Bayesian Methods for Sparse Signal Recovery

Chandra R. Murthy
Dept. of ECE
Indian Institute of Science

cmurthy@ece.iisc.ernet.in

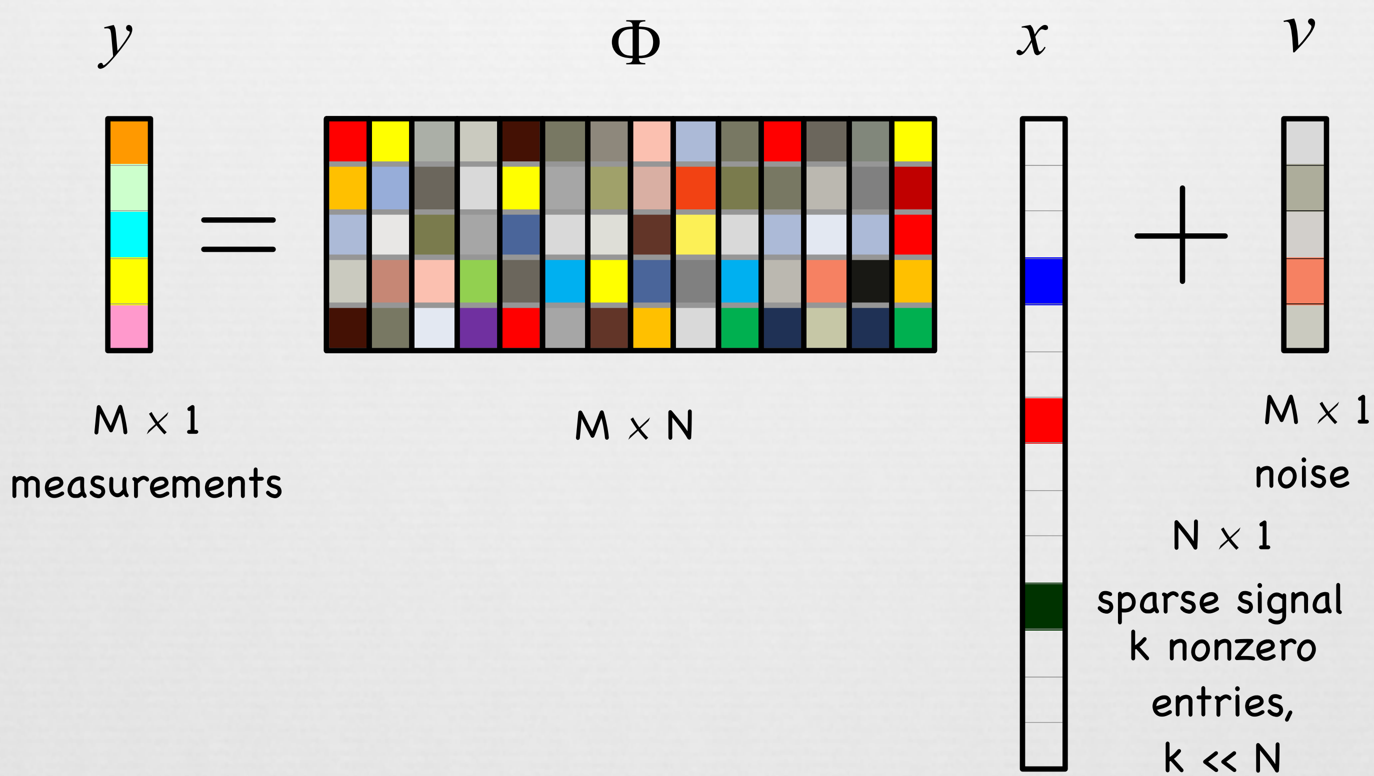
Outline

- Setting the stage
- Non convex methods for sparse recovery
- Sparse Bayesian learning
- Extensions
- Application to wireless communication
  - Channel estimation

Part 1: Setting the Stage

Motivation and background

Sparse Signal Recovery

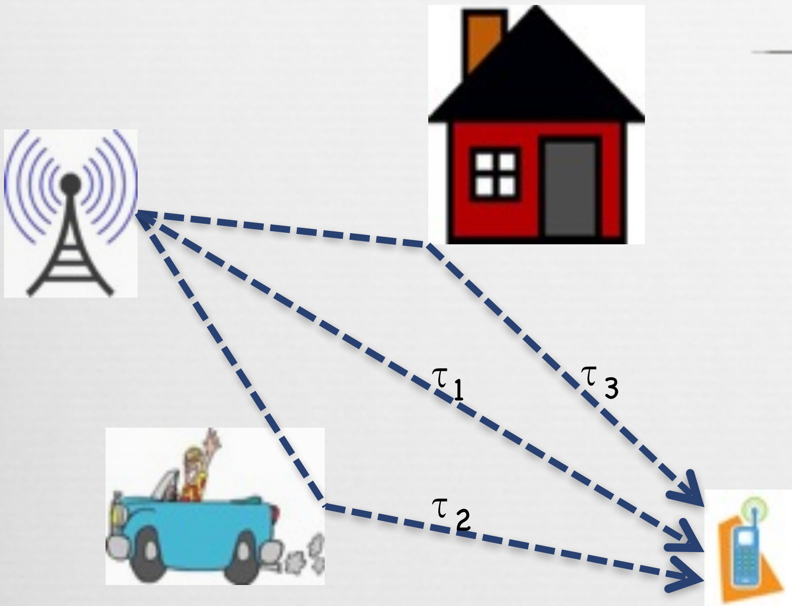


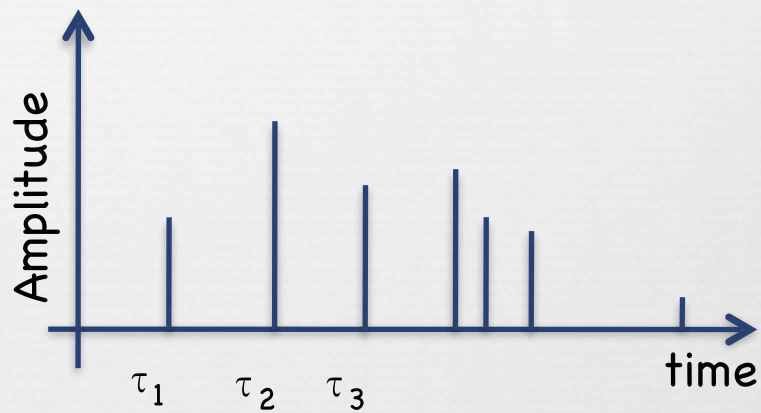
- **Goal:** Recover x from y
- M << N: infinitely many solutions

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 and M.,...)

Wireless Channel Estimation





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

The Problem

- Noiseless case: Given y and Φ, solve

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

- Noisy case: solve

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

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

Breakthrough 1: Uniqueness

- Underdetermined systems
  - Infinitely many solutions, but ...
  - Unique "sparse" solution if nullspace has no "sparse" vectors [Donoho, Elad '02]
  - Unique soln. with high probability, if $M \geq k+1$ [Bresler; Wakin etc]
- Sub-Nyquist sampling (compression) when:
  - Restrict to sparse signals
  - Sample in an "appropriate" basis

Breakthrough 2: Just Relax!

- l_1 min. instead of l_0 min.

$$\min \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = \Phi\mathbf{x}$$

  - Convex optimization problem
- Same solution as l_0 minimization!
  - If the measurement matrix is random
  - Use slightly larger number of measurements
  - Robust to measurement noise

$$M \approx K \log\left(\frac{N}{K}\right) \ll N$$

- See [Donoho; Candes, Romberg, Tao etc]

Recovery Algorithms

- Sequential recovery methods: Sequentially identify columns of Φ most aligned with the residual
  - Matching pursuit [Mallat, Zhang; Cotter, Rao]
  - Orthogonal matching pursuit [Tropp 03]
  - CoSAMP [Needell, Tropp]

$$\min_{\mathbf{x}} \|\mathbf{x}\|_p \text{ subject to } \mathbf{y} = \Phi\mathbf{x}$$

- Joint recovery methods: Use a cost function that encourages sparse solutions
  - Basis pursuit (l-p, with p=1) [Chen et al.]
  - FOCUSS (l-p, with p < 1) [Gordonitsky et al.]
  - Lasso (BPDN) [Tibshirani]
  - Dantzig selector [Candes, Tao]

$$\min_{\mathbf{x}} \tau\|\mathbf{x}\|_1 + \|\mathbf{y} - \Phi\mathbf{x}\|_2^2$$

Performance Guarantees

- Mutual coherence $\Phi = [\phi_1, \phi_2, \ldots, \phi_N]$ $\mu(\Phi) \triangleq \max_{1 \leq i,j \leq N, i \neq j} \frac{|\phi_i^T \phi_j|}{\|\phi_i\|_2 \|\phi_j\|_2}$
- Result (noiseless case): If $\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\Phi)}\right)$
  - OMP converges x after k iterations, where k = num. nonzeros in x [Tropp 03]
  - The sparse vector x_0 that generated y is the unique soln to [Donoho, Elad 03]

$$\min \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = \Phi\mathbf{x}$$

- Similar guarantees in the noisy case & in terms of restricted isometry constant etc.

Limitations of Greed & Relaxation

- Performance of BP and OMP depend on the form of the dictionary Φ
  - Poor performance when condns. violated
  - Hard to relate estimation error to the dictionary
- BP: perf. indep. of nonzero coeffs [Malioutov et al. 2004]
  - Performance does not improve when situation is favorable
- OMP: performance highly sensitive to magnitudes of nonzero coeffs
  - 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
- Correlated dictionary: disrupts l_0-l_1 equivalence
- Estimating embedded params (e.g., in Φ)

To Recap

- Sparse signal recovery
  - Basic problem, breakthroughs in CS
  - Algorithms
  - Guarantees
- Limitations
  - Scaling/shrinkage
  - Correlated dictionary
  - Embedded parameters

Part 2: Dont Relax!

A time and place for nonconvex methods?

Bayesian Methods

- MAP estmn. using a sparse linear model
  - Also a regression problem with sparsity promoting penalties (e.g., l_p-norm)
  - l_1-min (BP/LASSO) is a special case
- Algorithms:
  - Iterative reweighted l_1 [Candes et al. 2008]
  - Iterative reweighted l_2 [Chartrand & Yin 2008]
  - EM-based SBL [Tipping, 2001], [Wipf, Rao 2007]
  - AMP [Schniter 2008], [Rangan 2011]

MAP Estimation

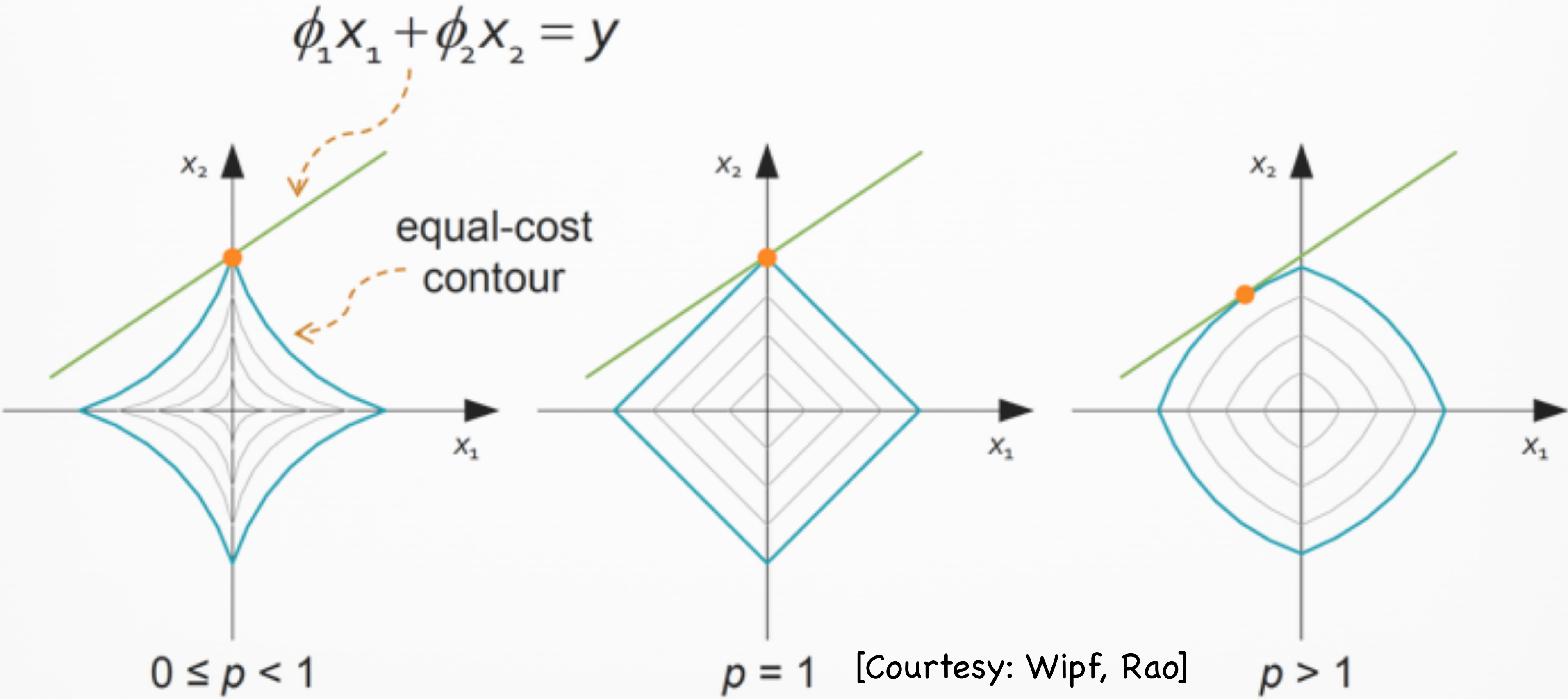
$$\begin{aligned}\hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \\ &= \arg\min_{\mathbf{x}} -\log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{x}) \quad \text{(Bayes' rule)} \\ &= \arg\min_{\mathbf{x}} \|\mathbf{y} - \Phi\mathbf{x}\|_2^2 + \lambda \sum_{i=1}^{N} g(|x_i|)\end{aligned}$$

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 estmn problem has at most M nonzeros [Rao et al., 99]

Why does it work?

- Min $|x_1|^p + |x_2|^p$ subject to $\phi_1 x_1 + \phi_2 x_2 = y$



The Optimization Problem

- To solve

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

  - g(x) concave, monotonically ↑ in |x|
  - G(x) convex + concave

Majorization-Minimization Approach

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

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

- Step 2: Set m <- m+1, update $g(x|x^{(m)})$, iterate
- **Works because**
  $G(x^{(m+1)}) \leq G(x^{(m+1)}|x^{(m)}) \leq G(x^{(m)}|x^{(m)}) = G(x^{(m)})$

Iterative Reweighted l_1

- Concavity: $g(x) \le g'(x^{(m)})(x-x^{(m)}) + g(x^{(m)})$
  - 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

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

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

Weighted l_1 minimization

Iterative Reweighted l_2

- $g(x)$ concave in x^2: $$g(x) \leq \left(\frac{\partial g(\sqrt{x^2})}{\partial (x^2)} \bigg|_{x=x_0} \right) (x^2 - x_0^2) + g(x_0)$$
- Optimization problem
$$\mathbf{x}^{(m+1)} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \Phi\mathbf{x}\|_2^2 + \lambda \sum_{i=1}^{N} w_i^{(m)} |x_i|^2$$
$$\|\mathbf{W}_m^{-\frac{1}{2}}\mathbf{x}\|_2^2$$
- Iterative reweighted l2 [Chartrand et al. 08]
  - Init: $m = 0$, $x^{(m)}$ = something convenient
  - Iterate:
    - Compute $$\mathbf{x}^{(m+1)} = \mathbf{W}_m \Phi^T (\lambda \mathbf{I} + \Phi \mathbf{W}_m \Phi^T)^{-1} \mathbf{y}$$
    - m <- m+1, update W_m
  - Until convergence

An Example

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

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

- 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 $p(x = x_o|y)$
  - 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

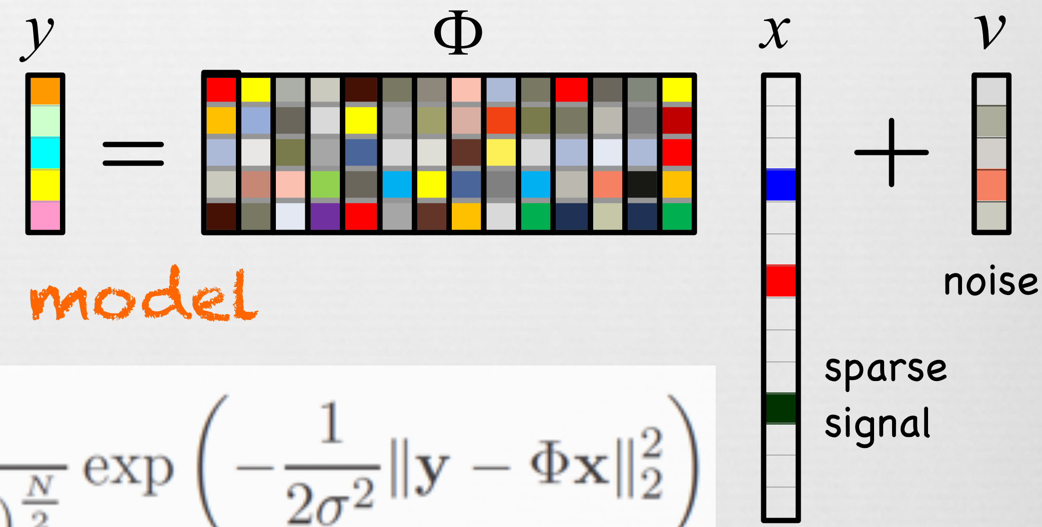
- Bayesian estimation
  - Basic MAP estimation
  - Majorization-minimization approach
  - Iterative reweighted algorithms
- Limitations
  - Many local minima
  - Posterior mean vs. posterior mode

Part 3: Sparse Bayesian Learning

Use lots of priors and pick the best one!

Setup

- Recall the canonical model



- Gaussian noise model

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

- General parameterized prior

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

Sparse Bayesian Methods

- Estimate γ_i from the data: Type-II ML

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

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

- SBL Cost function

$$\mathcal{L}(\Gamma) \propto -\log\det\left(\Sigma_{\mathbf{y}}\right) - \mathbf{y}^T\Sigma_{\mathbf{y}}^{-1}\mathbf{y}$$

A Simple Suboptimal Procedure

- Just maximize the integrand. Leads to

$$\min_{\mathbf{x},\Gamma} \frac{\|\mathbf{y}-\Phi\mathbf{x}\|^2}{2\sigma^2} + \sum_{i=1}^{n} \frac{|x_i|^2}{2\gamma_i} + \frac{1}{2}\log\gamma_i$$

- Alternating minimization:
  - Initialize $\Gamma = I$
  - Compute $\hat{\mathbf{x}} = \sigma^{-2}\left(\sigma^{-2}\Phi^T\Phi + \Gamma^{-1}\right)^{-1}\Phi^T\mathbf{y}$

    $\gamma_i = \hat{x}_i^2$
  - Repeat
- Will call this "Approximate MAP" or A-MAP estimation

The EM Iterations

- **E-step:** posterior distribution given $\Gamma^{(t)}$:

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

  - The posterior distribution is

$$p(\mathbf{x}|\mathbf{y};\Gamma^{(t)}) = \mathcal{N}(\mu, \Sigma)$$

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

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

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

The SBL Algorithm

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

Variational Interpretation

- Lower bound on L:

$$
\begin{aligned}
\mathcal{L}(\Gamma) &= \log \int q_{\mathbf{x}}(\mathbf{x}) \frac{p(\mathbf{x}, \mathbf{y}; \Gamma)}{q_{\mathbf{x}}(\mathbf{x})} \mathrm{d}\mathbf{x} \\
&\geq \int q_{\mathbf{x}}(\mathbf{x}) \log \left(\frac{p(\mathbf{x}, \mathbf{y}; \Gamma)}{q_{\mathbf{x}}(\mathbf{x})} \right) \mathrm{d}\mathbf{x} \\
&\triangleq \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}); \Gamma)
\end{aligned}
$$

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)
  - Unfortunately, fixed point not necessarily a local min or saddle point [Wipf and Nagarajan 09]
  - But, not found to be a problem in practice
- The global min of L occurs at the sparsest solution in the noiseless case [Wipf et al. 04]
- All local minima occur at sparse solutions in the noisy case [Wipf et al. 04]
- More properties [Wipf and Nagarajan 09]

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:

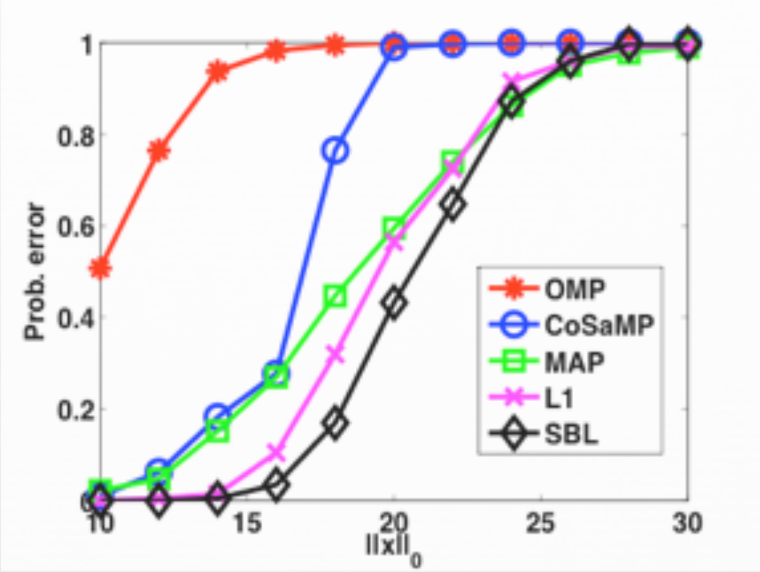
$$\mathbf{x}_{\text{opt}} = \arg\min_{\mathbf{x}} \|y - \Phi\mathbf{x}\|_2^2 + \sigma^2 g_{\text{SBL}}(\mathbf{x})$$

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

  - Facilitates iterative reweighted l_1 and l_2 algorithms [Wipf and Nagarajan, 09]
  - Overcomes some limitations of EM

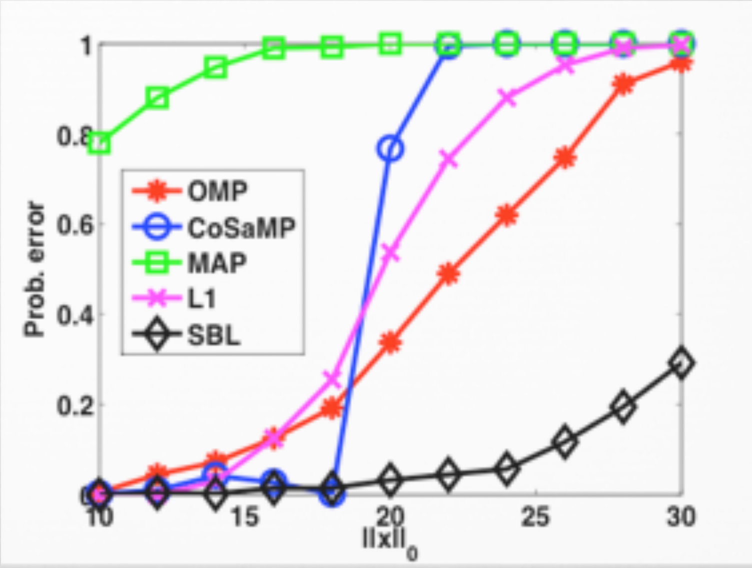
- Replace E-step with an approx. posterior computation: AMP-SBL [Al-Shoukairi and Rao 14]

Empirical Example



Unit magnitude entries

- Generate random 50 x 100 matrix A
- Generate sparse vector x_0
- Compute $y = Ax_0$
- Solve for x_0, average over 1000 trials
- Repeat for different sparsity values



Highly scaled entries

Advantages of SBL

- Averaging over x: fewer minima in $p(y;\gamma)$
- Versatile: γ can also be used to
  - Tie several parameters together - fewer parameters to estimate
  - Incorporate structure
    - Block/cluster sparsity
    - Intra/inter-vector correlation

Colored Noise

- In many applications, noise may be
  - Colored
  - Rank-deficient covariance matrix
- Example 1: interference with a known direction of arrival
- Example 2: Good cop, bad cop: expensive, noiseless meas. or cheap, noisy meas.?

Model

- Measurement model $$\mathbf{y} = \mathbf{\Phi x} + \mathbf{n}$$
- Noise model

$$\mathbb{E}[nn^T] = \mathbf{Q} = [\mathbf{V}_1\mathbf{V}_2]\begin{bmatrix} \mathbf{D} & \mathbf{0}_{p\times m-p} \\ \mathbf{0}_{m-p\times p} & \mathbf{0}_{m-p\times m-p} \end{bmatrix}\begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix}$$

- Equivalent model

$$\mathbb{E}[\mathbf{n}_1\mathbf{n}_1^T] = \mathbf{D}$$

$$\mathbf{y}_1 = \mathbf{\Phi}_1\mathbf{x} + \mathbf{n}_1$$

$$\mathbb{E}[\mathbf{n}_2\mathbf{n}_2^T] = \mathbf{0}_{m-p}$$

$$\mathbf{y}_2 = \mathbf{\Phi}_2\mathbf{x} + \mathbf{n}_2$$

- How to recover x from $\{y_1, y_2\}$?

CoNo-SBL

- E-Step:

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

- Posterior density

$$p\left(\mathbf{x}|\mathbf{y};\boldsymbol{\gamma}^{(r)}\right) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} = \Gamma^{(r)} - \Gamma^{(r)}\left(\sum_{m=1}^{2}\sum_{n=1}^{2}\boldsymbol{\Phi}_n^T\mathbf{B}_{nm}\boldsymbol{\Phi}_m\right)\Gamma^{(r)} \quad \mu = \boldsymbol{\Sigma}\boldsymbol{\Phi}_1^T\mathbf{D}^{-1}\mathbf{y}_1 + \sigma_2^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}_2^T\mathbf{y}_2$$

$$\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{D} + \boldsymbol{\Phi}_1^T\Gamma^{(r)}\boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_1\Gamma^{(r)}\boldsymbol{\Phi}_2^T \\ \boldsymbol{\Phi}_2\Gamma^{(r)}\boldsymbol{\Phi}_1^T & \sigma_2^2\mathbf{I}_{m-p} + \boldsymbol{\Phi}_2\Gamma^{(r)}\boldsymbol{\Phi}_2^T \end{bmatrix}^{-1}$$

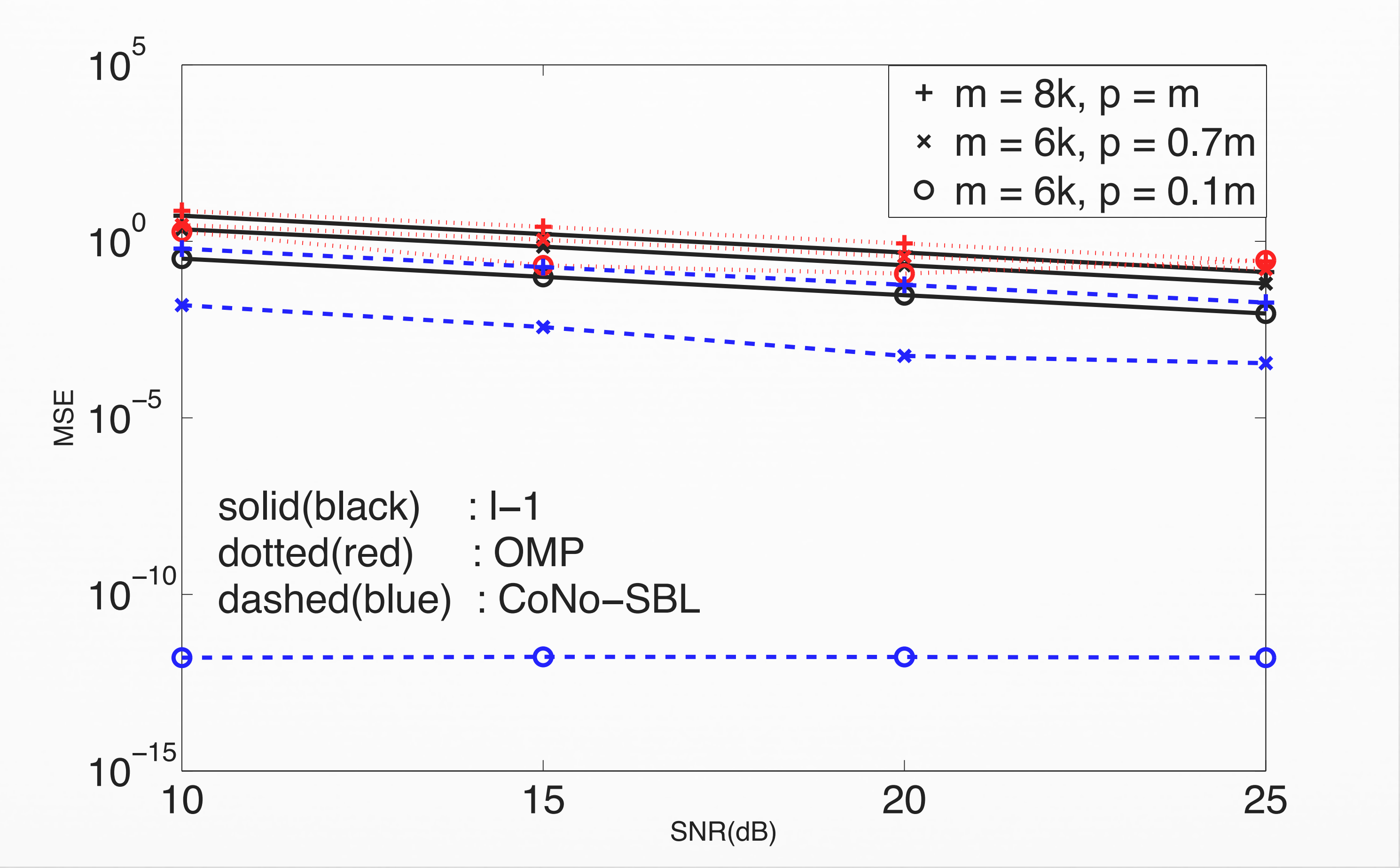
CoNo-SBL (contd)

- Can let $\sigma_2^2 \rightarrow 0$ by using easy results from block matrix inversion and Woodbury identity
  - For example: (Details: [Vinjamuri & M., ICASSP 15])

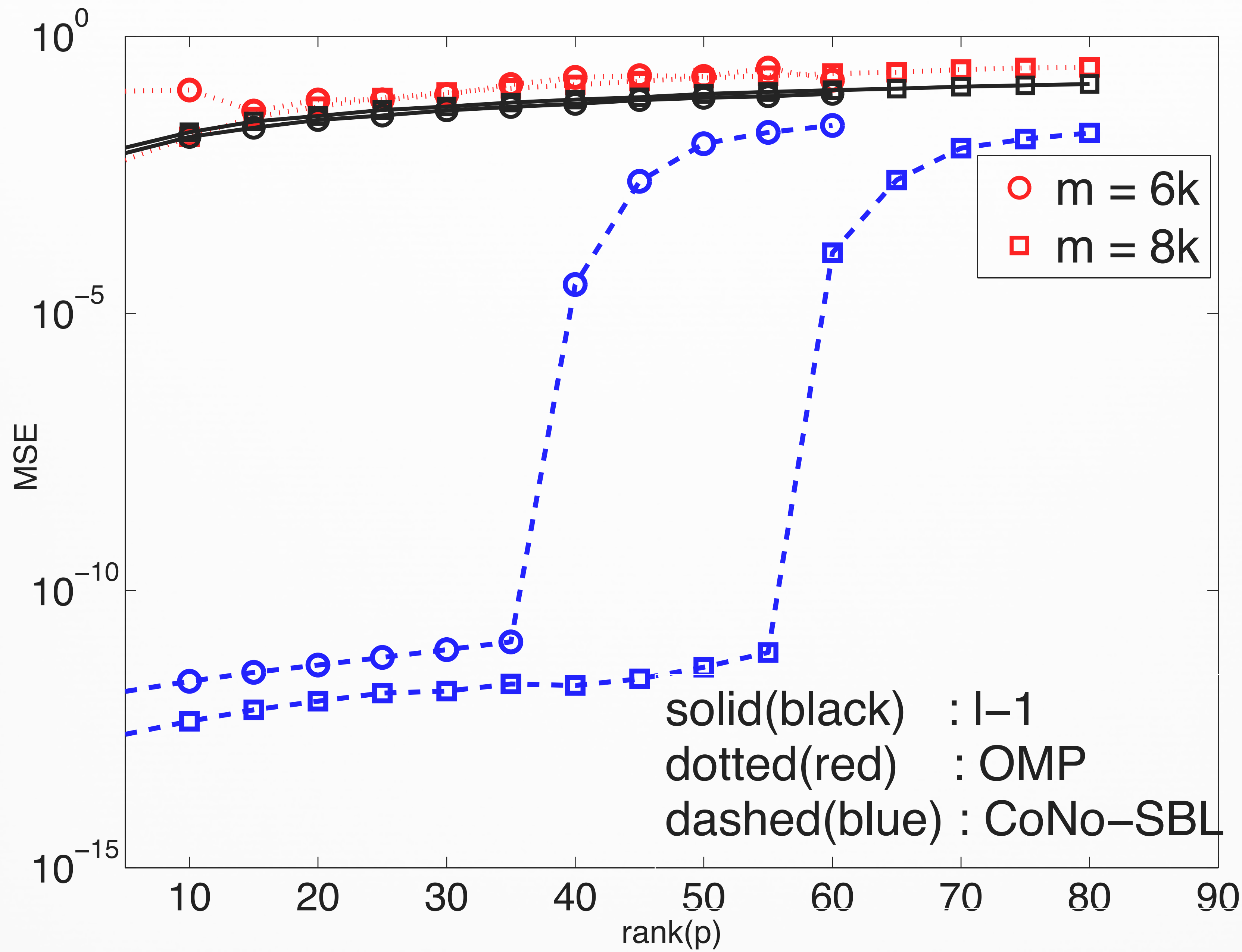
$$\mu = \mathbf{\Sigma}\mathbf{\Phi}_1^T\mathbf{D}^{-1}\mathbf{y}_1 + \Gamma^{(r)\frac{1}{2}}\mathbf{U}_2^{\frac{1}{2}}(\Theta_2\mathbf{U}_2^{\frac{1}{2}})^{\dagger}\mathbf{y}_2$$

- M-Step same as before: $\gamma_i^{(r+1)} \leftarrow |\boldsymbol{\mu}_i|^2 + \mathbf{\Sigma}_{ii}$

Empirical Example



N = 100
k = 10



MSE
rank(p)
10^0
10^−5
10^−10
10^−15
10
20
30
40
50
60
70
80
90
m = 6k
m = 8k
solid(black) : l−1
dotted(red) : OMP
dashed(blue) : CoNo−SBL

To Recap

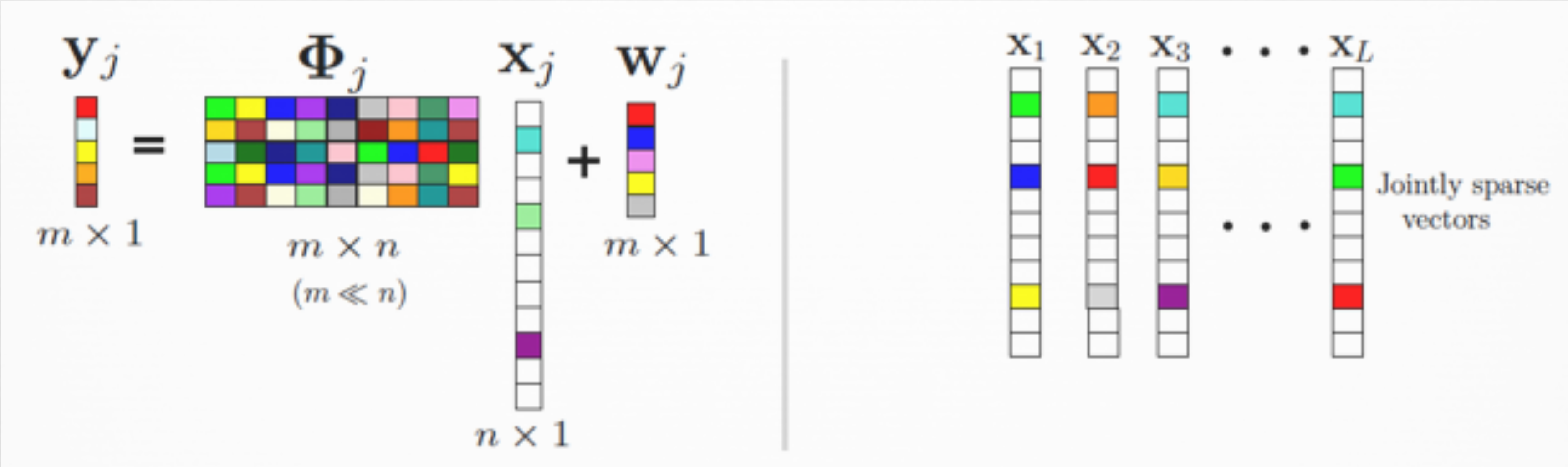
- Sparse Bayesian learning
  - Sparse vector recovery via estimating hyperparameter
  - Expectation-maximization iterations
  - Convergence properties
  - Alternative implementations
- Limitations
  - Computational complexity
    - More recent algos overcome this
  - Slow convergence
    - Fast versions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

Part 4: Extensions

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

Multiple Measurement Vectors: Joint Sparsity

- Observation Model



- Why? As L -> ∞, with m = 1,

  P(exact support recov.) -> 1 [Baron et al. 09]

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

Algos for Joint Sparse Recovery

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

Sparse vector dimension

Num. measurements

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

- M-Jeffreys

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

- M-FOCUSS

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

The M-SBL Algo

- Cost function

$$p(\mathbf{Y};\boldsymbol{\gamma}) = \int p(\mathbf{Y},\mathbf{X};\boldsymbol{\gamma})\mathrm{d}\mathbf{X} = \prod_{j\in\mathcal{J}} \int p(\mathbf{y}_j|\mathbf{x}_j)p(\mathbf{x}_j;\boldsymbol{\gamma})\mathrm{d}\mathbf{x}_j$$

$$\mathcal{J} = \{1,2,\ldots,L\}$$

- EM Iterations

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

$$\text{M-step:} \quad \boldsymbol{\gamma}^{k+1} = \underset{\boldsymbol{\gamma}\in\mathbb{R}_+^n}{\arg\max}\; Q\left(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k\right)$$

- Posterior distribution $p(\mathbf{x}_j|\mathbf{y}_j;\boldsymbol{\gamma}^k) \sim \mathcal{N}\left(\boldsymbol{\mu}_j^{k+1}, \boldsymbol{\Sigma}_j^{k+1}\right)$

E & M Steps

- E Step:

$$\boldsymbol{\Sigma}_j^{k+1} = \boldsymbol{\Gamma}^k - \boldsymbol{\Gamma}^k \boldsymbol{\Phi}_j^T \left(\sigma_j^2 \mathbf{I}_m + \boldsymbol{\Phi}_j \boldsymbol{\Gamma}^k \boldsymbol{\Phi}_j^T \right)^{-1} \boldsymbol{\Phi}_j \boldsymbol{\Gamma}^k$$

$$\boldsymbol{\mu}_j^{k+1} = \sigma_j^{-2} \boldsymbol{\Sigma}_j^{k+1} \boldsymbol{\Phi}_j^T \mathbf{y}_j$$

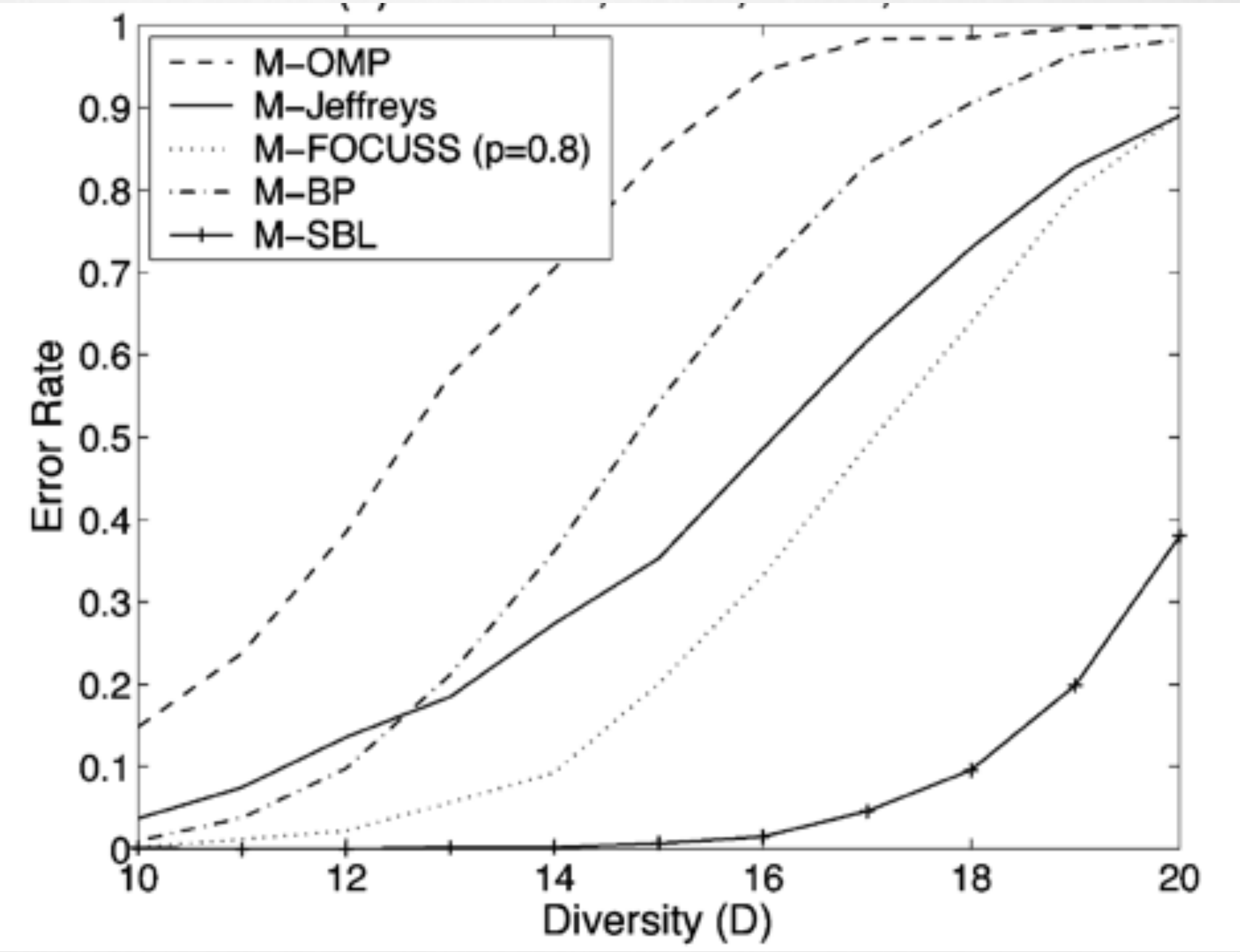
- M Step:

$$\gamma^{k+1}(i) = \frac{1}{L} \sum_{j \in \mathcal{J}} \left(\boldsymbol{\Sigma}_j^{k+1}(i,i) + \boldsymbol{\mu}_j^{k+1}(i)^2 \right)$$

  - Average of the individual estimates of γ_i across measurements

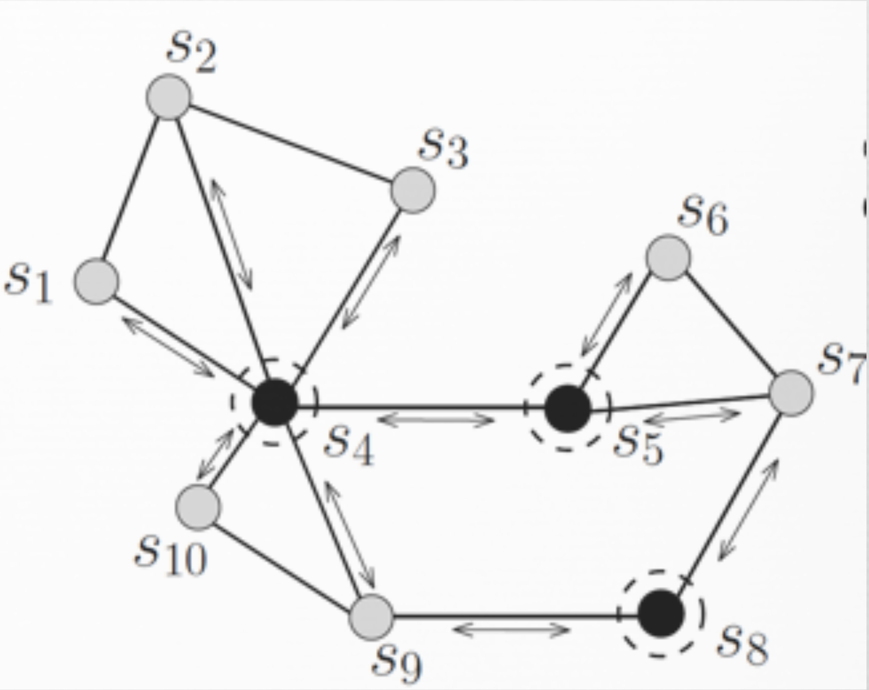
Empirical Example

- M = 25
  N = 50
  L = 3
- Source:
  [Wipf & Rao, TSP Aug. 04]



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



SBL for Joint Sparse Recovery

- EM Iterations:
  - E-step: $\mathbf{\Sigma}_j^{k+1} = \mathbf{\Gamma}^k - \mathbf{\Gamma}^k \mathbf{\Phi}_j^T \left(\sigma_j^2 \mathbf{I}_m + \mathbf{\Phi}_j \mathbf{\Gamma}^k \mathbf{\Phi}_j^T \right)^{-1} \mathbf{\Phi}_j \mathbf{\Gamma}^k$

    $\boldsymbol{\mu}_j^{k+1} = \sigma_j^{-2} \mathbf{\Sigma}_j^{k+1} \mathbf{\Phi}_j^T \mathbf{y}_j$

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

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

A Simple Trick

- Equivalent problems

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

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

- For distributed implementation

$$\arg \min_{\gamma_j, j \in [L]} \sum_{j \in [L]} |\gamma_j - a_j|^2$$

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

Bridge nodes
Linear constraints

Alternating Directions Method of Multipliers

- General problem

$$\min_{\{\mathbf{x},\mathbf{y}\}} f(\mathbf{x}) + g(\mathbf{y})$$

$$\text{subject to } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{c}$$

- Augmented Lagrangian

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

- ADMM iterations

Convex problems, easy to solve

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

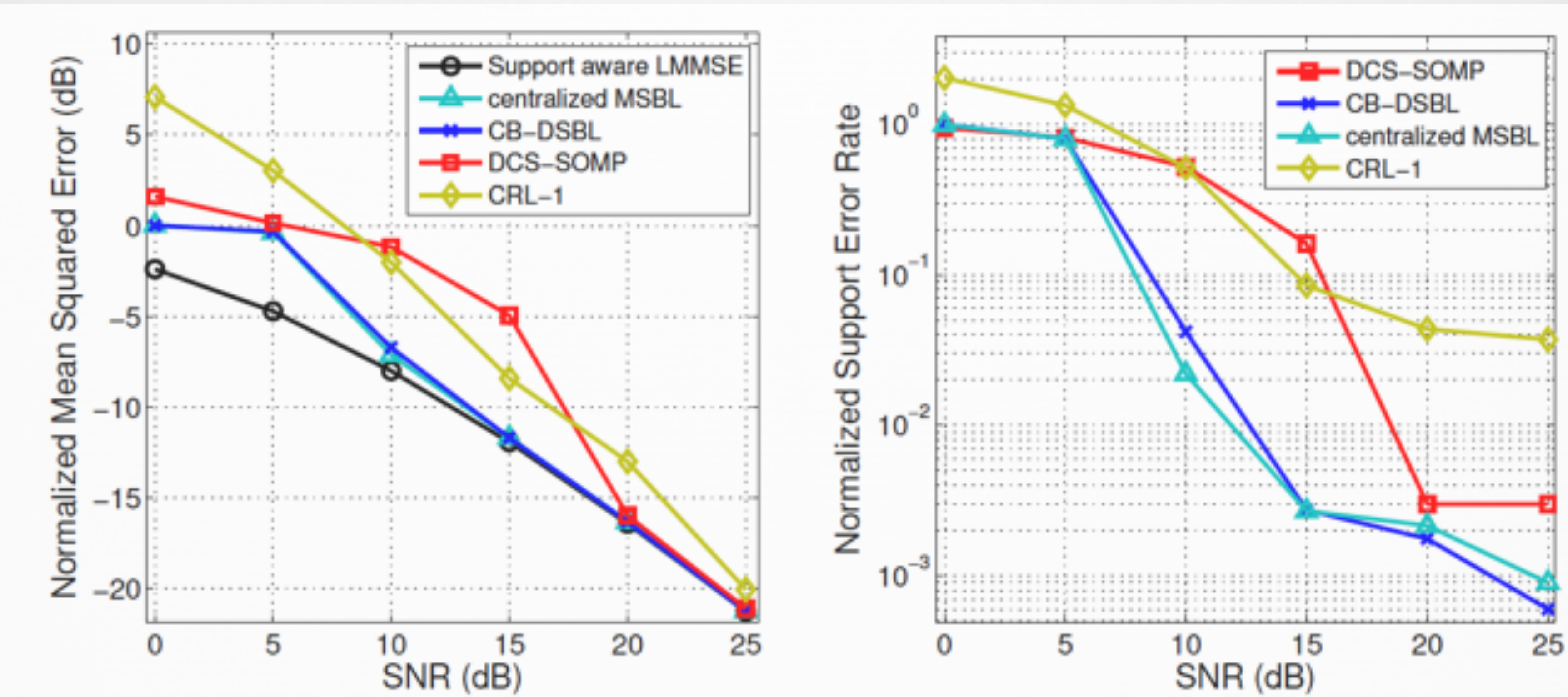
Dual update

$$\lambda^{(k+1)} = \lambda^{(k)} + \rho(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} - \mathbf{c})$$

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 $\rho_{\text{opt}} = \left(\frac{1}{\text{min. no. of bridge nodes per node}} \right)$

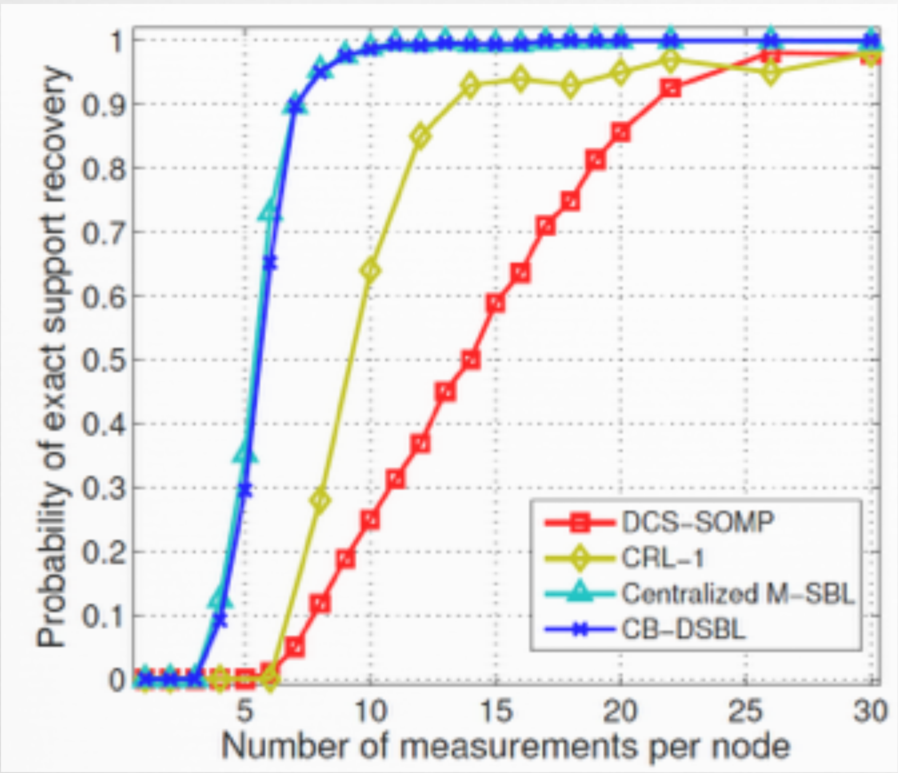
Simulation Result: Mean Squared Error

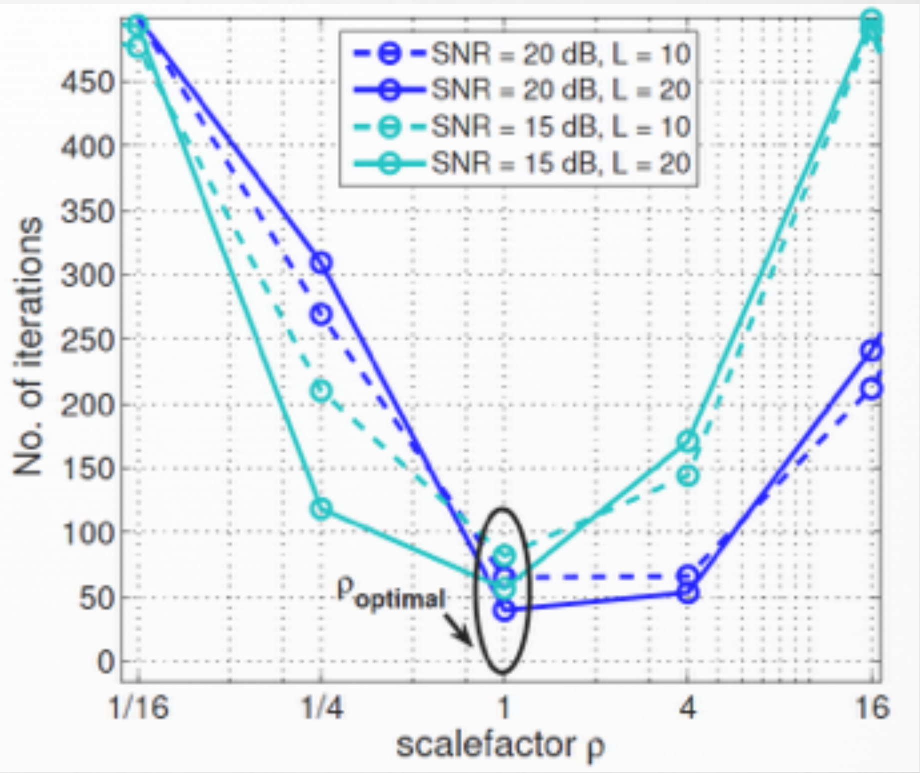


L = 10 nodes, n = 50, m = 10, 10% sparsity

[S. Khanna, C. R. Murthy, Globecom 2014]

Support Recovery & ADMM Parameter ρ





L = 10 nodes, n = 50, SNR = 15dB (L), m = 10 (R), 10% sparsity

[S. Khanna, C. R. Murthy, Globecom 2014]

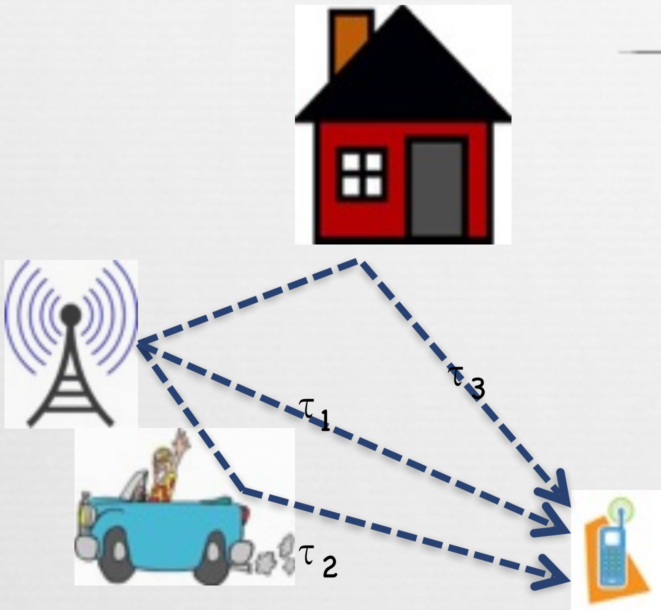
To Recap

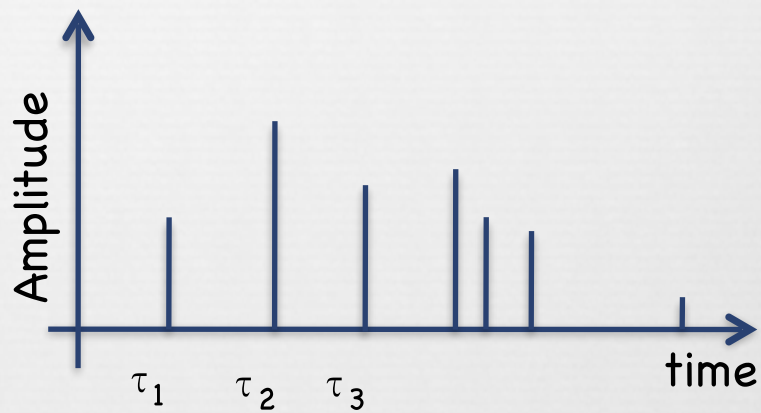
- Multiple measurement vectors
  - M-SBL algorithm
  - Exploits joint sparsity
- Distributed sparse signal recovery
  - ADMM iterations
  - Simulation examples

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

- Time-varying channel:

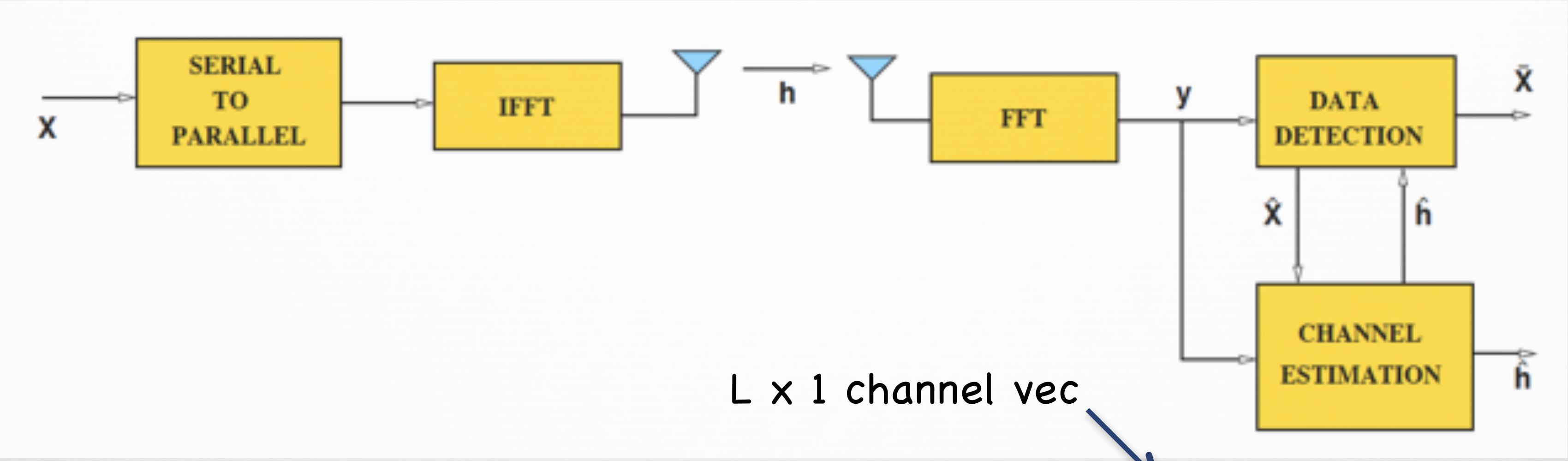
  Channel varies from symbol-to-symbol

  - Want to exploit temporal correlation (group-sparse estimation)

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

Support-Aware EM

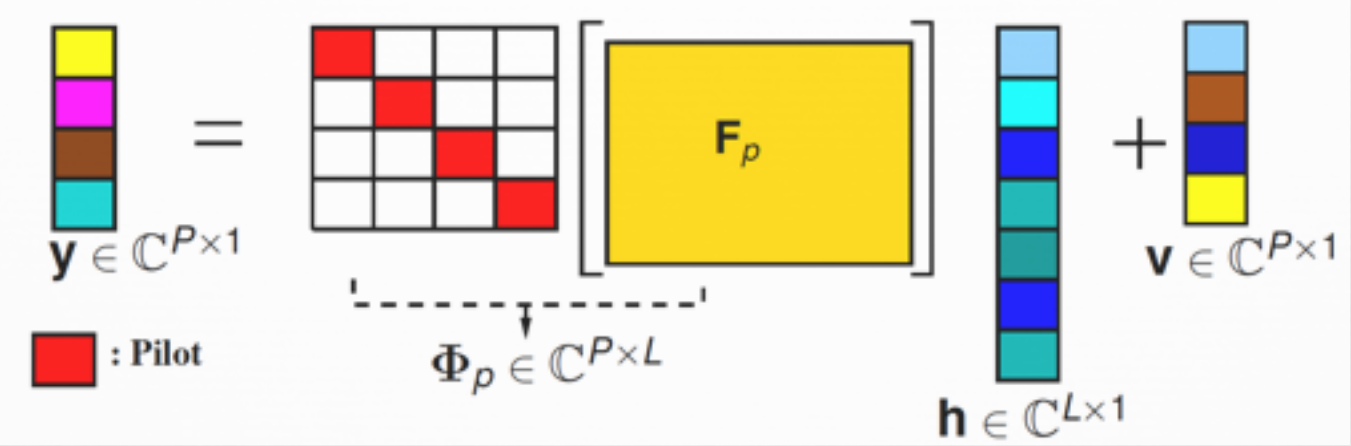
- Joint channel estimation and data detection
- E-Step: $$Q\left(\mathbf{X}|\mathbf{X}^{(t)}\right) = E_{\mathbf{h}|\mathbf{y},\mathbf{X}^{(t)}}\left(\log p(\mathbf{y},\mathbf{h}|\mathbf{X})\right)$$

$$\mathbf{X}^{(t+1)} = \arg\max_{\mathbf{X}} Q\left(\mathbf{X}|\mathbf{X}^{(t)}\right)$$

- M-Step:

$$\log p(\mathbf{y},\mathbf{h}|\mathbf{X}) = \underbrace{\log p(\mathbf{y}|\mathbf{h},\mathbf{X})}_{\text{Log Likelihood, func. of }\mathbf{X}} + \underbrace{\log p(\mathbf{h})}_{\text{not a func. of }\mathbf{X}}$$

Sparse Channel Estimation from Pilot Symbols

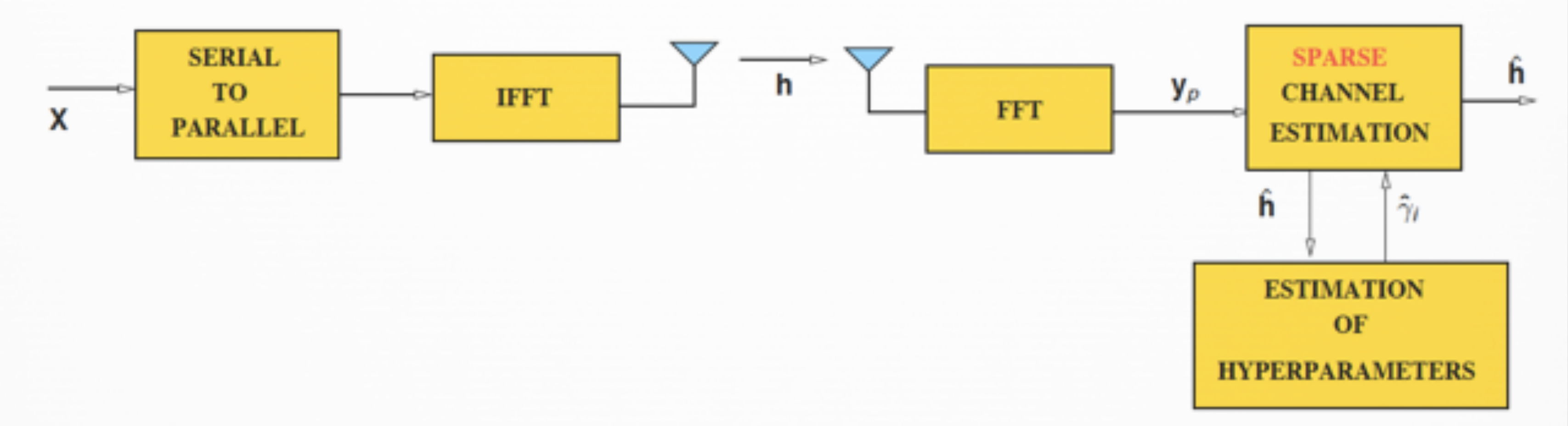


- h sparse in time (lag) domain
- Hierarchical prior: $\mathbf{h}(i) \sim \mathcal{CN}(0, \gamma_i)$
  γ_i deterministic, unknown hyperparams
- Goal:
  Given y, X, estimate h & sparsity profile

SBL for Basis Selection

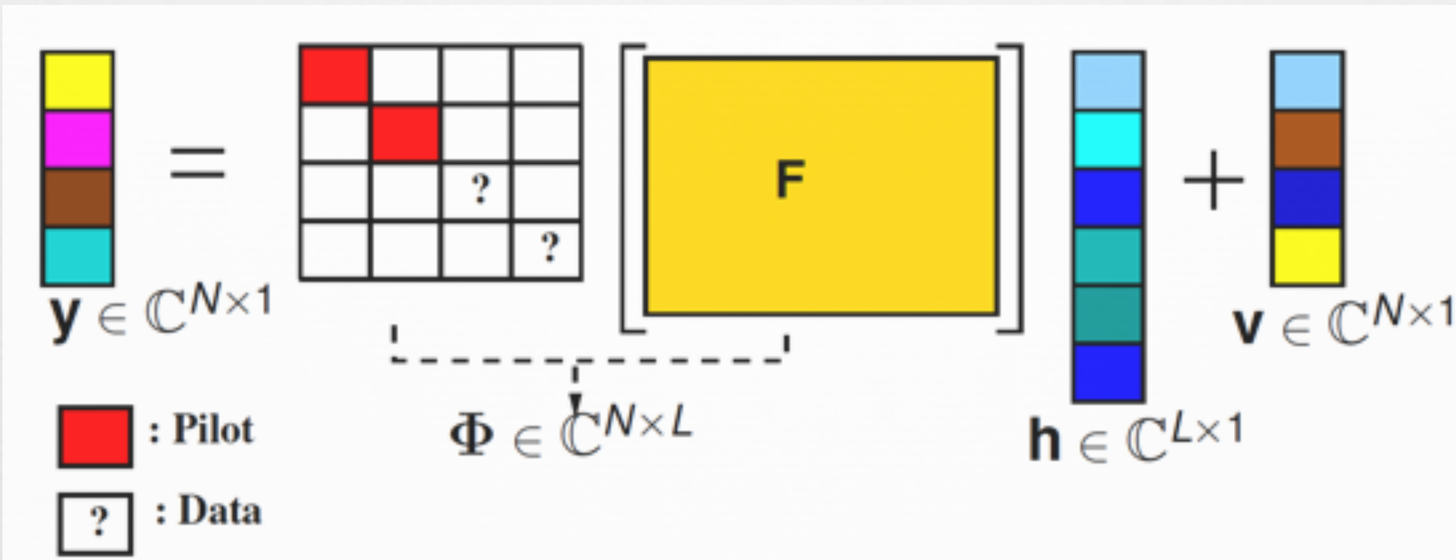
- E-Step: $$Q(\Gamma|\Gamma^{(t)}) = E_{\mathbf{h}|\mathbf{y};\Gamma^{(t)}}(\log p(\mathbf{y},\mathbf{h};\Gamma))$$
$$p\left(\mathbf{h}|\mathbf{y};\Gamma^{(t)}\right) = \mathcal{N}\left(\mu, \Sigma_h\right), \quad \mu \triangleq \sigma^{-2}\Sigma_h \mathbf{A}^H \mathbf{y}$$
$$\Sigma_h \triangleq \left(\sigma^{-2}\mathbf{A}^H\mathbf{A} + \left(\Gamma^{(t)}\right)^{-1}\right)^{-1}, \quad \mathbf{A} \triangleq \mathbf{XF}$$
- M-Step: $$\Gamma^{(t+1)} = \arg\max_{\gamma_i > 0} Q(\Gamma|\Gamma^{(t)})$$
$$\log p(\mathbf{y},\mathbf{h};\Gamma) = \underbrace{\log p(\mathbf{y}|\mathbf{h})}_{\text{not a func. of } \gamma_i} + \underbrace{\log p(\mathbf{h};\Gamma)}_{\text{func. of } \gamma_i}$$

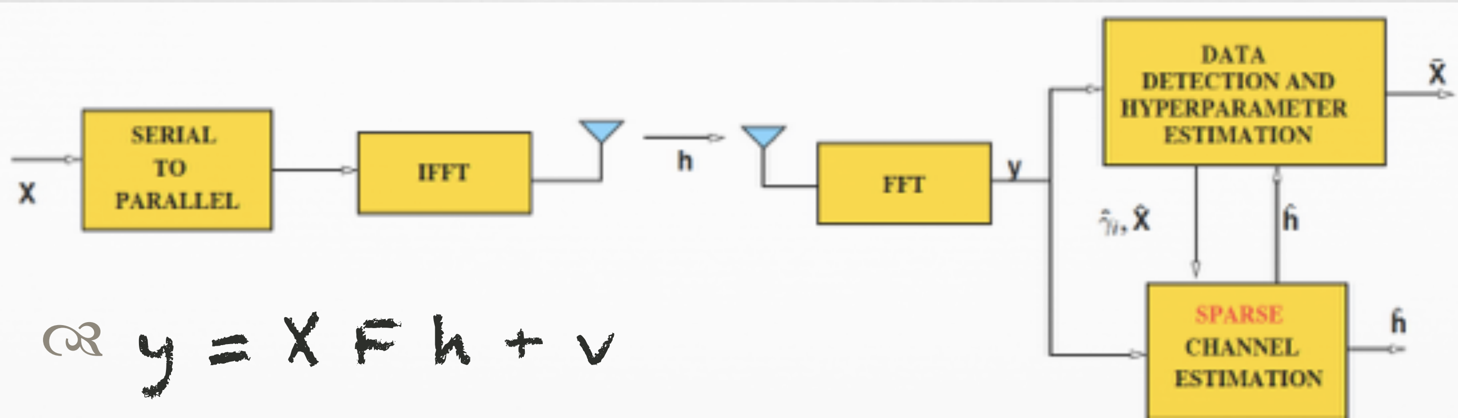
Basis Selection to Channel Estimation



- Upon convergence, many of the γ_i -> 0
  - If $\gamma_i = 0$, then $h(i) = 0$
- Obtain channel estimate as a by-product of the EM iterations

Joint Channel, Support Estmn. & Data Detn.





- $y = X F h + v$

Joint Channel, Support Estmn. & Data Detn.

E–step: $E_{\mathbf{h}/\mathbf{y},\mathbf{X}^{(p)},\Gamma^{(p)}}[\log p(\mathbf{y},\mathbf{h};\mathbf{X},\Gamma)]$

M–step: $\arg\max_{\Gamma,\mathbf{X}}\{\text{E-step}\}$

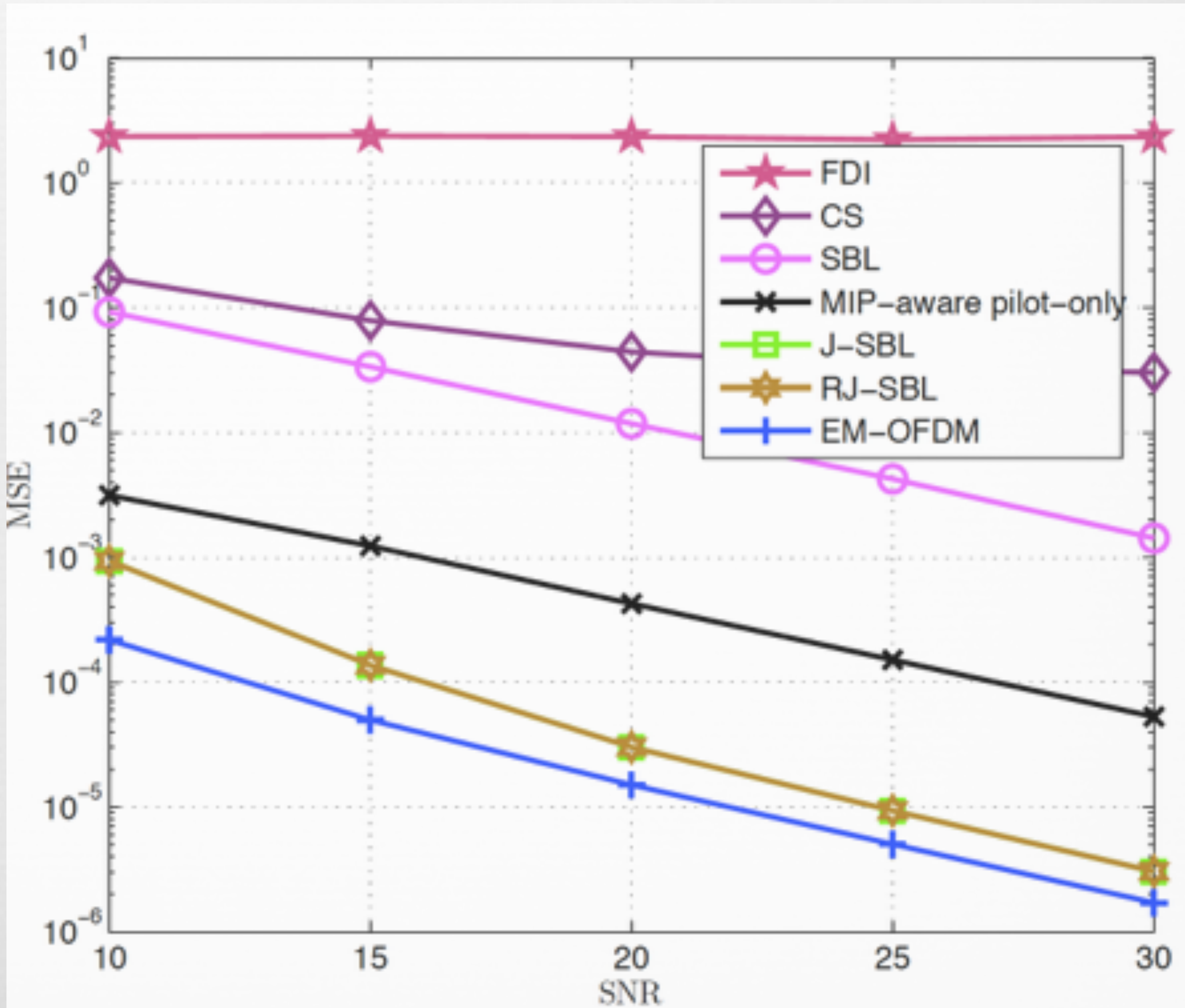
$\arg\max_{\Gamma} E_{\mathbf{h}/\mathbf{y},\mathbf{X}^{(p)},\Gamma^{(p)}}[\log p(\mathbf{h};\Gamma)]$ $\quad$ $\arg\max_{\mathbf{X}} E_{\mathbf{h}/\mathbf{y},\mathbf{X}^{(p)},\Gamma^{(p)}}[\log p(\mathbf{y}/\mathbf{h};\mathbf{X})]$

Γ_{ML} $\quad$ $\mathbf{X}_{ML}$

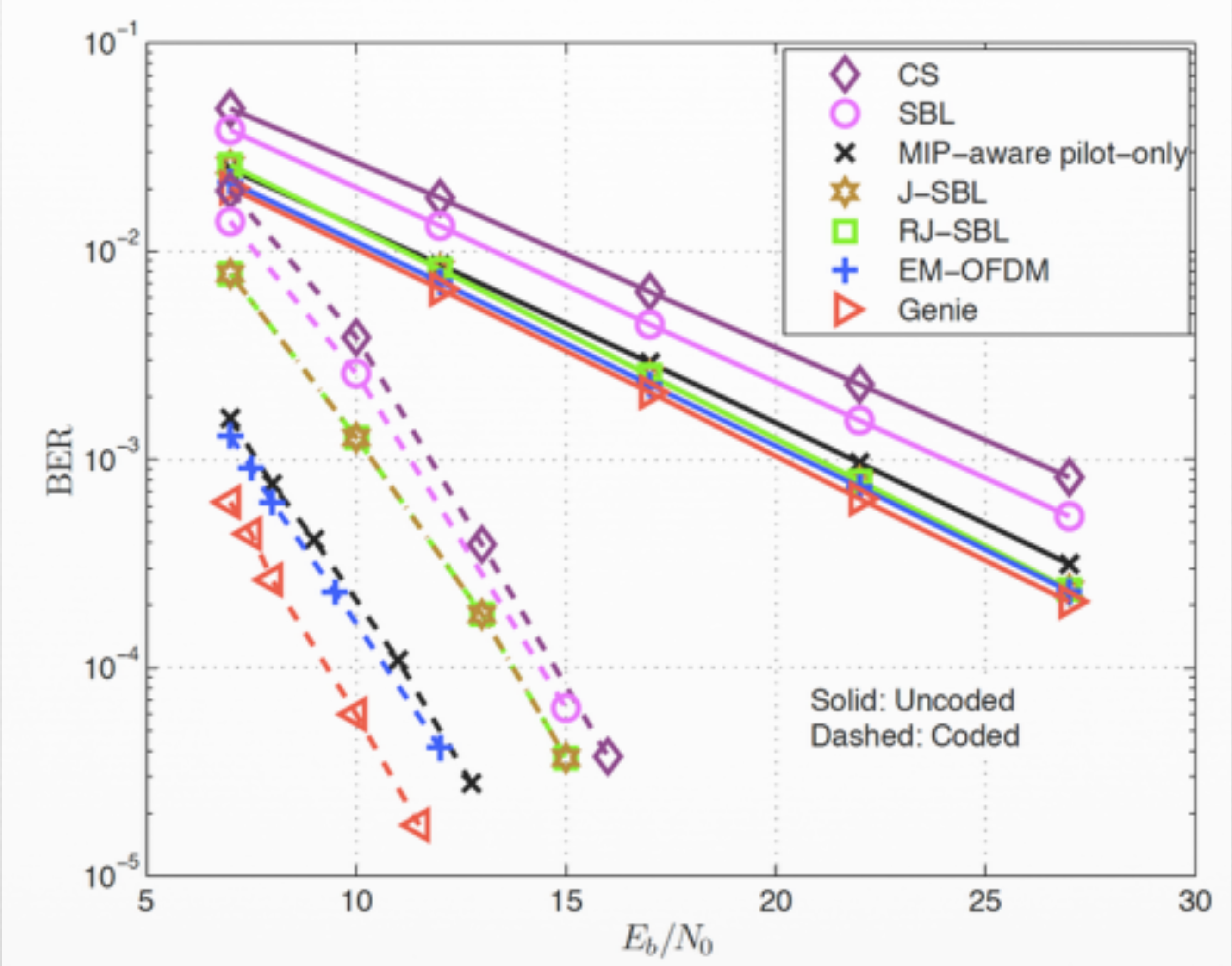
- Get h as a by-product of the E-step

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

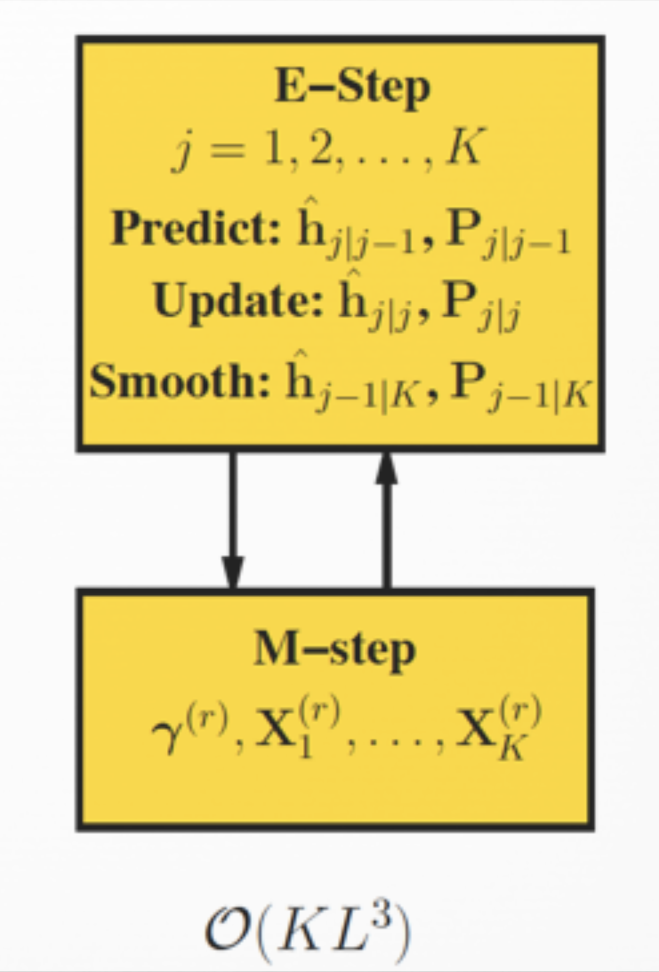


Time-Varying Channels

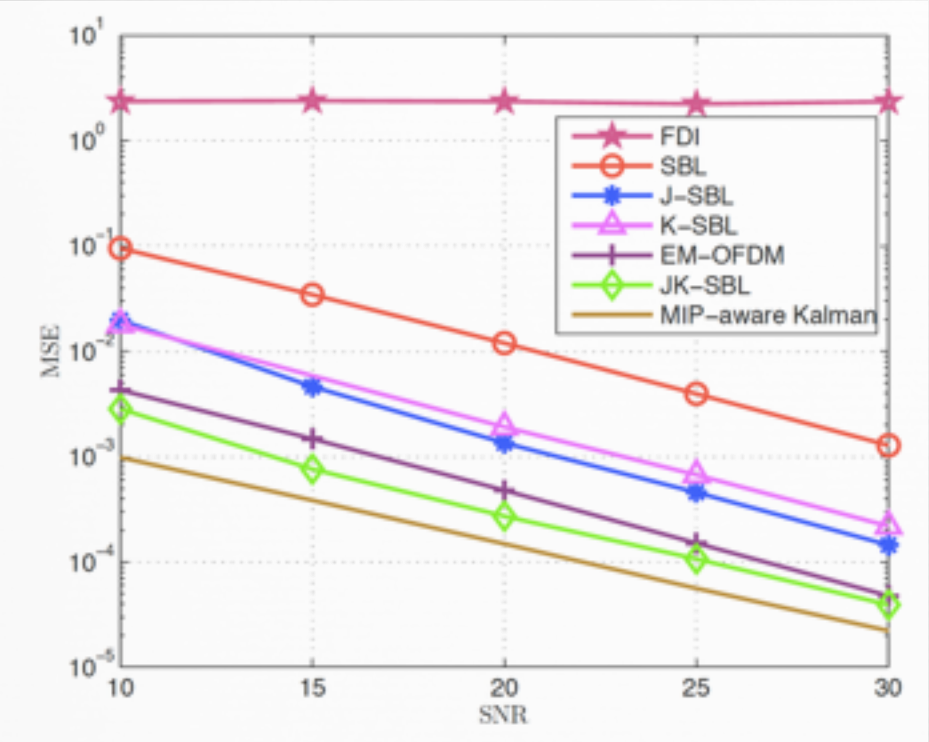
- Channel correlated from symbol-to-symbol
- AR model: $\mathbf{h}_k = \rho \mathbf{h}_{k-1} + \mathbf{u}_k$
- The factor ρ depends on the normalized doppler freq, which in turn depends on the speed of the mobile
- SBL framework can be extended to incorporate the temporal correlation

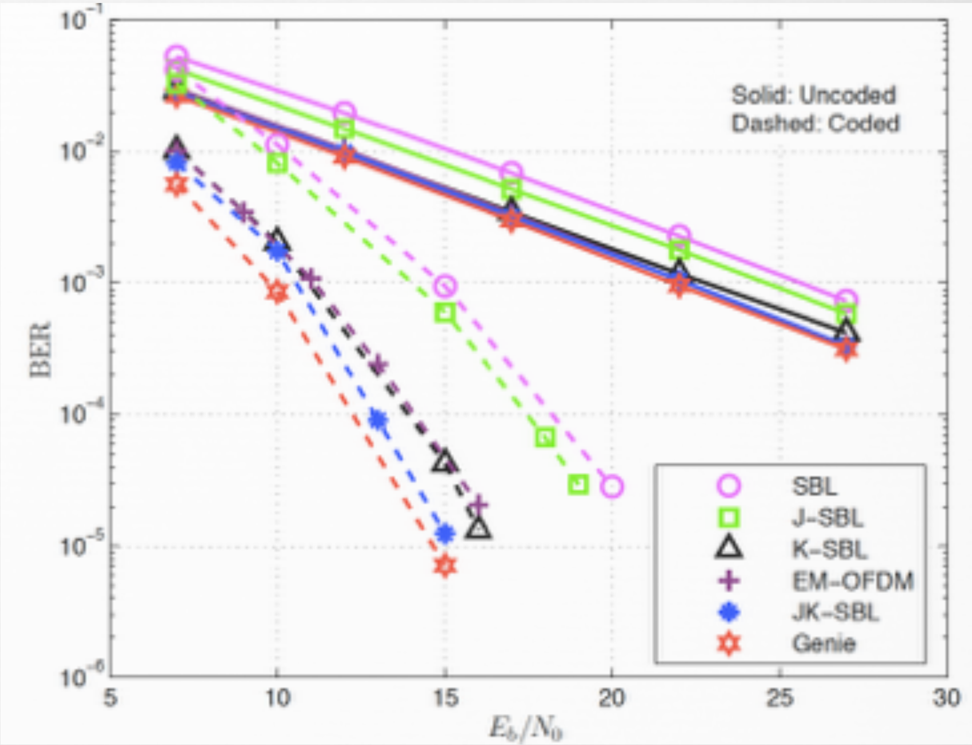
Joint Kalman SBL (JK-SBL)

- 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



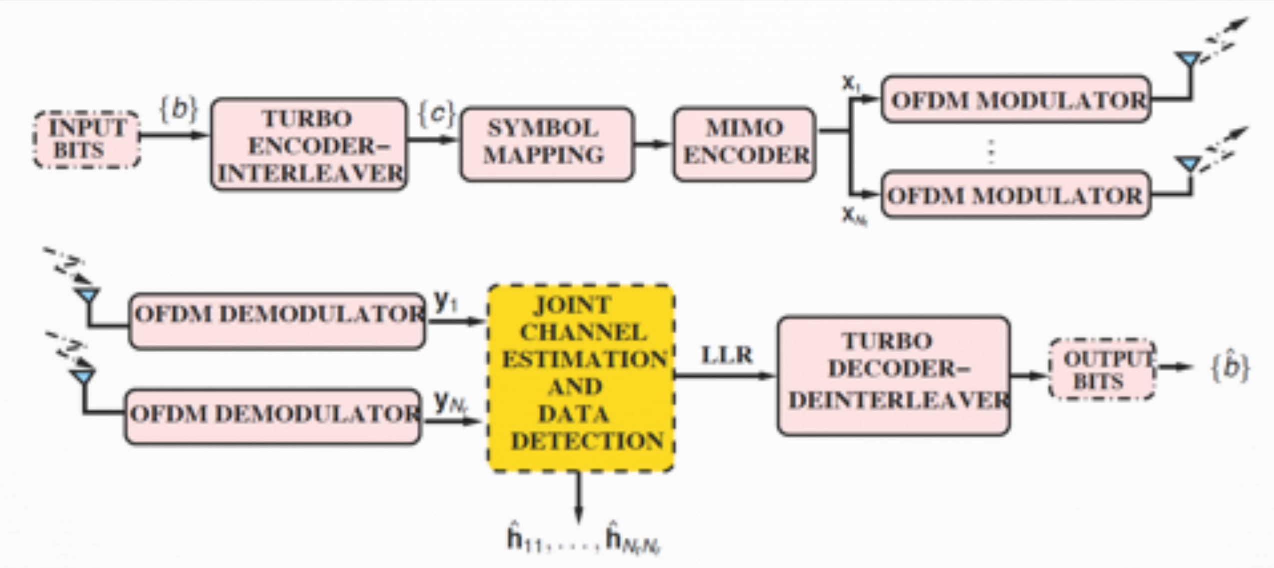
Simulation Result





- $f_d T_s$ = 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 $y_1 \ldots y_{Nr}$

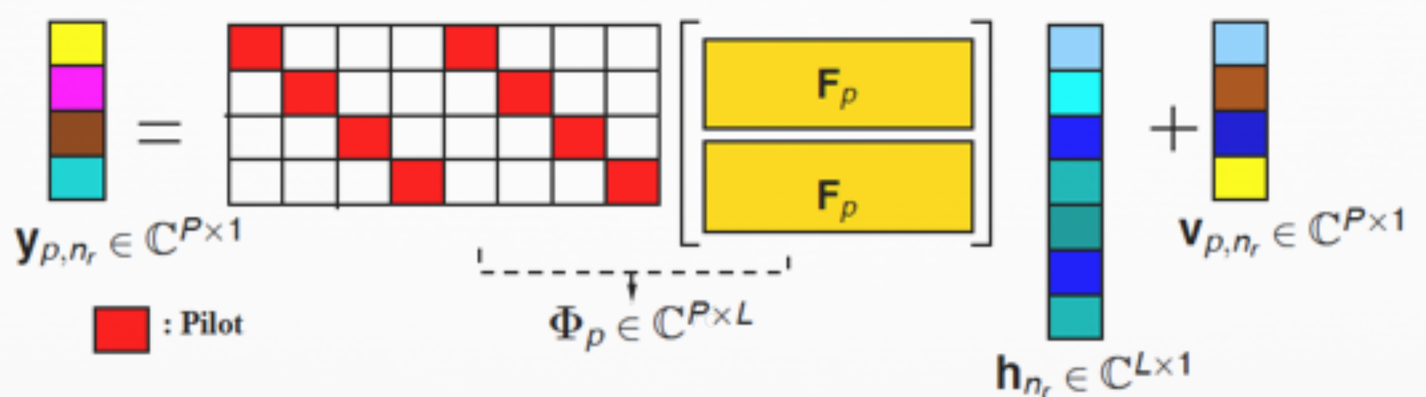
- [Prasad & M., NCC 2014]

MMV Framework

- Measurement model

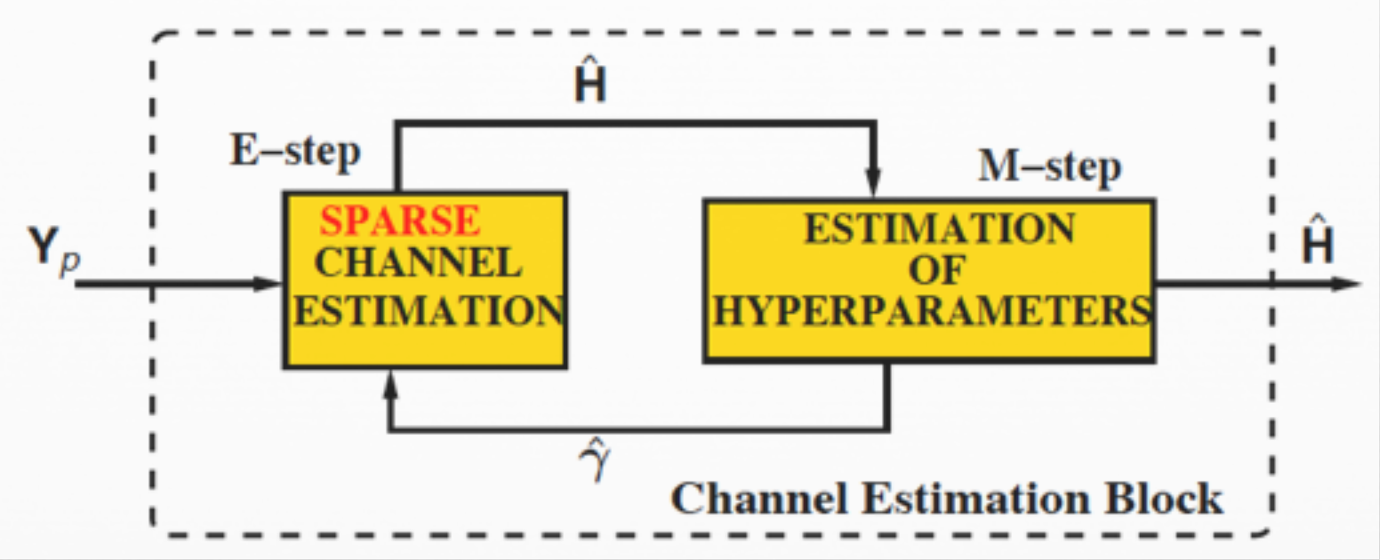
$$\underbrace{[\mathbf{y}_1,\ldots,\mathbf{y}_{N_r}]}_{\mathbf{Y}\in\mathbb{C}^{N\times N_r}} = \underbrace{\mathbf{X}(\mathbf{I}_{N_t}\otimes\mathbf{F})}_{\Phi\in\mathbb{C}^{N\times LN_t}} \underbrace{\begin{pmatrix} \mathbf{h}_{11} & \ldots & \mathbf{h}_{1N_r}\\ \vdots & \vdots & \\ \mathbf{h}_{N_t1}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

- Pilot subcarriers



The M-SBL Algorithm

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



The E and M Steps

- E-Step: Posterior distribution $\mathcal{CN}(\boldsymbol{\mu}_{n_r}, \boldsymbol{\Sigma})$

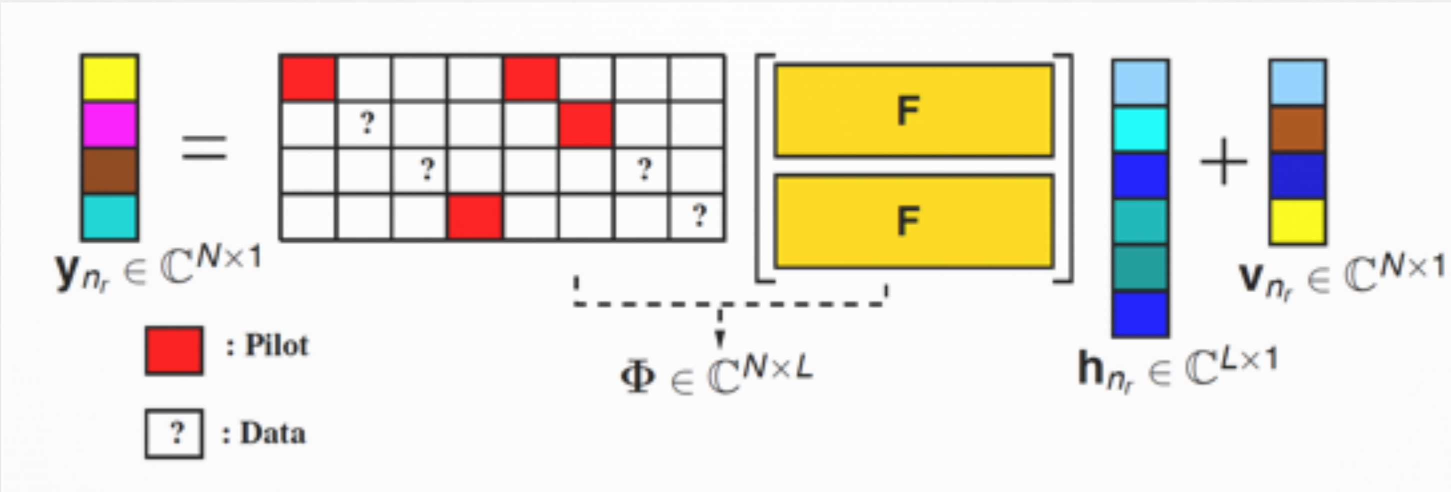
$$\boldsymbol{\mu}_{n_r} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}_p^H \mathbf{y}_{p,n_r} \quad \boldsymbol{\Sigma} = \left(\frac{\boldsymbol{\Phi}_p^H \boldsymbol{\Phi}_p}{\sigma^2} + {\Gamma_b^{(r)}}^{-1} \right)^{-1}$$

- M-Step:

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

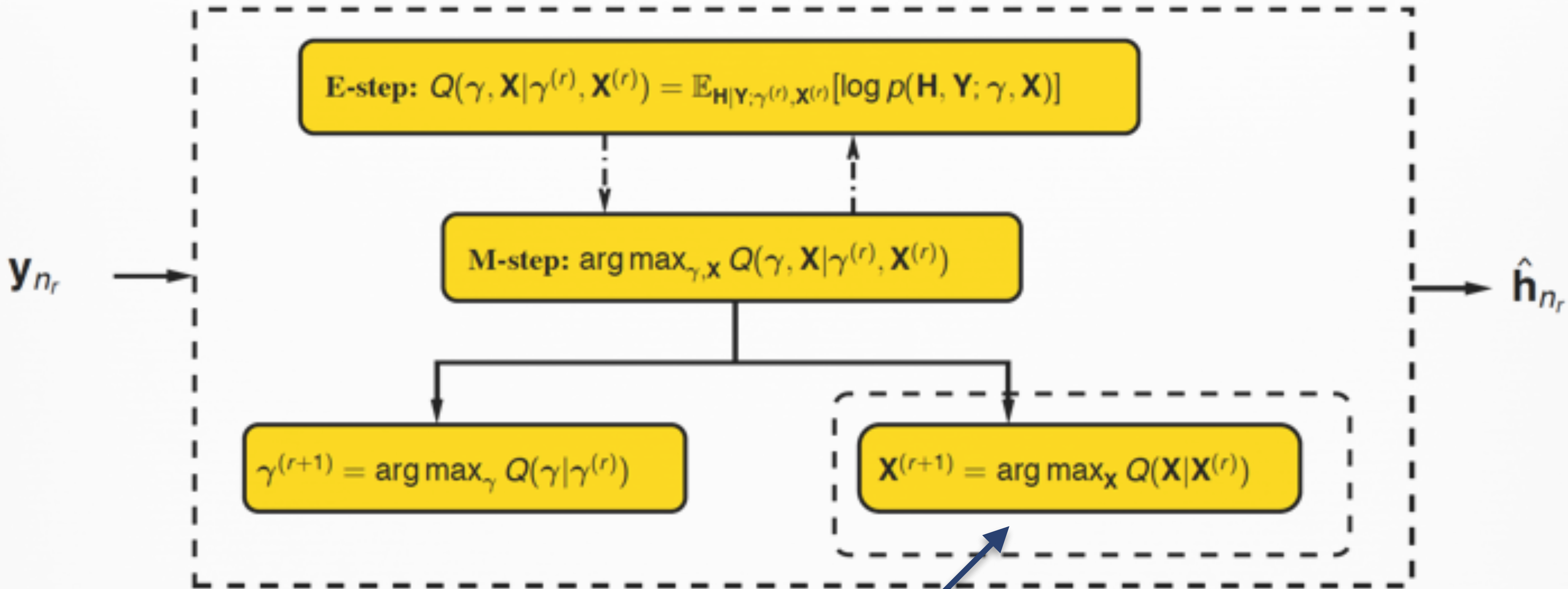
$$\gamma^{(r+1)}(i) = \frac{1}{N_t N_r} \sum_{n_r=1}^{N_r} \sum_{n_t=0}^{N_t-1} \left(\|\mathbf{M}(i + n_t L, n_r)\|_2^2 + \boldsymbol{\Sigma}(i + n_t L, i + n_t L) \right)$$

Joint Channel Estmn. & Data Detection



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

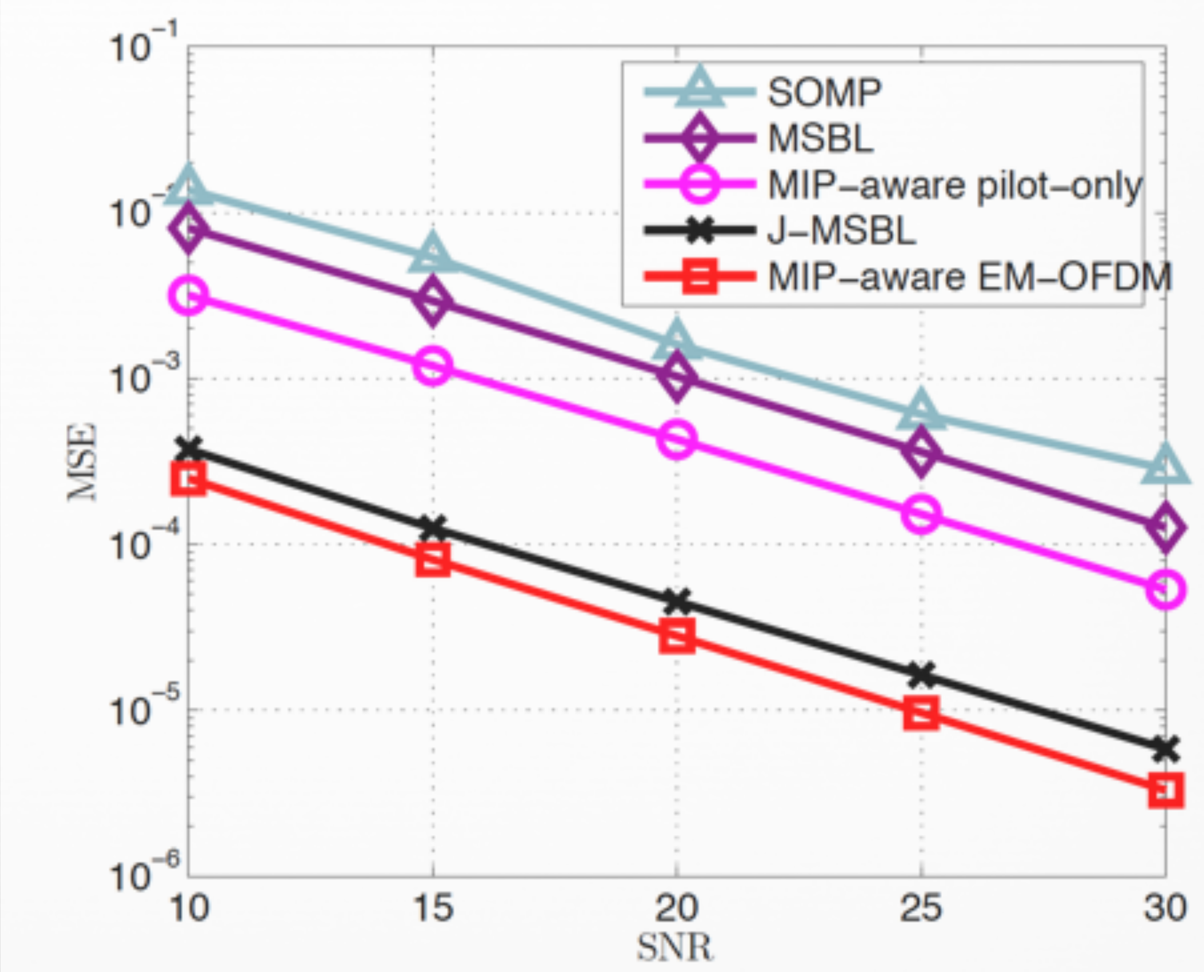
The M Step Splits as Two Separate Problems



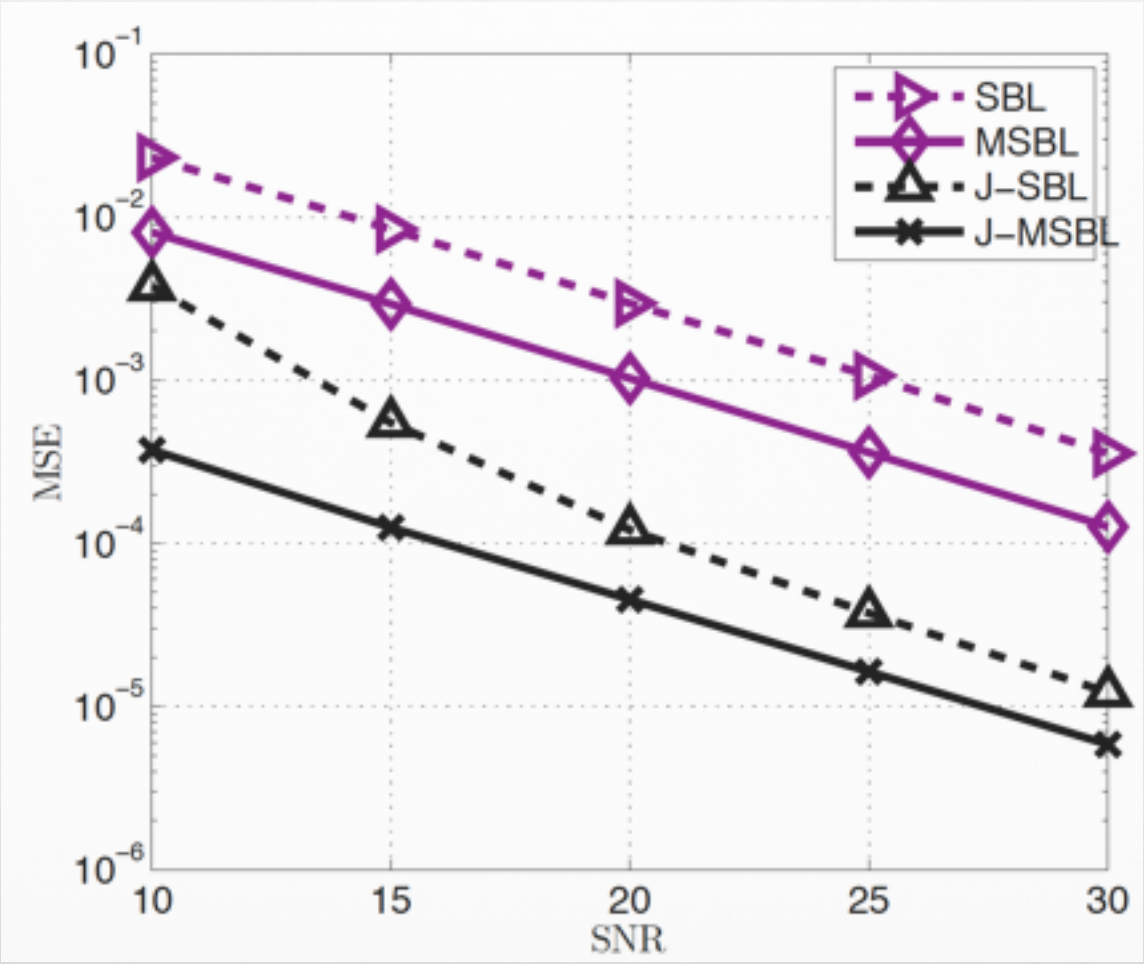
Can use, e.g., sphere decoding to update $\mathbf{X}$

MSE Performance

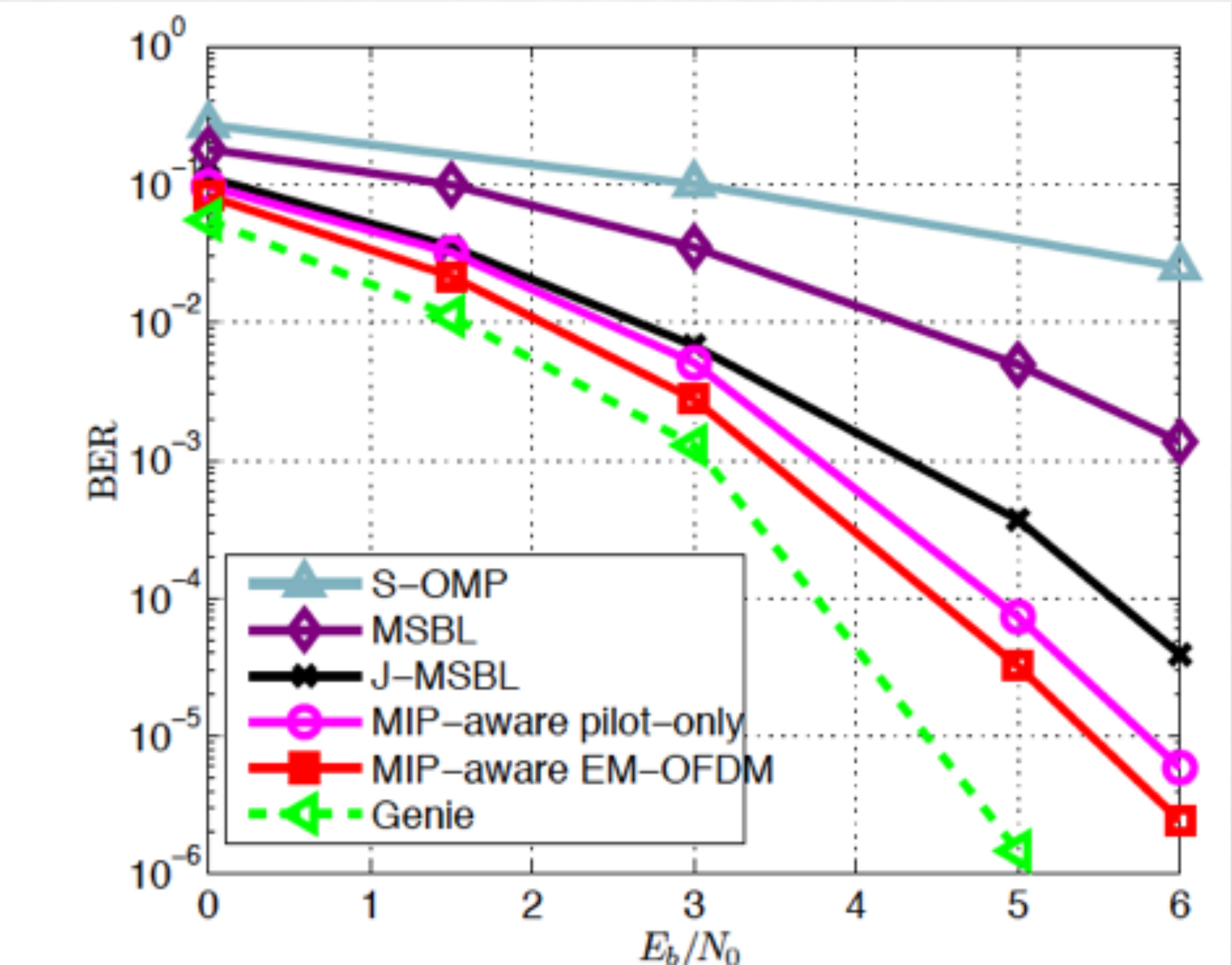
- 2 x 2 MIMO-OFDM System
- 256 subcarriers
- CP length 64
- 44 pilot subcarriers
- PedB PDP
- QPSK constellation



Exploiting Structure Helps!



BER Performance



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!

Extensions

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

Summary

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

References - Our Work

- 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, C. R. Murthy, and B. Rao, **Nested Sparse Bayesian Learning for Block-Sparse Signals with Intra-Block Correlation**, ICASSP 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

Acknowledgements




- Students:
  - Geethu Joseph
  - Saurabh Khanna
  - Ranjitha Prasad
  - Vinuthna Vinjamuri
- Prof. Bhaskar Rao, UC San Diego
- Dr. David Wipf, Microsoft Research Beijing

Thank you!

Contact: cmurthy@ece.iisc.ernet.in