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
Goal: Recover signal $X$ and parameters $\theta_A$ from noisy measurements

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

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: \( Y = A(\theta_A)X + W \)

Assumptions:
- \( A(\cdot) : \mathbb{R}^Q \rightarrow \mathbb{R}^{M \times N} \)  \hspace{1cm} \text{measurement operator}
- \( X \in \mathbb{R}^{N \times L} \) with \( x_{nl} \overset{\text{i.i.d.}}{\sim} p_X(\cdot; \theta_X) \)  \hspace{1cm} \text{random signal}
- \( W \in \mathbb{R}^{M \times L} \) with \( w_{ml} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \theta_w^{-1}) \)  \hspace{1cm} \text{AWGN}
- \( \Theta \triangleq \{ \theta_A, \theta_X, \theta_w \} \)  \hspace{1cm} \text{unknown deterministic parameters}

Goal: compute...
- \( \hat{\Theta}_{\text{ML}} = \arg \max_{\Theta} p(Y; \Theta) \)  \hspace{1cm} \text{maximum likelihood}
- \( \hat{X}_{\text{MMSE}} = \mathbb{E}[X|Y; \hat{\Theta}_{\text{ML}}] \)  \hspace{1cm} \text{“empirical Bayes”}
Related Work: $\Theta = \{ \theta_A, \theta_X, \theta_w \}$ Known

Consider the case where $\Theta \triangleq \{ \theta_A, \theta_X, \theta_w \}$ (and thus $A, p_X, p_W$) are known.

- The “vector AMP” (VAMP) algorithm\(^5\) can be applied.

- When $A$ 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 $A$.

---

\(^5\)Rangan,Schniter,Fletcher’16 \(^6\)Tulino,Caire,Verdú,Shamai’13 \(^7\)Donoho,Maleki,Montanari’10
Related Work: $\theta_A$ Known, $\{\theta_X, \theta_w\}$ Unknown

Now consider case where $\theta_A$ (and thus $A$) is known, but $\theta_X$ and $\theta_w$ are not.

- The EM-VAMP algorithm\(^8\) can be applied.

- When $A$ 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 $\{\hat{\theta}_X, \hat{\theta}_w\}$ are asymptotically consistent.\(^4\)

\(^8\)Fletcher, Rangan, Schniter’17

\(^4\)Schniter & Sarkar (OSU)
Finally, consider the case of interest, where $\theta_A$, $\theta_X$ and $\theta_w$ are unknown.

- Suppose we have $Y = A(\theta_A)X + W$ with affine linear map $A(\cdot)$.

- The **Parametric Bilinear Generalized AMP** (P-BiG-AMP) algorithm\textsuperscript{910} can be used to recover $\theta_A$ and $X \sim p_x(\cdot; \theta_X)$.

- But P-BiG-AMP is based on an i.i.d. Gaussian model for $A(\cdot)$, and it may not perform well when $A(\cdot)$ deviates from that model!

\textsuperscript{9}Parker, Schniter'16 \hspace{1cm} \textsuperscript{10}Schulke, Schniter, Zdeborova'16
Outline

This Work: $\{\theta_A, \theta_X, \theta_w\}$ 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 $\Theta$ via Expectation Maximization (EM)
4. Joint inference & learning via EM-VAMP
Variational Inference

For the moment, let’s suppose that $\Theta = \{\theta_A, \theta_X, \theta_w\}$ is known.

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

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

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

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

$$p(X|Y; \Theta) = \arg\min_b D\left(b(X)\|p(X|Y; \Theta)\right) \quad \text{where} \quad D(\cdot\|\cdot) \text{ is KL divergence}$$

$$= \arg\min_b \underbrace{D\left(b(X)\|p(X; \Theta)\right) + D\left(b(X)\|p(Y|X; \Theta)\right)}_{\text{Gibbs free energy}} + H\left(b(X)\right)$$

$$= \arg\min_{b, \bar{b}, q} \underbrace{D\left(b(X)\|p(X; \Theta)\right) + D\left(\bar{b}(X)\|p(Y|X; \Theta)\right)}_{\text{s.t.} \quad b = \bar{b} = q,} + H\left(q(X)\right)$$

$$\triangleq J_{\text{Gibbs}}(b, \bar{b}, q; \Theta)$$

but the density constraint keeps the problem difficult.
**Expectation Consistent Approximation**

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

\[
p(X|Y; \Theta) \approx \arg\min_{b, \bar{b}, q} J_{\text{Gibbs}}(b, \bar{b}, q; \Theta)
\]

subject to:

\[
\begin{align*}
\mathbb{E}\{x_l|b\} &= \mathbb{E}\{x_l|\bar{b}\} = \mathbb{E}\{x_l|q\} \quad \forall l \\
\text{tr}(\text{Cov}\{x_l|b\}) &= \text{tr}(\text{Cov}\{x_l|\bar{b}\}) = \text{tr}(\text{Cov}\{x_l|q\}) \quad \forall l.
\end{align*}
\]

- The stationary points of EC are the densities

\[
\begin{align*}
b(X) &\propto \prod_{l=1}^{L} p(x_l; \Theta) \mathcal{N}(x_l; \mathbf{r}_l, \mathbf{I}/\gamma_l) \\
\bar{b}(X) &\propto \prod_{l=1}^{L} p(y_l|x_l; \Theta) \mathcal{N}(x_l; \overline{\mathbf{r}}_l, \mathbf{I}/\overline{\gamma}_l) \\
q(X) &= \prod_{l=1}^{L} \mathcal{N}(x_l; \hat{x}_l, \mathbf{I}/\eta_l)
\end{align*}
\]

with parameters \( \{R, \gamma, \overline{R}, \overline{\gamma}, \hat{X}, \eta\} \) such that

\[
\begin{align*}
\mathbb{E}\{x_l|b\} &= \mathbb{E}\{x_l|\bar{b}\} = \hat{x}_l \quad \forall l \\
\frac{1}{N} \text{tr}(\text{Cov}\{x_l|b\}) &= \frac{1}{N} \text{tr}(\text{Cov}\{x_l|\bar{b}\}) = 1/\eta_l \quad \forall l.
\end{align*}
\]

\(^{11}\text{Opper, Winther'04}\)
The VAMP Algorithm

An iterative approach to finding \( \{R, \gamma, \overline{R}, \overline{\gamma}, \hat{X}, \eta\} \):

Initialize \( \{R, \gamma\} \) and select the estimation functions

\[
g(r_l; \gamma_l) = \mathbb{E}\{x_l | b; r_l, \gamma_l\} \\
\overline{g}(\overline{r}_l; \overline{\gamma}_l) = \mathbb{E}\{x_l | b; \overline{r}_l, \overline{\gamma}_l\}.
\]

For \( t = 1, 2, 3, \ldots \)

\[
\hat{x}_l \leftarrow g(r_l; \gamma_l), \ \forall l \quad \text{MMSE estimation}
\]

\[
\eta_l \leftarrow \gamma_l N / \text{tr} [\partial g(r_l; \gamma_l)/\partial r_l], \ \forall l
\]

\[
\overline{r}_l \leftarrow (\eta_l \hat{x}_l - \gamma_l r_l) / (\eta_l - \gamma_l), \ \forall l \quad \text{pseudo-measurement}
\]

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

\[
\overline{x}_l \leftarrow \overline{g}(\overline{r}_l; \overline{\gamma}_l), \ \forall l \quad \text{LMMSE estimation}
\]

\[
\overline{\eta}_l \leftarrow \overline{\gamma}_l N / \text{tr} [\partial \overline{g}(\overline{r}_l; \overline{\gamma}_l)/\partial \overline{r}_l], \ \forall l
\]

\[
r_l \leftarrow (\overline{\eta}_l \overline{x}_l - \overline{\gamma}_l \overline{r}_l) / (\overline{\eta}_l - \overline{\gamma}_l), \ \forall l \quad \text{pseudo-prior}
\]

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

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

We now return to the case where $\Theta = \{\theta_A, \theta_X, \theta_w\}$ is unknown.

- The EM algorithm is a well-known iterative approach to maximum-likelihood estimation of $\Theta$.

- The EM algorithm can be written in terms of the Gibbs free energy as\(^\text{12}\)

\[
\hat{\Theta}^{(t+1)} = \arg \min_{\Theta} \left( D_{KL}(b^{(t)}(X)||p(X; \Theta)) + D_{KL}(b^{(t)}(X)||p(Y|X; \Theta)) + H(b^{(t)}) \right) \\
= J_{\text{Gibbs}}(b^{(t)}, b^{(t)}, b^{(t)}; \Theta)
\]

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

- Thus EM and VAMP can be combined to solve

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

\(^{12}\)Neal,Hinton'98
The EM-VAMP Algorithm

Initialize \( \{ R, \gamma, \hat{\Theta} \} \) and select \( g(\cdot) \) & \( \bar{g}(\cdot) \) as before.

For \( t = 1, 2, 3, \ldots \)

\[
\hat{x}_l \leftarrow g(r_l; \gamma_l, \hat{\Theta}), \quad \forall l
\]

\( \eta_l \leftarrow \gamma_l N/ \text{tr} \left[ \partial g(r_l; \gamma_l, \hat{\Theta})/\partial r_l \right], \quad \forall l \)

\( \bar{r}_l \leftarrow (\eta_l \hat{x}_l - \gamma_l r_l)/(\eta_l - \gamma_l), \quad \forall l \)

\( \bar{\gamma}_l \leftarrow \eta_l - \gamma_l, \quad \forall l \)

\( \hat{\Theta} \leftarrow \arg \max_{\Theta} \mathbb{E}\{ \ln p(Y|X; \Theta) | \bar{R}; \bar{\gamma}, \hat{\Theta} \} \quad \text{EM update} \)

\[
\bar{x}_l \leftarrow \bar{g}(\bar{r}_l; \bar{\gamma}_l, \hat{\Theta}), \quad \forall l
\]

\( \bar{\eta}_l \leftarrow \bar{\gamma}_l N/ \text{tr} \left[ \partial \bar{g}(\bar{r}_l; \bar{\gamma}_l, \hat{\Theta})/\partial \bar{r}_l \right], \quad \forall l \)

\( r_l \leftarrow (\bar{\eta}_l \bar{x}_l - \bar{\gamma}_l \bar{r}_l)/(\bar{\eta}_l - \bar{\gamma}_l), \quad \forall l \)

\( \gamma_l \leftarrow \bar{\eta}_l - \bar{\gamma}_l, \quad \forall l \)

\( \hat{\Theta} \leftarrow \arg \max_{\Theta} \mathbb{E}\{ \ln p(X; \Theta) | R; \gamma, \hat{\Theta} \} \quad \text{EM update} \)
Problem: The precisions \( \{\gamma, \gamma\} \) of the pseudo-\{measurement,prior\} are imperfect when \( \hat{\Theta} \) is imperfect.

Thus, at each iteration, we estimate these precisions jointly with the unknown parameters \( \Theta \). For example, with the prior parameters \( \theta_X \): 

\[
(\hat{\gamma}, \hat{\theta}_X) \leftarrow \arg\max_{\gamma,\theta_X} p(R; \gamma, \theta_X)
\]

under \( r_l = x_l + \mathcal{N}(0, I/\gamma_l), \quad x_l \sim p_X(\cdot; \theta_X), \quad \forall l. \)

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

\[
(\hat{\gamma}, \hat{\theta}_X) = \arg\max_{\gamma,\theta_X} \mathbb{E}_{\gamma,\theta_X}[\ln p(X, R; \gamma, \theta_X) | R; \hat{\gamma}, \hat{\theta}_X].
\]

This “variance auto-tuning” procedure\(^{13}\) leads to asymptot. consistent \( \hat{\theta}_X \).

\(^{13}\)Fletcher, Rangan, Schniter’17
State-Evolution and Consistency

Suppose that

- \( \mathbf{Y} = \mathbf{A}(\theta_A)\mathbf{X} + \mathcal{N}(\mathbf{0}, \mathbf{I}/\theta_w) \in \mathbb{R}^{M \times L}, \quad \mathbf{X} \overset{\text{i.i.d}}{\sim} p_X(\cdot; \theta_X) \)
- \( \mathbf{A}(\theta) = \mathbf{A}_0 + \sum_{q=1}^{Q} \theta_q \mathbf{A}_q \) 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 \( M = O(N), \quad N \to \infty \), 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: $N = 256$, $\mathbf{A}_0 \sim \text{i.i.d. } \mathcal{N}(0, 10)$, $\mathbf{A}_{q>1} \sim \text{i.i.d. } \mathcal{N}(0, 1)$, SNR=40dB, 10 trials.

\(^{14}\)Zhu, Leus, Giannakis’11
Numerical Results

How many snapshots $L$ are needed?

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

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

NMSE (dB) versus $N$ and $L$:

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

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\textsuperscript{15} and SPAMS\textsuperscript{16}.

\textsuperscript{15}Aharon, Elad, Bruckstein’06
\textsuperscript{16}Mairal, Bach, Ponce, Sapiro’10
Dictionary Learning: Robustness to Condition Number

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

**Numerical Results**

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 $A$.

- 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.