
Supplementary Material

Learning Higher-Order Graph Structure with Features by Structure Penalty

Shilin Ding¹, Grace Wahba^{1,2,3}, and Xiaojin Zhu²

Department of ¹Statistics, ²Computer Sciences, ³Biostatistics and Medical Informatics}
 University of Wisconsin-Madison, WI 53705
 {sding, wahba}@stat.wisc.edu, jerryzhu@cs.wisc.edu

1 Multivariate Bernoulli model

The multivariate Bernoulli (MVB) model of K random variables has $2^K - 1$ natural parameters [1]. Given the predictive variable X , these parameters are functions of X , called conditional log odds ratios. From the distribution of the MVB, $f^\omega(X)$ can be written as:

$$f^\omega(x) = \log OR(Y_i, i \in \omega | Y_j = 0, j \notin \omega; X = x) \quad (1)$$

Here, the odds ratios are calculated recursively

$$OR(Y_i | X = x) = \frac{P(Y_i = 1 | X = x)}{1 - P(Y_i = 1 | X = x)}, \quad (2)$$

$$OR(Y_i, i \in \omega \cup \{k\} | X = x) = \frac{OR(Y_i, i \in \omega | Y_k = 1, X = x)}{OR(Y_i, i \in \omega | Y_k = 0, X = x)}, \text{ with } k \notin \omega \quad (3)$$

The following two notations are useful in optimization and parameter tuning:

$$S^\omega(y; x) = \sum_{\kappa \subseteq \omega} y^\kappa f^\kappa(x); \quad S^\omega(x) = \sum_{\kappa \subseteq \omega} f^\kappa(x); \quad (4)$$

It follows from the definition of the conditional log odds ratio in (1) that

$$\exp(S^\omega(x)) = \frac{P(Y_i = 1, i \in \omega, \text{ and } Y_j = 0, j \in \Omega \setminus \omega | X = x)}{P(Y_i = 0, i \in \omega | X = x)} \quad (5)$$

Then the normalization factor is:

$$\exp(b(f(x))) = 1 + \sum_{\omega \in \Psi_K} \exp(S^\omega(x)) \quad (6)$$

In practice, the $\exp(b(f(x)))$ is calculated by the junction tree algorithm to avoid enumerating 2^K possible values of Y , which is intractable in large graphs.

2 Dual of the proximal linearization problem

To solve the following objective of the proximal linearization problem

$$\min_c L_k + \nabla L_k^T(c - c_k) + \frac{\alpha_k}{2} \|c - c_k\|^2 + \lambda J(c) \quad (7)$$

we solve its dual problem as suggested in Liu and Ye [2]. Let $Z = \{v \in \Psi_K | \|c^{T_v}\| = 0\}$, and $\bar{Z} = \Psi_K - Z$ be the complement. Define $s_v, v \in \Psi_K$ as:

$$s_v \in \mathbb{S}_v = \{s = (s^\omega)_{\omega \in \Psi_K} \mid s \in \mathbb{R}^{\bar{P}}, \|s\| \leq \lambda p_v, s^\omega = 0 \text{ if } \omega \in T_v\} \quad (8)$$

then the subgradient of (7) is:

$$\nabla L + \alpha_k(c - c_k) + \sum_{v \in Z} s_v + \sum_{u \in \bar{Z}} r_u \quad (9)$$

where s_v is the subgradient of $\lambda p_v \|c^{T_v}\|$ for $v \in Z$ and r_u is the subgradient for $u \in \bar{Z}$:

$$r_u = \arg \max_{s_u} \langle s_u, c \rangle, \text{ for } u \in \bar{Z} \quad (10)$$

The subgradient s_v is in a unit ball of certain subspace of $\mathbb{R}^{\bar{p}}$. These subspaces are not perpendicular to each other. Thus, s_v 's are not separable, and closed form solution of (7) cannot be obtained. We solve the proximal subproblem (7) by its smoothing and convex dual problem. Note (7) is equivalent to:

$$\min_{c \in \mathbb{R}^{\bar{p}}} \max_{S \in \mathbb{S}} \phi(c, S) = \nabla L_k^T(c - c_k) + \frac{\alpha_k}{2} \|c - c_k\|^2 + \sum_{v \in \Omega} \langle s_v, c \rangle \quad (11)$$

where S is a $\tilde{p} \times |\Psi_K|$ matrix whose columns are s_v . $\mathbb{S} = \{S | S = (s_1, \dots, s_v, \dots, s_\Omega), s_v \in \mathbb{S}_v \text{ for } v \in \Psi_K\}$ is the feasible region of S . Since $\phi(\cdot, S)$ is lower semicontinuous and $\phi(c, \cdot)$ is upper semicontinuous, there exists a saddle point and the max and min are exchangeable. The solution of minimizing $\phi(c, S)$ is:

$$\tilde{c} = \arg \min_c \phi(c, S) = c_k - \frac{1}{\alpha_k} \nabla L_k - \frac{1}{\alpha_k} \sum_v s_v \quad (12)$$

Substitute \tilde{c} back into (11), we have the dual problem of (7) as:

$$\max_{S \in \mathbb{S}} \eta(S) = -\frac{1}{2} \left\| \sum_v s_v \right\|^2 + (\alpha_k c_k - \nabla L_k)^T \sum_v s_v \quad (13)$$

Following the proof in Liu and Ye [2], we can show that $\eta(S)$ is convex and Lipschitz continuous. The differential is $\alpha_k \tilde{c} e^T$ where $e \in \mathbb{R}^{\bar{p}}$ is a vector of ones. Hence, (13) can be solved by existing gradient methods.

3 B-spline

Given m knots, $t_0 \leq t_1 \leq \dots \leq t_{m-1}$, the B-spline basis functions of degree d are defined recursively [3]:

$$b_{k,0} = \begin{cases} 1; & \text{if } t_k \leq t < t_{k+1} \\ 0; & \text{otherwise} \end{cases}, \text{ for } k = 0, \dots, m-2$$

$$b_{k,l} = \frac{t - t_k}{t_{k+l} - t_k} b_{k,l-1}(t) + \frac{t_{k+l+1} - t}{t_{k+l+1} - t_{k+1}} b_{k+1,l-1}(t), \text{ for } k = 0, \dots, m-d-2; l = 0, \dots, d$$

Let $B_k(\cdot) = b_{k,d}(\cdot)$, then $\{B_k, k = 0, \dots, m-d-2\}$ are $m-d-1$ basis functions, which span the functional space \mathcal{B} . The B-spline curve in \mathcal{B} is:

$$g(t) = \sum_{k=0}^{m-d-2} c_k B_k(t) \quad (14)$$

where c_k 's are the control points to be estimated. In our simulation studies, c_k 's are assumed to be one dimensional scalars for simplicity.

We let each $f^\omega(x)$ where $x = (x_1, \dots, x_p)'$ be in $\mathcal{B}_0 \oplus \mathcal{B}_1 \oplus \dots \oplus \mathcal{B}_p$. Here, \mathcal{B}_0 is a space of constant functions and $\mathcal{B}_j, j = 1 \dots, p$ is a B-spline functional space on domain $x_j \in \mathcal{X}_j$. Therefore,

$$f^\omega(x) = c_0^\omega + \sum_{j=1}^p g_j(x_j) \quad (15)$$

where $g_j(x_j) \in \mathcal{B}_j$ are defined in (14).

4 Tuning

For i -th data point $(y(i), x(i))$, denote $S_i^\omega = S^\omega(x(i))$, then the normalization factor of the i -th data is $b_i = b(f(x(i))) = \log(1 + \sum_\omega \exp S_i^\omega)$. The mean of the augmented response $\mathcal{Y}(i)$ in the MVB model is:

$$\mu(i) = E[\mathcal{Y}(i)|x(i), f] = (\mu^1(i), \dots, \mu^\kappa(i), \dots, \mu^\Omega(i)) \quad (16)$$

$$\text{where } \mu^\kappa(i) = \frac{\partial b_i}{\partial f^\kappa} = \frac{\sum_{\omega \in T_\kappa} \exp S_i^\omega}{\exp b_i} \quad (17)$$

The $|\Psi_K| \times |\Psi_K|$ covariance matrix of the augmented response is:

$$W(i) = \text{var}(\mathcal{Y}(i)|x(i), f) \quad (18)$$

where the (α, β) -th element of $W(i)$ is:

$$W_{\alpha, \beta}(i) = \frac{\partial^2 b_i}{\partial f^\alpha (\partial f^\beta)^T} = \frac{\sum_{\omega \in T_\alpha \cap T_\beta} \exp S_i^\omega}{\exp b_i} - \mu^\alpha(i) \cdot \mu^\beta(i) \quad (19)$$

Let R_v be a $\tilde{p} \times \tilde{p}$ diagonal matrix whose (i, i) -th element is 1 if $c_i \neq 0$. Then, the v -th group penalty $J(f^{T_v})$ can be written as:

$$J(f^{T_v}) = p_v \sqrt{\sum_{\omega \in T_v} \|f^\omega\|^2} = p_v \|R_v c\| \quad (20)$$

Note R_v is symmetric and $R_v \cdot R_v = R_v$, direct calculation yields the derivative and Hessian of the penalty term:

$$\frac{\partial J}{\partial c} = \sum_{v: R_v c \neq 0} p_v \frac{R_v c}{\|R_v c\|} \quad (21)$$

$$\frac{\partial^2 J}{\partial c \partial c^T} = \sum_{v: R_v c \neq 0} p_v J_v = \sum_{v: R_v c \neq 0} p_v \frac{R_v (\|R_v c\|^2 I - c \cdot c^T) R_v}{\|R_v c\|^3} \quad (22)$$

where $J_v \doteq (R_v (\|R_v c\|^2 I - c \cdot c^T) R_v) / \|R_v c\|^3$. Denote the grand design matrix as:

$$D = (D(1)^T \quad \dots \quad D(n)^T)^T \quad (23)$$

$$\text{where } D(i) = \begin{pmatrix} x(i)^T & 0 & \dots & 0 \\ 0 & x(i)^T & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x(i)^T \end{pmatrix} \quad (24)$$

Suppose there are N non-zero elements of c at location $\{a_1, \dots, a_N\}$. Let \tilde{D} be the matrix composed by the a_1, \dots, a_N th column of D . Then, the Hessian matrix of I_λ is:

$$\frac{\partial^2 I_\lambda}{\partial c \partial c^T} = \frac{\partial^2 L}{\partial c \partial c^T} + \lambda \frac{\partial^2 J}{\partial c \partial c^T} = \tilde{D}^T W \tilde{D} + \lambda \sum_{v: R_v c \neq 0} p_v J_v \quad (25)$$

Let H be the $n|\Psi_K| \times n|\Psi_K|$ influence matrix that implies

$$f_{\lambda, \epsilon} - f_\lambda \approx H \epsilon \quad (26)$$

where ϵ is a small perturbation on \mathcal{Y} ; $f_\lambda = D c_\lambda$ is the estimated function value with tuning parameter λ ; and $f_{\lambda, \epsilon}$ is the estimated function value with the perturbation. Then, the analysis of the first order Taylor expansion of $\frac{\partial I_\lambda}{\partial c}(\mathcal{Y} + \epsilon, c_{\lambda, \epsilon})$ leads to the formulation of H as follows (refer to Xiang and Wahba [4] and Ma [5] Chapter 3 for more details)

$$H = \tilde{D} \left(\frac{\partial^2 I_\lambda}{\partial c \partial c^T} \right)^{-1} \tilde{D}^T = \tilde{D} \left(\tilde{D}^T W \tilde{D} + \lambda \sum_{v: R_v c \neq 0} p_v J_v \right)^{-1} \tilde{D}^T \quad (27)$$

The (i, j) -th $q \times q$ submatrix of H is

$$H(i, j) = \tilde{D}(i)^T \left(\frac{\partial^2 I_\lambda}{\partial c \partial c^T} \right)^{-1} \tilde{D}(j) \quad (28)$$

Let $Q(i) = I - H(i, i)W(i)$ for $i = 1, \dots, n$, define the generalized average matrix, denoted as \bar{Q} , of $\{Q(i), i = 1, \dots, n\}$ as follows

$$\bar{Q} = (\delta - \gamma)I_{q \times q} + \gamma \cdot ee^T = \begin{pmatrix} \delta & \gamma & \cdots & \gamma \\ \gamma & \delta & \cdots & \gamma \\ \vdots & \vdots & \ddots & \vdots \\ \gamma & \gamma & \cdots & \delta \end{pmatrix} \quad (29)$$

where e is the unit vector of length q and

$$\delta = \frac{1}{nq \sum_{i=1}^n \text{tr}(Q(i))}, \quad \gamma = \frac{1}{nq(q-1)} [e^T Q(i)e - \text{tr}(Q(i))] \quad (30)$$

Let \bar{H} be the generalized average of $\{H(i, i), i = 1, \dots, n\}$, the GACV score is

$$GACV(\lambda) = OBS(\lambda) + \frac{1}{n} \sum_{i=1}^n \mathcal{Y}(i)^T \bar{Q}^{-1} \bar{H} (\mathcal{Y}(i) - \mu(i)) \quad (31)$$

where

$$OBS(\lambda) = \frac{1}{n} \left[-\mathcal{Y}(i)^T f_\lambda(x(i)) + b(f_\lambda(x(i))) \right] \quad (32)$$

is the observed log-likelihood.

The degrees of freedom of multivariate Bernoulli data is generally difficult to obtain. But we can have a good approximation from GACV [6] as

$$\hat{df}(\lambda) = \sum_{i=1}^n \mathcal{Y}(i)^T \bar{Q}^{-1} \bar{H} (\mathcal{Y}(i) - \mu(i)) \quad (33)$$

So the BGACV score can be defined as

$$BGACV(\lambda) = OBS(\lambda) + \frac{1}{n} \frac{\log n}{2} \sum_{i=1}^n \mathcal{Y}(i)^T \bar{Q}^{-1} \bar{H} (\mathcal{Y}(i) - \mu(i)) \quad (34)$$

For the model selection criteria AIC, the degree of freedom is approximated by the number of non-zero c_{jk} 's in the group penalty.

References

- [1] J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley (Chichester England and New York), 1990.
- [2] J. Liu and J. Ye. Fast overlapping group lasso. *arXiv:1009.0306v1*, 2010.
- [3] C. De Boor. *A practical guide to splines*. Applied Mathematical Sciences, 1978.
- [4] D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, 6:675–692, 1996.
- [5] Xiwen Ma. *Penalized Regression in Reproducing Kernel Hilbert Spaces With Randomized Covariate Data*. PhD thesis, Department of Statistics, University of Wisconsin-Madison, 2010.
- [6] W. Shi, G. Wahba, S. Wright, K. Lee, R. Klein, and B. Klein. LASSO-Patternsearch algorithm with application to ophthalmology and genomic data. *Statistics and its Interface*, 1(1):137, 2008.