Supplementary of "Nonlinear Statistical Learning with Truncated Gaussian Graphical Models"

1. Training TGGM for Classification

Similar to the regression model, the derivatives of $\mathcal{Q}(\cdot)$ can be derived as

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{W}_0} = -\frac{1}{\sigma_0^2}\left(\mathbb{E}[\mathbf{H}|\mathbf{X}] - \mathbb{E}[\mathbf{H}|\mathbf{Y},\mathbf{X}]\right)\mathbf{X}^T, \tag{1}$$

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{b}_0} = -\frac{1}{\sigma_0^2}\left(\mathbb{E}[\mathbf{H}|\mathbf{X}] - \mathbb{E}[\mathbf{H}|\mathbf{Y},\mathbf{X}]\right)\mathbf{1}_N, \tag{2}$$

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{W}_1} = -\Big(\mathbf{W}_1\mathbb{E}\big[\mathbf{H}\mathbf{H}^T|\mathbf{Y},\mathbf{X}\big] - \left(\mathbb{E}\left[\mathbf{Z}|\mathbf{Y},\mathbf{X}\right] - \mathbf{b}_1\mathbf{1}_N^T\right)\mathbb{E}[\mathbf{H}^T|\mathbf{Y},\mathbf{X}]\Big), \tag{3}$$

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{b}_1} = -\left(N\mathbf{b}_1 - \left(\mathbb{E}[\mathbf{Z}|\mathbf{Y},\mathbf{X}] - \mathbf{W}_1\mathbb{E}[\mathbf{H}|\mathbf{Y},\mathbf{X}]\right)\mathbf{1}_N\right) \tag{4}$$

where $\mathbf{Z} \triangleq [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N]$. With the gradients, we can update the model parameters $\mathbf{\Theta}$ using appropriate optimization algorithms, such as SGD and its variants.

The prior expectation $\mathbb{E}[\cdot|\mathbf{X}]$ can be computed easily due to $p(\mathbf{H}|\mathbf{X})$ comprising of univariate truncated normals (Johnson et al., 1994). For the posterior expectation $\mathbb{E}[\cdot|\mathbf{Y},\mathbf{X}]$, we resort to the mean-field VB approximation. Define $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N]$ with $\mathbf{s}_i \triangleq \mathbf{T}_i\mathbf{z}_i$. Suppose a fully factorized distribution $q(\mathbf{H},\mathbf{S}) = \prod_{i=1}^N\prod_{k=1}^K q(\mathbf{h}_i(k))q(\mathbf{s}_i(k))$. Then, we minimize the KL-divergence between $q(\mathbf{H},\mathbf{S})$ and the true posterior $p(\mathbf{Y},\mathbf{S},\mathbf{H}|\mathbf{X}) = \prod_{i=1}^N \mathcal{N}_T\left(\mathbf{h}_i\left|\mathbf{W}_0\mathbf{x}_i + \mathbf{b}_0, \sigma_0^2\mathbf{I}_M\right.\right) \times \mathcal{N}\left(\mathbf{s}_i\middle|\mathbf{T}_i(\mathbf{W}_1\mathbf{h}_i + \mathbf{b}_1), \mathbf{T}_i\mathbf{T}_i^T\right) \times \prod_{k\neq y_i} I(\mathbf{s}_i(k) \geq 0)$, with the KL-divergence expressed as

$$\begin{aligned} KL = &-\sum_{i=1}^N \frac{1}{2\sigma_0^2}\left\langle \|\mathbf{h}_i - \mathbf{W}_0\mathbf{x}_i - \mathbf{b}_0\|^2\right\rangle_q + \langle I(\mathbf{h}_i \geq \mathbf{0})\rangle_q \\ &-\sum_{i=1}^N \ln Z_i - \sum_{i=1}^N \frac{1}{2}\left\langle \left\|\mathbf{T}_i^{-1}\mathbf{s}_i - \mathbf{W}_1\mathbf{h}_i - \mathbf{b}_1\right\|^2\right\rangle_q \\ &+\sum_{i=1}^N\sum_{k\neq y_i}\langle \ln\left(\mathbf{s}_i(k) \geq 0\right)\rangle_q - \frac{MN\ln 2\pi}{2} + \mathcal{H}(q), \end{aligned} \tag{5}$$

where $\langle\cdot\rangle$ means expectation taken w.r.t. $q(\mathbf{H},\mathbf{S})$; and $\mathcal{H}(q)$ is the entropy of $q(\mathbf{H},\mathbf{S})$. For convenience of presentation, denote $\mathbf{v}_i = [\mathbf{h}_i^T, \mathbf{s}_i^T]^T$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N]$. Thereby, $q(\mathbf{H},\mathbf{S})$ can now be denoted as $q(\mathbf{V})$. It is known that when all $q(\mathbf{v}_s(\ell))$ except $(\ell, s) = (k, i)$ are known, the KL-divergence is minimized if $\ln q(\mathbf{v}_i(k)) = \langle \ln p(\mathbf{y}_i, \mathbf{v}_i|\mathbf{x}_i)\rangle_{\neq(k,i)} + const$ (Jordan et al., 1999). Following the similar procedures and arrangements in regression, it can be obtained that

$$\begin{aligned} &\ln p(\mathbf{y}_i, \mathbf{v}_i|\mathbf{x}_i) \\ &= -\frac{1}{2}\mathbf{P}_i(k,k)\mathbf{v}_i^2(k) + \sum_{k\neq M+y_i}\ln I(\mathbf{v}_i(k) \geq 0) + C_3 \\ &\quad + \left(\boldsymbol{\gamma}_i(k) - \mathbf{P}_i(k,-k)\mathbf{v}_i(-k)\right)\mathbf{v}_i(k), \end{aligned} \tag{6}$$

where C_3 represents all terms without reliance on $\mathbf{v}_i(k)$; and

$$\mathbf{P}_i \triangleq \begin{bmatrix} \frac{1}{\sigma_0^2}\mathbf{I}_M + \mathbf{W}_1^T\mathbf{W}_1 & -\mathbf{W}_1^T \\ -\mathbf{W}_1 & \left(\mathbf{T}_i\mathbf{T}_i^T\right)^{-1}, \end{bmatrix}, \tag{7}$$

$$\boldsymbol{\gamma}_i \triangleq \begin{bmatrix} \frac{1}{\sigma_0^2}(\mathbf{W}_0\mathbf{x}_i + \mathbf{b}_0) - \mathbf{W}_1^T\mathbf{b}_1 \\ (\mathbf{T}_i^{-1})^T\mathbf{b}_1. \end{bmatrix}. \tag{8}$$

From the fact $\ln q(\mathbf{v}_i(k)) = \langle \ln p(\mathbf{y}_i, \mathbf{v}_i|\mathbf{x}_i)\rangle_{\neq(k,i)} + const$, it can be derived that

$$q(\mathbf{v}_i(k)) = \begin{cases} \mathcal{N}_T\left(\mathbf{v}_i(k)\left|\boldsymbol{\varsigma}_i(k), \frac{1}{\mathbf{P}_i(k,k)}\right.\right), & \text{if } k \neq M + y_i, \\ \mathcal{N}\left(\mathbf{v}_i(k)\left|\boldsymbol{\varsigma}_i(k), \frac{1}{\mathbf{P}_i(k,k)}\right.\right), & \text{otherwise,} \end{cases} \tag{9}$$

where $\boldsymbol{\varsigma}_i$ is defined as $\boldsymbol{\varsigma}_i(k) = \frac{\boldsymbol{\gamma}_i(k) - \tilde{\mathbf{P}}_i(k,:)\langle\mathbf{v}_i\rangle_q}{\mathbf{P}_i(k,k)}$ with $\tilde{\mathbf{P}}_i = \mathbf{P}_i - \mathrm{diag}(\mathbf{P}_i)$. From the distribution $q(\mathbf{v}_i(k))$, the expectation $\mathbb{E}[\mathbf{v}_i(k)|\mathbf{y}_i,\mathbf{x}_i]$ and variance $\mathrm{Var}[\mathbf{v}_i(k)|\mathbf{y}_i,\mathbf{x}_i]$ using the truncated normal properties. With the fact $\mathbf{v}_i = [\mathbf{h}_i^T, \mathbf{s}_i^T]^T$, the expectations $\mathbb{E}[\mathbf{H}|\mathbf{Y},\mathbf{X}]$, $\mathbb{E}[\mathbf{H}\mathbf{H}|\mathbf{Y},\mathbf{X}]$ and $\mathbb{E}[\mathbf{S}|\mathbf{Y},\mathbf{X}]$ required in the gradient computation can be obtained directly. For $\mathbf{s}_i \triangleq \mathbf{T}_i\mathbf{z}_i$, we have $\mathbb{E}\left[\mathbf{z}_i|\mathbf{y}_i,\mathbf{x}_i\right] = \mathbf{T}_i^{-1}\mathbb{E}\left[\mathbf{s}_i|\mathbf{y}_i,\mathbf{x}_i\right]$, and thus $\mathbb{E}[\mathbf{Z}|\mathbf{Y},\mathbf{X}]$ can be computed easily.

2. Training Deep TGGM

The training algorithms for deep regression and classification TGGMs are almost the same, thus we only present that for classification only. With similar transformation in single layer model, we can represent deep classification TGGM as $p(\mathbf{Y},\mathbf{S},\mathbf{H}|\mathbf{X};\mathbf{\Theta}) = \prod_{i=1}^N p\left(\mathbf{h}_i\left|\mathbf{x}_i\right.\right) \times$

$\mathcal{N}(\mathbf{s}_i|\mathbf{T}_i(\mathbf{W}_2\mathbf{h}_i+\mathbf{b}_2), \mathbf{T}_i\mathbf{T}_i^T) \times \prod_{k\neq y_i} I(\mathbf{s}_i(k) \geq 0)$, where $p(\mathbf{h}_i|\mathbf{x}_i)$ is truncated normal distribution defined as $p(\mathbf{h}_i|\mathbf{x}_i) \triangleq \frac{1}{Z_i}\exp\{-\frac{\|\mathbf{h}_i^{(1)}-\mathbf{W}_0\mathbf{x}_i-\mathbf{b}_0\|^2}{2\sigma_0^2} - \frac{\|\mathbf{h}_i^{(2)}-\mathbf{W}_1\mathbf{h}_i^{(1)}-\mathbf{b}_1\|^2}{2\sigma_0^2}\} I\{\mathbf{h}_i \geq \mathbf{0}\}$ with $\mathbf{h}_i \triangleq [\mathbf{h}_i^{(1)T}, \mathbf{h}_i^{(2)T}]^T$ and $\mathbf{H} \triangleq [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_N]$. The lengths of $\mathbf{h}_i^{(1)}$ and $\mathbf{h}_i^{(2)}$ are denoted as M_1 and M_2, respectively. It can be seen that the whole model is very closely related to TGGM and preserves most of the properties of truncated normal, thus can be trained efficiently similar to above models. The derivatives of $\mathcal{Q}$ can be derived as

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{W}_0} = -\frac{1}{\sigma_0^2}\Big(\mathbb{E}[\mathbf{H}^{(1)}|\mathbf{X}] - \mathbb{E}[\mathbf{H}^{(1)}|\mathbf{Y},\mathbf{X}]\Big)\mathbf{X}^T; \quad (10)$$

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{b}_0} = -\frac{1}{\sigma_0^2}\Big(\mathbb{E}[\mathbf{H}^{(1)}|\mathbf{X}] - \mathbb{E}[\mathbf{H}^{(1)}|\mathbf{Y},\mathbf{X}]\Big)\mathbf{1}_N; \quad (11)$$

$$\begin{aligned}\frac{\partial \mathcal{Q}}{\partial \mathbf{W}_1} =& -\frac{1}{\sigma_0^2}\Big(\mathbf{b}_1\mathbf{1}_N^T(\mathbb{E}[\mathbf{H}^{(1)T}|\mathbf{Y},\mathbf{X}] - \mathbb{E}[\mathbf{H}^{(1)T}|\mathbf{X}]) \\ &+ \mathbf{W}_1(\mathbb{E}[\mathbf{H}^{(1)}\mathbf{H}^{(1)T}|\mathbf{Y},\mathbf{X}] - \mathbb{E}[\mathbf{H}^{(1)}\mathbf{H}^{(1)T}|\mathbf{X}]) \\ &- \mathbb{E}[\mathbf{H}^{(2)}\mathbf{H}^{(1)T}|\mathbf{Y},\mathbf{X}] + \mathbb{E}[\mathbf{H}^{(2)}\mathbf{H}^{(1)T}|\mathbf{X}]\Big) \quad (12)\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{Q}}{\partial \mathbf{b}_1} =& -\frac{1}{\sigma_0^2}\Big(\mathbf{W}_1(\mathbb{E}[\mathbf{H}^{(1)}|\mathbf{Y},\mathbf{X}] - \mathbb{E}[\mathbf{H}^{(1)}|\mathbf{X}])\mathbf{1}_N \\ &- (\mathbb{E}[\mathbf{H}^{(2)}|\mathbf{Y},\mathbf{X}] - \mathbb{E}[\mathbf{H}^{(2)}|\mathbf{X}])\mathbf{1}_N\Big); \quad (13)\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{Q}}{\partial \mathbf{W}_2} =& -\mathbf{W}_2\mathbb{E}[\mathbf{H}^{(2)}\mathbf{H}^{(2)T}|\mathbf{Y},\mathbf{X}] + \mathbb{E}[\mathbf{Z}\mathbf{H}^{(2)T}|\mathbf{Y},\mathbf{X}] \\ &- \mathbf{b}_2\mathbf{1}_N^T\mathbb{E}[\mathbf{H}^{(2)T}|\mathbf{Y},\mathbf{X}]; \quad (14)\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{Q}}{\partial \mathbf{b}_2} =& -\mathbf{W}_2\mathbb{E}[\mathbf{H}^{(2)}|\mathbf{Y},\mathbf{X}]\mathbf{1}_N - N\mathbf{b}_2 + \mathbb{E}[\mathbf{Z}|\mathbf{Y},\mathbf{X}]\mathbf{1}_N, \\ & \quad (15)\end{aligned}$$

where $\mathbf{H}^{(\ell)} \triangleq [\mathbf{h}_1^{(\ell)}, \mathbf{h}_2^{(\ell)}, \cdots, \mathbf{h}_N^{(\ell)}]$ for $\ell = 1, 2$. With the gradients, we can update the model parameters $\mathbf{\Theta}$ using appropriate optimization algorithms, such as SGD and its variants.

In deep models, since the prior is also a multivariate truncated normal, it is expensive to compute the prior expectation $\mathbb{E}[\cdot|\mathbf{X}]$ analytically as that in one-layer case. For the efficiency of training, we resort to mean-field VB for the estimation of both prior and posterior expectations $\mathbb{E}[\cdot|\mathbf{X}]$ and $\mathbb{E}[\cdot|\mathbf{Y},\mathbf{X}]$. The prior distribution $p(\mathbf{h}_i|\mathbf{x}_i)$ can be equivalently written as

$$p(\mathbf{h}_i|\mathbf{x}_i) = \mathcal{N}_T\left(\mathbf{h}_i|\mathbf{Q}^{-1}\boldsymbol{\beta}_i, \mathbf{Q}^{-1}\right), \quad (16)$$

where

$$\mathbf{Q} \triangleq \frac{1}{\sigma_0^2}\begin{bmatrix} \mathbf{I}_{M_1} + \mathbf{W}_1^T\mathbf{W}_1 & -\mathbf{W}_1^T \\ -\mathbf{W}_1 & \mathbf{I}_{M_2} \end{bmatrix}, \quad (17)$$

$$\boldsymbol{\beta}_i \triangleq \frac{1}{\sigma_0^2}\begin{bmatrix} \mathbf{W}_0\mathbf{x}_i + \mathbf{b}_0 - \mathbf{W}_1^T\mathbf{b}_1 \\ \mathbf{b}_1 \end{bmatrix}. \quad (18)$$

Suppose a fully factorized distribution $q(\mathbf{h}_i) = \prod_{k=1}^M q(\mathbf{h}_i(k))$ with $M \triangleq M_1 + M_2$. We now minimize the KL-divergence between $q(\mathbf{h}_i)$ and the true posterior $p(\mathbf{h}_i|\mathbf{x}_i)$. It is known that when all $q(\mathbf{h}_s(\ell))$ except $(\ell, s) = (k, i)$ are known, the KL-divergence is minimized if $\ln q(\mathbf{h}_i(k)) = \langle \ln p(\mathbf{h}_i|\mathbf{x}_i)\rangle_{\neq(k,i)} + const.$ By rearranging the terms in $\ln p(\mathbf{h}_i|\mathbf{x}_i)$, it can be easily obtained that $p(\mathbf{h}_i|\mathbf{x}_i) = -\frac{1}{2}\mathbf{Q}(k,k)\mathbf{h}_i^2(k) + (\boldsymbol{\beta}_i(k) - \mathbf{Q}(k,-k)\mathbf{h}_i(-k))\mathbf{h}_i(k) + \ln I(\mathbf{h}_i(k) \geq 0) + C_4$. Thus, we have

$$q(\mathbf{h}_i(k)) = \mathcal{N}_T\left(\mathbf{h}_i(k)\left|\frac{\boldsymbol{\beta}_i(k) - \mathbf{Q}_0(k,:)\langle\mathbf{h}_i\rangle_q}{\mathbf{Q}(k,k)}, \frac{1}{\mathbf{Q}(k,k)}\right.\right), \quad (19)$$

where $\mathbf{Q}_0 \triangleq \mathbf{Q} - \mathrm{diag}(\mathbf{Q})$. From the univariate truncated normal $q(\mathbf{h}_i)$, we can estimate the prior expectation $\mathbb{E}[\mathbf{H}|\mathbf{X}]$ easily.

For posterior expectation $\mathbb{E}[\cdot|\mathbf{Y},\mathbf{X}]$, we also suppose a fully factorized distribution $q(\mathbf{v}_i)$ with $\mathbf{v}_i = [\mathbf{h}_i^T, \mathbf{s}_i^T]^T$ and then minimize the KL-divergence. First, we express the log-likelihood as

$$\begin{aligned}&\ln p(\mathbf{y}_i, \mathbf{v}_i|\mathbf{x}_i) \\ &= -\frac{1}{2}\mathbf{P}_i(k,k)\mathbf{v}_i^2(k) + \sum_{k\neq M+y_i} \ln I(\mathbf{v}_i(k) \geq 0) \\ &\quad + (\boldsymbol{\gamma}_i(k) - \mathbf{P}_i(k,-k)\mathbf{v}_i(-k))\mathbf{v}_i(k) + C_4, \quad (20)\end{aligned}$$

where in deep models $\mathbf{P}_i$ and $\boldsymbol{\gamma}_i$ are defined as

$$\mathbf{P}_i = \frac{1}{\sigma_0^2}\begin{bmatrix} \mathbf{I}_{M_1}+\mathbf{W}_1^T\mathbf{W}_1 & -\mathbf{W}_1^T & \mathbf{0} \\ -\mathbf{W}_1 & \mathbf{I}_{M_2}+\mathbf{W}_2^T\mathbf{W}_2 & -\mathbf{W}_2^T\mathbf{T}_i^{-1} \\ 0 & -\mathbf{T}_i^{-1T}\mathbf{W}_2 & (\mathbf{T}_i\mathbf{T}_i^T)^{-1} \end{bmatrix}; \quad (21)$$

$$\boldsymbol{\gamma}_i = \frac{1}{\sigma_0^2}\begin{bmatrix} \mathbf{W}_0\mathbf{x}_i + \mathbf{b}_0 - \mathbf{W}_1^T\mathbf{b}_1 \\ \mathbf{b}_1 - \mathbf{W}_2^T\mathbf{b}_2 \\ \mathbf{T}_i^{-1T}\mathbf{b}_2 \end{bmatrix}. \quad (22)$$

Then, it can be known that the KL-divergence is minimized with $q(\mathbf{v}_i(k))$ being the same form as (9). The only difference are the expressions of precision matrix $\mathbf{P}_i$ and linear vector $\boldsymbol{\gamma}_i$. With the factorized truncated normal distribution, the posterior expectations $\mathbb{E}[\mathbf{H}|\mathbf{Y},\mathbf{X}]$ and covariance $\mathbb{E}[\mathbf{H}\mathbf{H}^T|\mathbf{Y},\mathbf{X}]$ can be estimated easily using truncated normal properties.

References

Johnson, Norman L, Kotz, Samuel, and Balakrishnan, Narayanaswamy. Continuous univariate distributions, vol. 1-2, 1994.

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.