Bilinear Recovery using EM Vector-AMP

Phil Schniter and Subrata Sarkar

The Ohio State University

Collaborators: **Sundeep Rangan** @ NYU and **Allie Fletcher** @ UCLA

With support from: NSF CCF-1527162 and NSF CCF-1716388

ITA Workshop (San Diego) — Feb 12, 2018

Bilinear Recovery

Goal: Recover signal $\mathbf{X}$ and parameters $\boldsymbol{\theta}_A$ from noisy measurements

$$\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}_A)\mathbf{X} + \mathbf{W} \quad \text{with affine linear map } \mathbf{A}(\cdot).$$

Applications:

- Self calibration[1]
- Compressed sensing with matrix uncertainty[2]
- Blind deconvolution[3]
- Dictionary learning[4]
- Joint channel estimation and symbol detection

[1] Ling,Strohmer'15 [2] Zhu,Leus,Giannakis'11 [3] Ahmed,Recht,Romberg'12 [4] Aharon,Elad,Bruckstein'06

Statistical Model

Measurements: $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}_A)\mathbf{X} + \mathbf{W}$

Assumptions:

- $\mathbf{A}(\cdot) : \mathbb{R}^Q \to \mathbb{R}^{M\times N}$ measurement operator
- $\mathbf{X} \in \mathbb{R}^{N\times L}$ with $x_{nl} \overset{\text{i.i.d}}{\sim} p_X(\cdot;\boldsymbol{\theta}_X)$ random signal
- $\mathbf{W} \in \mathbb{R}^{M\times L}$ with $w_{ml} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \theta_w^{-1})$ AWGN
- $\boldsymbol{\Theta} \triangleq \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_X, \theta_w\}$ unknown deterministic parameters

Goal: compute...

- $\widehat{\boldsymbol{\Theta}}_{\text{ML}} = \arg\max_{\boldsymbol{\Theta}} p(\mathbf{Y};\boldsymbol{\Theta})$ maximum likelihood
- $\widehat{\mathbf{X}}_{\text{MMSE}} = \mathbb{E}[\mathbf{X}|\mathbf{Y};\widehat{\boldsymbol{\Theta}}_{\text{ML}}]$ "empirical Bayes"

Related Work: $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_X, \theta_w\}$ Known

Consider the case where $\boldsymbol{\Theta} \triangleq \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_X, \theta_w\}$ (and thus $\mathbf{A}$, p_X, p_W) are known.

- The "vector AMP" (VAMP) algorithm[5] can be applied.
- When $\mathbf{A}$ is a large right-rotationally invariant random matrix, the macroscopic behavior of VAMP is rigorously characterized by a scalar state-evolution.[5]
- When the state-evolution has a unique fixed point, it is "good" in the sense that VAMP's MSE agrees with the replica prediction[6] of the MMSE.
- VAMP is more robust than AMP[7], which requires large i.i.d sub-Gaussian $\mathbf{A}$.

[5] Rangan,Schniter,Fletcher'16 [6] Tulino,Caire,Verdú,Shamai'13 [7] Donoho,Maleki,Montanari'10

Related Work: $\boldsymbol{\theta}_A$ Known, $\{\boldsymbol{\theta}_X, \theta_w\}$ Unknown

Now consider case where $\boldsymbol{\theta}_A$ (and thus $\mathbf{A}$) is known, but $\boldsymbol{\theta}_X$ and θ_w are not.

- The EM-VAMP algorithm[8] can be applied.
- When $\mathbf{A}$ is a large right-rotationally invariant random matrix, the macroscopic behavior of EM-VAMP is rigorously characterized by a state-evolution.[4]
- For certain classes of p_X (e.g., exponential family), EM-VAMP's parameter estimates $\{\widehat{\boldsymbol{\theta}}_X, \widehat{\theta}_w\}$ are asymptotically consistent.[4]

[8] Fletcher,Rangan,Schniter'17

Related Work: $\{\boldsymbol{\theta}_A, \boldsymbol{\theta}_X, \theta_w\}$ Unknown

Finally, consider the case of interest, where $\boldsymbol{\theta}_A$, $\boldsymbol{\theta}_X$ and θ_w are unknown.

- Suppose we have $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}_A)\mathbf{X} + \mathbf{W}$ with affine linear map $\mathbf{A}(\cdot)$.
- The Parametric Bilinear Generalized AMP (P-BiG-AMP) algorithm[9][10] can be used to recover $\boldsymbol{\theta}_A$ and $\mathbf{X} \overset{\text{i.i.d}}{\sim} p_x(\cdot; \boldsymbol{\theta}_X)$.
- But P-BiG-AMP is based on an i.i.d. Gaussian model for $\mathbf{A}(\cdot)$, and it may not perform well when $\mathbf{A}(\cdot)$ deviates from that model!

[9] Parker,Schniter'16 [10] Schulke,Schniter,Zdeborova'16

This Work: $\{\boldsymbol{\theta}_A, \boldsymbol{\theta}_X, \theta_w\}$ Unknown

In the next few slides, we will outline the EM-VAMP methodology:

1. Inference via Expectation Consistent Approximation (EC)
2. Algorithmic implementation via VAMP
3. Learning $\boldsymbol{\Theta}$ via Expectation Maximization (EM)
4. Joint inference & learning via EM-VAMP

Variational Inference

For the moment, let's suppose that $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_X, \theta_w\}$ is known.

- Ideally, we would like to compute the exact posterior density

$$p(\mathbf{X}|\mathbf{Y};\boldsymbol{\Theta}) = \frac{p(\mathbf{X};\boldsymbol{\Theta})p(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta})}{Z(\boldsymbol{\Theta})} \quad \text{for} \quad Z(\boldsymbol{\Theta}) \triangleq \int p(\mathbf{X};\boldsymbol{\Theta})p(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta})\,\mathrm{d}\mathbf{X},$$

  but the high-dimensional integral in $Z(\boldsymbol{\Theta})$ is difficult to compute.

- We can avoid computing $Z(\boldsymbol{\Theta})$ through variational optimization:

$$\begin{aligned}
p(\mathbf{X}|\mathbf{Y};\boldsymbol{\Theta}) &= \arg\min_b D\big(b(\mathbf{X})\big\|p(\mathbf{X}|\mathbf{Y};\boldsymbol{\Theta})\big) \text{ where } D(\cdot\|\cdot) \text{ is KL divergence} \\
&= \arg\min_b \underbrace{D\big(b(\mathbf{X})\big\|p(\mathbf{X};\boldsymbol{\Theta})\big) + D\big(b(\mathbf{X})\big\|p(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta})\big) + H\big(b(\mathbf{X})\big)}_{\text{Gibbs free energy}} \\
&= \arg\min_{b,\bar{b},q} \underbrace{D\big(b(\mathbf{X})\big\|p(\mathbf{X};\boldsymbol{\Theta})\big) + D\big(\bar{b}(\mathbf{X})\big\|p(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta})\big) + H\big(q(\mathbf{X})\big)}_{\triangleq J_{\text{Gibbs}}(b,\bar{b},q;\boldsymbol{\Theta})} \\
&\quad \text{s.t. } b = \bar{b} = q,
\end{aligned}$$

  but the density constraint keeps the problem difficult.

Expectation Consistent Approximation

- In expectation-consistent approximate inference (EC),[11] the density constraint $b = \bar{b} = q$ is relaxed to moment-matching constraints:

$$p(\mathbf{X}|\mathbf{Y};\boldsymbol{\Theta}) \approx \underset{b,\bar{b},q}{\arg\min}\, J_{\mathsf{Gibbs}}(b,\bar{b},q;\boldsymbol{\Theta})$$

$$\text{s.t.} \begin{cases} \mathbb{E}\{\mathbf{x}_l|b\} = \mathbb{E}\{\mathbf{x}_l|\bar{b}\} = \mathbb{E}\{\mathbf{x}_l|q\} \ \forall l \\ \mathrm{tr}(\mathrm{Cov}\{\mathbf{x}_l|b\}) = \mathrm{tr}(\mathrm{Cov}\{\mathbf{x}_l|\bar{b}\}) = \mathrm{tr}(\mathrm{Cov}\{\mathbf{x}_l|q\}) \ \forall l. \end{cases}$$

- The stationary points of EC are the densities

$$\begin{aligned} b(\mathbf{X}) &\propto \textstyle\prod_{l=1}^{L} p(\mathbf{x}_l;\boldsymbol{\Theta})\mathcal{N}(\mathbf{x}_l;\mathbf{r}_l,\mathbf{I}/\gamma_l) \\ \bar{b}(\mathbf{X}) &\propto \textstyle\prod_{l=1}^{L} p(\mathbf{y}_l|\mathbf{x}_l;\boldsymbol{\Theta})\mathcal{N}(\mathbf{x}_l;\bar{\mathbf{r}}_l,\mathbf{I}/\bar{\gamma}_l) \\ q(\mathbf{X}) &= \textstyle\prod_{l=1}^{L} \mathcal{N}(\mathbf{x}_l;\widehat{\mathbf{x}}_l,\mathbf{I}/\eta_l) \end{aligned}$$

  with parameters $\{\mathbf{R},\boldsymbol{\gamma},\overline{\mathbf{R}},\overline{\boldsymbol{\gamma}},\widehat{\mathbf{X}},\boldsymbol{\eta}\}$ such that

$$\mathbb{E}\{\mathbf{x}_l|b\} = \mathbb{E}\{\mathbf{x}_l|\bar{b}\} = \widehat{\mathbf{x}}_l \ \forall l$$

$$\tfrac{1}{N}\,\mathrm{tr}(\mathrm{Cov}\{\mathbf{x}_l|b\}) = \tfrac{1}{N}\,\mathrm{tr}(\mathrm{Cov}\{\mathbf{x}_l|\bar{b}\}) = 1/\eta_l \ \forall l.$$

[11] Opper,Winther'04

The VAMP Algorithm

An iterative approach to finding $\{\mathbf{R}, \boldsymbol{\gamma}, \overline{\mathbf{R}}, \overline{\boldsymbol{\gamma}}, \widehat{\mathbf{X}}, \boldsymbol{\eta}\}$:

Initialize $\{\mathbf{R}, \boldsymbol{\gamma}\}$ and select the estimation functions

$$\mathbf{g}(\mathbf{r}_l; \gamma_l) = \mathbb{E}\{\mathbf{x}_l | b;\ \mathbf{r}_l, \gamma_l\}$$

$$\overline{\mathbf{g}}(\overline{\mathbf{r}}_l; \overline{\gamma}_l) = \mathbb{E}\{\mathbf{x}_l | \bar{b};\ \overline{\mathbf{r}}_l, \overline{\gamma}_l\}.$$

For $t = 1, 2, 3, \ldots$

$$\widehat{\mathbf{x}}_l \leftarrow \mathbf{g}(\mathbf{r}_l; \gamma_l),\ \forall l \qquad \text{MMSE estimation}$$

$$\eta_l \leftarrow \gamma_l N / \operatorname{tr}\left[\partial \mathbf{g}(\mathbf{r}_l; \gamma_l) / \partial \mathbf{r}_l\right],\ \forall l$$

$$\overline{\mathbf{r}}_l \leftarrow (\eta_l \widehat{\mathbf{x}}_l - \gamma_l \mathbf{r}_l)/(\eta_l - \gamma_l),\ \forall l \qquad \text{pseudo-measurement}$$

$$\overline{\gamma}_l \leftarrow \eta_l - \gamma_l,\ \forall l$$

$$\overline{\mathbf{x}}_l \leftarrow \overline{\mathbf{g}}(\overline{\mathbf{r}}_l; \overline{\gamma}_l),\ \forall l \qquad \text{LMMSE estimation}$$

$$\overline{\eta}_l \leftarrow \overline{\gamma}_l N / \operatorname{tr}\left[\partial \overline{\mathbf{g}}(\overline{\mathbf{r}}_l; \overline{\gamma}_l) / \partial \overline{\mathbf{r}}_l\right],\ \forall l$$

$$\mathbf{r}_l \leftarrow (\overline{\eta}_l \overline{\mathbf{x}}_l - \overline{\gamma}_l \overline{\mathbf{r}}_l)/(\overline{\eta}_l - \overline{\gamma}_l),\ \forall l \qquad \text{pseudo-prior}$$

$$\gamma_l \leftarrow \overline{\eta}_l - \overline{\gamma}_l,\ \forall l$$

Note: this specialization of VAMP is equivalent to expectation propagation (EP).

Expectation Maximization

We now return to the case where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_X, \theta_w\}$ is unknown.

- The EM algorithm is a well-known iterative approach to maximum-likelihood estimation of $\boldsymbol{\Theta}$.
- The EM algorithm can be written in terms of the Gibbs free energy as[12]

$$\widehat{\boldsymbol{\Theta}}^{(t+1)} = \arg\min_{\boldsymbol{\Theta}} \underbrace{D_{KL}(b^{(t)}(\mathbf{X})||p(\mathbf{X};\boldsymbol{\Theta})) + D_{KL}(b^{(t)}(\mathbf{X})||p(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta})) + H(b^{(t)})}_{= J_{\text{Gibbs}}(b^{(t)}, b^{(t)}, b^{(t)}; \boldsymbol{\Theta})}$$

  using the belief $b^{(t)} \triangleq p(\mathbf{X}|\mathbf{Y}; \widehat{\boldsymbol{\Theta}}^{(t)})$

- Thus EM and VAMP can be combined to solve

$$\min_{\boldsymbol{\Theta}} \min_{b,\bar{b},q} J_{\text{Gibbs}}(b, \bar{b}, q; \boldsymbol{\Theta}) \text{ s.t. } \begin{cases} \mathbb{E}\{\mathbf{x}_l|b\} = \mathbb{E}\{\mathbf{x}_l|\bar{b}\} = \mathbb{E}\{\mathbf{x}_l|q\} \ \forall l \\ \text{tr}[\text{Cov}\{\mathbf{x}_l|b\}] = \text{tr}[\text{Cov}\{\mathbf{x}_l|\bar{b}\}] = \text{tr}[\text{Cov}\{\mathbf{x}_l|q\}] \ \forall l. \end{cases}$$

[12] Neal,Hinton'98

The EM-VAMP Algorithm

Initialize $\{\mathbf{R}, \boldsymbol{\gamma}, \widehat{\boldsymbol{\Theta}}\}$ and select $\mathbf{g}(\cdot)$ & $\overline{\mathbf{g}}(\cdot)$ as before.

For $t = 1, 2, 3, \ldots$

$\widehat{\mathbf{x}}_l \leftarrow \mathbf{g}(\mathbf{r}_l; \gamma_l, \widehat{\boldsymbol{\Theta}}), \ \forall l$ — MMSE estimation

$\eta_l \leftarrow \gamma_l N / \operatorname{tr}\left[\partial \mathbf{g}(\mathbf{r}_l; \gamma_l, \widehat{\boldsymbol{\Theta}}) / \partial \mathbf{r}_l\right], \ \forall l$

$\overline{\mathbf{r}}_l \leftarrow (\eta_l \widehat{\mathbf{x}}_l - \gamma_l \mathbf{r}_l)/(\eta_l - \gamma_l), \ \forall l$ — pseudo-measurement

$\overline{\gamma}_l \leftarrow \eta_l - \gamma_l, \ \forall l$

$\widehat{\boldsymbol{\Theta}} \leftarrow \arg\max_{\boldsymbol{\Theta}} \mathbb{E}\{\ln p(\mathbf{Y}|\mathbf{X}; \boldsymbol{\Theta}) \,|\, \overline{\mathbf{R}}; \overline{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\Theta}}\}$ — EM update

---

$\overline{\mathbf{x}}_l \leftarrow \overline{\mathbf{g}}(\overline{\mathbf{r}}_l; \overline{\gamma}_l, \widehat{\boldsymbol{\Theta}}), \ \forall l$ — LMMSE estimation

$\overline{\eta}_l \leftarrow \overline{\gamma}_l N / \operatorname{tr}\left[\partial \overline{\mathbf{g}}(\overline{\mathbf{r}}_l; \overline{\gamma}_l, \widehat{\boldsymbol{\Theta}}) / \partial \overline{\mathbf{r}}_l\right], \ \forall l$

$\mathbf{r}_l \leftarrow (\overline{\eta}_l \overline{\mathbf{x}}_l - \overline{\gamma}_l \overline{\mathbf{r}}_l)/(\overline{\eta}_l - \overline{\gamma}_l), \ \forall l$ — pseudo-prior

$\gamma_l \leftarrow \overline{\eta}_l - \overline{\gamma}_l, \ \forall l$

$\widehat{\boldsymbol{\Theta}} \leftarrow \arg\max_{\boldsymbol{\Theta}} \mathbb{E}\{\ln p(\mathbf{X}; \boldsymbol{\Theta}) \,|\, \mathbf{R}; \boldsymbol{\gamma}, \widehat{\boldsymbol{\Theta}}\}$ — EM update

Variance Auto-Tuning

- Problem: The precisions $\{\gamma, \overline{\gamma}\}$ of the pseudo-{measurement,prior} are imperfect when $\widehat{\boldsymbol{\Theta}}$ is imperfect.
- Thus, at each iteration, we estimate these precisions jointly with the unknown parameters $\boldsymbol{\Theta}$. For example, with the prior parameters $\boldsymbol{\theta}_X$:

$$(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}_X) \leftarrow \arg\max_{\boldsymbol{\gamma}, \boldsymbol{\theta}_X} p(\mathbf{R}; \boldsymbol{\gamma}, \boldsymbol{\theta}_X)$$

$$\text{under} \quad \mathbf{r}_l = \mathbf{x}_l + \mathcal{N}(\mathbf{0}, \mathbf{I}/\gamma_l), \quad \mathbf{x}_l \sim p_X(\cdot; \boldsymbol{\theta}_X), \quad \forall l.$$

- In practice, inner iterations of EM can be used to solve the above, e.g.,

$$(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}_X) = \arg\max_{\boldsymbol{\gamma}, \boldsymbol{\theta}_X} \mathbb{E}\Big[\ln p(\mathbf{X}, \mathbf{R}; \boldsymbol{\gamma}, \boldsymbol{\theta}_X) \,\Big|\, \mathbf{R}; \widehat{\gamma}, \widehat{\boldsymbol{\theta}}_X \Big].$$

- This "variance auto-tuning" procedure[13] leads to asymptot. consistent $\widehat{\boldsymbol{\theta}}_X$.

[13] Fletcher,Rangan,Schniter'17

State-Evolution and Consistency

Suppose that

- $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}_A)\mathbf{X} + \mathcal{N}(\mathbf{0}, \mathbf{I}/\theta_w) \in \mathbb{R}^{M\times L}, \quad \mathbf{X} \overset{\text{i.i.d}}{\sim} p_X(\cdot;\boldsymbol{\theta}_X)$
- $\mathbf{A}(\boldsymbol{\theta}) = \mathbf{A}_0 + \sum_{q=1}^{Q} \theta_q \mathbf{A}_q$ with right-rotationally invariant $\mathbf{A}_q \in \mathbb{R}^{M\times N}$.

Conjecture: the behavior of the proposed EM-VAMP algorithm is rigorously characterized by a state-evolution with $M = O(N)$, $N \to \infty$, and either

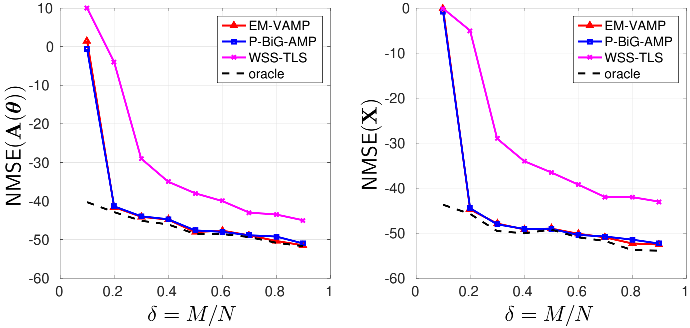
1. fixed Q and $L = 1$ (CS with matrix uncertainty)
2. $Q = O(N)$ and L fixed (self-calibration)
3. $Q = O(N^2)$ and $L = O(N \log N)$ (dictionary learning)

Technical conditions include:

- all vectors converge empirically with second-order moments to random variables
- singular values of $\mathbf{A}_q$ converge empirically with second-order moments to a bounded positive random variable
- Lipschitz $g(\cdot)$ and $g'(\cdot)$.
- exponential-family p_X
- etc...

Compressed Sensing with Matrix Uncertainty

Problem: Recover 10-sparse $\mathbf{x} \in \mathbb{R}^N$ from $\mathbf{y} = (\mathbf{A}_0 + \sum_{q=1}^{10} \theta_q \mathbf{A}_q)\mathbf{x} + \mathbf{w} \in \mathbb{R}^M$.



EM-VAMP performs similarly to P-BiG-AMP and much better than WSS-TLS.[14]

Details: $N = 256$, $\mathbf{A}_0 \sim$ i.i.d. $\mathcal{N}(0, 10)$, $\mathbf{A}_{q>1} \sim$ i.i.d. $\mathcal{N}(0, 1)$, SNR=40dB, 10 trials.

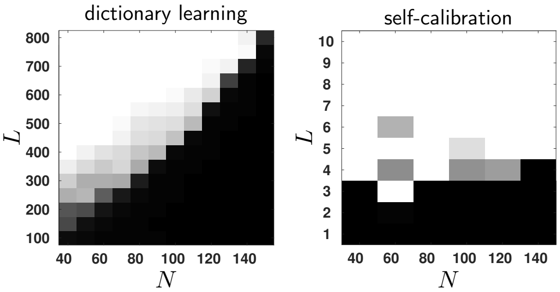
[14] Zhu,Leus,Giannakis'11

How many snapshots L are needed?

Problem: Recover $\mathbf{A} \in \mathbb{R}^{N\times N}$ and sparse $\mathbf{X} \in \mathbb{R}^{N\times L}$ from $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W}$ with

1. $\mathbf{A} \sim$ i.i.d. $\mathcal{N}(0,1)$ (dictionary learning)
2. $\mathbf{A} = \sum_{q=1}^{N} \theta_q \mathbf{A}_q$ with known $\mathbf{A}_q \sim$ i.i.d. $\mathcal{N}(0,1)$ (self-calibration)

NMSE (dB) versus N and L:

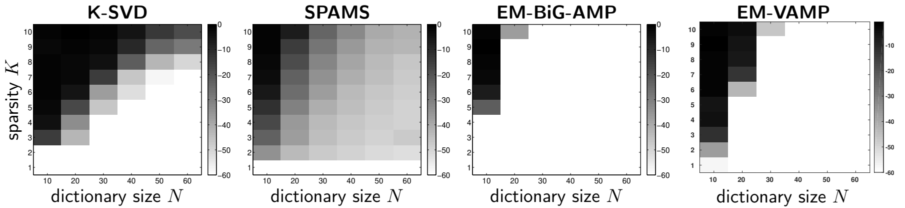


Details: i.i.d. Bernoulli-Gaussian $\mathbf{X}$, $K = \lceil 0.2N \rceil$, SNR$= 40$ dB.

Dictionary Learning: Sparsity vs Size

Problem: Recover i.i.d. Gaussian $\mathbf{A} \in \mathbb{R}^{N \times N}$ and Bernoulli-Gaussian $\mathbf{X} \in \mathbb{R}^{N \times L}$ with K-sparse columns from $\mathbf{Y} = \mathbf{AX}$.

NMSE (in dB) over 10 realizations for $L = 5N \log N$:

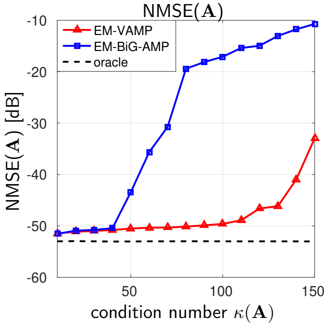


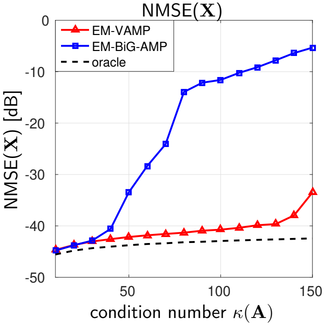
EM-VAMP has slightly worse phase-transition than BiG-AMP, but better than K-SVD[15] and SPAMS[16].

[15] Aharon,Elad,Bruckstein'06 [16] Mairal,Bach,Ponce,Sapiro'10

Dictionary Learning: Robustness to Condition Number

Same problem, but now with geometrically spaced singular values in $\mathbf{A}$.





EM-VAMP is more robust than EM-BiG-AMP to condition number $\kappa(\mathbf{A})$.

Details: $N \times N$ dictionary, $N = 64$, $K = 13$-sparse $\mathbf{x}_l$, $L = 5N \log N$, SNR=40dB, median NMSE over 100 realizations.

Conclusions

- We propose a bilinear recovery algorithm, with applications in self-calibration, CS with matrix uncertainty, blind deconvolution, dictionary learning, and joint channel-estimation and symbol detection.
- Broadly speaking, our approach is empirical-Bayesian and uses a combination of EC and EM.
- More specifically, our approach builds on the recently proposed EM-VAMP algorithm, by extending the set of unknown parameters to those that describe the measurement matrix $\mathbf{A}$.
- Numerical results suggest performance that is similar to P-BiG-AMP but more robust to non-iid matrices.
- We are currently working on proving the state-evolution conjectures.