Generalized PCA Summary

Alex Katopodis

July 2023

1 Motivation

Consider a data matrix $X \in \mathbb{R}^{n \times d}$, where the i^{th} row corresponds to the i^{th} observation, and the j^{th} column corresponds to the j^{th} feature. We are interested in computing the deviance between the original data X and Θ, the new matrix of natural parameters after applying generalized PCA. It is possible that features in X do not follow a Gaussian distribution, and minimizing the MSE is not appropriate for finding the best projection of X onto a lower dimension. We instead would like to minimize the deviance associated with the exponential family distribution that each column follows. For example, if the data are Bernoulli distributed, minimizing the Bernoulli deviance is more appropriate than minimizing the MSE.

2 Bernoulli Example

Consider $X_{ij} \sim \text{Bernoulli}(p_{ij})$. We are interested in finding the projection matrix by minimizing Bernoulli deviance. The natural parameter for X_{ij} is $\theta_{ij} = \text{logit}(p_{ij})$. The saturated model occurs when $p_{ij} = X_{ij}$, meaning

$$\hat{\Theta}_{ij} = \begin{cases} -\infty & \text{if } p_{ij} = 0 \\ \infty & \text{if } p_{ij} = 1 \end{cases}$$

Let $Q = 2X - \mathbf{1}_n \mathbf{1}_d^T$ be the matrix defining X with the values $\{-1, 1\}$ in lieu of $\{0, 1\}$. This gives values of 0 a negative natural parameter. We can now represent the saturated natural parameter model $\hat{\Theta} \approx mQ$ with $\sigma(\hat{\Theta}) \approx X$ for a sufficiently large $m \in \mathbb{R}$. Now, we can define $\Theta = \mathbf{1}_n \boldsymbol{\mu} + (\hat{\Theta} - \mathbf{1}_n \boldsymbol{\mu}) U U^T$, where $\boldsymbol{\mu}$ is a row vector with $\mu_i = \text{mean}(\hat{\Theta}_i)$ for the i^{th} column of $\hat{\Theta}$.

The goal of Logistic PCA is to find the projection matrix U that minimizes the Bernoulli Deviance $D(X, \Theta)$, subject to $U^T U = I_k$, where k is the dimension we are reducing to.

$$D(X; \Theta) = 2(\ln[p(X \mid \hat{\Theta})] - \ln[p(X \mid \Theta)])$$

$\hat{\Theta}$ is fixed by the parameter m, and thus the value that minimizes the deviance $D(X;\Theta)$ is the value of Θ that maximizes the log likelihood $\ln[p(X \mid \Theta)]$.

$$\begin{aligned}
\ln[p(X;\Theta)] &= \sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij}\ln(p_{ij}) + (1-X_{ij})\ln(1-p_{ij}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij}\ln(\sigma(\Theta_{ij})) + (1-X_{ij})\ln(1-\sigma(\Theta_{ij})) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij}\ln(\frac{\sigma(\Theta_{ij})}{1-\sigma(\Theta_{ij})}) + \ln(1-\sigma(\Theta_{ij})) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij}\text{logit}(\sigma(\Theta_{ij})) + \ln(1-\sigma(\Theta_{ij})) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij}\Theta_{ij} + \ln(1-\sigma(\Theta_{ij})) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij}\Theta_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{d}\ln(1-\sigma(\Theta_{ij})) \\
&= \text{trace}(X^T\Theta) + \sum_{i=1}^{n}\sum_{j=1}^{d}\ln(1-\sigma(\Theta_{ij}))
\end{aligned}$$

The trace operator is a trick to compute the sum of the elementwise products. The i^{th} diagonal of $X^T\Theta$ is the dot product of $\mathbf{x}_i$ and $\boldsymbol{\theta}_i$, and thus the trace is the sum of all products.

3 Credit

This is a summary and implementation of the paper: Dimensionality Reduction for Binary Data through the Projection of Natural Parameters (Landgraf and Lee). The original paper can be found at:
`https://doi.org/10.48550/arXiv.1510.06112`