CrossMark

Gaussian conditional random fields extended for directed graphs

Tijana Vujicic[1] · **Jesse Glass**[2] · **Fang Zhou**[2] ·
Zoran Obradovic[2]

Abstract For many real-world applications, structured regression is commonly used for predicting output variables that have some internal structure. Gaussian conditional random fields (GCRF) are a widely used type of structured regression model that incorporates the outputs of unstructured predictors and the correlation between objects in order to achieve higher accuracy. However, applications of this model are limited to objects that are symmetrically correlated, while interaction between objects is asymmetric in many cases. In this work we propose a new model, called Directed Gaussian conditional random fields (DirGCRF), which extends GCRF to allow modeling asymmetric relationships (e.g. friendship, influence, love, solidarity, etc.). The DirGCRF models the response variable as a function of both the outputs of unstructured predictors and the asymmetric structure. The effectiveness of the proposed model is characterized on six types of synthetic datasets and four real-world applications where DirGCRF was consistently more accurate than the standard GCRF model and baseline unstructured models.

Keywords Structured regression · Gaussian conditional random fields · Asymmetric structure · Directed Gaussian conditional random fields

✉ Zoran Obradovic
zoran.obradovic@temple.edu

Tijana Vujicic
tijana.vujicic@unimediteran.net

Jesse Glass
jesse.glass@temple.edu

Fang Zhou
fang.zhou@temple.edu

[1] Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia

[2] Department of Computer and Information Sciences, Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA

Springer

1 Introduction

Structured regression models are designed to use relationships between objects for predicting output variables. In other words, structured regression models are using the given attributes and dependencies between the outputs to make predictions. This prior knowledge about relationships among the outputs is application-specific. For example relationships between hospitals can be based on similarity of their specialization (Polychronopoulou and Obradovic 2014), relationships between pairs of scientific papers can be presented as the similarity of sequences of citation (Slivka et al. 2014), relationships between documents can be quantified based on similarity of their contents (Radosavljevic et al. 2014), etc. The Gaussian conditional random fields (GCRF) model is a type of structured regression model that incorporates the outputs of unstructured predictors (based on the given attribute values) and the correlation between output variables in order to achieve higher prediction accuracy. This model was first applied in computer vision (Liu et al. 2007), but since then it has been used in different applications (Polychronopoulou and Obradovic 2014; Radosavljevic et al. 2010; Uversky et al. 2013), and extended for various purposes (Glass et al. 2015; Slivka et al. 2014; Stojkovic et al. 2016). A main assumption in the GCRF model is that if two objects are closely related, they should be more similar to each other and they should have similar values of the output variable. The similarity considered in GCRF is symmetric. However, in many real-world networks objects are asymmetrically linked (Beguerisse-Díaz et al 2014). Therefore, one limitation of the GCRF model is that the direction of link is neglected.

Networked data (such as social networks, traffic networks, information networks, etc.) are naturally modeled as graphs, where objects are represented as nodes, and relations are represented as edges between nodes. Many of these objects have directed links. For example, friendship strength is often not symmetric. In empirical studies (Michell and Amos 1997; Snijders et al. 2010) of friendship networks, participants are typically asked to identify their friends and to mark how close friends they are, which results in a directed graph in which friendships often run in only one direction between a pair of individuals. Another example is in social networks, such as Twitter or GitHub, where a user could follow all tweets posted by another user, or a developer could follow the work conducted by another developer. Also, in the email system, each individual communicates with one or more individuals by sending and receiving email messages, which results in a directed graph in which each edge has the number of sent emails as its weight.

The similarity matrix that quantifies the connections among the nodes of the graphs presented in these examples is an asymmetric matrix, and GCRF could not be directly applied on it since this model requires a symmetric matrix. One possible solution for this problem is to convert the similarity matrix from asymmetric to symmetric, which will probably cause loss of accuracy. To elaborate on this problem, we give an example of a relationship network at Fig. 1. Figure 1a presents the graph in which three nodes (marked A, B, and C) are linked with edges for which weights represent influence, in the sense that higher weight means higher influence. The edge from A to B means that A is influenced by B with weight 25. From this figure, we can conclude that node A is influenced by node C much more than by node B. On other hand node C is very much influenced by B, and not influenced at all by A. The influences from B to A, and from A to B, are the same. Converting this graph to an undirected one using the average approach results in the graph that is presented at Fig. 1b. If we look at this graph we will come to very different conclusions. Now the influence is bidirectional, which implies that connected nodes are mutually influenced with the same weight. Influence values on the relations B−A and C−A now have the same value, and nodes B and C
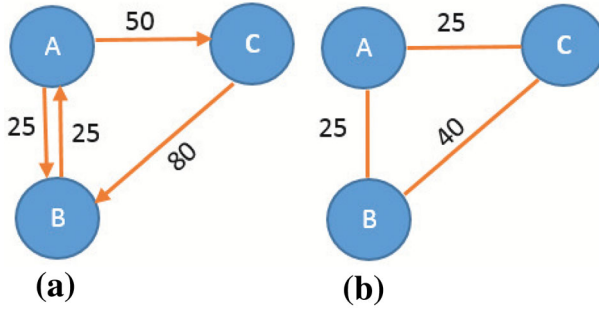
Fig. 1 An illustrative example of a graph that represents influence between objects. **a** Directed graph and **b** Undirected graph
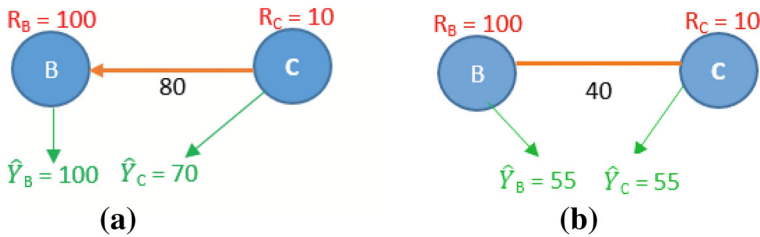

Fig. 2 An illustrative example of graph that represents influence between objects B and C, with corresponding values of the predicted output ($\hat{y}_B$ and $\hat{y}_C$). R_B and R_C are values predicted by unstructured predictor for nodes B and C. **a** Directed graph and **b** Undirected graph

are mutually influenced with a high weight. At Fig. 2 we presented how these structures can affect the predicted values, on the example of nodes B and C. First we will assume that values predicted by unstructured predictor (R) for nodes B and C are 100 and 10, respectively. At Fig. 2a these nodes are asymmetrically influenced, which means that predicted output ($\hat{\mathbf{y}}$) for the node C will get close to the output for the node B. On other hand, at Fig. 2b these nodes are symmetrically influenced, which means that predicted outputs for the nodes C and B will get close to each other. This example clearly illustrates how direction of the relation can affect the predicted value of the output.

In this work, we propose a new model, called Directed Gaussian Conditional Random Fields (DirGCRF), which extends the GCRF model by considering asymmetric similarity. The DirGCRF models the response variable as a function of both the outputs of unstructured predictors and the asymmetric structure. To evaluate the proposed model, we tested it on both synthetic and real-world datasets and compared its accuracy with standard GCRF, as well as with unstructured predictors Neural Networks and Linear Regression and simple Last and Average methods. All datasets and codes are publicly available.

We summarize contributions of this work as follows:

1. This is the first work that considers asymmetric links between objects in GCRF-based structured regression.
2. The proposed model considers both asymmetric structure and the outputs of unstructured predictor.
3. The effectiveness of the proposed directed model is characterized by experiments on six types of synthetic datasets and four real-world applications.

Table 1 Survey of graph models literature

Method	Network	Year	Purpose	Cold start
GLS (Altken 1935)	Discriminative	1935	Bias reduction	Yes, $O(f)$
GCRF (Radosavljevic et al. 2010)	Discriminative	2010	Multiple output	Yes, $O(n^3)$
SpGCRF (Wytock and Kolter 2013)	Generative	2013	Multiple output	No
Network lasso (Hallac et al. 2015)	Discriminative	2015	Multiple output	Yes, $O(nf)$

f is number of features, n is number of nodes

In the following, we first review related work in Sect. 2, followed by the details of the proposed method in Sect. 3. In Sect. 4 we provide the details about the datasets used and the experimental setup, as well as present experimental results. Finally, Sect. 5 consists of a summary of our findings, as well as future directions we intend to undertake with this project.

2 Related work

There exists a large corpus of research on regression and classification using graph based models (Table 1). Each approach takes different inputs and has various benefits and drawbacks. Some of these methods (Wytock and Kolter 2013) learn relationships between nodes from attributes. These are referred to as generative networks. On the other hand, discriminative network requires inputing the network structure (Radosavljevic et al. 2010; Altken 1935; Hallac et al. 2015). The origins of Gaussian conditional random fields (GCRF) model (Radosavljevic et al. 2010) lie in generalized least squares (GLS) (Altken 1935). In that model relationships between outputs are observed and affect the Mahalanobis distance in order to reduce training bias. GCRF leverages the same idea for multiple output regression.

None of the above models can handle asymmetric link weights. However, this work is focused on advancing the GCRF model because it produces high accuracy and it is the most scalable learning approach of all listed above (Glass et al. 2015). GCRF has been used on a broad set of applications: climate (Radosavljevic et al. 2010, 2014; Djuric et al. 2015), energy forecasting (Wytock and Kolter 2013; Guo 2013), healthcare (Gligorijevic et al. 2015; Polychronopoulou and Obradovic 2014), speech recognition (Khorram et al. 2014), computer vision (Tappen et al. 2007; Wang et al. 2014), etc. There are other works that capture asymmetric dependencies, such as Asym-MRF model (Heesch and Petrou 2010). Since it is out of scope of this paper, for more details, please refer to Heesch and Petrou (2010) and Wang et al. (2005). Below we give a brief description of CRF and GCRF.

In a conditional random field (CRF) model, the observables $\mathbf{x}$ interact with each of the targets y_i directly and independently of one another. For a general network structure, the outputs $\mathbf{y}$ also have independent pairwise interaction functions. Thus, the CRF probability function can be represented by an equation of the form:

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

There are two sets of feature functions, association potential (A) and interaction potential (I). The larger the value of A, the more y_i is related to attributes $\mathbf{x}$. The larger the value of I, the more y_i is related to y_j. Restricting these feature functions to be quadratic differences

between a function of observables $R(\mathbf{x})$ and targets $\mathbf{y}$ produces a convex ensemble method:

$$A(\alpha, y_i, \mathbf{x}) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_k (y_i - R_{i,k}(\mathbf{x}))^2,$$

where $R_{i,k}$ represents output of unstructured predictor R_k for node i, and K is the number of unstructured predictors, and N is the number of nodes. When incorporating quadratic pairwise interaction functions among outputs $\mathbf{y}$, a general graph structure ensemble method is obtained:

$$I(\beta, y_i, y_j) = -\sum_{l=1}^{L}\sum_{i \sim j} \beta_l S_{ij}^l (y_i - y_j)^2,$$

where L represents the number of similarity functions, and S_{ij} represents similarity between nodes i and j. This similarity is symmetric, which means that $S_{ij} = S_{ji}$.

The GCRF model is a CRF model with both quadratic feature and quadratic interaction functions that can be transposed directly onto a Gaussian multivariate probability distribution:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{2\pi |\Sigma|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{y} - \mu)^T Q(\mathbf{y} - \mu)\right).$$

When setting these two conditional probability models equal to one another, we get a precision matrix (Q) defined in terms of the confidence of input predictors and the pairwise interaction structure, measured by α and β respectively. Let denote L_j as the Laplacian matrix of pairwise interaction structure matrix S_j for brevity:

$$Q = \sum_k \alpha_k I + \sum_j \beta_j L_j.$$

Representing input predictions as a matrix R, the formula for the final prediction can be concisely written as:

$$\mu = Q^{-1} R \alpha.$$

The only remaining constraint is that Q is positive semi-definite, which is a bound on convexity, but also a by-product of the multivariate Gaussian assumption. As long as the positive semi-definite constraint is satisfied, the model is convex.

In this work, the restrictions on symmetric link weights is relaxed, which alters the model in a way that is no longer capable of using a precision matrix. Additionally, convexity is no longer guaranteed. We will show convexity in special cases and demonstrate it empirically in Sect. 4.4.

3 Methodology

The proposed model DirGCRF is described in this section. Since asymmetric influence between objects violates some of the fundamental assumptions of the GCRF model (Radosavljevic et al. 2010), we re-derive the pseudo-Gaussian form and explain where the new formulation differs from the original. Below are the details of the derivation of a new matrix Q.

We start by showing that Gaussian normal form (GNF) can be equivalent to a conditional random field (CRF) model under certain conditions. The CRF is represented as:

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

The following is the exact formulation for the CRF as mentioned above. The summations are re-arranged so that the CRF can be shown to be equivalent with the GNF.

$$\sum_{i=1}^{N} A(\alpha, y_i, \mathbf{x}) + \sum_{j \sim i} I(\beta, y_i, y_j) = -\sum_{i}^{N} \sum_{k}^{K} \alpha_k (y_i - R_{i,k}(X))^2 - \sum_{i \sim j}^{L} \sum_{l}^{L} \beta_l S_{ij}^l (y_i - y_j)^2,$$

(1)

where $\sim$ means that i is connected to j. Since the summation of all link weights is unchanged, by assuming that all nodes have a link weight of zero if not otherwise specified, we can rewrite Eq. 1 in a form which requires no outside information about the structure of S.

$$\text{Eq. (1)} = -\sum_{i}^{N} \sum_{k}^{K} \alpha_k (y_i - R_{i,k}(X))^2 - \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} \sum_{l}^{L} \beta_l S_{ij}^l (y_i - y_j)^2.$$

Then the quadratic feature functions are expanded out. This allows us to group summations of independent linear and quadratic components.

$$\text{Eq. (1)} = -\sum_{i}^{N} \sum_{k}^{K} \alpha_k y_i^2 + \sum_{i}^{N} \sum_{k}^{K} 2\alpha_k y_i R_{i,k}(X) - \sum_{i}^{N} \sum_{k}^{K} \alpha_k (R_{i,k}(X))^2$$
$$+ \sum_{i}^{N} \sum_{j}^{N} \sum_{l}^{L} \beta_l S_{ij}^l y_i y_j - \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} \sum_{l}^{L} \beta_l (S_{ij}^l + S_{ij}^l) y_i^2.$$

The main difference between GCRF and DirGCRF is that the matrix S row sum, $rowsum(S)$, is not equal to column sum, $colsum(S)$. The equivalent conditions for the Conditional Random Field and Gaussian Normal Form Probability distributions are solved by segmenting the equation into quadratic, linear, and constant components. This is because the coefficients across the models differ but these variable degrees are the same across both models. This equivalence can then be solved as three different equivalences: quadratic coefficients, linear coefficients, and a constant component. In order to make this as clear as possible, the GNF is written using summations rather than matrix notation:

$$P(\mathbf{y}) = -\frac{1}{2} \mathbf{y}^T Q \mathbf{y} + \mathbf{y}^T b + c = \sum_{i}^{N} \sum_{j}^{N} Q_{ij} y_i y_j + \sum_{i}^{N} y_i b_i + c,$$

where Q, b, and c are an arbitrary matrix, vector, and scalar for GNF. Equivalent conditions for the quadratic component are:

$$\sum_{i}^{N} \sum_{j}^{N} Q_{ij} y_i y_j = \left(\sum_{k}^{K} \alpha_k\right) \sum_{i}^{N} y_i^2 + \sum_{i}^{N} \sum_{l}^{L} (rowsum(S_l) + colsum(S_l)) \cdot \beta_l y_i^2$$

(2)

$$- \sum_{i}^{N} \sum_{j}^{N} \sum_{l}^{L} \beta_l S_{ij}^l y_i y_j,$$

where summations over the coefficients y_i^2 are diagonal elements of Q and coefficients for $y_i y_j$ are off-diagonal elements of Q. Thus, the entries of Q can be segmented into a diagonal matrix, D (Eq. 3), plus a weighted adjacency matrix, A (Eq. 4):

$$Q = D + A$$

$$D_{ii} = \left(\sum_{k}^{K} \alpha_k \right) + \sum_{l}^{L} \beta_l \frac{1}{2} (rowsum(S_l) + colsum(S_l)) \tag{3}$$

$$A_{ij} = - \sum_{l}^{L} \beta_l S_{ij}^l \tag{4}$$

Let the Laplacian of an asymmetric similarity matrix be defined as:

$$L = -\frac{1}{2}(diag(rowsum(S) + colsum((S)) - 2S),$$

then the new derived "precision" matrix (Q) can be concisely defined as:

$$Q = \left(\sum_{k}^{K} \alpha_k \right) I + \sum_{l}^{L} \beta_l L_l.$$

When mapping the linear coefficients from CRF to GNF, one finds that the asymmetric network does not change the mappings as found in the original GCRF:

$$b_i = \sum_{k}^{K} \alpha_K R_{i,k}(X),$$

or, more concisely, $b = R\alpha$. Since the constant does not affect the marginalized likelihood, it can be omitted.

The Multinormal Likelihood Function ($P_2(\epsilon)$, Eq. 5) is equivalent to a Gaussian Normal Form ($P_1(\mathbf{y})$, Eq. 6) given certain conditions.

$$P_2(\epsilon) = \frac{1}{Z} exp \left(-\frac{1}{2} \epsilon^T \Sigma^{-1} \epsilon \right) \tag{5}$$

$$P_1(\mathbf{y}) = \frac{1}{Z} exp \left(-\mathbf{y}^T \Sigma^{-1} \mathbf{y} + b^T \mathbf{y} + c \right) \tag{6}$$

The equivalent conditions are:

$$c = -\mu^T \Sigma^{-1} \mu, \quad \mu = \Sigma b,$$

where μ is the optimal prediction given the covariance matrix (Σ) and the linear component of the Gaussian Normal Form (b).

DirGCRF uses the above formulas and gradient ascent in order to find the optimal values for parameters α_i and β_i. The only remaining step is to find the first order derivatives of log-likelihood function and updates of α_i and β_i in gradient ascent. Equation for the log-likelihood function (l) is:

$$l = -\frac{1}{2} \epsilon^T Q \epsilon - Z.$$

The partial derivatives of the precision matrix (Q) with respect to α_i and β_i can be found as:

$$\frac{\partial Q}{\partial \alpha_i} = I \quad \frac{\partial Q}{\partial \beta_i} = L_i.$$

Recall from the mapping of Gaussian Normal Form to Multinormal Likelihood Function, μ can be presented as:

$$\mu = Q^{-1}b. \tag{7}$$

Since μ is in the log-likelihood function via $\epsilon = \mathbf{y} - \mu$, its partial derivatives with respect to α_i and β_i are:

$$\frac{\partial \mu}{\partial \alpha_i} = -Q^{-1}IQ^{-1}b + Q^{-1}R_i \quad \frac{\partial \mu}{\partial \beta_i} = -Q^{-1}L_iQ^{-1}b.$$

Fully elucidated form of the log-likelihood function is:

$$l =^+ -\frac{1}{2}(\mathbf{y}^T Q\mathbf{y} - \mathbf{y}^T Q\mu - \mu^T Q\mathbf{y} + \mu^T Q\mu) - log|Q^{-1}|^2.$$

Minor steps in the remaining derivation of the partial derivatives with respect to the parameters α_i and β_i are omitted. The final result is below, and is not hard to verify:

$$\frac{\partial l}{\partial \alpha_i} = -\frac{1}{2}\left[(\mathbf{y} - \mu)^T(\mathbf{y} - \mu) + (R_i - \mu)^T(I + Q^{-1}Q)(\mu - \mathbf{y})\right] + \frac{1}{2}Tr(Q^{-1})$$

$$\frac{\partial l}{\partial \beta_i} = -\frac{1}{2}\left[\mathbf{y}^T L_i\mathbf{y} - (-Q^{-1}L_i\mu)^T Q\mathbf{y} - \mu^T L_i\mathbf{y} + (-Q^{-1}L_i\mu)^T Q\mu\right] + \frac{1}{2}Tr(L_iQ^{-1})$$

4 Experiments

4.1 Datasets and experimental setup

4.1.1 Synthetic datasets

The purpose of experiments on synthetic data was to investigate the proposed model under controlled conditions on different types of graphs. Below are the descriptions of each type of graph and their node attributes and asymmetric similarities.

- *Fully connected directed graph*: Each pair of distinct nodes is connected by a pair of edges (one in each direction) with different weights.
- *Directed graph with edge probability p*: Directed graphs with different density. For each pair of distinct nodes, a random number between 0 and 1 is generated. If the number exceeds p, then the selected node pair will be connected with an edge.
- *Directed graph without direct loop*: Each pair of distinct nodes is connected by a single edge, which direction is chosen randomly. For example, if there is an edge from node A to node B, there could not be an edge from node B to node A.
- *Directed acyclic graph*: A graph with no cycles. For example, there is no path that starts from a node A and follows a consistently-directed sequence of edges that loops back to node A.
- *Chain*: All nodes are connected in a single sequence, from one node to another.
- *Binary tree*: A graph with a tree structure in which each node could have at most two children.

All these graph types are unlabeled and unweighted. Therefore, we randomly generated edge weights S and unstructured values R. The generated S and R were used to calculate the actual value of response variable $\mathbf{y}$ for each node, in accordance to the Eq. 7, with some added random noise. For calculation of $\mathbf{y}$ we needed to choose values for α and β parameters. For

GCRF based models, when there is only one α and only one β, only the ratio between values of α and β parameters matters, not their actual values. A greater value of α means that the model is putting more emphasis on values that are provided by the unstructured predictor (R), while a greater value of β means that the model is putting more emphasis on structure (S). We chose three different combinations in order to compare the performance of the model: (1) larger value of α parameter ($\alpha = 5$, $\beta = 1$); (2) larger value of β parameter ($\alpha = 1$, $\beta = 5$); (3) same value of both parameters ($\alpha = 1$, $\beta = 1$).

In all experiments on generated synthetic datasets one graph is used for training and five graphs for testing. For evaluating accuracy, experiments were conducted on graphs with 200 nodes. For testing run time, experiments were conducted on fully connected directed graphs with 500, 1K, 5K, 10K and 15K nodes.

4.1.2 Real-world datasets

We also evaluated our model on four real-world datasets: *Delinquency* (Snijders et al. 2010), *Teenagers* (Michell and Amos 1997), *Glasgow* (Bush et al. 1997) and *Geostep* (Scepanovic et al. 2015). The first three datasets contain data about habits of students (e.g. tobacco and alcohol consumption) and friendship networks at different observation time points. *Geostep* dataset contains data about treasure hunt games. Node attributes, edge weights, and response variables are extracted from data. All values were normalized to fit in range from 0 to 1. The experimental procedure and the obtained results are described in more detail in Sects. 4.3.1–4.3.4.

4.1.3 Baselines

The accuracy performance of DirGCRF was compared with the standard GCRF (Polychronopoulou and Obradovic 2014), and four nonlinear and linear unstructured baselines briefly described in this section: neural networks (NN) (Haykin 2009), linear regression (LR) or multivariate linear regression (MLR) (Weisberg 2005), average and last methods.

- *GCRF*: In order to apply the standard GCRF to the directed graphs, S matrix was converted from asymmetric to symmetric. In a symmetric matrix each pair of distinct nodes is connected by a single undirected edge, where the weight was calculated as an average of weights in the corresponding asymmetric matrix. The Neural Network unstructured predictor was used for both, DirGCRF and standard GCRF.
- *NN*: Neurons in feed-forward artificial neural networks are grouped in three layers: input, output and hidden layer. The number of neurons in the input layer was same as the number of features in the considered dataset. The number of neurons in the output layer was 1 for all datasets. The number of neurons in the hidden layer was selected based on the accuracy performance on the training data.
- *LR or MLR*: Linear regression or multivariate linear regression is used depending on the number of features in the considered dataset. Coefficients of predictors were trained on the features of all nodes on the training data, and then applied on the features on the test data to form the prediction.
- *Last*: In the real-world datasets, the graphs have evolved. Therefore, we consider one simple method, Last, which assigns values to the response variables using the same values as in the previous time point.
- *Average*: Another simple technique that calculates prediction of **y** value at each time stamp as the average of the **y** values in all previous time stamps.

Table 2 Average (± standard deviation) R^2 of DirGCRF and GCRF on different types of asymmetric structures with parameters values $\alpha = 5$ and $\beta = 1$

Graph type	DirGCRF	GCRF
Directed graphs	0.9176 (±0.00625)	0.5893 (±0.02680)
Directed graphs with $p = 0.5$	0.9799 (±0.00332)	0.6582 (±0.06063)
Directed graphs with $p = 0.2$	0.9951 (±0.00074)	0.8880 (±0.00846)
Directed graphs without direct loop	0.9865 (±0.00084)	0.4608 (±0.03497)
Acyclic graphs	0.9881 (±0.00019)	0.2580 (±0.03584)
Chains	0.9995 (±0.00001)	0.9987 (±0.00009)
Binary trees	0.9995 (±0.00004)	0.9988 (±0.00008)

To calculate the regression accuracy of all methods, we used R^2 coefficient of determination that measures how closely the output of the model matches the actual value of the data. A score of 1 indicates a perfect match, while a score of 0 indicates that the model simply predicts the output variable mean. R^2 of some poor predictors can even be worse than average and are characterized with negative coefficient of determination.

$$R^2 = 1 - \sum_i \frac{(y_i - \hat{y}_i)^2}{(y_i - y_{average})^2},$$

where $\hat{y}_i$ is the predicted value, y_i is the true value, and $y_{average}$ is the average of **y** values.

For DirGCRF and GCRF gradient ascent was used to find the optimal values for parameters α and β. Initial values of parameters were $\alpha = 1$ and $\beta = 1$, in each experiment. Learning rate was set to 0.01.

All methods are implemented in Java, and experiments were run on Windows with 32GB memory (28GB for JVM) and 3.4GHz CPU. All codes are publicly available.[1]

4.2 Performance on synthetic datasets

4.2.1 Effectiveness of DirGCRF

We first tested the accuracy of the DirGCRF model, and compared the performance against the standard GCRF model. Experiments were conducted on all synthetic datasets described in the Sect. 4.1.1. The outputs of unstructured predictor (R) and similarity matrix (S) are randomly generated. For each type of graph, one graph is used for training the model, and five graphs for testing. All graphs contain 200 nodes. α was set as 5 and β was set as 1 in this experiment. Average R^2 and standard deviations of both models are presented in Table 2.

The results show that the DirGCRF produces higher accuracy than the standard GCRF on all synthetic directed graphs. On the fully connected directed graph, DirGCRF has 0.33 larger R^2 value than GCRF. With decreasing probability of edge existence, the graphs become sparser. Thus, the difference between DirGCRF and GCRF in accuracy becomes smaller. For graphs that do not have a direct loop or cycle, DirGCRF performs much better than GCRF, 0.53 and 0.73 larger R^2 value, respectively, which indicates the superiority of DirGCRF on directed graphs. Also, we noticed that in all experiments DirGCRF has very low standard deviation (from 0.007 to 0.00004) of R^2 performance.
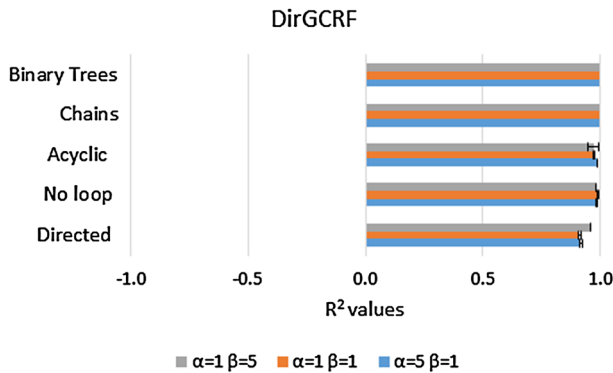
---

[1] https://github.com/vujicictijana/NewApp.

Fig. 3 Average R^2 of DirGCRF on different types of asymmetric structures with different α and β values

The only exceptions are the results on the chains and binary trees where both algorithms have similar accuracy. This is expected since these structures are very sparse where every node has a maximum of two nodes that directly affect its output.

4.2.2 Accuracy with respect to different α and β values

The purpose of this experiment is to find out how values of α and β parameters in data generation process affect the accuracy of DirGCRF and GCRF models. In this experiment, we tested three different setups to generate synthetic graphs. In the first one, α has higher value, $\alpha = 5$ and $\beta = 1$, which means that more emphasis is put on the unstructured predictor value and less on the structure. In the second one, both parameters have the same value: $\alpha = 1$ and $\beta = 1$. In the third one, the β parameter has higher value: $\alpha = 1$ and $\beta = 5$, that is, more emphasis is put on the structure.

From Fig. 3, we can notice that the variations in R^2 value for DirGCRF across three different settings in all types of graphs are minor. However, there is a big difference in R^2 value for GCRF, especially on directed graphs and on directed graphs without loop or cycle. For example, in directed graphs a larger value of β parameter caused a slight increase in accuracy of DirGCRF (from 0.92 to 0.96), but a large decrease in accuracy of GCRF (from 0.59 to -0.1) (Fig. 4). This indicates that the standard GCRF could not utilize asymmetric structure to provide good results, especially for datasets in which structure is more useful.

4.2.3 Run time

Time complexity of DirGCRF is same as time complexity of the standard GCRF (Radosavljevic et al. 2014). If the number of nodes in the training set is N and the learning process lasts T iterations, computation results in $O(TN^3)$ time to train the model. The main cost of computation is matrix inversion.

The following speed tests of the DirGCRF model were conducted on synthetically generated fully connected directed graphs with varying numbers of nodes: 500, 1K, 5K, 10K and 15K nodes. The time consumption is presented after 50 iterations and the results are shown in Table 3. Model takes more time due to Java's object-oriented nature, which requires more memory and more time to handle large matrix computations.
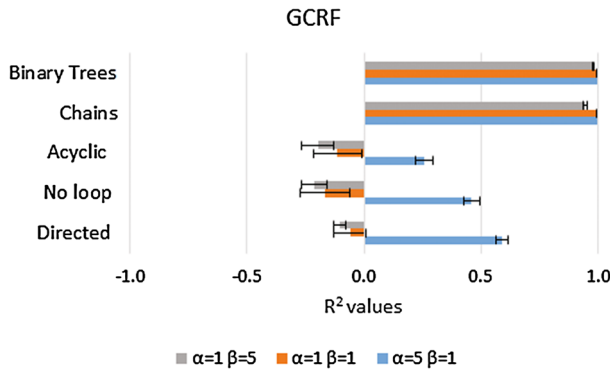
Fig. 4 Average R^2 of GCRF on different types of asymmetric structures with different α and β values

Table 3 Run time of DirGCRF for different number of nodes

No. of nodes	Speed
500	8 s
1000	48 s
5000	2 h
10,000	17 h
15,000	2.2 days

4.3 Performance on real-world datasets

We have conducted experiments on four real-world datasets and compared the performance of DirGCRF against all baselines. We chose NN as the unstructured predictor for DirGCRF and GCRF, as it produces better results than LR\MLR. Details about each dataset are provided in Table 4 and they will be described in the following sections.

4.3.1 Delinquency dataset

The *Delinquency* (Snijders et al. 2010) dataset[2] consists of four temporal observations of 26 students (aged between 11 and 13) in a Dutch school class between September 2003 and June 2004. For each observation, a friendship matrix is provided, as well as delinquency and alcohol scores. Both the delinquency and the alcohol scores are ranked from 1 to 5. The friendship networks were formed by allowing the students to name up to 12 best friends. The total number of edges in these matrices was between 88 and 133 (density from 13 to 20%). On average, 49% of students' friendships were one-directional. The similarity (S_{ij}^t) from the student i to the student j at the specific time point t was calculated based on the friendship existence in all previous time points and the current one, that is,

$$S_{ij}^t = \frac{\sum_{k=1}^t S_{ij}^k}{t}. \tag{8}$$

---

[2] https://www.stats.ox.ac.uk/~snijders/siena/tutorial2010_data.htm.

Table 4 Real-world dataset

| Dataset (nodes)	Time points	x	y	S
Delinquency (26 students)	4	1. Previous delinquency 2. Alcohol consumption	Delinquency level	Friendship network
Teenagers (50 teenagers)	4	1. Previous alcohol consumption	Alcohol consumption	Friendship network
Glasgow (129 students)	3	1. Alcohol 2. Cannabis consumption 3. Romantic relationship 4. Pocket money per month	Tobacco consumption	Friendship network
Geostep (50 games)	N/A	1. No. of clues in social category 2. No. of clues in business cat. 3. No. of clues in travel cat. 4. No. of clues in irrelevant cat. 5. Privacy scope 6. Duration	Relevance for touristic purposes	Games similarity
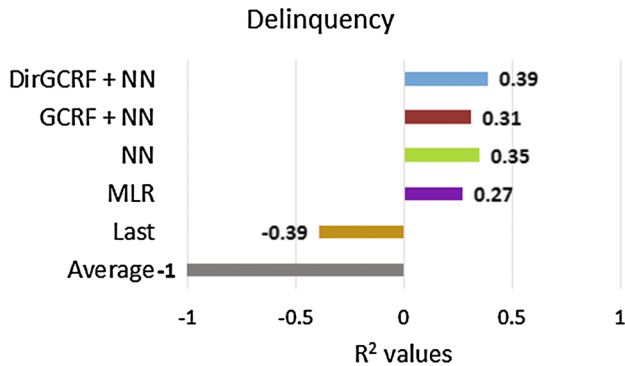
Fig. 5 Average R^2 for *Delinquency* dataset

The goal was to predict the delinquency level for each student. Training was performed on the observation points 2 and 3. Alcohol consumption and previous delinquency level were used as attribute values **x**. The models were tested on the observation point 4.

From the results presented at Fig. 5 we can see that the DirGCRF model outperforms all other competing models. The DirGCRF model has 8% larger accuracy than the standard GCRF model, and 4% larger accuracy than the Neural Network. Neural Network was the second best model. Multivariate Linear Regression was less accurate, but better than the Last and Average methods which produced negative R^2s. The GCRF model produces a lower R^2 than NN, which means that using converted symmetric friendship network was not helpful to improve the regression.
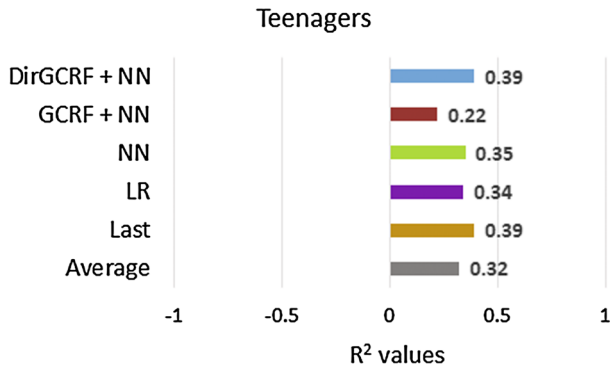
Teenagers

Fig. 6 Average R^2 for *Teenagers* dataset

4.3.2 Teenagers dataset

The *Teenagers* (Michell and Amos 1997) dataset[3] consists of three temporal observations of 50 teenagers (aged 13) in a school in the West of Scotland over a 3-year period (1995–1997). Just like in the *Delinquency* dataset the teenagers were asked to identify up to 12 best friends. The total number of edges in these observations was between 113 and 122 (density around 5%). On average 60% of teenagers' friendships were one-directional. The same approach (Eq. 8) as in the *Delinquency* dataset was used to calculate similarity matrix. Besides friendship networks, the dataset contains information about teenager's alcohol consumption (ranging from 1 to 5). The goal in this dataset was to predict alcohol consumption at the observation time point 3, based on two previous observations.

Figure 6 shows that the DirGCRF model has 17% larger accuracy than the standard GCRF model, and 4% larger accuracy than the Neural Network. Neural Network and Linear Regression have similar accuracy on this dataset, 0.35 and 0.34, respectively. The simple Last method has higher R^2 than both unstructured predictors and the same R^2 as DirGCRF, and the Average method also produced a high accuracy. This is due to the fact that in this application there are no additional features—only previous value of **y** was used to make predictions.

4.3.3 Glasgow dataset

The *Glasgow* (Bush et al. 1997) dataset[4] consists of three temporal observations of 160 students at a secondary school in Glasgow. Students were followed over a 2-year period starting in February 1995, when the students were aged 13, and ending in January 1997. We used data for 129 students who were present at all three measurement points. The friendship networks were formed by allowing the students to name up to six friends and to mark them from 0 to 2 as follows: 1—best friend, 2—just a friend, 0—no friend. The total number of edges in these matrices was around 362 (density 2%). On average 72% of students' friendships were one-directional. In order to predict tobacco consumption, the following features were used as attribute values **x**:

– Alcohol consumption (from 1 to 5).
– Cannabis consumption (from 1 to 4).

---

[3] https://www.stats.ox.ac.uk/~snijders/siena/s50_data.htm.

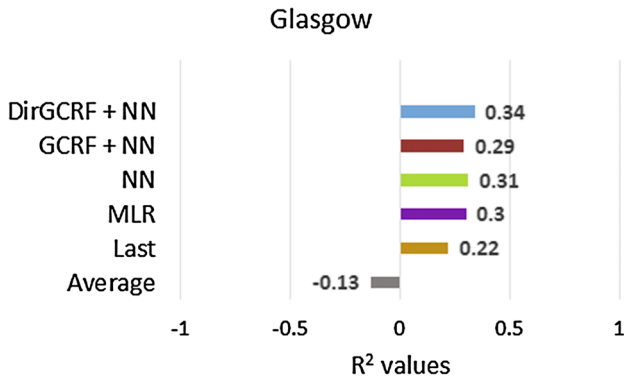[4] https://www.stats.ox.ac.uk/~snijders/siena/Glasgow_data.zip.

Fig. 7 Average R^2 for *Glasgow* dataset

- Romantic relationship (indicates whether the student had a romantic relation at the specific time point).
- Amount of pocket money per month.

Graphs from the first two observation points were used for training, and the graph from third one was used for testing. From Fig. 7, we can see that DirGCRF model outperforms all other competing models, but almost all other baselines (except the simple Last and Average methods) have produced close R^2 values. There is a noticeable difference between models that are using asymmetric and symmetric structure, that is, DirGCRF has 5% higher accuracy than the standard GCRF.

4.3.4 Geostep dataset

The *Geostep* (Scepanovic et al. 2015) dataset[5] consists of data about 50 treasure hunt games. Each game can have maximum 10 clues and each clue belongs to one of 4 categories. The goal is to predict probability that the game can be used for touristic purposes. Features that were used as **x** values are: the number of clues in each category (business, social, travel, and irrelevant), game privacy scope, and game duration. We randomly chose 25 games for training and the rest of them were used for testing. A similarity matrix was created based on the games' features. The similarity of game i to the game j (S_{ij}) is defined as the sum of the common number of clues in each category k in both games divided by total number of clues in the game i.

$$S_{ij} = \frac{\sum_{k=1}^{4} min(C_i^k, C_j^k)}{\sum_{k=1}^{4} C_i^k}, \tag{9}$$

where C_i^k is the number of clues in the category k for the game i.

From the results presented at Fig. 8, we can see that the use of this asymmetric structure significantly improved the result of Neural Network. On the other hand, the difference in accuracy between GCRF and NN is the highest on this dataset (accuracy of GCRF is 13% lower), which indicates that converting asymmetric similarity matrix to symmetric had negative impact on regression performance.
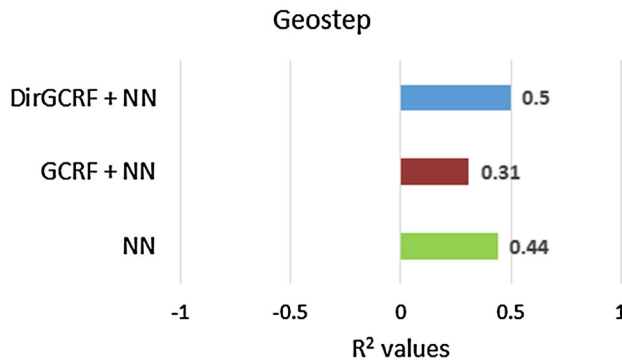
---

[5] http://www.geostep.me/.

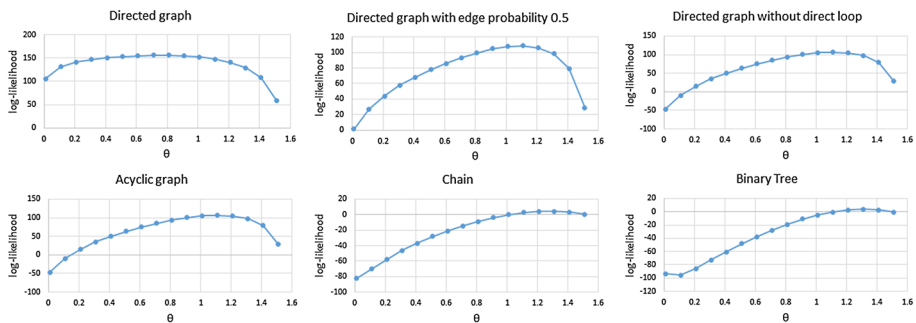**Fig. 8** Average R^2 for *Geostep* dataset

Fig. 9 Experimental demonstration of model convexity for synthetic datasets

It can be noticed that, from Figs. 5, 6, 7 and 8, the accuracies of DirGCRF, GCRF and NN are consistent for all four real-world datasets. In each dataset DirGCRF has the highest accuracy, while GCRF has lower accuracy than Neural Network.

4.4 Convexity

The experimental results presented in the previous sections were obtained for a specific value of hyper-parameter θ. The learning task was to choose the parameters α and β to maximize the conditional log-likelihood. The additional experiments were conducted in order to empirically demonstrate model convexity for all used synthetic and real-life datasets. Results are presented at Figs. 9 and 10. In this experiment we incrementally increase θ from 0 to $\frac{\pi}{2}$. For each θ we calculate α and β as $\alpha = sin(\theta)$ and $\beta = cos(\theta)$. Then we calculate log-likelihood with respect to these parameters and plot the values. These figures show that log-likelihood is a convex function of parameters α and β and that its optimization leads to globally optimal solution. The only exception is the Binary Tree dataset in Fig. 9, where the kink on the left hand side of the curve does not show convexity for this network structure. However, the optimization procedure still finds the global maximum even when starting close to the local maximum. The fact that we can plot the entire likelihood function guarantees that we are finding the global maximum.
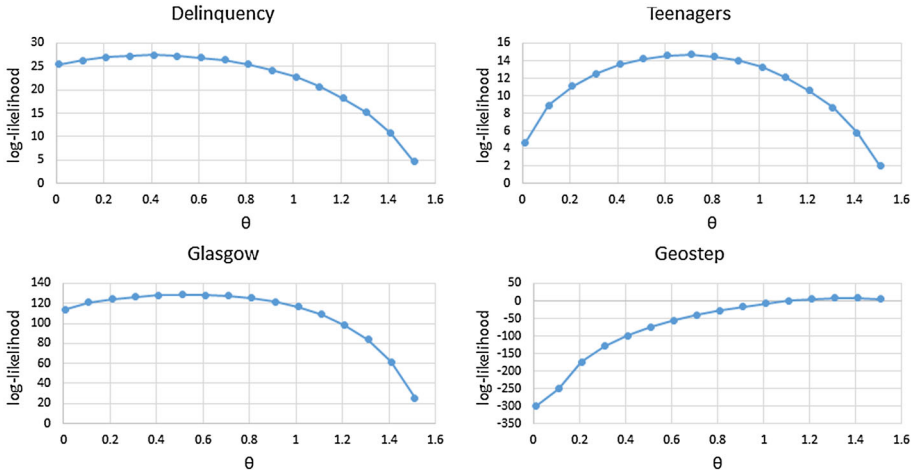
Fig. 10 Experimental demonstration of model convexity for real-world datasets

5 Conclusions

In this paper, we introduced a problem of using structured regression for predicting output variables that are asymmetrically linked. A new model, called directed Gaussian conditional random fields (DirGCRF), is proposed. This model extends the GCRF model by considering asymmetric similarities among objects. To evaluate the proposed model, we tested it on both synthetic and real-world datasets. A significant accuracy improvement is achieved compared to standard GCRF: from 5 to 19% for real-world datasets and in average 30% for synthetic datasets. If the data has more emphasis on structure than on values that are provided by the unstructured predictor, then the DirGCRF model even doubles the accuracy of GCRF for some types of directed graphs. Also, the experimental results confirmed that the simple approach of converting an asymmetric similarity matrix to a symmetric one for GCRF has negative impact on regression performance. Since this model is implemented in Java, which takes time to handle large matrix computations, our plan for future work is to implement the model in a procedural or functional programming language in order to speed it up and make it more efficient for large datasets. We also plan to apply the DirGCRF model to other real-world applications and to demonstrate that our model can use multiple unstructured predictors (multiple α parameters) and multiple graphs (multiple β parameters).

References

Altken, A. (1935). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, *55*, 42–48.

Beguerisse-Díaz, M., Garduno-Hernández, G., Vangelov, B., Yaliraki, S. N., & Barahona, M. (2014). Interest communities and flow roles in directed networks: The twitter network of the UK riots. *Journal of the Royal Society Interface*, *11*(101), 20140940.

Bush, H., West, P., & Michell, L. (1997). The role of friendship groups in the uptake and maintenance of smoking amongst pre-adolescent and adolescent children: Distribution of frequencies. Working Paper No. 62. MRC Medical Sociology Unit Glasgow.

Djuric, N., Radosavljevic, V., Obradovic, Z., & Vucetic, S. (2015). Gaussian conditional random fields for aggregation of operational aerosol retrievals. *IEEE Geoscience and Remote Sensing Letters*, *12*, 761–765.

Glass, J., Ghalwash, M., Vukicevic, M., & Obradovic, Z. (2015). Extending the modelling capacity of Gaussian conditional random fields while learning faster. In *Proceedings 30th AAAI conference on artificial intelligence (AAAI-16)*, pp. 1596–1602.

Gligorijevic, D., Stojanovic, J., & Obradovic, Z. (2015). Improving confidence while predicting trends in temporal disease networks. In *4th workshop on data mining for medicine and healthcare, SIAM international conference on data mining (SDM)*.

Guo, H. (2013). Modeling short-term energy load with continuous conditional random fields. In *European conference on machine learning and principles and practice of knowledge discovery in databases (ECML/PKDD)*, pp. 433–448.

Hallac, D., Leskovec, J., & Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 387–396). ACM.

Haykin, S. S. (2009). *Neural networks and learning machines* (Vol. 3). Upper Saddle River: Pearson.

Heesch, D., & Petrou, M. (2010). Markov random fields with asymmetric interactions for modelling spatial context in structured scene labelling. *Journal of Signal Processing Systems*, *61*(1), 95–103.

Khorram, S., Bahmaninezhad, F., & Sameti, H. (2014). Speech synthesis based on Gaussian conditional random fields. In *Artificial intelligence and signal processing*, pp. 183–193.

Liu, C., Adelson, E. H., & Freeman, W. T. (2007) Learning Gaussian conditional random fields for low-level vision. In *Proceedings of CVPR* (p. 7). Citeseer.

Michell, L., & Amos, A. (1997). Girls, pecking order and smoking. *Social Science & Medicine*, *44*(12), 1861–1869.

Polychronopoulou, A., & Obradovic, Z. (2014). Hospital pricing estimation by gaussian conditional random fields based regression on graphs. In *2014 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 564–567). IEEE.

Radosavljevic, V., Vucetic, S., & Obradovic, Z. (2010). Continuous conditional random fields for regression in remote sensing. In *ECAI*, pp. 809–814.

Radosavljevic, V., Vucetic, S., & Obradovic, Z. (2014). Neural Gaussian conditional random fields. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 614–629). Springer.

Scepanovic, S., Vujicic, T., Matijevic, T., & Radunovic, P. (2015). Game based mobile learning—Application development and evaluation. In *Proceedings of an 6th conference on e-learning*, pp. 142–147.

Slivka, J., Nikolić, M., Ristovski, K., Radosavljević, V., & Obradović, Z. (2014). Distributed Gaussian conditional random fields based regression for large evolving graphs. In *Proceedings of 14th SIAM international conference on data mining, workshop on mining networks and graphs*.

Snijders, T. A., Van de Bunt, G. G., & Steglich, C. E. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, *32*(1), 44–60.

Stojkovic, I., Jelisavcic, V., Milutinovic, V., & Obradovic, Z. (2016). Distance based modeling of interactions in structured regression, pp. 2032–2038.

Tappen, M. F., Liu, C., Adelson, E. H., & Freeman, W. T. (2007). Learning Gaussian conditional random fields for low-level vision. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1–8.

Uversky, A., Ramljak, D., Radosavljević, V., Ristovski, K., & Obradović, Z. (2013). Which links should i use?: A variogram-based selection of relationship measures for prediction of node attributes in temporal multigraphs. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 676–683). ACM.

Wang, S., Wang, S., Greiner, R., Schuurmans, D., & Cheng, L. (2005). Exploiting syntactic, semantic and lexical regularities in language modeling via directed Markov random fields. In *Proceedings of the 22nd international conference on machine learning* (pp. 948–955). ACM.

Wang, S., Zhang, L., Urtasun, R. (2014). Transductive Gaussian processes for image denoising. In *2014 IEEE international conference on computational photography (ICCP)* (pp. 1–8). IEEE.

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). Hoboken: Wiley.

Wytock, M., & Kolter, J. Z. (2013). Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. *ICML*, *3*, 1265–1273.