Bayesian-Inspired Non-Convex Methods for Sparse Signal Recovery

Chandra R. Murthy, Indian Institute of Science

and

David Wipf, Microsoft Research Beijing

{cmurthy1, davidwipf}@gmail.com

Outline

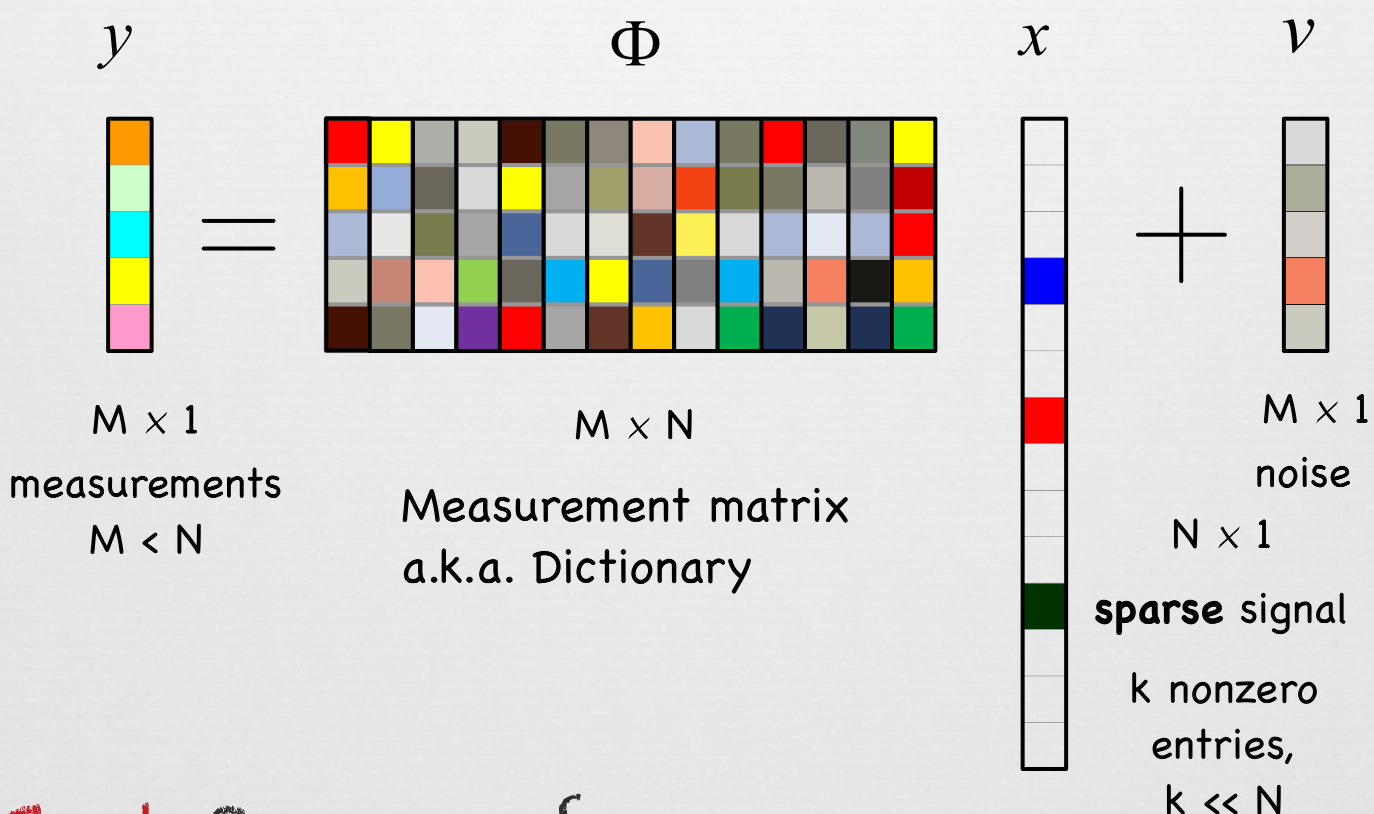
- Background and motivation
  - Algorithms for sparse signal recovery
  - Guarantees on sparse signal recovery
- Non convex methods:
  - MAP estimation
  - Sparse Bayesian learning
- Useful extensions
- Application to wireless communication

Part 1: Setting the Stage

Motivation and background

Basic results

Sparse Signal Recovery



- Goal: Recover x from y
- M << N: infinitely many solutions

Compressed Sensing

- Deals with two main questions:
  - Design of sensing matrices with recovery guarantees

$$\mathbf{\Phi}_{M\times N} = \mathbf{A}_{M\times N}\mathbf{\Psi}_{N\times N}$$

Sparsifying Basis

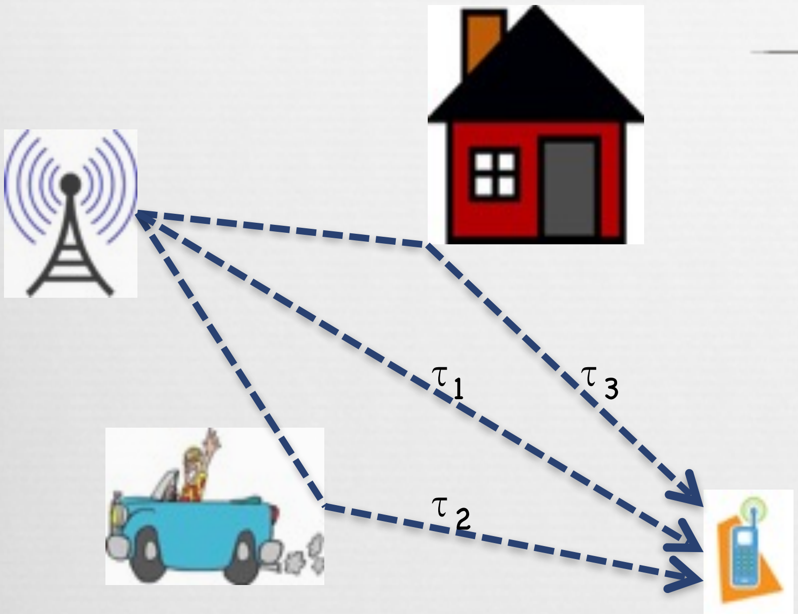
  - Computationally efficient recovery

- Our focus: sparse signal recovery from noisy linear underdetermined measurements

Applications

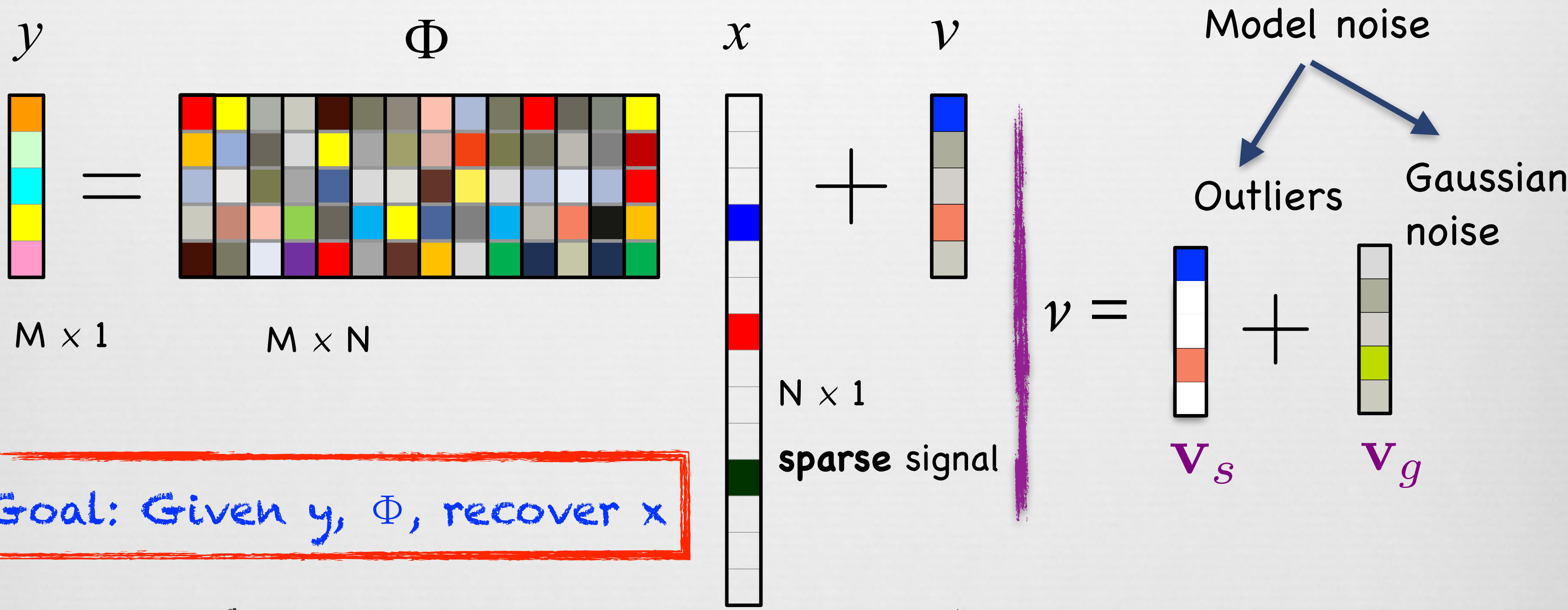
- Signal representation (Mallat, Coifman, Wickerhauser, Donoho, ...)
- Functional Approx. (Chen, Nagarajan, Cun, Hassibi, ...)
- Spectral estmn., cartography (Papoulis, Lee, Cabrera, Parks, ...)
- EEG/MEG (Leahy, Gordonitsky, Ioannides, ...)
- Medical imaging (Lustig, Pauly, ...)
- Speech SP (Ozawa, Ono, Kroon, Atal, ...)
- Sparse channel estimation (Fevrier, Greenstein, Proakis, Prasad et al.,...)
- Outlier removal and feature selection in machine learning

Wireless Channel Estimation



- Wireless channels exhibit multipath
  - Naturally sparse in the lag-domain
  - Need to estimate both support & channel
- Channel equalization & data detection

Robust Linear Regression: Underdetermined Case



Goal: Given y, Φ, recover x

- Transform into an overcomplete problem:

$$\mathbf{Y} = \mathbf{\Phi x} + \mathbf{\Psi v}_s + \mathbf{v}_g, \text{ where } \mathbf{\Psi} = \mathbf{I}$$

$$\text{or } \mathbf{Y} = [\mathbf{\Phi}, \mathbf{\Psi}] \begin{bmatrix} \mathbf{x} \\ \mathbf{v}_s \end{bmatrix} + \mathbf{v}_g$$

Sparse recovery algos are now applicable!

Robust Linear Regression: Overdetermined Case

- Measurement model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{E} + \mathbf{e}$$

$M \times N;$ $M \geq N$ — Outliers; sparse — Noise

- Use SVD: $\mathbf{A} = \mathbf{U}_1 \Sigma \mathbf{V}_1^T;\ \mathbf{U}_2^T \mathbf{A} = \mathbf{0}$
- Processed measurements:

$$\tilde{\mathbf{y}} = \mathbf{U}_2^T \mathbf{y} = \mathbf{U}_2^T \mathbf{E} + \mathbf{U}_2^T \mathbf{e}$$

- Can now directly apply sparse signal recovery algorithms to estimate and remove outliers!

The Problem

- Noiseless case: Given y and Φ, solve

$$\min \|\mathbf{x}\|_0 \text{ subject to } \mathbf{y} = \mathbf{\Phi}\mathbf{x}$$

- Noisy case: solve

$$\min \|\mathbf{x}\|_0 \text{ subject to } \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2 \leq \beta$$

- l_0 norm minimization
  - Combinatorial complexity
  - Not robust to noise

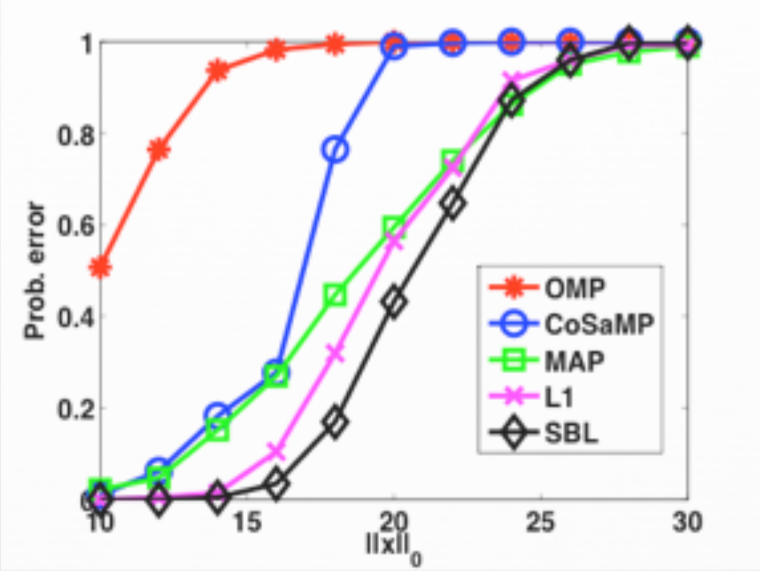
Recovery Algorithms

- Greedy algorithms:
  - Matching pursuit [Mallat, Zhang; Cotter, Rao]
  - Orthogonal matching pursuit [Tropp 03]
  - CoSAMP [Needell, Tropp]
- Relaxation based methods (minimize diversity meas.):
  - Basis pursuit (l-p, with p=1) [Chen et al.]
  - Lasso (BPDN) [Tibshirani]
  - Dantzig selector [Candes, Tao]
  - Homotopy based methods (e.g., LARS) [Garrigues et al. 09]
  - FOCUSS (l-p, with p < 1) [Gordonitsky et al.]
- Iterative methods:
  - Basic/Iterative hard thresholding
  - Hard thresholding pursuit

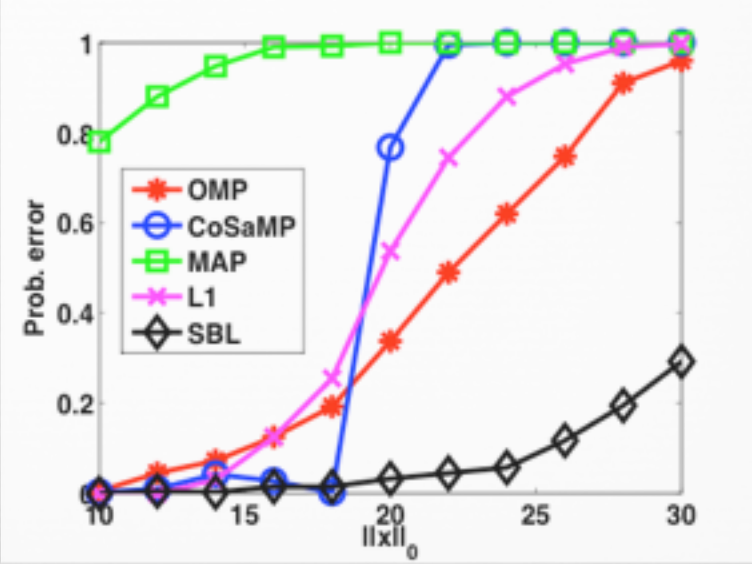
Recovery guarantees exist for most of these algorithms! See [Rauhut & Foucart]

Motivational Example

- Generate random 50 x 100 matrix Φ
- Generate sparse vector x_0
- Compute $y = \Phi x_0$
- Solve for x_0, average over 1000 trials
- Repeat for different sparsity values



Unit magnitude entries



Highly scaled entries

Limitations of Relaxation and Greed

- Performance of BP and OMP depend on Φ
  - Poor performance when conditions are violated
  - Hard to relate estimation error to the dictionary
  - Correlated dictionary: disrupts l_0-l_1 equivalence
- BP: performance independent of nonzero coeffs [Malioutov et al. 2004]
  - Cannot improve when situation is favorable
- OMP: performance highly sensitive to magnitudes of nonzero coefficients
  - Poor performance with unit magnitudes

Other Limitations of Convex Relaxation

- Scaling/shrinkage:
  - Noiseless: l_0 <-> l_1 <-> l_2. Shrinking large coeffs can reduce variance, but at the cost of sparsity
  - Noisy: The τ in lasso that minimizes the MSE could result in a much larger number of nonzero coeffs
- Estimating embedded params (e.g., in Φ)

To Recap

- Sparse signal recovery
  - Basic problem
  - Algorithms
- Limitations
  - Scaling/shrinkage
  - Correlated dictionary
  - Embedded parameters

Part 2: Dont Relax!

A time and place for nonconvex methods?

Bayesian Methods

- MAP estimation (Type I):
  - Also a regression problem with sparsity promoting penalties (e.g., l_p-norm)
  - l_1-min (BP/LASSO) is a special case
  - Iterative reweighted l_1 [Candes et al. 2008]
  - Iterative reweighted l_2 [Chartrand & Yin 2008]
- Hierarchical Bayesian methods (Type II):
  - EM-based SBL [Tipping, 2001], [Wipf, Rao 2007]
  - AMP-based methods [Schniter 2008], [Rangan 2011]

MAP Estimation

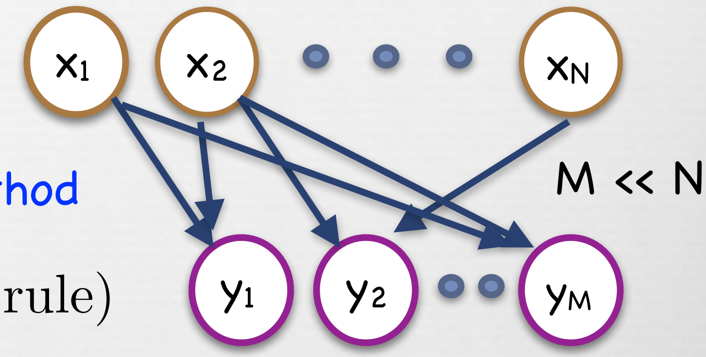
$$\boxed{\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})}$$

Type-I method

$$= \arg\min_{\mathbf{x}} -\log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{x}) \quad \text{(Bayes' rule)}$$

$$= \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \lambda \sum_{i=1}^{N} g(|x_i|)$$

Separable prior



- For sparse solutions, $g(|x_i|)$ should be a concave, nondecreasing function
  - Example: $g(|x_i|) = |x_i|^p$, $p \leq 1$
  - Lasso is a special case: p=1
- Any local min. of the MAP estimation problem has at most M nonzeros [Rao et al., 99]

The Optimization Problem

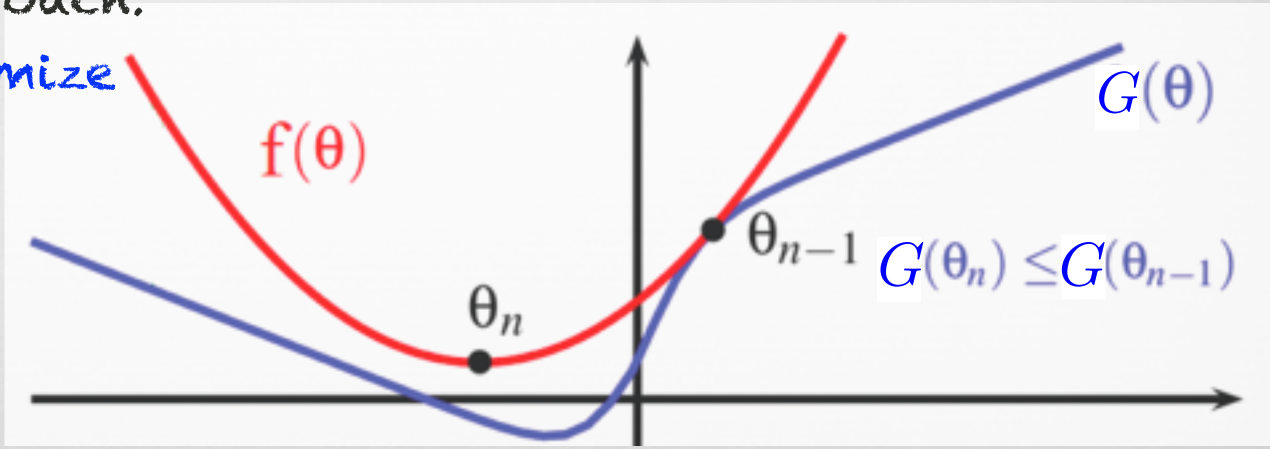
- To solve

$$\arg\min_{\mathbf{x}} G(\mathbf{x}) \triangleq \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda \sum_{i=1}^{N} g(|x_i|)$$

  - g(|x|) symmetric and concave, monotonically increasing for $\mathbf{x} \in \mathbb{R}^+$
  - G(x) convex + concave
- Many options for g(|x|) to promote sparsity
- Many options for solving the optz. problem

Sparsity-Promoting Penalties

- Concave penalty fns. promote sparsity
  - $g(|x|) = \log(|x|^2 + \varepsilon),\ \varepsilon > 0$ [Chartrand & Yin 2008]
  - $g(|x|) = \log(|x| + \varepsilon),\ \varepsilon > 0$ [Candes et al. 2008]
  - $g(|x|) = |x|^p,\ 0 < p < 1$ [Rao et al., 99]
- A general approach: majorize-minimize



Majorization-Minimization Approach

- **Find an upper bound** $g(x) \leq f(x|x^{(m)})$
  - Equality at $x = x^{(m)}$, convenient for opt.
- Step 1: Optimize

$$\arg\min_{\mathbf{x}} F\left(\mathbf{x}|\mathbf{x}^{(m)}\right) \triangleq \|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \lambda \sum_{i=1}^{N} f\left(|x_i||x_i^{(m)}\right)$$

- Step 2: Set m <- m+1, update $f(x|x^{(m)})$, iterate
- **Works because**

  $G(x^{(m+1)}) \leq F(x^{(m+1)}|x^{(m)}) \leq F(x^{(m)}|x^{(m)}) = G(x^{(m)})$

Iterative Reweighted l_1

- Concavity in $|x|$: $g(x) \le g'(x^{(m)})(x-x^{(m)}) + g(x^{(m)})$
  - Equality at x = x(m), linear in x
- Iterative reweighted l_1: [Candes et al. 08]
  - Init: $m = 0$, $x^{(m)}$ = something convenient
  - Iterate:
    - Optimize

$$\mathbf{x}^{(m+1)} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \lambda \sum_{i=1}^{N} g'(x_i^{(m)})|x_i|$$

    - m <- m+1, update $g'(x_i^{(m)})$
  - Until convergence

Weighted l_1 minimization

Iterative Reweighted L_2

- $g(x)$ concave in x^2: $g(x) \leq \left(\left. \frac{\partial g(\sqrt{x^2})}{\partial (x^2)} \right|_{x=x_0} \right) (x^2 - x_0^2) + g(x_0)$
- Optimization problem

$$\mathbf{x}^{(m+1)} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2 + \lambda \sum_{i=1}^{N} w_i^{(m)} |x_i|^2$$

$$\|\mathbf{W}_m^{-\frac{1}{2}}\mathbf{x}\|_2^2$$

- Iterative reweighted L2 [Chartrand et al. 08]
  - Init: m = 0, $x^{(m)}$ = something convenient
  - Iterate:
    - Compute $\mathbf{x}^{(m+1)} = \mathbf{W}_m \Phi^T \left(\lambda \mathbf{I} + \Phi \mathbf{W}_m \Phi^T\right)^{-1} \mathbf{y}$
    - m <- m+1, update W_m
  - Until convergence

Other Ways to Bound

- Taylor's expansion
- Jensen's inequality
- Concave conjugate inequality
- Good opportunity to innovate!

An Example

- Suppose $g(x) = \log(|x| + \varepsilon)$, $\varepsilon > 0$
  - Concave in $|x|$, x^2
- Iterative reweighted l1

$$g'\left(x_i^{(m)}\right) = \left[\left|x_i^{(m)}\right| + \epsilon\right]^{-1}$$

- Iterative reweighted l2

$$w_i^{(m)} = \left[\left(x_i^{(m)}\right)^2 + \epsilon\left|x_i^{(m)}\right|\right]^{-1}$$

Limitations of MAP

- Many local minima $O({}^{N}C_{M})$
  - May get stuck at a local minimum
- MAP only guarantees max $p(x = x_0|y)$
  - Probability mass, rather than mode, may be more relevant for continuous random vars
  - Perhaps posterior mean $E(x|y)$?
- Even with the true prior, MAP estimators do not minimize MSE: so MSE may be high!
  - In fact, using "true" statistics often does not lead to the lowest MSE!

To Recap

- Bayesian estimation
  - Basic MAP estimation
  - Majorization-minimization approach
  - Iterative reweighted algorithms
- Limitations
  - Many local minima
  - Posterior mean vs. posterior mode

Part 3: Sparse Bayesian Learning

Use lots of priors and pick the best one!

Point of Departure: Alternative Prior

- Need tractable representations for sparsity promoting priors
- Gaussian Scaled Mixtures (GSM)

$$\mathbf{x} = \sqrt{\Gamma} G; \ G \sim \mathcal{N}(\mathbf{g}; 0, 1)$$

$$p(\mathbf{x}) = \int p(\mathbf{x}|\gamma) p(\gamma) \mathrm{d}\gamma = \int \mathcal{N}(\mathbf{x}; 0, \gamma) p(\gamma) \mathrm{d}\gamma$$

  - γ: non-negative random variable, independent of G

Why GSMs?

- **Defn.:** A function $f(x)$ is **completely monotonic** on (a,b) if $(-1)^n f^{(n)}(x) \geq 0$, $n = 0, 1, \ldots$ where $f^{(n)}(x) = n^{th}$ order derivative

- **Theorem:** A density $p(x)$ can be represented by a GSM **iff** $p(x^{1/2})$ is **completely monotonic** on $(0, \infty)$

- Most sparse priors on x can be expressed using GSMs (incl. ones with concave g)
[Palmer et al., 2006]

Examples

- Laplacian density
  - We use: $$p(\gamma) = \frac{a^2}{2} \exp\left(-\frac{a^2}{2}\gamma\right),\ \gamma \geq 0$$
  - And get: $$p(x_i; a) = \frac{a}{2} \exp(-a|x_i|)$$
  - Which leads to the familiar LASSO problem
- Student's t distribution
  - We use: gamma distribution
  - And get:

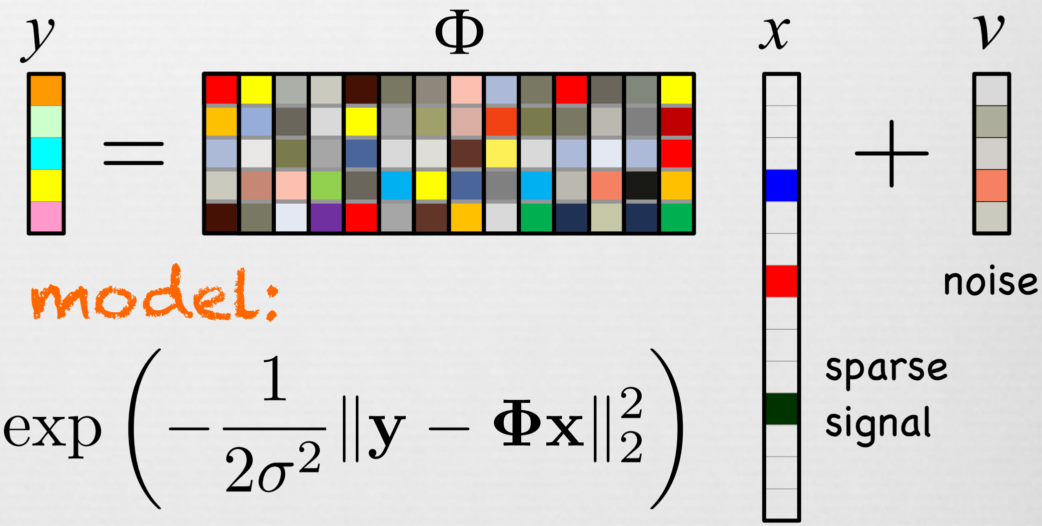
$$p(x_i; a, b) = \frac{b^a \Gamma(a + 1/2)}{\sqrt{2\pi}\Gamma(a)} \frac{1}{(b + x_i^2/2)^{a+1/2}}$$

Examples

- Generalized Gaussian
  - We use: positive alpha-stable density of order p/2
  - And get: $$p(x_i; p) = \frac{1}{2\Gamma\left(1 + \frac{1}{p}\right)} \exp(-|x_i|^p)$$
- Generalized logistic distribution
  - We use: A scale mixing density related to the Kolmogorov Smirnoff distance
  - And get: $$p(x_i; \alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \frac{\exp(-\alpha|x_i|)}{(1 + \exp(-|x_i|))^{2\alpha}}$$

Sparse Bayesian Learning

- Recall the canonical model



- Gaussian noise model:

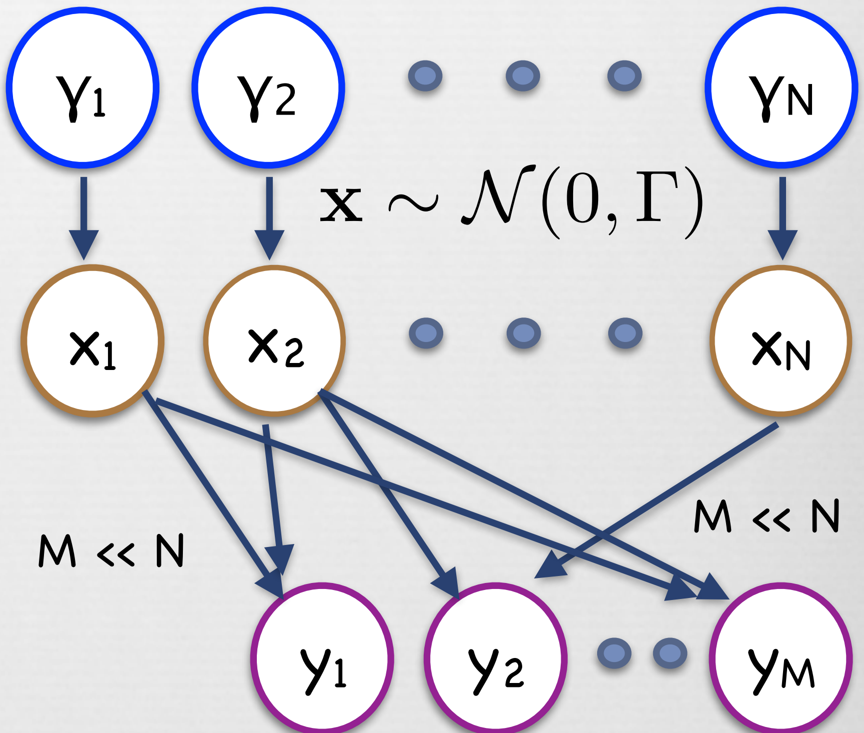
$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{\Phi x}\|_2^2\right)$$

- Parameterized Gaussian prior:

$$p(x_i; \gamma_i) = \frac{1}{\sqrt{2\pi\gamma_i}} \exp\left(-\frac{x_i^2}{2\gamma_i}\right), \; \gamma_i \geq 0$$

Graphical Model

- Markov chain: γ -> x -> y
- γ: nonnegative hyperparameters
- Potential advantages:
  - Given γ, $p(x|y;\gamma)$ is Gaussian: easy to find point estimates
  - Averaging over x -> fewer local minima in $p(\gamma|y)$
  - γ can be used to tie parameters together: fewer params. to estimate



$y = \Phi x + v$

Hierarchical Bayesian Framework

- First, estimate hyperparameters: $\hat{\gamma} = \arg\max_{\gamma} p(\gamma|\mathbf{y})$
  - γ : deterministic and unknown, or random with hyperprior distbn.
- Then, find posterior distribution p(x|y;$\hat{\gamma}$)

$$p(\mathbf{x}|\mathbf{y};\hat{\gamma}) = \mathcal{N}(\mu_x, \Sigma_x)$$

$$\mu_x = \hat{\Gamma}\Phi^T\left(\Phi\hat{\Gamma}\Phi^T + \lambda\mathbf{I}\right)^{-1}\mathbf{y}$$

$$\Sigma_x = \hat{\Gamma} - \hat{\Gamma}\Phi^T\left(\Phi\Gamma\Phi^T + \lambda\mathbf{I}\right)^{-1}\Phi\hat{\Gamma}$$

- For point estimates: e.g., posterior mean: $\mathbb{E}\left(\mathbf{x}|\mathbf{y};\hat{\gamma}\right)$

Sparse Bayesian Methods

- Estimate γ_i from the data: Type-II ML

$$\mathcal{L}(\Gamma) = \log p(\mathbf{y};\Gamma) = \log \int p(\mathbf{y}|\mathbf{x};\Gamma)p(\mathbf{x};\Gamma)\mathrm{d}\mathbf{x}$$

$$p(\mathbf{y};\Gamma) = \mathcal{N}\left(0, \underbrace{\sigma^2\mathbf{I} + \Phi\Gamma\Phi^T}_{\Sigma_{\mathbf{y}}}\right)$$

- When γ is random: can find MAP estimates
  - Just add $\sum_{i=1}^{N} \log p(\gamma_i)$ term to log likelihood fn
- SBL cost function: $\mathcal{L}(\Gamma) \propto -\log\det(\Sigma_{\mathbf{y}}) - \mathbf{y}^T\Sigma_{\mathbf{y}}^{-1}\mathbf{y}$

Optimization via EM

- Log likelihood of the complete data

$$-\log p(\mathbf{y},\mathbf{x};\gamma) = \underbrace{\frac{\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{x}\|_2^2}{2\sigma^2}}_{\substack{-\log p(\mathbf{y}|\mathbf{x};\gamma) \\ \text{indep. of } \gamma}} + \underbrace{\frac{1}{2}\left[\sum_{i=1}^{N}\frac{x_i^2}{\gamma_i} + \log\gamma_i\right]}_{\substack{-\log p(\mathbf{x};\gamma) \\ \text{func. of } \gamma}} \boxed{- \sum_{i=1}^{N}\log p(\gamma_i)}$$

Facilitates type-II algorithms

- E-Step: compute "Q-function"

$$Q\left(\Gamma|\Gamma^{(t)}\right) = \mathbb{E}_{\mathbf{x}|\mathbf{y};\Gamma^{(t)}}\left[-\log p(\mathbf{y},\mathbf{x};\Gamma)\right]$$

$$\doteq \sum_{i=1}^{N}\frac{\mathbb{E}(x_i^2|\mathbf{y};\Gamma^{(t)})}{\gamma_i} + \log\gamma_i$$

from previous iteration

- Easy to compute: $p(x_i|\mathbf{y};\Gamma^{(t)})$ is Gaussian

The EM Iterations

- **E-step (continued):** $p\left(\mathbf{x}|\mathbf{y};\Gamma^{(t)}\right) = \mathcal{N}(\mu, \Sigma)$

$$\mu = \sigma^{-2}\left(\sigma^{-2}\Phi^T\Phi + \left(\Gamma^{(t)}\right)^{-1}\right)^{-1}\Phi^T\mathbf{y} \qquad \Sigma = \left(\sigma^{-2}\Phi^T\Phi + \left(\Gamma^{(t)}\right)^{-1}\right)^{-1}$$

- **M-step:** maximize $Q(\Gamma|\Gamma^{(t)})$ given posteriors gathered in the E-step:

$\mathbb{E}(x_i^2|\mathbf{y};\Gamma^{(t)})$

$$\Gamma^{(t+1)} = \arg\max_{\gamma_i \geq 0} Q\left(\Gamma|\Gamma^{(t)}\right) = \boxed{\mathrm{diag}\left(\mu_i^2 + \Sigma_{ii}\right)}$$

- Component-wise updates

Can recover type-I methods by treating γ as hidden and taking expectation over γ instead of x

The SBL Algorithm

1. Initialize $\Gamma = I$
2. Compute $$\mu = \sigma^{-2}\left(\sigma^{-2}\Phi^T\Phi + \left(\Gamma^{(t)}\right)^{-1}\right)^{-1}\Phi^T\mathbf{y}$$ $$\Sigma = \left(\sigma^{-2}\Phi^T\Phi + \left(\Gamma^{(t)}\right)^{-1}\right)^{-1}$$
3. Update $\Gamma^{(t+1)} = \mathrm{diag}\left(\mu_i^2 + \Sigma_{ii}\right)$
4. Repeat steps 2 and 3
5. Output μ after convergence

Variational Interpretation

- Lower bound on L:

$$\mathcal{L}(\Gamma) = \log \int q_{\mathbf{x}}(\mathbf{x}) \frac{p(\mathbf{x}, \mathbf{y}; \Gamma)}{q_{\mathbf{x}}(\mathbf{x})} \mathrm{d}\mathbf{x}$$

$$\geq \int q_{\mathbf{x}}(\mathbf{x}) \log \left(\frac{p(\mathbf{x}, \mathbf{y}; \Gamma)}{q_{\mathbf{x}}(\mathbf{x})} \right) \mathrm{d}\mathbf{x}$$

$$\triangleq \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}); \Gamma)$$

Jensen's inequality

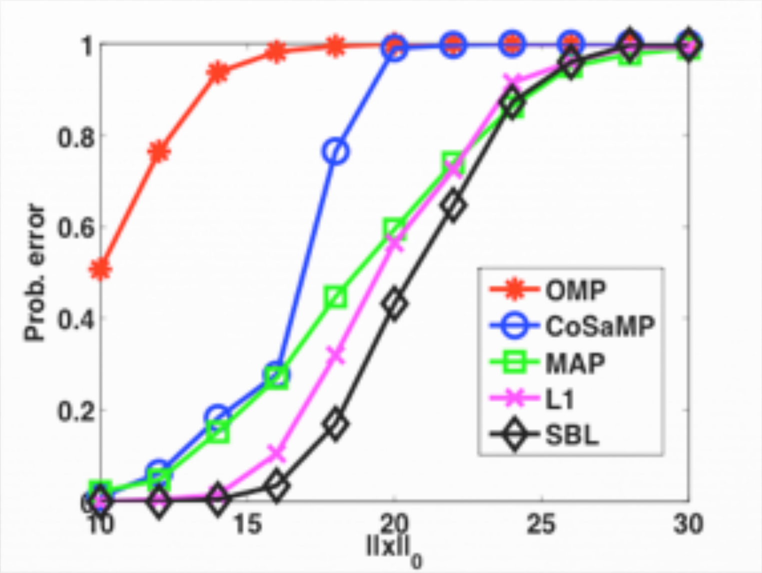
- In each iteration, EM maximizes the bound

Convergence

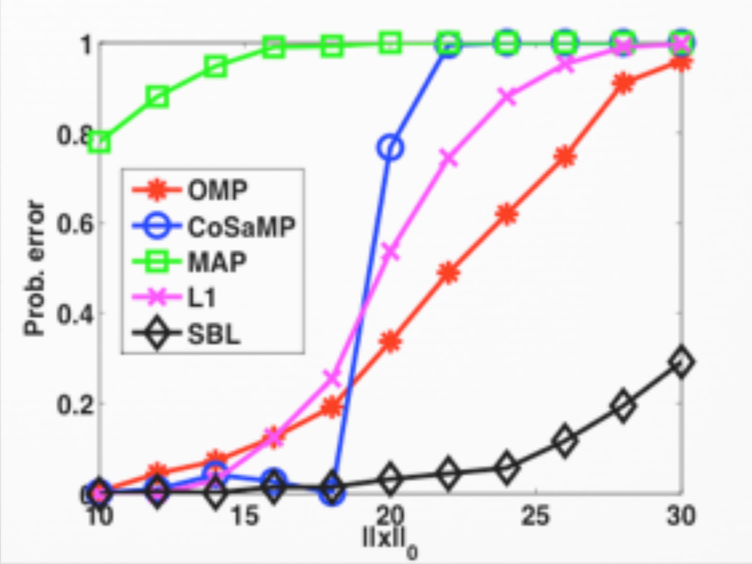
- Convergence guaranteed to a fixed pt. of L from any initialization (property of EM)
- The global min of L occurs at the sparsest solution in the noiseless case => no structural problems! [Wipf et al. 04]
- Attempts to estimate posterior p(x|y) in regions with significant mass
- All local minima occur at sparse solutions in the noisy case [Wipf et al. 04]
- Cost function much smoother than the associated MAP estimation: fewer local minima [Wipf and Nagarajan 09]

Recall Empirical Example

- Generate random 50 x 100 matrix Φ
- Generate sparse vector x_0
- Compute $y = \Phi x_0$
- Solve for x_0, average over 1000 trials
- Repeat for different sparsity values

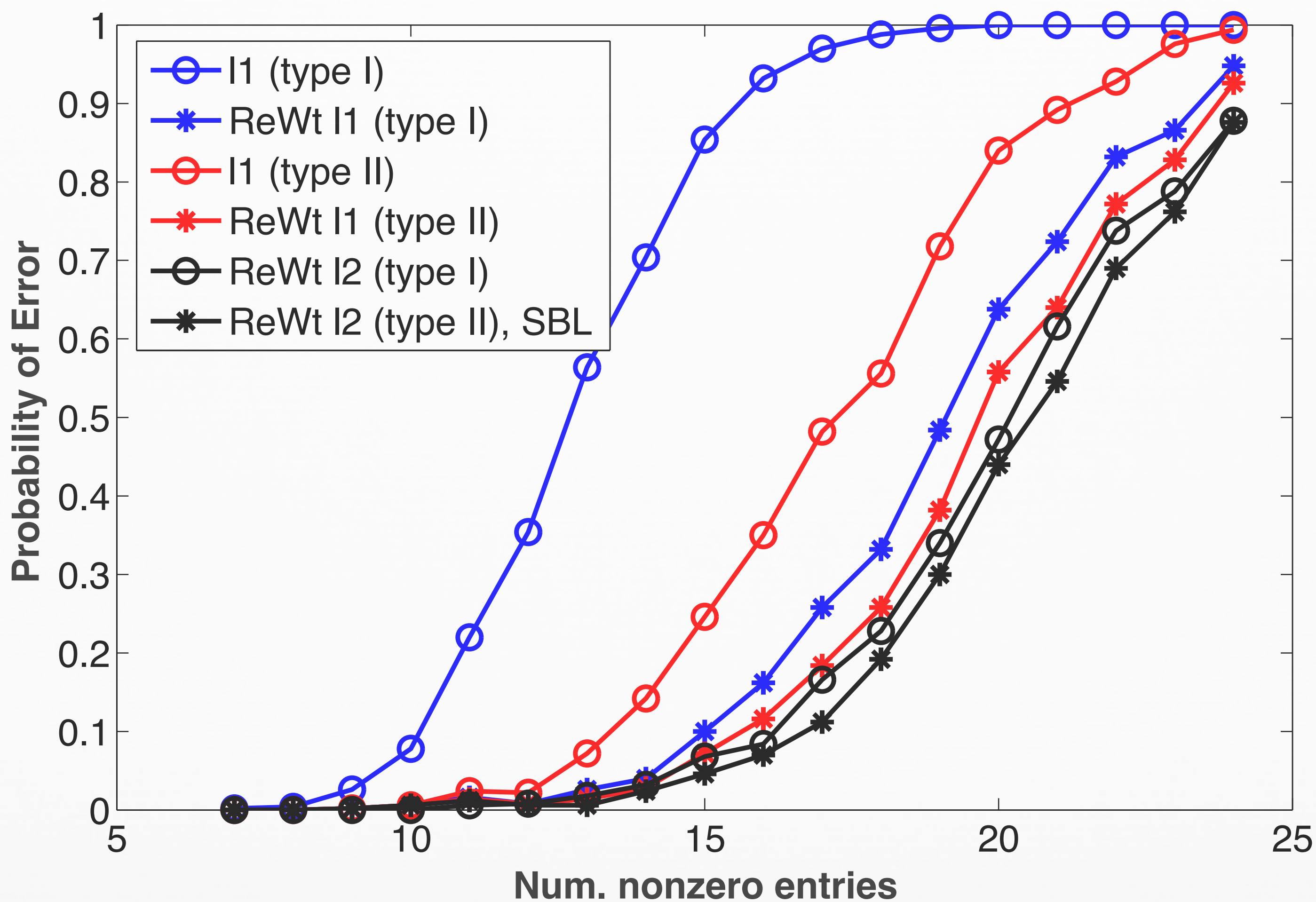


Unit magnitude entries



Highly scaled entries

Type I vs. Type II



Other Options for SBL Cost Min.

- McKay updates [Tipping, 2001]
  - Set gradient of SBL cost = 0
  - Faster convergence than EM
- Greedy approach:
  - Update hyperparams one at a time [Tipping & Faul, 2003]
    - Closed-form update for each hyperparam
    - Fast, but can get trapped in a local min.
  - Fast Bayesian matching pursuit [Schniter et al., 08]

Other Options for SBL Cost Min.

- Use dual-form of SBL. Cost function:

$$\mathbf{x}_{\text{opt}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \sigma^2 g_{\text{SBL}}(\mathbf{x})$$

$$g_{\text{SBL}}(\mathbf{x}) \triangleq \min_{\gamma \geq 0} \mathbf{x}^T \Gamma^{-1} \mathbf{x} + \log\det\left(\sigma^2 \mathbf{I} + \Phi\Gamma\Phi^T\right)$$

  - Facilitates iterative reweighted l_1 and l_2 algorithms [Wipf and Nagarajan, 09]
  - Overcomes some limitations of EM
- Replace E-step with an approx. posterior computation: AMP-SBL [Al-Shoukairi and Rao 14]

Approximate Message Passing

- AMP [Donoho, Maleki, Montanari 09]:
  - Uses loopy belief propagation + Gaussian approximations to solve LASSO
  - Key advantage: low complexity
- In SBL:
  - All Gaussian PDFs: approximation is not necessary
  - Only need to track means and variances
  - Can replace computationally expensive E-step with the AMP based iterations

Factor Graph

- In the E-Step, we're after

$$p\left(\mathbf{x}|\mathbf{y};\Gamma^{(t)}\right) \propto p(\mathbf{y}|\mathbf{x})p\left(\mathbf{x};\Gamma^{(t)}\right)$$

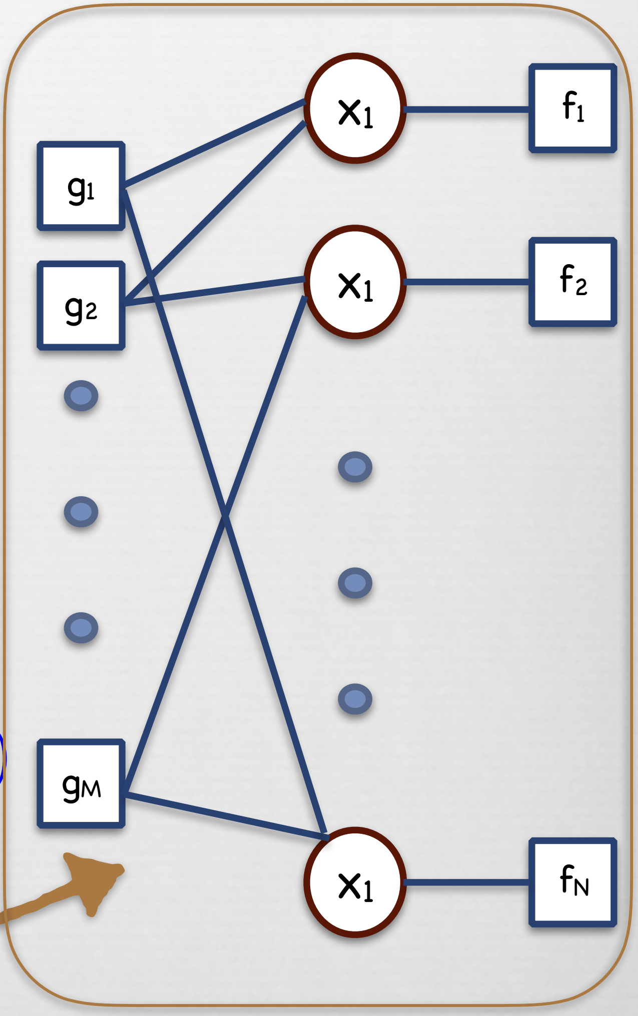
$$\propto \prod_{m=1}^{M} p(y_m|\mathbf{x}) \prod_{n=1}^{N} p(x_n;\gamma_n^{(t)})$$

- And we define

$$g_m(\mathbf{x}) \triangleq p(y_m|\mathbf{x}) = \mathcal{N}(y_m;\Phi_m^H\mathbf{x},\sigma^2)$$

$$f_n(x_n) \triangleq p(x_n;\gamma_n) = \mathcal{N}(x_n;0,\gamma_n)$$

- To get the factor graph



AMP-SBL

- General form of updates:

$$\hat{\mathbf{x}}^{t+1} = \eta_t\left(\Phi^H \mathbf{z}^t + \hat{\mathbf{x}}^t\right)$$

$$\mathbf{z}^t = \mathbf{y} - \Phi\hat{\mathbf{x}}^t + \boxed{\frac{1}{\delta}\mathbf{z}^{t-1}\langle {\eta'}_{t-1}\left(\Phi^H \mathbf{z}^{t-1} + \hat{\mathbf{x}}^{t-1}\right)\rangle}$$

Message passing term

- η_t: soft-thresholding function – linear for SBL

- O(M+N) msg updates: low computational cost!

Definitions:

$$F_n(K_n, c) = K_n\left(\frac{\gamma_n}{c+\gamma_n}\right)$$

$$G_n(K_n, c) = \frac{c\gamma_n}{c+\gamma_n}$$

$$F'_n(K_n, c) = \frac{\gamma_n}{c+\gamma_n}$$

Message Updates:

$$K_n = \sum_{m=1}^{M} \Phi^*_{mn} z_m + \mu_n$$

$$\mu_n = F_n(K_n, c)$$

$$v_n = G_n(K_n, c)$$

$$c = \sigma^2 + \frac{1}{M}\sum_{n=1}^{N} v_n$$

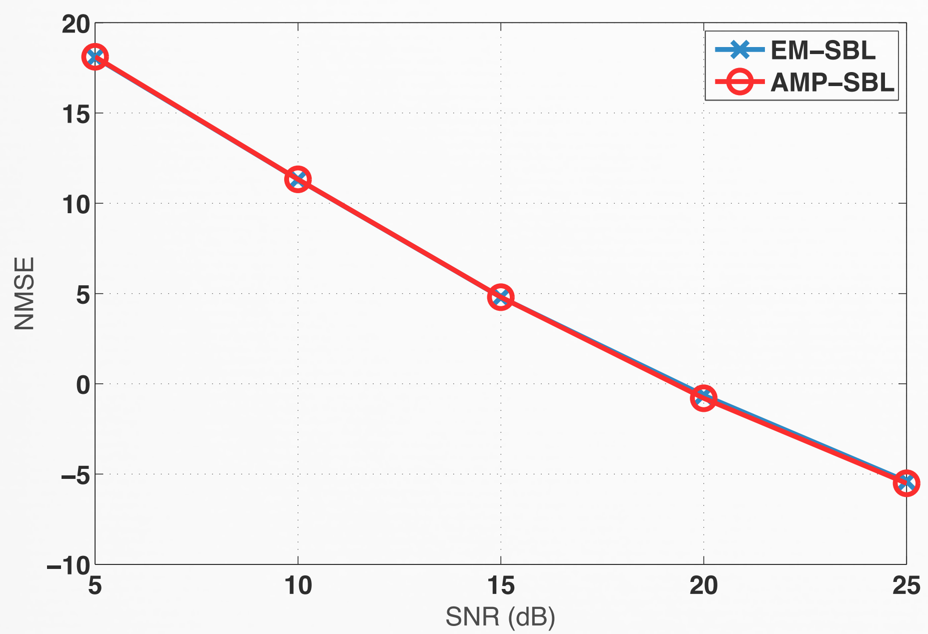
$$z_m = y_m - \sum_{n=1}^{N} \Phi_{mn}\mu_n + \frac{z_m}{M}\sum_{n=1}^{N} F'_n(\mu_n, c)$$

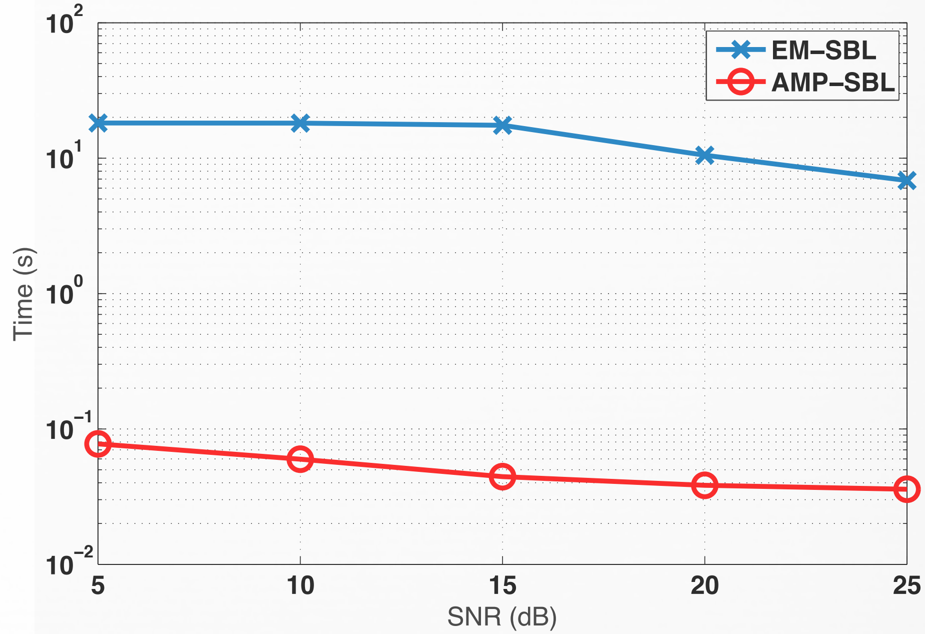
Parameter Update/M-Step:

$$\gamma_n = v_n + \mu_n^2$$

Empirical Example

- N = 200, M = 100, K = 20, Gaussian measurement matrix





Advantages of SBL

- Averaging over x: fewer minima in $p(y;\gamma)$
- Get an estimate of the error in recovery
- Allows for "exact inference"
- Versatile: γ can also be used to tie several params. together - easier to estimate
- Useful extensions: incorporate structure
  - Intra/inter-vector correlation
    - SBL allows the use of Kalman framework
  - Block/cluster sparsity
  - Colored noise (rank-deficient cov.)

To Recap

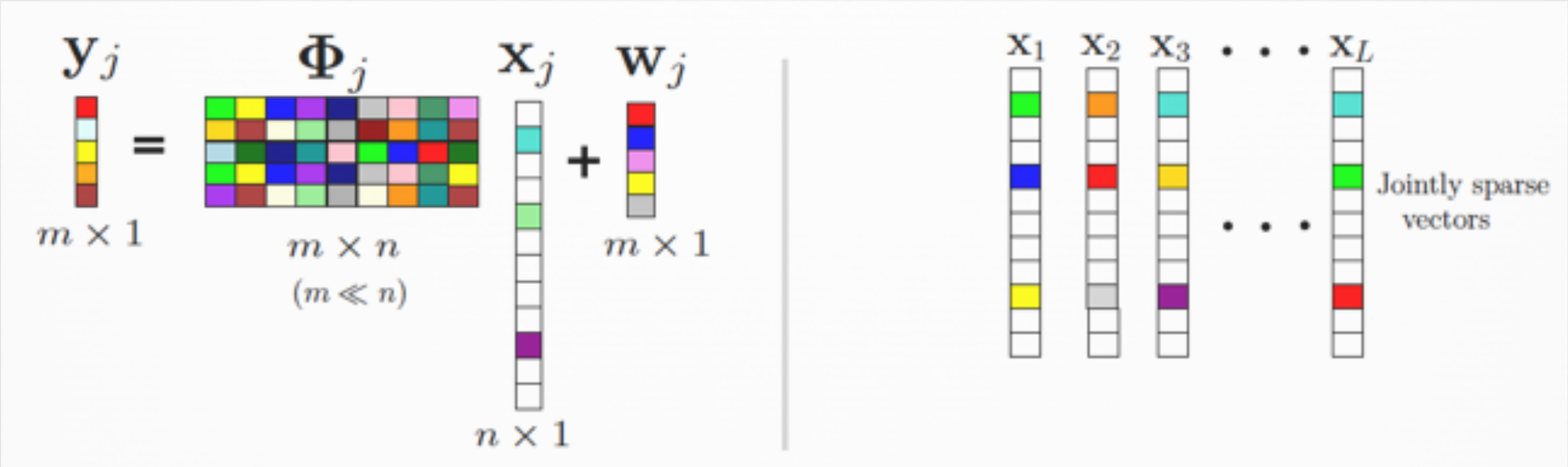
- Sparse Bayesian learning
  - Sparse vector recovery via estimating hyperparameters
  - Expectation-maximization iterations
  - Convergence properties
  - Alternative implementations
- Limitations
  - Computational complexity
    - More recent AMP-based algos overcome this
  - Slow convergence
    - Fast versions exist, but without the same convergence guarantees

Part 4: Extensions

1. Multiple measurement vectors
2. Cluster-sparsity, inter-vector correlation
3. Distributed sparse signal recovery
4. Deep learning

Multiple Measurement Vectors

- Observation Model



- Why? Multiple measurements can provide complementary information

- Joint Prior $p(\mathbf{x}_j;\Gamma) = \mathcal{N}(0,\Gamma),\ j = 1,\ldots,L$

Algos for Joint-Sparse Recovery

- M-OMP [Tropp et al., 06]
- M-BP [Cotter et al. 05, Malioutov et al. 05]

(l1 norm across rows)

Num. measurements

$$\min_{\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_L\}} \sum_{l=1}^{L} \|\mathbf{y}_l - \Phi_l \mathbf{x}_l\|_2^2 + \lambda \sum_{i=1}^{N} \|\mathbf{x}_i^T\|_2$$

(l2 norm of ith row)

- M-Jeffreys [Figueiredo 02, Rao et al. 97, Candes et al. 08]

$$\min_{\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_L\}} \sum_{l=1}^{L} \|\mathbf{y}_l - \Phi_l \mathbf{x}_l\|_2^2 + \lambda \sum_{i=1}^{N} \log \|\mathbf{x}_i^T\|_2$$

- M-FOCUSS [Rao et al. 03, Cotter et al. 05, Chen et al. 09]

$$\min_{\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_L\}} \sum_{l=1}^{L} \|\mathbf{y}_l - \Phi_l \mathbf{x}_l\|_2^2 + \lambda \sum_{i=1}^{N} \left(\|\mathbf{x}_i^T\|_2\right)^p, \; p < 1$$

The M-SBL Algo

- Cost function

$$p(\mathbf{Y};\gamma) = \int p(\mathbf{Y},\mathbf{X};\gamma)\mathrm{d}\mathbf{X} = \prod_{j=1}^{L}\int p(\mathbf{y}_j|\mathbf{x}_j)p(\mathbf{x}_j;\gamma)\mathrm{d}\mathbf{x}_j$$

- Key point: γ couples the sparsity pattern across x_j
  - Fewer parameters to estimate: N << (N x L)
- EM Iterations

$$\text{E-step: } Q(\gamma|\gamma^k) = \mathbb{E}_{\mathbf{X}|\mathbf{Y},\gamma^k}\left[\log p(\mathbf{Y},\mathbf{X};\gamma)\right]$$

$$\text{M-step: } \gamma^{k+1} = \arg\max_{\gamma\in\mathbb{R}_+^N} Q(\gamma|\gamma^k)$$

- Posterior distbn.: $p\left(\mathbf{x}_j|\mathbf{y}_j;\gamma^k\right) \sim \mathcal{N}\left(\mu_j^{k+1},\Sigma_j^{k+1}\right)$

E & M Steps

- E Step:

$$\Sigma_j^{k+1} = \Gamma^k - \Gamma^k \Phi_j^T \left(\sigma_j^2 \mathbf{I}_M + \Phi_j \Gamma^k \Phi_j^T\right)^{-1} \Phi_j \Gamma^k$$

$$\mu_j^{k+1} = \sigma_j^{-2} \Sigma_j^{k+1} \Phi_j^T \mathbf{y}_j$$

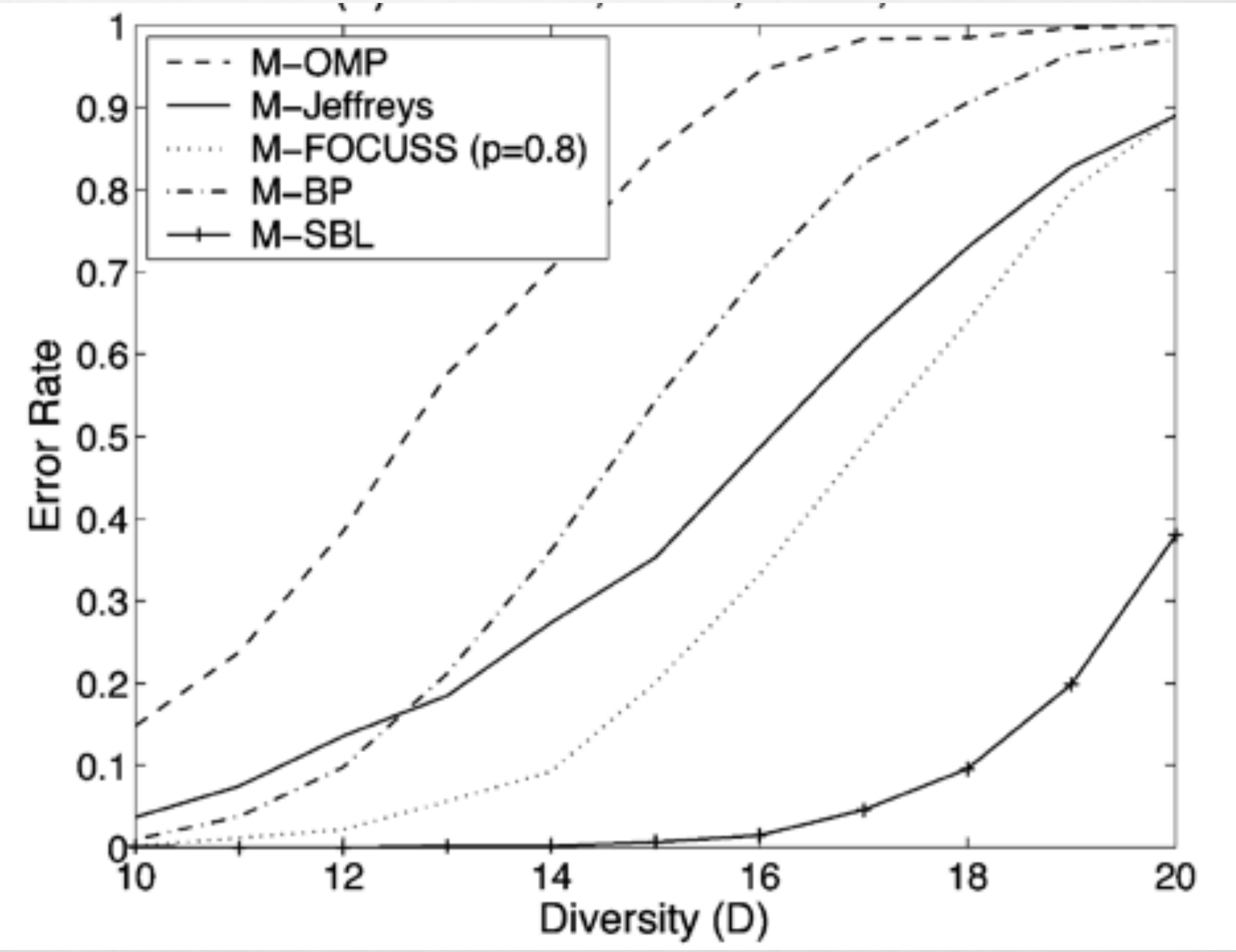
- M Step:

$$\gamma^{k+1}(i) = \frac{1}{L} \sum_{j=1}^{L} \mu_j^{k+1}(i)^2 + \Sigma_j^{k+1}(i,i)$$

  - Average of the individual estimates of γ_i across measurements

Empirical Example

- $M = 25$
  $N = 50$
  $L = 3$



[Wipf & Rao, 07]

Analysis: Failure of Standard Sparse Regression

- Let $\tilde{\mathbf{X}}_0 \in \mathbb{R}^{k \times L}$ = nonzero rows in X_0, and $\Phi_j = \Phi \, \forall \, j$
- Suppose $\tilde{\mathbf{X}}_0 \tilde{\mathbf{X}}_0^T$ is full rank ($L \geq k$), & $\mathbf{Y} = \mathbf{\Phi X}_0 = \tilde{\mathbf{\Phi}} \tilde{\mathbf{X}}_0$
- Lemma: [Wipf et al. 11]
  - There exist Φ, X_0 such that solving

$$\min_{\mathbf{X}} \sum_{i=1}^{N} g_i(\|\mathbf{x}_i^T\|_2) \text{ s. t. } \mathbf{Y} = \mathbf{\Phi X}_0 = \mathbf{\Phi X}$$

    for any possible g_i will have solutions NOT equal to X_0!
- Sparse regression can fail!

Analysis: Success of MUSIC

- When $\tilde{\mathbf{X}}_0\tilde{\mathbf{X}}_0^T$ is full rank, $\mathrm{span}[\mathbf{Y}] = \mathrm{span}[\tilde{\mathbf{\Phi}}]$
- MUSIC algorithm:
  - Compute $\epsilon_i = \min_{\alpha} \|\phi_i - \mathbf{Y}\alpha\|_2 \quad \forall \phi_i \in \mathbf{\Phi}$
  - Index i is in the support iff $\varepsilon_i = 0$
- Result: MUSIC is guaranteed to estimate the correct support whenever $\tilde{\mathbf{X}}_0\tilde{\mathbf{X}}_0^T$ is full rank!

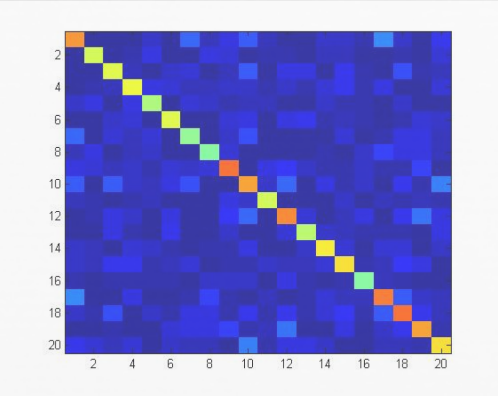
[Feng and Bresler 96]

Hybrid Algorithms

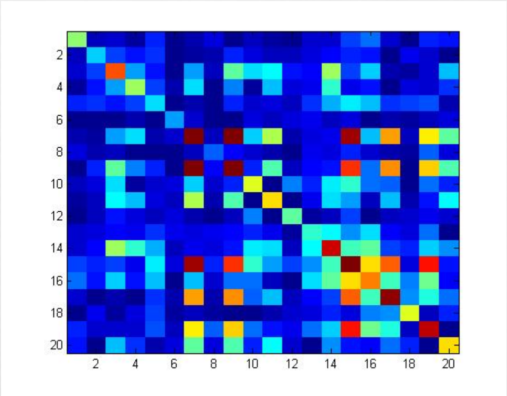
- Combine MUSIC and sparse recovery

  [Davies and Eldar, 2012; Kim et al., 2012; Lee et al., 2012]

- MUSIC only works if $L \geq k$
- Sparse recovery can sometimes work even if $L < k$
- Problem: correlated columns in Φ
-



Easy: $\Phi^T \Phi \approx \mathbf{I}$



Hard: $\Phi^T \Phi \neq \mathbf{I}$

Compensating for Dictionary Structure

- Simple example: building column norm invariance
  Let $\alpha_i \triangleq \|\Phi_i\|_2$ and $g(\mathbf{X}, \alpha) \triangleq \sum_{i=1}^{N} \alpha_i \|\mathbf{x}_i^T\|_2$
  Then, the problem
  $$\min_{\mathbf{x}} \|\mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}\|_2^2 + \lambda g(\mathbf{X}, \alpha)$$
  is invariant to dictionary column norms.

- So what about some fn. g that depends on the correlation structure $\Phi^T\Phi$?

Analysis of M-SBL Cost

- M-SBL is equivalent to solving

$$\min_{\mathbf{X}} g(\mathbf{X}; \mathbf{\Phi}^T \mathbf{\Phi}) \text{ s. t. } \mathbf{Y} = \mathbf{\Phi} \mathbf{X}_0 = \mathbf{\Phi} \mathbf{X}$$

Incorporates correlation structure into the cost function

- Result: Unique stationary point X_0 when:
  - Rows of X_0 sufficiently uncorrelated, OR
  - Sorted row norms of X_0 decay sufficiently fast

[Min & Wipf 15; Wipf et al. 15]

- True even under correlated dictionaries
- But failure still possible when MUSIC succeeds…

Augmented M-SBL Model

- Modified likelihood function:

$$p(\mathbf{Y}|\mathbf{X};\mathbf{\Psi}) \propto \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{\Phi}\mathbf{X}\mathbf{\Psi}\|_F^2\right]$$

- Posterior distribution is Gaussian with mean

$$\hat{\mathbf{X}} = \mathbb{E}_{p(\mathbf{X}|\mathbf{Y};\mathbf{\Gamma},\mathbf{\Psi})}[\mathbf{X}] \approx \mathbf{\Gamma}\mathbf{\Phi}^T\left(\mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T+\sigma^2\mathbf{I}\right)^{-1}\mathbf{Y}\mathbf{\Psi}$$

If Γ is sparse, so is the mean

- Estimate both Γ and Ψ via marginalization:

$$\max_{\mathbf{\Psi},\mathbf{\Gamma}\geq 0} \int p(\mathbf{Y}|\mathbf{X};\mathbf{\Psi})p(\mathbf{X};\mathbf{\Gamma})d\mathbf{X}$$

[Wipf et al. 15]

Analysis of A-SBL

- Augmented SBL is equivalent to solving

$$\min_{\mathbf{X},\mathbf{\Psi}} g_{\text{aug}}(\mathbf{X},\mathbf{\Psi};\mathbf{\Phi}^T\mathbf{\Phi}) \text{ s.t. } \mathbf{Y}=\mathbf{\Phi}\mathbf{X}_0=\mathbf{\Phi}\mathbf{X}\mathbf{\Psi}$$

for some g_{aug}. Moreover,

1. Have unique stationary point at $X^*\Psi^*$ if
$$\tilde{\mathbf{X}}_0\tilde{\mathbf{X}}_0^T = \text{full rank}$$
2. For any fixed Ψ, have unique stationary point at $X^*\Psi^* = X_0$ if sorted row norms of $X_0\Psi$ decay sufficiently fast

- Exploits both signal and dictionary correlation

Empirical Evaluation

- Generate correlated dictionary

$$\mathbf{\Phi} = \sum_{i=1}^{m} \frac{1}{i}\mathbf{a}_i\mathbf{b}_i^T; \quad \mathbf{a}_i, \mathbf{b}_i \rightarrow \text{ iid } \mathcal{N}(0,1)$$

- Generate correlated X_0, varying rank

$$\tilde{\mathbf{X}}_0 = \sum_{i=1}^{L} \frac{1}{i}\mathbf{u}_i\mathbf{v}_i^T \quad \mathbf{u}_i, \mathbf{v}_i \rightarrow \text{ iid } \mathcal{N}(0,1)$$

- Compute observations

$$\mathbf{Y} = \mathbf{\Phi}\mathbf{X}_0$$

- Compare algos as problem dimensions change

Results

Fixed:

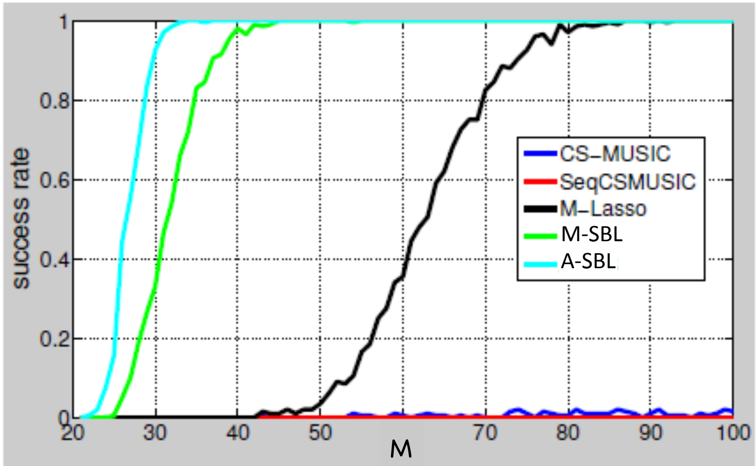
N = 200 (num. cols.)
k = 20 (row sparsity)

Varying:

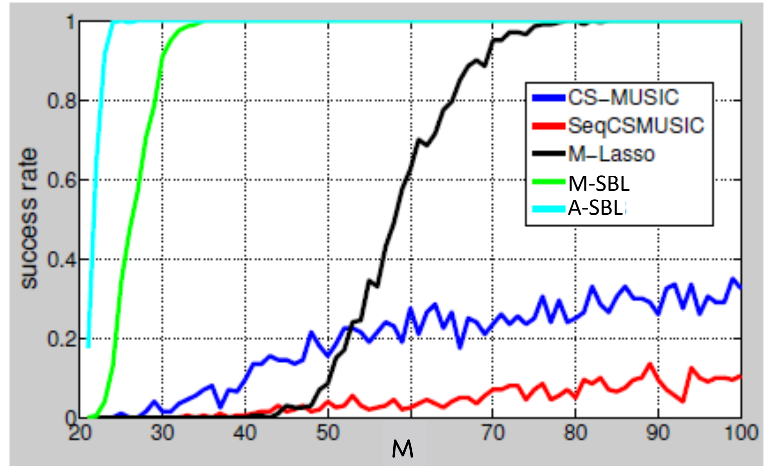
M = num. rows
L = num. cols in X_0

A-SBL outperforms existing algos, including MUSIC and convex LASSO based methods!

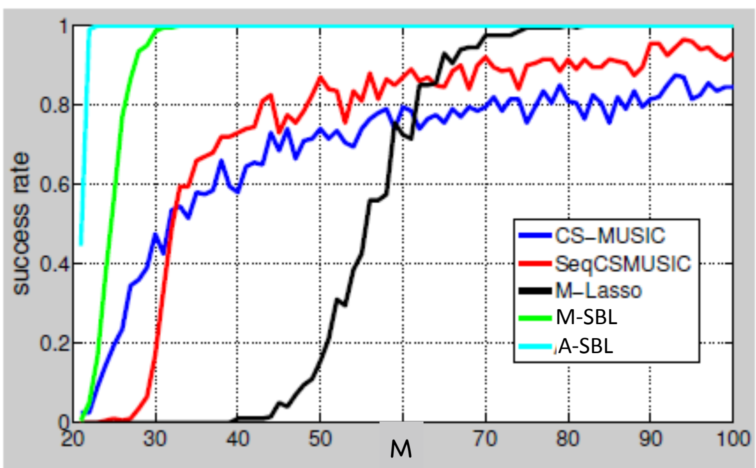
No other existing algo has similar guarantees.



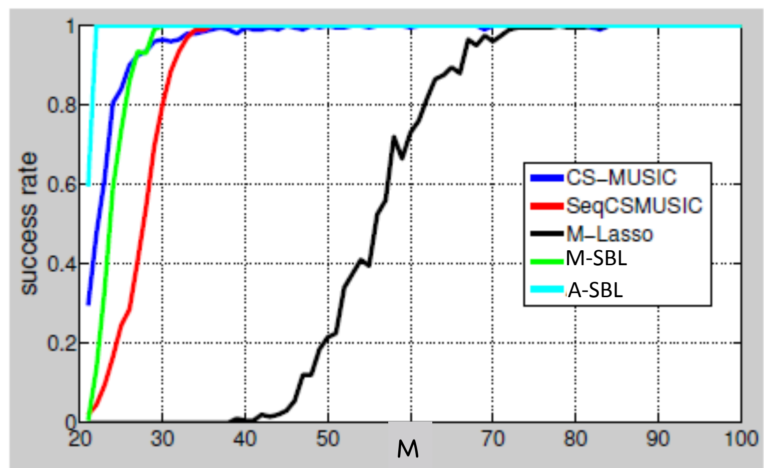
(a) L= 4



(b) L= 8



(c) L= 12



(d) L= 16

Clustered MMV Model

- Another twist: Suppose x_1, …, x_L (tasks) belong to K clusters, K < L
- Common support within each cluster
- Ω_k: column indices of X corresponding to cluster k, unknown
- Objective: Membership of each x_j?

Clustered SBL Model

- Gaussian likelihood: $p(\mathbf{Y}|\mathbf{X}) \propto \prod_j \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{y}_j - \mathbf{\Phi}_j\mathbf{x}_j\|_2^2\right]$
- Prior distribution: $p(\mathbf{X}|\mathbf{\Lambda},\mathbf{W}) \propto \prod_j \exp\left[-\frac{1}{2}\mathbf{x}_j^T\mathbf{\Gamma}_j^{-1}\mathbf{x}_j\right]$
  - Hyperparameters: $\mathbf{\Lambda} \in \mathbb{R}^{N\times K}$, $\mathbf{W} \in \mathbb{R}^{L\times K}$
  - W: rows lie in simplex $\mathcal{S} \triangleq \left\{\mathbf{w}_j^T : \sum_k w_{j,k} = 1,\ w_{j,k} \in [0,1]\right\}$
  - Covariance Γ_j diagonal: $\mathbf{\Gamma}_j^{-1} = \sum_k w_{j,k}\Lambda_k^{-1}$
    where Λ_k = diag(k^{th} column of Λ)

[Wang et al. 15]

Optimization Problem

- Posterior distbn.: Gaussian with mean

$$\hat{\mathbf{x}}_j = \mathbf{\Gamma}_j \mathbf{\Phi}_j^T \big(\underbrace{\sigma^2 \mathbf{I} + \mathbf{\Phi}_j \mathbf{\Gamma}_j \mathbf{\Phi}_j^T}_{\triangleq \Sigma_{y_j}}\big)^{-1} \mathbf{y}_j$$

- Can compute MAP estimates via

$$\max_{\Lambda > 0, \mathbf{W} \in \mathcal{S}} \int p(\mathbf{Y}|\mathbf{X}) p(\mathbf{X}; \Lambda, \mathbf{W}) p(\Lambda) p(\mathbf{W}) \mathrm{d}\mathbf{X}$$

Will design ρ to promote clustering

- Assuming $p(\Lambda) = 1; p(\mathbf{W}) \propto \exp(-\frac{1}{2}\rho(\mathbf{W}))$, equivalent problem

$$\max_{\Lambda > 0, \mathbf{W} \in \mathcal{S}} \sum_j \left[\mathbf{y}_j^T \Sigma_{\mathbf{y}_j}^{-1} \mathbf{y}_j + \log |\Sigma_{\mathbf{y}_j}| \right] + \sum_{j,k} \rho(w_{j,k})$$

Cost Function

- Determinant identities and Jensen's inequality: get upper bound on the cost function:

$$\mathcal{L}(\Lambda, \mathbf{W}) \triangleq \sum_j \left[\mathbf{y}_j^T \Sigma_{\mathbf{y}_j}^{-1} \mathbf{y}_j\right] + \sum_{j,k} \rho(w_{j,k}) + \sum_j \log \left|\sum_k w_{j,k} \Lambda_k^{-1} + \frac{1}{\sigma^2} \Phi_j^T \Phi_j\right| + \sum_{j,k} w_{j,k} \log |\Lambda_k|$$

- Can be optimized using majorization-minimization
- How to choose $\rho(w)$?
  - Examples: $\rho(w) = \beta w \log w, \quad \rho(w) = \beta |w|^2,$ etc.
  - Convex over [0,1]: favors sharing of basis functions along cols of W or merges Λ_k together - desirable

Low Noise Cost Function Behavior

- Assume that an optimal solution X* to

$$\min_{\mathbf{X}} \sum_j \|\mathbf{x}_j\|_0 \text{ s.t. } \mathbf{y}_j = \mathbf{\Phi}_j \mathbf{x}_j, \forall j$$

  exists with $\|\mathbf{x}_j^*\|_0 < N$ and $\text{spark}[\mathbf{\Phi}_j] = N+1, \forall j$

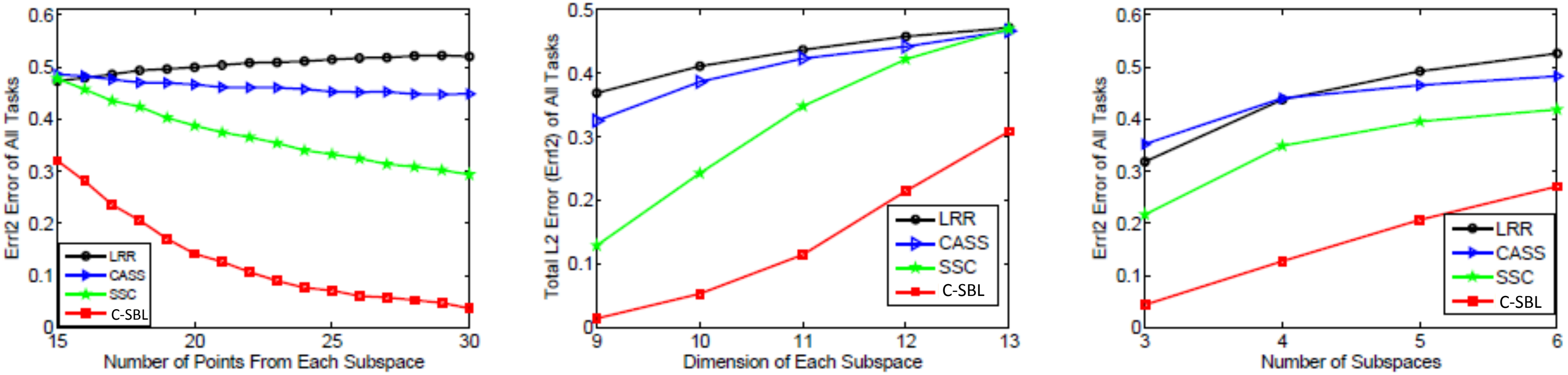
- Let Λ^*, W* denote any global solution to

$$\lim_{\sigma^2 \to 0} \inf_{\Lambda > 0, \mathbf{W} \in \mathcal{S}} \mathcal{L}(\Lambda, \mathbf{W})$$

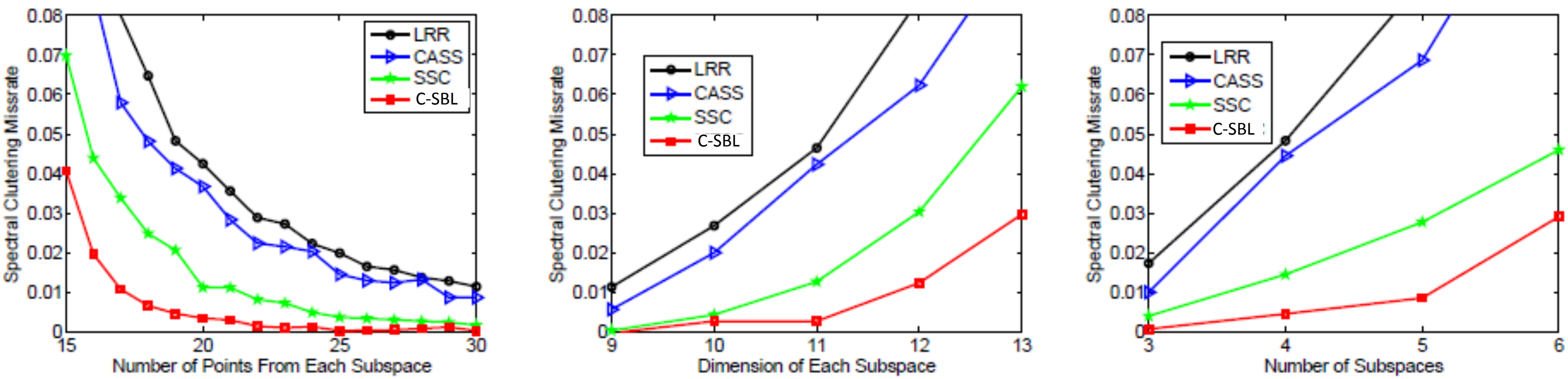
  Then, $\hat{\mathbf{x}}_j = \Gamma_j^* \mathbf{\Phi}_j (\mathbf{\Phi}_j \Gamma_j^* \mathbf{\Phi}_j^T)^\dagger \mathbf{y}_j$, with $\Gamma_j^* = \left(\sum_k w_{j,k}^* (\Lambda_k^*)^\dagger \right)^\dagger$ forms a globally optimal solution to

$$\min_{\mathbf{X}} \sum_j \|\mathbf{x}_j\|_0 \text{ s.t. } \mathbf{y}_j = \mathbf{\Phi}_j \mathbf{x}_j, \forall j$$

Experimental Results



Mean-squared reconstruction error



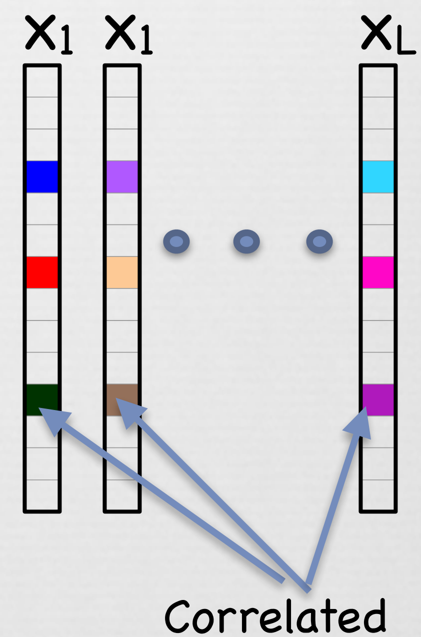
Clustering error

Remarks

- Demonstrated that M-SBL can be adapted for subspace segmentation
- A simple, novel, empirical prior is justified using properties of the resulting cost function
- The associated analysis promotes understanding of the central mechanisms that lead to successful subspace clustering

Inter-Vector Correlation

- Temporal correlation is usually present, and should be exploited
- Better, faster recovery
- Model correlation using a first order autoregressive process:

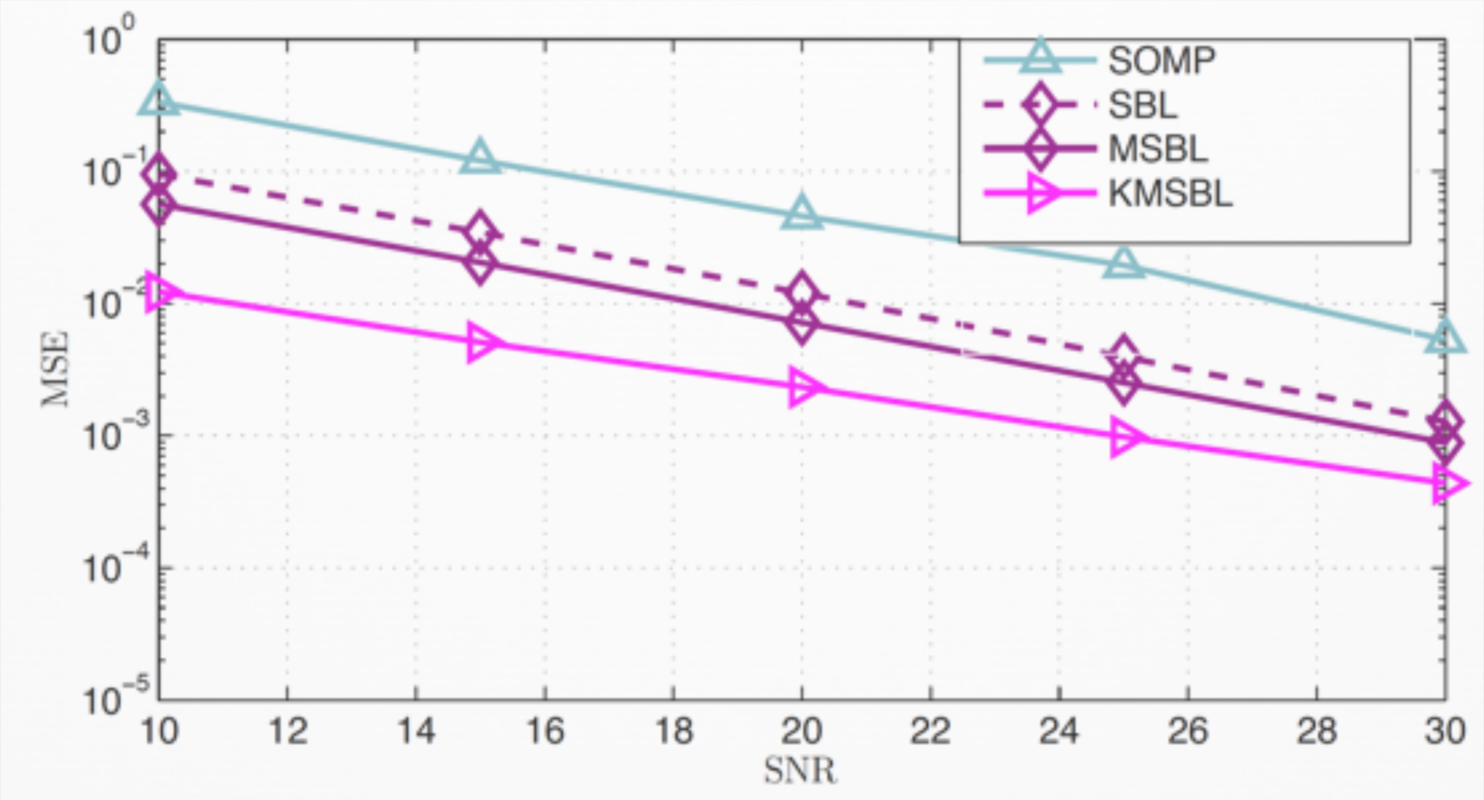


$$x_{(i,l+1)} = \sqrt{\gamma_i} h_{(i,l+1)} \text{ and } h_{(i,l+1)} = \rho h_{(i,l)} + \sqrt{1-\rho^2}\epsilon_{(i,l)},\ l = 1,\ldots,L$$

Inter-Vector Correlation: EM Algorithm

- E-Step: $$Q(\gamma|\gamma^r) = \mathbb{E}_{\mathbf{x}_1,\ldots\mathbf{x}_L|\mathbf{Y};\gamma^r}[\log p(\mathbf{Y},\mathbf{x}_1,\ldots\mathbf{x}_L;\gamma)]$$
$$= \mathbb{E}\left[\sum_{l=1}^{L}\log p(\mathbf{y}_l|\mathbf{x}_l) + \sum_{l=1}^{L}\log p(\mathbf{x}_l|\mathbf{x}_{l-1};\gamma)\right]$$
  - Requires computation of fixed-interval smoothed estimates
  - Efficient recursive implementation via Kalman smoothing [Prasad et al. TSP 2014]
- M-Step: Decouples as in the single measurement case: simple update rule

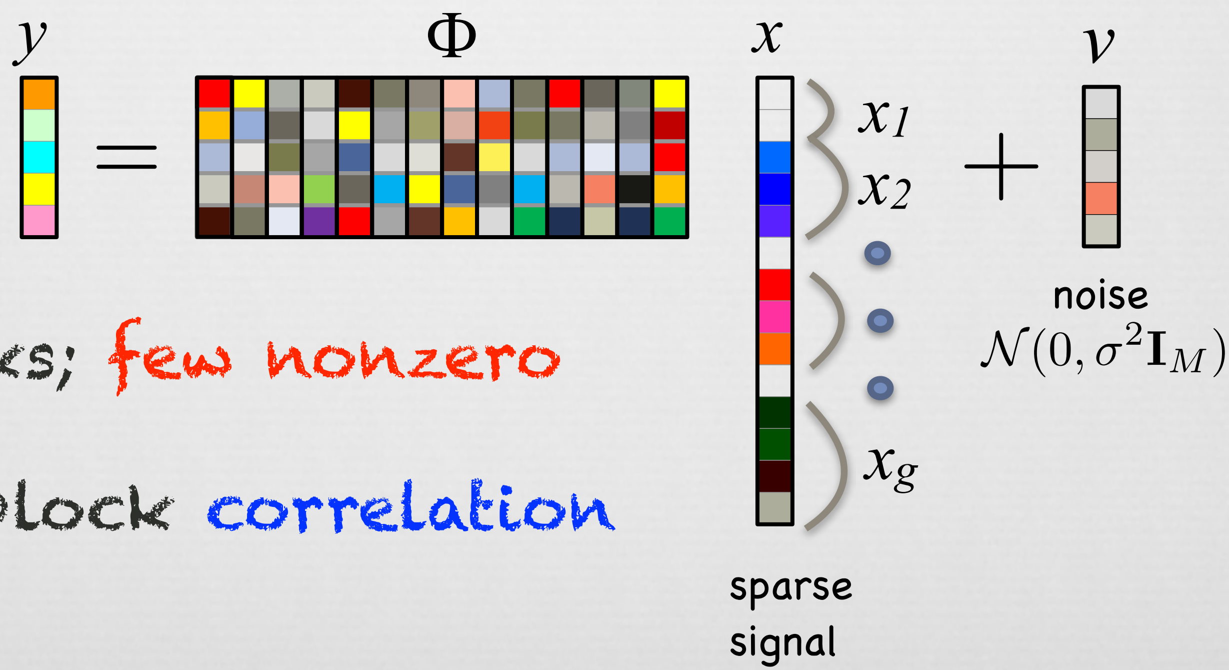
Simulation Result



$N = 64, M = 44, K = 30, L = 7, \rho = 0.999$

Block Sparsity & Intra-Block Correlation

- Intra-vector correlation is often present, and is important to model & exploit



- g blocks; few nonzero
- Intra-block correlation

Block-Sparse Bayesian Learning Framework

- Measurement model: $\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$

$$\mathbf{x} = [\underbrace{x_1, \dots x_{d_1}}_{\mathbf{x}_1^T}, \dots, \underbrace{x_{d_{g-1}+1}, \dots x_{d_g}}_{\mathbf{x}_g^T}]^T$$

- Parameterized prior

$$p(\mathbf{x}_i; \gamma_i, \mathbf{B}_i) \sim \mathcal{N}(0, \gamma_i \mathbf{B}_i),\ i = 1, 2, \dots, g$$

  - γ_i controls sparsity
  - $\mathbf{B}_i$ controls intra-block correlation

Optimization Problem

- Posterior distribution

$$p\left(\mathbf{x}|\mathbf{y};\sigma^2,(\gamma_i\mathbf{B}_i)_{i=1}^g\right) \sim \mathcal{N}(\mu_x,\Sigma_x)$$

  - where $\mu_x = \Sigma_0\Phi^T(\sigma^2\mathbf{I}+\Phi\Sigma_0\Phi^T)^{-1}\mathbf{y}$

$$\Sigma_x = \Sigma_0 - \Sigma_0\Phi^T(\sigma^2\mathbf{I}+\Phi\Sigma_0\Phi^T)^{-1}\Phi\Sigma_0$$

$$\Sigma_0 = \mathrm{diag}(\gamma_1\mathbf{B}_1,\ldots,\gamma_g\mathbf{B}_g)$$

- All params. can be estimated by maximizing:

$$\mathcal{L}(\Theta) = -2\log\int p(\mathbf{y}|\mathbf{x};\sigma^2)p(\mathbf{x};\Sigma_0)\mathrm{d}\mathbf{x}$$

$$= \log\det\left(\sigma^2\mathbf{I}+\Phi\Sigma_0\Phi^T\right) + \mathbf{y}^T\left(\sigma^2\mathbf{I}+\Phi\Sigma_0\Phi^T\right)^{-1}\mathbf{y}$$

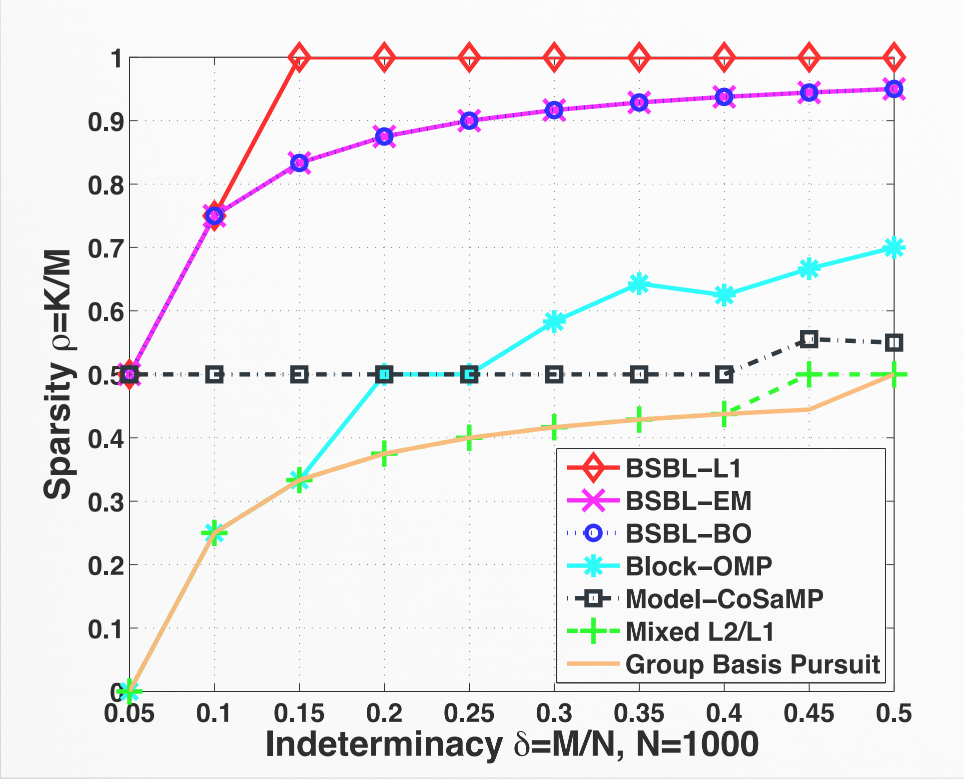
Several Options for Optimization

- BSBL-EM: Use expectation-maximization
- BSBL-BO: Use bounded optimization, i.e., majorization-minimization
- BSBL-L1: Use a reweighted L1 procedure (special case of BSBL-BO)
- Different strategies offer a variety of performance-complexity tradeoffs

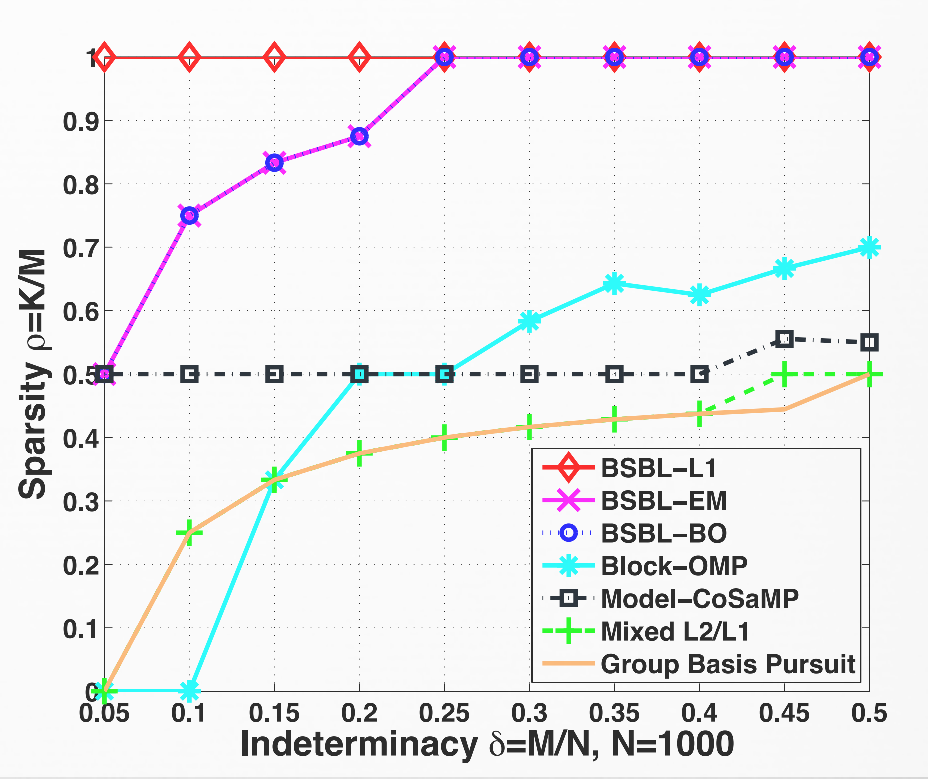
[Zhang et al. 2013]

Phase Transition

Correlation = 0



Correlation = 0.95



$N = 1000$, $M = \delta N$, $g = 40$, block size = 25

Curves indicate > 99% success

[Zhang et al. 2013]

Pattern-Coupled SBL

- Hierarchical model: $p(\mathbf{x}|\alpha) = \prod_{i=1}^{N} \mathcal{N}\left(x_i; 0, (\alpha_i + \beta\alpha_{i+1} + \beta\alpha_{i-1})^{-1}\right)$
- $0 \leq \beta \leq 1$ controls the coupling
- E-step almost the same as before:

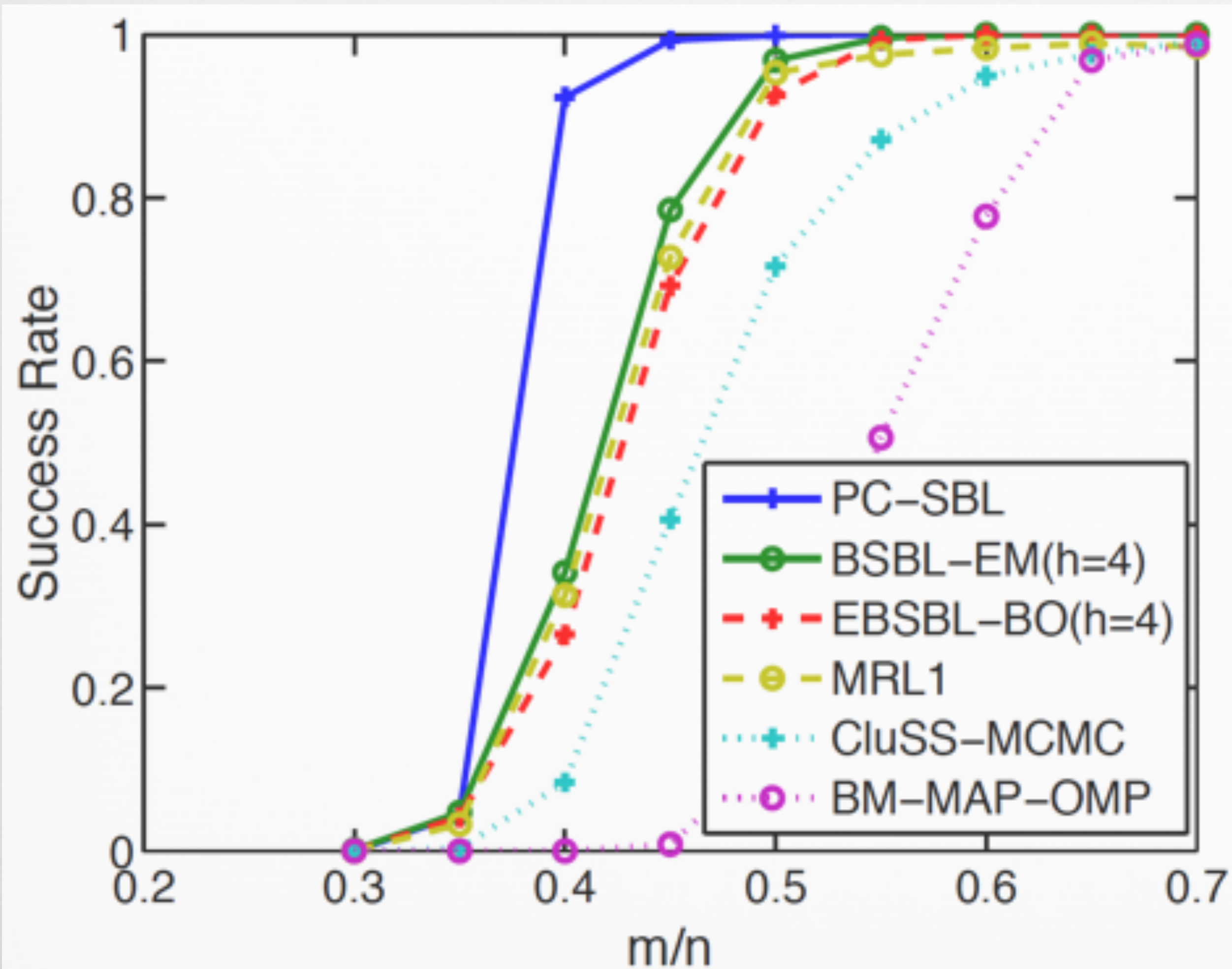
$$\mu = \sigma^{-2}\left(\sigma^{-2}\Phi^T\Phi + \left(\Gamma^{(t)}\right)^{-1}\right)^{-1}\Phi^T\mathbf{y} \quad \Sigma = \left(\sigma^{-2}\Phi^T\Phi + \left(\Gamma^{(t)}\right)^{-1}\right)^{-1}$$

  - $\Gamma^{(t)}$= diagonal$\left(\alpha_i^{(t)} + \beta\alpha_{i+1}^{(t)} + \beta\alpha_{i-1}^{(t)}\right)^{-1}$

- M-step: coupled equations. Approx. soln:

$$\alpha_i^{(t+1)} = \left(\mu_i^2 + \Sigma_{i,i} + \beta(\mu_{i-1}^2 + \Sigma_{i-1,i-1}) + \beta(\mu_{i+1}^2 + \Sigma_{i+1,i+1})\right)^{-1}$$

Empirical Performance

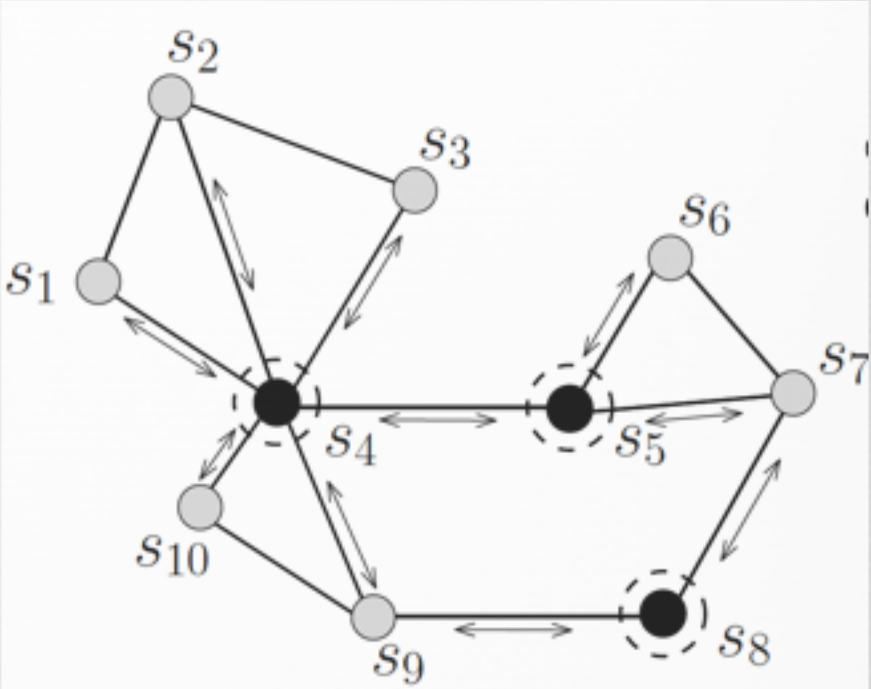


N = 100 entries
K = 25 nonzeros
L = 4 clusters

Source: J. Fang et al., "Pattern-Coupled Sparse Bayesian Learning for Recovery of Block-Sparse Signals", IEEE TSP Jan. 2015

Distributed Recovery: Learning Over a Network

- Network of L data centers
  - Node j has observation y_j
- Want to learn x_j:
  - Statistically related to y_j
- Centralized processing:
  - Optimal, but
  - Computationally demanding
- Distributed (in-network) processing:
  - Secure
  - Robust to node failures



Recap: SBL for Joint Sparse Recovery

- EM Iterations:
  - E-step:

$$\Sigma_j^{k+1} = \Gamma^k - \Gamma^k \Phi_j^T \left(\sigma_j^2 \mathbf{I}_M + \Phi_j \Gamma^k \Phi_j^T\right)^{-1} \Phi_j \Gamma^k$$

$$\mu_j^{k+1} = \sigma_j^{-2} \Sigma_j^{k+1} \Phi_j^T \mathbf{y}_j$$

    - Separable: x_j are independent given Γ
    - Can be computed locally at each node
  - M-step: not separable

$$\Gamma^{k+1} = \frac{1}{L} \sum_{j=1}^{L} \mathbf{a}_j^{(k+1)}$$

A Simple Trick

- Equivalent problems

$$\gamma^* = \frac{1}{L}\sum_{j=1}^{L} a_j \qquad \gamma^* = \arg\min_{\gamma} \sum_{j=1}^{L} |\gamma - a_j|^2$$

Can be computed
locally at each node!
Objective fn. separable

- For distributed implementation

$$\underset{\gamma_j,\, j\in[L]}{\arg\min} \sum_{j=1}^{L} |\gamma_j - a_j|^2$$

Bridge nodes
Linear constraints

$$\text{subject to } \gamma_j = \gamma_b, b \in \mathcal{B}_j, j \in [L]$$

Alternating Directions Method of Multipliers

- General problem: given convex fns. f and g

$$\min_{\{\mathbf{x},\mathbf{y}\}} f(\mathbf{x}) + g(\mathbf{y})$$
$$\text{subject to } \mathbf{Ax} + \mathbf{By} = \mathbf{c}$$

- Augmented Lagrangian

$$\mathcal{L}_\rho(\mathbf{x},\mathbf{y},\lambda) = f(\mathbf{x}) + g(\mathbf{y}) + \lambda^T(\mathbf{Ax} + \mathbf{By} - \mathbf{c}) + \frac{\rho}{2}\|\mathbf{Ax} + \mathbf{By} - \mathbf{c}\|_2^2$$

- ADMM iterations

Convex problems, easy to solve

$$\boxed{\begin{aligned}\mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x}} \mathcal{L}_\rho\left(\mathbf{x}, \mathbf{y}^{(k)}, \lambda^{(k)}\right)\\ \mathbf{y}^{(k+1)} &= \arg\min_{\mathbf{y}} \mathcal{L}_\rho\left(\mathbf{x}^{(k+1)}, \mathbf{y}, \lambda^{(k)}\right)\end{aligned}}$$

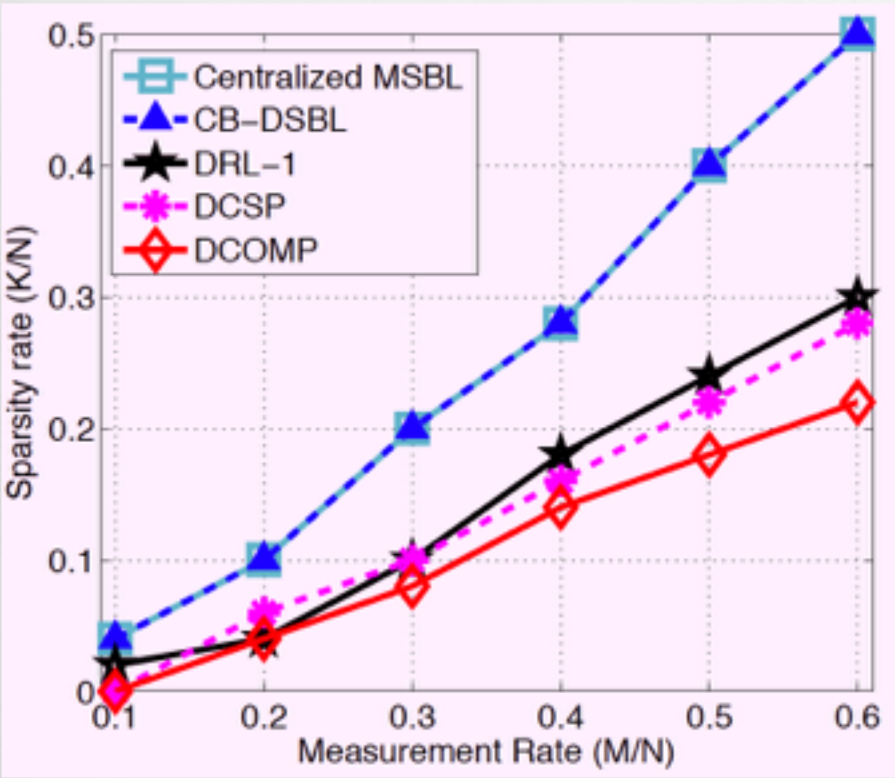
Dual update

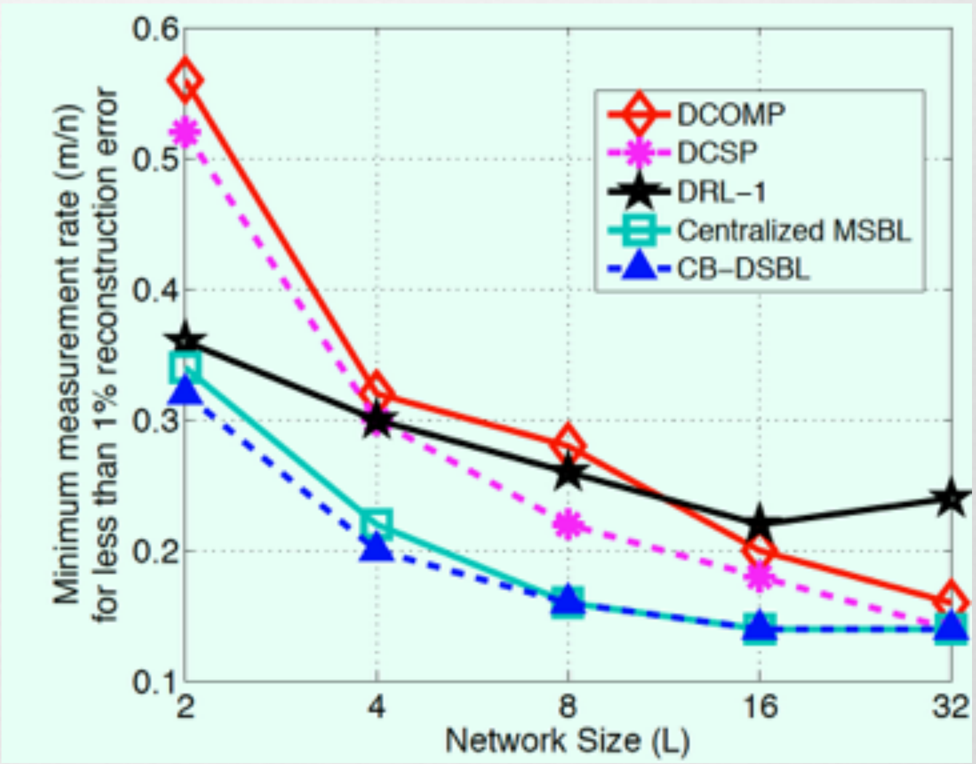
$$\lambda^{(k+1)} = \lambda^{(k)} + \rho(\mathbf{Ax} + \mathbf{By} - \mathbf{c})$$

Benefits of ADMM

- Facilitates distributed algorithms
  - Many rigorous convergence results exist
  - E.g., $\sum_{j=1}^{L} \|\gamma_j^{(r+1)} - \gamma_j^*\|_2 \leq c^{(r)}$ where $c^{(r)} \to 0$ monotonically as $r \to \infty$
- Can extend to many other nonseparable objective fns, e.g., the nuclear norm
- Fastest convergence $\rho_{\text{opt}} = (\text{min. no. of bridge nodes per node})^{-1}$

Simulation Result: NMSE Phase Transition

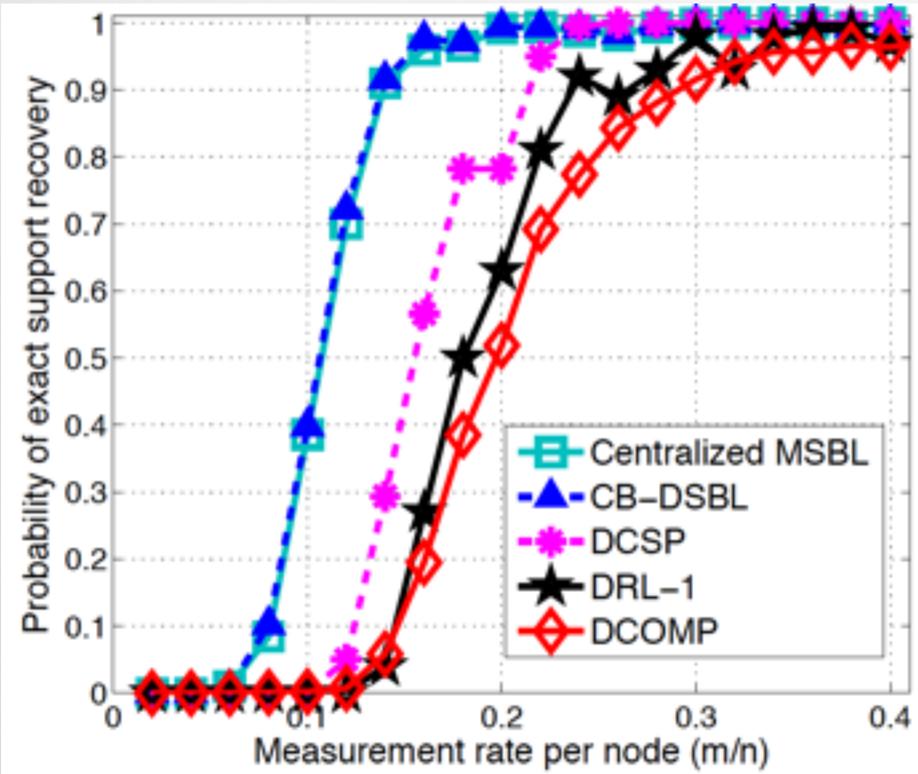


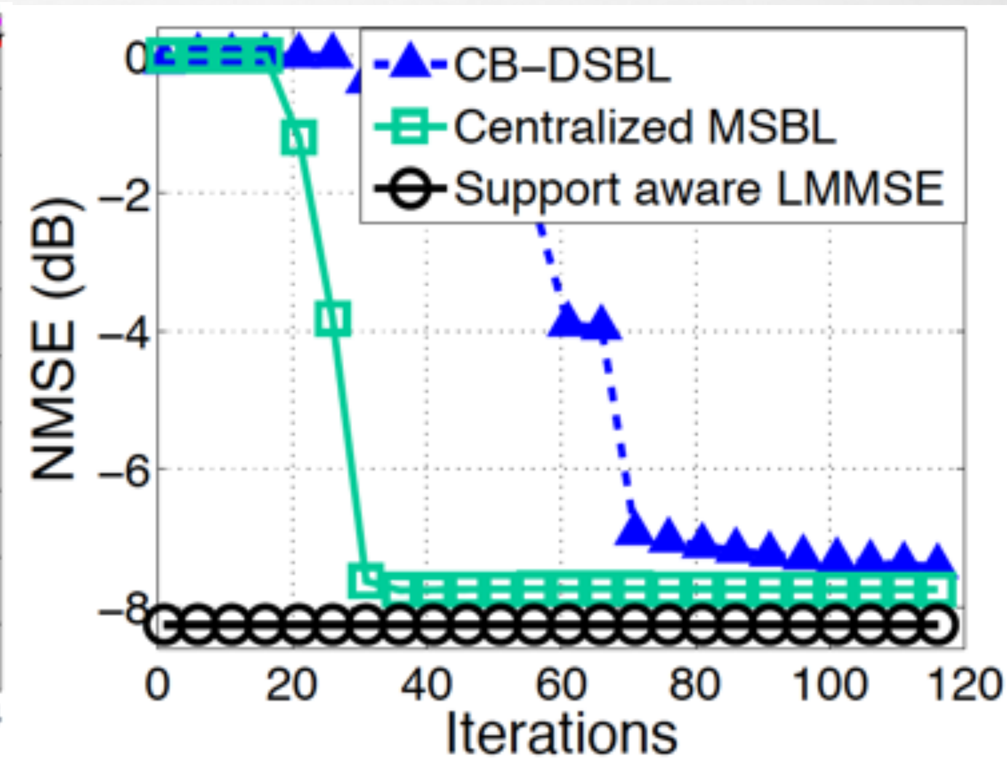


L = 5 nodes, n = 50, m = 10, 10% sparsity, SNR = 30 dB

[S. Khanna, C. R. Murthy, 2015 (under review)]

Support Recovery & Convergence Properties





L = 10 nodes, n = 50, SNR = 10dB, m = 10 (R), 10% sparsity

[S. Khanna, C. R. Murthy, 2015 (under review)]

Parameter Identifiability in SBL

- $y = \Phi x + v \Rightarrow p(y; \Theta)$
- Parameter Θ depends on the model:
  - Type I: x deterministic: $\Theta = \mathbf{x}$, $p^{(\mathrm{I})}(\mathbf{y}) = \mathcal{N}(\Phi\mathbf{x}, \sigma^2\mathbf{I})$
  - Type II: x random: $\mathbf{x} \sim \mathcal{N}(0, \Gamma); \Theta = \Gamma$, $p^{(\mathrm{II})}(\mathbf{y}) = \mathcal{N}(0, \Phi\Gamma\Phi^H + \sigma^2\mathbf{I})$
- Question: when is Θ identifiable?
  - Identifiable: $p(y;\Theta_1) \neq p(y,\Theta_2) \forall \Theta_1 \neq \Theta_2$.

[P. Pal and P. P. Vaidyanathan, ICASSP 14]

Type I Methods

- Lemma: without assuming sparsity, Θ is non-identifiable if N > M!
  - No consistent estimator exists in the underdetermined case
  - Need to constrain the parameter space for Type I estimation to be meaningful
- Under sparsity assumptions, Θ identifiable (depends on spark/Kruskal rank of Φ)

Type II Methods

- Thm. $\Gamma :\rightarrow p^{(\mathrm{II})}(\mathbf{y};\Gamma)$ is identifiable if $N = \mathrm{rank}(\Phi \odot \Phi)$
  - For suitable Φ, $\mathrm{rank}(\Phi \odot \Phi) = O(M^2)$
  - Remains identifiable till $N \approx O(M^2)$, without even assuming sparsity!
- Thm. If $N = \mathrm{rank}(\Phi \odot \Phi)$, the solution to the SBL cost function is consistent & asymptotically efficient
  - True even if Γ has > M nonzero values!

Recovery Guarantees for M-SBL: Noiseless Case

- If the cols of X are orthogonal, and
$$k < \mathrm{spark}(\Phi) - 1$$
there exists a unique stable fixed point $\hat{\gamma}$ of the M-SBL cost function such that
$$\mathrm{supp}(\hat{\gamma}) = \mathrm{supp}(\mathbf{X})$$
[Wipf & Rao, 07]

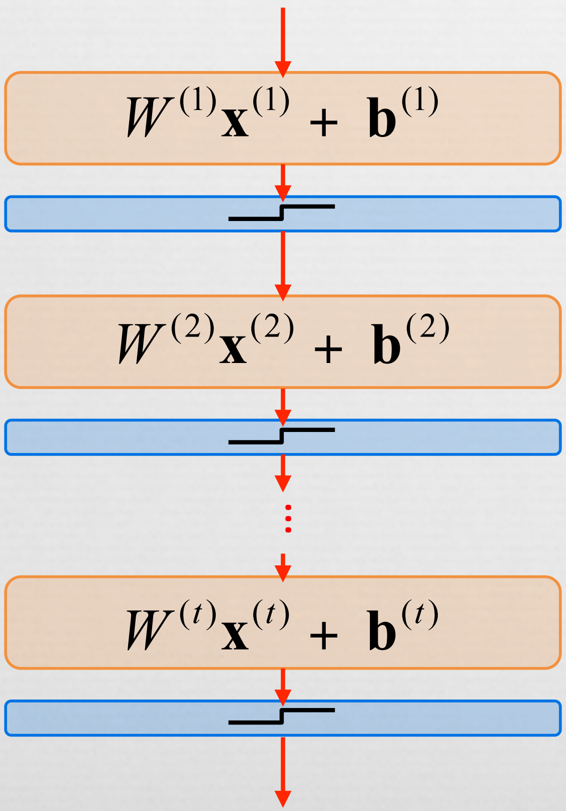
- If the cols of X are orthogonal and
$$\mathrm{rank}(\Phi \odot \Phi) = N$$
Not difficult to satisfy
then M-SBL correctly recovers the support, even if $m < k < N$!
[Balkan et al., 14]

To Recap

- Multiple measurement vectors
  - M-SBL algorithm and its extensions
    - Exploits joint sparsity
  - Intra- and inter-vector correlation
  - Pattern-coupled SBL
- Distributed M-SBL
- M-SBL under colored noise (did not cover)

Maximal Sparsity & Deep Networks?

- Basic DNN template



Linear filter

Nonlinearity/threshold

Observation:

Many common iterative algos follow exactly the same script

$$\mathbf{x}^{(t+1)} = f\left(\mathbf{W}\mathbf{x}^{(t)} + \mathbf{b}\right)$$

Examples: Compressive sensing, robust regression, sparse coding, …

Iterative Hard Thresholding

- Unconstrained gradient step

$$\mathbf{u} = \mathbf{x}^{\text{old}} - \mu \left. \frac{\partial \|\mathbf{y} - \Phi \mathbf{x}\|_2^2}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^{\text{old}}}$$

$$\frac{\partial \|\mathbf{y} - \Phi \mathbf{x}\|_2^2}{\partial \mathbf{x}} \propto \Phi^T \Phi \mathbf{x} - \Phi^T \mathbf{y}$$

- Projection/thresholding step

$$\mathbf{x}^{\text{new}} = \underbrace{H_k(\mathbf{u})}$$

$$u_i = \begin{cases} u_i : & |u_i| \text{ one of the } k \text{ largest elements} \\ 0 : & \text{otherwise} \end{cases}$$

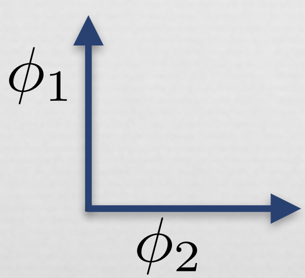
Restricted Isometry Property (RIP)

- A matrix Φ satisfies RIP with constant $\delta_k(\Phi) < 1$ if

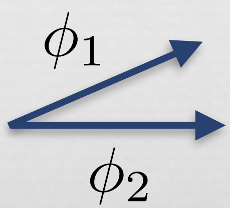
$$(1 - \delta_k[\Phi])\|\mathbf{x}\|_2^2 \leq \|\Phi\mathbf{x}\|_2^2 \leq (1 + \delta_k[\Phi])\|\mathbf{x}\|_2^2$$

holds for all $\{\mathbf{x} : \|\mathbf{x}\|_0 \leq k\}$

Small RIP constant $\delta_2[\Phi]$



Large RIP constant $\delta_2[\Phi]$



Recovery Guarantee with IHT

- Suppose there exists some x* such that

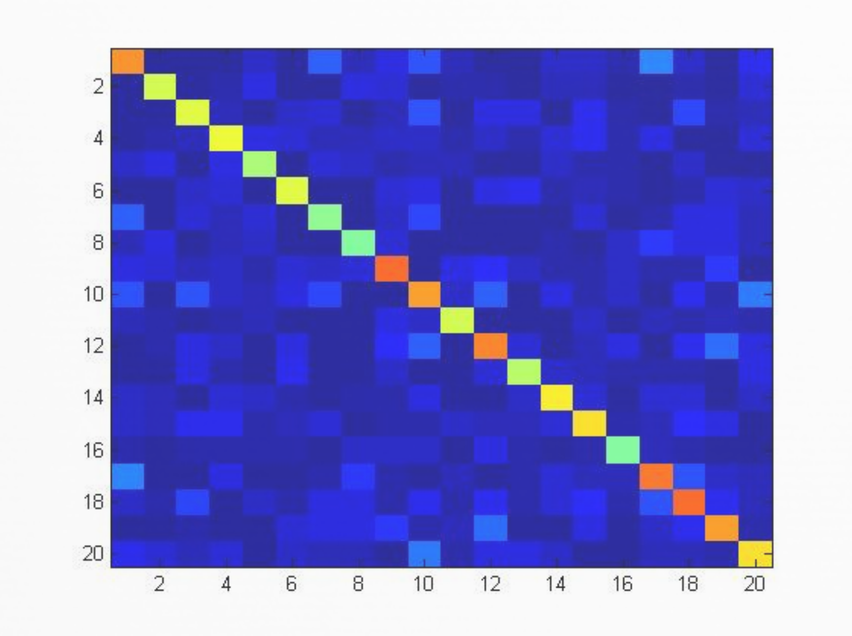
$$\mathbf{y} = \Phi \mathbf{x}^*$$
$$\|\mathbf{x}^*\|_0 \leq k$$
$$\delta_{3k}[\Phi] < \frac{1}{\sqrt{32}}$$

then the IHT iterations are guaranteed to converge to x*

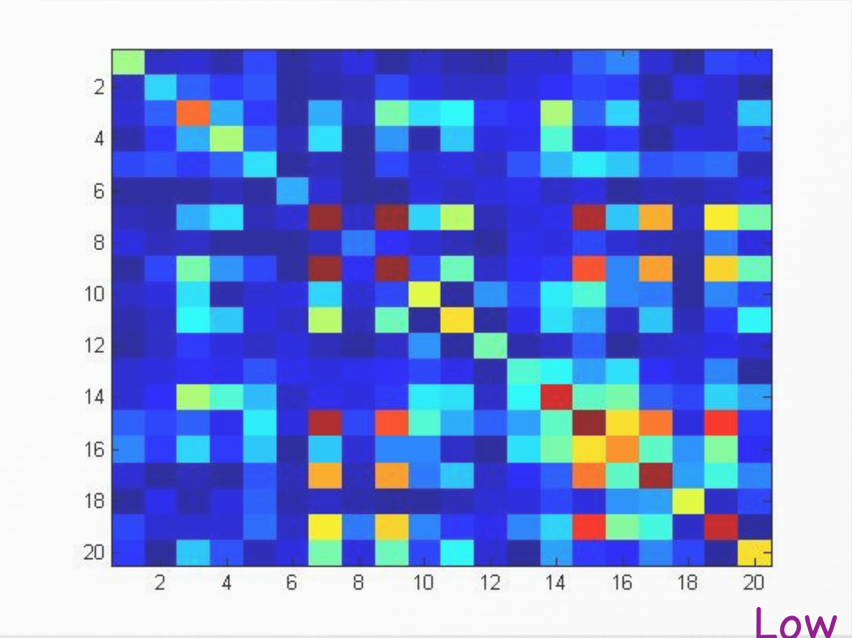
[Blumensath and Davies, 09]

Effects of Correlation Structure

Low correlation: easy



High correlation: hard



Example

$$\Phi_{(\text{uncor})} \rightarrow \text{ iid } \mathcal{N}(0, v) \text{ entries}$$

$$\delta_{3k}[\Phi] < \frac{1}{\sqrt{32}}$$ Small RIP constant

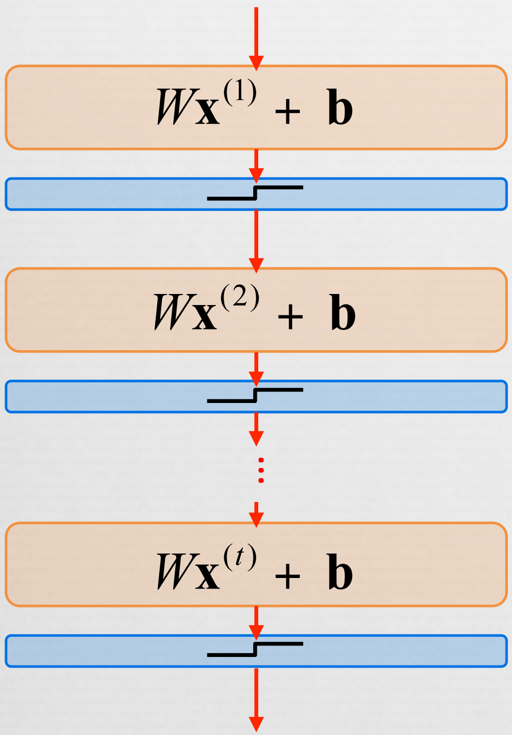
Example

Low rank

$$\Phi_{(\text{cor})} = \Phi_{(\text{uncor})} + \Delta$$

$$\delta_{3k}[\Phi] \gg \frac{1}{\sqrt{32}}$$ Large RIP constant

Unfolded IHT Iterations



Linear filter

Nonlinearity/threshold

$$\mathbf{W} = \mathbf{I} - \mu\Phi^T\Phi$$

$$\mathbf{b} = \mu\Phi^T\mathbf{y}$$

- Clear resemblance to the structure of a deep neural network
- So is there an advantage to learning the weights?

Performance Bound with Learned Layer Weights

- **Theorem**

  There will always exist layer weights W and bias b such that the effective RIP constant is reduced via

$$\delta^*_{3k}[\Phi] \triangleq \inf_{\mathbf{W},\mathbf{D}} \delta_{3k}[\mathbf{W}\Phi\mathbf{D}] < \delta_{3k}[\Phi]$$

Effective RIP constant — Original RIP constant

where W is arbitrary and D is diagonal

It is therefore possible to reduce high RIP constants!

[Xin and Wipf, 16]

Practical Consequences

- **Theorem**

  Suppose we have correlated dictionary formed via

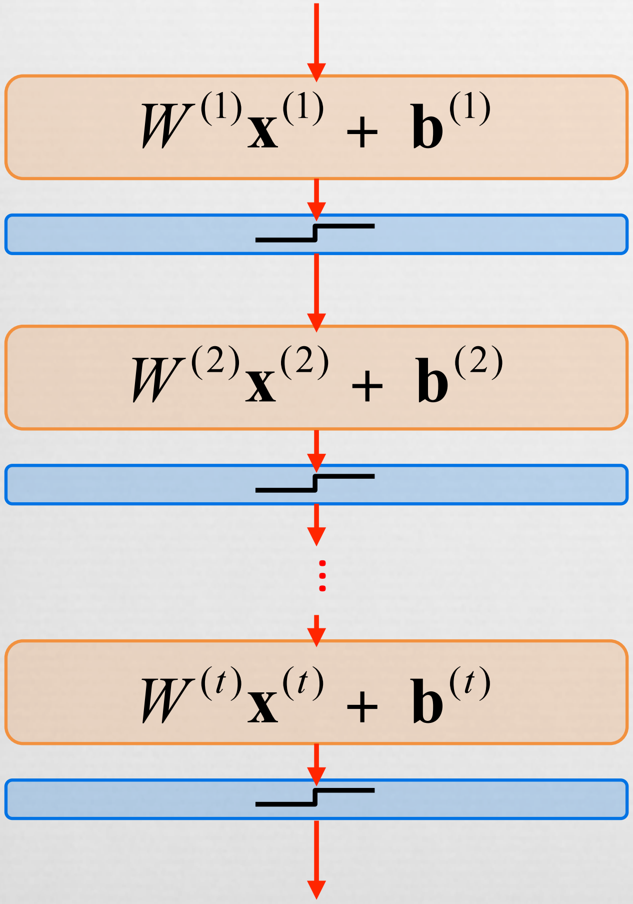
$$\Phi_{(\mathrm{cor})} = \Phi_{(\mathrm{uncor})} + \Delta$$

  with $\Phi_{(\mathrm{uncor})} \rightarrow$ iid $\mathcal{N}(0, v)$ entries and Δ low rank. Then $\mathbb{E}\left(\delta^*_{3k}[\Phi_{(\mathrm{cor})}]\right) \approx \mathbb{E}\left(\delta_{3k}[\Phi_{(\mathrm{uncor})}]\right)$

Can "undo" low rank correlations that would otherwise produce a high RIP constant …

[Xin and Wipf, 16]

Advantages of Independent Layer Weights & Activations



- **Theorem**
  Independent weights on each layer
  ⇓
  Often possible to obtain nearly ideal RIP even when **full rank** Δ is present

[Xin and Wipf, 16]

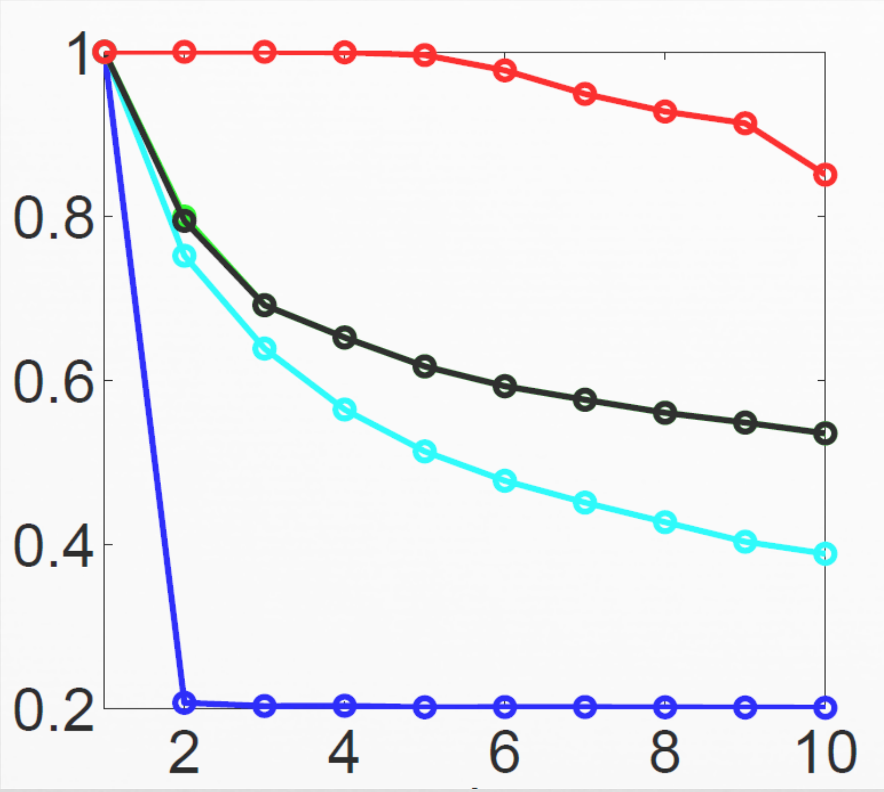
Alternative Learning-Based Strategy

- Thus far: idealized deep network weights exist that improve RIP constants
- Given access to feasible pairs

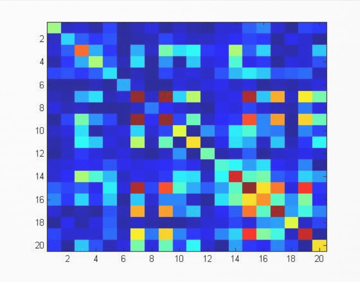
$$\{\mathbf{y}, \mathbf{x}^* : \mathbf{y} = \Phi \mathbf{x}^*, \|\mathbf{x}^*\|_0 \leq k\}$$

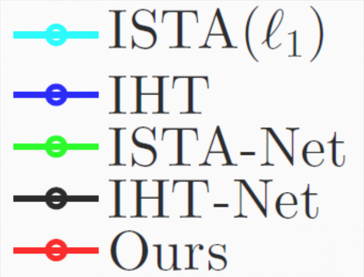
  can learn an approximation to weights
- Can treat as a multi-label DNN classification problem to estimate support of x^*
- Many other important training modifications are motivated by this analysis

Simulation Example



$$\Phi^T \Phi \neq I$$





[Gregor and LeCun, 10; Wang et al., 16]

Robust Surface Normal Estimation

- Input:

- Per-pixel model:

Specular reflections, shadows, etc. (outliers)

$$\mathbf{y} = \mathbf{L}\mathbf{n} + \mathbf{x}$$

Observations under different lightings

Lighting matrix

Raw unknown surface normal

- Can apply any sparse learning method to obtain outliers

Convert to Sparse Estimation Problem

$$\underbrace{\mathrm{Proj}_{\mathrm{Null}[\mathbf{L}^T]}(\mathbf{y})}_{\tilde{\mathbf{y}}} = \mathrm{Proj}_{\mathrm{Null}[\mathbf{L}^T]}(\mathbf{L}\mathbf{n} + \mathbf{x}) = \underbrace{\mathrm{Proj}_{\mathrm{Null}[\mathbf{L}^T]}}_{\Phi}(\mathbf{x})$$

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ s.t. } \tilde{\mathbf{y}} = \Phi \mathbf{x}$$

Once outliers are known, can estimate **n** via

$$\hat{\mathbf{n}} = \left(\Phi^T \Phi\right)^{-1} \Phi^T (\mathbf{y} - \mathbf{x})$$

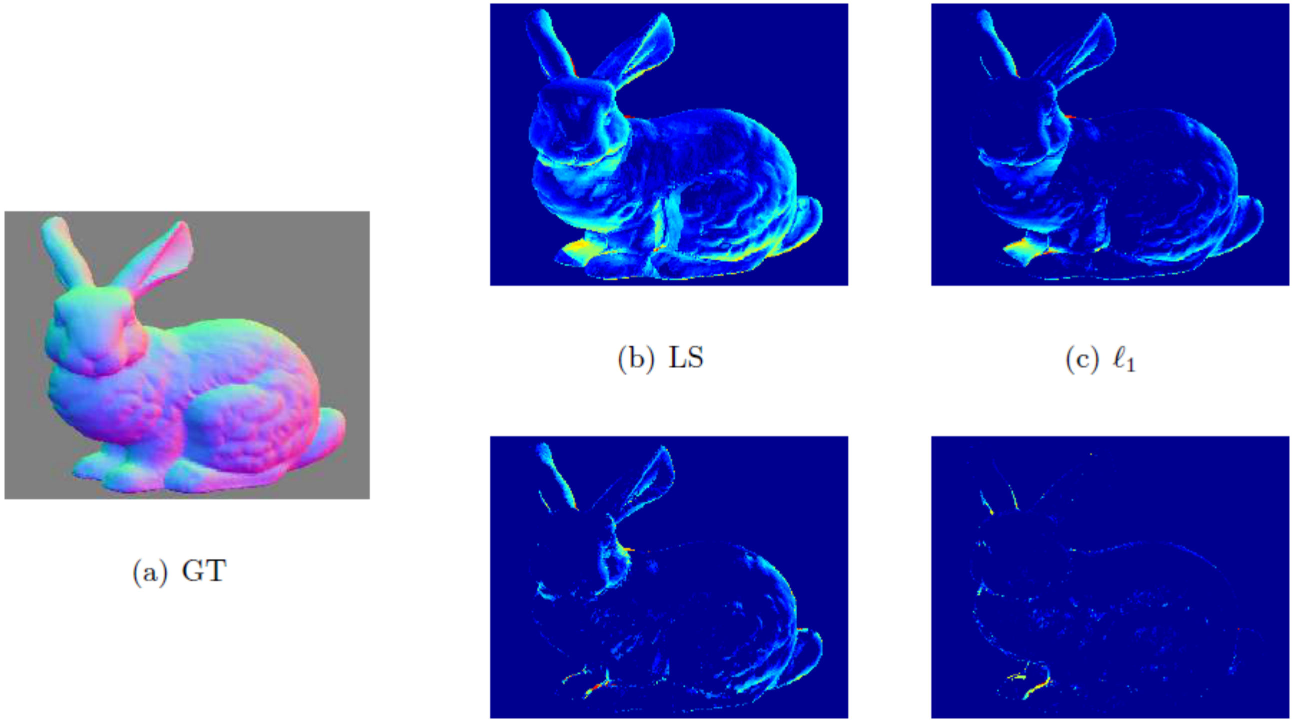
[Candes and Tao 04]

DNN Weakly-Supervised Training Setup

- Generated 600,000 synthetic training points:
  - Support patterns of x^* randomly generated
  - Nonzero values were generated iid from $N(\mu,\sigma^2)$ with (μ, σ^2) loosely fit to real-world imaging data
- Trained a 20 layer network using SGD and a softmax output layer
- Testing performed using imaging data with known ground truth

Results

Bunny Object, INRIA 3D Database



(a) GT

(b) LS

(c) ℓ_1

(d) SBL

(e) Ours

	LS	l_1	SBL	Ours
Angular	12.13	7.10	4.02	1.48
Time	4.10	33.7	59.1	1.17

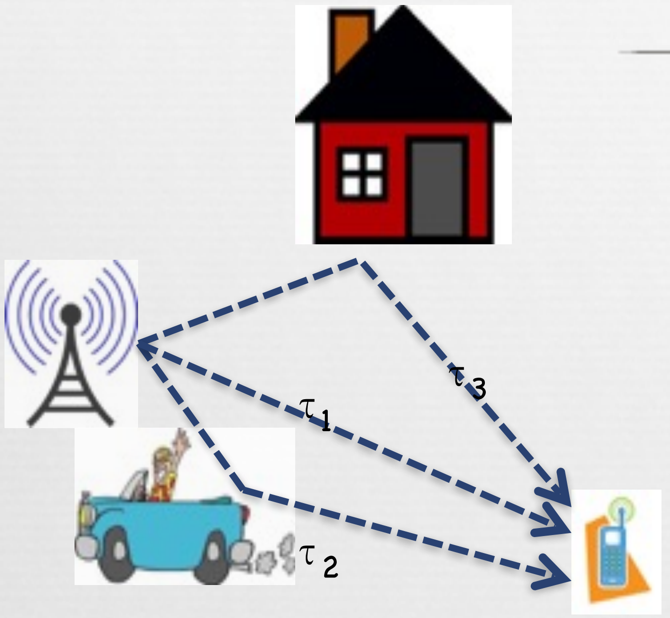
Summary

- First rigorous analysis of how unfolded iterative algorithms can be provably enhanced by learning
- Detailed characterization of how different architecture choices affect performance
- Narrow benefit: First ultra-fast method for obtaining optimal sparse representations with correlated designs (i.e., high RIP constants)
- Broad benefit: General insights into why DNNs can outperform hand-crafted algorithms

Part 5: Applications

Wireless channel estimation & data detection

Wireless Channels



- Wireless channels exhibit multipath
  - Naturally sparse in the lag-domain
- Channel equalization & data detection
  - Need to estimate both support & channel

Channel Models

- Block fading channel:

  Channel constant for the duration of a block (say, K symbols), changes i.i.d. from block-to-block (classic SMV-SBL)

- Time-varying channel:

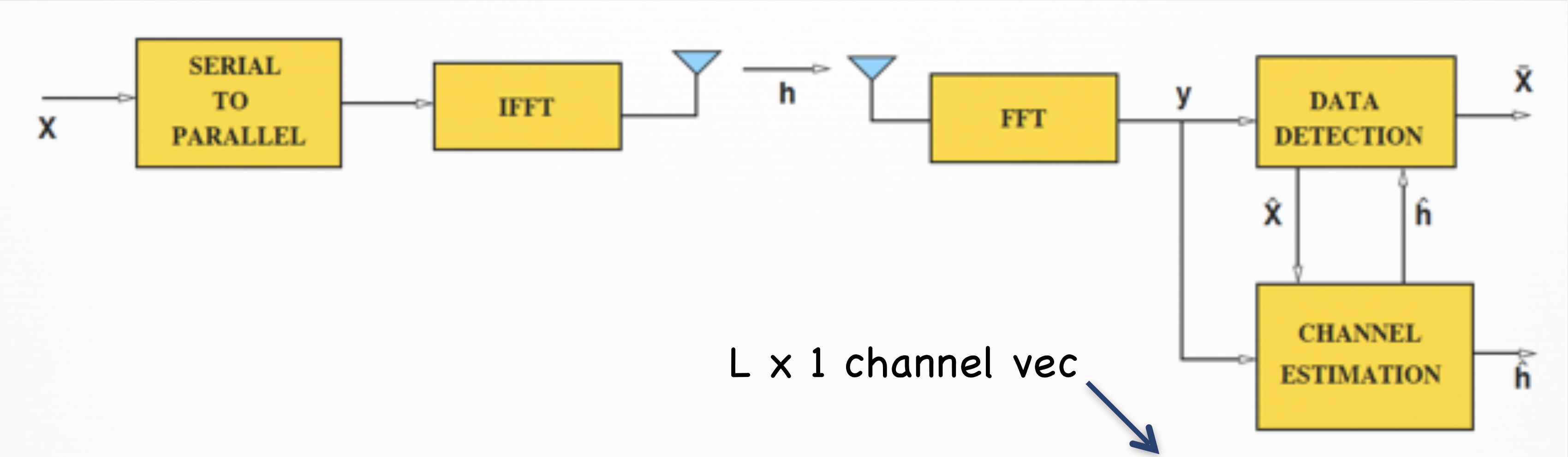
  Channel varies from symbol-to-symbol

  - Want to exploit temporal correlation and group-sparsity (MMV-SBL)

Outline

1. Block fading case:
   1. Known channel support: Joint channel estimation & data detection
   2. Unknown channel support: Channel and support estimation using pilot symbols
   3. Unknown data & support: Joint support, channel estimation & data detection
2. Time-varying case:
   1. AR model: Kalman-EM algo for joint support, channel estimation & data detn

OFDM with Block Fading Channel



L x 1 channel vec

- Received signal model $y = X F h + v$

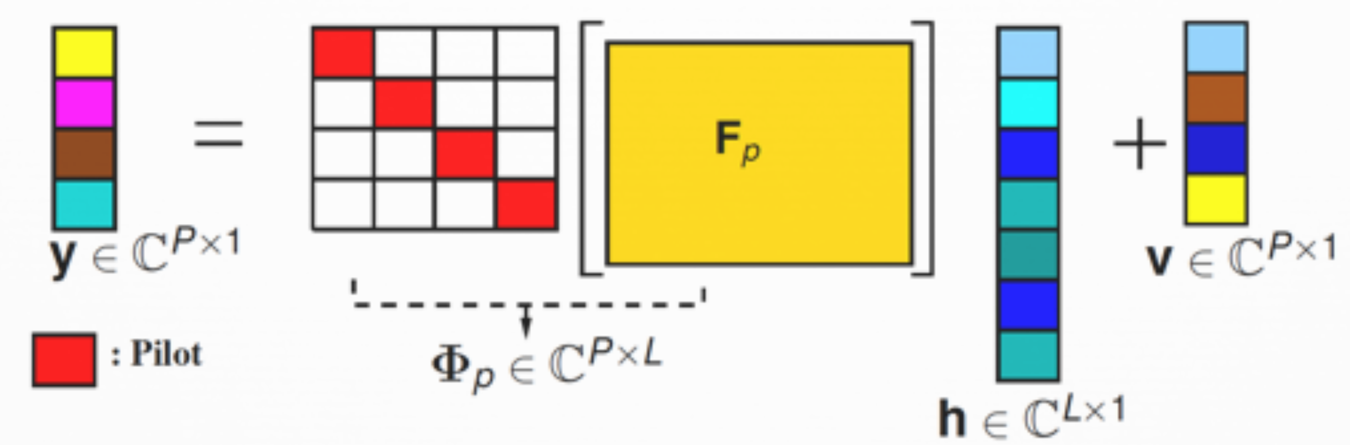
Diagonal data matrix; N x N
N: number of subcarriers

N x L DFT matrix, containing
first L cols of N x N DFT matrix
L: max channel delay spread

Noise

- Goal: Given y, jointly estimate X & h

Sparse Channel Estimation from Pilot Symbols



- h sparse in time (lag) domain
- Hierarchical prior: $\mathbf{h}(i) = \mathcal{CN}(0, \gamma_i)$
  γ_i deterministic, unknown hyperparams
- Goal:
  Given y, X, estimate h (& sparsity profile)

SBL for Basis Selection

- E-Step: $$Q\left(\Gamma|\Gamma^{(t)}\right) = \mathbb{E}_{\mathbf{h}|\mathbf{y};\Gamma^{(t)}} \log p(\mathbf{y}, \mathbf{h}; \Gamma)$$

$$p\left(\mathbf{h}|\mathbf{y};\Gamma^{(t)}\right) = \mathcal{N}(\mu, \Sigma_h), \;\; \mu \triangleq \sigma^{-2}\Sigma_h \mathbf{A}^H \mathbf{y}$$

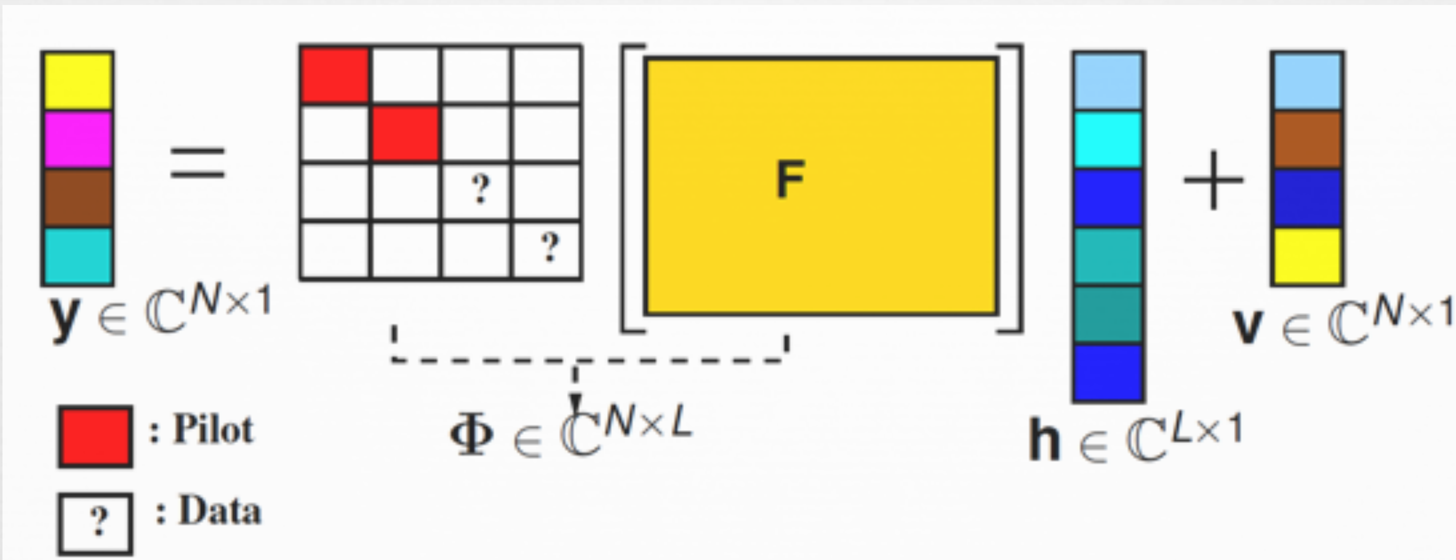
$$\Sigma_h \triangleq \left(\sigma^{-2}\mathbf{A}^H\mathbf{A} + \left(\Gamma^{(t)}\right)^{-1}\right)^{-1}, \;\; \mathbf{A} \triangleq \mathbf{XF}$$

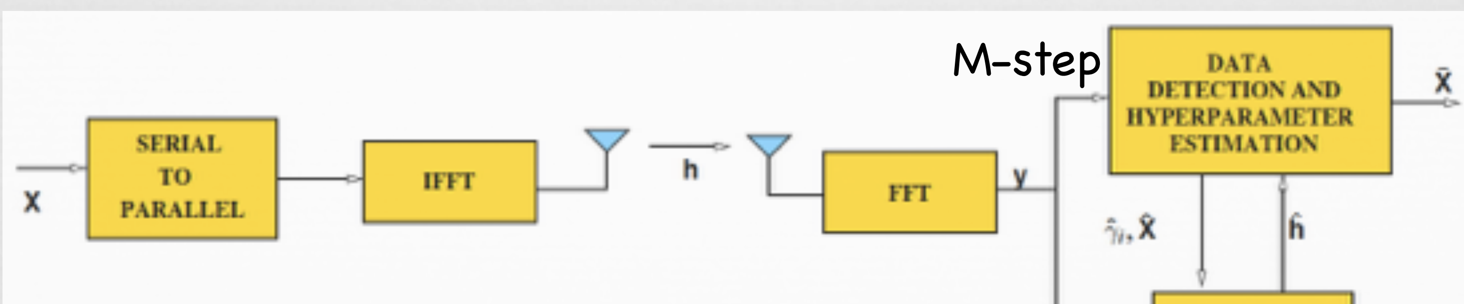
- M-Step: $$\Gamma^{(t+1)} = \arg\max_{\gamma_i \geq 0} Q\left(\Gamma|\Gamma^{(t)}\right)$$

$$\log p(\mathbf{y}, \mathbf{h}; \Gamma) = \log p(\mathbf{y}|\mathbf{h}) + \log p(\mathbf{h}; \Gamma)$$

not a function of γ_i function of γ_i

Joint Channel, Support Estmn. & Data Detn.

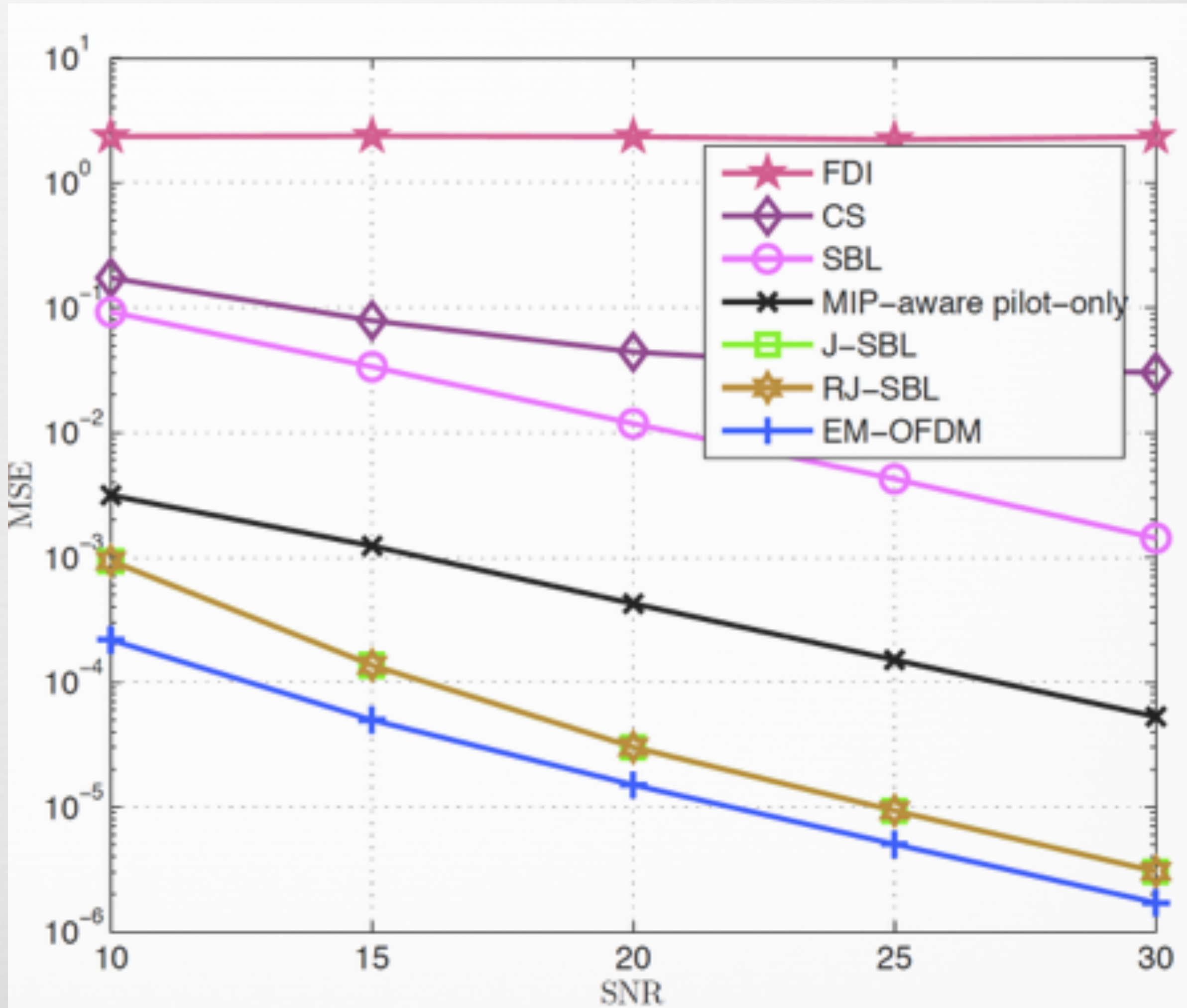




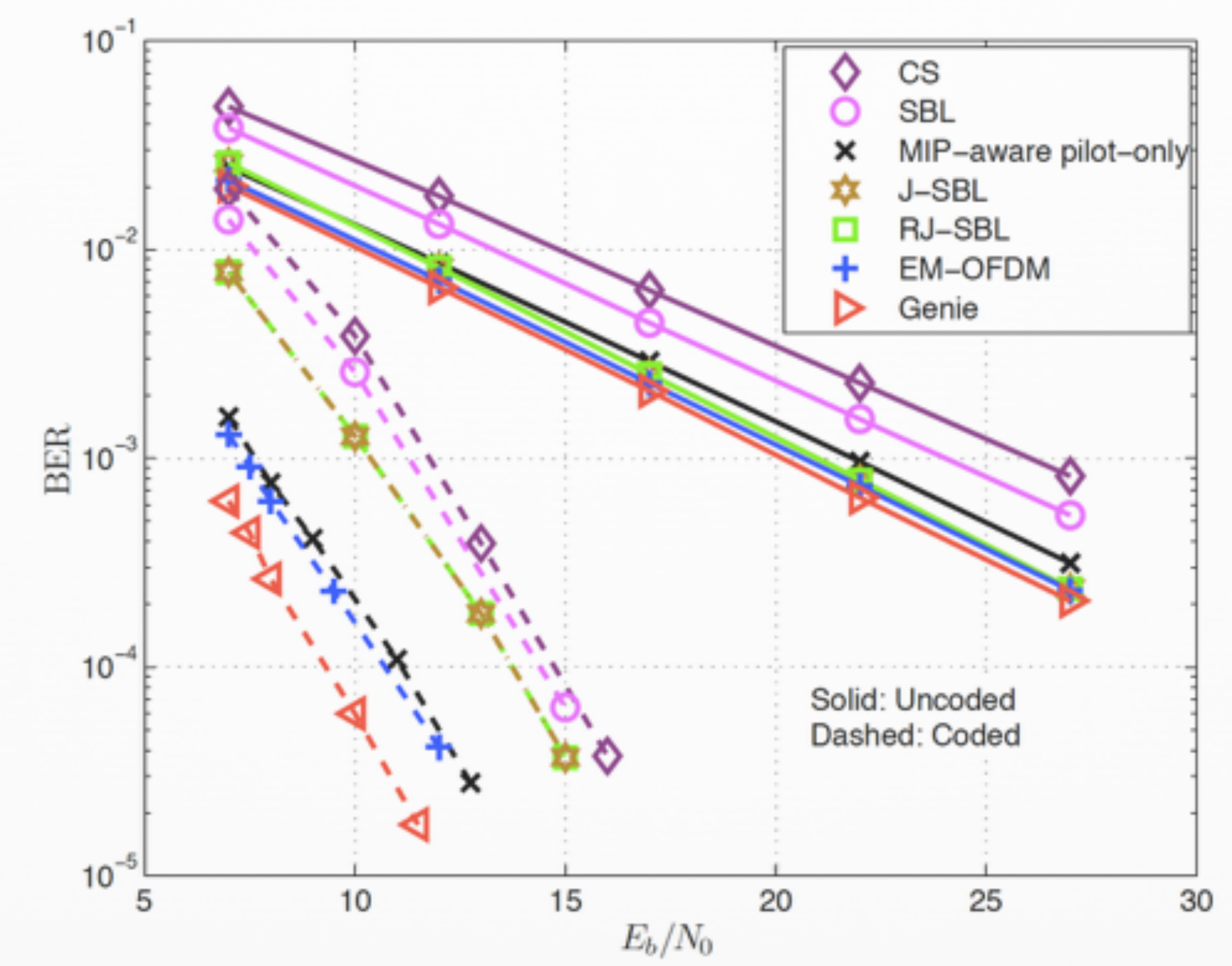
- $y = XFh + v$

Simulation Result

- OFDM system
- N=256 subcarriers,
- max delay spread L=64
- K=7 symbols/slot
- PedB PDP: 6 nonzero taps
- 44 pilot subcarriers
- Data: rate ½ turbo code, QPSK



BER Performance

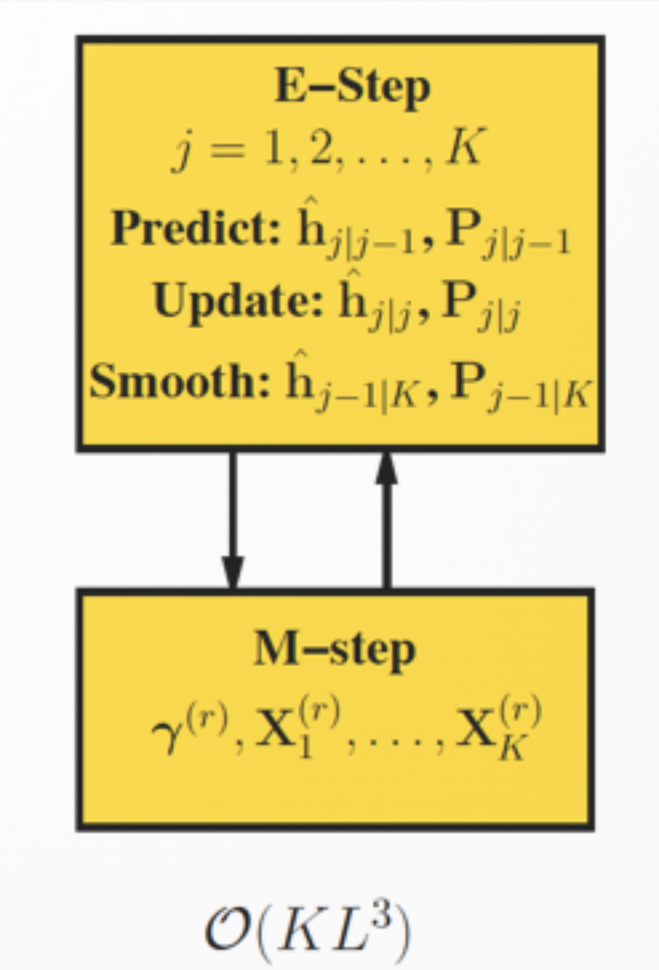


Time-Varying Channels

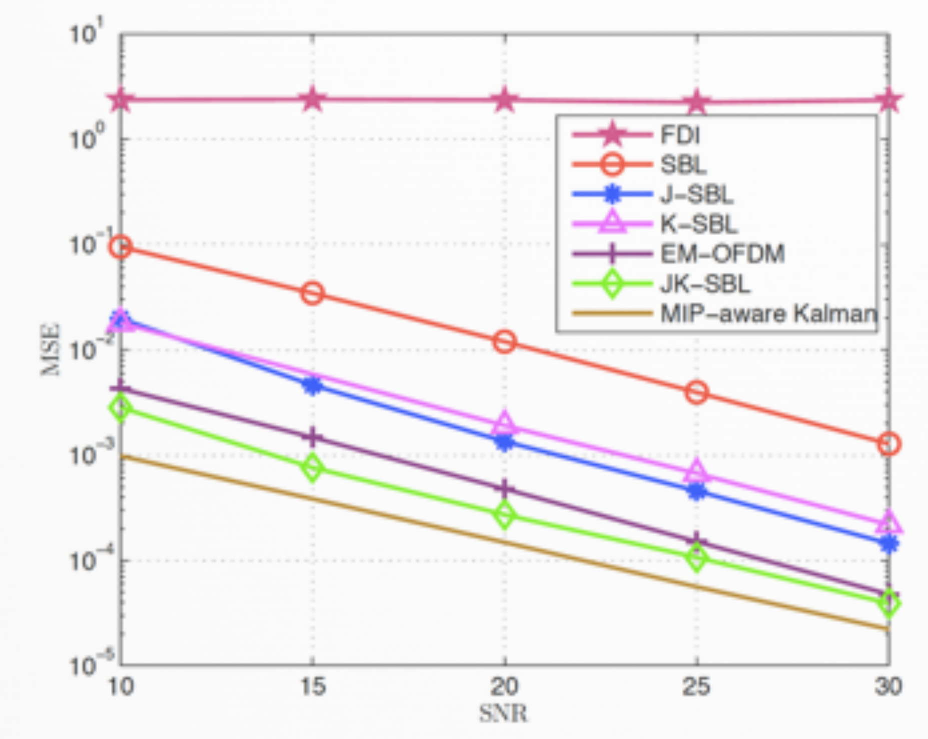
- Channel correlated from symbol-to-symbol
- AR model: $\mathbf{h}_k = \rho \mathbf{h}_{k-1} + \mathbf{u}_k$
- The factor ρ depends on the normalized doppler freq, which in turn depends on the speed of the mobile
- SBL framework can be extended to incorporate the temporal correlation

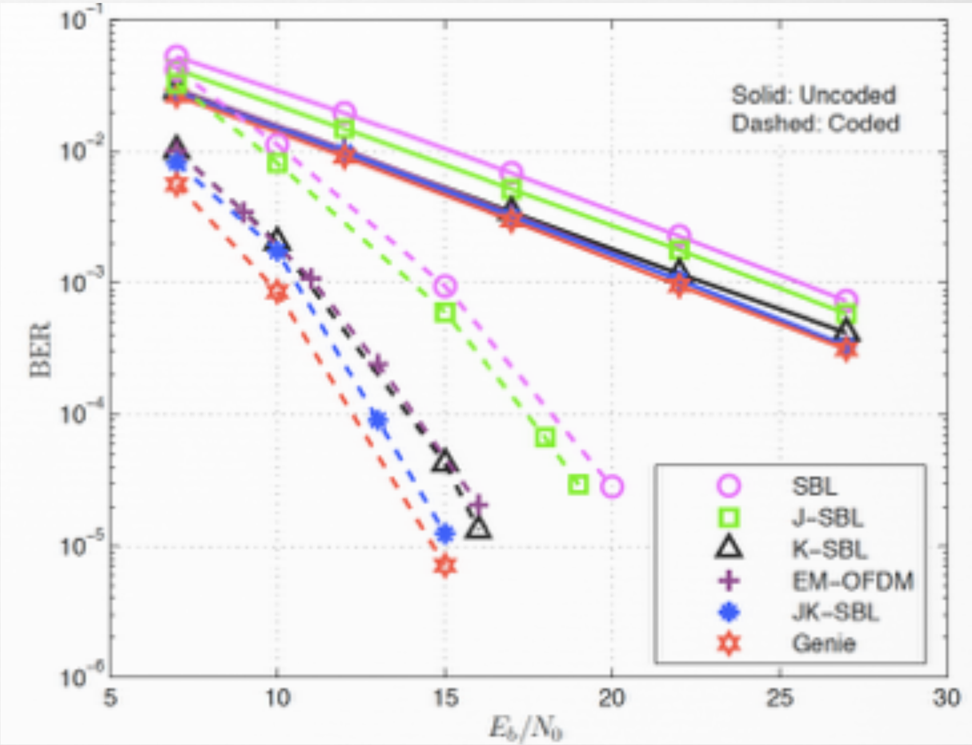
Joint Kalman SBL (JK-SBL)

- Complexity $O(KL^3)$: smaller than block-based methods $O(K^3L^3)$ [Zhang et al. 10]
  - (K = num. OFDM symbols used in joint estimation)
- In the block-fading case: get recursive, more computationally efficient versions of our algos



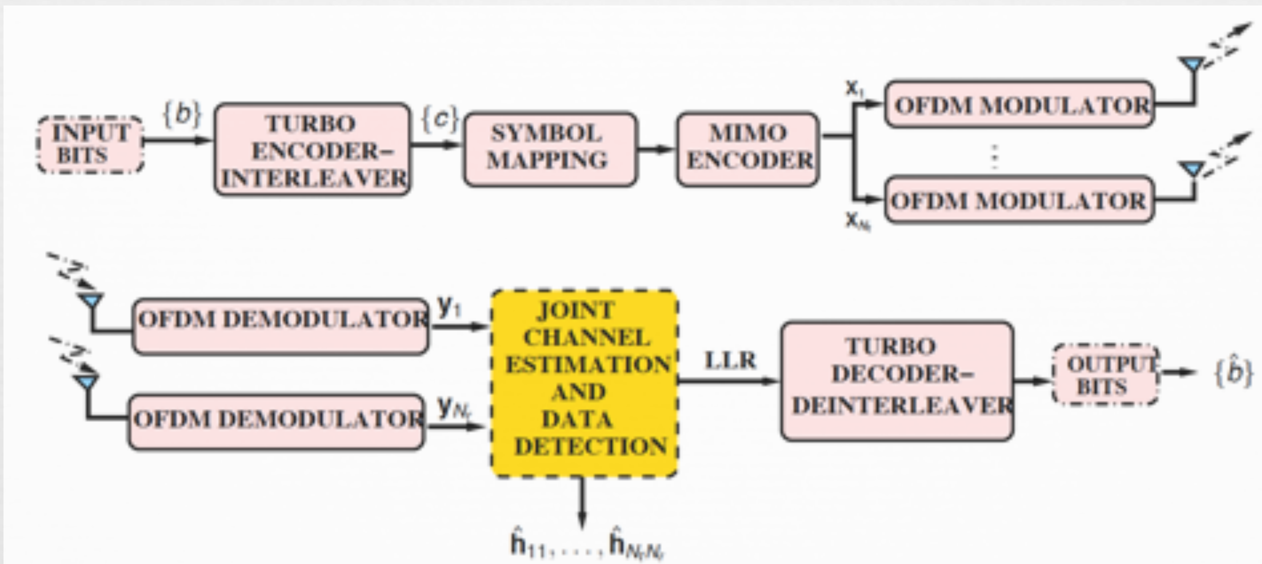
Simulation Result





- $f_d T_s = 0.001$ (slowly time-varying)

MIMO-OFDM



$$\mathbf{y}_{n_r} = \sum_{n_t=1}^{N_t} \mathbf{X}_{n_t}\mathbf{F}\mathbf{h}_{n_t n_r} + \mathbf{v}_{n_r},\ n_r = 1, \ldots, N_r$$

- Goal: Recover $h_1, \ldots, h_{Nr}$ from $y_1 \ldots y_{Nr}$

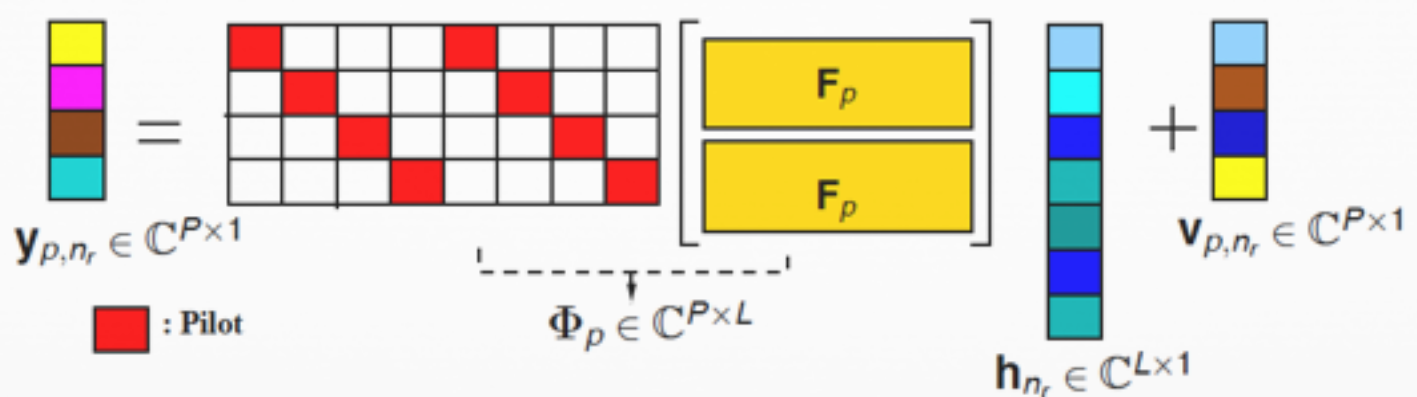
- [Prasad, M. & R., TSP 2015]

MMV Framework

- Measurement model

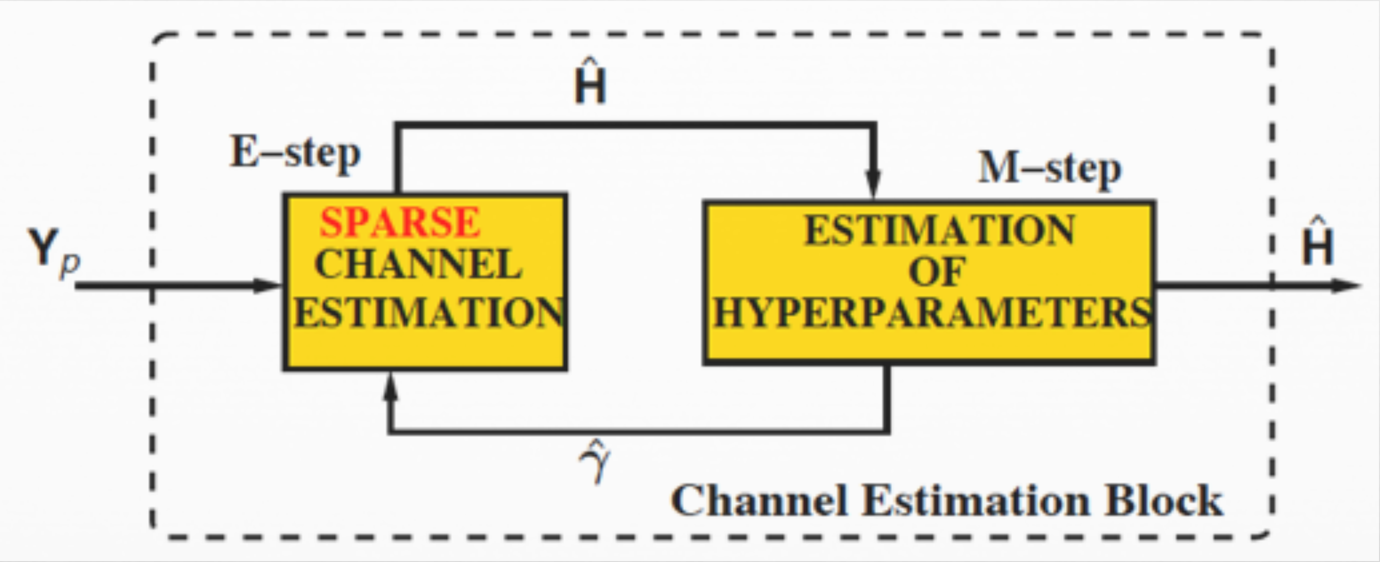
$$\underbrace{[\mathbf{y}_1,\ldots,\mathbf{y}_{N_r}]}_{\mathbf{Y}\in\mathbb{C}^{N\times N_r}} = \underbrace{\mathbf{X}(\mathbf{I}_{N_t}\otimes\mathbf{F})}_{\Phi\in\mathbb{C}^{N\times LN_t}} \underbrace{\begin{pmatrix} \mathbf{h}_{11} & \ldots & \mathbf{h}_{1N_r} \\ \vdots & \vdots & \\ \mathbf{h}_{N_t1} & \ldots & \mathbf{h}_{N_tN_r} \end{pmatrix}}_{\mathbf{H}\in\mathbb{C}^{LN_t\times N_r}} + \underbrace{[\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_{N_r}]}_{\mathbf{V}\in\mathbb{C}^{N\times N_r}}$$

- Pilot subcarriers



The M-SBL Algorithm

- E Step $$Q\left(\gamma|\gamma^{(r)}\right) = \mathbb{E}_{\mathbf{H}|\mathbf{Y}_p;\gamma^{(r)}} \log p(\mathbf{Y}_p, \mathbf{H}; \gamma)$$
- M Step $$\gamma^{(r+1)} = \arg\max_{\gamma \in \mathbb{R}_+^L} Q\left(\gamma|\gamma^{(r)}\right)$$



The E and M Steps

- E-Step: Posterior distribution $\mathcal{CN}(\mu_{n_r}, \mathbf{\Sigma})$

$$\mu_{n_r} = \sigma^{-2}\mathbf{\Sigma}\mathbf{\Phi}_p^H \mathbf{y}_{p,n_r} \qquad \mathbf{\Sigma} = \left(\frac{\Phi_p^H \Phi_p}{\sigma^2} + \left(\Gamma_b^{(r)}\right)^{-1} \right)^{-1}$$

- M-Step:

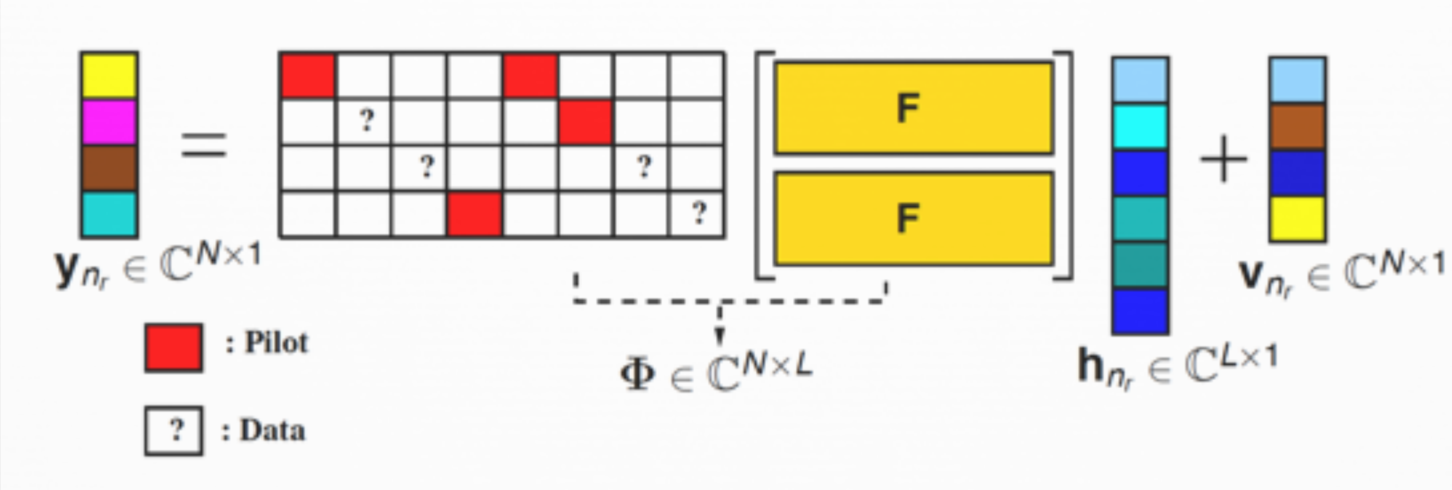
$$Q\left(\gamma|\gamma^{(r)}\right) = c' - \mathbb{E}_{\mathbf{H}|\mathbf{Y}_p}\left[\sum_{n_r=1}^{N_r}\sum_{n_t=1}^{N_t} \mathbf{h}_{n_t n_r}^H \Gamma^{-1} \mathbf{h}_{n_t n_r}\right]$$

Common γ

$$\gamma^{(r+1)}(i) = \frac{1}{N_t N_r}\sum_{n_r=1}^{N_r}\sum_{n_t=0}^{N_t-1} \|\mathbf{M}(i+n_t L, n_r)\|_2^2 + \mathbf{\Sigma}(i+n_t L, i+n_t L)$$

Averaging γ across antennas

Joint Channel Estmn. & Data Detection

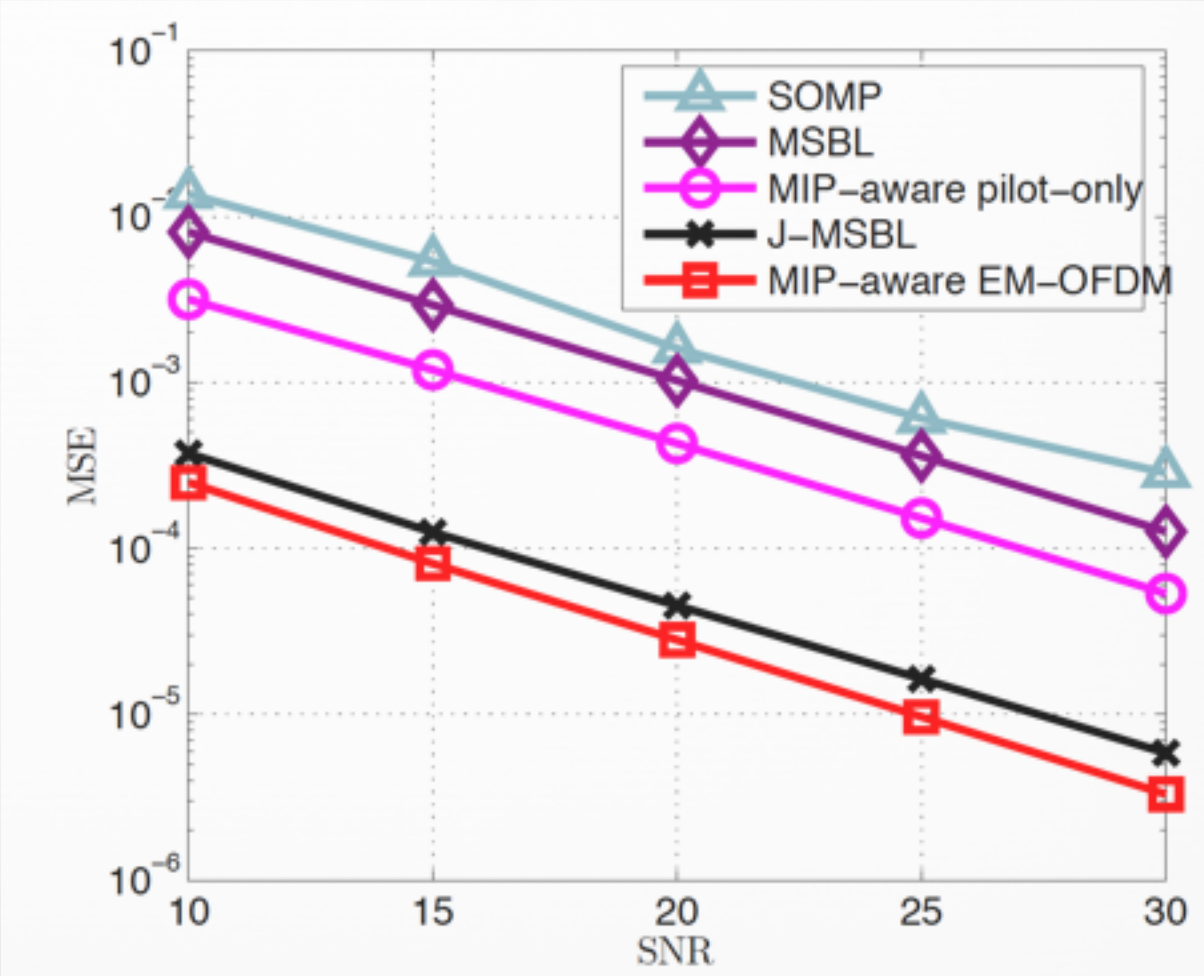


- E Step remains unchanged
- M Step: $\left(\gamma^{(r+1)}, \mathbf{X}^{(r+1)}\right) = \arg \max_{\gamma \in \mathbb{R}_+^L, \mathbf{X} \in \mathcal{S}} Q\left(\gamma, \mathbf{X} | \gamma^{(r)}, \mathbf{X}^{(r)}\right)$

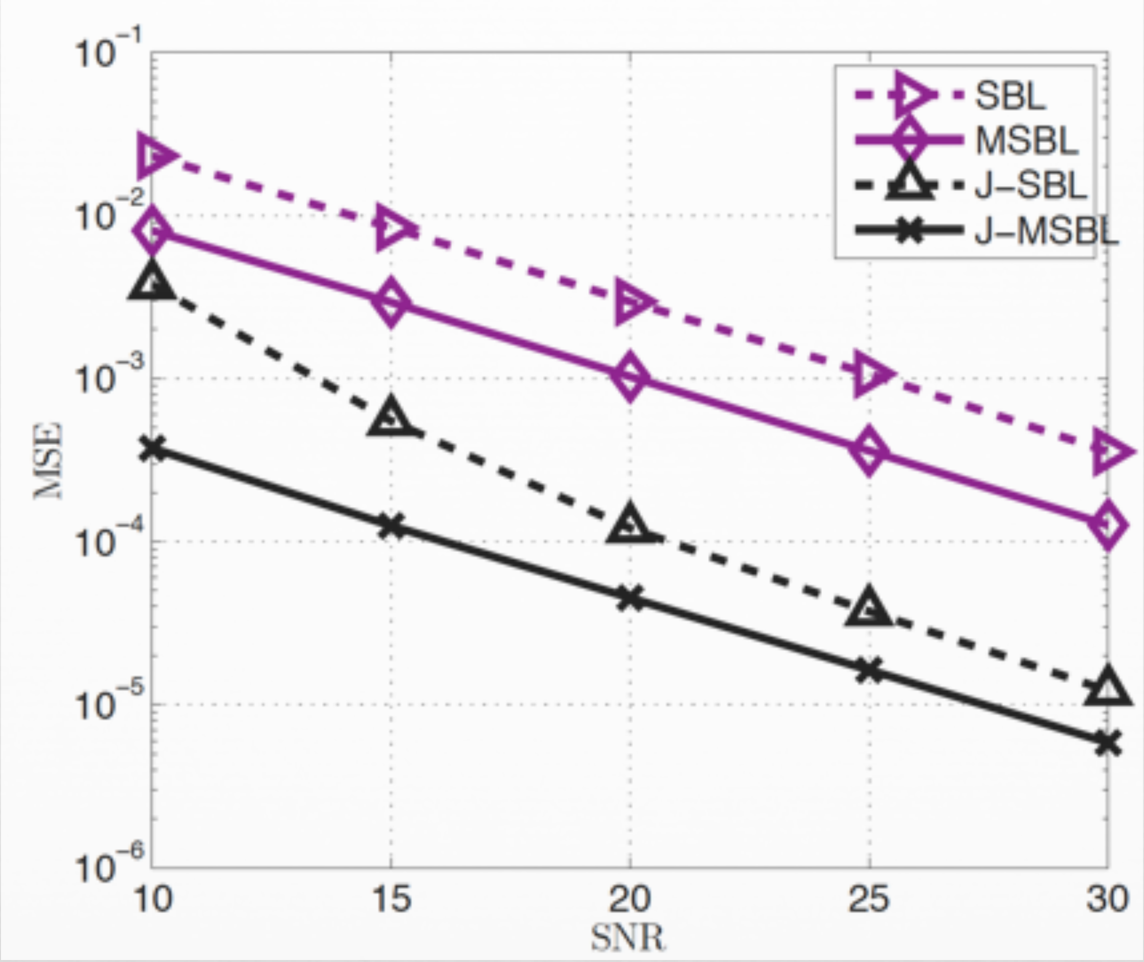
Splits as two separate sub-problems

MSE Performance

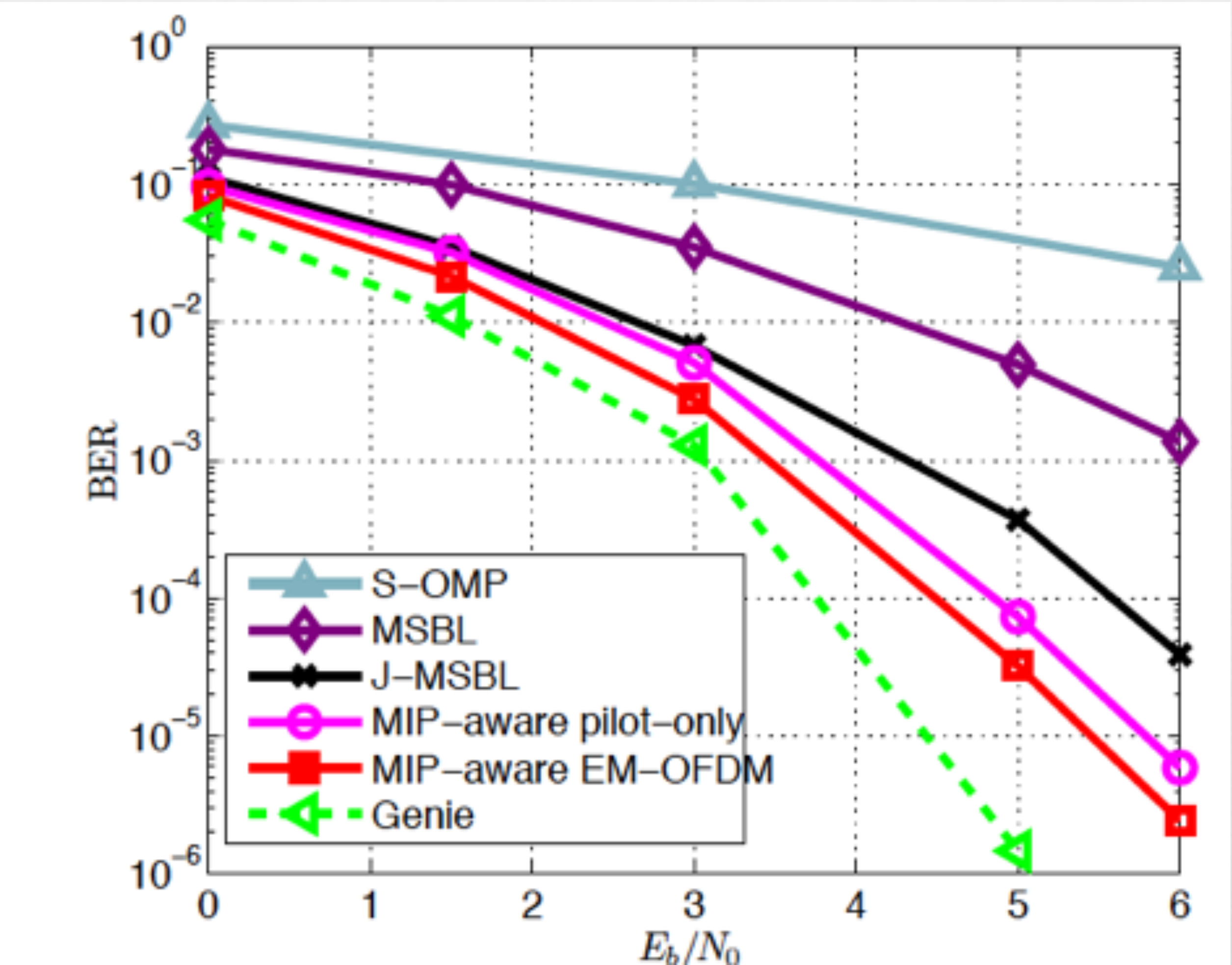
- 2 x 2 MIMO-OFDM System
- 256 subcarriers
- CP length 64
- 44 pilot subcarriers
- PedB PDP
- QPSK constellation



Exploiting Structure Helps!



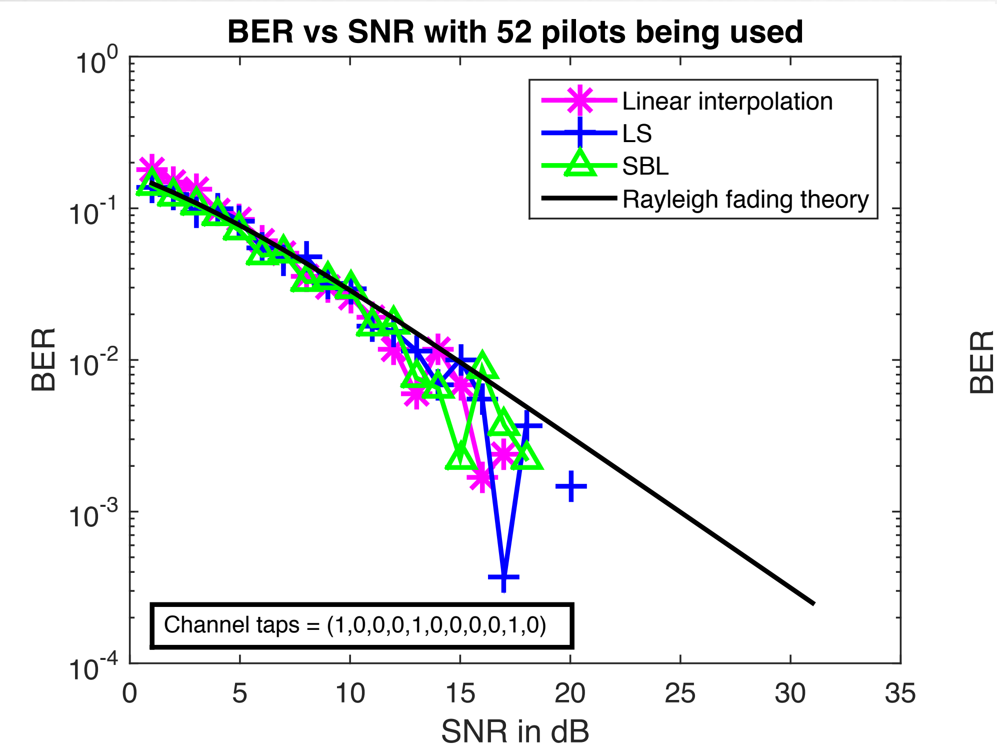
BER Performance

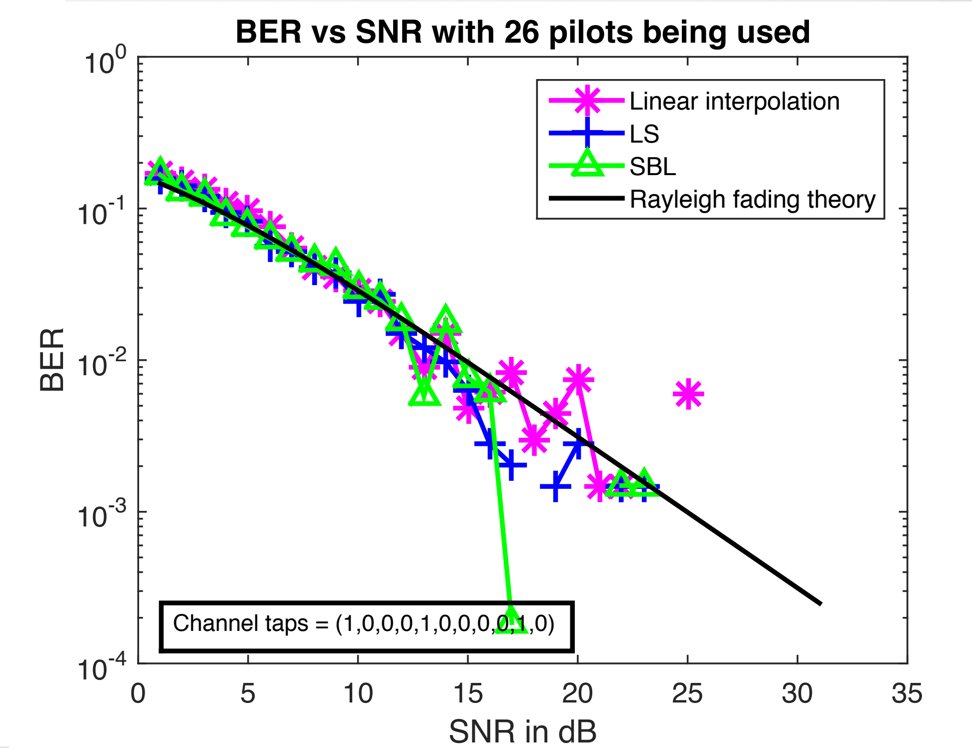


But Does it Work?

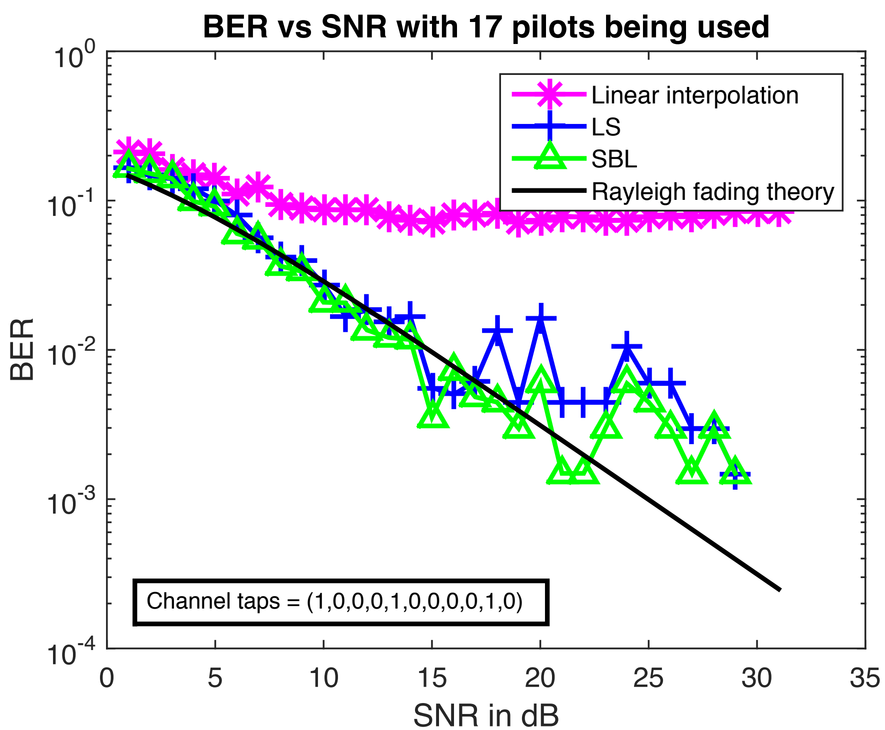
- Implementation on GNU Radio platform
  - In C++/Python
- Integrated into a USRP-based test setup
- Single-antenna OFDM, 64 subcarriers, CP length 16
- Channel estimation
  - Least-squares estimation
  - Sparse Bayesian Learning
  - Frequency-domain interpolation

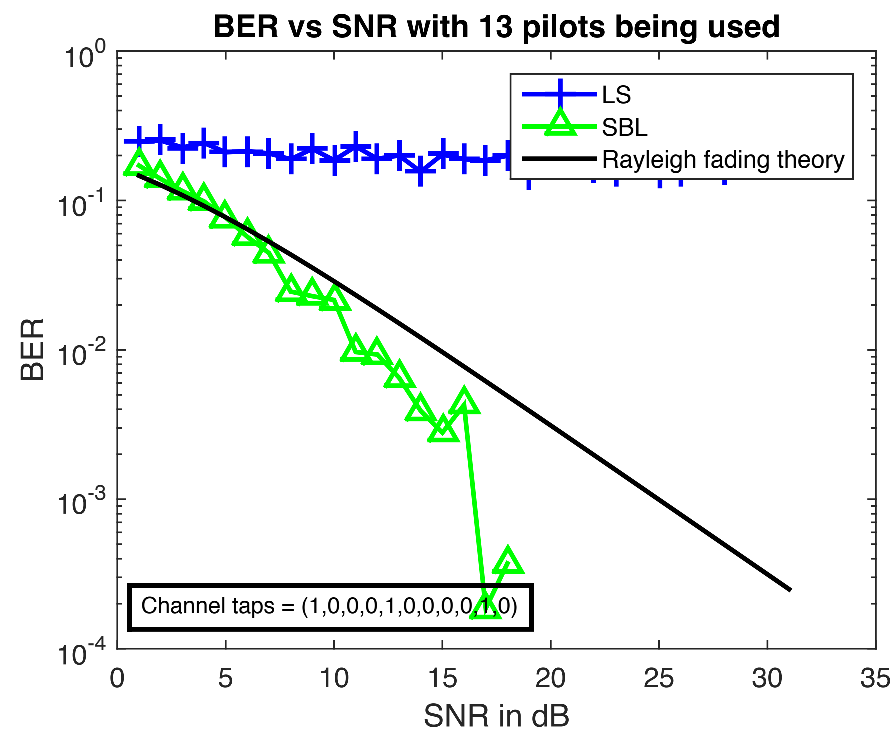
GNU-Radio Loopback-Mode Simulation Results



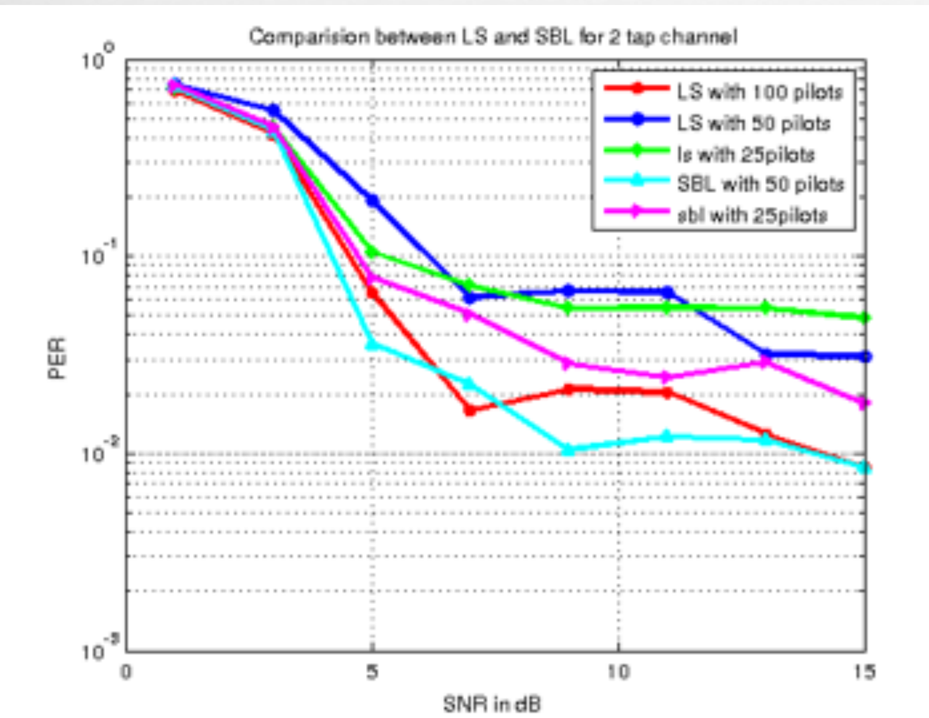


GNU-Radio Loopback-Mode Simulation Results

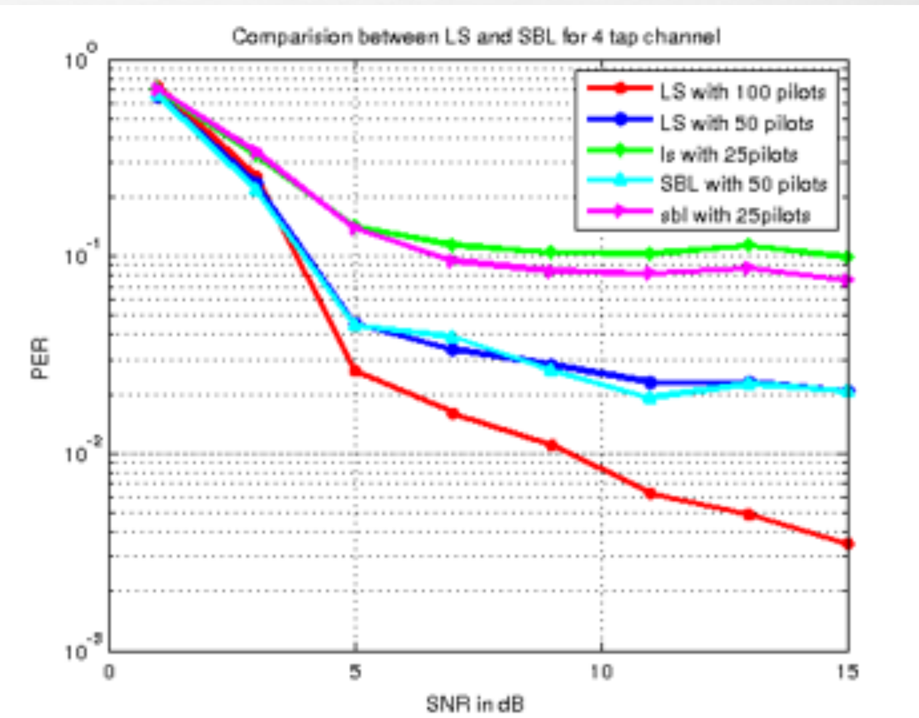




Over-the-air Results



2-tap channel



3-tap channel

OFDM system, 256 subcarriers, CP length 16, 4-QAM

To Recap

- SBL based OFDM channel estimation
- Block-fading case: proposed J-SBL and low-complexity recursive J-SBL for joint channel estmn & data detn
- Time-varying case: low-complexity K-SBL and JK-SBL proposed
  - Algos fully exploit channel correlation
- MIMO case: Estimation in MMV framework
- Take-home point: Exploit any known structure!

Further Extensions

- MIMO-OFDM: tracking time-varying channels using the Kalman framework [Prasad et al., TSP 2015]
- Cluster sparsity: paths occur in closely spaced clusters [Prasad et al., ICASSP 2014]
- Approximate sparsity due to transmit/receive pulse shaping, filtering, etc [Prasad et al., TSP Jul. 2014]

Summary

- Bayesian methods:
  - Simple updates
  - Promising performance
- Challenges:
  - Theoretical analysis
  - New algorithms
  - Novel applications
- Plenty of opportunities!

References

- J. M. Adler, B. Rao, and K. Kreutz-Delgado, **Comparison of basis selection methods**, Asilomar 1999
- S. F. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, **Sparse solutions to linear inverse problems with multiple measurement vectors**, IEEE Trans. Sig. Proc., 2005
- D. Wipf, B. Rao, and S. Nagarajan, **Latent variable Bayesian models for promoting sparsity,** IEEE Trans. on Inform. Theory, 2011
- D. Wipf and B. Rao, **An empirical bayesian strategy for solving the simultaneous sparse approximation problem,** IEEE Trans. Sig. Proc., 2007
- Z. Zhang and B. Rao, **Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning,** IEEE J-STSP, 2011
- Z. Zhang and B. Rao, **Recovery of block sparse signals using the framework of block sparse bayesian learning**, ICASSP 2012
- R. Giri, B. Rao, **Type I and Type II Bayesian Methods for Sparse Signal Recovery using Scale Mixtures,** submitted, IEEE Trans. Sig. Proc., 2015

References

- R. Prasad and C. R. Murthy, **Cramér-Rao-Type Bounds for Sparse Bayesian Learning**, IEEE Transactions on Sig. Proc., vol. 61, no. 3, pp. 622-632, Mar. 2013
- R. Prasad, C. R. Murthy and B. Rao, **Joint Approximately Sparse Channel Estimation and Data Detection in OFDM Systems using Sparse Bayesian Learning**, IEEE Trans. Sig. Proc., Jul. 2014
- R. Prasad and C. R. Murthy, **Joint Approximately Group Sparse Channel Estimation and Data Detection in MIMO-OFDM Systems Using Sparse Bayesian Learning**, NCC 2014 [best paper award!]
- S. Khanna and C. R. Murthy, **Decentralized Bayesian Learning of Jointly Sparse Signals**, Globecom 2014
- V. Vinuthna, R. Prasad, and C. R. Murthy, **Sparse signal recovery in the presence of colored noise and rank-deficient noise covariance matrix: an SBL approach**, ICASSP 2015
- R. Prasad, C. R. Murthy, and B. D. Rao, **Joint Channel Estimation and Data Detection in MIMO-OFDM Systems: A Sparse Bayesian Learning Approach**, IEEE Trans. on Sig. Proc., Oct. 2015

References

- Y. Wang, D. Wipf, J-M. Yun, W. Chen, I. Wassel, **Clustered Sparse Bayesian Learning**, UAI 2015
- D. Wipf, J-M. Yun, Q. Ling, **Augmented Bayesian Compressive Sensing**, DCC 2015
- B. Xin, Y. Wang, W. Gao and D. Wipf, **Maximal Sparsity with Deep Networks?** ArXiv:1605.01636v2, May 2016

Acknowledgements

Geethu Joseph

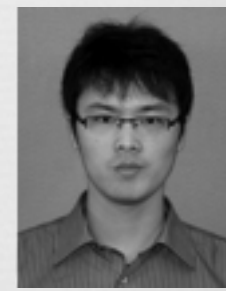
Ritwik Giri

Yu Wang

Saurabh Khanna

Jason Palmer

Bo Xin



Ranjitha Prasad

Bhaskar Rao

Thank you!