Bilinear Generalized Approximate Message Passing (BiG-AMP) for Matrix Completion

Jason T. Parker and Phil Schniter

Joint work with Jeremy Vila, Subhojit Som, and Volkan Cevher

(With support from NSF CCF-1218754, DARPA/ONR N66001-10-1-4090, NSF IIP-0968910, and an AFOSR lab task.)

Three Important Matrix Recovery Problems:

- **Matrix Completion** (MC):

  Recover low-rank matrix $\boldsymbol{X}$ from AWGN-corrupted incomplete observations $\boldsymbol{Y} = \mathcal{P}_\Omega\big(\boldsymbol{X} + \boldsymbol{W}\big)$.

- **Robust Principle Components Analysis** (RPCA):

  Recover low-rank matrix $\boldsymbol{X}$ and sparse matrix $\boldsymbol{S}$ from AWGN-corrupted observations $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{S} + \boldsymbol{W}$.

- **Dictionary Learning** (DL):

  Recover overcomplete dictionary $\boldsymbol{A}$ and sparse matrix $\boldsymbol{S}$ from AWGN-corrupted observations $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{W}$.

The following **extensions** may also be of interest:

- RPCA and DL with incomplete observations and/or structured sparsity.
- Any of the above with a non-additive noise model (e.g., quantized $\boldsymbol{Y}$).

Our contribution:

- We propose a novel unified approach to these matrix-recovery problems that leverages the recent framework of **approximate message passing** (AMP).

- While previous AMP algorithms have been proposed for the **linear model**:
  - Infer $\boldsymbol{s} \sim \prod_n p_S(s_n)$ from $\boldsymbol{y} = \boldsymbol{\Phi s} + \boldsymbol{w}$ with AWGN $\boldsymbol{w}$ and known $\boldsymbol{\Phi}$ [Donoho/Maleki/Montanari'10]

  or the **generalized linear model**:
  - Infer $\boldsymbol{s} \sim \prod_n p_S(s_n)$ from $\boldsymbol{y} \sim \prod_m p_{Y|X}(y_m|x_m)$ with hidden $\boldsymbol{x} = \boldsymbol{\Phi s}$ and known $\boldsymbol{\Phi}$ [Rangan'10]

  our new algorithm is formulated for the **generalized bilinear model**:
  - Infer $\boldsymbol{A} \sim \prod_{m,r} p_A(a_{mr})$ and $\boldsymbol{B} \sim \prod_{r,n} p_B(b_{rn})$ from $\boldsymbol{Y} \sim \prod_{m,n} p_{Y|X}(y_{mn}|x_{mn})$ with hidden $\boldsymbol{X} = \boldsymbol{AB}$ [Parker/Schniter/Cevher'11,12]

- Our work is still **in-progress**. Today we will focus on results for **Matrix Completion**. A journal submission with **RPCA** and **DL** examples is in preparation. Preliminary results are encouraging; stay tuned!

Outline:

1. **Brief review** of popular approaches to matrix-completion and robust PCA:
   - Convex
   - Greedy
   - Bayesian

2. **Bilinear Generalized AMP (BiG-AMP)**.
   - What is it?
   - What are AMP's approximations?
   - How to apply to MC, RPCA, DL?

3. **Preliminary results**:
   - Phase transition curves
   - NMSE and runtime
   - Practical examples: image completion, video surveillance

Convex-Optimization for Matrix-Completion & Robust PCA:

- Consider the combined MC-and-RPCA problem:

  Recover low-rank $\boldsymbol{X}$ and sparse $\boldsymbol{S}$ from AWGN-corrupted incomplete observations $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X}+\boldsymbol{S}+\boldsymbol{W})$.

- Optimization approach:

$$\min_{\boldsymbol{X},\boldsymbol{S}} \left\{ \mathrm{rank}(\boldsymbol{X}) + \gamma\|\boldsymbol{S}\|_0 \right\} \text{ s.t. } \|\mathcal{P}_\Omega(\boldsymbol{X}+\boldsymbol{S}) - \boldsymbol{Y}\|_F \leq \eta \quad \ldots \textbf{intractable}$$

$$\min_{\boldsymbol{X},\boldsymbol{S}} \left\{ \|\boldsymbol{X}\|_* + \gamma\|\boldsymbol{S}\|_1 \right\} \text{ s.t. } \|\mathcal{P}_\Omega(\boldsymbol{X}+\boldsymbol{S}) - \boldsymbol{Y}\|_F \leq \eta \quad \ldots \textbf{convex}!$$

- Convex relaxation yields **perfect noiseless** & **stable noisy** recovery when:
  - $\mathrm{rank}(\boldsymbol{X})$ is sufficiently small,
  - singular vectors of $\boldsymbol{X}$ are not too cross-correlated nor too spiky,
  - support of $\boldsymbol{S}$ is random and sufficiently sparse,
  - observation set Ω is random and sufficiently large.

  Details given in, e.g., [Candés/Recht'08], [Candés/Plan'09], [Candés/Li/Ma/Wright'09], [Zhou/Wright/Li/Candés/Ma'10], and [Chen/Jalali/Sanghavi/Caramanis'11].

Fast Algorithms for Convex Matrix-Completion & Robust PCA:

- A comparison of convex RPCA algorithms is given at Yi Ma's webpage:

  `http://perception.csl.uiuc.edu/matrix-rank/sample_code.html`

Algorithm	Error	Time (sec)
Singular Value Thresholding [Cai/Candes/Shen'08]	3.4e-4	**877**
Dual Method [Lin/Ganesh/Wright/Wu/Chen/Ma'09]	1.6e-5	**177**
Accelerated Proximal Gradient (partial SVD) [Lin/Ganesh/Wright/Wu/Chen/Ma'09]	1.8e-5	**8**
Alternating Direction Methods [Yuan/Yang'09]	2.2e-5	**5**
Exact Augmented Lagrange Method [Lin/Chen/Wu/Ma'09]	7.6e-8	**4**
Inexact Augmented Lagrange Method [Lin/Chen/Wu/Ma'09]	4.3e-8	**2**

  for the recovery of 400×400 rank-20 matrix $\boldsymbol{X}$ corrupted by 5%-sparse $\boldsymbol{S}$ with amplitudes uniform in $[-50, 50]$.

- Evidently a lot of progress has been made! Can one do better?

Greedy Approaches to Matrix-Completion & Robust PCA:

- First consider **matrix completion**, where we want to recover low-rank $\boldsymbol{X}$ from AWGN-corrupted incomplete observations $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X} + \boldsymbol{W})$.
- If we suppose that ...

  $\boldsymbol{X} \in \mathbb{R}^{M \times N}$ is square or tall (i.e., $M \geq N$) with $\mathrm{rank}(\boldsymbol{X}) = R$,

  then the difficult part of the MC problem is finding the column space of $\boldsymbol{X}$, leading to squared-error minimization on the **Grassmanian manifold** $\mathcal{G}_{M,R}$:

$$\min_{\boldsymbol{A} \in \mathcal{G}_{M,R}} \min_{\boldsymbol{B}} \|\mathcal{P}_\Omega(\boldsymbol{A}\boldsymbol{B}) - \boldsymbol{Y}\|_F^2$$

- Example algorithms:
  - **Optspace** [Keshavan/Montanari/Oh'09]: Grad-descent minimizing $(\boldsymbol{A}, \boldsymbol{B})$.
  - **SET** [Dai/Milenkovic'09]: Solves for $\boldsymbol{B}$, then takes gradient w.r.t $\boldsymbol{A}$.
  - **GROUSE** [Balzano/Nowak/Recht'10]: Grad-descent one column at a time.
- This greedy approach can also be extended to **RPCA**:
  - **GRASTA** [He/Balzano/Lui'11].

Bayesian Approaches to Matrix-Completion & Robust PCA:

- First consider **matrix completion**, where we want to recover low-rank $\boldsymbol{X}$ from AWGN-corrupted incomplete observations $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X} + \boldsymbol{W})$.
- The *basic* Bayesian approach decomposes $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}$ and assumes priors $\boldsymbol{A} \sim \mathcal{N}(\boldsymbol{0}, \sigma_A^2 \boldsymbol{I})$ and $\boldsymbol{B} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The log posterior then becomes

$$\ln p(\boldsymbol{A}, \boldsymbol{B}|\boldsymbol{Y}) = \frac{1}{2\sigma_W^2}\|\mathcal{P}_\Omega(\boldsymbol{A}\boldsymbol{B}) - \boldsymbol{Y}\|_F^2 + \frac{1}{2\sigma_A^2}\|\boldsymbol{A}\|_F^2 + \frac{1}{2}\|\boldsymbol{B}\|_F^2 + C.$$

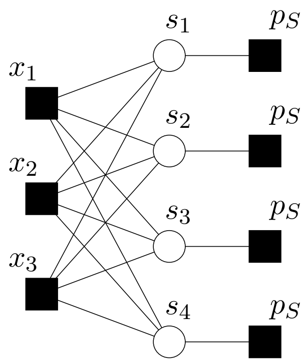
  To infer $(\boldsymbol{A}, \boldsymbol{B})$, various schemes have been proposed, e.g.,
  - **EM** ("Probabilistic PCA") [Tipping/Bishop'99]
  - **SDP** ("Maximum-Margin Matrix Factorization") [Srebro/Rennie/Jaakkola'04]
  - **VB** ("Variational Bayes") [Lim/Teh'07]
  - **MCMC** ("Probabilistic Matrix Factorization") [Salakhutdinov/Mnih'08]

  Each has their own way of estimating the hyperparameters $\{\sigma_W^2, \sigma_A^2\}$.
- This approach can be extended to **RPCA** by changing the noise model to a heavy-tailed one (e.g., [Luttinen/Ilin/Karhunen'09], [Ding/He/Carin'11]).

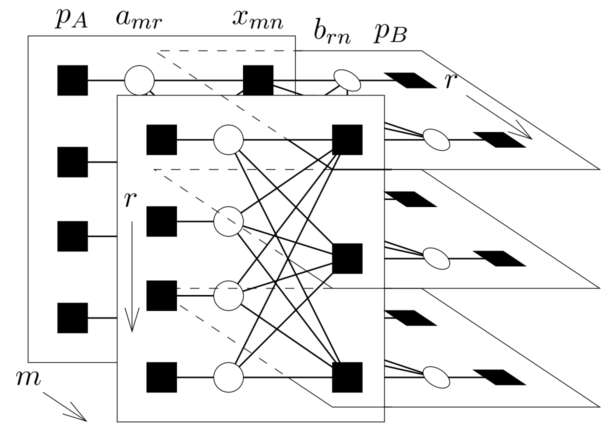
Bilinear Generalized AMP (BiG-AMP):

- **BiG-AMP** is a Bayesian approach that uses **approximate message passing (AMP)** strategies to infer $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{S})$.

Compressive Sensing (CS):

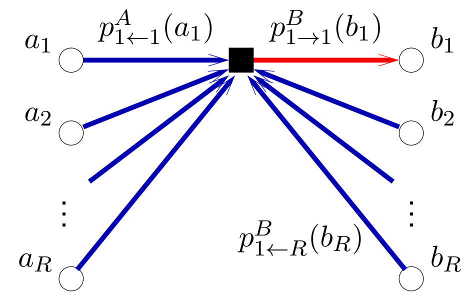


MC/RPCA/DL:



- In AMP, beliefs are propagated on a loopy factor graph using approximations that exploit the **blessings of dimensionality**:
  1. Gaussian message approximation (motivated by CLT),
  2. Taylor-series approximation of message differences.
- A rigorous large-system analysis of AMP for CS (with i.i.d Gaussian $\boldsymbol{\Phi}$) has established a number of optimalities [Bayati/Montanari'10],[Rangan'10].

BiG-AMP Approximations (sum-product version):



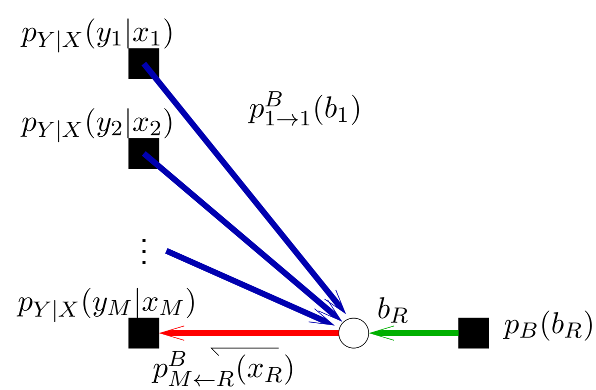
1. Message from i^{th} node of $\boldsymbol{X}$ to j^{th} node of $\boldsymbol{B}$:

$$p^B_{i\to j}(b_j) \propto \int_{\{a_r\}_{r=1}^R, \{b_r\}_{r\neq j}} p_{Y|X}\Big(y_i \Big| \overbrace{\textstyle\sum_r a_r b_r}^{x_i|b_j \approx \mathcal{N} \text{ via CLT!}}\Big) \Big(\textstyle\prod_r p^B_{i\leftarrow r}(b_r)\Big)\Big(\textstyle\prod_{r\neq j} p^A_{i\leftarrow r}(a_r)\Big)$$

$$\approx \int_{x_i} p_{Y|X}(y_i|x_i)\,\mathcal{N}\big(x_i; \hat{x}_i(b_j), \nu^x_i(b_j)\big) \approx \mathcal{N} \text{ (exact for AWGN!)}$$

To compute $\hat{x}_i(b_j), \nu^x_i(b_j)$, the means and variances of $p^B_{i\leftarrow r}, p^A_{i\leftarrow r}$ suffice, thus we have **Gaussian message passing!** (Same thing happens with $\boldsymbol{X} \to \boldsymbol{A}$ messages.)

2. Although Gaussian, we still have $4MNR$ messages to compute (too many!). Exploiting similarity among the messages $\{p^B_{i\leftarrow j}\}_{i=1}^M$, AMP employs a Taylor-series approximation whose error vanishes as $M \to \infty$. (Same for $\{p^A_{i\leftarrow j}\}_{i=1}^N$.) In the end, AMP only needs to compute $\mathcal{O}(MN)$ **messages**!



BiG-AMP for MC, RPCA, and DL:

BiG-AMP can be applied to a wide variety of matrix recovery problems:

- **Matrix Completion** (MC):

  Recover low-rank $\boldsymbol{AB}$ from $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{AB} + \boldsymbol{W})$.

  ...set $\boldsymbol{A} \sim \mathcal{N}(\boldsymbol{0}, \sigma_A^2 \boldsymbol{I})$ and $\boldsymbol{B} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

- **Robust PCA** (RPCA):

  Recover low-rank $\boldsymbol{AB}$ and sparse $\boldsymbol{S}$ from $\boldsymbol{Y} = \boldsymbol{AB} + \boldsymbol{S} + \boldsymbol{W}$.

  ...set $\boldsymbol{A} \sim \mathcal{N}(\boldsymbol{0}, \sigma_A^2 \boldsymbol{I})$, $\boldsymbol{B} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{S} \sim$ Bern(λ)-$\mathcal{N}(\boldsymbol{0}, \sigma_S^2 \boldsymbol{I})$.

- **Dictionary Learning** (DL):

  Recover overcomplete $\boldsymbol{A}$ and sparse $\boldsymbol{S}$ from $\boldsymbol{Y} = \boldsymbol{AS} + \boldsymbol{W}$.

  ...set $\boldsymbol{A} \sim \mathcal{N}(\boldsymbol{0}, \sigma_A^2 \boldsymbol{I})$ and $\boldsymbol{S} \sim$ Bern(λ)-$\mathcal{N}(\boldsymbol{0}, \sigma_S^2 \boldsymbol{I})$.

Moreover:

- **Non-Gaussian** (e.g., quantized) observations can be incorporated via $p_{Y|X}$.
- **Structured sparsity** can be incorporated via "**turbo-AMP**." [Schniter'10]
- Hyperparameters can be learned via **EM**. [Ziniel/Schniter'10],[Vila/Schniter'11,12]

BiG-AMP in Context:

Advantages:

- A **unified** approach to a wide range of problems, e.g., MC, RPCA, DL, ...
- Competitive with best algorithms for each application.
  - **Very fast** and **scaleable**: no SVDs, easily parallelizable.

    ...will see from runtime curves.
  - **Accurate**: in part due to flexibility of choice of priors.

    ...will see from phase transition and NMSE curves.

Relation to other message-passing algorithms for matrix completion:

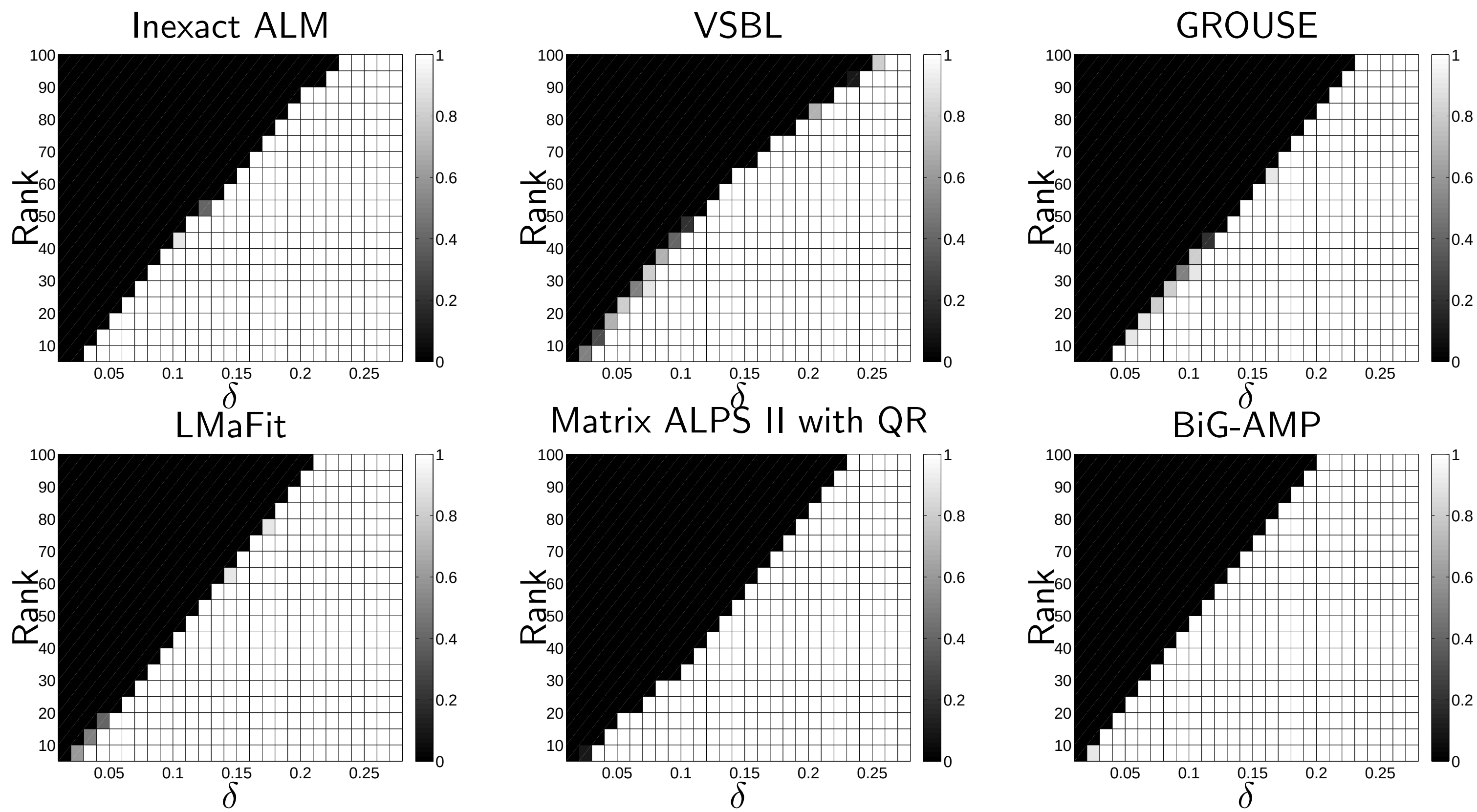
- [Kim/Yedla/Pfister'10]
  - All quantities are **discrete**.
- [Keshavan/Montanari'11] (1 page poster only!)
  - Variable nodes are vector-valued; updates involve **matrix inversion?**

BiG-AMP Comments:

- Low computational **cost**
  - Dominated by 8 matrix multiplies per iteration
  - Sparse matrix math $\longrightarrow$ cost per multiply $\mathcal{O}(R|\Omega|)$
  - Uniform variances $\longrightarrow$ eliminates 5 matrix multiplies per iteration
  - Sparse MM + Uniform variances + Gaussian priors $\longrightarrow$ **BiG-AMP Lite**
- Adaptive **stepsize** scheme based on **GAMP** work
- EM **hyperparameter learning** using BiG-AMP for the "E" step
- Many **extensions** to pursue:
  - quantized outputs (e.g., Netflix ratings)
  - non-negativity constraints (e.g., pmf)
  - structure (e.g., tree-structured dictionaries)
  - linear (not missing) observations
  - etc, etc, etc. . .
- **Theoretical analysis/guarantees?**

Matrix Completion — Phase Transitions:

For $M \times N = 1000 \times 1000$ matrices in the absence of noise, median over 10 trials:

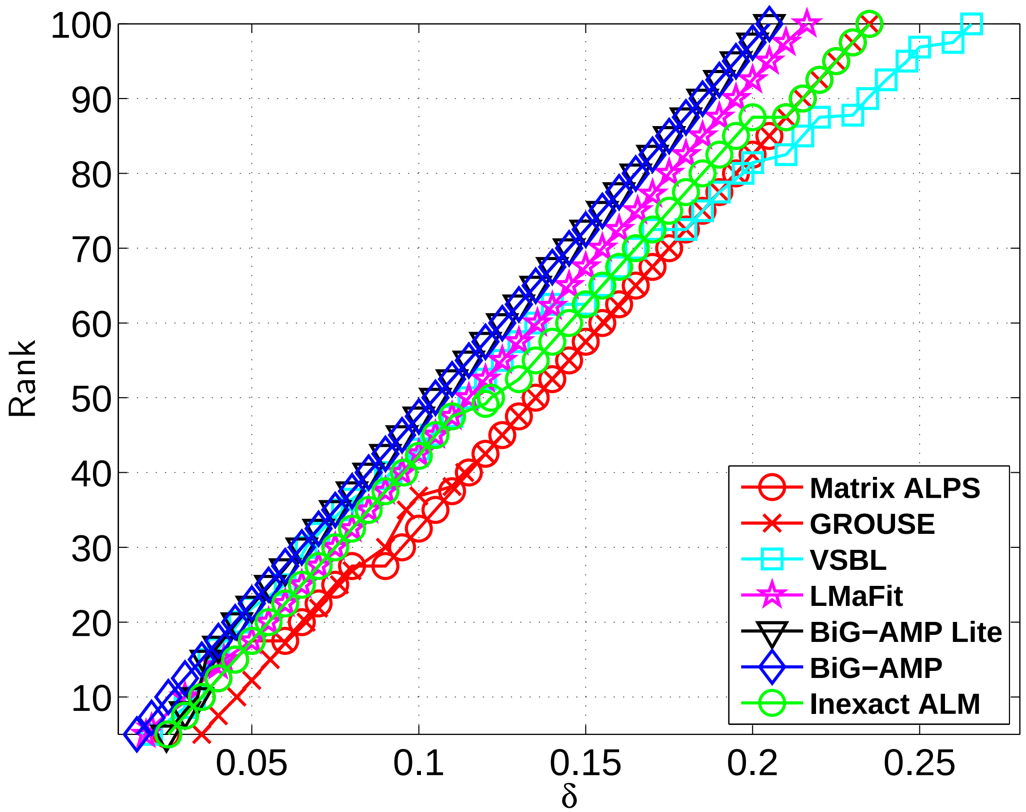


where

- $\delta \triangleq$ fraction of observed entries.

Matrix Completion — Phase Transitions, 50% Contours:

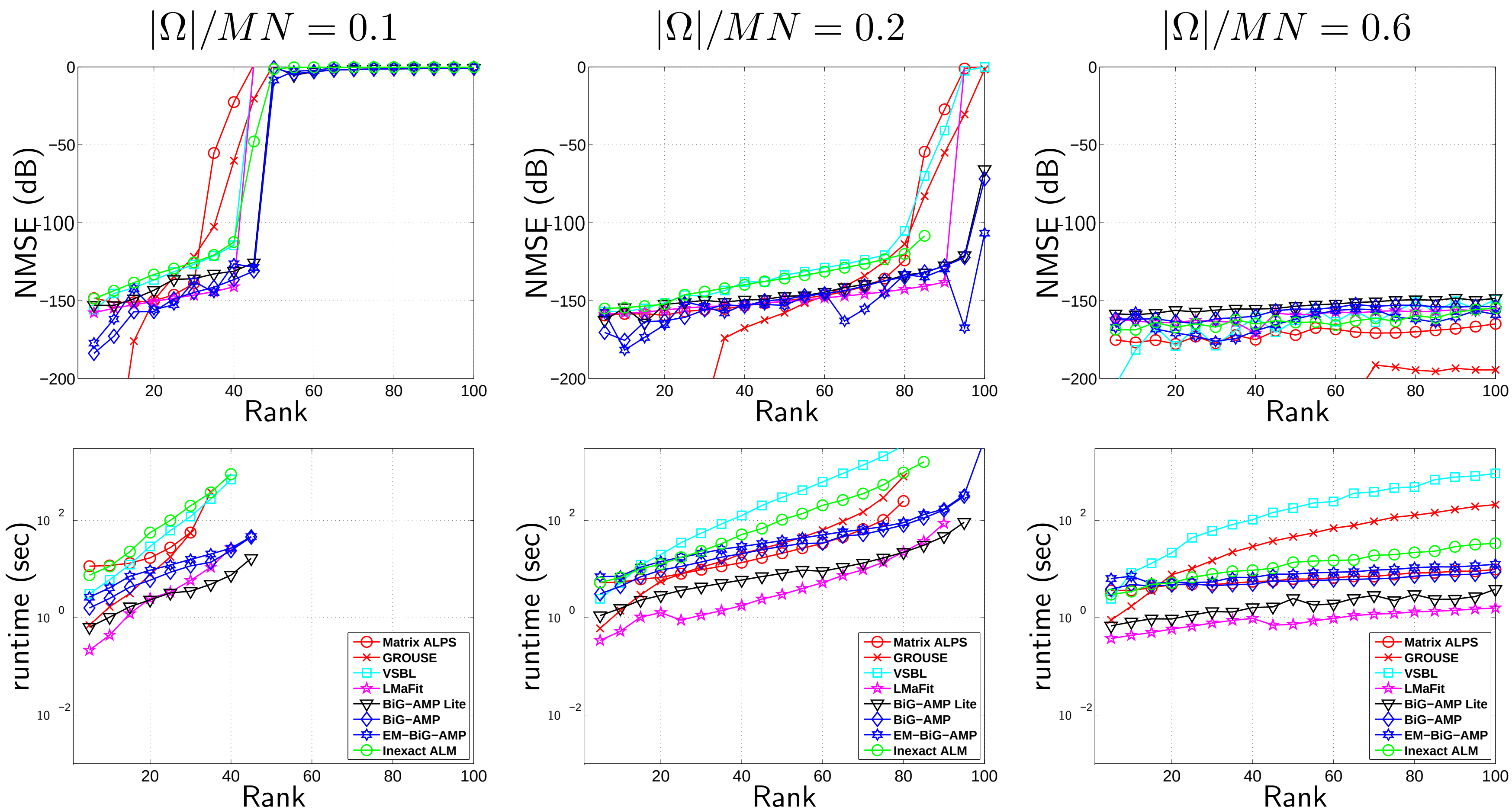
For $M \times N = 1000 \times 1000$ matrices in the absence of noise, median over 10 trials:



BiG-AMP achieves the best phase transition in this test

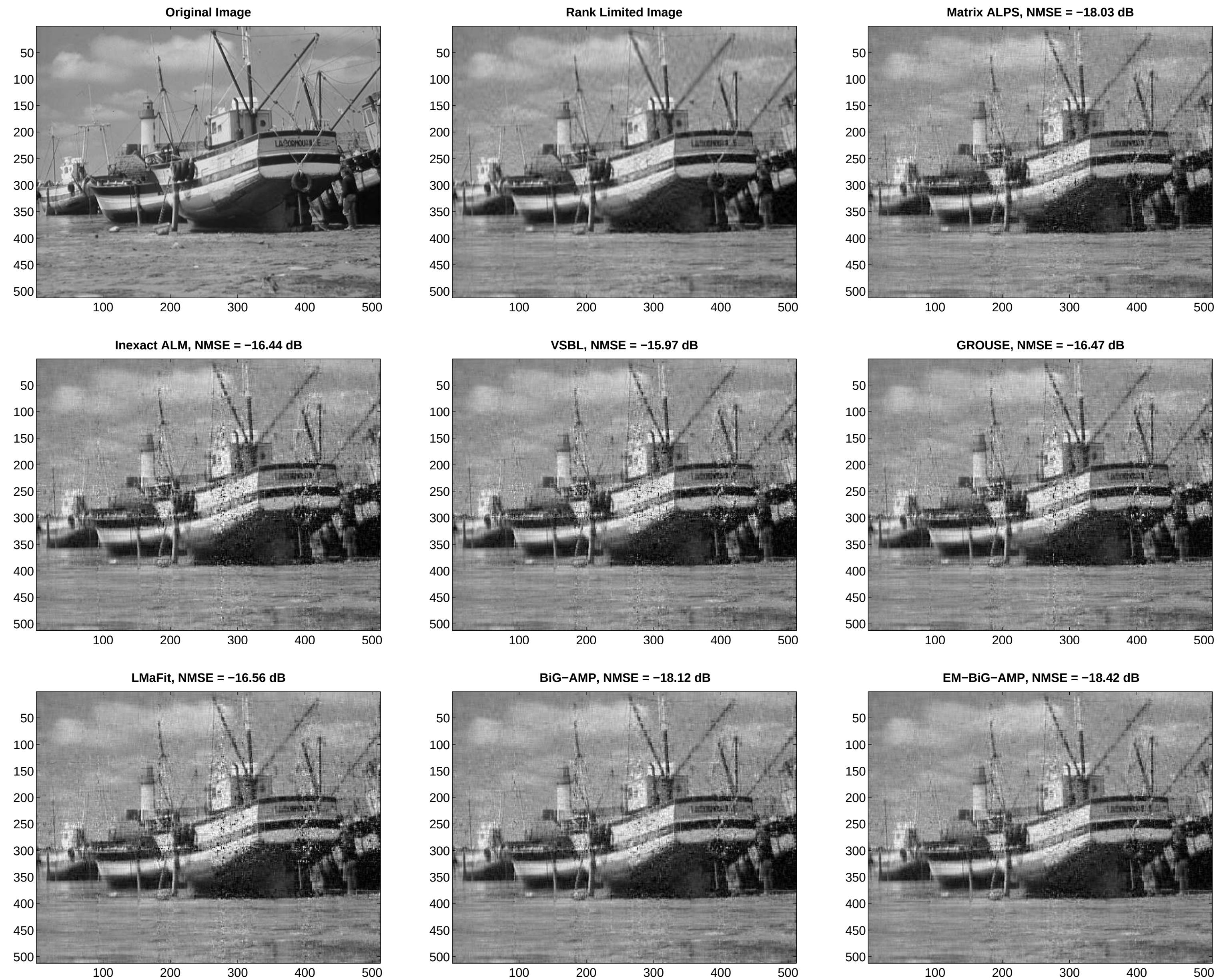
Matrix Completion — NMSE and Runtime (to -100 dB):

(vertical slices of phase plane)

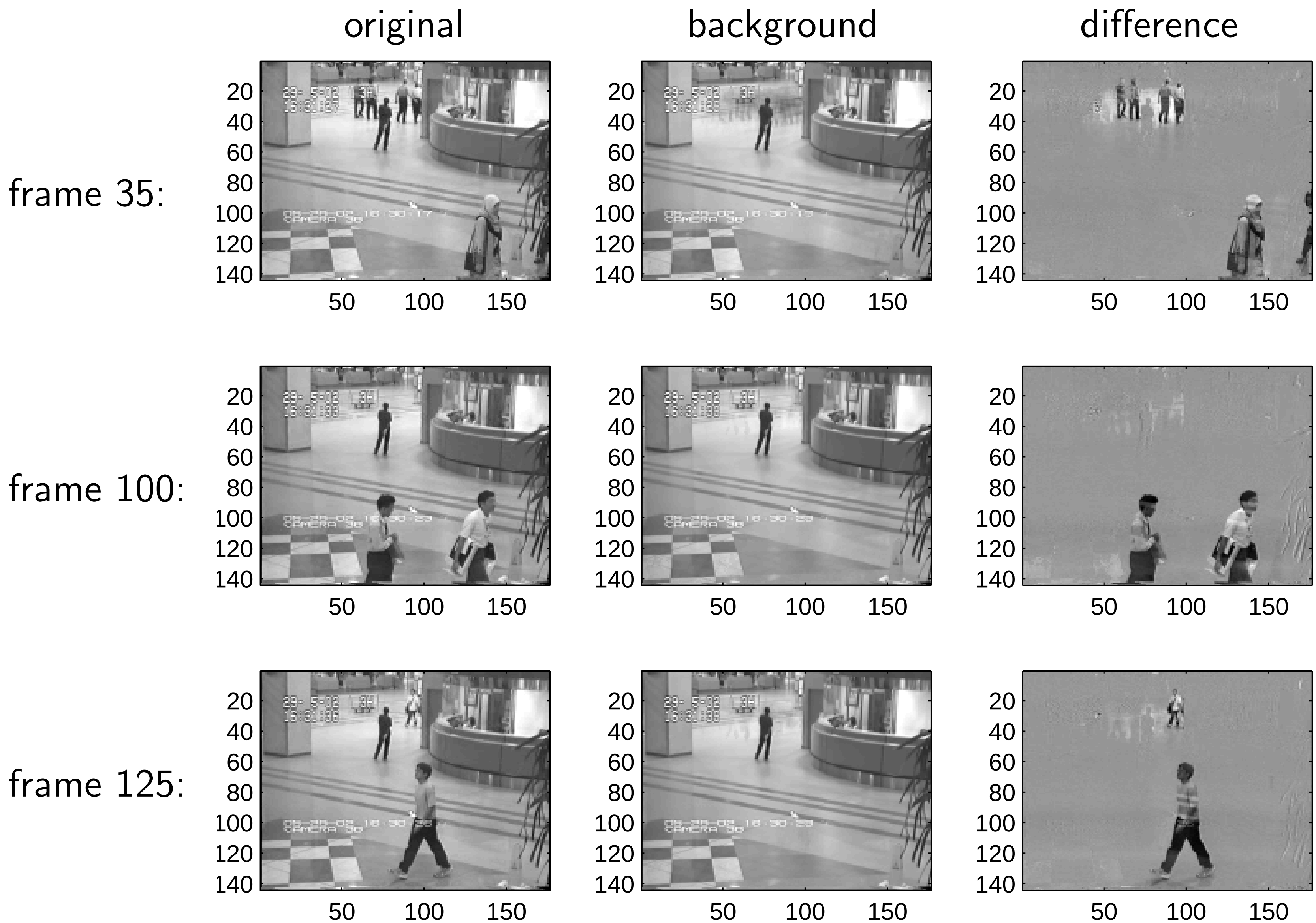


BiG-AMP achieves very high accuracy and is faster than most approaches.

BiG-AMP Lite is competitive with the fastest techniques.


Original Image
Rank Limited Image
Matrix ALPS, NMSE = −18.03 dB
Inexact ALM, NMSE = −16.44 dB
VSBL, NMSE = −15.97 dB
GROUSE, NMSE = −16.47 dB
LMaFit, NMSE = −16.56 dB
BiG−AMP, NMSE = −18.12 dB
EM−BiG−AMP, NMSE = −18.42 dB
50 100 150 200 250 300 350 400 450 500
100 200 300 400 500

Robust PCA — Video Surveillance (over 200 frames):



Conclusions:

BiG-AMP is . . .

- Approximate message passing (**AMP**) for the **generalized bilinear** model.
- A **unified** approach to many **matrix-recovery** problems (MC, RPCA, DL. . .)
- **Competitive** with the best algorithms for each application.

Ongoing Work

- Rank learning
- EM learning for RPCA and DL
- DL applications
  - Hyperspectral imaging (with J. Vila and J. Meola)
  - Topic modeling (with S. Som)
- Parametric BiG-AMP