Supplementary Material

Derivation of the First-Order Optimality Conditions (2.2-2.4)

To arrive at (2.2), we first obtain the gradient of the deviance with respect to $\mathbf{U}$ from the steps below.

$$\begin{aligned}\frac{1}{2}\frac{\partial D}{\partial \mathbf{U}} = &- \frac{\partial}{\partial \mathbf{U}} tr\left(\mathbf{X}^T\left(\mathbf{1}_n\boldsymbol{\mu}^T + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^T)\mathbf{U}\mathbf{U}^T\right)\right) \\ &+ \frac{\partial}{\partial \mathbf{U}} \sum_{i,j} b_j\left(\mu_j + \left[\mathbf{U}\mathbf{U}^T(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})\right]_j\right).\end{aligned}$$

By standard matrix derivative rules,

$$\frac{\partial}{\partial \mathbf{U}} tr(\mathbf{X}^T(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^T)\mathbf{U}\mathbf{U}^T) = \left(\mathbf{X}^T(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^T) + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^T)^T\mathbf{X}\right)\mathbf{U}.$$

Letting $\hat{\theta}_{ij} = \mu_j + [\mathbf{U}\mathbf{U}^T(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j$, note that each element in the second term of the gradient with b_j is given by

$$\frac{\partial}{\partial u_{kl}} \sum_{i,j} b_j(\hat{\theta}_{ij}) = \sum_{i,j} b_j'(\hat{\theta}_{ij}) \frac{\partial \hat{\theta}_{ij}}{\partial u_{kl}}.$$

Since

$$\frac{\partial[\mathbf{U}\mathbf{U}^T(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j}{\partial u_{kl}} = \begin{cases} (\tilde{\theta}_{ik} - \mu_k)u_{jl} & \text{if } k \neq j \\ (\tilde{\theta}_{ik} - \mu_k)u_{jl} + (\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})^T U_l & \text{if } k = j, \end{cases}$$

we have

$$\begin{aligned}\frac{\partial}{\partial u_{kl}} \sum_{i,j} b_j(\hat{\theta}_{ij}) &= \sum_{i,j} b_j'(\hat{\theta}_{ij})(\tilde{\theta}_{ik} - \mu_k)u_{jl} + \sum_i b_k'(\hat{\theta}_{ik})(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})^T U_l \\ &= (\tilde{\Theta}_k - \mathbf{1}_n\mu_k)^T b'(\hat{\boldsymbol{\Theta}})U_l + b_k'(\hat{\Theta}_k)^T(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^T)U_l,\end{aligned}$$

where $b'(\hat{\boldsymbol{\Theta}})$ is a matrix with its ijth element $b_j'(\hat{\theta}_{ij})$ and U_l is the lth column of $\mathbf{U}$. In matrix notation,

$$\frac{\partial}{\partial \mathbf{U}} \sum_{i,j} b_j\left(\mu_j + \left[\mathbf{U}\mathbf{U}^T(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})\right]_j\right) = \left(b'(\hat{\boldsymbol{\Theta}})^T(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^T) + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^T)^T b'(\hat{\boldsymbol{\Theta}})\right)\mathbf{U},$$

and the result in (2.2) follows.

The gradient of the deviance with respect to $\boldsymbol{\mu}$ in (2.3) is derived as follows.

$$\begin{aligned}\frac{1}{2}\frac{\partial D}{\partial \boldsymbol{\mu}} = &- \frac{\partial}{\partial \boldsymbol{\mu}} tr\left(\mathbf{X}^T\mathbf{1}_n\boldsymbol{\mu}^T\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^T\right)\right) \\ &+ \frac{\partial}{\partial \boldsymbol{\mu}}\sum_{i,j} b_j\left(\mu_j + \left[\mathbf{U}\mathbf{U}^T(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})\right]_j\right).\end{aligned}$$

Using standard vector differentiation,

$$\frac{\partial}{\partial \boldsymbol{\mu}} tr\left(\mathbf{X}^T\mathbf{1}_n\boldsymbol{\mu}^T\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^T\right)\right) = (\mathbf{I}_d - \mathbf{U}\mathbf{U}^T)\mathbf{X}^T\mathbf{1}_n$$

and

$$\frac{\partial}{\partial \boldsymbol{\mu}}\sum_{i,j} b_j\left(\mu_j + \left[\mathbf{U}\mathbf{U}^T(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})\right]_j\right) = \sum_{i,j} b_j'(\hat{\theta}_{ij})\left(\mathbf{e}_j - \mathbf{u}_{j\cdot}^T\mathbf{U}^T\right) = (\mathbf{I}_d - \mathbf{U}\mathbf{U}^T)b'(\hat{\boldsymbol{\Theta}})^T\mathbf{1}_n,$$

where $\mathbf{e}_j$ is a length d standard basis vector with 1 in the jth position and $\mathbf{u}_{j\cdot}$ is the jth row of $\mathbf{U}$.

Limiting Behavior of the Deviance of the Poisson Distribution with the Deviation Held Constant

Lemma 1. *Let x be an observation from a Poisson distribution with mean parameter λ. If the deviation $\Delta := (x - \lambda)$ is held constant, the deviance $D(x;\lambda)$ monotonically decreases to zero as either λ or x increases to ∞. That is, for fixed Δ*

i. $D(\Delta + \lambda; \lambda)$ decreases in λ, and $\lim_{\lambda\to\infty} D(\Delta + \lambda; \lambda) = 0$, and

ii. $D(x; x - \Delta)$ decreases in x, and $\lim_{x\to\infty} D(x; x - \Delta) = 0$.

Proof. The Poisson deviance is given by $D(x;\lambda) = 2\{x\log(x/\lambda) - (x - \lambda)\}$. When $(x - \lambda)$ is fixed at Δ, the deviance is proportional to $h(\lambda|\Delta) := (\Delta + \lambda)\log\left[(\Delta + \lambda)/\lambda\right] - \Delta$. Due to L'Hôpital's rule,

$$\lim_{\lambda\to\infty}(\Delta + \lambda)\log\left[(\Delta + \lambda)/\lambda\right] = \lim_{\lambda\to\infty}\frac{\log\left[(\Delta + \lambda)/\lambda\right]}{1/\lambda} = \lim_{\lambda\to\infty}\frac{\Delta\lambda^2}{(\Delta + \lambda)\lambda} = \Delta,$$

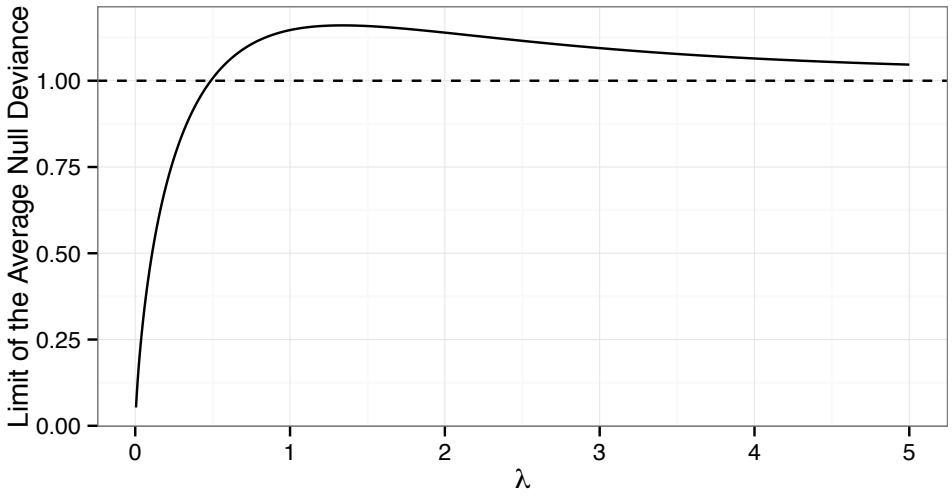

Figure 9: Numerical estimates of the limit of the average null deviance as a function of the mean parameter for Poisson(λ)

which proves that $\lim_{\lambda\to\infty} D(\Delta+\lambda;\lambda) = 0$. The derivative of $h(\lambda|\Delta)$ with respect to λ equals $\log\left[(\Delta+\lambda)/\lambda\right] - \Delta/\lambda = \log\left(\Delta/\lambda + 1\right) - \Delta/\lambda$, which is non-positive for all λ ($\geq -\Delta$) and Δ because $e^z \geq 1+z$ for all z. Hence, the deviance decreases monotonically in λ. The second statement with respect to x can be proved similarly. □

Normalization Factor for the Poisson Distribution

For Poisson variables, $\tau_j = -2\left(\bar{X}_j \log(\bar{X}_j) - \frac{1}{n}\sum_i x_{ij}\log(x_{ij})\right)$, which is not as familiar as it was for Gaussian or Bernoulli variables. To see the relation between the normalization factor and the Poisson mean parameter λ, we obtained numerical estimates of the limit of the average null deviance as n goes to ∞ for the Poisson distribution, which equals $-2\left[\lambda\log(\lambda) - E(X\log(X))\right]$. Figure 9 shows this limit as a function of λ. When the sample size is large, τ_j is expected to be smallest when λ tends towards 0 and to flatten out when λ is greater than 1. This implies that variables with means smaller than 1 get the most weight and that two variables with means much larger than 1 will get approximately the same weight. In

contrast to the Bernoulli distribution, the normalization factor is not monotonically related to the variance of a Poisson variable, which equals the mean.

Multi-Parameter Exponential Family Data

For discussion of generalized PCA formulation so far, we have implicitly assumed that data are from a one-parameter exponential family distribution. To expound on multi-parameter exponential family data, we consider categorical data. As an extension of a Bernoulli distribution, a multinomial distribution can be posited for a categorical variable with more than two categories. Suppose that the categorical variable has K possible outcomes. This categorical variable can be represented by $K-1$ cell counts or proportion variables, where each variable represents the count or proportion of times that category is taken.

The number of trials associated with each case may differ depending on the context. In many scenarios, each case may involve only one trial. A scenario with multiple trials naturally arises in modeling text data, where a case may be a document and the different categories would be the words used in the document. Such data can be arranged into a so-called document-term frequency matrix.

Suppose that $(x_{i1}, \ldots, x_{iK})$ is the ith case with x_{ij} representing the proportion of the jth category from n_i multinomial trials with cell probabilities p_{ij}. The log-likelihood based on the ith case is given by $n_i \sum_{j=1}^{K-1} (x_{ij}\theta_{ij}) - n_i \log\left(1 + \sum_{j=1}^{K-1} e^{\theta_{ij}}\right)$, where $\theta_{ij} = \log(p_{ij}/p_{iK}) \in \mathbb{R}$ and n_i, the number of trials for the ith case, acts as a weight.

With the multi-parameter distribution, $b_j(\cdot)$ is not a separate function for each of the $K-1$ categories. Instead, there is one $b(\cdot)$ function for the whole categorical variable,

$$b(\boldsymbol{\theta}_i) = \log\left(1 + \sum_{l=1}^{K-1} e^{\theta_{il}}\right).$$

The natural parameter from the saturated model, $\tilde{\theta}_{ij}$, is similar to the Bernoulli case, but now depends on both x_{ij} and $x_{iK} := 1 - \sum_{j=1}^{K-1} x_{ij}$. As for other distributions, we must

Table 3: The value of the jth partial derivative of $b(\boldsymbol{\theta})$ evaluated at the approximate saturated natural parameters for different possible combinations of x_{ij} and x_{iK}. N_0 represents the number of $x_{ij}, j = 1, \ldots, K-1$ that are 0. The last column gives the limit as m goes to infinity.

x_{ij}	x_{iK}	$b'_j(\tilde{\boldsymbol{\theta}}_i)$	$\lim_{m\to\infty} b'_j(\tilde{\boldsymbol{\theta}}_i)$
0	$[0,1]$	$e^{-m}/(1+\sum_{l=1}^{K-1} e^{\tilde{\theta}_{il}})$	0
$(0,1)$	$(0,1)$	$(x_{ij}/x_{iK})/(1/x_{iK} + N_0 e^{-m})$	x_{ij}
$(0,1)$	0	$x_{ij}e^m/(1+e^m+N_0e^{-m})$	x_{ij}
1	0	$e^m/(1+e^m+(K-2)e^{-m})$	1

approximate ∞ with a large number m, and specify

$$\tilde{\theta}_{ij} = \begin{cases} -m & \text{if } x_{ij} = 0 \\ \log(x_{ij}/x_{iK}) & \text{if } x_{ij} \in (0,1) \text{ and } x_{iK} \in (0,1) \\ m + \log(x_{ij}) & \text{if } x_{ij} \in (0,1) \text{ and } x_{iK} = 0 \\ m & \text{if } x_{ij} = 1 \end{cases}. \tag{6.1}$$

Lemma 2 shows that as m gets very large, these parameters will converge to the perfect fit: $\frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_i} \longrightarrow \mathbf{x}_i$.

Lemma 2. *For the approximate saturated model parameters $\tilde{\theta}_{ij}$ defined in (6.1),*

$$\lim_{m\to\infty} \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_i} = \mathbf{x}_i.$$

Proof. The jth partial derivative of $b(\boldsymbol{\theta})$ $(j = 1, \ldots, K-1)$ at $\tilde{\boldsymbol{\theta}}_i$ is $\frac{\partial b(\boldsymbol{\theta})}{\partial \theta_j}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_i} = \frac{\exp(\tilde{\theta}_{ij})}{1+\sum_{l=1}^{K-1}\exp(\tilde{\theta}_{il})}$.

The values of this partial derivative evaluated at the approximate saturated model parameters from (6.1) and their limits are given in Table 3. As m goes to infinity, the partial derivative converges to x_{ij} in all cases.

The limit in the first row is true because the denominator is always greater than 1, regardless of the other values of x_{il}. The second row relies on the fact that $1 - x_{iK} = \sum_{l=1}^{K-1} x_{il}$

and the third row relies on the fact that $\sum_{l=1}^{K-1} x_{il} = 1$ when $x_{iK} = 0$. The fourth row is straightforward. □

Proof of the Minimizers of the Majorizing Function

Holding $\boldsymbol{\mu}$ constant, we verify that

$$\begin{aligned}
&\arg\min_{\mathbf{U}^T\mathbf{U}=\mathbf{I}_k} M(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = \arg\min \left\| \left(\mathbf{V}^{(t)}\right)^{1/2} \tilde{\boldsymbol{\Theta}}_c^{(t)} \mathbf{U}\mathbf{U}^T - \left(\mathbf{V}^{(t)}\right)^{1/2} \mathbf{Z}_c^{(t)} \right\|_F^2 \\
&= \arg\min_{\mathbf{U}^T\mathbf{U}=\mathbf{I}_k} tr\left(\mathbf{U}\mathbf{U}^T(\tilde{\boldsymbol{\Theta}}_c^{(t)})^T\mathbf{V}^{(t)}\tilde{\boldsymbol{\Theta}}_c^{(t)}\mathbf{U}\mathbf{U}^T\right) - tr\left(\mathbf{U}\mathbf{U}^T(\tilde{\boldsymbol{\Theta}}_c^{(t)})^T\mathbf{V}^{(t)}\mathbf{Z}_c^{(t)}\right) - tr\left((\mathbf{Z}_c^{(t)})^T\mathbf{V}^{(t)}\tilde{\boldsymbol{\Theta}}_c^{(t)}\mathbf{U}\mathbf{U}^T\right) \\
&= \arg\min_{\mathbf{U}^T\mathbf{U}=\mathbf{I}_k} tr\left[\mathbf{U}^T\left((\tilde{\boldsymbol{\Theta}}_c^{(t)})^T\mathbf{V}^{(t)}\tilde{\boldsymbol{\Theta}}_c^{(t)} - (\tilde{\boldsymbol{\Theta}}_c^{(t)})^T\mathbf{V}^{(t)}\mathbf{Z}_c^{(t)} - (\mathbf{Z}_c^{(t)})^T\mathbf{V}^{(t)}\tilde{\boldsymbol{\Theta}}_c^{(t)}\right)\mathbf{U}\right] \\
&= \arg\max_{\mathbf{U}^T\mathbf{U}=\mathbf{I}_k} tr\left[\mathbf{U}^T\left(\left(\tilde{\boldsymbol{\Theta}}_c^{(t)}\right)^T\mathbf{V}^{(t)}\mathbf{Z}_c^{(t)} + \left(\mathbf{Z}_c^{(t)}\right)^T\mathbf{V}^{(t)}\tilde{\boldsymbol{\Theta}}_c^{(t)} - \left(\tilde{\boldsymbol{\Theta}}_c^{(t)}\right)^T\mathbf{V}^{(t)}\tilde{\boldsymbol{\Theta}}_c^{(t)}\right)\mathbf{U}\right].
\end{aligned}$$

The trace in the last line is maximized by the first k eigenvectors of $\left(\tilde{\boldsymbol{\Theta}}_c^{(t)}\right)^T\mathbf{V}^{(t)}\mathbf{Z}_c^{(t)} + \left(\mathbf{Z}_c^{(t)}\right)^T\mathbf{V}^{(t)}\tilde{\boldsymbol{\Theta}}_c^{(t)} - \left(\tilde{\boldsymbol{\Theta}}_c^{(t)}\right)^T\mathbf{V}^{(t)}\tilde{\boldsymbol{\Theta}}_c^{(t)}$, as shown by Fan (1949).

Given $\mathbf{U}^{(t)}$, minimizing $M(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})$ with respect to $\boldsymbol{\mu}$ is equivalent to minimizing

$$\begin{aligned}
&tr\left[\left(\mathbf{I} - \mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T\right)\boldsymbol{\mu}\left(\mathbf{1}_n^T\mathbf{V}^{(t)}\mathbf{1}_n\right)\boldsymbol{\mu}^T\left(\mathbf{I} - \mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T\right)\right] - \\
&2tr\left[\left(\mathbf{Z}^{(t)} - \tilde{\boldsymbol{\Theta}}\mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T\right)^T\mathbf{V}^{(t)}\mathbf{1}_n\boldsymbol{\mu}^T\left(\mathbf{I} - \mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T\right)\right].
\end{aligned}$$

The minimizer $\boldsymbol{\mu}$ can be found by differentiating $M(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})$ with respect to $\boldsymbol{\mu}$ and setting the gradient equal to zero, which leads to

$$\left(\mathbf{I} - \mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T\right)\boldsymbol{\mu}\left(\mathbf{1}_n^T\mathbf{V}^{(t)}\mathbf{1}_n\right) = \left(\mathbf{I} - \mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T\right)\left(\mathbf{Z}^{(t)} - \tilde{\boldsymbol{\Theta}}\mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T\right)^T\mathbf{V}^{(t)}\mathbf{1}_n.$$

The update rule produces $\boldsymbol{\mu}$ that satisfies the equation at each step since $\mathbf{I} - \mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T$ is a projection matrix and $\mathbf{1}_n^T\mathbf{V}^{(t)}\mathbf{1}_n$ is a scalar.

Stability of Algorithm 1 with Random Initialization

To demonstrate the stability of Algorithm 1, we used 100 random initializations of $\mathbf{U}^{(1)}$ and compared their fitted deviances using $k = 5$ on a simulated dataset with $P(X = 0) = 0.9$,

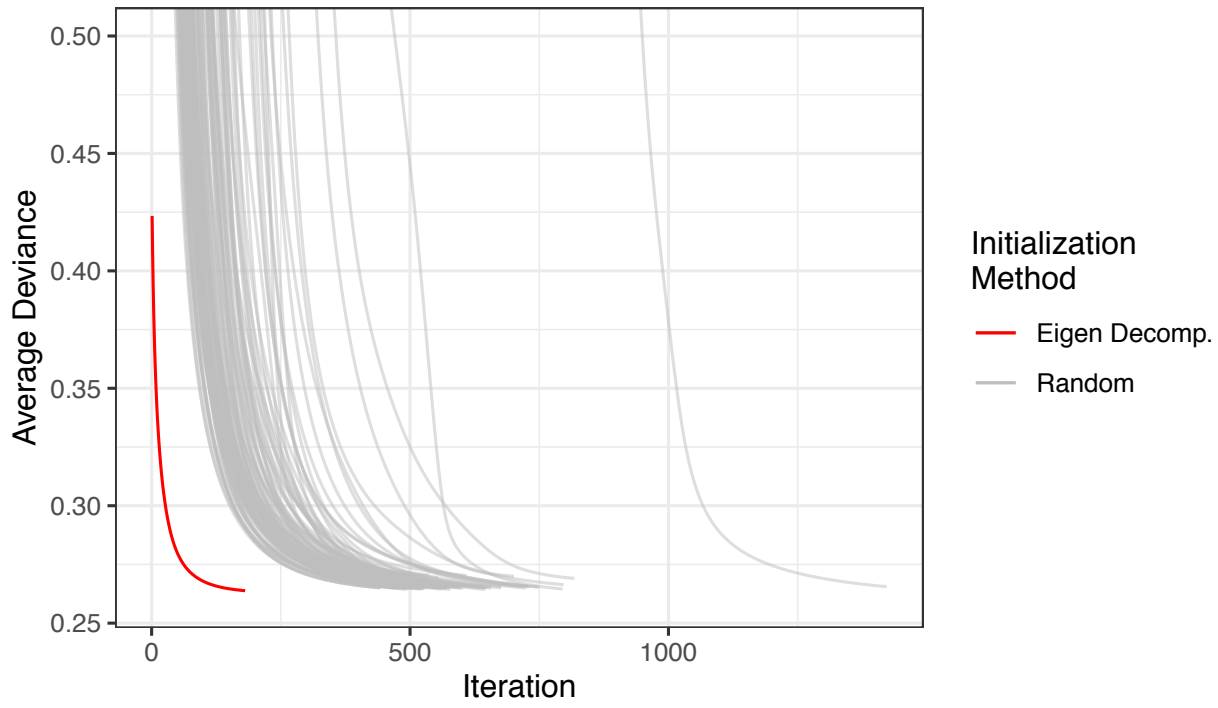

Figure 10: Average training deviance of solutions as a function of the number of iterations for 100 random initializations trained on a simulated dataset with $P(X = 0) = 0.9$, $E(X|X > 0) = 3$, $n = 100$, and $d = 50$ and fit with $k = 5$.

$E(X|X > 0) = 3$, $n = 100$, and $d = 50$ considered in Section 4. The results are shown in Figure 10, along with the result using the eigen-decomposition for initialization. $\mathbf{U}$ was randomly initialized by simulating independent standard normal elements in an $n \times k$ matrix and performing a QR decomposition to orthonormalize the matrix. All 100 replications converged to similar deviances, which explain between 89.58% and 89.82% of the null deviance. The random initializations are slightly worse than the eigen-decomposition based initialization, which explains 89.84% of the null deviance, and they take more iterations to converge.

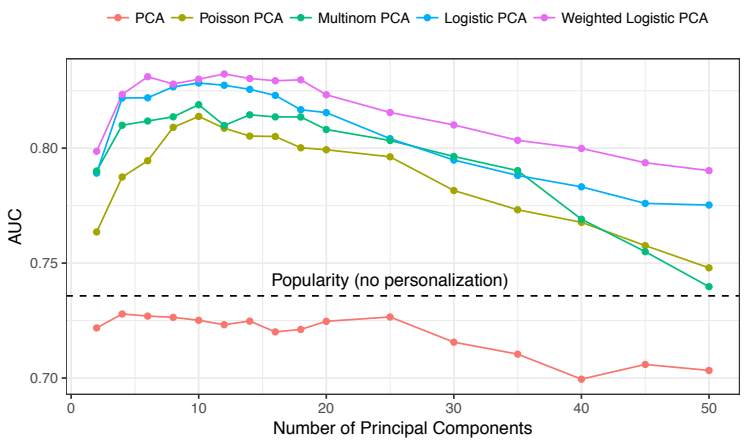

Figure 11: Comparison of different generalized PCA methods for recommending songs to users with the play count data

Additional Million Song Dataset Analysis

To highlight additional ways in which this data can be analyzed by generalized PCA, we have also binarized the subset of the MSD used for recommendations, converting all listen counts greater than 0 to 1. With this, we applied logistic PCA to the data. Further, we weighted the binary responses, giving higher weights to the larger counts. Figure 11 shows the results of these analyses. Binarization and weighting both improve the AUC, indicating that most important feedback in the data is whether or not the user listened to the song, and of less importance is the number of times.