Sparse Bayesian Learning using Underdetermined Linear Measurements with Application to Sparse Wireless Channel Estimation and Data Detection

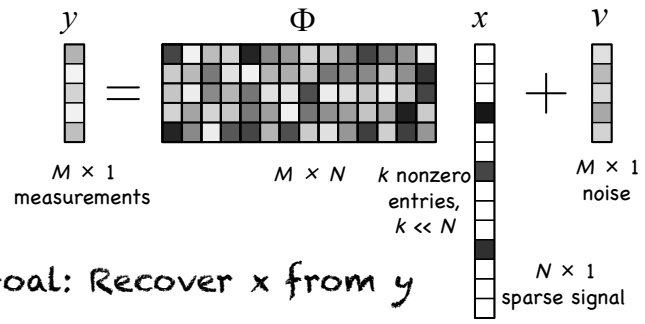
Chandra R. Murthy
Dept. of ECE
Indian Institute of Science

cmurthy@ece.iisc.ernet.in

Joint work with Ranjitha Prasad and Bhaskar Rao

Bayesian Methods for Sparse Signal Recovery

Sparse Signal Recovery



- Goal: Recover x from y
- In general, solution non-unique
- But when x is *sparse*, can find a unique solution, under certain conditions, using *M* << *N* measurements

Applications

- Signal representation (Mallat, Coifman, Wickerhauser, Donoho,...)
- EEG/MEG (Leahy, Gordonitsky, Ioannides,...)
- Functional Approx. (Chen,Nagarajan,Cun,Hassibi,...)
- Spectral estmn (Papoulis, Lee, Cabrera, Parks,...)
- Speech coding (Ozawa, Ono, Kroon, Atal,...)
- MRI (Lustig,...)
- Sparse channel estimation (Fevrier, Greenstein, Proakis, Prasad and Murthy!...)

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
  - Unique soln. with high probability, if *M ≥ k+1* [Bresler; Wakin etc]
  - Combinatorial complexity
  - Not robust to noise

Breakthrough: Just Relax!

- *l-1* minimization instead of *l-0* minimization

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

- Same solution as *l-0* minimization!
  - If the measurement matrix is random
  - Use slightly larger number of measurements
  - Robust to measurement noise $M \approx K \log\left(\frac{N}{K}\right) \ll N$
- Solution methods
  - Basis pursuit [Chen, Donoho, Sanders 1998]
  - Linear programming
  - Augmented Lagrangian method [Bertsekas 03]
- 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
  - CoSAMP [Needell, Tropp]
- 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]

Performance Guarantees

- Mutual coherence. Let $\Phi = [\phi_1, \phi_2, \ldots, \phi_N]$

$$\mu(\Phi) \triangleq \max_{1 \le i,j \le N, i \ne j} \frac{\left|\phi_i^T \phi_j\right|}{\|\phi_i\|_2 \|\phi_j\|_2}$$

- Result (noiseless case): If $\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\Phi)}\right)$
  - [Tropp 04] OMP converges x after k iterations, where k = num. nonzeros in x
  - [Donoho, Elad 03] The sparse vector x_o that generated y is the unique soln to

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

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

Limitations of Greed & Relaxation

- Perf. of BP and OMP depend on the form of the dictionary Φ
  - Poor performance when condns. violated
  - Hard to relate estimation error (e.g., covariance) to Φ
- BP: perf. indep. of nonzero coeffs [Malioutov et al. 2004]
  - Perf. does not improve when situation is favorable
- OMP: perf. highly sensitive to magnitudes of nonzero coeffs
  - Perf. poor with unit magnitudes

Other Limitations of Convex Relaxation

- Scaling/shrinkage:
  - Noiseless: l0 <-> l1 <-> L2. 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 l0-l1 equivalence
- Estimating embedded params (e.g., in Φ)

Don't Relax!

A time and place for nonconvex methods?

Bayesian Methods

- MAP estimation using a sparse linear model
- Can be viewed as a regression problem with sparsity promoting penalties (e.g., lp-norm)
  - l1-min (BP/LASSO) is a special case
- Can overcome some of the previous limitations
- Theory hard to come by, but results promising
- Algorithms:
  - Iterative reweighted l_2
  - EM-based SBL [Tipping, 2001], [Wipf, Rao 2007]
  - [Chartrand and Yin, 2008]
  - AMP [Schniter 2008], [Rangan 2011]

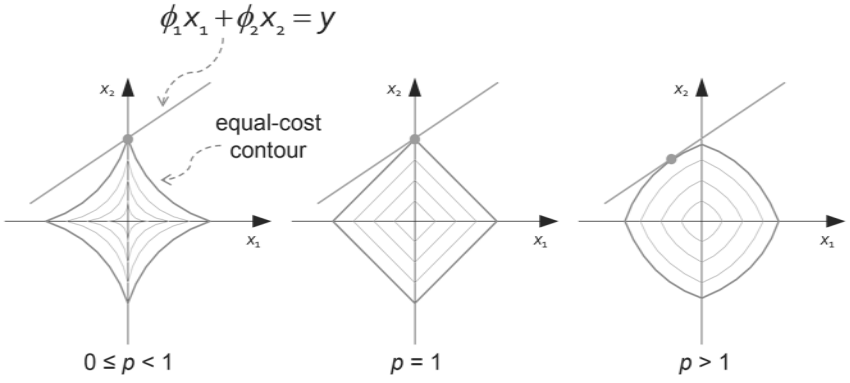
MAP Estimation

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

- For sparse solutions, $g(|x_i|)$ should be a concave, nondecreasing function
  - Example: $g(|x_i|) = |x_i|^p$, $p \leq 1$
  - When $g(|x_i|) = |x_i|$, get Lasso as a special case
- Any local min. of the MAP estmn problem has at most M nonzeros [Rao et al., 03]

Why does it work?

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



[Courtesy: Wipf, Rao]

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!

Other sparsity-inducing Priors?

- Consider a general parameterized prior

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

- If know γ_i, estimating x from y is easy
  - MAP estimate: just the conditional mean
- ML estimate of γ_i from the data: maximize:

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

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 $\hat{\mathbf{x}} = \sigma^{-2}\left(\sigma^{-2}\Phi^T\Phi + \Gamma^{-1}\right)^{-1}\Phi^T\mathbf{y}$
  - Compute $\gamma_i = \hat{x}_i^2$
  - Repeat steps 2 and 3
- Will call this "Approximate MAP" or A-MAP estimation

Maximizing L(Γ): EM

- 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:

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

The SBL Algorithm

- Initialize $\Gamma = I$
- Compute

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

- Update $\Gamma^{(t+1)} = \mathrm{diag}(\mu_i^2 + \Sigma_{ii}^2)$
- Repeat steps 2 and 3

A Variational Interpretation of EM

- 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

The EM Iterations

- **E-Step:** $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = \arg\max_{q_{\mathbf{x}}(\mathbf{x})} \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}); \Gamma^{(t)})$
- **M-Step:** $\Gamma^{(t+1)} = \arg\max_{\Gamma} \mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}); \Gamma)$

- **E-step solution:** $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}; \Gamma^{(t)})$
- **Proof:** $\mathcal{L}(\Gamma) \geq \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}); \Gamma)$

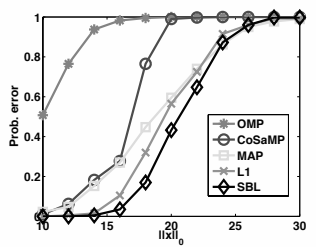
$$\begin{aligned}
\mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}); \Gamma^{(t)}) &= \int p(\mathbf{x}|\mathbf{y}; \Gamma^{(t)}) \log\left(\frac{p(\mathbf{x}, \mathbf{y}; \Gamma^{(t)})}{p(\mathbf{x}|\mathbf{y}; \Gamma^{(t)})}\right) \mathrm{d}\mathbf{x} \\
&= \int p(\mathbf{x}|\mathbf{y}; \Gamma^{(t)}) \log\left(\frac{p(\mathbf{y}; \Gamma^{(t)}) p(\mathbf{x}|\mathbf{y}; \Gamma^{(t)})}{p(\mathbf{x}|\mathbf{y}; \Gamma^{(t)})}\right) \mathrm{d}\mathbf{x} \\
&= \log p(\mathbf{y}; \Gamma^{(t)}) = \mathcal{L}(\Gamma^{(t)})
\end{aligned}$$

Convergence

- Convergence guaranteed from any initialization (property of EM)
- The global min of L occurs at the sparsest solution in the noiseless case [Wipf et al. 04]
- Convergence to a sparse local optimum guaranteed in the noisy case [Wipf et al. 04]

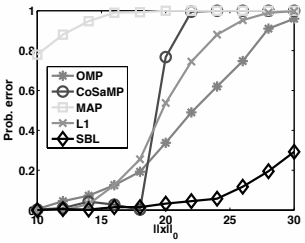
Empirical Example

- 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



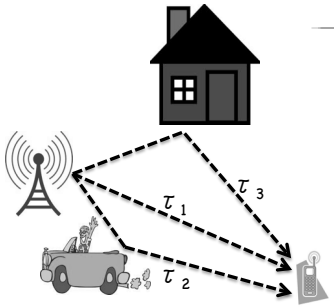
Unit magnitude entries

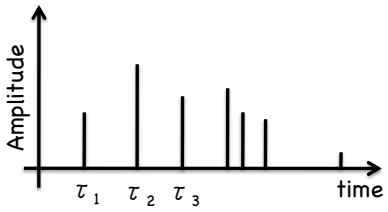
Highly scaled entries



Part 2: Wireless Channel Estimation

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:

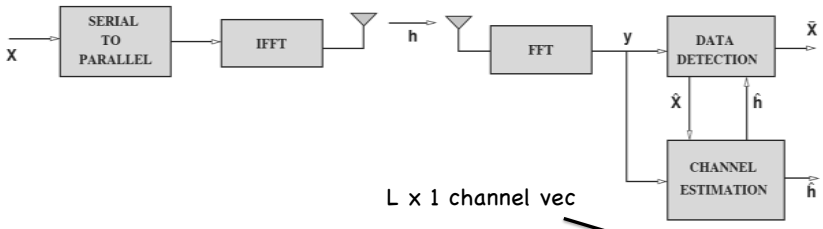
  Channel varies from symbol-to-symbol
  - Want to exploit temporal correlation (group-sparse estimation)

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

Block Fading Channel Estimation

OFDM System Model



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

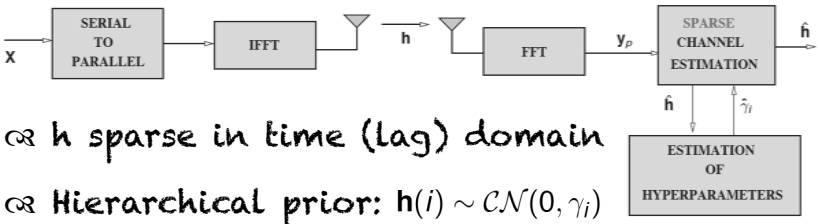
- Goal: Given y, jointly estimate X & h

Support-Aware EM

- E-Step: $Q\left(\mathbf{X}|\mathbf{X}^{(t)}\right) = E_{\mathbf{h}|\mathbf{y},\mathbf{X}^{(t)}}\left(\log p(\mathbf{y},\mathbf{h}|\mathbf{X})\right)$
- M-Step: $\mathbf{X}^{(t+1)} = \arg\max_{\mathbf{X}} Q\left(\mathbf{X}|\mathbf{X}^{(t)}\right)$

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

Sparse Channel Estimation: Known Data



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

SBL for Basis Selection

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

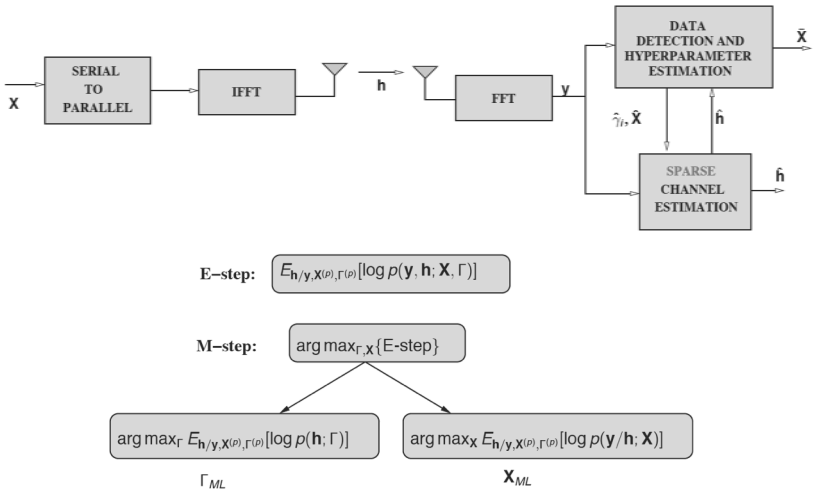
$$p(\mathbf{h}|\mathbf{y}; \Gamma^{(t)}) = \mathcal{N}(\mu, \Sigma_h), \quad \mu = \sigma^{-2}\Sigma_h \mathbf{A}^H \mathbf{y}$$
$$\Sigma_h = \left(\sigma^{-2}\mathbf{A}^H\mathbf{A} + \Gamma^{(p)^{-1}}\right)^{-1}, \mathbf{A} \triangleq \mathbf{XF}$$

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

$$\log p(\mathbf{y}, \mathbf{h}; \Gamma) = \underbrace{\log p(\mathbf{y}|\mathbf{h})}_{\text{not a func. of } \gamma_i} + \underbrace{\log p(\mathbf{h}; \Gamma)}_{\text{func. of } \gamma_i}$$

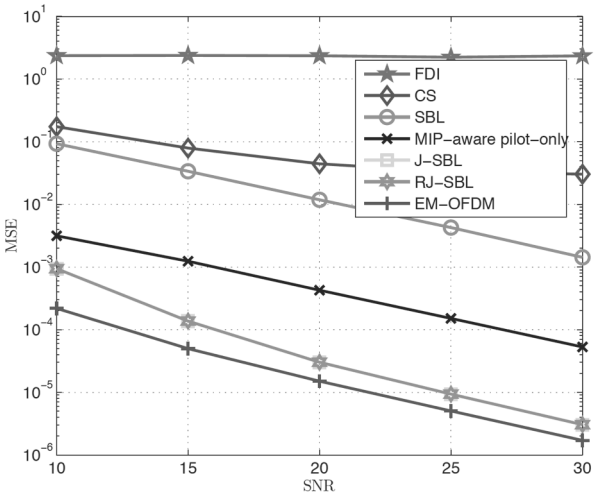
- Upon convergence, many of the γ_i -> 0

Joint Channel, Support Estmn. & Data Detn.

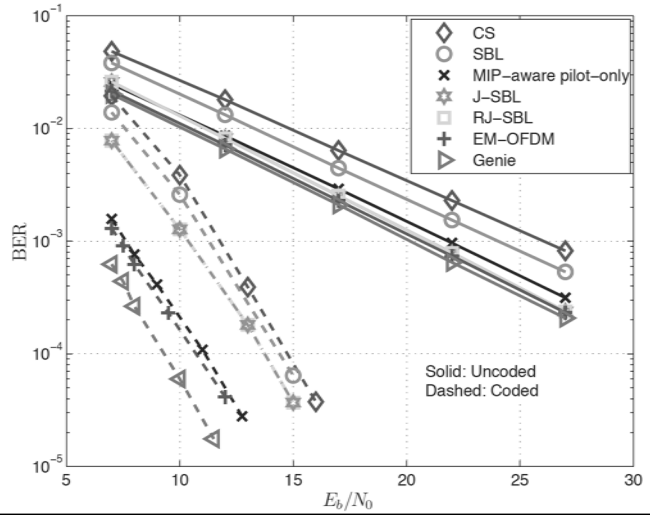


Simulation Result

- 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

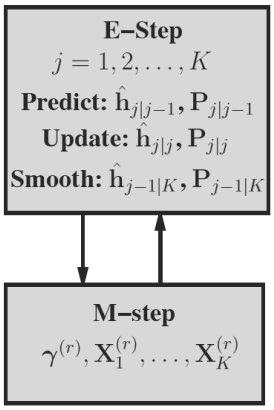


Time-Varying Channels

- 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

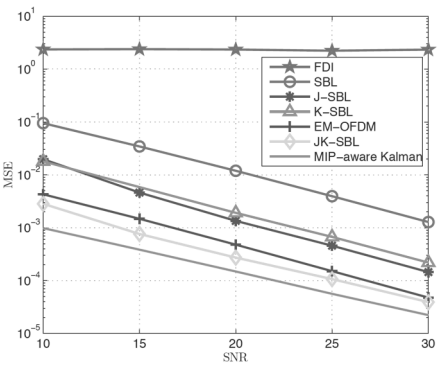
Joint Kalman SBL (JK-KSBL)

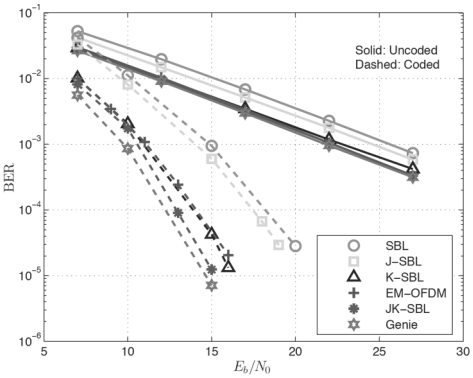
- Complexity $O(KL^3)$ smaller than block-based methods $O(K^3L^3)$ [Zhang et al. 10]
  - (K = num. OFDM symbols used in joint estimation)
- In the block-fading case, get recursive, more computationally efficient versions of our algos



$\mathcal{O}(KL^3)$

Simulation Result





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

Summary

- Used the SBL algorithm for 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
- In practice, algos work even if channel is only approximately sparse

Message

- Bayesian methods can address some limitations in BP/OMP type algos
  - E.g., when Φ has embedded parameters such as unknown data symbols
- Simple updates, promising performance in practical applications
- Many opportunities for new theoretical developments & novel applications
- Did not cover: approximate inference methods (e.g., AMP [Schniter 08])

References

- R. Prasad and M., Bayesian Learning for Joint Sparse OFDM Channel Estimation and Data Detection, Proc. Globecom, 2010
- R. Prasad and M., Cramér-Rao-Type Bounds for Sparse Bayesian Learning, IEEE Trans. Sig. Proc., Mar. 2013
- R. Prasad, M. and B. Rao, Joint Approximately Sparse Channel Estimation and Data Detection in OFDM Systems using Sparse Bayesian Learning, Submitted, IEEE Trans. Sig. Proc., Nov. 2012

Algorithms

- CS methods
  - Yall1: [www.caam.rice.edu/~optimization/L1/YALL1]
  - SpaRSA: [Wright et al., TSP 2009] http://www.lx.it.pt/~mtf/SpaRSA/
  - l1_ls: [Kim et al., JSTSP Dec. 2007]
  - OMP: [Tropp, Gilbert, TIT Dec. 2007]
  - FOCUSS: [Gordonitsky et al., 1997]
  - IRLS: [Chartrand and Yin, 2008]
  - SparseLab: [http://sparselab.stanford.edu/]
- Bayesian methods:
  - SBL: [Tipping, 2001]
  - AMP: [Schniter, 2008], [Rangan, 2011]

Acknowledgements

- Prof. Bhaskar Rao, UC San Diego
- Dr. David Wipf, Microsoft Research Beijing

End!