Abstract

In matrix completion, we are given a matrix where the values of only some of the entries are present, and we want to reconstruct the missing ones. Much work has focused on the assumption that the data matrix has low rank. We propose a more general assumption based on denoising, so that we expect that the value of a missing entry can be predicted from neighboring points. We propose a nonparametric version of denoising based on local, iterated averaging with mean-shift, possibly constrained to preserve local low-rank manifold structure. The few user parameters required (the denoising scale, number of neighbors and local dimensionality) and the number of iterations can be estimated by cross-validating the reconstruction error.

Denoising with (manifold) blurring mean-shift algorithms (GBMS/MBMS)

- Consider a (fully observed) dataset \( \{x_{ij}\}_{i,j} \subset \mathbb{R}^d \) and define a kernel density estimate \( \hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i) \), with Gaussian kernel \( G_k(x) = \exp\frac{-1}{2}(x - x_i)^T G_k(x - x_i) \), where \( G_k \) is the covariance matrix of the stationary point equation of \( \nabla \hat{p}(x) = 0 \) the iterative scheme \( x^{(n+1)} = \frac{1}{\hat{p}(x^{(n)})} \sum_{i=1}^{N} p(x - x_i) x_i + \frac{1}{\hat{p}(x^{(n)})} \sum_{i=1}^{N} p(x - x_i) x_i \).
- The blurring mean-shift algorithm (GBMS) applies one step of the previous scheme, initialized from every point, in parallel for all points, and replaces \( x \) with the updated dataset \( X \), which is a blurred (shrunken) version of \( X \). And the algorithm iterates this process to maximize the objective function \( E(X) = \sum_{i,j} |x_{ij} - \hat{p}(x_{ij})|^2 \) by taking a mean-shift step for every point in parallel.

GBMS (\( k, \sigma \)) with full or \( k \)-nn graph: given \( X_{D \times M} \)

\[
\text{repeat for } n = 1, \ldots, N
\]
\[
X_{\tau} \leftarrow \{1, \ldots, N\} \quad \text{(full graph)}
\]
\[
\text{or } \{ \text{k nearest neighbors of } x_i \} \quad \text{(k-nn graph)}
\]
\[
\begin{align*}
\Delta x_i &\leftarrow x_i - \frac{1}{\sum_{j=1}^{n} p(x_i - x_j)} \sum_{j=1}^{n} p(x_i - x_j) x_j \\
&\quad \text{mean shift step for all points, and replaces } x_i \\
&\quad \text{return } X
\end{align*}
\]

MBMS (\( k, \sigma \)) with full or \( k \)-nn graph: given \( X_{D \times M} \)

\[
\text{repeat for } n = 1, \ldots, N
\]
\[
X_{\tau} \leftarrow \{1, \ldots, N\} \quad \text{(full graph)}
\]
\[
\text{or } \{ \text{k nearest neighbors of } x_i \} \quad \text{(k-nn graph)}
\]
\[
\begin{align*}
\Delta x_i &\leftarrow x_i - \frac{1}{\sum_{j=1}^{n} p(x_i - x_j)} \sum_{j=1}^{n} p(x_i - x_j) x_j \\
&\quad \text{mean shift step for all points, and replaces } x_i \\
&\quad \text{return } X
\end{align*}
\]

Convergence and stopping criterion

- Running the algorithm to convergence would equalize all values; instead, we want to achieve just enough denoising and stop the algorithm, as was the case with GBMS clustering.
- We determine the optimal number of iterations and all other parameters by cross-validation: select a held-out set by picking a random subset of the present entries and considering them as missing; this allows us to evaluate an error between our completion for them and the ground truth. We stop iterating when this error increases.

5 Experimental results

- Compare with representative methods:
  - low-rank matrix completion method—singular value projection (SVP); fitting the data with a low-r dimensional Gaussian model with EM and imputing the missing values of each \( x \), as the conditional mean; nonlinear PCA (nPCA) (Scholz 2005).
- We initialize our algorithms from them.
- Train on 90% present entries and cross validate user parameters on the remaining 10% present entries. Then run the algorithms with optimal parameters values on the entire present data and report the test error with the ground truth.

Mocap data

\[
\text{Name 1 (seg distance)} \quad \text{Name 10 (foot pose)} \quad \text{Name 147 (leg pose)}
\]

\[
\% \text{ of present data}
\]

- Left: mean of errors of 5 runs obtained by different algorithms for varying percentage of missing values. Right: sample reconstructions when 85% percent data is missing.

Row 2: 1 initialization + MBMS. Color indicates different initialization: black, original data; red, nPCA; blue, SVP; green, Gaussian.

100D Swissroll

\[
\begin{align*}
\text{SVP} \quad &\text{r = 0} \\
\text{SVP + GBMS} \quad &r = 1 \\
\text{SVP + MBMS} \quad &r = 2 \\
\text{Gaussian} \quad &r = 0 \\
\text{Gaussian + GBMS} \quad &r = 1 \\
\text{Gaussian + MBMS} \quad &r = 2
\end{align*}
\]

Denoising effect of different algorithms over iterations. The initializations given by SVP and Gaussian model are both quite noisy.

Reconstruction error of GBMS over iterations (each curve is a different \( r \) value).

MNIST digit ‘7’

\[
\begin{align*}
\text{Organ} \quad &\text{Stat.} \\
\text{Org.} \quad &\text{mean} \\
\text{SVP} \quad &6.99 \\
\text{SVP + MBMS} \quad &6.54 \\
\text{SVP + GBMS} \quad &6.03
\end{align*}
\]

Selected reconstructions of MNIST block-occuded digits ‘7’ (50% of the pixels are missing) with different methods. We use rank 10 for SVP and \( L = 1 \) for MBMS.

Reconstruction errors of different algorithms along with their optimal parameters.

6 Discussion

- A special case of our algorithm \( (k = N \text{ and } \sigma = \infty) \) is directly related to low-rank matrix completion algorithms (alternate between SVD projection and resetting values).
- The idea of averaging values of bordering points is similar to one category of collaborative filtering methods that essentially use similar users/items to predict missing values.
- The MBMS-based algorithm bridges the gap between pure denoising (GBMS) and local low rank. Other definitions of denoising should be possible.