Decomposition Based Reparametrization for Efficient Estimation of Sparse Gaussian Conditional Random Fields

Ivan Stojkovic[*][1] **Vladisav Jelisavcic**[*][2] **Jelena Gligorijevic**[1] **Djordje Gligorijevic**[1] **Zoran Obradovic**[3]

Abstract

Simultaneously estimating multi-output regression model, while recovering dependency structure among variables, from high-dimensional observations is an interesting and useful exercise in contemporary statistical learning applications. A prominent approach is to fit a Sparse Gaussian Conditional Random Field by optimizing regularized maximum likelihood objective, where the sparsity is induced by imposing L_1 norm on the entries of a precision and transformation matrix. We studied how reparametrization of the original problem may lead to more efficient estimation procedures. Particularly, instead of representing problem through precision matrix, we used its Cholesky factor, which attractive properties allowed inexpensive coordinate descent based optimization algorithm, that is highly parallelizable.

1. Introduction

Conditional Random Fields (CRF) are a class of probabilistic graphical models, that model the probability distribution of a number of random variables conditioned on a set of observations (Sutton & McCallum, 2012). CRFs are commonly used for approaching multi-output/structured-prediction tasks like named entity recognition in Natural Language Processing (Tran et al., 2017), gene finding in Bioinformatics (Chang et al., 2015), or image segmentation in Computer Vision (Orlando et al., 2017), to name a few.

Although very flexible and descriptive framework, CRFs are also computationally expensive and generally intractable, due to fact that normalization (a.k.a. partition) function needs to be integrated over the input space, for which there is no closed form solution, except in few special cases. Therefore, additional modeling assumptions are typically introduced to make estimation and inference tractable.

A popular and useful one is to assume a particular distribution, like Multivariate Normal Distribution, which leads to convenient mathematical properties and efficient estimation and inference algorithms for the model (Krähenbühl & Koltun, 2011). Gaussian assumption makes estimation of CRF models from data tractable, and is often sound for common regression tasks. However, contemporary applications are comprised of increasingly higher number of variables (tens of thousands of biomarkers measured, millions of pixels in the images, etc.) and when cardinality of target variables set is large, model parameters' space grows polinomially, thus increasing storage requirements, computational time, and tendency of model to overfit. Additional assumptions make sense to lead to more parsimonious models and more efficient computation, for example that some (or majority) of variable pairs are conditionally independent given all others. That leads to sparsity in problem parametrization, and hence the name Sparse Gaussian Conditional Random Fields (SGCRF) (Wytock & Kolter, 2013).

State of the art SGCRF methods are fitting the inverse covariance (i.e. precision) matrix and transformation matrix against the data, in a L_1 regularized maximum likelihood objective. Such formulation contains logarithm of determinant of the precision matrix which is expensive to compute. We propose reparametrization of the objective, in terms of Cholesky factor of the precision matrix. Such parametrized problem greatly relieves the burden of calculating the logdet term, while learning the sparse Cholesky factor still produces sparse precision (in most circumstances). Moreover, columns of Cholesky factor can be calculated independently of each other, which allows trivially parallel computation.

2. Background

2.1. Problem formulation

The target vector $y \in \mathbb{R}^{p\times 1}$ (a p-dimensional column vector), that is conditioned (dependent) on vector of observed values $x \in \mathbb{R}^{q\times 1}$ (a q-dimensional column vector of random variables), is modeled as a particular type of log-linear

[*]Equal contribution [1]Yahoo! Research, Sunnyvale, California, USA [2]Mathematical Institute, National Academy of Sciences and Arts, Belgrade, Serbia [3]Computer and Information Sciences, Temple University, Philadelphia, Pennsylvania, USA. Correspondence to: Ivan Stojkovic <ivan.stojkovic@temple.edu>.

3rd Tractable Probabilistic Modeling Workshop, 36th International Conference on Machine Learning, Long Beach, California, 2019. Copyright 2019 by the author(s).

model, i.e. Gaussian CRF. Conditional probability density function of y in Gaussian CRF has a canonical form of a multivariate normal distribution:

$$\begin{gathered} d(y|x) = (2\pi)^{-p/2}\ \det(\Sigma)^{-1/2} \\ \exp\left(-\frac{1}{2}(y-\mu(x))^{\mathrm{T}}\Sigma^{-1}(y-\mu(x))\right) \end{gathered} \tag{1}$$

where μ is the mean function mapping: $\mathbb{R}^{q\times 1} \rightarrow \mathbb{R}^{p\times 1}$ and Σ is nondegenerate covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. We adopt μ as a parameterized function (Radosavljevic et al., 2010), linear in input space x, of a form of $\mu(x) = \Sigma\Theta^T x$.

Equation 1 models both mapping between the input and output spaces Θ ($\Theta \in \mathbb{R}^{q\times p}$), as well as dependencies among the random variables in the output space Σ. Task is to estimate parameter matrices Σ and Θ, by fitting the conditional probability function using the observed data D, which is in form of n IID pairs of input-output (or explanatory-dependent) variables $D = \{\{x_1, y_1\}, \{x_2, y_2\}, ..., \{x_n, y_n\}\}$. Joint probability is obtained as the product of individual conditional density functions (1) for each of the observations:

$$\begin{gathered} p(y|x;D) = (2\pi)^{-np/2}\ \det(\Sigma^{-1})^{n/2} \\ \prod_{i=1}^{n}\exp\left(-\frac{1}{2}(y_i^T\Sigma^{-1}y_i - 2x_i^T\Theta y_i + x_i^T\Theta\Sigma\Theta^T x_i)\right) \end{gathered} \tag{2}$$

Transformation with -log (negative log likelihood) is performed, as well as vectorization of the expression (stack n observations in matrix $Y \in \mathbb{R}^{p\times n}$, and matrix $X \in \mathbb{R}^{q\times n}$). Logdet is written as a trace of a scalar, and logarithm of the constant term is dropped, as it doesn't affect the extremum:

$$\begin{gathered} p(y|x;D) = (2\pi)^{-np/2}\ \det(\Sigma^{-1})^{n/2} \\ \exp\left(-\frac{1}{2}tr(Y^T\Sigma^{-1}Y - 2X^T\Theta Y + X^T\Theta\Sigma\Theta^T X)\right) \end{gathered} \tag{3}$$

Subsequently, we seek parameters' values that minimize the negative log-likelihood (known result that the trace is invariant to cyclic permutations is used):

$$\begin{gathered} l(\Sigma,\Theta) = -\frac{n}{2}log(det(\Sigma^{-1})) + \frac{1}{2}tr\left(\Sigma^{-1}YY^T\right) \\ -tr(\Theta YX^T) + \frac{1}{2}tr(\Theta\Sigma\Theta^T XX^T) \end{gathered} \tag{4}$$

To make the equations more concise, following substitutions are introduced: $\frac{1}{n}YY^T$=$S^{yy} \in \mathbb{R}^{p\times p}$, $\frac{1}{n}YX^T$=$S^{yx} \in \mathbb{R}^{p\times q}$ and $\frac{1}{n}XX^T$=$S^{xx} \in \mathbb{R}^{q\times q}$. Finally, the SGCRF model objective is obtained by imposing an L_1 norm on matrices Σ^{-1} and Θ (λ_Σ and λ_Θ are the respective weights):

$$\begin{gathered} l(\Sigma,\Theta) = -\frac{1}{2}log|\Sigma^{-1}| + \frac{1}{2}tr\left(\Sigma^{-1}S^{yy}\right) - tr(\Theta S^{yx}) \\ +\frac{1}{2}tr(\Theta\Sigma\Theta^T S^{xx}) + \lambda_\Sigma\|\Sigma^{-1}\|_1 + \lambda_\Theta\|\Theta\|_1 \end{gathered} \tag{5}$$

2.2. Related work

Objective in eq. (5) is originally formulated in (Sohn & Kim, 2012), dubbed Conditional Gaussian Graphical Model, and solved using Orthant-Wise Quasi-Newton algorithm (Andrew & Gao, 2007). Later work by (Wytock & Kolter, 2013) referred to the same problem formulation more specifically as Gaussian Conditional Random Field, recognizing it as a discriminative extension of well studied sparse inverse covariance problem, and accordingly adopted second order active set approach from (Hsieh et al., 2011) to solve it. Similarly, block-wise coordinate descent algorithm for inverse covariance learning (Hsieh et al., 2013) was subsequently utilized for scaling the approach to large SGCRF problems in (McCarter & Kim, 2016).

Beside improving learning efficiency and scalabillity of algorithms, there are efforts to improve modeling capabilities of SGCRF models. Neural Conditional Random Fields were proposed to allow nonlinear mappings from observations (Radosavljevic et al., 2014), and marginalization based approach was used for working on partially observed data (Stojanovic et al., 2015). Copula approach was used to enable distributions other than Gaussian (Kim, 2016), and a method robust to outliers was proposed in (Hirose et al., 2017). In addition, there are efforts to merge SGCRF with recurrent neural networks for use on time series data (Wang et al., 2018), as well as with variational autoencoders for modeling noise in images (Dorta et al., 2018).

In the reminder of paper, our focus is on improving the computational efficiency and scalability of SGCRF estimation algorithm, through convenient objective reparametrization.

3. Method

First, we reparametrize eq. (4) in terms of Cholesky factors L, where $\Sigma^{-1} = LL^T$ (scaled to lose the positive factor n):

$$\begin{gathered} l(L,\Theta) = -\frac{1}{2}log(det(LL^T)) + \frac{1}{2}tr\left(LL^T S^{yy}\right) \\ -tr(\Theta S^{yx}) + \frac{1}{2}tr(\Theta {L^{-1}}^T L^{-1}\Theta^T S^{xx}) \end{gathered} \tag{6}$$

Cholesky reparametrization was previously proposed for estimating sparse inverse covariance matrix in (Stojkovic et al., 2017), however applying this trick for the SGCRF model is not trivial. It is computationally very inconvenient to have both L and its inversion L^{-1} in the expression of objective (6), as cost of matrix inversion is roughly cubic in size of the problem. To further alleviate such inconvenience, we are resorting to introduction of an additional reparametrization $L^{-1}\Theta^T = W^T$, where $W \in \mathbb{R}^{q\times p}$ is new resulting parameter matrix which we aim to estimate. The new-reparametrized objective, with substitute $\Theta = WL^T$, boils down to:

$$
\begin{aligned}
l(L,W) = -\frac{1}{2}log|LL^T| + \frac{1}{2}tr(LL^TS^{yy}) \\
-tr(WL^TS^{yx}) + \frac{1}{2}tr(WW^TS^{xx})
\end{aligned}
\tag{7}
$$

The first half of equation (7), the "logdet" and the "empirical covariance" terms, are identical as in Gaussian Markov Random Field (GMRF) introduced in (Stojkovic et al., 2017; Jelisavcic et al., 2018). However, since this is a conditional extension of GMRF, aimed for the discriminative supervised task of regression, there are additional terms. The fourth term is the "conditional" part, i.e. it is the quadratic form of the features, and the third term accounts for interactions between conditional and target variables.

Objective in equation (7) is our original contribution and starting point for further study. In the reminder we present favorable properties of formulated objective, and how to exploit them for achieving better computational efficiency.

3.1. Properties

Lemma 1. *Log-det term in equation (7) is separable by columns of Cholesky factor L.*

Proof. We refer to the known result that determinant of LL^T is just sum of diagonal elements of L. □

Log-det term is the most expensive part of equation for computation ($O(p^3)$), but according to Lemma 1 in the new reparametrization of the objective function, the log-det term can be efficiently computed in linear ($O(p)$) time.

Lemma 2. *Trace terms in equation (7) are separable by columns of L and W matrices.*

Proof. L can be observed as a sum of p "rank 1" $p \times p$ matrices L_{*j}, where each matrix contains the j_{th} column of the L matrix and the rest of entries are zero ($L = \sum_j L_j$):

$$
tr(LL^TS^{yy}) = \sum_{j=1}^{p} tr(L_{*j}L_{*j}^TS^{yy}) \tag{8}
$$

Similarly for the columns of W and feature covariance term:

$$
tr(WW^TS^{yy}) = \sum_{j=1}^{p} tr(W_{*j}W_{*j}^TS^{xx}) \tag{9}
$$

Mixed term $tr(WL^TS^{yx})$ is also separable over columns of L and W since:

$$
tr(WL^TS^{yx}) = tr(S^{xy}LW^T) = \sum_{j=1}^{p} tr(S^{xy}L_{*j}W_{*j}^T) \tag{10}
$$

□

Theorem 1. *Objective $l(L,W)$ defined in eq. (7) is completely separable by columns of Cholesky factor L, in other words, contribution of column L_i ($i \in \{1,...,p\}$) to the objective l is independent of other columns L_j ($j \neq i$).*

Proof. We use the claims from Lemma 1 and Lemma 2 to write column-wise separable objective:

$$
\begin{aligned}
l(L,W) = \sum_{j=1}^{p}(-log(L_{jj}) + \frac{1}{2}tr(L_{*j}{L_{*j}}^TS^{yy}) \\
-tr(W_{*j}{L_{*j}}^TS^{yx}) + \frac{1}{2}tr(W_{*j}{W_{*j}}^TS^{xx}))
\end{aligned}
\tag{11}
$$

□

Corollary 1. *Arguments that minimize objective l in eq. (11) can be obtained by finding partial arguments for each of the columns separately, which makes this problem trivially parallelizible for up to p tasks.*

Now we can derive the expressions for the derivatives of the differentiable function $l(L,W)$. In case of L parameters, we differentiate between two separate types of variables: off-diagonal elements (when $i = j$), and diagonal elements (when $i \neq j$), so we will have two sets of update equations. We continue by using the standard matrix calculus:

$$
\begin{aligned}
\frac{\partial tr(L_{*j}{L_{*j}}^TS)}{\partial L_{ij}} &= tr((\frac{\partial L_{*j}}{\partial L_{ij}}{L_{*j}}^T + L_{*j}\frac{\partial {L_{*j}}^T}{\partial L_{ij}})S) \\
&= 2tr(\frac{\partial L_{*j}}{\partial L_{ij}}{L_{*j}}^TS) = 2L_j^TS_i
\end{aligned}
\tag{12}
$$

Here, S_i denotes the i-th row of matrix S, and L_j is the j-th column vector of the matrix L. To get the expression (12) we used the fact that matrix $\frac{\partial L_{*}j}{\partial L_{ij}}$ is actually a zero matrix with a single 1 on the i-th row and j-th column, and the cyclic property of the trace operator for symmetric matrices.

We use sparsity inducing absolute norm to penalize L and W. Differentials over L and W are respectivelly:

$$vec(\frac{\partial g(L,W)}{\partial L_{ij}})^T vec(\Delta_L) = tr(L{\Delta_L}^T S^{yy}) - tr(W{\Delta_L}^T S^{yx}) \tag{13}$$

$$vec(\frac{\partial g(L,W)}{\partial W_{lj}})^T vec(\Delta_W) = tr(\Delta_W W^T S^{xx}) - tr(\Delta_W L^T S^{yx}) \tag{14}$$

Δ_X is matrix where all elements are zeros, except particular column j which has values of j-th column in matrix X, and vec is an operator which reorders matrix into vector ("vectorization" operator). For the diagonal elements ($i = j$) differential is the same as in eq. (13), just with an addition of $-\frac{1}{L_{ii}}$ term, due to a log det contribution.

From there, we get derivatives, as well as the optimality conditions:

$$\frac{\partial g(L,W)}{\partial L_{ij}} = \sum_{k=1}^{p} {S^{yy}}_{ik} L_{kj} - \sum_{k=1}^{m} {S^{yx}}_{ik} W_{kj} \tag{15}$$

$$\frac{\partial g(L,W)}{\partial W_{lj}} = \sum_{k=1}^{m} {S^{xx}}_{lk} W_{kj} - \sum_{k=1}^{p} {S^{xy}}_{lk} L_{kj} \tag{16}$$

Separability over the columns is obvious (gradient expression, for each column, is dependent only on the elements from that same column), and the presented objective (11) is equivalent to objective (5). If we penalized Σ^{-1} and Θ matrices, or LL^T and WL^T, the results would be the same as original problem (Sohn & Kim, 2012; Wytock & Kolter, 2013; Sojoudi, 2016), but such penal would destroy attractive separability property of new objective.

We propose penalizing newly introduced parameter matrices L and W (except for the diagonal elements of L) so that objective stays separable, which is definitely computationally convenient. As for the justification from the modeling prior perspective, further investigation is needed.

As model has a large number of parameters, and may easily overfit, we add sparsity inducing regularization terms to the objective (7): $\lambda_L |L|_1 + \lambda_W |W|_1$. Now derivative equations take form:

$$\frac{\partial g(L,W)}{\partial L_{ij}} = \sum_{k=1}^{p} {S^{yy}}_{ik} L_{kj} - \sum_{k=1}^{m} {S^{yx}}_{ik} W_{kj} + \lambda_L sign(L_{ij}) \tag{17}$$

$$\frac{\partial g(L,W)}{\partial W_{lj}} = \sum_{k=1}^{m} {S^{xx}}_{lk} W_{kj} - \sum_{k=1}^{p} {S^{xy}}_{lk} L_{kj} + \lambda_W sign(W_{lj}) \tag{18}$$

3.2. Column Reordering

Even though L_1 norm is inducing sparsity in newly introduced parameter space L and W, it is not guaranteed that it will necessarily result in sparsity in original space of Σ^{-1} and Θ. We can observe an example where product of sparse Cholesky factors results into a dense matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix} \tag{19}$$

However, another Cholesky factor with same sparsity level as previous example, results into a sparse precision matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 4 \end{bmatrix} \tag{20}$$

In this situation ordering of columns plays a crucial role, and columns can be rearranged to control the mapping of sparsity from Cholesky factor to a precision matrix.

Optimal solution to reordering of the matrix to produce the sparsest Cholesky is known to be NP-complete problem. We suggest the use of the existing heuristic algorithm, the approximate minimum degree ordering (Amestoy et al., 1996), which produces a fractal-like structure with big blocks of zeros. Another alternative is to use the reverse Cuthill-McKee ordering, which reduces the bandwidth and concentrates all the nonzero elements near diagonal. Mentioned approaches have about linear computational complexity (in the number of nonzero parameters), and would be performed only once at the pre-processing stage to conveniently reorder empirical covariance matrices.

3.3. Coordinate Descent Optimization Algorithm

Lemma 3. *Optimization function (7) is convex for any feasible point $L_{ii} > 0, \forall i < N$.*

Proof. We prove convexity by showing that the Schur complement of the objective Hessian matrix must be positive definite. First, we calculate the second derivatives:

$$H_{LL} = \frac{\partial^2 g(L,W)}{\partial L_{ij} \partial L_{mn}} = \begin{cases} 0, & n \neq j \\ S^{yy}_{im}, & n = j \neq i \\ S^{yy}_{ii} + \frac{1}{L^2_{ii}}, & n = m = j = i \end{cases} \tag{21}$$

$$H_{LW} = \frac{\partial^2 g(L,W)}{\partial L_{ij} \partial W_{mn}} = \begin{cases} 0, & n \neq j \\ -S^{yx}{}_{im}, & n = j \end{cases} \tag{22}$$

$$H_{WW} = \frac{\partial^2 g(L,W)}{\partial W_{ij} \partial W_{mn}} = \begin{cases} 0, & n \neq j \\ S^{xx}{}_{im}, & n = j \end{cases} \tag{23}$$

Hessian H is positive definite if and only if matrices H_{LL} and Schur complement $H/H_{LL} = H_{WW} - H_{WL}{H_{LL}}^{-1}H_{LW}$ are positive definite.

It is worth noticing that matrix H_{LL} can be rearranged as a block matrix, with blocks corresponding to leading principal minors of S^{yy} matrix with additional positive diagonal terms corresponding to $\frac{1}{L^2_{ii}}$. Therefore, matrix H_{LL} is positive definite (it is a sum of Hermitian matrix and a diagonal matrix with non-negative entries).

Similar observations can be made for H_{WW} and H_{LW}, as they are also block-diagonal. Positive definiteness of Schur complement is defined as:

$$H/H_{LL} = H_{WW} - H_{WL}{H_{LL}}^{-1}H_{LW} > 0 \tag{24}$$

Since each matrix in equation (24) is block diagonal, by utilizing the direct sum property of the block diagonal matrices, equation (24) can be partitioned and considered separately for each block; it is sufficient to show that individual blocks of the H/H_{LL} matrix are positive definite:

$$\begin{aligned} det(H_c/H_{LL}) &= |S^{xx}{}_c - S^{xy}{}_c(S^{yy}{}_c + D_{Lc})^{-1}S^{yx}{}_c| \\ &> |S^{xx}{}_c - S^{xy}{}_c(S^{yy}{}_c)^{-1}S^{yx}{}_c| \\ &= |(x^Tx)_c - (x^Ty)_c{(y^Ty)_c}^{-1}(y^Tx)_c| \\ &= |(x^Tx)_c - (x^Tx)_c| = 0 \end{aligned} \tag{25}$$

where c in $S^{yy}{}_c$ corresponds to subset of rows and columns, and D_{Lc} is the corresponding positive diagonal term.

□

Algorithm 1 Coordinate Descent for SGCRF

```
**Preprocess:** $S^{yy}$, $S^{yx}$, $S^{xx}$ column reordering
**Initial conditions:** $L^0 = I$, $W^0 = 0$
**repeat**
  **for** $i, j$ **in** $L$ **do**
    $L_{ij} \leftarrow argmin(f_{ij}(L_j, \lambda))$ (26), (28)
  **end for**
  **for** $i, j$ **in** $W$ **do**
    $W_{ij} \leftarrow argmin(f_{ij}(W_j, \lambda))$ (27),
  **end for**
**until** $||\delta L||_2 < \epsilon$
```

Equating expressions (17) and (18) with zero, gives us first order coordinate descent update equations:

$$L_{ij} = -\frac{\sum_{k \neq i}^{p} S^{yy}{}_{ik}L_{kj} - \sum_{k=1}^{m} S^{yx}{}_{ik}W_{kj} + \lambda_L sign(L_{ij})}{S_{ii}{}^{yy}} \tag{26}$$

$$W_{lj} = -\frac{\sum_{k \neq l}^{m} S^{xx}{}_{lk}W_{kj} - \sum_{k=1}^{p} S^{xy}{}_{lk}L_{kj} + \lambda_W sign(W_{lj})}{S^{xx}_{ll}} \tag{27}$$

And for the diagonal elements of L, $(i = j)$, there is also a log det term derivative, hence the update is a positive solution of the quadratic equation (and no L_1 penal):

$$L_{ii} = \frac{-(\sum_{k \neq i} L_{ki}S_{ik} + \lambda_{ii}) + \sqrt{(\sum_{k \neq i} L_{ki}S_{ik} + \lambda_{ii})^2 + 4S_{ii}}}{2S_{ii}} \tag{28}$$

Final procedure for learning SGCRF using the derived update formulas is described by pseudocode in Algorithm (1). Since objective is convex (Lemma (3)), and regularization is separable over optimization variables, described coordinate descent Algorithm (1) converges to optimal solution.

4. Discussion and future work

We studied how Cholesky decomposition can be utilized for estimating Sparse Gaussian Conditional Random Field in a more efficient way. Observations are that proposed parameters transformations are leading to very convenient formulas with attractive properties, like linear time calculation of log det term, column-wise separable objective, and closed form solution for coordinate-wise optima. The approach can be improved on multiple points, e.g. fewer optimization steps through block-coordinate descent, or second order Newton method, and using thresholding approach (Zhang et al., 2018) to focus only on smaller set of active coordinates. In this paper we have analyzed the approach only from theoretical perspective. In the future work, we will perform thorough empirical evaluation on synthetic and real word data, to verify if the modeling assumptions are justified.

References

Amestoy, P. R., Davis, T. A., and Duff, I. S. An approximate minimum degree ordering algorithm. *SIAM Journal on Matrix Analysis and Applications*, 17(4):886–905, 1996.

Andrew, G. and Gao, J. Scalable training of l 1-regularized log-linear models. In *Proc. of the 24th international conference on Machine learning*, pp. 33–40. ACM, 2007.

Chang, K. Y., Lin, T.-p., Shih, L.-Y., and Wang, C.-K. Analysis and prediction of the critical regions of antimicrobial peptides based on conditional random fields. *PloS one*, 10(3):e0119490, 2015.

Dorta, G., Vicente, S., Agapito, L., Campbell, N. D., and Simpson, I. Training vaes under structured residuals. *arXiv preprint arXiv:1804.01050*, 2018.

Hirose, K., Fujisawa, H., and Sese, J. Robust sparse gaussian graphical modeling. *Journal of Multivariate Analysis*, 161:172–190, 2017.

Hsieh, C.-J., Dhillon, I. S., Ravikumar, P. K., and Sustik, M. A. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in neural information processing systems*, pp. 2330–2338, 2011.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., and Poldrack, R. Big & quic: Sparse inverse covariance estimation for a million variables. In *Advances in neural information processing systems*, pp. 3165–3173, 2013.

Jelisavcic, V., Stojkovic, I., Milutinovic, V., and Obradovic, Z. Fast learning of scale-free networks based on cholesky factorization. *International Journal of Intelligent Systems*, 33(6):1322–1339, 2018.

Kim, M. Sparse conditional copula models for structured output regression. *Pattern Recognition*, 60:761–769, 2016.

Krähenbühl, P. and Koltun, V. Efficient inference in fully connected crfs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems*, pp. 109–117, 2011.

McCarter, C. and Kim, S. Large-scale optimization algorithms for sparse conditional Gaussian graphical models. *Artificial Intelligence and Statistics*, pp. 528–537, 2016.

Orlando, J. I., Prokofyeva, E., and Blaschko, M. B. A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE transactions on Biomedical Engineering*, 64 (1):16–27, 2017.

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

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

Sohn, K.-A. and Kim, S. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Artificial Intelligence and Statistics*, pp. 1081–1089, 2012.

Sojoudi, S. Equivalence of graphical lasso and thresholding for sparse graphs. *The Journal of Machine Learning Research*, 17(1):3943–3963, 2016.

Stojanovic, J., Jovanovic, M., Gligorijevic, D., and Obradovic, Z. Semi-supervised learning for structured regression on partially observed attributed graphs. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 217–225. SIAM, 2015.

Stojkovic, I., Jelisavcic, V., Milutinovic, V., and Obradovic, Z. Fast sparse Gaussian markov random fields learning based on Cholesky factorization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence IJCAI-17*, pp. 2758–2764, 2017.

Sutton, C. and McCallum, A. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

Tran, V. C., Nguyen, N. T., Fujita, H., Hoang, D. T., and Hwang, D. A combination of active learning and self-learning for named entity recognition on twitter using conditional random fields. *Knowledge-Based Systems*, 132:179–187, 2017.

Wang, X., Zhang, M., and Ren, F. Sparse gaussian conditional random fields on top of recurrent neural networks. In *Thirty-Second AAAI*, 2018.

Wytock, M. and Kolter, Z. Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *Proc. of the 30th International Conference on Machine Learning*, pp. 1265–1273, 2013.

Zhang, R., Fattahi, S., and Sojoudi, S. Large-scale sparse inverse covariance estimation via thresholding and max-det matrix completion. In *International Conference on Machine Learning*, pp. 5761–5770, 2018.