Vector-Space Markov Random Fields via Exponential Families

Wesley Tansey TANSEY@CS.UTEXAS.EDU
Department of Computer Science, The University of Texas, Austin, TX 78712, USA

Oscar Hernan Madrid Padilla OSCAR.MADRID@UTEXAS.EDU
Department of Statistics and Data Sciences, The University of Texas, Austin, TX 78712, USA

Arun Sai Suggala ARUNSAI@CS.UTEXAS.EDU
Pradeep Ravikumar PRADEEPR@CS.UTEXAS.EDU
Department of Computer Science, The University of Texas, Austin, TX 78712, USA

Abstract

We present Vector-Space Markov Random Fields (VS-MRFs), a novel class of undirected graphical models where each variable can belong to an arbitrary vector space. VS-MRFs generalize a recent line of work on scalar-valued, uni-parameter exponential family and mixed graphical models, thereby greatly broadening the class of exponential families available (e.g., allowing multinomial and Dirichlet distributions). Specifically, VS-MRFs are the joint graphical model distributions where the node-conditional distributions belong to generic exponential families with general vector space domains. We also present a sparsistent M-estimator for learning our class of MRFs that recovers the correct set of edges with high probability. We validate our approach via a set of synthetic data experiments as well as a real-world case study of over four million foods from the popular diet tracking app MyFitnessPal. Our results demonstrate that our algorithm performs well empirically and that VS-MRFs are capable of capturing and highlighting interesting structure in complex, real-world data. All code for our algorithm is open source and publicly available.

1. Introduction

Undirected graphical models, also known as Markov Random Fields (MRFs), are a popular class of models for probability distributions over random vectors. Popular parametric instances include Gaussian MRFs, Ising, and Potts models, but these are all suited to specific data-types: Ising models for binary data, Gaussian MRFs for thin-tailed continuous data, and so on. Conversely, when there is prior knowledge of the graph structure but limited information otherwise, nonparametric approaches are available (Sudderth et al., 2010). A recent line of work has considered the challenge of specifying classes of MRFs targeted to the data-types in the given application, when the structure is unknown. For the specific case of homogeneous data, where each variable in the random vector has the same data-type, (Yang et al., 2012) proposed a general subclass of homogeneous MRFs. In their construction, they imposed the restriction that each variable conditioned on other variables belong to a shared exponential family distribution, and then performed a Hammersley-Clifford-like analysis to derive the corresponding joint graphical model distribution, consistent with these node-conditional distributions. As they showed, even classical instances belong to this sub-class of MRFs; for instance, with Gaussian MRFs and Ising models, the node-conditional distributions follow univariate Gaussian and Bernoulli distributions respectively.

Yang et al. (2014) then proposed a class of *mixed MRFs* that extended this construction to allow for random vectors with variables belonging to different data types, and allowing each node-conditional distribution to be drawn from a different univariate, uni-parameter exponential family member (such as a Gaussian with known variance or a Bernoulli distribution). This flexibility in allowing for different univariate exponential family distributions yielded a class of mixed MRFs over heterogeneous random vectors that were capable of modeling a much wider class of distributions than was previously feasible, opening up an entirely new suite of possible applications.

To summarize, the state of the art can specify MRFs over heterogeneous data-typed random vectors, under the re-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

striction that each variable conditioned on others belong to a uni-parameter, univariate exponential family distribution. But in many applications, such a restriction would be too onerous. For instance, a discrete random variable is best modeled by a categorical distribution, but this is a *multi-parameter* exponential family distribution, and does not satisfy the required restriction above. Other multi-parameter exponential family distributions popular in machine learning include gamma distributions with unknown shape parameter and Gaussian distributions with unknown variance. Another restriction above is that the variables be scalar-valued; but in many applications the random variables could belong to more general vector spaces, for example a Dirichlet distribution.

As modern data modeling requirements evolve, extending MRFs beyond such restrictive paradigms is becoming increasingly important. In this paper, we thus extend the above line of work in (Yang et al., 2012; 2014). As opposed to other approaches which merely cluster scalar variables (Vats & Moura, 2012), we allow node-conditional distributions to belong to a generic exponential family with a general vector space domain. We then perform a subtler Hammersley-Clifford-like analysis to derive a novel class of vector-space MRFs (VS-MRFs) as joint distributions consistent with these node-conditional distributions. This class of VS-MRFs provides support for the many modelling requirements outlined above, and could thus greatly expand the potential applicability of MRFs to new scientific analyses.

We also introduce an M-estimator for learning this class of VS-MRFs based on the sparse group lasso, and show that it is sparsistent, and that it succeeds in recovering the underlying edges of the graphical model. To solve the M-estimation problem, we also provide a scalable optimization algorithm based on Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011). We validate our approach empirically via synthetic experiments measuring performance across a variety of scenarios. We also demonstrate the usefulness of VS-MRFs by modeling a real-world dataset of over four million foods from the MyFitnessPal food database.

The remainder of this paper is organized as follows. Section 2 provides background on mixed MRFs in the uni-parameter, univariate case. Section 3 details our generalization of the mixed MRF derivations to the vector-space case. Section 4 introduces our M-estimator and derives its sparsistency statistical guarantees. Section 5 contains our synthetic experiments and the MyFitnessPal case study. Finally, Section 6 presents concluding remarks and potential future work.

2. Background: Scalar Mixed Graphical Models

Let $X = (X_1, X_2, \cdots X_p)$ be a p-dimensional random vector, where each variable X_r has domain $\mathcal{X}_r$. An undirected graphical model or a Markov Random Field (MRF) is a family of joint distributions over the random vector X that is specified by a graph $G = (V, E)$, with nodes corresponding to each of the p random variables $\{X_r\}_{r=1}^p$, and edges that specify the factorization of the joint as:

$$\mathbb{P}(X) \propto \prod_{C \in \mathcal{C}(G)} \psi_C(X_C),$$

where $\mathcal{C}(G)$ is the set of fully connected subgraphs (or cliques) of the graph G, $X_C = \{X_s\}_{s \in C}$ denotes the subset of variables in the subset $C \subseteq V$, and $\{\psi_C(X_C)\}_{C \in \mathcal{C}(G)}$ are *clique-wise* functions, each of which is a "local function" in that it only depend on the variables in the corresponding clique, so that $\psi_C(X_C)$ only depends on the variable subset X_C.

Gaussian MRFs, Ising MRFs, etc. make particular parametric assumptions on these clique-wise functions, but a key question is whether there exists a more flexible specification of the form of these clique-wise functions that is targeted to the data-type and other characteristics of the random vector X.

For the specific case where the variables are scalars, so that the domains $\mathcal{X}_r \subseteq \mathbb{R}$, in a line of work, (Yang et al., 2012; 2014) used the following construction to derive a subclass of MRFs targeted to the random vector X. Suppose that for variables $X_r \in \mathcal{X}_r$, the following (single-parameter) univariate exponential family distribution $P(X_r) = \exp\{\theta_r B_r(X_r) + C_r(X_r) - A_r(\theta_r)\}$, with natural parameter scalar θ, sufficient statistic scalar $B_r(X_r)$, base measure $C_r(X_r)$ and log normalization constant $A_r(\theta)$, serves as a suitable statistical model. Suppose that we use these univariate distributions to specify *conditional* distributions:

$$P(X_r|X_{-r}) = \exp\{ \begin{array}{l} E_r(X_{-r})B_r(X_r) + \\ C_r(X_r) - A_r(X_{-r})\} \end{array}, \quad (1)$$

where $E_r(\cdot)$ is an arbitrary function of the rest of the variables X_{-r} that serves as the natural parameter. Would these node-conditional distributions for each node $r \in V$ be consistent with some joint distribution for some specification of these functions $\{E_r(\cdot)\}_{r \in V}$? Theorem 1 from Yang et al. (2014) shows that there does exist a unique joint MRF

distribution with the form:

$$P(X;\theta) = \exp\Bigg\{ \sum_{r\in V} \theta_r B_r(X_r) + \sum_{r\in V} \sum_{t\in N(r)} \theta_{rt} B_t(X_t) B_r(X_r) + \dots + \sum_{(t_1,\dots,t_k)\in\mathcal{C}} \theta_{t_1\dots t_k}(X) \prod_{j=1}^{k} B_{t_j}(X_{t_j}) + \textstyle\sum_{r\in V} C_r(X_r) - A(\theta) \Bigg\}, \tag{2}$$

where $A(\theta)$ is the log-normalization constant. Their proof followed an analysis similar to the Hammersley-Clifford Theorem (Lauritzen, 1996), and entailed showing that for a consistent joint, the only feasible conditional parameter functions $E_r(\cdot)$ had the following form:

$$E_r(X_{-r}) = \theta_r + \sum_{t\in N(r)} \theta_{rt} B_t(X_t) + \dots + \sum_{t_2,\dots,t_k\in N(r)} \theta_{rt_2\dots t_k}(X) \prod_{j=2}^{k} B_{t_j}(X_{t_j}), \tag{3}$$

where $\theta_{r\cdot} := \{\theta_r, \theta_{rt}, \dots, \theta_{rt_2\dots t_k}\}$ is a set of parameters, and $N(r)$ is the set of neighbors of node r.

While their construction allows the specification of targeted classes of graphical models for heterogeneous random vectors, the conditional distribution of each variable conditioned on the rest of the variables is assumed to be a single-parameter exponential family distribution with a scalar sufficient statistic and natural parameter. Furthermore, their Hammersley-Clifford type analysis and sparsistency proofs relied crucially on that assumption. However in the case of multi-parameter and multivariate distributions, the sufficient statistics are a vector; indeed the random variables need not be scalars at all but could belong to a more general vector space. Could one construct classes of MRFs for this more general, but prevalent, setting? In the next section, we answer in the affirmative, and present a generalization of mixed MRFs to the vector-space case, with support for more general exponential families.

3. Generalization to the Vector-space Case

Let $X = (X_1, X_2, \cdots X_p)$ be a p-dimensional random vector, where each variable X_r belongs to a vector space $\mathcal{X}_r$. As in the scalar case, we will assume that a suitable statistical model for variables $X_r \in \mathcal{X}_r$ is an exponential family distribution

$$P(X_r) = \exp\{\sum_{j=1}^{m_r} \theta_{rj} B_{rj}(X_r) + C_r(X_r) - A_r(\theta)\}, \tag{4}$$

with natural parameters $\{\theta_{rj}\}_{j=1}^{m_r}$, and sufficient statistics $\{B_{rj}\}_{j=1}^{m_r}$, base measure $C_r(X_r)$ and log normalization constant $A_r(\theta)$. We assume the sufficient statistics $B_{rj} : \mathcal{X}_r \mapsto \mathbb{R}$ lie in some Hilbert space $\mathcal{H}_s$, and moreover specify a minimal exponential family so that:

$$\sum_{j=1}^{m_r} \alpha_j B_{rj}(X_r) \neq c, \tag{5}$$

for any constant c and any vector $\alpha \neq \mathbf{0}$. We note that even though the variables $\{X_r\}$ could lie in general vector spaces, the exponential family distribution above is finite-dimensional. However, it has multiple parameters, which is the other facet that distinguishes it from the single-parameter univariate setting of (Yang et al., 2012; 2014). We defer a generalization of our framework to infinite-dimensional exponential families to future work.

Suppose we use these general exponential family distributions to specify node-conditional distributions of variables X_r conditioned on the rest of the random variables:

$$P(X_r|X_{-r}) = \exp\{\textstyle\sum_{j=1}^{m_r} E_{rj}(X_{-r}) B_{rj}(X_r) + C_r(X_r) - A_r(X_{-r})\}, \tag{6}$$

where $\{E_{rj}(X_{-r})\}_{j=1}^{m_r}$ are arbitrary functions of the rest of the variables that serve as natural parameters for the conditional distribution of X_r. As before, we ask the question whether these node-conditional distributions can be consistent with some joint distribution for some specification of the parameter functions $\{E_{rj}(X_{-r})\}_{j=1}^{m_r}$; the following theorem addresses this very question.

Theorem 1. *Let $X = (X_1, X_2, \dots, X_p)$ be a p-dimensional random vector with node-conditional distribution of each random vector X_r conditioned on the rest of random variables as defined in (6). These node-conditionals are consistent with a joint MRF distribution over the random vector X, that is, Markov with respect to a graph $G = (V, E)$ with clique-set $\mathcal{C}$, and with factors of size at most k, **if and only if** the functions $\{E_r()\}_{r\in V}$ specifying the node-conditional distributions have the form:*

$$E_{ri}(X_{-r}) = \theta_{ri} + \sum_{t\in N(r)} \sum_{j=1}^{m_t} \theta_{ri;tj} B_{tj}(X_t) + \dots + \sum_{\substack{t_2,\dots,t_k \\ \in N(r)}} \sum_{\substack{i_2=1\dots m_{t_2} \\ \dots \\ i_k=1\dots m_{t_k}}} \theta_{ri;\dots;t_k i_k} \prod_{j=2}^{k} B_{t_j i_j}(X_{t_j}), \tag{7}$$

where $\theta_{r\cdot} = \{\theta_{ri}, \theta_{ri;tj}, \theta_{ri;\dots;t_k i_k}\}$ is a set of parameters, m_t is the dimension of the sufficient statistic vector for the t^{th} node-conditional distribution, and $N(r)$ is the set of neighbors of node r in graph G. The corresponding consistent joint MRF distribution has the following form:

$$P(X|\theta) = \exp\Bigg\{\sum_{r\in V}\sum_{i=1}^{m_r}\theta_{ri}B_{ri}(X_r) + \dots + \sum_{t_1,\dots,t_k\in C}\sum_{\substack{i_1=1\dots m_{t_1}\\ \dots \\ i_k=1\dots m_{t_k}}}\theta_{t_1 i_1;\dots;t_k i_k}\prod_{j=1}^{k}B_{t_j i_j}(X_{t_j}) + \sum_{r\in V}C_r(X_r) - A(\theta)\Bigg\} \quad (8)$$

We provide a Hammersley-Clifford type analysis as proof of this theorem in the supplementary material, which however has subtleties not present in (Yang et al., 2012; 2014), due to the arbitrary vector space domain of X_r, and the multiple parameters in the exponential families, which consequently entailed leveraging the geometry of the corresponding Hilbert spaces $\{\mathcal{H}_s\{s \in V\}$ underlying the sufficient statistics $\{B_{sj}\}$.

The above Theorem 1 provides us with a general class of vector-space MRFs (VS-MRFs), where each variable could belong to more general vector space domains, and whose conditional distributions are specified by more general finite-dimensional exponential families. Consequently, many common distributions can be incorporated into VS-MRFs that were previously unsupported or lacking in (Yang et al., 2012; 2014). For instance, gamma and Gaussian nodes, though univariate, require vector-space parameters in order to be fully modeled. Additionally, multivariate distributions that were impossible to use with previous methods, such as the multinomial and Dirichlet distributions are now also available.

3.1. Pairwise conditional and joint distributions

Given the form of natural parameters in (7), the conditional distribution of a node X_r given all other nodes X_{-r} for the special case of pairwise MRFs (i.e. $k = 2$) has the form

$$\begin{aligned} P(X_r|X_{-r},\theta_r,\theta_{rt}) &= \exp\Bigg\{\sum_{i=1}^{m_r}\theta_{ri}B_{ri}(X_r) + \sum_{t\in N(r)}\sum_{i=1}^{m_r}\sum_{j=1}^{m_t}\theta_{ri;tj}B_{tj}(X_t)B_{ri}(X_r) \\ &\quad + C_r(X_r) - A_r(X_{-r},\theta_{r\cdot})\Bigg\} \\ &= \exp\Bigg\{\left\langle B_r(X_r), \theta_r + \sum_{t\in N(r)}\theta_{rt}B_t(X_t)\right\rangle \\ &\quad + C_r(X_r) - A_r\left(\theta_r + \sum_{t\in N(r)}\theta_{rt}B_t(X_t)\right)\Bigg\} \end{aligned} \quad , \quad (9)$$

where θ_r is a vector formed from scalars $\{\theta_{ri}\}_{i=1}^{m_r}$, θ_{rt} is a matrix of dimension $m_r \times m_t$ obtained from scalars $\theta_{ri;tj}$ and $\langle ., . \rangle$ represents dot product between two vectors. Thus, the joint distribution has the form

$$P(X|\theta) = \exp\Bigg\{\sum_{r\in V}\left\langle B_r(X_r), \theta_r + \sum_{t\in N(r)}\theta_{rt}B_t(X_t)\right\rangle + \sum_{r\in V}C_r(X_r) - A(\theta)\Bigg\} \quad , \quad (10)$$

with the log-normalization constant $A(\theta) = \log\int_{\mathcal{X}}\exp\{\sum_{r\in V}\langle B_r(X_r), \theta_r + \sum_{t\in N(r)}\theta_{rt}B_t(X_t)\rangle + \sum_{r\in V}C_r(X_r)\}$. Since $A(\theta)$ is generally intractable to calculate, we next present an efficient approach to learning the structure of VS-MRFs.

4. Learning VS-MRFs

To avoid calculation of the log-normalization constant, we approximate the joint distribution in (10) with the independent product of node conditionals, also known as the *pseudo-likelihood*,

$$P(X|\theta) \approx \prod_r P(X_r|X_{-r},\theta_r,\theta_{rt})\,. \quad (11)$$

Let $\theta_{r\cdot} = \{\theta_r, \theta_{\backslash r}\}$ be the set of parameters related to the node-conditional distribution of node r, where $\theta_{\backslash r} = \{\theta_{rt}\}_{t\in V\backslash r}$. Since $A_r()$ is convex for all exponential families (Wainwright & Jordan, 2008), this gives us a loss function that is convex in $\theta_{r\cdot}$:

$$\ell(\theta_{r\cdot};\mathcal{D}) = -\tfrac{1}{n}\sum_{i}^{n}\left(\left\langle B_r(X_r^{(i)}), \theta_r + \sum_{t\in V\backslash r}\theta_{rt}B_t(X_t^{(i)})\right\rangle - A_r\left(\theta_r + \sum_{t\in V\backslash r}\theta_{rt}B_t(X_t^{(i)})\right)\right) . \quad (12)$$

We then seek to find a sparse solution in terms of both edges and individual parameters by employing the sparse group lasso regularization penalty (Friedman et al., 2010; Simon et al., 2013):

$$R(\theta_{r\cdot}) = \lambda_1\sum_{t\in V\backslash r}\sqrt{\nu_{rt}}\,\|\theta_{rt}\|_2 + \lambda_2\left\|\theta_{\backslash r}\right\|_1\,, \quad (13)$$

where $\nu_{rt} = m_r \times m_t$ is the number of parameters in the *pseudo-edge* from node r to node t (i.e., the edge (r, t) in the r^{th} node-conditional). This yields a collection of independent convex optimization problems, one for each node-conditional.

$$\underset{\theta_{r\cdot}}{\text{minimize}} \quad \ell(\theta_{r\cdot};\mathcal{D}) + R(\theta_{r\cdot}) \quad (14)$$

We next present an approach to solving this problem based on Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011).

4.1. Optimization Procedure

We first introduce a slack variable z into (14) to adhere to the canonical form of ADMM. For notational simplicity, we omit the data parameter $\mathcal{D}$ from the loss function and the subscripts in $\theta_{r\cdot}$ and A_r since it is clear we are dealing with the optimization of a single node-conditional.

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \ell(\theta) + R(z) \\ \text{subject to} & \theta = z \end{array}, \tag{15}$$

where $length(\theta) = \tau$. The augmented Lagrangian is

$$L_\alpha(\theta, z, \rho) = \ell(\theta) + R(z) + \rho^T(\theta - z) + (\alpha/2)\left\|\theta - z\right\|_2^2 . \tag{16}$$

Defining the residual of the slack $r = \theta - z$, we instead use the scaled form with $u = (1/\alpha)\rho$. ADMM proceeds in an alternating fashion, performing the following updates at each iteration:

$$\theta^{k+1} = \underset{\theta}{\text{argmin}} \left(\ell(\theta) + (\alpha/2)\left\|\theta - z^k + u^k\right\|_2^2\right) \tag{17}$$

$$z^{k+1} = \underset{z}{\text{argmin}} \left(R(z) + (\alpha/2)\left\|\theta^{k+1} - z + u^k\right\|_2^2\right) \tag{18}$$

$$u^{k+1} = u^k + \theta^{k+1} - z^{k+1} \tag{19}$$

Updating θ^{k+1}. The j^{th} subgradient of θ is $g_j(\theta) = -\overline{B}_j + \nabla_j \overline{A}(\theta) + \alpha(\theta_j + z_j^k - u_j^k)$. Note that the log-partition function, $A(\eta)$, over the natural parameters, $\eta = B\theta$, is available in closed form for most commonly-used exponential families. Thus, $\nabla^2 \overline{A}(\theta)$ is a weighted sum of rank-one matrices. In cases where the number of samples is much less than the total number of parameters (i.e. $n << \tau$), we can efficiently calculate an exact Newton update in $\mathcal{O}(\tau)$ by leveraging the matrix inversion lemma (Boyd & Vandenberghe, 2009). Otherwise, we use a diagonal approximation of the Hessian and perform a quasi-Newton update.

Updating z^{k+1}. We can reformulate (18) as the *proximal operator* (Parikh & Boyd, 2013) of $R(z)$:

$$\mathbf{prox}_{R/\alpha}(y) = \underset{z}{\text{argmin}} \left(R(z) + (\alpha/2)\left\|z - y\right\|_2^2\right) , \tag{20}$$

where $y = \theta^{k+1} + u^k$. From Friedman et al. (2010), it is straightforward to show that the update has a closed-form solution for each j^{th} block of edge parameters,

$$z_j^{k+1} = \frac{\left(\left\|S(\alpha(y_j), \lambda_2)\right\|_2 - \sqrt{\nu_j}\lambda_1\right)_+ S(\alpha(y_j), \lambda_2)}{\alpha\left\|S(\alpha(y_j), \lambda_2)\right\|_2 + \sqrt{\nu_j}\lambda_1(1-\alpha)} , \tag{21}$$

where $S(x, \lambda)$ is the soft-thresholding operator on x with cutoff at λ.

Updating u^{k+1}. Per ADMM, closed-form is given in (19).

We iterate each of the above update steps in turn until convergence, then AND pseudo-edges when stitching the graph back together.

4.2. Domain constraints

Many exponential family distributions require parameters with bounded domain. These bounds correspond to affine constraints on subsets of θ in the ADMM algorithm.[1] Often these constraints are simple implicit restrictions to $\mathbb{R}^+$ or $\mathbb{R}^-$. In these cases the log-normalization function $A(\eta)$ serves as a built-in *log-barrier* function. For instance, a normal distribution with unknown mean μ and unknown variance σ^2 has natural parameters $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = -\frac{1}{2\sigma^2}$, implying $\eta_2 < 0$. However, since $A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)$, this constraint will be effectively enforced so long as we are given a feasible starting point for η. Such a feasible point can always be discovered using a standard phase I method (Boyd & Vandenberghe, 2009). In the case of equality requirements, such as categorical and multinomial distributions, we can directly incorporate the constraints into the ADMM algorithm and solve an equality-constrained Newton's method when updating θ.

4.3. Sparsistency

We next provide the mathematical conditions that ensure with high probability our learning procedure recovers the true graph structure underlying the joint distribution. Our results rely on similar sufficient conditions to those imposed in papers analyzing the Lasso (Wainwright, 2009) and the l_1/l_2 penalty in (Jalali et al., 2011). Before stating the assumptions, we introduce the notation used in the proof.

4.3.1. Notation

Let $N(r) = \{t : \theta^*_{rt} \neq 0\}$ be the true neighbourhood of node r and let d_r be the degree of r, *i.e,* $d_r = |N(r)|$. And S_r be the index set of parameters $\{\theta^*_{rj;tk} : t \in N(r)\}$ and similarly S_r^c be the index of parameters $\{\theta^*_{rj;tk} : t \notin N(r)\}$. From now on we will overload the notation and simply use S and S^c instead of S_r and S_r^c. Let $S_r^{(ex)} = \{\theta^*_{rj;tk} : \theta^*_{rj;tk} \neq 0 \ \wedge \ t \in N(r)\}$.

[1] Note that these subsets are different than the edge-wise groups that are L_2-penalized. Rather, these constraints apply to the sum of the i^{th} value of each edge parameter and the i^{th} bias weight.

Let $Q_r^n = \nabla^2 \ell(\theta_{r.}^*; \mathcal{D})$ be the sample Fischer Information matrix at node r. As before, we will ignore subscript r and use Q^n instead of Q_r^n. Finally, we write Q_{SS}^n for the sub-matrix indexed by S.

We use the group structured norms defined in (Jalali et al., 2011) in our analysis. The group structured norm $\|u\|_{\mathcal{G},a,b}$ of a vector u with respect to a set of disjoint groups $\mathcal{G} = \{G_1, \ldots, G_T\}$ is defined as $\|(\|u_{G_1}\|_b, \ldots, \|u_{G_T}\|_b)\|_a$. We ignore the group $\mathcal{G}$ and simply use $\|u\|_{a,b}$ when it is clear from the context. Similarly the group structured norm $\|M\|_{(a,b),(c,d)}$ of a matrix $M_{p\times p}$ is defined as $\left\|(\|M^1\|_{c,d}, \ldots, \|M^p\|_{c,d})\right\|_{a,b}$. In our analysis we always use $b = 2$, $d = 2$ and to minimize the notation we use $\|M\|_{a,c}$ to denote $\|M\|_{(a,2),(c,2)}$. And we define $\|M\|_{max}$ as $\max_{i,j} |M_{i,j}|$, i.e, element wise maximum of M.

4.3.2. Assumptions

Let us begin by imposing assumptions on the sample Fisher Information matrix Q^n.

Assumption 1. *Dependency condition:* $\Lambda_{min}(Q_{SS}^n) \geq C_{min}$.

Assumption 2. *Incoherence condition:* $\left\|Q_{S^cS}^n (Q_{SS}^n)^{-1}\right\|_{\infty,2} \leq \frac{m_{min}}{m_{max}} \frac{(1-\alpha)}{\sqrt{d_r}}$ *for some* $\alpha \in (0, 1]$, *where* $m_{max} = \max_t m_t$, $m_{min} = \min_t m_t$.

Assumption 3. *Boundedness:* $\Lambda_{max}(E[B\left(X_{V\setminus r}\right) B\left(X_{V\setminus r}\right)^T]) \leq D_{max} < \infty$, *where* $B\left(X_{V\setminus r}\right)$ *is a vector such that* $B\left(X_{V\setminus r}\right) = \{B_t(X_t)\}_{t\in V\setminus r}$.

Note that the sufficient statistics $\{B_{ri}(X_r)\}_{i=1}^{m_r}$ of node r need not be bounded. So to analyze the M-estimation problem, we make the following assumptions on log-partition functions of joint and node-conditional distributions. These are similar to the conditions imposed for sparsistency analysis of GLMs.

Assumption 4. *The log partition function of the joint distribution satisfies the following conditions: for all* $r \in V$ *and* $i \in [m_r]$

1. *there exists constants* k_m, k_v *such that* $E[B_{ri}(X_r)] \leq k_m$ *and* $E[B_{ri}(X_r)^2] \leq k_v$,
2. *there exists constant* k_h *such that* $max_{u:|u|\leq 1} \frac{\partial^2 A(\theta)}{\partial \theta_{ri}^2}(\theta_{ri}^* + u, \theta_{r.}^*) \leq k_h$,
3. *for scalar variable* η *, we define a function* $\bar{A}_{r,i}$ *as:*

$$\bar{A}_{r,i}(\eta;\theta) = \log \int_{\mathcal{X}_p} \exp\Big\{\eta B_{ri}(X_r)^2 + \sum_{s\in V} C_s(X_s) + \sum_{s\in V} \Big\langle B_s(X_s), \theta_s + \sum_{t\in N(s)} \theta_{st} B_t(X_t)\Big\rangle\Big\} d(x) \tag{22}$$

Then, there exists a constant k_h *such that* $max_{u:|u|\leq 1} \frac{\partial^2 A_{r,i}(\eta;\theta_{r.}^*)}{\partial \eta^2}(u) \leq k_h$.

Assumption 5. *For all* $r \in V$, *the log-partition function* $A_r(.)$ *of the node wise conditional distribution satisfy that there exists functions* $k_1(n,p)$ *and* $k_2(n,p)$ *such that for all feasible pairs* θ *and* X, $\left\|\nabla^2 A_r(a)\right\|_{max} \leq k_1(n,p)$ *where* $a \in [b, b + 4\, k_2(n,p) \max\{\log(n), \log(p)\}\, \mathbf{1}]$ *for* $b := \theta_r + \sum_{t\in V\setminus r} \theta_{rt} B_t(X_t)$, *where for vectors* u *and* v *we define* $[u, v] := \otimes_i [u_i, v_i]$. *Moreoever, we assume that* $\left\|\nabla^3 A_r(b)\right\|_{max} \leq k_3(n,p)$ *for all feasible pairs* X *and* θ.

4.3.3. Sparsistency Theorem

Given these assumptions in 4.3.2 we are now ready to state our main sparsistency result.

Theorem 2. *Consider the vector space graphical model distribution in* (10) *with true parameters* θ^*, *edge set* E *and vertex set* V *such that the assumptions 1-5 hold. Suppose that* θ^* *satisfies* $\min_{(r,t)\in E} \|\theta_{rt}^*\|_2 \geq \frac{10\, m_{max}}{C_{min}}(\lambda_1 + \lambda_2)$ *and regularization parameters* λ_1, λ_2 *satisfy* $M_1 \frac{2-\alpha}{\alpha} \frac{m_{max}}{m_{min}} \sqrt{k_1(n,p)} \sqrt{\frac{log(pm_{max}^2)}{n}} \leq \lambda_1 + \lambda_2 \leq M_2 \frac{2-\alpha}{\alpha} k_1(n,p)\, k_2(n,p)$ *for positive constants* M_1 *and* M_2 *and* $\lambda_2 < \left(\frac{\alpha}{2-\alpha+2\, m_{max}/m_{min}}\right)\lambda_1$. *Then, there exists constants* L, c_1, c_2 *and* c_3 *such that if* $n \geq max\Big\{L \frac{m_{max}^9}{m_{min}} d^2\, k_1(n,p)\, (k_3(n,p))^2\, (logp')^2\, log\left(p\, m_{max}^2\right), \frac{4\, log\left(p\, m_{max}^2\right)}{k_1(n,p)\, k_4\, k_2(n,p)^2}, \frac{8\, k_h^2}{k_4^2} log\left(\sum_t m_t\right)\Big\}$, *with probability at least* $1 - c\,(p')^{-3}\left(\sum_t m_t\right) - exp(-c_2\, n) - exp(-c_3\, n)$, *the following statements hold.*

- *For each node* $r \in V$, *the solution of the M-estimation problem* (14) *is unique*
- *Moreover, for each node* $r \in V$ *the M-estimation problem recovers the true neighbourhood exactly.*

where $m_{max} = max\, m_t$, $m_{min} = min\, m_t$, $p' = max(n, p)$.

The proof of Theorem 2 follows along similar lines to the sparsistency proof in (Yang et al., 2014), albeit with a subtler analysis to support general vector-spaces. It is based on the primal dual witness proof technique and relies on the previous results. We refer the interested reader to the supplementary material for additional details regarding the proofs.

5. Experiments

We demonstrate the effectiveness of our algorithm on both synthetic data and a real-world dataset of over four million foods logged on the popular diet app, MyFitnessPal.

5.1. Synthetic experiments

The synthetic experiments were run on a vector-space mixed MRF consisting of eight Bernoulli, eight gamma

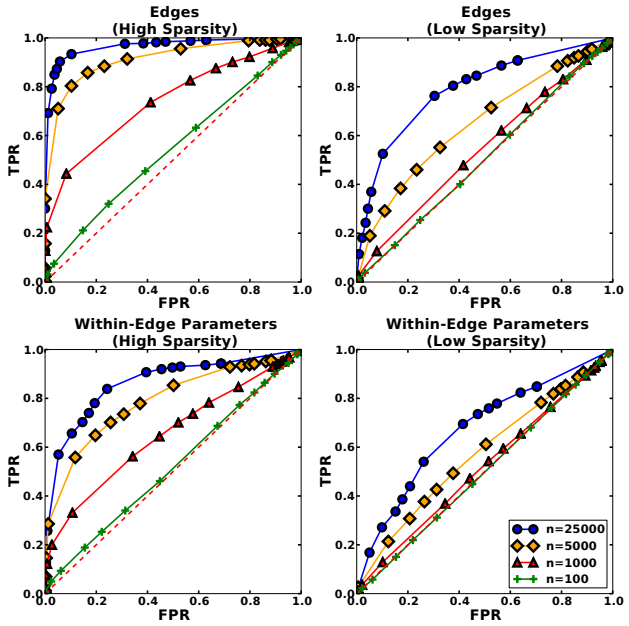

Figure 1. ROC curves for our synthetic experiments. The top left and bottom left plots show both edge as well as within-edge-parameter recovery performance respectively, for graphs with a high degree of sparsity. The two right plots show the same performance measures, but for graphs with a relatively low degree of sparsity. The low sparsity scenario is more challenging, requiring more data to recover the majority of the graph.

(with unknown shape and rate), eight Gaussian (with unknown mean and variance), and eight Dirichlet (k=3) nodes. The choice of these node-conditional distributions is meant to highlight the ability of VS-MRFs to model many different types of distributions. Specifically, the Bernoulli represents a univariate, uni-parameter distribution that would still be possible to incorporate into existing mixed models. The gamma and Gaussian distributions are both multi-parameter, univariate distributions which would have required fixing one parameter (e.g. fixing the Gaussians' variances) to be compatible with previous approaches. Finally, the Dirichlet distribution is multi-parameter *and* multivariate, thereby making VS-MRFs truly unique in their ability to model this joint distribution.

For each experiment, we conducted 30 independent trials by generating random weights and sampling via Gibbs sampling with a burn-in of 2000 and thinning step size of 10. We consider two different sparsity scenarios: high (90% edge sparsity, 50% intra-edge parameter sparsity) and low (50% edge sparsity, 10% intra-edge parameter sparsity). Edge recovery capability is examined by fixing λ_2 to a small value and varying λ_1 over a grid of values in the range $[0.0001, 0.5]$; parameter recovery is examined analogously by fixing λ_1 and varying λ_2. We use `AND` graph stitching and measure the true positive rate (TPR) and false positive rate (FPR) as the number of samples increases from 100 to 25K.

Figure 5 shows the ROC curves at both the edge and parameter levels. The results demonstrate that our algorithm improves well as the dataset size scales. They also illustrate that graphs with a higher degree of sparsity are easier to recover with fewer samples. In both the high and low sparsity graphs, the algorithm is better able to recover the coarse-grained edge structure than the more fine-grained within-edge parameter structure, though both improve favourably with the size of the data.

5.2. MyFitnessPal Food Dataset

MyFitnessPal[2] (MFP) is one of the largest diet-tracking apps in the world, with over 80M users worldwide. MFP has a vast crowd-sourced database of food data, where each food entry contains a description, such as "Trader Joe's Organic Carrots," and a vector of sixteen macro- and micro-nutrients, such as fat and vitamin C.

We treat these foods entries as random vectors with an underlying VS-MRF distribution, which we learn treating the food entries in the database as samples from the underlying VS-MRF distribution. The text descriptions are tokenized, resulting in a dictionary of approximately 2650 words; we use a Bernoulli distribution to model the conditional distribution of each word. The conditional distribution of each nutrient (on a per-calorie basis) is generally gamma distributed, but contains spikes at zero[3] and large outlier values.[4] The gamma distribution is undefined at zero, and the outlier values can result in numerical instability during learning, which thus suggests using a distribution other than the vanilla gamma distribution. Such zero-inflated data are common in many biostatistics applications, and are typically modeled via a mixture model density of the form $p(Z) = \pi\,\delta_0 + (1-\pi)\,g(z)$, where δ_0 is the dirac delta at zero, and $g(z)$ is the density of the non-zero-valued data. Unfortunately, such mixture models are not generally representable as exponential families.

To overcome this, we introduce the following class of **point-inflated exponential family** distributions. For any random variable $Z \in \mathcal{Z}$, consider any exponential family $P(Z) = \exp(\eta^T B(Z) + C(Z) - A(\eta))$, with sufficient statistics $B(\cdot)$, base measure $C(\cdot)$, and log-normalization constant $A(\cdot)$. We consider an inflated variant of this random variable, inflated at some value j; note that this could potentially lie outside the domain $\mathcal{Z}$, in which case the domain of the inflated random variable would become $\mathcal{Z} \cup \{j\}$. We then define the corresponding point-inflated

[2] http://myfitnesspal.com

[3] This is common in foods since many dishes are marketed as "fat free" or contain low nutrient density (e.g. soda).

[4] This occurs when foods contain few calories but a large amount of some micro-nutrient (e.g. multi-vitamins)

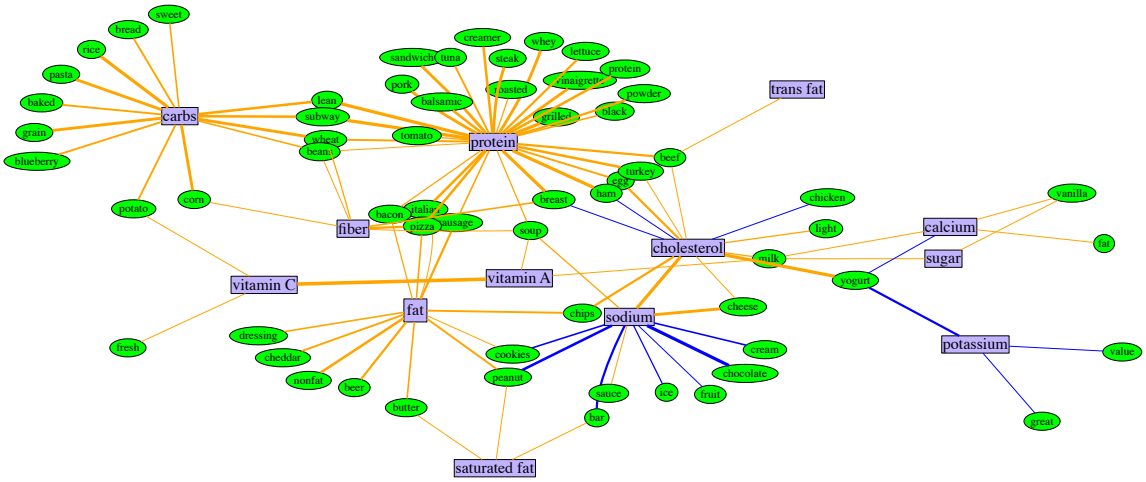

Figure 2. The top 100 edges in the MyFitnessPal food graph. Purple rectangular nodes correspond to macro- and micro-nutrients; green oval nodes correspond to food description terms. Edge color is determined by the approximate effect of the edge on the means of the node-conditionals: darker, blue edges represent lower means; brighter, orange edges represent higher means; thickness corresponds to the norm of the edge weight.

exponential family distribution as:

$$P_{\text{infl}}(Z) = \exp\left\{\eta_0 \mathbb{I}(Z = j) + \eta_1^T B(z) + C(z) - A_{\text{infl}}(\eta)\right\},$$

where $A_{\text{infl}}(\eta)$ is the log-normalization constant of the point-inflated distribution which can be expressed in terms of the log-normalization constant $A(\cdot)$ of the uninflated exponential family distribution: $A_{\text{infl}}(\eta) = \log\left(\exp\{\eta_0\} - \exp\{\eta_1^T B(j)\mathbb{I}(j \in \mathcal{Z})\} + \exp\{A(\eta_1)\}\right)$. Thus, as long as we have a closed form $A(\cdot)$ for the log-partition function of the base distribution, we can efficiently calculate $P_{\text{infl}}(Z)$. The definition also permits an arbitrary number of inflated points by recursively specifying the base distribution as another point-inflated model. We model each of the MFP nutrients via a two-point-inflated gamma distribution, with points at zero and a winsorized outlier bucket.

Due to the size of the overall graph, presenting it in closer detail here is not feasible. To give a qualitative perspective of the relationships captured by our algorithm, we selected the top 100 edges in the MFP food graph by ranking the edges based on their L2-norm. We then calculated their approximate contribution to the mean of their corresponding node-conditionals to determine edge color and thickness. Figure 2 shows the results of this process, with edges that contribute positively colored in orange and edges that contribute negatively colored in blue; edge thickness corresponds to the magnitude of the contribution. A high-level view of the entire learned graph is available in the supplementary materials.

Several interesting relationships can be discovered, even from just this small subset of the overall model. For instance, the negative connection between "peanut" and sodium may seem counter-intuitive, given the popularity of salted nuts, yet on inspection of the raw database it appears that indeed many peanut-based foods are actually very low in sodium on a per-calorie basis. As another example, "chips" are often thought of as a high-carb food, but the graph suggests that they actually tend to be a bigger indicator of high fat. In general, we believe there is great potential for wide-ranging future uses of VS-MRFs in nutrition and other scientific fields, with the MFP case study only scratching the surface of what can be achieved.

6. Conclusion

We have presented vector-space MRFs as a flexible and scalable approach to modeling complex, heterogeneous data. In particular, we generalize the concept of mixed MRFs to allow for node-conditional distributions to be distributed according to a generic exponential family distribution, that is potentially multi-parameter and even multivariate. Our VS-MRF learning algorithm has reassuring sparsistency guarantees and was validated against a variety of synthetic experiments and a real-world case study. We believe that the broad applicability of VS-MRFs will make them a valuable addition to the scientific toolbox. All code for our VS-MRF implementation is publicly available.[5]

7. Acknowledgments

We acknowledge the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1320894, IIS-1447574, and DMS-1264033.

References

Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2009.

Boyd, Stephen, Parikh, Neal, Chu, Eric, Peleato, Borja, and

[5] https://github.com/tansey/vsmrfs

Eckstein, Jonathan. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010. URL http://arxiv.org/abs/1001.0736.

Jalali, Ali, Ravikumar, Pradeep, Vasuki, Vishvas, and Sanghavi, Sujay. On learning discrete graphical models using group-sparse regularization. *AI STAT*, 2011.

Lauritzen, Steffen L. *Graphical models*. Oxford University Press, 1996.

Parikh, Neal and Boyd, Stephen. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

Simon, Noah, Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Sudderth, Erik B, Ihler, Alexander T, Isard, Michael, Freeman, William T, and Willsky, Alan S. Nonparametric belief propagation. *Communications of the ACM*, 53(10): 95–103, 2010.

Vats, Divyanshu and Moura, José MF. Finding non-overlapping clusters for generalized inference over graphical models. *Signal Processing, IEEE Transactions on*, 60(12):6368–6381, 2012.

Wainwright, M. J. Sharp thresholds for noisy and high-dimensional recovery of sparsity using l1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, pp. 2183–2202, 2009.

Wainwright, Martin J and Jordan, Michael I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2): 1–305, 2008.

Yang, Eunho, Allen, Genevera, Liu, Zhandong, and Ravikumar, Pradeep. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pp. 1358–1366, 2012.

Yang, Eunho, Baker, Yulia, Ravikumar, Pradeep, Allen, Genevera, and Liu, Zhandong. Mixed graphical models via exponential families. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 1042–1050, 2014.