol{\mu}_j^T}{\partial \alpha_k} \Sigma_j^{-1} \boldsymbol{\mu}_j - \frac{1}{2} \boldsymbol{\mu}_j^T \frac{\partial \Sigma_j^{-1}}{\partial \alpha_k} + \frac{1}{2} \text{Tr} \left( \Sigma_j \frac{\partial \Sigma_j^{-1}}{\partial \alpha_k} \right) \end{aligned} \quad (25)$$

where:

$$\frac{\partial S_j^{-1}}{\partial \alpha_k} = \frac{\partial \Sigma_j^{-1}}{\partial \alpha_k} = \begin{cases} 2, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \quad (26)$$

$$\frac{\partial \mathbf{m}_j^T}{\partial \alpha_k} = - \left( \mathbf{y}_j - \frac{1}{2}\mathbf{I} + \boldsymbol{\mu}_j^T \Sigma_j^{-1} \right) S_j \frac{\partial S_j^{-1}}{\partial \alpha_k} S_j + \frac{\partial \boldsymbol{\mu}_j^T}{\partial \alpha_k} \Sigma_j^{-1} S_j + \boldsymbol{\mu}_j^T \frac{\partial \Sigma_j^{-1}}{\partial \alpha_k} S_j \quad (27)$$

$$\frac{\partial \boldsymbol{\mu}_j^T}{\partial \alpha_k} = \left( 2\alpha_k R_k(\mathbf{x}) - \frac{\partial \Sigma_j^{-1}}{\partial \alpha_k} \boldsymbol{\mu}_j \right)^T \Sigma_j^T \quad (28)$$

Similarly partial derivatives with respect to  $\boldsymbol{\beta}$  can be defined as:

$$\begin{aligned} \frac{\partial \underline{\mathcal{L}}_j(\mathbf{y}_j | \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\xi}_j)}{\partial \beta_l} &= -\frac{1}{2} \text{Tr} \left( S_j \frac{\partial S_j^{-1}}{\partial \beta_l} \right) + \frac{\partial \mathbf{m}_j^T}{\partial \beta_l} S_j^{-1} \mathbf{m}_j + \frac{1}{2} \mathbf{m}_j^T \frac{\partial S_j^{-1}}{\partial \beta_l} \mathbf{m}_j \\ &\quad - \frac{\boldsymbol{\mu}_j^T}{\partial \beta_l} \Sigma_j^{-1} \boldsymbol{\mu}_j - \frac{1}{2} \boldsymbol{\mu}_j^T \frac{\partial \Sigma_j^{-1}}{\partial \beta_l} + \frac{1}{2} \text{Tr} \left( \Sigma_j \frac{\partial \Sigma_j^{-1}}{\partial \beta_l} \right) \end{aligned} \quad (29)$$

where:

$$\frac{\partial S_j^{-1}}{\partial \beta_l} = \frac{\partial \Sigma_j^{-1}}{\partial \beta_l} = \begin{cases} \sum_{n=1}^N e_{in}^l S_{in}^l(x), & \text{if } i = j \\ -e_{ij}^l S_{ij}^l(x), & \text{if } i \neq j \end{cases} \quad (30)$$

$$\frac{\partial \mathbf{m}_j^T}{\partial \beta_l} = - \left( \mathbf{y}_j - \frac{1}{2}\mathbf{I} + \boldsymbol{\mu}_j^T \Sigma_j^{-1} \right) S_j \frac{\partial S_j^{-1}}{\partial \beta_l} S_j + \frac{\partial \boldsymbol{\mu}_j^T}{\partial \beta_l} \Sigma_j^{-1} S_j + \boldsymbol{\mu}_j^T \frac{\partial \Sigma_j^{-1}}{\partial \beta_l} S_j \quad (31)$$

$$\frac{\partial \boldsymbol{\mu}_j^T}{\partial \beta_l} = \left( -\frac{\partial \Sigma_j^{-1}}{\partial \beta_l} \boldsymbol{\mu}_j \right)^T \Sigma_j^T \quad (32)$$

In the same manner partial derivatives of conditional log likelihood with respect to  $\xi_{ji}$  are:

$$\begin{aligned} \frac{\partial \underline{\mathcal{L}}_j(\mathbf{y}_j|\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\xi}_j)}{\partial \xi_{ji}} &= -\frac{1}{2} \text{Tr} \left( 2\mathcal{S}_j \frac{\partial \Lambda_j}{\partial \xi_{ji}} \right) - \left[ 2 \left( \mathbf{y}_j - \frac{1}{2} \mathbf{I} \right) \mathcal{S}_j \frac{\partial \Lambda_j}{\partial \xi_{ji}} \mathcal{S}_j \right] \mathcal{S}_j^{-1} \mathbf{m}_j \\ &+ \mathbf{m}_j^T \frac{\partial \Lambda_j}{\partial \xi_{ji}} \mathbf{m}_j + \sum_{i=1}^N \left( \left( \frac{1}{\sigma(\xi_{ji})} + \frac{1}{2} \xi_{ji} \right) \frac{\partial \sigma(\xi_{ji})}{\partial \xi_{ji}} + \frac{1}{2} \left( \sigma(\xi_{ji}) - \frac{3}{4} \right) \right) \end{aligned} \quad (33)$$

where:

$$\frac{\partial \Lambda_j}{\partial \xi_{ji}} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \frac{\partial \lambda(\xi_{ji})}{\partial \xi_{ji}} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \quad (34)$$

$$\frac{\partial \sigma(\xi_{ji})}{\partial \xi_{ji}} = \sigma(\xi_{ji})(1 - \sigma(\xi_{ji})) \quad (35)$$

$$\frac{\partial \lambda(\xi_{ji})}{\partial \xi_{ji}} = \frac{1}{2\xi_{ji}} \frac{\partial \sigma(\xi_{ji})}{\partial \xi_{ji}} - \frac{1}{2} \left( \sigma(\xi_{ji}) - \frac{1}{2} \right) \frac{1}{\xi_{ji}^2} \quad (36)$$

Gradient ascent algorithm cannot be directly applied to constrained optimization problems. There are several procedures that can be applied in constrained problem optimization. The first one involves log transformation and it was presented in [Radosavljevic et al., 2010]. The procedure can be further extended by some of the adaptive learning parameter methods. However, in this paper due to large number of parameters, the truncated Newton algorithm for constrained optimization (TNC) was used. More details about TNC can be found in [Nocedal and Wright, 2006] and [Facchinei et al., 2002]. It is necessary to emphasize that the conditional log likelihood is not convex function of parameters  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$  and  $\boldsymbol{\xi}$ . Because of this, finding a global optimum cannot be guaranteed.

### 3.2.2 Learning in GCRFBCnb Model

Learning in GCRFBCnb model is simpler compared to the GCRFBCb algorithm, because instead of marginalization, the mode of posterior distribution of continuous latent variable  $\mathbf{z}$  is evaluated directly so there is no need for approximation technique. The conditional log likelihood can be expressed as:

$$\underline{\mathcal{L}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\mu}) = \log P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\mu}) = \sum_{j=1}^M \sum_{i=1}^N \log P(y_{ji}|\mathbf{x}, \boldsymbol{\theta}, \mu_{ji}) = \sum_{j=1}^M \sum_{i=1}^N \underline{\mathcal{L}}_{ji}(y_{ji}|\mathbf{x}, \boldsymbol{\theta}, \mu_{ji}) \quad (37)$$

$$\underline{\mathcal{L}}_{ji}(y_{ji}|\mathbf{x}, \boldsymbol{\theta}, \mu_{ji}) = y_{ji} \log \sigma(\mu_{ji}) + (1 - y_{ji}) \log (1 - \sigma(\mu_{ji})) \quad (38)$$

The derivatives of the conditional log likelihood with respect to  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are defined as, respectively:

$$\frac{\partial \underline{\mathcal{L}}_{ji}(y_{ji}|\mathbf{x}, \boldsymbol{\theta}, \mu_{ji})}{\partial \alpha_k} = (y_{ji} - \sigma(\mu_{ji})) \frac{\partial \mu_{ji}}{\partial \alpha_k} \quad (39)$$

$$\frac{\partial \underline{\mathcal{L}}_{ji}(y_{ji}|\mathbf{x}, \boldsymbol{\theta}, \mu_{ji})}{\partial \beta_l} = (y_{ji} - \sigma(\mu_{ji})) \frac{\partial \mu_{ji}}{\partial \beta_l} \quad (40)$$

where  $\frac{\partial \mu_{ji}}{\partial \alpha_k}$  and  $\frac{\partial \mu_{ji}}{\partial \beta_l}$  are elements of the vectors  $\frac{\partial \mu_j}{\partial \alpha_k}$  and  $\frac{\partial \mu_j}{\partial \beta_l}$  and can be obtained by Eqs. 28 and 32, respectively.

In a similar manner TNC or log transformation gradient ascent algorithms can be used. Additionally, an iterative sequential quadratic programming for constrained nonlinear optimization can be used, as a result of small number of optimization parameters [Boggs and Tolle, 1995].

## 4 Experimental Evaluation

Both proposed models were tested and compared on synthetic data. All methods are implemented in Python and experiments were run on Ubuntu server with 128 GB of memory and Intel Xeon 2.9 GHz CPU. All used codes are publicly available.<sup>1</sup>

To calculate classification performance of all presented classifiers, the area under ROC curve (AUC) score was used. The AUC score assumes that the classifier outputs a real value for each instance and estimates a probability that for two randomly chosen instances from two different classes the instance from the positive class will have higher value than the instance from the negative class [Mohri et al., 2018]. A score of 1 indicates perfect classification, whereas score of 0.5 indicates random prediction performance. Aside of AUC score, the lower bound (in case of GCRFBCb) of conditional log likelihood  $\underline{\mathcal{L}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\mu})$  and actual value (in case of GCRFBCnb) of conditional log likelihood  $\mathcal{L}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$  of obtained values on synthetic test dataset were also reported.

### 4.1 Synthetic Dataset

The main goal of experiments on synthetic datasets was to examine models under various controlled conditions and show advantages and disadvantages of each. In all experiments on synthetic datasets two different graphs were used (hence  $\boldsymbol{\beta} \in \mathbb{R}^2$ ) and two unstructured predictors (hence  $\boldsymbol{\alpha} \in \mathbb{R}^2$ ). In order to generate and label nodes in graph, edge weights  $S$  and unstructured predictor values  $R$  were randomly generated from uniform distribution. Besides, it was necessary to choose values of parameters  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ . Greater values of  $\boldsymbol{\alpha}$  indicate that model is more confident about performance of unstructured predictors, whereas for the larger value of  $\boldsymbol{\beta}$  model is putting more emphasis on the dependence structure of output variables.

For generated  $S$ ,  $R$ , and given parameters  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ , probabilities of outputs are obtained and labeling is performed according to the threshold of 0.5. The complete dataset with unstructured predictors, dependence structure and labeled nodes is used for optimizing parameters  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ . Additionally, 20% of all data was used for testing and 80% for training procedure.

#### 4.1.1 Prediction Performance Evaluation

The main goal of this experiment is to evaluate how the selection of parameters  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  in data generating process affects prediction performance of GCRFBCb and GCRFBCnb. Six different values of parameters  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  were used. The values of parameters were separated in three distinct group:

1. The first group, in which  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  have similar values. Hence, unstructured predictors and dependence structure between outputs have similar importance.
2. The second group, in which  $\boldsymbol{\alpha}$  has higher values compared to  $\boldsymbol{\beta}$ , which means that model is putting more emphasis on unstructured predictors in comparison with dependence structure.

---

<sup>1</sup><https://github.com/andrijaster>

Table 1: Comparison of GCRFBCb and GCRFBCnb prediction performance for different values of  $\alpha$  and  $\beta$ , as measured by AUC, log likelihood, and norm of diagonal elements of the covariance matrix

No.	Parameters	GCRFBCb			GCRFBCnb	
		AUC	$\underline{\mathcal{L}}(\mathbf{Y} \mathbf{X}, \boldsymbol{\theta})$	$\ \boldsymbol{\sigma}\ _2$	AUC	$\mathcal{L}(\mathbf{Y} \mathbf{X}, \boldsymbol{\theta})$
1	$\alpha = [5, 4]$ $\beta = [5, 22]$	0.812	-71.150	0.000	0.812	-71.151
2	$\alpha = [1, 18]$ $\beta = [1, 18]$	0.903	-75.033	0.001	0.902	-75.033
3	$\alpha = [22, 21]$ $\beta = [5, 22]$	0.988	-83.957	0.000	0.988	-83.957
4	$\alpha = [22, 21]$ $\beta = [0.1, 0.67]$	0.866	-83.724	0.000	0.886	-83.466
5	$\alpha = [0.8, 0.5]$ $\beta = [5, 22]$	0.860	-83.353	34.827	0.817	-84.009
6	$\alpha = [0.2, 0.4]$ $\beta = [1, 18]$	0.931	-70.692	35.754	0.821	-70.391

- The third group, in which  $\beta$  has higher values compared to  $\alpha$ , thus model is putting more emphasis on dependence structure and less on unstructured predictors.

Along with the AUC and conditional log likelihood, norm of the variances of latent variables (diagonal elements in the covariance matrix) is evaluated and presented in Table 1. It can be noticed, in cases where norm of the variances of latent variables is insignificant, both models have equal performance considering AUC and conditional log likelihood  $\underline{\mathcal{L}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$ . This is the case when values of parameters  $\alpha$  used in data generating process are greater or equal than values of parameters  $\beta$ . Therefore, conditional distribution  $P(\mathbf{y}, \mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$  is highly concentrated around mean value and MAP estimate is a satisfactory approximation. However, when data were generated from distribution with significantly higher values of  $\beta$ , compared to  $\alpha$  the GCRFBCb performs significantly better than GCRFBCnb. For the larger values of variance norm this difference is also large. It can be concluded that GCRFBCb has at least equal prediction performance as GCRFBCnb. Also, it can be argued that the models were generally able to utilize most of the information (from both features and the structure between outputs), which can be seen through AUC values.

#### 4.1.2 Runtime Evaluation

The computational and memory complexity of GCRFBCnb during learning and inference is same as time complexity of standard GCRF [Radosavljevic et al., 2014]. If the training lasts  $T$  iterations, overall complexity of GCRF is  $O(TN^3)$ . However, this is the worst case performance and in case of sparse precision matrix, this can be reduced to  $O(TN^2)$ . The additional memory complexity of GCRF is negligible, which holds for GCRFBCnb, too.

However, in the case of GCRFBCb memory complexity during training is  $O(M)$  due to dependency of variational parameters on the number of instances. Computational complexity is also higher –  $O(TMN^3)$ , which can also be reduced to  $O(TMN^2)$  in case of sparse precision matrix.

The following speed tests of GCRFBCb and GCRFBCnb were conducted on synthetically generated data with varying numbers of parameters and nodes. In Figs. 4(a) and 4(b) the computation time of both models with respect to number of instances is presented. The number of nodes in both models is 4. The larger number of instances have significant impact on increase

of computation time. Figs. 4(c) and 4(d) present computation time with respect to number of nodes, while holding constant value of product of number of instances and nodes (i.e. total number of values of  $y$ ). It can be seen that while holding constant value of products of number of instances and nodes, that computational time increases faster with larger number of instances compared to larger number of nodes.

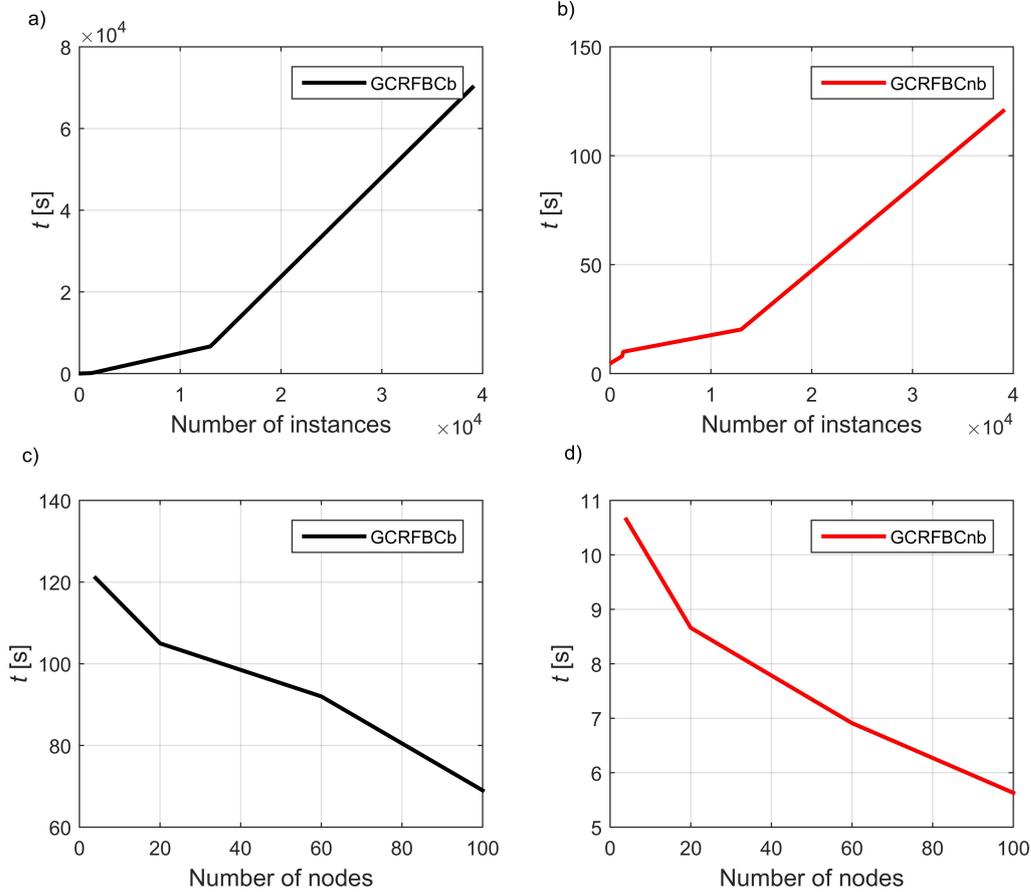


Figure 4: The computational time of GCRFBCb and GCRFBCnb with respect to number of instances and nodes

## 5 Conclusion

In this paper, a new model, called Gaussian conditional random fields for binary classification (GCRFBC) is presented. The model is based on latent GCRF structure, which means that intractable structured classification problem can become tractable and efficiently solved. Moreover, improvements previously applied to regression GCRF can be easily extended to GCRFBC. Two different variants of GCRFBC were derived: GCRFBCb and GCRFBCnb. Empirical Bayes (marginalization of latent variables) by local variational methods is used in optimization procedure of GCRFBCb, whereas MAP estimate of latent variables is applied in GCRFBCnb. Based on presented methodology and obtained experimental results on synthetic datasets, several key findings can be summarized:

- Both models GCRFBCb and GCRFBCnb have better prediction performance compared to the unstructured predictors

- GCRFBCb has better performance considering AUC score and lower bound of conditional log likelihood  $\underline{\mathcal{L}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$  compared to GCRFBCnb, in cases where norm of the variances of latent variables is high. However, in cases where norm of the variances is close to zero, both models have equal prediction performance.
- Due to high memory and computational complexity of GCRFBCb compared to GCRFBCnb, in cases where norm of the variances is close to zero, it is reasonable to use GCRFBCnb. Additionally, the trade off between complexity and accuracy can be made in situation where norm of the variances is high.

Further studies should address extending GCRFBC to structured multi-label classification problems, and lower computational complexity of GCRFBCb by considering efficient approximations.

## Acknowledgements

This research is partially supported by the Ministry of Science, Education and Technological Development of the Republic of Serbia grants OI174021, TR35011 and TR41008. The authors would like to express gratitude to company Saga d.o.o Belgrade, for supporting this research.

## References

- [Boggs and Tolle, 1995] Boggs, P. T. and Tolle, J. W. (1995). Sequential quadratic programming. *Acta numerica*, 4:1–51.
- [Boutell et al., 2004] Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. (2004). Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771.
- [Chen et al., 2016] Chen, G., Li, Y., and Srihari, S. N. (2016). Word recognition with deep conditional random fields. *arXiv preprint arXiv:1612.01072*.
- [Cotterell and Duh, 2017] Cotterell, R. and Duh, K. (2017). Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 91–96.
- [Davis and Rabinowitz, 2007] Davis, P. J. and Rabinowitz, P. (2007). *Methods of numerical integration*. Courier Corporation.
- [Elisseeff and Weston, 2002] Elisseeff, A. and Weston, J. (2002). A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687.
- [Facchinei et al., 2002] Facchinei, F., Lucidi, S., and Palagi, L. (2002). A truncated newton algorithm for large scale box constrained optimization. *SIAM Journal on Optimization*, 12(4):1100–1125.
- [Fox, 2015] Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.
- [Frot et al., 2018] Frot, B., Jostins, L., and McVean, G. (2018). Graphical model selection for gaussian conditional random fields in the presence of latent variables. *Journal of the American Statistical Association*, (just-accepted).

- 
- [Glass et al., 2016] Glass, J., Ghalwash, M. F., Vukicevic, M., and Obradovic, Z. (2016). Extending the modelling capacity of gaussian conditional random fields while learning faster. In *AAAI*, pages 1596–1602.
- [Jaakkola and Jordan, 2000] Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.
- [Kim, 2017] Kim, M. (2017). Mixtures of conditional random fields for improved structured output prediction. *IEEE transactions on neural networks and learning systems*, 28(5):1233–1240.
- [Kosov et al., 2018] Kosov, S., Shirahama, K., Li, C., and Grzegorzec, M. (2018). Environmental microorganism classification using conditional random fields and deep convolutional neural networks. *Pattern Recognition*, 77:248–261.
- [Maaten et al., 2011] Maaten, L., Welling, M., and Saul, L. (2011). Hidden-unit conditional random fields. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 479–488.
- [Masada and Bunescu, 2017] Masada, K. and Bunescu, R. C. (2017). Chord recognition in symbolic music using semi-markov conditional random fields. In *ISMIR*, pages 272–278.
- [Mohri et al., 2018] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.
- [Nocedal and Wright, 2006] Nocedal, J. and Wright, S. J. (2006). *Numerical optimization* 2nd.
- [Radosavljevic, 2011] Radosavljevic, V. (2011). *Gaussian conditional random fields for regression in remote sensing*. Temple University.
- [Radosavljevic et al., 2010] Radosavljevic, V., Vucetic, S., and Obradovic, Z. (2010). Continuous conditional random fields for regression in remote sensing. In *ECAI*, pages 809–814.
- [Radosavljevic et al., 2014] Radosavljevic, V., Vucetic, S., and Obradovic, Z. (2014). Neural gaussian conditional random fields. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 614–629. Springer.
- [Silverman, 2018] Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.
- [Stojanovic et al., 2015] Stojanovic, J., Jovanovic, M., Gligorijevic, D., and Obradovic, Z. (2015). Semi-supervised learning for structured regression on partially observed attributed graphs. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 217–225. SIAM.
- [Strang et al., 1993] Strang, G., Strang, G., Strang, G., and Strang, G. (1993). *Introduction to linear algebra*, volume 3. Wellesley-Cambridge Press Wellesley, MA.
- [Su, 2015] Su, H. (2015). *Multilabel Classification through Structured Output Learning - Methods and Applications*. Aalto University.
- [Sun and Ma, 2018] Sun, X. and Ma, S. (2018). Conditional random fields with decode-based learning: Simpler and faster.
- [Sutton and McCallum, 2006] Sutton, C. and McCallum, A. (2006). *An introduction to conditional random fields for relational learning*, volume 2. Introduction to statistical relational learning. MIT Press.

- 
- [Tan et al., 2010] Tan, C., Tang, J., Sun, J., Lin, Q., and Wang, F. (2010). Social action tracking via noise tolerant time-varying factor graphs. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1049–1058. ACM.
- [Tappen et al., 2007] Tappen, M. F., Liu, C., Adelson, E. H., and Freeman, W. T. (2007). Learning gaussian conditional random fields for low-level vision. In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE.
- [Trohidis et al., 2008] Trohidis, K., Tsoumakas, G., Kalliris, G., and Vlahavas, I. P. (2008). Multi-label classification of music into emotions. In ISMIR, volume 8, pages 325–330.
- [Wytock and Kolter, 2013] Wytock, M. and Kolter, Z. (2013). Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In International conference on machine learning, pages 1265–1273.
- [Zhang et al., 2015] Zhang, P., Li, M., Wu, Y., and Li, H. (2015). Hierarchical conditional random fields model for semisupervised sar image segmentation. IEEE Transactions on Geoscience and Remote Sensing, 53(9):4933–4951.
- [Zia et al., 2018] Zia, H. B., Raza, A. A., and Athar, A. (2018). Urdu word segmentation using conditional random fields (crfs). arXiv preprint arXiv:1806.05432.