Bilinear Generalized Approximate Message Passing (BiG-AMP) for Dictionary Learning

Phil Schniter

THE OHIO STATE UNIVERSITY

Collaborators: Jason Parker @OSU, Jeremy Vila @OSU, and Volkan Cehver @EPFL

With support from NSF CCF-1218754, NSF CCF-1018368, NSF IIP-0968910, and DARPA/ONR N66001-10-1-4090

ITA — February 2014

Dictionary Learning

Problem objective:

Recover (possibly overcomplete) dictionary $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ and sparse matrix $\boldsymbol{X} \in \mathbb{R}^{N \times L}$ from (possibly noise-corrupted) observations $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{W}$.

Possible generalizations:

- non-additive corruption (e.g., one-bit or phaseless $\boldsymbol{Y}$)
- incomplete/missing observations
- structured sparsity
- non-negative $\boldsymbol{A}$ and $\boldsymbol{X}$, or simplex-constrained

Contributions

- We propose a unified approach to these dictionary-learning problems that leverages the recent framework of approximate message passing (AMP).

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

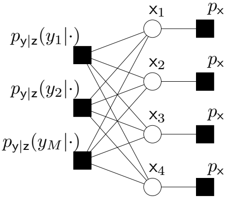
  or the generalized linear model:
  - Infer $\boldsymbol{x} \sim \prod_n p_{\mathsf{x}}(x_n)$ from $\boldsymbol{y} \sim \prod_m p_{\mathsf{y}|\mathsf{z}}(y_m|z_m)$ with hidden $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{x}$ and known $\boldsymbol{\Phi}$. [Rangan'10]

  our work tackles the generalized *bilinear* model:
  - Infer $\boldsymbol{A} \sim \prod_{m,n} p_{\mathsf{a}}(a_{mn})$ and $\boldsymbol{X} \sim \prod_{n,l} p_{\mathsf{x}}(x_{nl})$ from $\boldsymbol{Y} \sim \prod_{m,l} p_{\mathsf{y}|\mathsf{z}}(y_{ml}|z_{ml})$ with hidden $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X}$. [Schniter/Cevher'11]

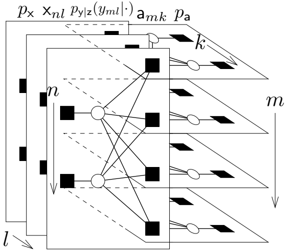
- In addition, we propose methods to select the rank of $\boldsymbol{Z}$, to estimate the parameters of $p_{\mathsf{a}}, p_{\mathsf{x}}, p_{\mathsf{y}|\mathsf{z}}$, and to handle non-separable priors on $\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}$.

Bilinear Generalized AMP (BiG-AMP)

Generalized Linear:



Generalized Bilinear:



- In AMP, beliefs are propagated on a loopy factor graph using approximations that exploit certain blessings of dimensionality:
  1. Gaussian message approximation (motivated by central limit theorem),
  2. Taylor-series approximation of message differences.

- Rigorous analyses of GAMP for CS (with large iid sub-Gaussian Φ) reveal a state evolution whose fixed points are optimal when unique. [Javanmard/Montanari'12]

Adaptive Damping

- The heuristics used to derive BiG-AMP hold in the large system limit: $M, N, L \to \infty$ with $\frac{M}{N} \to \delta$ and $\frac{M}{L} \to \gamma$ for constants $\delta, \gamma \in (0, 1)$.
- In practice, M, N, L are finite and the rank N is often very small!
- To prevent divergence, we damp the updates using an adjustable parameter $\beta \in (0, 1]$.
- Moreover, we adapt β by monitoring (an approximation to) the cost function minimized by BiG-AMP and adjusting β as needed to ensure decreasing cost.

$$\begin{aligned}\hat{J}(t) = &\sum_{n,l} D\Big(\hat{p}_{\mathsf{x}_{nl}|\mathbf{Y}}(\cdot \mid \boldsymbol{Y}) \Big\| p_{\mathsf{x}_{nl}}(\cdot)\Big) \quad \leftarrow \text{KL divergence between posterior \& prior} \\ &+ \sum_{m,n} D\Big(\hat{p}_{\mathsf{a}_{mn}|\mathbf{Y}}(\cdot \mid \boldsymbol{Y}) \Big\| p_{\mathsf{a}_{mn}}(\cdot)\Big) \\ &- \sum_{m,l} \mathrm{E}_{\mathcal{N}(\mathsf{z}_{ml};\bar{p}_{ml}(t);\nu^p_{ml}(t))}\big\{\log p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml} \mid \mathsf{z}_{ml})\big\}.\end{aligned}$$

Parameter Tuning via EM

- AMP methods assume $p_{\mathsf{x}}, p_{\mathsf{a}}, p_{\mathsf{y}|\mathsf{z}}$ are known, which is rarely true in practice.
- We assume families for these priors (e.g., Gaussian mixture) and estimate the associated parameters $\boldsymbol{\theta}$ using expectation-maximization (EM), as done for GAMP in [Vila/Schniter'13].

- Taking $\boldsymbol{X}$, $\boldsymbol{A}$, and $\boldsymbol{Z}$ to be the hidden variables, the EM recursion becomes

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{k+1} &= \arg\max_{\boldsymbol{\theta}} \mathrm{E}\left\{ \log p_{\boldsymbol{X},\boldsymbol{A},\boldsymbol{Z},\boldsymbol{Y}}(\boldsymbol{X},\boldsymbol{A},\boldsymbol{Z},\boldsymbol{Y};\boldsymbol{\theta}) \,\middle|\, \boldsymbol{Y};\hat{\boldsymbol{\theta}}^{k} \right\} \\
&= \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{n,l} \mathrm{E}\left\{ \log p_{\mathsf{x}_{nl}}(\mathsf{x}_{nl};\boldsymbol{\theta}) \,\middle|\, \boldsymbol{Y};\hat{\boldsymbol{\theta}}^{k} \right\} \right. \\
&\qquad + \sum_{m,n} \mathrm{E}\left\{ \log p_{\mathsf{a}_{mn}}(\mathsf{a}_{mn};\boldsymbol{\theta}) \,\middle|\, \boldsymbol{Y};\hat{\boldsymbol{\theta}}^{k} \right\} \\
&\qquad \left. + \sum_{m,l} \mathrm{E}\left\{ \log p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml} \,|\, \mathsf{z}_{ml};\boldsymbol{\theta}) \,\middle|\, \boldsymbol{Y};\hat{\boldsymbol{\theta}}^{k} \right\} \right\}
\end{aligned}
$$

- For tractability, the $\boldsymbol{\theta}$-maximization is performed one variable at a time.

Numerical Results for Dictionary Learning

We compared against several state-of-the-art techniques

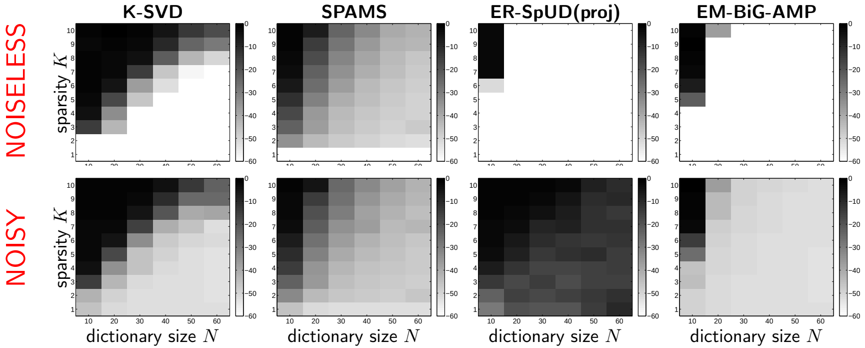
- K-SVD [Aharon/Elad/Bruckstein'06]
  - the standard; a generalization of K-means clustering
- SPAMS [Mairal/Bach/Ponce/Sapiro'10]
  - a highly optimized online approach
- ER-SpUD [Spielman/Wang/Wright'12]
  - the recent breakthrough on provable square-dictionary recovery

to our proposed technique:

- EM-BiG-AMP
  - BiG-AMP under AWGN, $\mathcal{BG}$ signal, and EM-adjusted $\lambda, \mu_{\mathsf{x}}, v_{\mathsf{x}}, v_{\mathsf{w}}$.

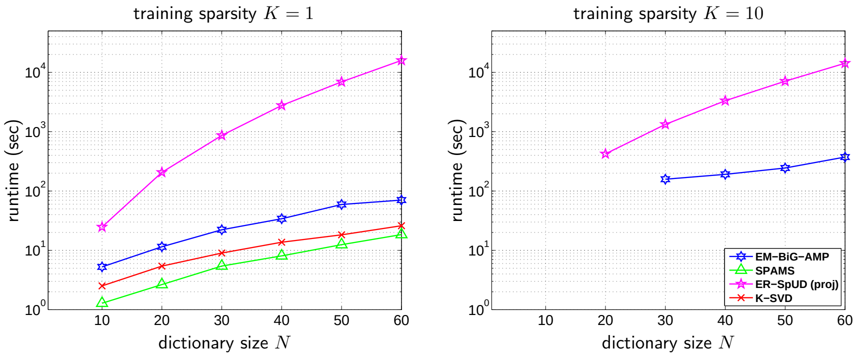
Square Dictionary Recovery: Phase Transitions

Mean NMSE over 10 realizations for recovery of an $N \times N$ dictionary from $L = 5N \log N$ examples with sparsity K:



- Noiseless case: EM-BiG-AMP's phase transition curve is much better than that of K-SVD and SPAMS and almost as good as ER-SpUD(proj)'s.
- Noisy case: EM-BiG-AMP is robust to noise, while ER-SpUD(proj) is not.

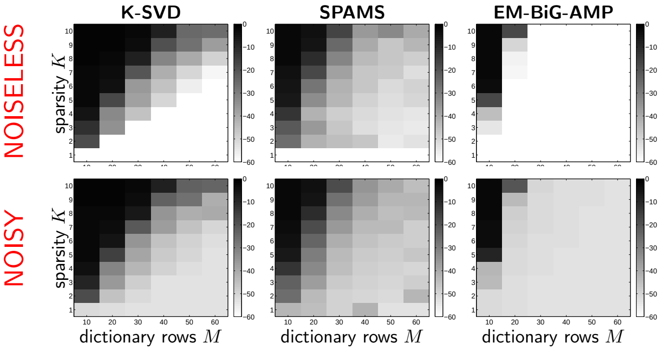
Square Dictionary Recovery: Runtime to NMSE=-60 dB



- EM-BiG-AMP runs within a factor-of-5 from the fastest approach (SPAMS).
- EM-BiG-AMP runs orders-of-magnitude faster than ER-SpUD(proj).

Overcomplete Dictionary Recovery: Phase Transitions

Mean NMSE over 10 realizations for recovery of an $M \times (2M)$ dictionary from $L = 5N \log N$ examples with sparsity K:



- Noiseless case: EM-BiG-AMP's phase transition curve is much better than that of K-SVD and SPAMS. Note: ER-SpUD not applicable when $M \neq N$.
- Noisy case: EM-BiG-AMP is again robust to noise.

Application: Hyperspectral Unmixing

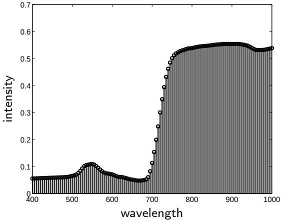
- In Hyperspectral unmixing, a sensor captures M wavelengths per pixel, over a scene of L pixels comprised of N materials.

- The received HSI data $\boldsymbol{Y}$ is modeled as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{W} \in \mathbb{R}_+^{M \times L},$$

where the nth column of $\boldsymbol{A} \in \mathbb{R}_+^{M \times N}$ is the spectrum of the nth material, the lth column of $\boldsymbol{X} \in \mathbb{R}_+^{N \times L}$ describes the abundance of materials at the lth pixel (and thus must sum to one), and $\boldsymbol{W}$ is additive noise.

spectrum at one pixel

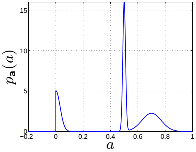


- The goal is to jointly estimate $\boldsymbol{A}$ and $\boldsymbol{X}$.
  - Standard NMF-based unmixing algs (e.g., VCA [Nascimento'05], FSNMF [Gillis'12]) assume pure-pixels, which may not occur in practice.
  - Furthermore, they do *not* exploit spectral coherence, spatial coherence, and sparsity, which do occur in practice.
  - Recent Bayesian approaches to unmixing (e.g., SCU [Mittelman'12]) exploit spatial coherence using Dirichlet processes, albeit at very high complexity.

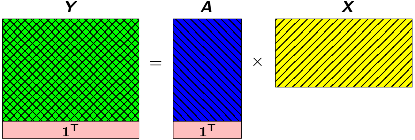
EM-BiG-AMP for HSI Unmixing

- To enforce non-negativity we place non-negative Gaussian Mixture (NNGM) prior on a_{mn}, and to encourage sparsity a Bernoulli-NNGM prior on x_{nl}.
  - We then use EM to learn the (B)NNGM parameters.
- To enforce the sum-to-one constraint on each column of $\boldsymbol{X}$, we augment both $\boldsymbol{Y}$ and $\boldsymbol{A}$ with a row of random variables with mean one and variance zero.
- To exploit spectral coherence we employ a hidden Gauss-Markov chain across each column in $\boldsymbol{A}$, and to exploit spatial coherence we employ an Ising model to capture the support across each row in $\boldsymbol{X}$.
  - We use EM to learn the Gauss-Markov and Ising parameters.

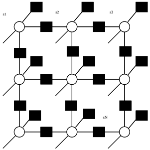
NNGM prior



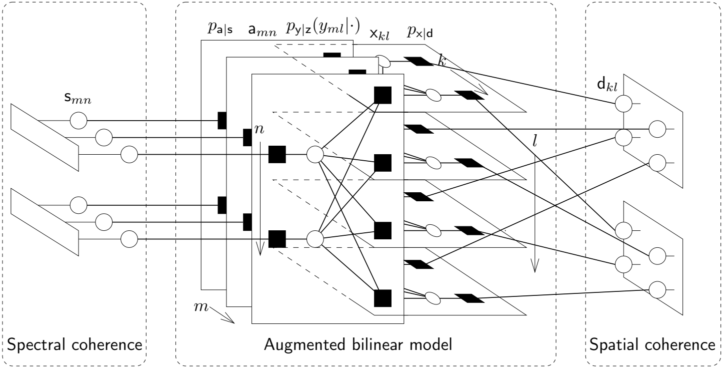
Augmented model:



Ising model



EM-BiG-AMP for HSI Unmixing



- Inference on the bilinear sub-graph is tackled using the BiG-AMP algorithm.
- Inference on the Gauss-Markov and Ising subgraphs are tackled using standard soft-input/soft-output belief propagation methods.
- Messages are exchanged between the three sub-graphs according to the sum-product algorithm, akin to "turbo" decoding in modern communication receivers [Schniter'10].

Numerical Results: Pure-Pixel Synthetic Data

- Pure pixel abundance maps $\boldsymbol{X}$ of size $L = 50 \times 50$ were generated with $N = 5$ materials residing in equal-sized spatial strips.
- Endmember spectra $\boldsymbol{A}$ were taken from a reflectance library.
- AWGN observations with SNR $= 30$ dB.

RGB view of data in 2D

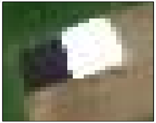
- Averaging performance over 10 realizations . . .

	Runtime	NMSE$_{\boldsymbol{S}}$	NMSE$_{\boldsymbol{A}}$
EM-BiG-AMP	5.57 sec	**-57.4** dB	**-108.6 dB**
VCA + FCLS	4.13 sec	-39.6 dB	-30.5 dB
FSNMF + FCLS	**3.97 sec**	-25.3 dB	-12.5 dB
SCU	2808 sec	-30.6 dB	-20.5 dB

EM-BiG-AMP gives significantly better NMSE than competing algorithms.

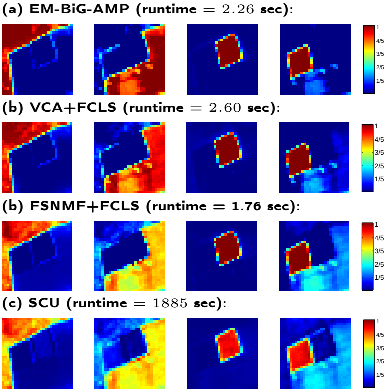
- EM-BiG-AMP's gives runtime comparable to the fastest algorithms and 3 orders-of-magnitude faster than SCU.

Results: SHARE 2012 dataset



RGB image from the SHARE 2012 dataset.

- Experiment constructed to provide pure pixels.
- EM-BiG-AMP yields the purest abundances (right).
- EM-BiG-AMP yields the best spectral angles (below).
- EM-BiG-AMP's runtime is on par with the fastest algorithm, FSNMF+FCLS.



$N = 4$ material abundance maps.

	grass	dry sand	white TyVek	black felt
EM-BiG–AMP	**0.999**	**0.999**	**1.000**	**0.998**
VCA + FCLS	**0.999**	**0.999**	0.999	0.981
FSNMF + FCLS	**0.999**	0.997	**1.000**	0.977
SCU	**0.999**	**0.999**	0.999	0.859

Spectral Angle Distance (SAD) between recovered and ground truth endmembers.

Conclusion

- BiG-AMP = approximate message passing for the generalized bilinear model.
- A novel approach to matrix completion, robust PCA, dictionary learning, etc.
- Includes mechanisms for adaptive damping, parameter tuning, non-separable priors, and model-order selection.
- Competitive with state-of-the-art algorithms for each application.
  - Best phase transitions for MC, RPCA, overcomplete DL.
  - Runtimes not far from the fastest algorithms.
- Currently working on generalizations of BiG-AMP to parametric models (e.g., Toeplitz matrices), as well as various applications.

References

1. J. T. Parker, P. Schniter and V. Cevher, "Bilinear Generalized Approximate Message Passing," *arXiv:1310.2632*, 2013.
2. D.L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," *ITW*, 2010.
3. S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," *ISIT*, 2011. (See also *arXiv:1010.5141*).
4. P. Schniter and V. Cevher, "Approximate message passing for bilinear models," *SPARS*, 2011.
5. A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *arXiv:1211.5164*, 2012.
6. J. P. Vila and P. Schniter, "Expectation-Maximization Gaussian-Mixture Approximate Message Passing," *IEEE Trans. Signal Process.*, 2013.
7. P. Schniter, "Turbo reconstruction of structured sparse signals," *Proc. CISS*, 2010.
8. M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Sig. Process.*, 2006.
9. J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, 2010.
10. D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," *J. Mach. Learn. Res.*, 2012.
11. J. Nascimento and J. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. GeoSci. Remote Sens.*, 2005.
12. N. Gillis and S.A. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *arXiv:1208.1237*, 2012.
13. R. Mittelman, N. Dobigeon, and A. Hero, "Hyperspectral image unmixing using a multiresolution sticky HDP," *IEEE Trans. Signal Process.*, 2012.
14. J. Vila, P. Schniter, and J. Meola, "Hyperspectral Image Unmixing via Bilinear Generalized Approximate Message Passing," Proc. SPIE, 2013.