Statistical Signal Processing

Lecture 12

Statistical Signal Modeling, Learning and Processing

chapter 6: Learning Topics

- Expectation-Maximization (EM) algorithm
- compressed sensing and sparsity
- compressed sensing algorithms: LASSO etc
- Sparse Bayesian Learning (SBL), Empirical Bayes
- Variational Bayes (VB)

Outline

Main Messages

- There are many Bayesian estimation problems, many of which are LMMSE (Wiener, Kalman), which contain hyperparameters to be tuned, using various approaches.
- Information combining: from weighted least-squares to message passing in a more general overall Bayesian formulation (e.g. cooperative location estimation)
- Empirical Bayes (EB) as principled framework for bias-variance trade-off
- but not necessarily using empirical Bayes for hyperparameter estimation: SURE, Cross Validation
- compressive sensing, sparse models, generalization of model order selection to support region, model complexity and structure
- Sparse Bayesian Learning (SBL) is one EB instance, allowing to exploit (approximate) sparsity for compressed sensing
  - can be extended to time-varying scenarios with sparse variations also
  - can be extended to dictionary learning, in particular with Kronecker structured dictionaries
- message passing (approximate iterative) inference techniques: easy to get the mean (estimate) correct but more difficult to get correct posterior variances

Kalman Filter

Linear state-space model:

state update equation:

$$\boldsymbol{x}_{k+1} = \boldsymbol{F}_k(\boldsymbol{\theta})\, \boldsymbol{x}_k + \boldsymbol{G}_k(\boldsymbol{\theta})\, \boldsymbol{w}_k$$

measurement equation:

$$\boldsymbol{y}_k = \mathbf{H}_k(\boldsymbol{\theta})\, \boldsymbol{x}_k + \boldsymbol{v}_k$$

for $k = 1, 2, \ldots$, with uncorrelated

- initial state $\boldsymbol{x}_0 \sim \mathcal{N}(\widehat{\boldsymbol{x}}_0, \boldsymbol{P}_0)$,
- measurement noise $\boldsymbol{v}_k \sim \mathcal{N}(0, \boldsymbol{R}_k(\boldsymbol{\theta}))$,
- state noise $\boldsymbol{w}_k \sim \mathcal{N}(0, \boldsymbol{Q}_k(\boldsymbol{\theta}))$.

State model known up to some parameters $\boldsymbol{\theta}$.

Often $\boldsymbol{F}_k(\boldsymbol{\theta})$, $\boldsymbol{G}_k(\boldsymbol{\theta})$, $\mathbf{H}_k(\boldsymbol{\theta})$ linear in $\boldsymbol{\theta}$: bilinear case.

Numerous Applications

- LMMSE wireless channel estimation:

  x_k = FIR filter response, θ: Power Delay Profile, AR(1) dynamics in e.g. diagonal $\boldsymbol{F}$ and $\boldsymbol{Q}$

- Bayesian adaptive filtering (BAF):
  analogous to LMMSE channel estimation, except measurement equation: only one 1D measurement is available per instance. An extremely simplified form of BAF is the so-called Proportionate LMS (P-LMS) algorithm.

- Position tracking (GPS):

$$\mathbf{x}_{t+1} = \begin{bmatrix} 1 & \Delta t & \frac{1}{2}\Delta t^2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{bmatrix} \cdot \mathbf{x}_t = \begin{bmatrix} x_t + \Delta t \cdot v_t + \frac{1}{2}\Delta t^2 a \\ v_t + \Delta t \cdot a \\ a \end{bmatrix}$$

  θ: acceleration model parameters (e.g. white noise, AR(1))

- Blind Audio Source Separation (BASS):

  x_k = source signals,
  θ: (short+long term) AR parameters, reverb filters

Static LMMSE (Wiener) Applications

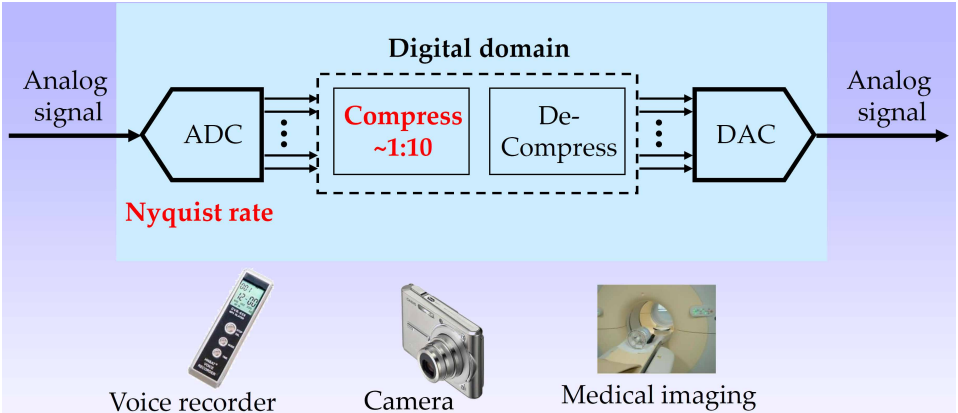
- Direction of Arrival (DoA) estimation: $\boldsymbol{x}$ = decorrelated sources, apart from the DoA parameters there could also be antenna array calibration parameters or source and noise covariance parameters.
- Blind and semi-blind channel estimation. In the techniques that exploit the (white) second-order statistics of $\boldsymbol{x}$, (the unknown part of) $\boldsymbol{x}$ gets modeled as Gaussian. Numerous variations: single-carrier, OFDM and CDMA versions, single- and multi-user, single- and multi-stream, with black box FIR channel models or propagation based parameterized channel models.
  Image Deblurring, NMRI Imaging
- LMMSE receiver (Rx) design: $\boldsymbol{x}$ = Tx symbol sequence to be recovered on the basis of Rx signal, in single- or multi-user settings and other variations as in the channel estimation case. The crosscorrelation between Tx and Rx signals depends on the channel response, which is part of the parameters. The Rx signal covariance matrix on the other hand can be modeled in various ways, non-parametric or parametric. Account for the channel estimation error in the LMMSE Rx design. Another approach: consider the LMMSE filter directly as the parameters.
  LMMSE Tx design, partial CSIR/CSIT.

Adaptive Kalman Filter solutions

- Extended Kalman Filter (EKF)
- other generic nonlinear Kalman Filter extensions:
  Unscented Kalman Filter (UKF), Cubature Kalman Filter (CKF), Gaussian Sum Filter, Particle Filter (PF)
- Recursive Prediction Error Method (RPEM) Kalman Filter
- Second-Order Extended Kalman Filter (SOEKF)
- Expectation-Maximization (EM)/Variational Bayes (VB) Kalman Filter

Outline

Compressed Sensing



Compressed Sensing

In some applications, measurements are costly:

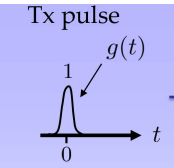
- Magnetic resonance imaging:
  - scan time $\approx$ 30 minutes
  - scan time proportional to # samples taken

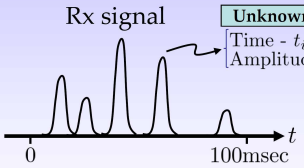
- Imaging outside visible spectrum:
  - CMOS doesn't work
  - high cost per pixel

- Wireless communication:
  - pilots inserted to measure channel
  - more pilots means less payload

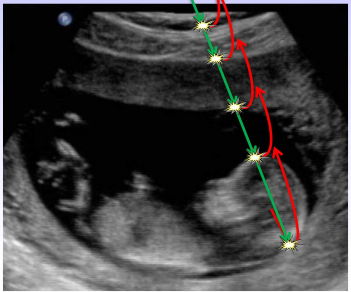
Ultra-Sound Imaging

- High sampling rates
- High digital processing rates





L echoes



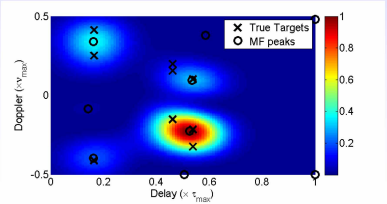
- Echoes result from scattering in the tissue
- The image is formed by identifying the scatterers

Radar

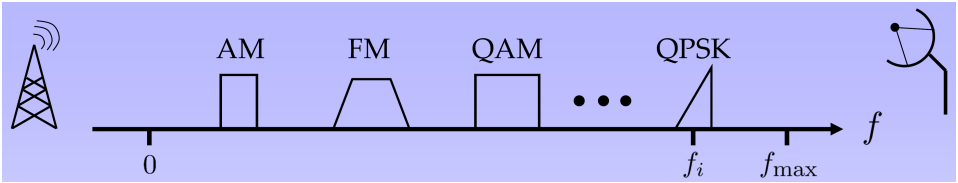
- Principle:
  - A known pulse is transmitted
  - Reflections from targets are received
  - Target's ranges and velocities are identified

- Challenge:
  - All processing is done digitally
  - Targets can lie on an arbitrary grid
  - Process of digitizing
    → loss of resolution in range-velocity domain

- **Subspace methods:**



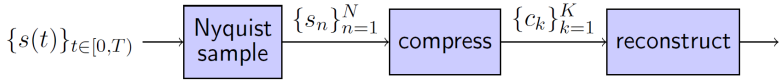
Cognitive Radio



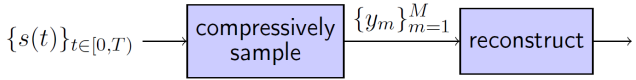
- The spectrum occupation by legacy/primary users is sparse.
- Unlicensed secondary users can insert in spectral holes.
- However, have to find the spectral holes, or the occupied spectrum portions.
- Generalized Nyquist says that one can sample at a rate exceeding the sum of the bandwidth, however here also (only) the spectral support needs to be estimated.

Compressive vs Classical Sampling

- Classical approach:



- New approach:



Nyquist rate $\frac{N}{T}$ $\gg$ compressive sampling rate $\frac{M}{T}$ $\gtrsim$ information rate $\frac{K}{T}$

Linear Measurements Model

1. For now, assume noiseless linear measurements, e.g.,

$$y_m = \int_0^T \phi_m(t)\, s(t)\, dt, \quad m = 1, \ldots, M$$

2. Also assume signal $s(t)$ is bandlimited, in which case Nyquist says

$$s(t) = \sum_{n=1}^{N} s_n \, \mathrm{sinc}\left(\frac{t}{T_s} - n + 1\right), \quad t \in [0, T).$$

Putting these together, we get the convenient discrete representation

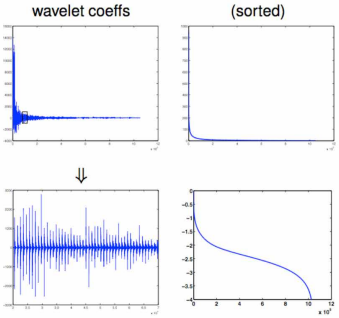
$$y_m = \sum_{n=1}^{N} s_n \underbrace{\int_0^T \phi_m(t) \, \mathrm{sinc}\left(\frac{t}{T_s} - n + 1\right) dt}_{\triangleq \Phi_{m,n}}$$

or, in matrix/vector form, $\boxed{\boldsymbol{y} = \boldsymbol{\Phi s}}$ for $\boldsymbol{s} \in \mathbb{R}^N$ and $\boldsymbol{y} \in \mathbb{R}^M$.

Sparsity

- Many real-world signals are approximately sparse in a known basis.
- For example, natural images are sparse in the discrete wavelet transform (DWT) basis:

1 megapixel image



Typically: 99% signal energy captured by only 2.5% of DWT coefficients!

Sparsity will be captured by prior information in a Bayesian approach.

Ksparse in a Dictionary Ψ

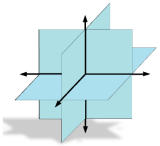
- We say that a signal class $\mathcal{S}$ is K-sparse in the dictionary $\boldsymbol{\Psi}$ if each $\boldsymbol{s} \in \mathcal{S}$ can be written as

$$\boldsymbol{s} = \boldsymbol{\Psi}\boldsymbol{x}$$

  for some K-sparse vector $\boldsymbol{x}$ (i.e., $\boldsymbol{x}$ has at most K nonzero elements).

- Usually orthonormal dictionaries $\boldsymbol{\Psi}$ are used (e.g., DWT, DCT, DFT), but overcomplete dictionaries may also be considered.

- Geometrically, a K-sparse vector $\boldsymbol{x} \in \mathbb{R}^N$ lives in a union of $\binom{N}{K}$ subspaces, each of dimension K:



Combining Sparsity with Compressed Measurements

Recall...

- Linear measurement model: $\boldsymbol{y} = \boldsymbol{\Phi s}$ for $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$
- Sparse signal model: $\boldsymbol{s} = \boldsymbol{\Psi x}$ for K-sparse $\boldsymbol{x} \in \mathbb{R}^{N}$

Together...

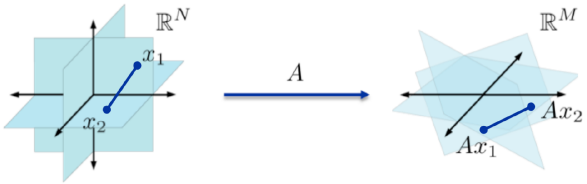
- Compressive sensing model: $\boldsymbol{y} = \underbrace{\boldsymbol{\Phi\Psi}}_{\triangleq \boldsymbol{A}} \boldsymbol{x}$ for $\boldsymbol{A} \in \mathbb{R}^{M \times N}$

Questions:

1. What properties of $\boldsymbol{A}$ ensure the recovery of $\boldsymbol{x}$?
2. Given dictionary $\boldsymbol{\Psi}$, how can we design $\boldsymbol{\Phi}$ to ensure agood $\boldsymbol{A}$?

Restricted Isometry Property

- Recall model: $\boxed{\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \text{ for } \boldsymbol{A} \in \mathbb{R}^{M \times N} \text{ and } K\text{-sparse } \boldsymbol{x} \in \mathbb{R}^N.}$
- Note: if signals $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ map to the same $\boldsymbol{y}$, they can't be recovered!



- In general, for our measurement system to be information preserving, we want that $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \approx \|\boldsymbol{A}\boldsymbol{x}_1 - \boldsymbol{A}\boldsymbol{x}_2\|_2$ for all K-sparse $\boldsymbol{x}_1, \boldsymbol{x}_2$, or

$$1 - \delta \leq \frac{\|\boldsymbol{A}\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_2^2} \leq 1 + \delta \quad \text{for all } 2K\text{-sparse } \boldsymbol{d}. \qquad \text{"RIP"}$$

RIP from Randomness

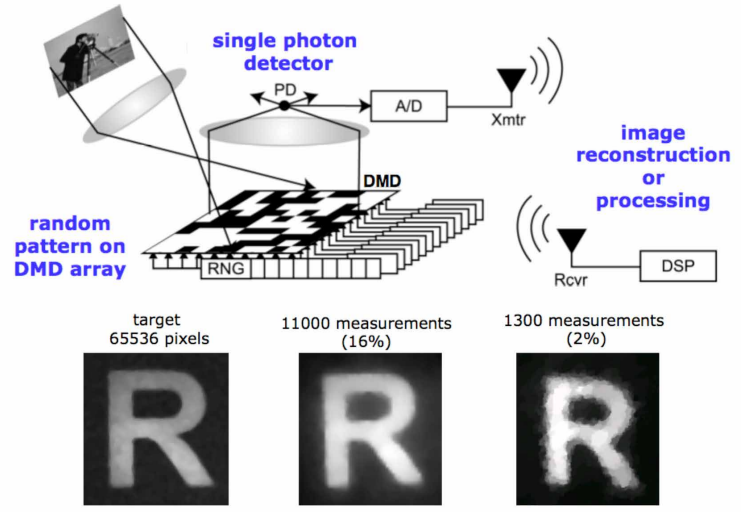
- Testing a given matrix for RIP is an NP-hard (combinatorial) problem.
- Fortunately, if $\boldsymbol{A}$ is randomly drawn with independent zero-mean sub-Gaussian entries (e.g., normal, ± 1), then *with high probability* it will satisfy RIP if

$$M \geq O\Big(K \log \frac{N}{K}\Big).$$

- Similarly, if $\boldsymbol{\Phi}$ is constructed randomly in the same way, then $\boldsymbol{A} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ will satisfy RIP for *any* orthonormal $\boldsymbol{\Psi}$.
- In practice, semi-random $\boldsymbol{\Phi}$ are preferable, e.g.,

  Create $\boldsymbol{\Phi} = \boldsymbol{J}\boldsymbol{F}\boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix with random ± 1s, $\boldsymbol{F}$ is the N-FFT matrix, and $\boldsymbol{J}$ randomly selects M outputs.

Single Pixel Camera (Rice U.)



Best Sparse Fit - the l_0 technique

Find the sparsest $\boldsymbol{x}$ that explains $\boldsymbol{y}$ up to a specified tolerance of ϵ:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \underbrace{\|\boldsymbol{x}\|_0}_{\#\text{ nonzero coefs}} \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon.$$

Unfortunately, this is NP-hard; we'd need to check all $\binom{N}{K} \approx N^K$ possible supports!

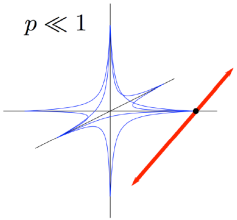
Let's think about this problem geometrically...

Geometry of constrained l_p minimization

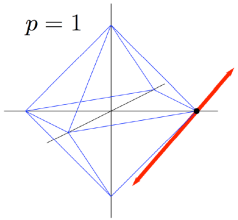
Now consider, for some fixed $p > 0$, the optimization problem:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \underbrace{\|\boldsymbol{x}\|_p}_{\sqrt[p]{\sum_n |x_n|^p}} \quad \text{s.t. } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon.$$

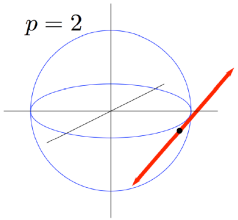
The solution can be found by growing the ℓ_p-ball until it touches the ϵ-rod:



Solution definitely sparse but problem is **NP hard**.



Solution usually sparse and problem is **convex**!



Solution is **not sparse**; $\Leftrightarrow$ LS when $\epsilon = 0$.

This suggests to use the ℓ_1 norm as a surrogate for the ℓ_0 norm!

LASSO

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{s.t. } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon$$

- Convex! Can be solved very efficiently.
- For $\boldsymbol{A}$ satisfying $2K$-RIP, LASSO guarantees that

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq \frac{C_1}{\sqrt{K}} \|\boldsymbol{x} - \boldsymbol{x}_K\|_1 + C_2 \|\boldsymbol{w}\|_2$$

  where $\boldsymbol{x}_K$ is the best K-sparse approximation of $\boldsymbol{x}$ and C_1, C_2 are constants that depend on the RIP δ. Wow!

- In the special case when $\boldsymbol{x}$ is K-sparse, this simplifies to

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq C_2 \|\boldsymbol{w}\|_2.$$

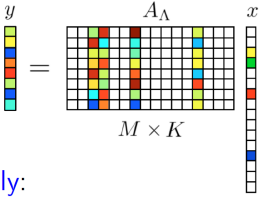
Greedy Search

Main ideas:

- If we can correctly recover the support Λ of $\boldsymbol{x}$ (i.e., the locations of nonzeros), then determining the non-zero amplitudes is easy, e.g.,

$$\boldsymbol{x}_\Lambda = (\boldsymbol{A}_\Lambda^H \boldsymbol{A}_\Lambda)^{-1} \boldsymbol{A}_\Lambda^H \boldsymbol{y}$$

(least squares)



- Estimate the support sequentially:
  - Find the column of $\boldsymbol{A}$ most "similar" to $\boldsymbol{y}$ and store its index.
  - Subtract the effect of this column from $\boldsymbol{y}$.
  - Repeat (until residual is sufficiently small)!

Famous algorithms include MP, OMP, IHT, CoSaMP, Subspace Pursuit

Bayesian Methods

In the Bayesian approach, one ...

- models the signal using a prior pdf $p(\boldsymbol{x})$,
- models the measurement process using a likelihood function $p(\boldsymbol{y}|\boldsymbol{x})$,
- performs inference via Bayes rule, yielding the posterior pdf

$$p(\boldsymbol{x}|\boldsymbol{y}) = Z^{-1} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}) \quad \text{where } Z \text{ is a scaling constant,}$$

- often summarizes the posterior pdf by a point estimate like

$$\widehat{\boldsymbol{x}} = \int \boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{y})\, d\boldsymbol{x} \qquad \text{MMSE estimate}$$

$$\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) \qquad \text{MAP estimate}$$

and possible other statistics that quantify estimate uncertainty.

LASSO as Bayesian Method

If we assume ...

- additive white Gaussian noise of variance σ^2
- i.i.d Laplacian signal with rate λ/σ^2

then

- likelihood: $p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2))$
- prior: $p(\boldsymbol{x}) = \frac{1}{(2\sigma^2/\lambda)^M} \exp(-\frac{\lambda}{\sigma^2}\|\boldsymbol{x}\|_1)$

for which the maximum aposteriori (MAP) estimate is

$$\begin{aligned}\widehat{\boldsymbol{x}} &= \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} \log\left(Z^{-1} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x})\right) \\ &= \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1\end{aligned}$$

which is an unconstrained version of the LASSO problem.

Relevance Vector Machine - Sparse Bayesian Learning

- The RVM is based on the *conditionally* Gaussian priors

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \prod_{n=1}^{N} \mathcal{N}(x_n; 0, \alpha_n^{-1}) \quad \text{and} \quad p(\boldsymbol{\alpha}) = \prod_{n=1}^{N} \Gamma(\alpha_n; 0, 0)$$
$$p(\boldsymbol{w}|\beta) \sim \prod_{m=1}^{M} \mathcal{N}(w_m; 0, \beta^{-1}) \quad \text{and} \quad \beta \sim \Gamma(0, 0)$$

  Note that, as "precision" $\alpha_n \to \infty$, the coefficient x_n is zeroed.

- The *conditional* posterior is (due to Gaussianity) simply

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

- In practice, $\{\boldsymbol{\alpha}, \beta\}$ are estimated using the EM algorithm and then plugged into $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to *approximate* the posterior $p(\boldsymbol{x}|\boldsymbol{y})$.

- The RVM (also known as "SBL" and "BCS") is relatively slow.

Other Bayesian Methods

- Bayesian matching pursuits:
  - Greedy methods that use probabilistic support selection.

- Approximate message passing (AMP):
  - Inspired by methods from statistical physics and information theory.
  - Near-optimal in terms of speed and accuracy if $\boldsymbol{A}$ is large & random.

Alternating Minimization

- also called "cyclic minimization" or "block coordinate descent"
- Cost function to be minimized: $f(\boldsymbol{\theta})$
- Partition $\boldsymbol{\theta} = \boldsymbol{\theta}_{1:m} = \{\theta_1, \cdots, \theta_m\} = \{\boldsymbol{\theta}_{1:k-1}, \theta_k, \boldsymbol{\theta}_{k+1:m}\}$.
- Sweep through the partition. In sweep i:

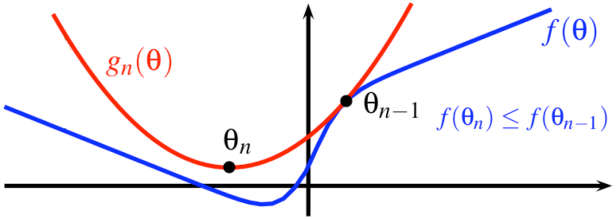
$$\theta_k^{(i)} = \arg\min_{\theta_k} f(\boldsymbol{\theta}_{1:k-1}^{(i)}, \theta_k, \boldsymbol{\theta}_{k+1:m}^{(i-1)}) \quad k = 1, \ldots, m$$

  Partitioning done in such a way that these reduced optimization problems are "easy".
- Guaranteed to converge to a local minimum:

$$f(\boldsymbol{\theta}_{1:k}^{(i)}, \boldsymbol{\theta}_{k+1:m}^{(i-1)}) \leq f(\boldsymbol{\theta}_{1:k-1}^{(i)}, \boldsymbol{\theta}_{k:m}^{(i-1)})$$

- Not necessary to do regular sweeps through the partition, can minimize in any order.

Majorization Minimization



- Cost function to be minimized: $f(\boldsymbol{\theta})$. At iteration $n-1$: have $\boldsymbol{\theta}^{(n-1)}$.
- A majorizer at iteration $n-1$ is a function $g_n(\boldsymbol{\theta})$ such that

$$g_n(\boldsymbol{\theta}^{(n-1)}) = f(\boldsymbol{\theta}^{(n-1)})$$
$$g_n(\boldsymbol{\theta}) \geq f(\boldsymbol{\theta}), \ \forall \boldsymbol{\theta}$$

  where $g_n(\boldsymbol{\theta})$ is possibly convex or in any case "easy" to minimize or just decrease.
- Then

$$\boldsymbol{\theta}^{(n)} = \arg\min_{\boldsymbol{\theta}} g_n(\boldsymbol{\theta})$$
$$f(\boldsymbol{\theta}^{(n)}) \leq g_n(\boldsymbol{\theta}^{(n)}) \leq g_n(\boldsymbol{\theta}^{(n-1)}) = f(\boldsymbol{\theta}^{(n-1)})$$

  Hence guaranteed to converge to a local minimum.

Expectation Maximization (EM) Algorithm

- Maximum Likelihood estimation: $f(\boldsymbol{\theta}) = -\ln p(\boldsymbol{y}|\boldsymbol{\theta})$
- Often resulting from eliminating random $\boldsymbol{x}$ in $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$. $\boldsymbol{\theta}$: deterministic (actual) parameters, $\boldsymbol{x}$: random, "unobserved" data, $\{\boldsymbol{x}, \boldsymbol{y}\}$: "complete" data.
  Estimating $\boldsymbol{\theta}$ from $\{\boldsymbol{x}, \boldsymbol{y}\}$ easier than from $\boldsymbol{y}$ alone.
- Majorizer: $g_i(\boldsymbol{\theta}) = f(\boldsymbol{\theta}^{(i-1)}) - \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(i-1)}}\{\ln\left(\frac{p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}^{(i-1)})}\right)\}$, $g_i(\boldsymbol{\theta}^{(i-1)}) = f(\boldsymbol{\theta}^{(i-1)})$

$$
\begin{aligned}
g_i(\boldsymbol{\theta}) &\geq f(\boldsymbol{\theta}^{(i-1)}) - \ln\left(\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(i-1)}}\{\tfrac{p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}^{(i-1)})}\}\right) \quad \text{Jensen's inequality} \\
&= f(\boldsymbol{\theta}^{(i-1)}) - \ln\left(\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(i-1)}}\{\tfrac{p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(i-1)})\,p(\boldsymbol{y}|\boldsymbol{\theta}^{(i-1)})}\}\right) \\
&= f(\boldsymbol{\theta}^{(i-1)}) - \ln\left(\int p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(i-1)})\tfrac{p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(i-1)})\,p(\boldsymbol{y}|\boldsymbol{\theta}^{(i-1)})}\,d\boldsymbol{x}\right) \\
&= f(\boldsymbol{\theta}^{(i-1)}) - \ln\left(\tfrac{1}{p(\boldsymbol{y}|\boldsymbol{\theta}^{(i-1)})}\underbrace{\int p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\,d\boldsymbol{x}}_{p(\boldsymbol{y}|\boldsymbol{\theta})}\right) = f(\boldsymbol{\theta}^{(i-1)}) - \ln\left(\tfrac{p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{\theta}^{(i-1)})}\right) = -\ln\left(p(\boldsymbol{y}|\boldsymbol{\theta})\right) = f(\boldsymbol{\theta})
\end{aligned}
$$

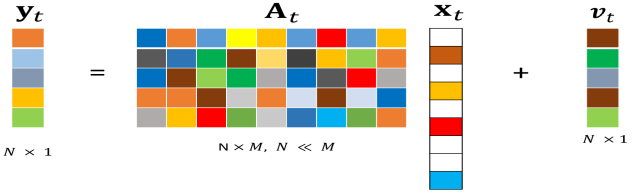
- EM algorithm: at iteration i:

$$
\begin{aligned}
\text{Expectation step:} \quad & \overline{g}_i(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(i-1)}}\{\ln p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\} \\
\text{Maximization step:} \quad & \boldsymbol{\theta}^{(i)} = \arg\max_{\boldsymbol{\theta}} \overline{g}_i(\boldsymbol{\theta})
\end{aligned}
$$

where $g_i(\boldsymbol{\theta}) = \text{constant} - \overline{g}_i(\boldsymbol{\theta})$. EM converges to (local) Maximum Likelihood estimate.

Time Varying Sparse State Tracking

Sparse signal $\boldsymbol{x}_t$ is modeled using an AR(1) process with diagonal correlation coefficient matrix $\boldsymbol{F}$.



Define: $\boldsymbol{\Xi} = \text{diag}(\boldsymbol{\xi})$, $\boldsymbol{F} = \text{diag}(\mathbf{f})$.

f_i : correlation coefficient and $x_{i,t} \sim \mathcal{CN}(\boldsymbol{x}_{i,t}; 0, \frac{1}{\xi_i})$. Further, $\boldsymbol{w}_t \sim \mathcal{CN}(\boldsymbol{w}_t; \mathbf{0}, \boldsymbol{\Gamma}^{-1} = \boldsymbol{\Xi}^{-1}(\mathbf{I} - \boldsymbol{F}\boldsymbol{F}^H))$ and $\boldsymbol{v}_t \sim \mathcal{CN}(\boldsymbol{v}_t; \mathbf{0}, \gamma^{-1}\mathbf{I})$. VB leads to Gaussian SAVE-Kalman Filtering (GS-KF).

Applications: Localization, Adaptive Filtering.

Compressed Sensing Problem

- Noiseless case: Given underdetermined $\boldsymbol{y}$, $\mathbf{A}$, the optimization problem is

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \text{ subject to } \boldsymbol{y} = \mathbf{A}\boldsymbol{x}.$$

  Can recover $\boldsymbol{x}$ and its support for small $N - \|\boldsymbol{x}\|_0$
  (small overdetermination if support were known)
- Noisy case:

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \text{ subject to } \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2 \leq \epsilon.$$

- l_0 norm minimization: an NP-hard problem.
- Constrained problem $\Rightarrow$ Lagrangian, Convex Relaxation (using p norm, $p > 1$):

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_p .$$

  Restricted Isometry Property (RIP): $\mathbf{A}^T\mathbf{A}$ sufficiently diagonally dominant
- Most identifiability work considered noiseless data & exact sparsity

Sparse Signal Recovery Algorithms

Convex Relaxation based Methods:

- Basis pursuit (l_1 norm) [1].
- LASSO(l_1 norm)[2]
- Dantzig selector[3]
- FOCUSS (l_p norm, with $p < 1$).

Greedy Algorithms:

- Matching Pursuit[4]
- Orthogonal Matching Pursuit (OMP)[5]
- CoSaMP[6]

Iterative Methods:

- Iterative Shrinkage and Thresholding Algorithm (ISTA)[7]or Fast ISTA.
- Approximate Message Passing variants (xAMP)- more details to follow.
- Very recent: Deep learning based methods such as Learned ISTA (LISTA)[8], Learned AMP (LAMP) and Learned Vector AMP (LVAMP)[9].

[1]Chen, Donoho, Saunders'99, [2]Tibshirani'96, [3]Candes, Tao'07
[4]Mallat, Zhang'93, [5]Tropp, Gilbert'07, [6]Needell, Tropp'09
[7]Daubechies, Defrise, Mol'04, [8]Gregor, Cun'10, [9]Borgerding, Schniter, Rangan'17

James-Stein Estimator

- Bayesian interpretation of (possibly overdetermined) Compressed Sensing:

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2 - 2\sigma_v^2 \ln p(\boldsymbol{x})$$

- Stein and James[10] showed that for i.i.d. linear Gaussian model $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \mathbf{0}, \xi^{-1}\mathbf{I})$, it is possible to construct a nonlinear estimate of $\boldsymbol{x}$ with lower MSE than that of ML for all values of the true unknown $\boldsymbol{x}$.
- A popular design strategy: is to minimize Stein's unbiased risk estimate (SURE), which is an unbiased estimate of the MSE.
- SURE directly approximates the MSE of an estimate from the data, without requiring knowledge of the hyperparameters (ξ), it is an instance of empirical Bayes.
- Stein's landmark discovery lead to the study of biased estimators that outperform minimum variance unbiased estimators (MVU) in terms of MSE, e.g. work by Yonina Eldar[11].
- Shrinkage estimators and penalized maximum likelihood (PML) estimators.

[10] James, Stein'61
[11] Eldar'08

Kernel Methods in Automatic Control

- Kernel methods in linear system identification[12] ($\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{v}$, $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{v}; \mathbf{0}, \gamma^{-1}\mathbf{I})$).
- Traditional methods: maximum likelihood (ML) or prediction error methods (PEM)
- ML/PEM optimal in the large data limit.
- Questions: Model structure design for ML/PEM. Achieving a good bias-variance trade off.
- Solution: Parameterized Kernel design and hyperparameter estimation. Methods for hyperparameter estimation include cross-validation (CV), empirical Bayes (EB), C_p statistics and Stein's unbiased risk estimate (SURE).
- Regularized least square estimator ($\boldsymbol{P}$ is symmetric and +ve semidefinite kernel matrix):

$$\widehat{\boldsymbol{x}} = \arg \min_{\boldsymbol{x} \in \mathcal{R}^M} \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2 + \frac{1}{\gamma} \boldsymbol{x}^T \boldsymbol{P}^{-1} \boldsymbol{x}.$$

- Parameterized family of matrices, $\boldsymbol{P}(\boldsymbol{\eta})$, where $\boldsymbol{\eta} \in \mathcal{R}^p$. $\boldsymbol{\eta}$ are the hyperparameters.
- Can be interpreted as a constrained form of SBL, with a zero-mean Gaussian prior for $\boldsymbol{x}$ of which the covariance matrix is a linear combination of some fixed matrices (SBL being a special case with fixed matrices $e_i e_i^T$).
- A good overview of Kernel methods, connection with machine learning[13].

[12] Pillonetto, Nicolao'10

[13] Pillonetto, Dinuzzo, Chen, Nicolao, Ljung'14

Kernel Hyperparameter Estimation

- Empirical Bayes (EB=Type II ML):

$$\widehat{\boldsymbol{\eta}}_{EB} = \arg\min_{\boldsymbol{\eta}} f_{EB}(\boldsymbol{P}(\boldsymbol{\eta})),$$
$$f_{EB}(\boldsymbol{P}(\boldsymbol{\eta})) = \boldsymbol{y}^T \boldsymbol{Q}^{-1} \boldsymbol{y} + \ln\det(\boldsymbol{Q}) \text{ with } \boldsymbol{Q} = \mathbf{A}\boldsymbol{P}\mathbf{A}^T + \tfrac{1}{\gamma}\mathbf{I}_N.$$

- Stein's Unbiased Risk Estimator (SURE) method:
- SURE: MSE of signal reconstruction ($MSE_x(\boldsymbol{P}) = \mathrm{E}(\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|^2)$):

$$SURE_x : \widehat{\boldsymbol{\eta}}_{Sx} = \arg\min_{\boldsymbol{\eta}} f_{Sx}(\boldsymbol{P}(\boldsymbol{\eta})), \text{ with}$$
$$f_{Sx}(\boldsymbol{P}(\boldsymbol{\eta})) = \tfrac{1}{\gamma^2}\boldsymbol{y}^T \boldsymbol{Q}^{-T} \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-2}\mathbf{A}^T\boldsymbol{Q}^{-1}\boldsymbol{y} + \tfrac{1}{\gamma}\mathrm{tr}\{2\boldsymbol{R}^{-1} - (\mathbf{A}^T\mathbf{A})^{-1}\},$$
$$\boldsymbol{R} = \mathbf{A}^T\mathbf{A} + \tfrac{1}{\gamma}\boldsymbol{P}^{-1}.$$

- The SURE estimator converge to the best possible hyperparameter in terms of MSE in the asymptotic limit, "asymptotically consistent".
- EB estimator converges to another best hyperparameter minimizing the expectation of the EB estimation criterion (EEB).
- Convergence of EB is faster than that of the SURE estimator.

[13] Mu, Chen, Ljung'18

Outline

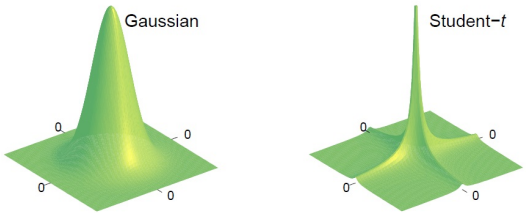
Sparse Bayesian Learning - SBL

- Bayesian Compressed Sensing: 2-layer hierarchical prior for $\boldsymbol{x}$ as in [14], inducing sparsity for $\boldsymbol{x}$ (conjugate priors: posterior pdf of same family as prior pdf) :

$$p_x(x_{i,t}|\xi_i) = \mathcal{N}(x_{i,t}; 0, \xi_i^{-1}),\ p(\xi_i|a,b) = \Gamma^{-1}(a) b^a \xi_i^{a-1} e^{-b\xi_i}$$

$\Rightarrow$ sparsifying Student-t marginal

$$p_x(x_{i,t}) = \frac{b^a \Gamma(a+\frac{1}{2})}{(2\pi)^{\frac{1}{2}}\Gamma(a)} (b + x_{i,t}^2/2)^{-(a+\frac{1}{2})}$$



- Sparsification of the Innovation Sequence: we apply the (Gamma) prior not to the precision of the state $\boldsymbol{x}$ but of it's innovation $\boldsymbol{w}$, allowing to sparsify at the same time the components of x AND their variation in time (innovation).
- The inverse of the noise variance γ is also assumed to have a Gamma prior, $p_\gamma(\gamma|c,d) = \Gamma^{-1}(c) d^c \gamma^{c-1} e^{-d\gamma}$.

[14] Tipping'01

Original SBL Algorithm (Type II ML)

- Original SBL[15], for a fixed estimate of the hyperparameters $(\widehat{\boldsymbol{\xi}}, \widehat{\gamma})$, the posterior of $\boldsymbol{x}$ is Gaussian, i.e.

$$p_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\xi}}, \widehat{\gamma}) = \mathcal{N}(\boldsymbol{x}; \widehat{\boldsymbol{x}}, \boldsymbol{\Sigma}_L)$$

leading to the (Linear) MMSE estimate for $\boldsymbol{x}$

$$\begin{aligned} \widehat{\boldsymbol{x}} &= \widehat{\gamma}(\widehat{\gamma}\mathbf{A}^T\mathbf{A} + \widehat{\boldsymbol{\Xi}})^{-1}\mathbf{A}^T\boldsymbol{y}, \\ \boldsymbol{\Sigma}_L &= (\widehat{\gamma}\mathbf{A}^T\mathbf{A} + \widehat{\boldsymbol{\Xi}})^{-1}. \end{aligned} \tag{1}$$

- The hyperparameters are estimated from the likelihood function by marginalizing over the sparse coefficients $\boldsymbol{x}$, the marginalized likelihood being denoted as $p_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{\xi}, \gamma)$. $\boldsymbol{\xi}, \gamma$ are estimated by maximizing $p_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{\xi}, \gamma)$ and this procedure is called as Type-II ML. Type-II ML is solved using EM, which leads to the following updates for the hyperparameters.

$$\widehat{\xi}_i = \frac{a+\frac{1}{2}}{\left(\frac{<x_i^2>}{\frac{1}{2}} + b\right)}, \text{ where } < x_i^2 >= \widehat{x}_i^2 + \sigma_i^2. \quad < \gamma >= \frac{c+\frac{N}{2}}{\left(\frac{<||\boldsymbol{y} - \mathbf{A}\boldsymbol{x}||^2>}{2} + d\right)},$$

$$\text{where, } < ||\boldsymbol{y} - \mathbf{A}\boldsymbol{x}||^2 >= ||\boldsymbol{y}||^2 - 2\boldsymbol{y}^T\mathbf{A}\widehat{\boldsymbol{x}} + \text{tr}\left(\mathbf{A}^T\mathbf{A}(\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}^T + \boldsymbol{\Sigma})\right),$$

$$\boldsymbol{\Sigma} = diag(\boldsymbol{\Sigma}_L) = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_M^2), \ \widehat{\boldsymbol{x}} = [\widehat{x}_1, \widehat{x}_2, ..., \widehat{x}_M]^T.$$

[15] Tipping'01, Wipf,Rao'04

Type I vs Type II ML

- Type I $\Longrightarrow$ standard MAP estimation (involves integrating out the hyperparameters)

$$\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} [\log p_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{x}) + p_{\boldsymbol{x}}(\boldsymbol{x})].$$

- Type II $\Longrightarrow$ hyperparameters ($\boldsymbol{\Psi} = \{\boldsymbol{\xi}, \gamma\}$) are estimated using an evidence maximization approach

$$\widehat{\boldsymbol{\Psi}} = \arg\max_{\boldsymbol{\Psi}} p_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}|\boldsymbol{y}) = \arg\max_{\boldsymbol{\Psi}} p_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}) \int p_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{\Psi}) = \arg\max_{\boldsymbol{\Psi}} p_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}) \int p_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{x}, \gamma) p_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{\xi}) d\boldsymbol{x}.$$

- Why Type II is better than Type I? [16] In the evidence maximization framework instead of looking for the mode of the true posterior $p_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{y})$, the true posterior is approximated as $p_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{y}; \widehat{\boldsymbol{\Psi}})$, where $\widehat{\boldsymbol{\Psi}}$ is obtained by maximizing the true posterior mass over the subspaces spanned by the non zero indexes.
- Type I methods seek the mode of the true posterior and use that as the point estimate of the desired coefficients. Hence, if the true posterior distribution has a skewed peak, then the Type I estimate (Mode) is not a good representative of the whole posterior.

[16] Giri, Rao'16

Variational Bayesian (VB) Inference

- The computation of the posterior distribution of the parameters is usually intractable. As in SAGE, SAVE is simply VB with partitioning of the unknowns at the scalar level. Define $\boldsymbol{\theta} = \{\boldsymbol{x}, \boldsymbol{\xi}, \gamma\}$, θ_i represents each scalar and $\boldsymbol{\theta}_{\bar{i}}$ denotes $\boldsymbol{\theta}$ excluding θ_i.

$$q(\boldsymbol{\theta}) \;=\; q_\gamma(\gamma) \prod_{i=1}^{M} q_{x_i}(x_i) \prod_{i=1}^{M} q_{\xi_i}(\xi_i).$$

- VB compute the factors q by minimizing the Kullback-Leibler distance between the true posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ and the $q(\boldsymbol{\theta})$. From [17],

$$KLD_{VB} \;=\; D_{KL}\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y})\right) = \int q(\boldsymbol{\theta}) ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta}.$$

- The KL divergence minimization is equivalent to maximizing the evidence lower bound (ELBO)[18].

$$\ln p(\boldsymbol{y}) = L(q) + KLD_{VB} = -D_{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}, y)) + D_{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y})), \quad \text{where,}$$

  $\ln p(\boldsymbol{y})$ is the evidence, and min KLD_{VB} becomes equivalent to max $L(q)$, the ELBO.

- We get for the element-wise VB recursions: (Expectation Maximization (EM) = special case: $\boldsymbol{\theta} = \{\boldsymbol{\theta}_s, \boldsymbol{\theta}_d\}$, $\boldsymbol{\theta}_s$ random, hidden $\boldsymbol{\theta}_d$ deterministic)

$$\ln(q_i(\theta_i)) = < \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_{\boldsymbol{\theta}_{\bar{i}}} + c_i,$$

[18] Beal'03, [19] Tzikas, Likas, Galatsanos'08

Low Complexity-Space Alternating Variational Estimation (SAVE)

- Mean Field (MF) approximation: VB partitioned to scalar level (MF vs VB // SAGE vs EM), results in a SBL algorithm without any matrix inversions.
- The resulting alternating optimization of the posteriors for each scalar in $\boldsymbol{\theta}$ leads to

$$\ln(q_i(\theta_i)) = < \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_{k \neq i} + c_i,$$
$$p(\boldsymbol{y}, \boldsymbol{\theta}) = p_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}, \gamma) p_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{\xi}) p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) p_{\gamma}(\gamma).$$

where $\boldsymbol{\theta} = \{\boldsymbol{x}, \boldsymbol{\xi}, \gamma\}$ and θ_i represents each scalar in $\boldsymbol{\theta}$.

$$\ln p(\boldsymbol{y}, \boldsymbol{\theta}) = \frac{N}{2} \ln \gamma - \frac{\gamma}{2} ||\boldsymbol{y} - \mathbf{A}\boldsymbol{x}||^2 + \sum_{i=1}^{M} \left(\frac{1}{2} \ln \xi_i - \frac{\xi_i}{2} x_i^2 \right) +$$
$$\sum_{i=1}^{M} ((a-1) \ln \xi_i + a \ln b - b\xi_i) + (c-1) \ln \gamma + c \ln d - d\gamma + \text{constants}.$$

- Gaussian approximate posterior for x_i:

$$\ln q_{x_i}(x_i) = -\frac{<\gamma>}{2} \Big\{ < ||\boldsymbol{y} - \mathbf{A}_{\bar{i}} \boldsymbol{x}_{\bar{i}}||^2 > - (\boldsymbol{y} - \mathbf{A}_{\bar{i}} < \boldsymbol{x}_{\bar{i}} >)^T \mathbf{A}_i x_i -$$
$$x_i \mathbf{A}_i^T (\boldsymbol{y} - \mathbf{A}_{\bar{i}} < \boldsymbol{x}_{\bar{i}} >) + ||\mathbf{A}_i||^2 x_i^2 \Big\} - \frac{<\xi_i>}{2} x_i^2 + c_{x_i} = -\frac{1}{2\sigma_i^2} (x_i - \widehat{x}_i)^2 + c'_{x_i}.$$

SAVE Iterations Continued...

- The SAVE iterations for $\boldsymbol{x}$ get obtained as

$$\sigma_i^2 = \frac{1}{<\gamma>||\mathbf{A}_i||^2+\xi_i}, \quad \widehat{x}_i = \sigma_i^2 \mathbf{A}_i^T (\boldsymbol{y} - \mathbf{A}_{\bar{i}} \widehat{\boldsymbol{x}}_{\bar{i}}) < \gamma > .$$

- Hyperparameter estimates: same as EM iterations. Gamma posterior for ξ_i and γ.
- No matrix inversions.
- Update of all the variables, $\boldsymbol{x}, \xi_i, \gamma$, requires simple addition and multiplication operations
- $\boldsymbol{y}^T\mathbf{A}$, $\mathbf{A}^T\mathbf{A}$ and $||\boldsymbol{y}||^2$ can be precomputed, so only need to be computed once.
- **Remarks:** From LMMSE expression in (1), i^{th} row of $\gamma \mathbf{A}^T\mathbf{A}\widehat{\boldsymbol{x}} + \boldsymbol{\Xi}\widehat{\boldsymbol{x}} = \gamma \mathbf{A}^T\boldsymbol{y}$:

  $\gamma \mathbf{A}_i^T\mathbf{A}\widehat{\boldsymbol{x}} + \xi_i \widehat{x}_i = \gamma \mathbf{A}_i^T\boldsymbol{y}$ or $(\gamma \|\mathbf{A}_i\|^2 + \xi_i)\widehat{x}_i = \gamma \mathbf{A}_i^T(\boldsymbol{y} - \mathbf{A}_{\bar{i}}\widehat{\boldsymbol{x}}_{\bar{i}})$

  Hence SAVE does linear PIC iterations to compute the LMMSE estimate.
- However, for the posterior variances : $\sigma_i^2 = ((\boldsymbol{\Sigma}_L^{-1})_{i,i})^{-1} \leq (\boldsymbol{\Sigma}_L)i,i$ with equality only for diagonal $\boldsymbol{\Sigma}_L$

Convergence of SAVE

Theorem 1

The convergence condition for the sparse coefficients x_i of the SAVE algorithm [20] can be written as $\rho(\mathbf{D}^{-1}\mathbf{H}) < 1$, where $\mathbf{D} = diag(\widehat{\gamma}\mathbf{A}^T\mathbf{A} + \widehat{\boldsymbol{\Xi}}), \mathbf{H} = offdiag(\widehat{\gamma}\mathbf{A}^T\mathbf{A})$. $\rho(\cdot)$ denotes the spectral radius. Moreover, if SAVE converges, assuming the estimate of hyperparameters are consistent, the posterior mean (point estimate) always converges to the exact value (LMMSE). However, the predicted posterior variance is quite suboptimal.

Remark: To fix the convergence of SAVE (when $\rho(\mathbf{D}^{-1}\mathbf{H}) > 1$), we can use the diagonal loading method[21]. The modified iterations (with a diagonal loading factor matrix $\boldsymbol{\Lambda}$) can be written as,

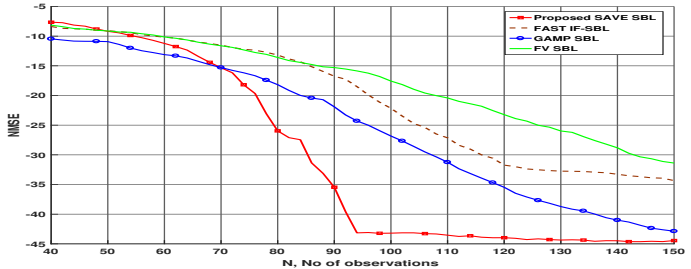
$$\begin{aligned}(\mathbf{D} + \widetilde{\boldsymbol{\Xi}})\boldsymbol{x}^{(t+1)} &= -(\mathbf{H} - \widetilde{\boldsymbol{\Xi}})\boldsymbol{x}^{(t)} + \widehat{\gamma}\mathbf{A}^T\boldsymbol{y}, \implies \\ \boldsymbol{x}^{(t+1)} &= -(\mathbf{D} + \widetilde{\boldsymbol{\Xi}})^{-1}(\mathbf{H} - \boldsymbol{\Xi})\boldsymbol{x}^{(t)} + (\mathbf{D} + \widetilde{\boldsymbol{\Xi}})^{-1}\widehat{\gamma}\mathbf{A}^T\boldsymbol{y}.\end{aligned}$$

The convergence condition gets modified as $\rho((\mathbf{D} + \widetilde{\boldsymbol{\Xi}})^{-1}(\mathbf{H} - \widetilde{\boldsymbol{\Xi}})) < 1$. Another point worth noting here is that, if the power delay profile $\boldsymbol{\Xi}$ is also estimated using MF, $\widehat{\gamma}\text{diag}(\mathbf{A}^T\mathbf{A}) + \widehat{\boldsymbol{\Xi}}$, where $\widehat{\boldsymbol{\Xi}} = \boldsymbol{\Xi} + \widetilde{\boldsymbol{\Xi}}$, with $\widetilde{\boldsymbol{\Xi}} > 0$. In this case, $\widetilde{\boldsymbol{\Xi}}$ may represent an automatic correction factor (diagonal loading) to force convergence of SAVE for cases where $\rho(\mathbf{D}^{-1}\mathbf{H}) > 1$.

[20] Thomas,Slock'18

[21] Johnson, Bickson, Dolev'09

NMSE Results



NMSE vs the number of observations M ($N = 200, K = 40$).

- For sufficient amount of data, SAVE has significantly lower MSE than the other fast algorithms.
- Why? The resulting problem of alternating optimization of $\boldsymbol{x}$ and the hyperparameters $\boldsymbol{\xi}$ and γ appears to be characterized by many local optima. Apparently, the component-wise VB approach appears to allow to avoid a lot of bad local optima, explaining the better performance, apart from lower complexity.
- At very low amount of data, suboptimal approaches such as AMP which don't introduce individual hyper parameters per $\boldsymbol{x}$ component and assume that the x_i behave i.i.d, behave better because of the lower number of hyper parameters to be estimated.

An Overview of Fast SBL Algorithms

- Fast SBL using Type II ML by Tipping [22]: greedy approach handling one x_i at a time, plus replacing precisions by their convergence values, leading to pruning of the small x_i components, i.e. explicit sparsity.
- Fast SBL using VB by Shutin et. al.[23]: Shutin uses VB while Tipping is Type II ML as in original SBL. They do both replace precisions by their convergence values. Shutin also added some extra view points in terms of the pruning condition being interpreted as relating between sparsity properties of SBL and a measure of SNR. Main message of the both being faster convergence compared to original SBL, not much reduction in per iteration complexity.
- BP-SBL[24]: In SBL, with fixed hyperparameters, MAP or MMSE estimate (follows from the Gaussian posterior) of $\boldsymbol{x}$ can be efficiently computed using belief propagation (BP), since all the messages involved are Gaussian (without any approx.).
- Inverse Free SBL (IF-SBL)[25]: Optimization using a relaxed ELBO.
- Hyperparameter free sparse estimation[26]: Does not require hyperparameter tuning compared to SBL. Uses covariance matching, equivalent to square root LASSO.

[21] Tipping, Faul'03, [22] Shutin, Buchgraber, Kulkarni, Poor'11
[23] Tan, Li'10, [24] Duan, Yang, Fang, Li'17
[25] Zachariah, Stoica'15

Complexity Comparisons-SBL Algorithms

Algorithm	Complexity per Iteration	Convergence (No of iterations)	Sparsity	Optimization function	Local Optimum
Type I	$O(M^3)$		Exact sparsity	Type I ML (Depending upon the prior used, type 1 ML corresponds to LASSO/Re-weighted l1/l2 min. problems)	
Type II SBL	$O(M^3)$		Exact sparsity (\alpha_i converges to \infty)	Type II ML solved using EM	
Fast SBL using Type II ML (Tipping,Faul'03) (Focus more on Convergence speed)	$O(L^3),\ L \leq M$	<< L	Exact sparsity (Using an entry dependent thresholding condition which follows from the computation of stationary point of \alpha_i)	Type II ML (stationary points of \alpha_i are computed to accelerate convergence)	Convergence to a local optimum.
Fast SBL (using VB) by Shutin (Focus more on Convergence speed)	$O(L^3),\ L \leq M$	<< L	Exact sparsity (Using a pruning condition similar as in Tipping's)	Maximization of ELBO in VB	Convergence to a local optimum of ELBO (Mean field free energy)
Hyperparameter free SBL (Zachariah, Stoica'15)	$O(M\hat{}2)$	<< M	The final objective function is a weighted square root LASSO. So the sum of l2 norm of (y and Ax) and weighted l1 norm of x which promotes sparsity here.	LMMSE estimator for x with Covariance matching for PDP, finally giving rise to an objective function which can be interpreted as weighted square root LASSO.	Convergence to a local optimum
BP-SBL (Tan, Li'10)	$O(MN)$ (Similar complexity as xAMP, see matrix form of the BP-SBL in the upcoming slides)	$\log(MN)$	Does not give exact sparsity	Posterior of x computed using BP and EM for hyper-parameters	Convergence to local optimum of Bethe Free Energy (BFE)
GAMP-SBL (Shoukairi, Schniter, Rao'18)	$O(MN)$	<< M	Does not give exact sparsity	Using GAMP for posterior of x, EM for hyperparameters	Convergence to local optimum of LSL-BFE
SAVE	$O(MN)$	<< M	Does not give exact sparsity	Maximization of ELBO in VB	Convergence to a local optimum of ELBO
Inverse Free SBL (Duan, Yang, Fang, Li'17)	$O(MN)$	<< M (similar to GAMP SBL)	Does not give Exact sparsity	Maximization of an approximate ELBO in VB	Convergence to a local optim the approx ELBO

References I

M. Bayati, A. Montanari, "The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing," in *IEEE Trans. on Info. Theo.*, Feb. 2011.

M. J. Beal, "Variational Algorithms for Approximate Bayesian Inference," in *Thesis, Univeristy of Cambridge, UK*, May 2003.

M. Borgerding, P. Schniter, S. Rangan, "AMP-Inspired Deep Networks for Sparse Linear Inverse Problems," *IEEE Trans. on Sig. Process.*, Aug. 2017.

B. Çakmak, M. Opper, "Expectation Propagation for Approximate Inference: Free Probability Framework," in *IEEE Inter. Sympo. On Info. Theo. (ISIT)*, 2018.

B. Çakmak, O. Winther, B. H. Fleury, "S-AMP: Approximate Message Passing for General Matrix Ensembles," in *IEEE Intl. Sympo. Info. Theo.*, 2014.

E. Candes, T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n.," *The annals of Statistics*, 2007.

S. S. Chen, D. L. Donoho, M. A. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM J. Sci. Comput.*, 1999.

R. Couillet, J. Hoydis, M. Debbah", "Random Beamforming over Quasi-Static and Fading Channels: A Deterministic Equivalent Approach," in *IEEE Trans. On Info. Theo.*, 2012.

I. Daubechies, M. Defrise, C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint.," *Comm. on Pure and Applied Mathematics*, 2004.

J. Du, S. Ma, Y-C. Wu, S. Kar, J. M. F. Moura, "Convergence Analysis of Distributed Inference with Vector-Valued Gaussian Belief Propagation," in *Jrnl. of Mach. Learn. Res.*, Apr. 2018.

H. Duan, L. Yang, J. Fang, H. Li, "Fast Inverse-Free Sparse Bayesian Learning via Relaxed Evidence Lower Bound Maximization," in *IEEE Sig. Process. Lett.*, Jun. 2017.

Y. C. Eldar "Rethinking Biased Estimation: Improving Maximum Likelihood and the CramérRao Bound," *Found. and Tren. in Sig. Process.*, 2008.

References II

S. Fortunati, F. Gini, M. S. Greco, C. D. Richmond, "Performance Bounds for Parameter Estimation under Misspecified Models: Fundamental Findings and Applications," in *IEEE Sig. Proc. Mag.*, Nov. 2017.

A. E. Gelfand, S. K. Sahu, "Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models," in *Journ. of the Americ. Stat. Assoc.*, Mar. 1999.

R. Giri, B. Rao, "Type I and Type II Bayesian Methods for Sparse Signal Recovery Using Scale Mixtures," in *IEEE Trans. Sig. Process.*, July 2016.

K. Gregor, Y. LeCun, "Learning Fast Approximations of Sparse Coding," *Intl. Conf. on Mach. Learn.*, 2010.

W. James, C. Stein "Estimation with quadratic loss," *Proc. 4th Berkeley Symp. Mathematical Statistics Probability*, 1961.

J. K. Johnson, D. Bickson, D. Dolev, "Fixing Convergence of Gaussian Belief Propagation," in *IEEE Intl. Symp. on Info. Theo.*, 2009.

M. Luo, Q. Guo, D. Huang, J. Xi, "Sparse Bayesian Learning using Approximate Message Passing with Unitary Transformation," in *IEEE VTS Asia Pac. Wire. Commun. Symp., APWCS*, Aug. 2019.

D. M. Malioutov, J. K. Johnson, A. S. Willsky, "Walk-Sums and Belief Propagation in Gaussian Graphical Models," in *Jrnl. of Mach. Learn. Res.*, Oct. 2006.

S. Mallat, Z. Zhang, "Matching Pursuits with time-frequency dictionaries.," *IEEE Trans. Sig. Process.*, 1993.

J. Ma, L. Ping "Orthogonal AMP," in *IEEE Access*, Mar. 2017.

T. P. Minka, "Expectation propagation for approximate Bayesian inference ," in *Proc. of Conf. on Uncert. in Art. Intell. (UAI)*, 2001.

B. Mu, T. Chen, L. Ljung, "On asymptotic properties of hyperparameter estimators for kernel-based regularization methods," in *Automatica*, May. 2018.

D. Needell, J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples.," *Applied and computational harmonic analysis* , 2009.

References III


G. Pillonetto, F. Dinuzzo, T. Chen, G. D Nicolao, L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, Feb. 2014.


G. Pillonetto, G. D. Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, 2010.


S. Rangan, "Generalized Approximate Message Passing for Estimation with Random Linear Mixing," in *IEEE Intl. Sympo. Info. Theo.*, 2011.


S. Rangan, A. K. Fletcher, P. Schniter, U. S. Kamilov, "Inference for Generalized Linear Models via Alternating Directions and Bethe Free Energy Minimization," in *IEEE Trans. Inf. Theory*, Jan. 2017.


S. Rangan, P. Schniter, A. K. Fletcher, "Vector Approximate Message Passing," in *IEEE Trans. Inf. Theory*, Oct. 2019.


C. D. Richmond and L. L. Horowitz, "Parameter bounds on estimation accuracy under model misspecification," in *IEEE Trans. on Sig. Process.*, May. 2015.


E. Riegler, G. E. Kirkelund, C. N. Manchón, B. H. Fleury, "Merging Belief Propagation and the Mean Field Approximation: a Free Energy Approach," in *IEEE Trans. on Info. Theo.*, Jan. 2013.


P. Rusmevichientong, B. Van Roy, "An analysis of belief propagation on the turbo decoding graph with Gaussian densities," *IEEE Trans. Inf. Theory*, 2001.


M. Al-Shoukairi, P. Schniter, B. D. Rao, "GAMP-Based Low Complexity Sparse Bayesian Learning Algorithm," in *IEEE Trans. on Sig. Process.*, Jan. 2018.


D. Shutin, T. Buchgraber, S. R. Kulkarni, H. V. Poor, "Fast Variational Sparse Bayesian LearningWith Automatic Relevance Determination for Superimposed Signals," in *IEEE Trans. Sig. Process.*, Dec. 2011.


V. Šmídl, A. Quinn "The Variational Bayes Method in Signal Processing," in *Springer Series on Sig. and Comm. Tech.*, 2005.


K. Takeuchi, "Rigorous Dynamics of Expectation- Propagation-Based Signal Recovery from Unitarily Invariant Measurements," in *IEEE Trans. Inf. Theory*, Jan. 2020.


X. Tan, J. Li, "Computationally Efficient Sparse Bayesian Learning via Belief Propagation," in *IEEE Trans. on Sig. Proc.*, Apr. 2010.

References IV

C. K. Thomas, K. Gopala, D. Slock, "Sparse Bayesian learning for a bilinear calibration model and mismatched CRB," in *EUSIPCO*, 2019.

C. K Thomas, D. Slock, "SAVE - space alternating variational estimation for sparse Bayesian learning," in *IEEE Data Sci. Wkshp.*, Jun. 2018.

C. K. Thomas, D. Slock, "Space Alternating Variational Estimation and Kronecker Structured Dictionary Learning," in *IEEE ICASSP*, 2019.

C. K. Thomas and D. Slock, "Convergence Analysis Of Sparse Bayesian Learning under Approximate Inference Techniques," in *Asilomar Conf. on Sig., Sys., and Comp.*, Nov. 2019.

C. K. Thomas and D. Slock, "Low Complexity Static and Dynamic Sparse Bayesian Learning Combining BP, VB and EP Message Passing," in *IEEE Asilomar*, Nov. 2019.

R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc. Series B (Methodol.)*, 1996.

M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learn. Res.*, 2001.

M. E. Tipping, A. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *AISTATS*, 2003.

J. A. Tropp, A. C. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Trans. Info. Theo.*, 2007.

D. G. Tzikas, A. C. Likas, N. P. Galatsanos, "The variational approximation for Bayesian inference," in *IEEE Sig. Process. Mag.*, Nov. 2008.

S. Wagner, R. Couillet, M. Debbah, D. Slock, "Large System Analysis of Linear Precoding in MISO Broadcast Channels with Limited Feedback," in *IEEE Trans. Inf. Theory*, July. 2012.

Y. Weiss, W. T. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *NIPS*, 2000.

D. P. Wipf, B. D. Rao, "Sparse Bayesian Learning for Basis Selection," *IEEE Trans. Sig. Process.*, Aug 2004.

J. S. Yedidia, W. T. Freeman, Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," in *IEEE Trans. on Info. Theo.*, Jun. 2005.

D. Zachariah, P. Stoica, "Online Hyperparameter-Free Sparse Estimation Method," in *IEEE Trans. on Sig. Proc.*, July. 2015.