Sparse Reconstruction via Bayesian Variable Selection and Bayesian Model Averaging

Phil Schniter, Lee Potter, and Subhojit Som

(With support from the Air Force Research Laboratory)

ITA, February 2009

The Sparse Reconstruction Problem:

From the M-length observation

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e},$$

where

$$\begin{array}{ll} \boldsymbol{A} & \text{is known and} \\ \boldsymbol{e} & \text{is AWGN,} \end{array}$$

we desire to estimate the N-length signal $\boldsymbol{x}$, which is

1. *underdetermined*: $\boldsymbol{x}$ has $N > M$ coefficients, and
2. *sparse*: $\boldsymbol{x}$ has $K < M$ non-zero coefficients (K unknown).

The Variable Selection Problem:

If we knew the active-coefficient indices S, we could write

$$\boldsymbol{y} = \boldsymbol{A}_S \boldsymbol{x}_S + \boldsymbol{e},$$

in which case estimation of the nonzero coefficients $\boldsymbol{x}_S$ becomes trivial, e.g.,

$$\begin{aligned}\hat{\boldsymbol{x}}_{\mathsf{LS}|S} &= (\boldsymbol{A}_S^T \boldsymbol{A}_S)^{-1} \boldsymbol{A}_S^T \boldsymbol{y} \\ \hat{\boldsymbol{x}}_{\mathsf{MMSE}|S} &= (\boldsymbol{A}_S^T \boldsymbol{A}_S + \sigma_e^2 \boldsymbol{I})^{-1} \boldsymbol{A}_S^T \boldsymbol{y}\end{aligned}$$

This motivates the problem of *Variable Selection*:

From $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$, estimate the active-coefficient indices S.

Variable Selection is the "difficult" part of sparse reconstruction and a long-standing problem in statistics!

[1] Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, 1976.

Bayesian Variable Selection:

The MAP model estimate is

$$\begin{aligned}\hat{S}_{\text{MAP}} &= \arg\max_S p(S|\boldsymbol{y}) \\ &= \arg\max_S p(\boldsymbol{y}|S)p(S) \\ &= \arg\max_S \int_{\boldsymbol{x}} \underbrace{p(\boldsymbol{y}|S,\boldsymbol{x})}_{\mathcal{N}} p(\boldsymbol{x}|S)d\boldsymbol{x} \cdot p(S)\end{aligned}$$

which then depends entirely on the assumed priors $p(\boldsymbol{x}|S)$ and $p(S)$.

[1] Lempers, *Posterior probabilities of alternative linear models*, Rotterdam: Rotterdam Univ. Press, 1971

[2] Mitchell & Beauchamp, "Bayesian variable selection in linear regression," *J. Amer. Statist. Assoc.*, 1988.

[3] George & McCulloch, "Variable selection via Gibbs sampling," *J. Amer. Statist. Assoc.*, 1993.

[4] Smith & Kohn, "Nonparametric regression using Bayesian variable selection," *J. Econometrics*, 1996.

[5] George & McCulloch, "Approaches for Bayesian variable selection," *Statist. Sinica*, 1997.

[6] George, "The variable selection problem," *J. Amer. Statist. Assoc.*, 2000.

Typical Priors in BVS:

- iid Bernoulli coefficient-activity:

$$p(S) \;=\; \lambda^{|S|}(1-\lambda)^{(N-|S|)} \quad \text{where } \lambda < 0.5 \text{ induces sparsity,}$$

- Gaussian $\boldsymbol{x}_S$:

$$p(\boldsymbol{x}_S|S) \;\sim\; \mathcal{N}(\mu \mathbf{1}_{|S|}, \boldsymbol{R}_S)$$

$$\text{for} \begin{cases} \boldsymbol{R}_S = \sigma_x^2 \boldsymbol{I}_{|S|}, \quad \mu \in \mathbb{R} & \text{“iid”} \\ \boldsymbol{R}_S = \sigma_x^2 (\boldsymbol{A}_S^T \boldsymbol{A}_S)^{-1}, \; \mu = 0 & \text{“Zellner”} \end{cases}$$

where the hyperparameters $\{\mu, \sigma_x^2, \lambda, \sigma_e^2\}$ could be treated as. . .

1. *random*: assign non-informative conjugate priors & integrate out unknowns.
2. *deterministic*: use the EM-algorithm to estimate hyperparameters.

[1] Cui & George, “Empirical Bayes vs. fully Bayes variable selection,” *J. Statist. Planning Infer.*, 2008.

BVS Posteriors:

Fixing $\{\mu, \sigma_x^2, \lambda, \sigma_e^2\}$, we get the model posterior

$$\ln p(S|\boldsymbol{y}) \;=\; -\tfrac{1}{2}\big\|\boldsymbol{y} - \mu \boldsymbol{A}_S \mathbf{1}_{|S|}\big\|^2_{\boldsymbol{\Phi}_S^{-1}} - \tfrac{1}{2}\ln\det(\boldsymbol{\Phi}_S) - |S|\ln(\tfrac{1-\lambda}{\lambda}) + C,$$

where $\boldsymbol{\Phi}_S$ denotes the observation covariance matrix conditioned on model S,

$$\boldsymbol{\Phi}_S \;=\; \begin{cases} \sigma_x^2 \boldsymbol{A}_S \boldsymbol{A}_S^T + \sigma_e^2 \boldsymbol{I}_{|S|} & \text{(iid)} \\ \sigma_x^2 \boldsymbol{A}_S (\boldsymbol{A}_S^T \boldsymbol{A}_S)^{-1} \boldsymbol{A}_S^T + \sigma_e^2 \boldsymbol{I}_{|S|} & \text{(Zellner)} \end{cases}.$$

We also get the S-conditional coefficient posterior

$$p(\boldsymbol{x}_S|\boldsymbol{y}, S) \;\sim\; \mathcal{N}\big(\hat{\boldsymbol{x}}_{\mathsf{MMSE}|S}, \boldsymbol{\Sigma}_S\big)$$

where

$$\begin{aligned} \hat{\boldsymbol{x}}_{\mathsf{MMSE}|S} \;&=\; \mu \mathbf{1}_{|S|} + \boldsymbol{R}_S \boldsymbol{A}_S^T \boldsymbol{\Phi}_S^{-1}(\boldsymbol{y} - \mu \boldsymbol{A}_S \mathbf{1}_{|S|}) \\ \boldsymbol{\Sigma}_S \;&=\; \boldsymbol{R}_S - \boldsymbol{R}_S \boldsymbol{A}_S^T \boldsymbol{\Phi}_S^{-1} \boldsymbol{A}_S \boldsymbol{R}_S. \end{aligned}$$

Connection to AIC/BIC/RIC:

Under the Zellner prior, it can be shown that

$$\hat{S}_{\mathsf{MAP}} \;=\; \arg\min_{S}\Big\{\tfrac{1}{\sigma_e^2}\big\|\boldsymbol{y}-\boldsymbol{A}_S\hat{\boldsymbol{x}}_{\mathsf{LS}|S}\big\|_2^2 + |S|\cdot\ln\Big((1+\tfrac{\sigma_x^2}{\sigma_e^2})(\tfrac{1-\lambda}{\lambda})^2\Big)\tfrac{\sigma_x^2+\sigma_e^2}{\sigma_x^2}\Big\}.$$

Thus there are strong connections between BVS and "information theoretic" model selection methods, e.g.,

$$\begin{aligned}
\hat{S}_{\mathsf{AIC}} &= \arg\min_{S}\Big\{\tfrac{1}{\sigma_e^2}\big\|\boldsymbol{y}-\boldsymbol{A}_S\hat{\boldsymbol{x}}_{\mathsf{LS}|S}\big\|_2^2 + |S|\cdot 2\Big\}\\
\hat{S}_{\mathsf{BIC}} &= \arg\min_{S}\Big\{\tfrac{1}{\sigma_e^2}\big\|\boldsymbol{y}-\boldsymbol{A}_S\hat{\boldsymbol{x}}_{\mathsf{LS}|S}\big\|_2^2 + |S|\cdot \ln M\Big\}\\
\hat{S}_{\mathsf{RIC}} &= \arg\min_{S}\Big\{\tfrac{1}{\sigma_e^2}\big\|\boldsymbol{y}-\boldsymbol{A}_S\hat{\boldsymbol{x}}_{\mathsf{LS}|S}\big\|_2^2 + |S|\cdot 2\ln N\Big\}.
\end{aligned}$$

[1] George & Foster, "Calibration and empirical Bayes variable selection," *Biometrika*, 2000.

Bayesian Model Averaging:

- Previously we motivated Bayesian variable selection, e.g.,

$$\hat{S}_{\mathsf{MAP}} = \arg\max_{S} p(S|\boldsymbol{y})$$

for subsequent use in a *conditional* estimation strategy, e.g.,

$$\hat{\boldsymbol{x}}_{\mathsf{MMSE}|\hat{S}_{\mathsf{MAP}}} = \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y}, \hat{S}_{\mathsf{MAP}}\}.$$

- But having access to the "soft information" $\{p(S|\boldsymbol{y})\}$ allows more sophisticated *unconditional* estimates, e.g.,

$$\hat{\boldsymbol{x}}_{\mathsf{MMSE}} = \sum_{\hat{S}} \hat{\boldsymbol{x}}_{\mathsf{MMSE}|\hat{S}}\; p(\hat{S}|\boldsymbol{y})$$

that are well approximated by summing over the few most probable $\hat{S}$.

This approach is known as *Bayesian Model Averaging*.

[1] Leamer, *Specification Searches*, New York: Wiley 1978.

[2] Raftery, Madigan, & Hoeting, "Bayesian model averaging for linear regression models," *J. Amer. Statist. Assoc.*, 1997.

[3] Clyde and George, "Model Uncertainty," *Statist. Sci.*, 2004.

BMA Implementation:

- The statistical literature focuses on random search based on Gibbs Sampling or Markov Chain Monte Carlo.
- We instead proposed a fast $\mathcal{O}(NM)$ update/downdate which can be used in a (non-exhaustive) tree search:
  - iid Gaussian $\boldsymbol{x}_S$: "Fast Bayesian Matching Pursuit" [1]
  - Zellner Gaussian $\boldsymbol{x}_S$: "Optimized OMP" [2] plus penalty term $|\hat{S}|\ln(\frac{1-\lambda}{\lambda})$

  with a total complexity of $\mathcal{O}(MNK)$.
- The 4 hyperparameters $\{\mu, \sigma_x^2, \sigma_e^2, \lambda\}$ can be determined using the EM algorithm, or a simplification thereof [3].

[1] Schniter, Potter, and Ziniel, "Fast Bayesian matching pursuit," *ITA*, 2008.

[2] Rebollo-Neira and Lowe, "Optimized orthogonal matching pursuit," *IEEE Sig. Proc. Letters*, 2002.

[3] Schniter, Potter, and Ziniel, "Fast Bayesian matching pursuit: Model uncertainty and parameter estimation for sparse linear models," Preprint, 2008.

Tipping's Relevance Vector Machine (RVM):

The RVM is another approach to Bayesian sparse reconstruction:

- For coefficient activity, RVM uses continuous "precisions" $\boldsymbol{\alpha} \in (\mathbb{R}^+)^N$:

$$\boldsymbol{x}|\boldsymbol{\alpha} \sim \text{independent } \mathcal{N}(0, \alpha_n^{-1}) \quad \text{and} \quad \boldsymbol{\alpha} \sim \text{iid } \Gamma(0,0)$$
$$\boldsymbol{e}|\beta \sim \mathcal{N}(\mathbf{0}, \beta^{-1}\boldsymbol{I}) \quad \text{and} \quad \beta \sim \Gamma(0,0)$$

- The RVM's gamma hyperpriors lead to the convenient posterior

$$p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\alpha},\beta) \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \quad \text{for} \quad \begin{cases} \boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{y} \\ \boldsymbol{\Sigma} = \big(\beta\boldsymbol{A}^T\boldsymbol{A} + \mathcal{D}(\boldsymbol{\alpha})\big)^{-1} \end{cases}$$

  and thus $\hat{\boldsymbol{x}}_{\text{MMSE}} = \boldsymbol{\mu}$.

- The EM algorithm can be used to estimate $\{\boldsymbol{\alpha},\beta\}$ jointly with $\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$. Can implement with an $\mathcal{O}(NK^2)$ recursion after an $O(N^2M)$ initialization.

[1] Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Res.*, 2001.

[2] Tipping & Faul, "Fast likelihood marginal maximization for sparse Bayesian models," *IWAIS*, 2003.

[3] Wipf and Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Processing*, 2004.

BMA versus RVM:

- Both are Bayesian approaches to sparse parameter estimation.
- For coefficient activity, RVM uses the continuous parameterization $\boldsymbol{\alpha}$, while BMA uses the discrete parameterization S.
- Implementations require roughly the same complexity.
- Upon termination, the RVM posterior is Gaussian

$$p(\boldsymbol{x}|\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

whereas the BMA posterior is a Gaussian mixture:

$$p(\boldsymbol{x}|\boldsymbol{y}) \sim \sum_{\hat{S}} \mathcal{N}\big(\hat{\boldsymbol{x}}_{\mathsf{MMSE}|\hat{S}}, \boldsymbol{\Sigma}_{\hat{S}}\big)\ p(\hat{S}|\boldsymbol{y}).$$

Thus, the BMA posterior can be more informative.

Numerical Experiments — "Compressible" Signal:

Setup:

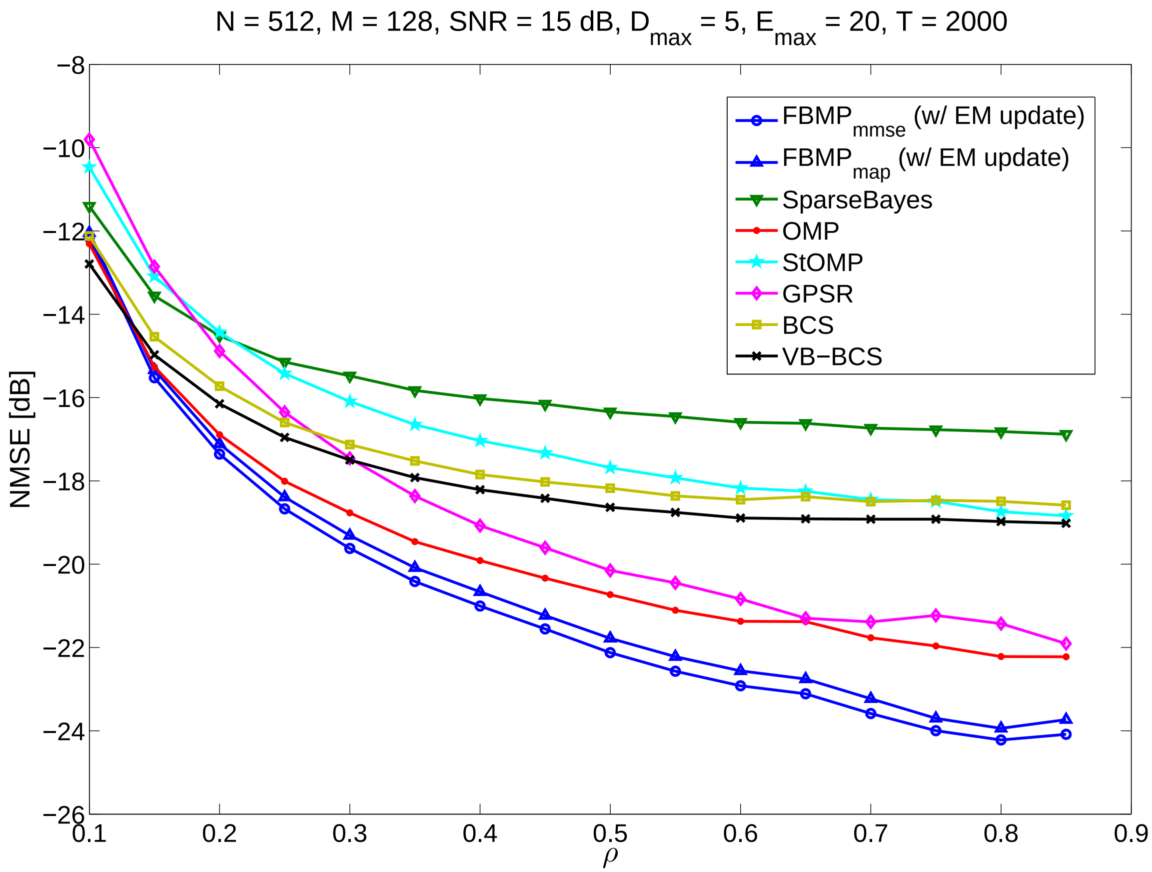
$$\begin{aligned} N &= 512 \\ M &= 128 \\ \boldsymbol{A} &: \text{ i.i.d. } \mathcal{N}(0,1) \quad \text{with columns scaled to unit norm} \\ \boldsymbol{x} &: \text{ sorted } x_n = e^{-\rho n} \text{ for decay rate } \rho \in (0,1) \\ \text{SNR} &= 15\text{dB} \end{aligned}$$

Algorithms:

OMP	– Tropp & Gilbert
StOMP	– Donoho, Tsaig, Drori & Starck
GPSR-Basic	– Figueiredo, Nowak & Wright ($\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \tau\|\boldsymbol{x}\|_1$)
SparseBayes	– Wipf & Rao (RVM)
BCS	– Ji & Carin (RVM)
FBMP	– Schniter, Potter & Ziniel (BMA)

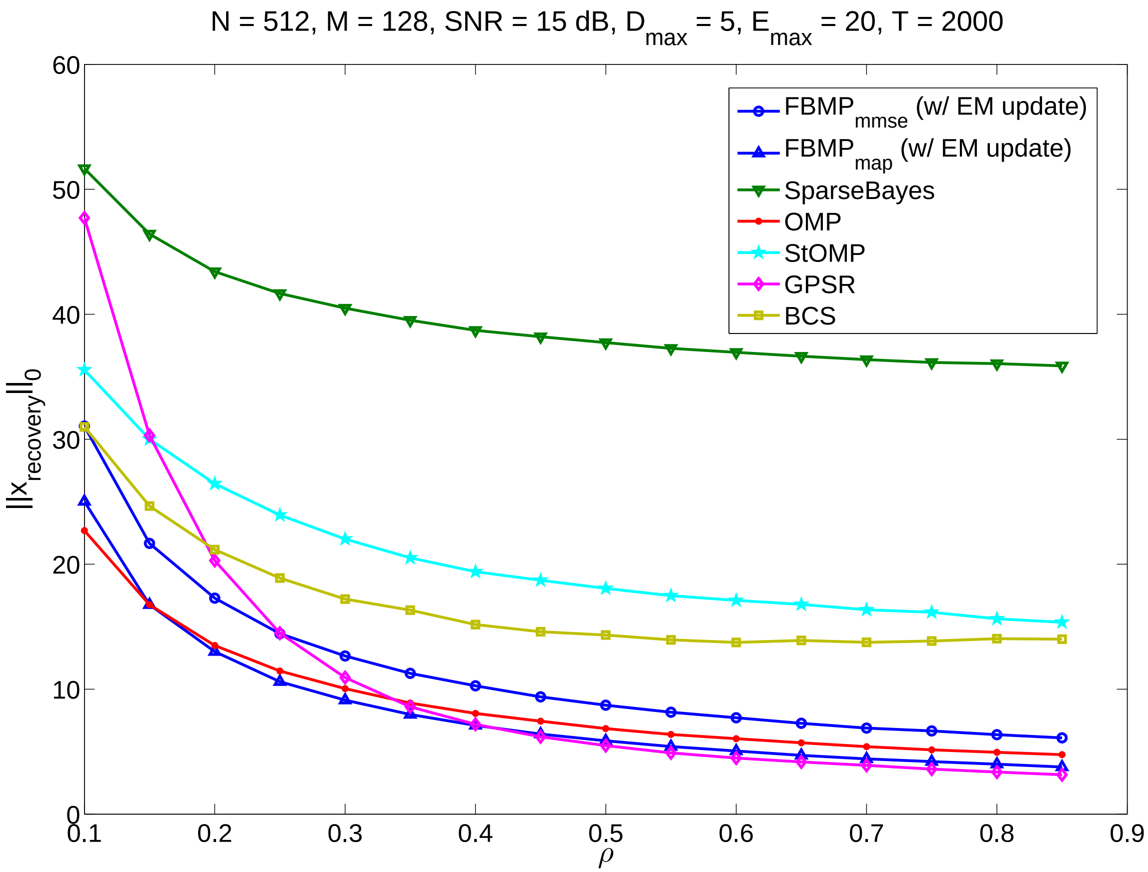
Performance: $\text{NMSE} \triangleq \text{Avg}\left\{\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\right\}$ over 2500 random trials.

NMSE versus decay rate ρ:



FBMP outperformed GPSR and OMP by 2 dB and others by much more.
Note: The signal priors favor GPSR.

Sparsity of estimate versus decay rate ρ:



The estimates returned by FBMP are among the sparsest.

Performance Guarantees for MAP Variable Selection:

Assuming that $\boldsymbol{A}$ that satisfies a Restricted Isometry Property (RIP), we've recently shown that the following properties hold *with high probability* for reasonably small constants K_1, K_2, K_3, K_4:

1. The energy of the missed signal coefficients is upper bounded by $K_1 M \sigma_e^2$.
2. No active coefficients are missed when $|\mu| > 4\sigma_1 + K_2\sqrt{M}\sigma_e^2$.
3. No coefficients are falsely detected when $|\mu| > K_3\sqrt{M}\sigma_1 + K_4\sqrt{M}\sigma_e^2$.

Pair-Wise Error Probability Analysis:

- We've recently shown that the probability of BVS-MAP incorrectly choosing $\hat{S}$ over correct S, i.e.,

$$P_{\hat{S}|S} \;=\; \Pr\left\{p(\hat{S}|\boldsymbol{y}) > p(S|\boldsymbol{y}) \;\middle|\; S\right\}$$

  has the following upper bound (in the Zellner case):

$$P_{\hat{S}|S} \;\leq\; \Pr\left\{\tfrac{\sigma_x^2}{\sigma_x^2+\sigma_e^2} Z_{\mathsf{fa}} - \tfrac{\sigma_x^2}{\sigma_e^2}(1-\epsilon) Z_{\mathsf{m}} > \tau\right\}$$

  where

$$\begin{aligned} \tau &= \big(|\hat{S}| - |S|\big) \ln\!\big((1+\tfrac{\sigma_x^2}{\sigma_e^2})\big(\tfrac{1-\lambda}{\lambda}\big)^2\big) \\ \epsilon &= \text{RIP constant} \\ Z_{\mathsf{fa}} &\sim \chi^2_{|\hat{S}_{\text{false alarm}}|} \\ Z_{\mathsf{m}} &\sim \chi^2_{|\hat{S}_{\text{miss}}|} \end{aligned}$$

- A Chernoff bound or saddle-point approximation can then be applied to characterize error probability.

Conclusion:

- Bayesian variable selection (BVS) and Bayesian model averaging (BMA) are well established statistical methods for sparse reconstruction, typically implemented via Gibbs sampling or MCMC.
- There are close connections between BVS and AIC/BIC/RIC.
- There are similarities & differences between BMA and Tipping's RVM.
- We proposed novel BVS/BMA implementations based on tree-search that lead to fast "matching pursuit"-like algorithms.
- Numerical experiments suggest that BMA yields excellent NMSE relative to other state-of-the-art algorithms.
- We presented preliminary results on BVS performance guarantees and error rate analyses based on the restricted isometry property (RIP).