Bilinear Recovery using Adaptive Vector-AMP

Subrata Sarkar (Ohio State)
Philip Schniter (Ohio State)

(supported by NSF 1716388)

THE OHIO STATE UNIVERSITY

Asilomar 2019

Alyson K. Fletcher (UCLA)
Sundeep Rangan (NYU)

Bilinear Recovery Problem

- Observations:

$$\boldsymbol{Y} = \sum_{i=1}^{Q} b_i \boldsymbol{A}_i \boldsymbol{X} + \boldsymbol{W}$$

where,

$\boldsymbol{X}$: unknown random matrix in $\mathbb{R}^{N\times L}$
$\boldsymbol{A}_1, \dots, \boldsymbol{A}_Q$: known matrices in $\mathbb{R}^{M\times N}$
$b_1, \dots, b_Q$: unknown deterministic parameters
$\boldsymbol{W}$: white Gaussian noise.

- Prior:

$X_{nl} \overset{\text{i.i.d}}{\sim} p_X(\cdot;\boldsymbol{\theta}_X)$ deterministic unknown parameters $\boldsymbol{\theta}_X$.
$W_{ml} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,\sigma_w^2)$ unknown variance σ_w^2.

Goal: jointly infer $\boldsymbol{X}$ and estimate $\boldsymbol{\theta} \triangleq \{\boldsymbol{b}, \boldsymbol{\theta}_X, \sigma_w^2\}$
Approach: combine variational inference with ML estimation.
Applications: Self-calibration, CS+matrix uncertainty, dictionary learning, ...

Variational Inference

- For now, let's suppose that $\boldsymbol{\theta}$ is known.
- We would like to compute the posterior density

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

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

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

$$\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{Y}) &= \arg\min_b D\big(b(\boldsymbol{X})\big\|p(\boldsymbol{X}|\boldsymbol{Y})\big) \text{ where } D(\cdot\|\cdot) \text{ is KL divergence}\\
&= \arg\min_b \underbrace{D\big(b(\boldsymbol{X})\big\|p(\boldsymbol{X};\boldsymbol{\theta})\big) + D\big(b(\boldsymbol{X})\big\|p(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})\big) + H\big(b(\boldsymbol{X})\big)}_{\text{Gibbs free energy}}\\
&= \arg\min_{b_1,b_2,q} \underbrace{D\big(b_1(\boldsymbol{X})\big\|p(\boldsymbol{X};\boldsymbol{\theta})\big) + D\big(b_2(\boldsymbol{X})\big\|p(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})\big) + H\big(q(\boldsymbol{X})\big)}_{\triangleq J(b_1,b_2,q;\boldsymbol{\theta})}
\end{aligned}$$

such that $b_1 = b_2 = q$,

but the density constraint keeps the problem difficult.

- Expectation consistent approximation (EC) [1] relaxes the density constraint to moment-matching constraints:

$$p(\boldsymbol{X}|\boldsymbol{Y}) \approx \arg\min_{b_1,b_2,q} J(b_1,b_2,q;\boldsymbol{\theta})$$

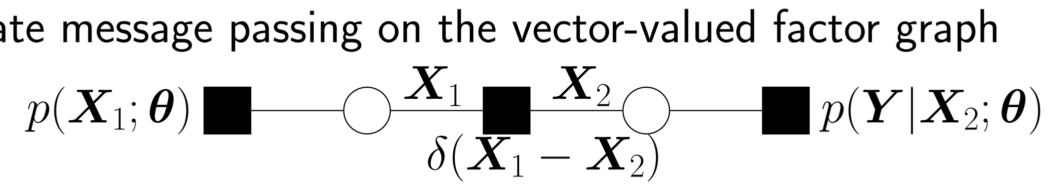
such that $\forall l \; \begin{cases} \mathbb{E}\{\boldsymbol{x}_l|b_1\} = \mathbb{E}\{\boldsymbol{x}_l|b_2\} = \mathbb{E}\{\boldsymbol{x}_l|q\} \\ \mathrm{tr}[\mathrm{Cov}\{\boldsymbol{x}_l|b_1\}] = \mathrm{tr}[\mathrm{Cov}\{\boldsymbol{x}_l|b_2\}] = \mathrm{tr}[\mathrm{Cov}\{\boldsymbol{x}_l|q\}]. \end{cases}$

- The stationary points of EC are the densities

$$\begin{aligned}
b_1(\boldsymbol{X}) &\propto \textstyle\prod_{l=1}^L p(\boldsymbol{x};\boldsymbol{\theta})\mathcal{N}(\boldsymbol{x}_l;\boldsymbol{r}_{1,l},\boldsymbol{I}/\gamma_{1,l})\\
b_2(\boldsymbol{X}) &\propto \textstyle\prod_{l=1}^L p(\boldsymbol{y}_l|\boldsymbol{x}_l;\boldsymbol{\theta})\mathcal{N}(\boldsymbol{x}_l;\boldsymbol{r}_{2,l},\boldsymbol{I}/\gamma_{2,l})\\
q(\boldsymbol{X}) &= \textstyle\prod_{l=1}^L \mathcal{N}(\boldsymbol{x}_l;\widehat{\boldsymbol{x}}_l,\boldsymbol{I}/\eta_l)
\end{aligned}
\quad \text{s.t.} \quad
\begin{cases}
\mathbb{E}\{\boldsymbol{x}_l|b_1\} = \mathbb{E}\{\boldsymbol{x}_l|b_2\} = \widehat{\boldsymbol{x}}_l\\
\mathrm{tr}[\mathrm{Cov}\{\boldsymbol{x}_l|b_1\}]\\
= \mathrm{tr}[\mathrm{Cov}\{\boldsymbol{x}_l|b_2\}] = N/\eta_l.
\end{cases}$$

Vector AMP (VAMP)

- There exist several algorithms (e.g., EC, ADATAP [2], S-AMP [3]) whose fixed points coincide with the EC stationary points, but often they don't converge.
- An exception is Vector AMP [4], which can be derived using a form of approximate message passing on the vector-valued factor graph



In particular, VAMP is provably convergent under either
1) strictly log-concave prior $p(\boldsymbol{X};\boldsymbol{\theta})$ and arbitrary $\boldsymbol{A}$ (after damping),
2) iid prior $p(\boldsymbol{X};\boldsymbol{\theta})$ and large, right-rotationally invariant $\boldsymbol{A}$.

VAMP algorithm

Initialize $\{\boldsymbol{R}_1, \gamma_1\}$ and define the estimation functions

$\boldsymbol{g}_1(\boldsymbol{r}_{1,l},\gamma_{1,l}) \triangleq \mathbb{E}\{\boldsymbol{x}_l|b_1;\boldsymbol{r}_{1,l},\gamma_{1,l}\}$ or any Lipschitz function
$\boldsymbol{g}_2(\boldsymbol{r}_{2,l},\gamma_{2,l}) \triangleq \mathbb{E}\{\boldsymbol{x}_l|b_2;\boldsymbol{r}_{2,l},\gamma_{2,l}\}$

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

$\widehat{\boldsymbol{x}}_{1,l} \leftarrow \boldsymbol{g}_1(\boldsymbol{r}_{1,l},\gamma_{1,l}), \forall l$ — denoising
$\eta_{1,l} \leftarrow \gamma_{1,l}N/\,\mathrm{tr}[\partial\boldsymbol{g}_1(\boldsymbol{r}_{1,l};\gamma_{1,l})/\partial\boldsymbol{r}_{1,l}], \forall l$
$\boldsymbol{r}_{2,l} \leftarrow (\eta_{1,l}\widehat{\boldsymbol{x}}_{1,l} - \gamma_{1,l}\boldsymbol{r}_{1,l})/(\eta_{1,l} - \gamma_{1,l}), \forall l$ — pseudo-measurement
$\gamma_{2,l} \leftarrow \eta_{1,l} - \gamma_{1,l}, \forall l$

$\widehat{\boldsymbol{x}}_{2,l} \leftarrow \boldsymbol{g}_2(\boldsymbol{r}_{2,l},\gamma_{2,l}), \forall l$ — LMMSE estimation
$\eta_{2,l} \leftarrow \gamma_{2,l}N/\,\mathrm{tr}[\partial\boldsymbol{g}_2(\boldsymbol{r}_{2,l};\gamma_{2,l})/\partial\boldsymbol{r}_{2,l}], \forall l$
$\boldsymbol{r}_{1,l} \leftarrow (\eta_{2,l}\widehat{\boldsymbol{x}}_{2,l} - \gamma_{2,l}\boldsymbol{r}_{2,l})/(\eta_{2,l} - \gamma_{2,l}), \forall l$ — pseudo-prior
$\gamma_{1,l} \leftarrow \eta_{2,l} - \gamma_{2,l}, \forall l$

Expectation maximization (EM)

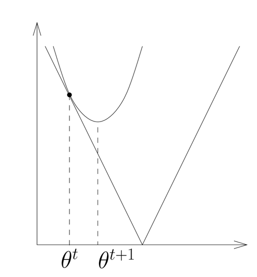
- We now return to the case where $\boldsymbol{\theta} = \{\boldsymbol{b}, \boldsymbol{\theta}_X, \sigma_w^2\}$ is unknown.
- The maximum-likelihood (ML) estimate is

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{Y};\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \{-\ln p(\boldsymbol{Y};\boldsymbol{\theta})\}.$$

- EM algorithm iteratively minimizes a tight upper bound on $-\ln p(\boldsymbol{Y};\boldsymbol{\theta})$:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{t+1} &= \arg\min_{\boldsymbol{\theta}} \mathbb{E}\big\{-\ln p(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\theta})\big|\boldsymbol{Y};\widehat{\boldsymbol{\theta}}^t\big\}\\
&= \arg\min_{\boldsymbol{\theta}} \big\{-\ln p(\boldsymbol{Y};\boldsymbol{\theta}) + \underbrace{D\big(b^t(\boldsymbol{X})\big\|p(\boldsymbol{X}|\boldsymbol{Y};\boldsymbol{\theta})\big)}_{\geq 0}\big\}
\end{aligned}$$

with $b^t(\boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{Y};\widehat{\boldsymbol{\theta}}^t)$



- The upper bound can also be rewritten in terms of Gibbs free energy

$$\begin{aligned}
Q(\boldsymbol{\theta}, b^t) &\triangleq -\ln p(\boldsymbol{Y};\boldsymbol{\theta}) + D\big(b^t(\boldsymbol{X})\big\|p(\boldsymbol{X}|\boldsymbol{Y};\boldsymbol{\theta})\big)\\
&= D\big(b^t(\boldsymbol{X})\big\|p(\boldsymbol{X};\boldsymbol{\theta})\big) + D\big(b^t(\boldsymbol{X})\big\|p(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})\big) + H\big(b^t(\boldsymbol{X})\big)\\
&= J(b^t, b^t, b^t;\boldsymbol{\theta})
\end{aligned}$$

which yields a variational interpretation of EM [5].

Variance Auto-Tuning

- In VAMP, the precisions $\{\gamma_{1,l},\gamma_{2,l}\}_{l=1}^L$ are imperfect when $\widehat{\boldsymbol{\theta}}$ is imperfect.
- So we estimate these precisions jointly with $\boldsymbol{\theta}$. E.g., for parameters $\boldsymbol{\theta}_X$:

$$(\gamma_1, \widehat{\boldsymbol{\theta}}_X) \leftarrow \arg\max_{\gamma_1,\boldsymbol{\theta}_X} p(\boldsymbol{R}_1;\gamma_1,\boldsymbol{\theta}_X) \text{ under } \boldsymbol{r}_{1,l} = \boldsymbol{x}_l + \mathcal{N}(\boldsymbol{0},\boldsymbol{I}/\gamma_{1,l}),\; \boldsymbol{x}_l \sim p(\cdot;\boldsymbol{\theta}_X)$$

- In practice, inner iterations of EM are used to solve the above "variance auto-tuning" problem.
- Under identifiability conditions, this leads to asymptotically consistent $\widehat{\boldsymbol{\theta}}_X$ [6].

The proposed Bilinear Adaptive (BAd)-VAMP Algorithm

- Recall that VAMP iteratively computes a posterior approximation $b^t(\boldsymbol{X})$ by minimizing $J(b_1, b_2, q;\boldsymbol{\theta})$ (under moment constraints) with known $\boldsymbol{\theta}$.
- Likewise, EM iteratively estimates $\boldsymbol{\theta}$ by minimizing $J(b^t, b^t, b^t;\boldsymbol{\theta})$, assuming the posterior approximation $b^t(\boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{Y};\boldsymbol{\theta}^t)$ is available.
- In BAd-VAMP, we combine VAMP, EM, and variance auto-tuning.
- In the denoising ($i=1$) and LMMSE ($i=2$) steps of VAMP, we infer $\boldsymbol{X}$ and jointly estimate $(\boldsymbol{\gamma},\boldsymbol{\theta})$ by running several inner iterations of

$$\forall l: \widehat{\boldsymbol{x}}_{i,l} \leftarrow \boldsymbol{g}_i(\boldsymbol{r}_{i,l},\gamma_{i,l};\widehat{\boldsymbol{\theta}}_i), \quad \eta_{i,l} \leftarrow \gamma_{i,l}N/\,\mathrm{tr}[\partial\boldsymbol{g}_i(\boldsymbol{r}_{i,l},\gamma_{i,l};\widehat{\boldsymbol{\theta}}_i)/\partial\boldsymbol{r}_{i,l}] \quad \text{(denoisin}$$

$$\forall l: \gamma_{i,l} \leftarrow \left(\tfrac{1}{N}\|\widehat{\boldsymbol{x}}_{i,l} - \boldsymbol{r}_{i,l}\|^2 + 1/\eta_{i,l}\right)^{-1} \quad \text{(auto-tunin}$$

$$q_i(\boldsymbol{X}) \propto \prod_{l=1}^L f_i(\boldsymbol{x}_{i,l};\boldsymbol{\theta}_i)\mathcal{N}(\boldsymbol{x}_{i,l};\boldsymbol{r}_{i,l},\boldsymbol{I}/\gamma_{i,l}) \quad \text{(belief updat}$$

$$\widehat{\boldsymbol{\theta}}_i \leftarrow \arg\max_{\boldsymbol{\theta}_i} \sum_{l=1}^L \mathbb{E}\left[\ln f_i(\boldsymbol{x}_{i,l},\boldsymbol{\theta}_i)|q_i\right] \quad \text{(EM update)}$$

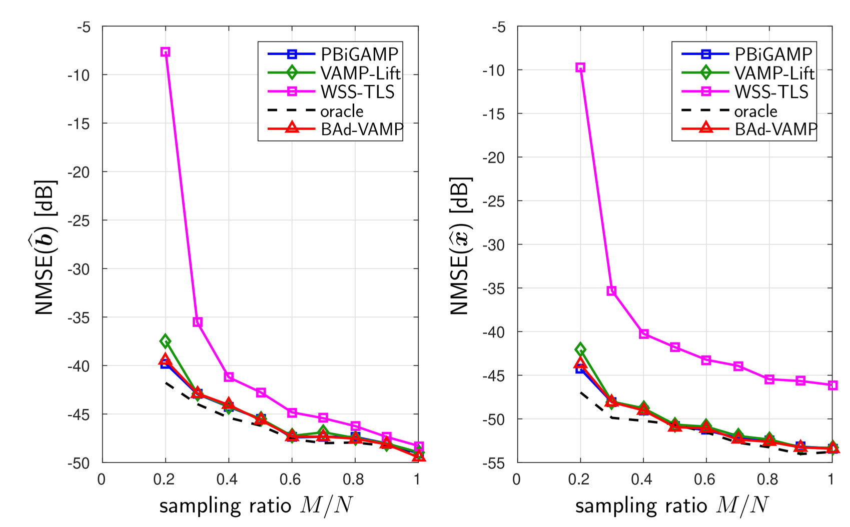
with $i \in \{1,2\}$, $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_X, \boldsymbol{\theta}_2 = \{\boldsymbol{b},\sigma_w^2\}$ and

$$f_1(\boldsymbol{x},\boldsymbol{\theta}_1) = p(\boldsymbol{x};\boldsymbol{\theta}_X), \quad f_2(\boldsymbol{x},\boldsymbol{\theta}_2) = \mathcal{N}\left(\boldsymbol{y};\sum_{j=1}^Q b_j\boldsymbol{A}_j\boldsymbol{x},\sigma_w^2\boldsymbol{I}\right).$$

Numerical Experiments

CS with Matrix Uncertainty: Recover $N = 256$-length sparse $\boldsymbol{x}$ and $b_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ from M-length $\boldsymbol{y} = (\boldsymbol{A}_0 + \sum_{i=1}^{10} b_i\boldsymbol{A}_i)\boldsymbol{x} + \boldsymbol{w}$, where $[\boldsymbol{A}_i]_{m,n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\boldsymbol{x}$ is sampled from the Bernoulli-Gaussian distribution, $\epsilon = 0.04$,

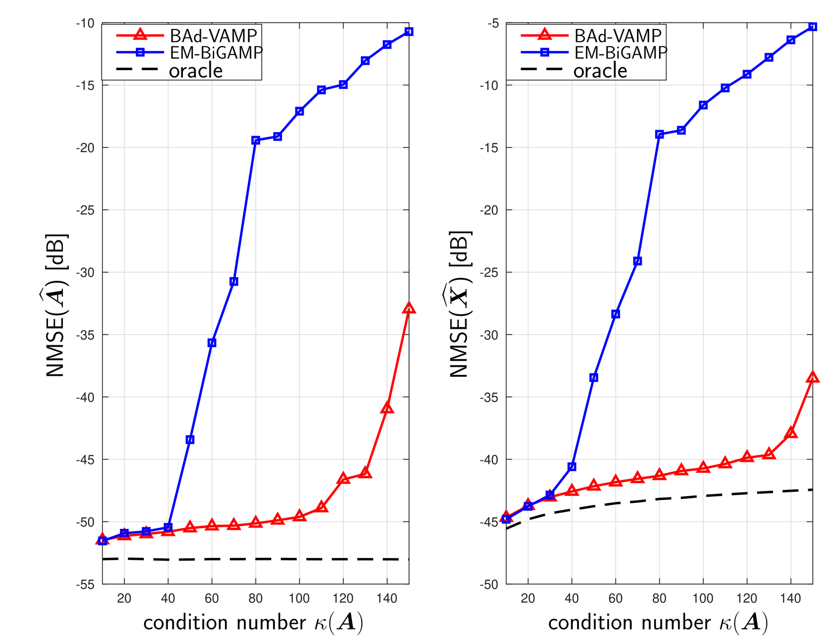
$$x_i \sim (1-\epsilon)\delta(x) + \epsilon\mathcal{N}(x;0,1)$$



Plots show median NMSE on signal $\widehat{\boldsymbol{x}}$ and parameter $\widehat{\boldsymbol{b}}$ estimates versus sampling ratio M/N.

BAd-VAMP performs much better than WSS-TLS [7] and on par with EM-PBiGAMP [8] and VAMP-Lift [9].

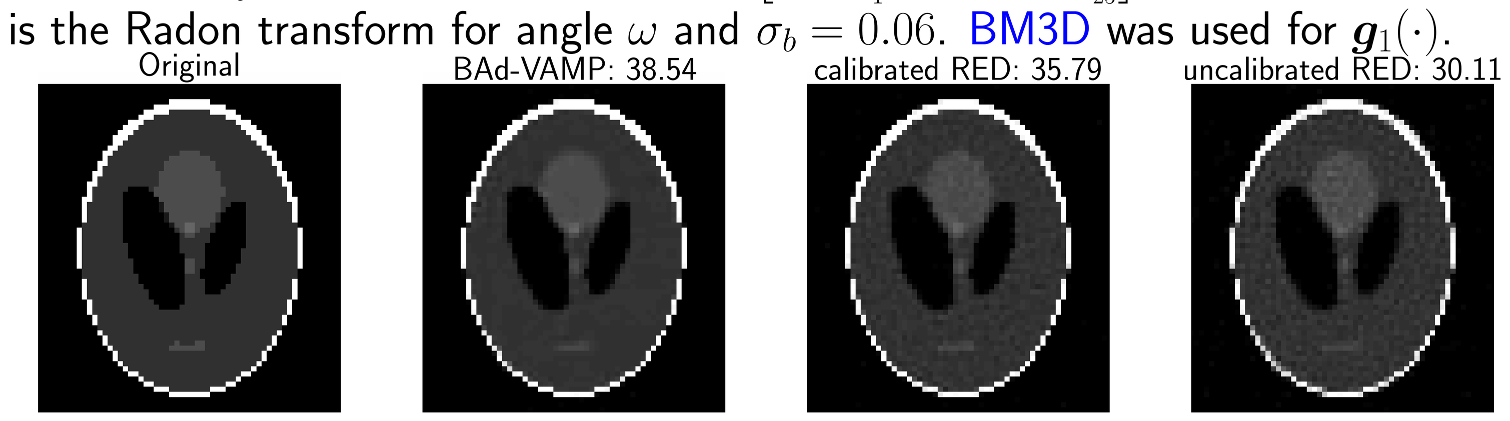
Dictionary Learning: Given $\boldsymbol{Y}$, the goal is to estimate $\boldsymbol{A}$ & $\boldsymbol{X}$ s.t. $\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X}$.



Median NMSE versus condition number $\kappa(\boldsymbol{A})$ for $\boldsymbol{A} \in \mathbb{R}^{64\times 64}$, i.i.d. Bernoulli-Gaussian $\boldsymbol{X}$ with $\epsilon = 0.2$, training length $L = 1331$, and SNR $= 40$ dB.

BAd-VAMP is much more robust to $\kappa(\boldsymbol{A})$ than EM-BiGAMP [10].

Self-Calibration in Tomography: Reconstruct image $\boldsymbol{x}$ and calibration parameters $b_i \sim \mathcal{N}(1,\sigma_b^2)$ from measurements $\boldsymbol{y} = \left[b_1\boldsymbol{\Psi}_{\omega_1}^\mathsf{T} \dots b_{25}\boldsymbol{\Psi}_{\omega_{25}}^\mathsf{T}\right]^\mathsf{T}\boldsymbol{x} + \boldsymbol{w}$, where $\boldsymbol{\Psi}_\omega$ is the Radon transform for angle ω and $\sigma_b = 0.06$. BM3D was used for $\boldsymbol{g}_1(\cdot)$.



PSNR (dB) of 64×64 Shepp-Logan phantom from 25 equally spaced tomographic projections.

References

1. M. Opper and O. Winther, "Expectation consistent approximate inference," *J. Mach. Learning Res.*, 2005. .
2. M. Opper and O. Winther, "Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling," *Phy. Rev. E*, 2001. .
3. B. Çakmak, O. Winther, and B.H. Fleury, "S-AMP: Approximate message passing for general matrix ensembles," *ISIT* 2014.
4. S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *arXiv:1610.03082*.
5. R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, 1998.
6. A. K. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter, "Rigorous dynamics and consistent estimation in arbitrarily conditioned linear systems," *Proc. NeurIPS*, 2017.
7. H. Zhu, G. Leus, and G. B. Giannakis, "Sparsity-cognizant total least-squares for perturbed compressive sampling," *IEEE TSP*, 2011. .
8. J. T. Parker, P. Schniter, "Parameteric bilinear generalized approximate message passing", *IEEE JSTSP*, 2016.
9. A. K. Fletcher, P. Pandit, S. Rangan, S. Sarkar, and P. Schniter, "Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis," *Proc. NeurIPS*, 2018.
10. J. T. Parker, P. Schniter and V. Cevher, "Bilinear generalized approximate message passing", *IEEE TSP*, 2014.
11. Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," SIAM J. Imaging Science, 2017.