A Denoising View of Matrix Completion

Weiran Wang, Miguel Á. Carreira-Perpiñán
EECS, University of California, Merced

Zhengdong Lu
Microsoft Research Asia

1 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 the values of 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.

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

- Consider a (fully observed) dataset $\{\mathbf{x}_n\}_{n=1}^N \subset \mathbb{R}^D$ and define a kernel density estimate $p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^N G_\sigma(\mathbf{x}, \mathbf{x}_n)$, with Gaussian kernel $G_\sigma(\mathbf{x}, \mathbf{x}_n) \propto \exp\left(-\frac{1}{2}(\|\mathbf{x}-\mathbf{x}_n\|/\sigma)^2\right)$. The mean-shift (GMS) algorithm rearranges the stationary point equation of $\nabla p(\mathbf{x}) = \mathbf{0}$ into the iterative scheme $\mathbf{x}^{(\tau+1)} = \mathbf{f}(\mathbf{x}^{(\tau)}) = \sum_{n=1}^N p(n|\mathbf{x}^{(\tau)})\mathbf{x}_n$ with $p(n|\mathbf{x}^{(\tau)}) = \frac{G_\sigma(\mathbf{x}^{(\tau)}, \mathbf{x}_n)}{\sum_{n'=1}^N G_\sigma(\mathbf{x}^{(\tau)}, \mathbf{x}_{n'})}$.
- The blurring mean-shift algorithm (GBMS) applies one step of the previous scheme, initialized from every point, in parallel for all points, and replaces $\mathbf{X}$ with the updated dataset $\tilde{\mathbf{X}}$, which is a blurred (shrunk) version of $\mathbf{X}$. And the algorithm iterates this process to maximize the objective function $E(\mathbf{X}) = \sum_{n=1}^N p(\mathbf{x}_n) = \frac{1}{N}\sum_{n,m=1}^N G_\sigma(\mathbf{x}_n, \mathbf{x}_m)$ by taking a mean-shift step for every point in parallel.
- GBMS can be seen as a data-dependent low-pass filter which denoises equally in all directions. When the data lies on a low-dimensional manifold, denoising orthogonally to it removes out-of-manifold noise, while denoising tangentially to it perturbs intrinsic degrees of freedom and causes shrinkage.
- The manifold blurring mean-shift algorithm (MBMS) first computes a predictor averaging step with GBMS, and then for each point $\mathbf{x}_n$ a corrector projective step removes the step direction lying in the tangent space of $\mathbf{x}_n$ (estimated locally with PCA). Both GBMS and MBMS must be stopped early to prevent excessive denoising and distortions.

GBMS (k, σ) with full or k-nn graph: given $\mathbf{X}_{D\times N}$, $\mathcal{M}$

repeat
  for $n = 1, \dots, N$
    $\mathcal{N}_n \leftarrow \{1, \dots, N\}$ (full graph) or
      k nearest neighbors of $\mathbf{x}_n$ (k-nn graph)
    $\partial\mathbf{x}_n \leftarrow -\mathbf{x}_n + \sum_{m\in\mathcal{N}_n} \frac{G_\sigma(\mathbf{x}_n,\mathbf{x}_m)}{\sum_{m'\in\mathcal{N}_n} G_\sigma(\mathbf{x}_n,\mathbf{x}_{m'})}\mathbf{x}_m$ — mean-shift step
  end
  $\mathbf{X}_\mathcal{M} \leftarrow \mathbf{X}_\mathcal{M} + (\partial\mathbf{X})_\mathcal{M}$ — move points' missing entries
until validation error increases
return $\mathbf{X}$

MBMS (L, k, σ) with full or k-nn graph: given $\mathbf{X}_{D\times N}$, $\mathcal{M}$

repeat
  for $n = 1, \dots, N$
    $\mathcal{N}_n \leftarrow \{1, \dots, N\}$ (full graph) or
      k nearest neighbors of $\mathbf{x}_n$ (k-nn graph)
    $\partial\mathbf{x}_n \leftarrow -\mathbf{x}_n + \sum_{m\in\mathcal{N}_n} \frac{G_\sigma(\mathbf{x}_n,\mathbf{x}_m)}{\sum_{m'\in\mathcal{N}_n} G_\sigma(\mathbf{x}_n,\mathbf{x}_{m'})}\mathbf{x}_m$ — mean-shift step
    $\mathcal{X}_n \leftarrow k$ nearest neighbors of $\mathbf{x}_n$
    $(\boldsymbol{\mu}_n, \mathbf{U}_n) \leftarrow \text{PCA}(\mathcal{X}_n, L)$ — estimate L-dim tangent space at $\mathbf{x}_n$
    $\partial\mathbf{x}_n \leftarrow (\mathbf{I} - \mathbf{U}_n\mathbf{U}_n^T)\partial\mathbf{x}_n$ — subtract parallel motion
  end
  $\mathbf{X}_\mathcal{M} \leftarrow \mathbf{X}_\mathcal{M} + (\partial\mathbf{X})_\mathcal{M}$ — move points' missing entries
until validation error increases
return $\mathbf{X}$

LTP (L, k) with k-nn graph: given $\mathbf{X}_{D\times N}$, $\mathcal{M}$

repeat
  for $n = 1, \dots, N$
    $\mathcal{X}_n \leftarrow k$ nearest neighbors of $\mathbf{x}_n$
    $(\boldsymbol{\mu}_n, \mathbf{U}_n) \leftarrow \text{PCA}(\mathcal{X}_n, L)$ — estimate L-dim tangent space at $\mathbf{x}_n$
    $\partial\mathbf{x}_n \leftarrow (\mathbf{I} - \mathbf{U}_n\mathbf{U}_n^T)(\boldsymbol{\mu}_n - \mathbf{x}_n)$ — project point onto tangent space
  end
  $\mathbf{X}_\mathcal{M} \leftarrow \mathbf{X}_\mathcal{M} + (\partial\mathbf{X})_\mathcal{M}$ — move points' missing entries
until validation error increases
return $\mathbf{X}$

3 GBMS/MBMS for matrix completion

Consider the natural extension of GBMS to matrix completion by adding the constraints given by present values:

$$\max_{\mathbf{X}} E(\mathbf{X}) = \sum_{n,m=1}^N G_\sigma(\mathbf{x}_n, \mathbf{x}_m) \quad \text{s.t.} \quad \mathbf{X}_\mathcal{P} = \overline{\mathbf{X}}_\mathcal{P},$$

where $\mathbf{X}_\mathcal{M}$ and $\mathbf{X}_\mathcal{P}$ indicate selection of the missing or present values and $\overline{\mathbf{X}}_\mathcal{P}$ are the present matrix entries.

- Similar to low-rank formulations for matrix completion that have the same constraints but use as objective function the reconstruction error, e.g. $\|\mathbf{X} - \mathbf{ABX}\|^2$ with $\mathbf{A}_{D\times L}$, $\mathbf{B}_{L\times D}$ and $L < D$.
- We initialize $\mathbf{X}_\mathcal{M}$ to the output of other matrix completion method, and apply gradient projected method, with gradient $\nabla_{\mathbf{x}_n} E(\mathbf{X}) \propto \frac{2}{\sigma^2} p(\mathbf{x}_n)\left(-\mathbf{x}_n + \sum_{m=1}^N p(m|\mathbf{x}_n)\mathbf{x}_m\right)$ and projection $\mathbf{X}^{(\tau+1)} = \mathbf{X}^{(\tau)} + \alpha\Pi_\mathcal{P}(\nabla_\mathbf{X} E(\mathbf{X}^{(\tau)}))$.
- We omit the line search and perform GBMS within the missing value space: apply a denoising step, refill the present values, iterate until the validation error increases.
- We also generalize the GBMS step to MBMS step in the algorithm and obtain superior results.
- Computational cost per iteration is, with fixed k-nn graph, $\mathcal{O}(NkD)$ for GBMS and $\mathcal{O}(NkD + N(D+k)\min(D,k)^2)$ for MBMS, times the proportion of missing data.

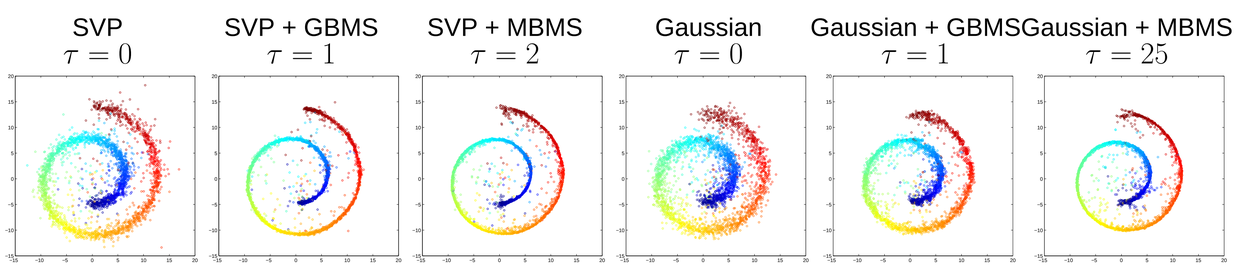
4 Convergence and stopping criterion

- 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.

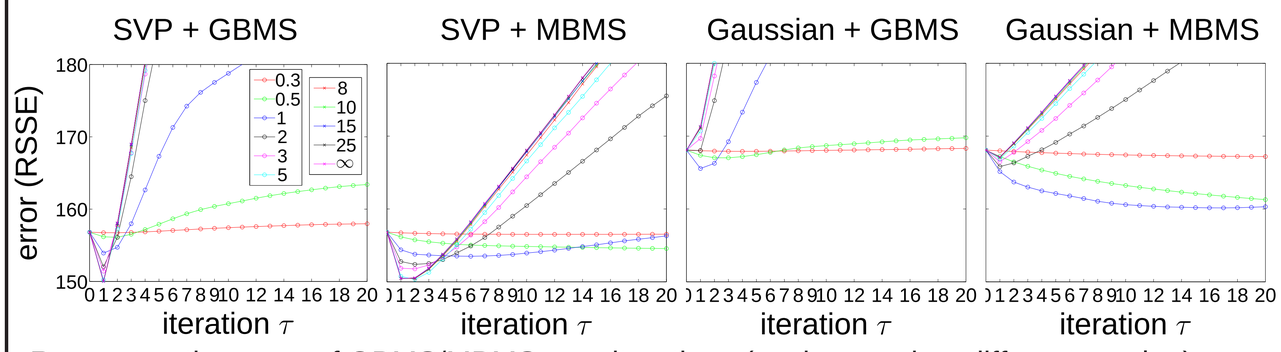
5 Experimental results

- Compare with representative methods:
  - low-rank matrix completion method—singular value projection (SVP);
  - fitting the data with a D-dimensional Gaussian model with EM and imputing the missing values of each $\mathbf{x}_n$ as the conditional mean;
  - nonlinear PCA (nlPCA) (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.

100D Swissroll



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



Reconstruction error of GBMS/MBMS over iterations (each curve is a different σ value).

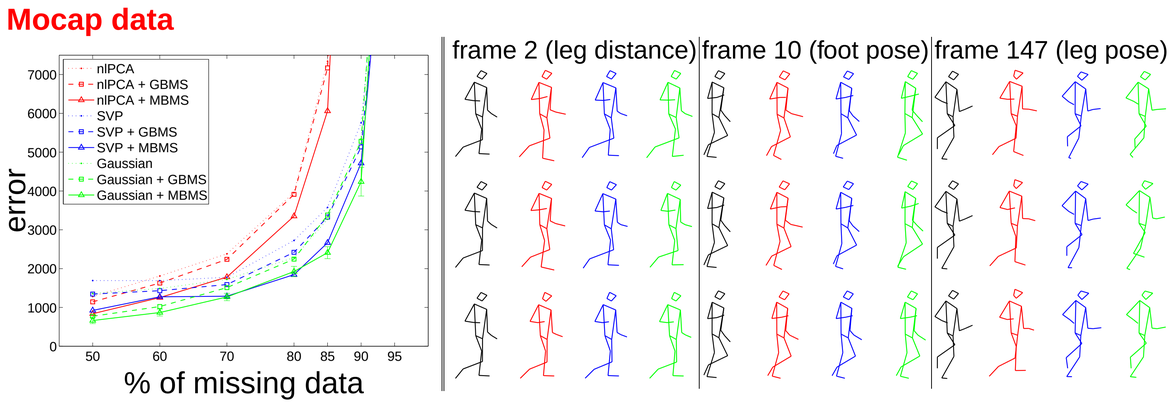
MNIST digit '7'

Selected reconstructions of MNIST block-occluded digits '7' (50% of the pixels are missing) with different methods. We use rank 10 for SVP and $L = 9$ for MBMS.

Methods	RSSE	mean	stdev
nlPCA	7.77	26.1	42.6
SVP	6.99	21.8	39.3
+ GBMS (400,140,0,1)	6.54	18.8	37.7
+ MBMS (500,140,9,5)	6.03	17.0	34.9

Reconstruction errors ($\times 10^{-4}$) and optimal parameters.

Mocap 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 1**: initialization. **Row 2**: init+GBMS. **Row 3**: init+MBMS. Color indicates different initialization: black, original data; red, nlPCA; blue, SVP; green, Gaussian.

Methods	RSSE	mean	stdev
Gaussian	168.1	2.63	1.59
+ GBMS (∞, 10, 0, 1)	165.8	2.57	1.61
+ MBMS (1, 20, 2, 25)	157.2	2.36	1.63
SVP	156.8	1.94	2.10
+ GBMS (3, 50, 0, 1)	151.4	1.89	2.02
+ MBMS (3, 50, 2, 2)	151.8	1.87	2.05

Reconstruction errors obtained by different algorithms along with their optimal parameters (σ, k, L, no. iterations τ).

6 Discussion

- A special case of our algorithm ($k = N$ 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 neighboring 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.

NIPS 2011, Granada, Spain.

Work supported by NSF CAREER award IIS0754089.