Supplementary Materials for “Unsupervised Learning with Truncated Gaussian Graphical Models”

Qinliang Su, Xuejun Liao, Chunyuan Li, Zhe Gan and Lawrence Carin
Department of Electrical & Computer Engineering
Duke University
Durham, NC 27708-0291

Gradient Computation

From the expression of energy function, it can be seen that $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial w_{ij}} = x_i h_j$, $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial a_i} = \frac{1}{2}x_i^2$, $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial b_i} = x_i$, $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial c_j} = h_j$ and $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial d_j} = \frac{1}{2}h_j^2$, where w_{ij} is the (i,j)-th element of $\mathbf{W}$; and x_i, h_i, a_i, b_i, c_i and d_i are the i-th element of corresponding vectors. Then, the deritives of log-likelihood can easily be obtained as

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta};\mathbf{x})}{\partial w_{ij}} = -\left(\mathbb{E}\left[x_i h_j\right] - x_i \mathbb{E}\left[h_j|\mathbf{x}\right]\right), \tag{1}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta};\mathbf{x})}{\partial a_i} = \frac{1}{2}\left(\mathbb{E}\left[x_i^2\right] - x_i^2\right), \tag{2}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta};\mathbf{x})}{\partial b_i} = -\left(\mathbb{E}\left[x_i\right] - x_i\right), \tag{3}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta};\mathbf{x})}{\partial c_j} = -\left(\mathbb{E}\left[h_j\right] - \mathbb{E}\left[h_j|\mathbf{x}\right]\right), \tag{4}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta};\mathbf{x})}{\partial d_j} = \frac{1}{2}\Big(\mathbb{E}\left[h_j^2\right] - \mathbb{E}\left[h_j^2|\mathbf{x}\right]\Big), \tag{5}$$

where expectations $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot|\mathbf{x}]$ are taken w.r.t. $p(\mathbf{x},\mathbf{h})$ and $p(\mathbf{h}|\mathbf{x})$, respectively. Due to the difficulties of obtaining $\mathbf{E}[\cdot]$, we resort to contrastive divergence (CD) algorithm to estimate the gradients. To this end, we sample $\mathbf{h}$ given $\mathbf{x}$ from $p(\mathbf{h}|\mathbf{x})$, and then sample $\mathbf{x}$ from the sampled $\hat{\mathbf{h}}$ according to $p(\mathbf{x}|\hat{\mathbf{h}})$. We denote this sample as $\mathbf{x}^{(1)}$. By repeating this process k steps, we obtain a sample denoted as $\mathbf{x}^{(k)}$. Then, we can estimate the derivatives as

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta};\mathbf{x})}{\partial w_{ij}} = -\Big(x_i^{(k)}\mathbb{E}\left[h_j|\mathbf{x}^{(k)}\right] - x_i^{(0)}\mathbb{E}\left[h_j|\mathbf{x}^{(0)}\right]\Big), \tag{6}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial a_i} = \frac{1}{2}\left(x_i^{(k)2} - x_i^{(0)2}\right), \tag{7}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta};\mathbf{x})}{\partial b_i} = -\left(x_i^{(k)} - x_i^{(0)}\right), \tag{8}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta};\mathbf{x})}{\partial c_j} = -\left(\mathbb{E}\left[h_j|\mathbf{x}^{(k)}\right] - \mathbb{E}\left[h_j|\mathbf{x}^{(0)}\right]\right), \tag{9}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta};\mathbf{x})}{\partial d_j} = \frac{1}{2}\Big(\mathbb{E}\left[h_j^2|\mathbf{x}^{(k)}\right] - \mathbb{E}\left[h_j^2|\mathbf{x}^{(0)}\right]\Big), \tag{10}$$

where $\mathbf{x}^{(0)} \triangleq \mathbf{x}$.

Derivation of $p^*(\mathbf{x};\boldsymbol{\Theta})$

From the joint distribution $p(\mathbf{x},\mathbf{h};\boldsymbol{\Theta})$, we have

$$\begin{aligned} p^*&(\mathbf{x};\boldsymbol{\Theta}) \\ &= \frac{1}{Z}e^{\mathbf{b}^T\mathbf{x}}\prod_{j=1}^{m}\int_0^{+\infty} e^{-\frac{1}{2}\left(d_j h_j^2 - 2\left[\mathbf{W}^T\mathbf{x}+\mathbf{c}\right]_j h_j\right)} dh_j \\ &= \frac{1}{Z}e^{\mathbf{b}^T\mathbf{x}}\prod_{j=1}^{m} e^{\frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j^2}{2d_j}}\int_0^{+\infty} e^{-\frac{d_j}{2}\left(h_j - \frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j}{d_j}\right)^2} dh_j. \end{aligned} \tag{11}$$

After some simple manipulations, we obtain

$$\begin{aligned} p^*(\mathbf{x};\boldsymbol{\Theta}) &= \frac{1}{Z}e^{\mathbf{b}^T\mathbf{x}}\prod_{j=1}^{m} e^{\frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j^2}{2d_j}}\left(\frac{2\pi}{d_j}\right)^{\frac{1}{2}}\Phi\left(\frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j}{\sqrt{d_j}}\right) \\ &= \frac{1}{Z}e^{\mathbf{b}^T\mathbf{x}}\prod_{j=1}^{m}\frac{1}{\sqrt{d_j}}\frac{\Phi\left(\frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j}{\sqrt{d_j}}\right)}{\phi\left(\frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j}{\sqrt{d_j}}\right)}, \end{aligned} \tag{12}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ means the PDF and CDF of standard normmal distribution.

Partition Function Estimation with Count Data

Without loss of generality, the count model is described in the context of bag-of-words topic model. Following (**?**), we use a matrix $\mathbf{X}$ to describe a document, in whch each column is a column hot-vector, with the non-zero element representing the corresponding word in vocabulary appearing once. Now, we have the joint pdf describing bag-of-words document $\mathbf{X}$ as

$$\begin{aligned} p(\mathbf{X},\mathbf{h};\boldsymbol{\Theta}) = \frac{1}{Z}&e^{-\frac{1}{2}\left(\|\mathbf{D}^{\frac{1}{2}}\mathbf{h}\|^2 - 2\hat{\mathbf{x}}^T\mathbf{W}\mathbf{h} - 2\mathbf{b}^T\hat{\mathbf{x}} - 2K\mathbf{c}^T\mathbf{h}\right)} \\ &I(\mathbf{x}_k \in \mathcal{I})I(\mathbf{h} \geq \mathbf{0}), \end{aligned} \tag{13}$$

where

$$\hat{\mathbf{x}} \triangleq \sum_{k=1}^{K}\mathbf{x}_k \tag{14}$$

with $\mathbf{x}_k$ being the k-th column of $\mathbf{X}$; and K means the number of words in the document; $\mathcal{I}$ is the set of one-hot vectors. Similar to operations in (**??**) and (**??**), we obtain

$$p(\mathbf{X};\mathbf{\Theta}) = \frac{1}{Z} e^{\mathbf{b}^T\hat{\mathbf{x}}} \times \prod_{j=1}^{m} \frac{1}{\sqrt{d_j}} \frac{\Phi\left(\frac{[\mathbf{W}^T\hat{\mathbf{x}}+K\mathbf{c}]_j}{\sqrt{d_j}}\right)}{\phi\left(\frac{[\mathbf{W}^T\hat{\mathbf{x}}+K\mathbf{c}]_j}{\sqrt{d_j}}\right)}. \quad (15)$$

To estimate the partition function Z, define $p_A(\mathbf{X},\mathbf{h}^A;\mathbf{\Theta}) = \frac{1}{Z_A} e^{-\frac{1}{2}||\mathbf{D}^{\frac{1}{2}}\mathbf{h}^A||^2} I(\mathbf{x}_k \in \mathcal{I}) I(\mathbf{h}^A \geq 0)$ and $p_B(\mathbf{X},\mathbf{h}^B;\mathbf{\Theta}) = p(\mathbf{X},\mathbf{h}^B;\mathbf{\Theta})$. It is known from (**?**) that the partition function Z can be estimated as

$$Z \approx \frac{\sum_{i=1}^{M} w^{(i)}}{M} Z_A; \quad (16)$$

where Z_A can be easily computed as

$$Z_A = \prod_{j=1}^{m} \frac{1}{\sqrt{d_j}} \frac{\Phi(0)}{\phi(0)}, \quad (17)$$

and coefficient $w^{(i)}$ is constructed from a Markov chain that is simulated to gradually transit from $p_A(\mathbf{x},\mathbf{h}^A)$ to $p_B(\mathbf{x},\mathbf{h}^B)$, with the transition realized via a sequence of intermediate distributions

$$\begin{aligned} &p_s(\mathbf{X},\mathbf{h}^A,\mathbf{h}^B;\mathbf{\Theta}) \\ &= \frac{1}{Z_s}\exp\Big\{ -\frac{1}{2}(1-\beta_s)||\mathbf{D}^{\frac{1}{2}}\mathbf{h}^A||^2 - \frac{1}{2}\beta_s||\mathbf{D}^{\frac{1}{2}}\mathbf{h}^B||^2 \\ &\quad + \beta_s\hat{\mathbf{x}}^T\mathbf{W}\mathbf{h}^B + \beta_s\mathbf{b}^T\hat{\mathbf{x}} + \beta_s K\mathbf{c}^T\mathbf{h}^B\Big\} \\ &\quad \times I(\mathbf{x}_k \in \mathcal{I}) I(\mathbf{h}^A,\mathbf{h}^B \geq \mathbf{0}) \end{aligned} \quad (18)$$

for $\beta_s \in [0,1]$. Specifically, $w^{(i)}$ is computed as

$$w^{(i)} = \frac{p_1^*({\mathbf{x}_i}^{(0)})}{p_0^*({\mathbf{x}_i}^{(0)})} \frac{p_2^*({\mathbf{x}_i}^{(1)})}{p_1^*({\mathbf{x}_i}^{(1)})} \cdots \frac{p_K^*({\mathbf{x}_i}^{(K-1)})}{p_{K-1}^*({\mathbf{x}_i}^{(K-1)})}, \quad (19)$$

where $p_s^*(\mathbf{X};\mathbf{\Theta})$ is the unnormalized term in $p_s(\mathbf{X};\mathbf{\Theta}) = \frac{1}{Z_s} p_s^*(\mathbf{X};\mathbf{\Theta}) I(\mathbf{x}_k \in \mathcal{I})$. To compute $w^{(i)}$, we integrate out $\mathbf{h}^A$ and $\mathbf{h}^B$ in $p_s(\mathbf{X},\mathbf{h}^A,\mathbf{h}^B;\mathbf{\Theta})$ and obtain

$$\begin{aligned} \log p_s^*(\mathbf{X};\mathbf{\Theta}) = &-\frac{1}{2}\sum_{j=1}^{m}\log((1-\beta_s)d_j) - m\log\frac{\phi(0)}{\Phi(0)} \\ &+ \beta_s\mathbf{b}^T\hat{\mathbf{x}} - \frac{1}{2}\sum_{j=1}^{m}\log(\beta_s d_j) \\ &- \sum_{j=1}^{m}\log\frac{\phi\left(\frac{\sqrt{\beta_s}[\mathbf{W}^T\hat{\mathbf{x}}+K\mathbf{c}]_j}{\sqrt{d_j}}\right)}{\Phi\left(\frac{\sqrt{\beta_s}[\mathbf{W}^T\hat{\mathbf{x}}+K\mathbf{c}]_j}{\sqrt{d_j}}\right)}. \end{aligned} \quad (20)$$

In (**??**), the sequence $\mathbf{x}^{(k)}$ is simulated from a Markov chain $(\mathbf{x}_i^{(0)}, \mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(K)})$ as $\mathbf{x}_i^{(0)} \sim p_0(\mathbf{x}_i,\mathbf{h}^A,\mathbf{h}^B)$, $(\mathbf{h}^A,\mathbf{h}^B) \sim p_1(\mathbf{h}^A,\mathbf{h}^B|{\mathbf{x}_i}^{(0)})$, $\mathbf{x}_i^{(1)} \sim p_1(\mathbf{x}_i|\mathbf{h}^A,\mathbf{h}^B)$, $\cdots$, $(\mathbf{h}^A,\mathbf{h}^B) \sim p_K(\mathbf{h}^A,\mathbf{h}^B|{\mathbf{x}_i}^{(K-1)})$ and $\mathbf{x}_i^{(K)} \sim p_K(\mathbf{x}_i|\mathbf{h}^A,\mathbf{h}^B)$, with the conditional pdfs equal to

$$p(\mathbf{h}^B|\mathbf{x};\mathbf{\Theta}) = \mathcal{N}_T\left(\mathbf{h}^B \middle| \mathbf{D}^{-1}(\mathbf{W}^T\mathbf{x}+K\mathbf{c}), \frac{1}{\beta_s}\mathbf{D}^{-1}\right), \quad (21)$$

$$p(\mathbf{x}|\mathbf{h}^B;\mathbf{\Theta}) = Multinomial\left(\mathbf{x};K,\boldsymbol{\xi}\right), \quad (22)$$

where $\xi_i \triangleq \frac{\exp\{\beta_s[\mathbf{W}\mathbf{h}^B+\mathbf{b}]_i\}}{\sum_{i=1}^{N}\exp\{\beta_s[\mathbf{W}\mathbf{h}^B+\mathbf{b}]_i\}}$ is the success probability of the i-th word; and K is the number of words in the document. Moreover, we can also derive that

Missing Data Prediction

Suppose the data $\mathbf{x}$ is composed of the observed part $\mathbf{x}_o$ and the unobserved part $\mathbf{x}_u$. With the help of already trained RTGGM model, we will recover the missing part $\mathbf{x}_u$ from the observed data $\mathbf{x}_o$. Obviously, the conditional pdf $p(\mathbf{x}_u,\mathbf{h}|\mathbf{x}_o)$ constitutes a new RTGGM, with its energy function expressed as

$$\begin{aligned} E(\mathbf{x}_u,\mathbf{h}) = \frac{1}{2}\Big(&||\mathbf{A}_u^{\frac{1}{2}}\mathbf{x}_u||^2 + ||\mathbf{D}^{\frac{1}{2}}\mathbf{h}||^2 - 2\mathbf{x}_u^T\mathbf{W}_u\mathbf{h} \\ &- 2\mathbf{b}_u^T\mathbf{x}_u - 2\tilde{\mathbf{c}}^T\mathbf{h}\Big) \end{aligned} \quad (23)$$

where $\tilde{\mathbf{c}} \triangleq \mathbf{c} + \mathbf{W}_o^T\mathbf{x}_o$ and $\mathbf{A}_u \triangleq \text{diag}(\mathbf{a}_u)$; and $\mathbf{a}_u$, $\mathbf{W}_u$ and $\mathbf{b}_u$ are composed of partial rows of $\mathbf{a}$, $\mathbf{W}$ and $\mathbf{b}_u$, respectively, with the selected row indexes corresponding to those of $\mathbf{x}_u$. Hence, under the Gaussian output case, we have the pdf $p(\mathbf{x}_u,\mathbf{h}|\mathbf{x}_o)$ as

$$p(\mathbf{x}_u,\mathbf{h}|\mathbf{x}_o) \propto e^{-E(\mathbf{x}_u,\mathbf{h})}\mathbb{I}(\mathbf{h} \geq \mathbf{0}). \quad (24)$$

The conditional pdfs can be further derived as

$$p(\mathbf{x}_u|\mathbf{h},\mathbf{x}_o) = \mathcal{N}\left(\mathbf{x}_u|\mathbf{A}_u^{-1}(\mathbf{W}_u\mathbf{h}+\mathbf{b}_u),\mathbf{A}_u^{-1}\right), \quad (25)$$

$$p(\mathbf{h}|\mathbf{x}_u,\mathbf{x}_o) = \mathcal{N}_T\left(\mathbf{h}|\mathbf{D}^{-1}(\mathbf{W}_u^T\mathbf{x}_u+\tilde{\mathbf{c}}),\mathbf{D}^{-1}\right). \quad (26)$$

By resorting to Gibbs sampling, the expectation $\mathbb{E}[\mathbf{x}_u|\mathbf{x}_o]$ can be estimated efficiently thanks to the conditional independence among $\mathbf{x}_u$ and $\mathbf{h}$. Notice that here we only considered the model with Gaussian observation, but it can be easily extended to binary observation models.

Restricted Gaussian Graphical Model

When we restrict the GGM having a bipartite structure and binary output, the joint pdf can be represented as

$$p(\mathbf{x},\mathbf{h};\mathbf{\Theta}) \propto e^{-E(\mathbf{x},\mathbf{h})}\mathbb{I}\left(\mathbf{x} \in \{0,1\}^N\right), \quad (27)$$

where

$$E(\mathbf{x},\mathbf{h}) \triangleq \frac{1}{2}\left(\mathbf{h}^T\text{diag}(\mathbf{d})\mathbf{h} - 2\mathbf{x}^T\mathbf{W}\mathbf{h} - 2\mathbf{b}^T\mathbf{x} - 2\mathbf{c}^T\mathbf{h}\right). \quad (28)$$

The model can be trained efficiently using contrastive divergence (CD), with the conditional pdfs given by

$$p(\mathbf{h}|\mathbf{x}) = \mathcal{N}\left(\text{diag}^{-1}(\mathbf{d})\left(\mathbf{W}^T\mathbf{x}+\mathbf{c}\right), \text{diag}^{-1}(\mathbf{d})\right), \quad (29)$$

$$p(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{N}\frac{[\mathbf{W}\mathbf{h}+\mathbf{b}]_i\, x_i}{1+\exp\{[\mathbf{W}\mathbf{h}+\mathbf{b}]_i\}}. \quad (30)$$

After training the model, we use AIS to evaluate its performance. The derivation process is almost the same as that in RTGGM models except the marginal pdf has the form

$$\begin{aligned}
p(\mathbf{x};\boldsymbol{\Theta}) &= \frac{1}{Z}e^{\mathbf{b}^T\mathbf{x}}\prod_{j=1}^{m}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(d_j h_j^2 - 2\left[\mathbf{W}^T\mathbf{x}+\mathbf{c}\right]_j h_j\right)} dh_j \\
&= \frac{1}{Z}e^{\mathbf{b}^T\mathbf{x}}\prod_{j=1}^{m} e^{\frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j^2}{2d_j}} \int_{-\infty}^{+\infty} e^{-\frac{d_j}{2}\left(h_j - \frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j}{d_j}\right)^2} dh_j \\
&= \frac{1}{Z}e^{\mathbf{b}^T\mathbf{x}}\prod_{j=1}^{m} e^{\frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j^2}{2d_j}} \left(\frac{2\pi}{d_j}\right)^{\frac{1}{2}} \\
&= \frac{1}{Z}e^{\mathbf{b}^T\mathbf{x}}\prod_{j=1}^{m} \frac{1}{\sqrt{d_j}\,\phi\left(\frac{[\mathbf{W}^T\mathbf{x}+\mathbf{c}]_j}{\sqrt{d_j}}\right)}. \qquad (31)
\end{aligned}$$

Generation with Two-layer RTGGM

Yale Face data set is considered, which contains faces of 15 persons, with each person having 11 images of size 32×32 under different expressions and lighting. Two-hidden-layer RTGGM (10-100) is used, with its bottom layer set to be Gaussian. We train the model layer by layer using the method developed in the paper. The learning rate is set to be 10^{-3} . After training, samples were drawn from this deep models using Gibbs sampling. Specifically, we first use Gibbs sampler to draw samples from the distribution constituted by the first two top layers. Then, we pass down the samples through the hierarchical graphical model, and treat the samples at the bottom layer as the generated images. From Figure **??**, we can see that the generated faces looks like true face images.

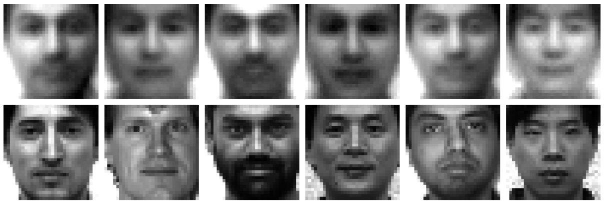

Figure 1: (Top) Faces drawn from the deep RTGGM with two-layer hidden structure 10-100. (Bottom) The corresponding most similar faces in the Yale Face data set.