Note on Disentanglement

Anonymous Author(s)
Affiliation
Address
email

Abstract

1 Problem formulation

Assume we have data points $\{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x}_i \in R^d$. $\mathbf{W}$ is the transformation matrix (weight matrix) such that $\mathbf{W} \in R^{k \times d}$ (k is the number of hidden units). Our goal is to put constraints on the hidden layer such that the covariance of hidden units has block-diagonal structure. The overall objective function can be written as:

$$L(\mathbf{X}, \mathbf{\Theta}) + \lambda_1 ||\mathbf{W}\mathbf{S}\mathbf{W}^\top||_1 + \lambda_2 ||\mathbf{W}\mathbf{S}\mathbf{W}^\top||_* \tag{1}$$

where $\mathbf{\Theta}$ contains all the parameters of the model (i.e weights, bias terms for RBM etc), $\mathbf{W}$ is the weight matrix(or transformation). S is the sample covariance matrix of the input computed as $\mathbf{S} = \frac{1}{N-1}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top$. $||.||_1$ is the L_1 norm, and $||.||_*$ is the nuclear norm. λ_1, λ_2 are the hyperparameters of the model. $L(.)$ is the loss function (i.e reconstruction error for autoencoder, negative log likelihood for RBM.etc). The reason that we put L_1 and nuclear norms is to make the covariance of hidden layer block-diagonal, which is most widely used approach. Please note that even we do add sigmoid function on top of each hidden layer, the regularization still holds from the following corollary.(need to verify)

Corollary 1.0.1. *The mutual information between random variables* $\mathbf{h}_i, \mathbf{h}_j$ *satisfies MI*$(\mathbf{h}_i, \mathbf{h}_j) =$ *MI*$(\sigma(\mathbf{h}_i), \sigma(\mathbf{h}_j))$

Proof. This is straightforward since $\sigma(.)$ is invertible function that has 1-1 mappings. □

2 Simple case : Linear Autencoder

For the proof of concept, we first use "linear" (can be easily extended to nonlinear case) autoencoder, then the objective function is:

$$\sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{W}^\top \mathbf{W}\mathbf{x}_i||_2^2 + \lambda_1 ||\mathbf{W}\mathbf{S}\mathbf{W}^\top||_1 + \lambda_2 ||\mathbf{W}\mathbf{S}\mathbf{W}^\top||_* \tag{2}$$

Optimizing this equation is challenging, since it contains both L_1 norm and nuclear norm, although there are bunches of previous work solves these problem separately. We could solve this optimization problems using (1). ADMM as discussed in [2]. (2). as discussed in [1] (need to figure this out).

After solving for $\mathbf{W}$, we can easily compute the covariance matrix first, then do row/column reorder to get the block-diagonal matrix. Some possible ways to get this block-diagonal matrix: (1). use spectral clustering(?). (2). apply stochastic blockmodel(?)..

2.1 Optimization

We learn the model parameters under ADMM framework. We can re-write the objective function using auxiliary variable $\mathbf{Z}_1$ and $\mathbf{Z}_2$:

$$\min_{\mathbf{W},\mathbf{Z}_1,\mathbf{Z}_2} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{W}^\top \mathbf{W}\mathbf{x}_i||_2^2 + \lambda_1 ||\mathbf{Z}_1||_* + \lambda_2 ||\mathbf{Z}_2||_1 \tag{3}$$
$$s.t.\ \mathbf{W}\mathbf{S}\mathbf{W}^\top = \mathbf{Z}_1, \mathbf{W}\mathbf{S}\mathbf{W}^\top = \mathbf{Z}_2$$

After using augmented Lagrangian multiplier, the objective function can be written as:

$$\begin{aligned}\mathbf{L}(\mathbf{W}, \mathbf{Z_1}, \mathbf{Z_2}, \mathbf{U_1}, \mathbf{U_2}) = &\sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{W}^\top \mathbf{W}\mathbf{x}_i||_2^2 + \lambda_1 ||\mathbf{Z}_1||_* + \lambda_2 ||\mathbf{Z}_2||_1 + \text{Tr}(\mathbf{U}_1^\top (\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1)) \\ &+ \text{Tr}(\mathbf{U}_2^\top (\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_2)) + \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1||^2 + \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_2||^2\end{aligned} \tag{4}$$

where ρ is the penalty parameter, $\mathbf{U}_1$ and $\mathbf{U}_2$ are the dual variables, Tr is the trace operator. The update is separated into four steps, where each step updates $\mathbf{W}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}_1, \mathbf{U}_2$ respectively.

2.1.1 Update $\mathbf{Z}_1$

After extracting relevant terms, we have

$$\begin{aligned}L(\mathbf{Z}_1) =&\lambda_1 ||\mathbf{Z}_1||_* + \text{Tr}(\mathbf{U}_1^\top (\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1)) + \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1||^2 \\ =&\lambda_1 ||\mathbf{Z}_1||_* + \frac{\rho}{2}\text{Tr}\left[\frac{2}{\rho}\mathbf{U}_1^\top (\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1) + (\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1)^\top (\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1)\right] \\ =&\lambda_1 ||\mathbf{Z}_1||_* + \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1 + \frac{1}{\rho}\mathbf{U}_1||^2 + \text{const}\end{aligned} \tag{5}$$

Thus, we have

$$\hat{\mathbf{Z}_1} = \text{argmin}_{\mathbf{Z}_1} \lambda_1 ||\mathbf{Z}_1||_* + \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1 + \frac{1}{\rho}\mathbf{U}_1||^2$$

Theorem 2.1. $\hat{\mathbf{Z}} = argmin_{\mathbf{Z}} \lambda ||\mathbf{Z}||_* + \rho ||\mathbf{Y} - \mathbf{Z}||$ *has the closed form solution*

$$\hat{\mathbf{Z}} = S_{\lambda/2\rho}(\mathbf{Y})$$

where $S_\alpha(\mathbf{Y})$ *is a soft-thresholding function defined as* $S_\alpha(\mathbf{Y}) = \mathbf{U}diag((\sigma_i - \alpha)_+)\mathbf{V}^\top$*, for matrix* $\mathbf{Y}$ *via SVD* $Y = \mathbf{U}diag(\sigma_i)\mathbf{V}^\top$

Following this theorem, we can compute $\mathbf{Z}_1$ as follows:

$$\hat{\mathbf{Z}_1} = S_{\alpha_1/\rho}(\mathbf{W}\mathbf{S}\mathbf{W}^\top + \frac{1}{\rho}\mathbf{U}_1)$$

where $\mathbf{W}\mathbf{S}\mathbf{W}^\top + \frac{1}{\rho}\mathbf{U}_1 = U\text{diag}(\sigma_i)\mathbf{V}^\top$

2.1.2 Update $\mathbf{Z}_2$

After extracting relevant terms for $\mathbf{Z}_2$, we have

$$L(\mathbf{Z}_2) = \lambda_2 ||\mathbf{Z}_2||_1 + \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_2 + \frac{1}{\rho}\mathbf{U}_2||^2$$

By taking the derivative, we have

$$\frac{\partial L(\mathbf{Z}_2)}{\partial (\mathbf{Z}_2)_{ij}} = \lambda_2 \text{sign}(\mathbf{Z}_2)_{ij} + \rho(\mathbf{Z}_2)_{ij} - \rho(\mathbf{W}\mathbf{S}\mathbf{W}^\top + \frac{1}{\rho}\mathbf{U}_2)_{ij}$$

Thus, we have the following update equations (derivation is the same as LASSO)

$$\hat{(\mathbf{Z}_2)}_{ij} = \begin{cases} (\mathbf{W}\mathbf{S}\mathbf{W}^\top + \frac{1}{\rho}\mathbf{U}_2)_{ij} - \frac{\lambda_2}{\rho} & (\mathbf{W}\mathbf{S}\mathbf{W}^\top + \frac{1}{\rho}\mathbf{U}_2)_{ij} \geq \frac{\lambda_2}{\rho} \\ (\mathbf{W}\mathbf{S}\mathbf{W}^\top + \frac{1}{\rho}\mathbf{U}_2)_{ij} + \frac{\lambda_2}{\rho} & (\mathbf{W}\mathbf{S}\mathbf{W}^\top + \frac{1}{\rho}\mathbf{U}_2)_{ij} \leq -\frac{\lambda_2}{\rho} \\ 0 \quad \text{otherwise} \end{cases}$$

2.1.3 Update $\mathbf{U}_1, \mathbf{U}_2$

$\mathbf{U}_1, \mathbf{U}_2$ are both dual variables, so the updates are simple:

$$\hat{\mathbf{U}_1} = \mathbf{U}_1 + (\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1)$$
$$\hat{\mathbf{U}_2} = \mathbf{U}_2 + (\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_2)$$

2.1.4 Update $\mathbf{W}$

After extracting relevant terms, we have

$$\begin{aligned} L(\mathbf{W}) =& \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{W}^\top\mathbf{W}\mathbf{x}_i||_2^2 + \text{Tr}(\mathbf{U}_1^\top(\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1)) + \text{Tr}(\mathbf{U}_2^\top(\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_2)) \\ &+ \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1||^2 + \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_2||^2 \end{aligned} \tag{6}$$

It is equivalent to minimize

$$L(\mathbf{W}) = \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{W}^\top\mathbf{W}\mathbf{x}_i||_2^2 + \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1 + \frac{1}{\rho}\mathbf{U}_1||^2 + \frac{\rho}{2}||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_2 + \frac{1}{\rho}\mathbf{U}_2||^2$$

Assume $\Delta_1 = ||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1 + \frac{1}{\rho}\mathbf{U}_1||^2$

$$\begin{aligned} \Delta_1 =& \text{Tr}((\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1 + \frac{1}{\rho}\mathbf{U}_1)^\top(\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_1 + \frac{1}{\rho}\mathbf{U}_1)) \\ =& \text{Tr}((\mathbf{W}\mathbf{S}\mathbf{W}^\top)^\top(\mathbf{W}\mathbf{S}\mathbf{W}^\top)) + \text{Tr}((\frac{1}{\rho}\mathbf{U}_1 - \mathbf{Z}_1)^\top\mathbf{W}\mathbf{S}\mathbf{W}^\top) + \text{Tr}((\mathbf{W}\mathbf{S}\mathbf{W}^\top)^\top(\frac{1}{\rho}\mathbf{U}_1 - \mathbf{Z}_1)) + \text{const} \\ =& \text{Tr}(\mathbf{W}^\top\mathbf{W}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S}) + 2\text{Tr}(\mathbf{W}\mathbf{S}\mathbf{W}^\top(\frac{1}{\rho}\mathbf{U}_1 - \mathbf{Z}_1)) + \text{const} \end{aligned} \tag{7}$$

Thus,

$$\frac{\partial \Delta_1}{\partial \mathbf{W}} = 4\mathbf{W}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S} + 2(\frac{1}{\rho}\mathbf{U}_1 - \mathbf{Z}_1)^\top\mathbf{W}\mathbf{S} + 2(\frac{1}{\rho}\mathbf{U}_1 - \mathbf{Z}_1)\mathbf{W}\mathbf{S} \tag{8}$$

Similarly, for $\Delta_2 = ||\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{Z}_2 + \frac{1}{\rho}\mathbf{U}_2||^2$ we have

$$\frac{\partial \Delta_2}{\partial \mathbf{W}} = 4\mathbf{W}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S} + 2(\frac{1}{\rho}\mathbf{U}_2 - \mathbf{Z}_2)^\top\mathbf{W}\mathbf{S} + 2(\frac{1}{\rho}\mathbf{U}_2 - \mathbf{Z}_2)\mathbf{W}\mathbf{S} \tag{9}$$

For $\Delta_3 = \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{W}^\top\mathbf{W}\mathbf{x}_i||_2^2$, we have

$$\frac{\partial \Delta_3}{\partial \mathbf{W}} = \sum_{i=1}^{N} 4\mathbf{W}\mathbf{W}^\top\mathbf{W}\mathbf{x}_i\mathbf{x}_i^\top - 4\mathbf{W}\mathbf{x}_i\mathbf{x}_i^\top \tag{10}$$

If we assume x_i is centered around zero mean, then $N\mathbf{S} = \sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top$, so we have

$$\frac{\partial \Delta_3}{\partial \mathbf{W}} = 4N\mathbf{W}\mathbf{W}^\top\mathbf{W}\mathbf{S} - 4N\mathbf{W}\mathbf{S} \tag{11}$$

Finally, we can compute the derivative $L(\mathbf{W})$ respect to $\mathbf{W}$ as:

$$\begin{aligned}\frac{\partial L(\mathbf{W})}{\partial \mathbf{W}} =& 4\mathbf{W}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S} + 2(\frac{1}{\rho}\mathbf{U}_1 - \mathbf{Z}_1)^\top\mathbf{W}\mathbf{S} + 2(\frac{1}{\rho}\mathbf{U}_1 - \mathbf{Z}_1)\mathbf{W}\mathbf{S} + 4\mathbf{W}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S} \\ &+ 2(\frac{1}{\rho}\mathbf{U}_2 - \mathbf{Z}_2)^\top\mathbf{W}\mathbf{S} + 2(\frac{1}{\rho}\mathbf{U}_2 - \mathbf{Z}_2)\mathbf{W}\mathbf{S} + 4N\mathbf{W}\mathbf{W}^\top\mathbf{W}\mathbf{S} - 4N\mathbf{W}\mathbf{S} \qquad (12) \\ =& (4\mathbf{W}\mathbf{S}\mathbf{W}^\top + 2(\frac{1}{\rho}\mathbf{U}_1 - \mathbf{Z}_1)^\top + 2(\frac{1}{\rho}\mathbf{U}_1 - \mathbf{Z}_1) + 4\mathbf{W}\mathbf{S}\mathbf{W}^\top + 2(\frac{1}{\rho}\mathbf{U}_2 - \mathbf{Z}_2)^\top \\ &+ 2(\frac{1}{\rho}\mathbf{U}_2 - \mathbf{Z}_2) + 4N\mathbf{W}\mathbf{W}^\top - 4N)\mathbf{W}\mathbf{S} \qquad (13)\end{aligned}$$

References

[1] Richard, Emile, Pierre-Andr Savalle, and Nicolas Vayatis. "Estimation of simultaneously sparse and low rank matrices." arXiv preprint arXiv:1206.6474 (2012).

[2] Zhou, Ke, Hongyuan Zha, and Le Song. "Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes." Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics. 2013.

[3] Avron, Haim, et al. "Efficient and practical stochastic subgradient descent for nuclear norm regularization." arXiv preprint arXiv:1206.6384 (2012).

[4] Feng, Jiashi, et al. "Robust Subspace Segmentation with Block-Diagonal Prior." Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014.

[5] Cands, Emmanuel J., et al. "Robust principal component analysis?." Journal of the ACM (JACM) 58.3 (2011): 11.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.