EMPIRICAL BAYESIAN INFERENCE USING JOINT SPARSITY

Anne Gelb

Research supported by ONR MURI #N00014-20-1-2595, AFOSR9550-18-1-0316 and NSF-DMS 1502640, and all in collaboration with Dr. Theresa Scarnati at the Wright Patterson Air Force Research Lab.

SIMDA Meeting October 13 2020.

Department of Mathematics, Dartmouth College, Hanover, NH 03755.
annegelb@math.dartmouth.edu

Empirical Bayesian Inference using Joint Sparsity

- Let $\mathbf{X}, \mathbf{Y}, \mathbf{E}$ be independent random variables sampled from a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.
- We consider the linear forward model of the form

$$\mathbf{Y} = A\mathbf{X} + \mathbf{E}.$$

  $\mathbf{X}$ represents the underlying, unknown we seek to recover, $\mathbf{Y}$ corresponds to the distribution of the observable data, $A \in \mathbb{C}^{n\times n}$ is a known forward operator, and $\mathbf{E} \sim \mathcal{N}(0, \sigma^2\mathbb{I})$ models the noise.
- A particular observation of is given by $y = Ax + e$.
- We consider the case where we have multiple observable measurements $y^j = A^j x + e^j \in Y$, $1 \leq j \leq J$.
- The collection of measurements are commonly referred to as multiple measurement vectors (MMVs) for $J > 1$.
- For now we consider the same source of measurements, but the goal is to use multiple sources.
- In fact we only need one measurement. We can generate multiple data sets by adding noise.

Assumptions made for signal recovery

- The underlying signal is *sparse* in some domain (e.g. gradient, edge, or wavelet domain).
- For given observation $y = Ax + e$, *compressive sensing* (CS) algorithms are often employed:

$$x^* = \arg\min_{x \in \mathbb{R}^n} \left\{ \lambda ||\mathcal{L}x||_1 + \frac{1}{2}||y - Ax||_2^2 \right\}.$$

- $\mathcal{L}$ is a pre-determined sparsifying transform.
- The ℓ_1 norm is typically viewed as a surrogate for the ℓ_0 pseudo-norm since the latter yields an intractable problem.
- The regularization parameter λ is chosen to balance the contribution from each term.
- Smaller λ means more confidence in the data (fidelity term) $||y - Ax||_2^2$ and vice versa.

Drawbacks of ℓ_1 regularization and potential fixes

- Lack of robustness in parameter selection – prevents automation.
- The standard sparsity assumption affects the *global* solution – it does not capture any *local* features in the sparse domain. (This is because the regularization enforces small ℓ_1 norm.)
- It would be better to have a spatially varying parameter $\lambda(x)$ that only penalizes sparse regions while allowing true signal to "pass through".
- Various *iterative reweighting* methods have been developed for this purpose. However (i) they have issues with robustness since an incorrect weight will be subsequently reinforced in later iterations (ii) not efficient since a new weighting parameter has to be calculated with each iteration.

Drawbacks of ℓ_1 regularization and potential fixes

How to address these issues:

- The additional information from the MMVs can be further exploited by employing the joint sparsity assumption – the sparse domain of the underlying signal should be similar across all collected measurements (for that same signal).
- We can extract this joint sparsity by looking at the variability of the multiple measurements in the sparse domain.
- The weights are essentially designed to be inversely proportional to the spatial variation – low variability suggests true sparsity, so the regularization term should be heavily penalized. In contrast, high variability suggests that we are in a region of support.
- The weights can be calculated offline instead of iteratively: (1) more efficient and (2) more robust. It will not reinforce "bad" solutions.
- More accurate information about support in the underlying sparse domain can also inform posterior estimates and reduce the uncertainty.

Definitions and notations surrounding joint sparsity

Let $v \in \mathbb{R}^n$.

- We write $\|\cdot\|_0$ for the ℓ^0-'norm', i.e.

$$\|v\|_0 = |\mathrm{supp}(v)|,$$

where $\mathrm{supp}(v)$ is the support of $v = (v_i)_{i=1}^n$ defined by $\mathrm{supp}(v) = \{i : v_i \neq 0\}$.

- Given a vector $\boldsymbol{w} = (w_i)_{i=1}^n$ of positive weights, we define the weighted $\ell^1_{\boldsymbol{w}}$-norm as

$$\|v\|_{1,\boldsymbol{w}} = \sum_{i=1}^n w_i |v_i| = \|Wv\|_1$$

where $W = diag(w_1, \ldots, w_n)$.

- If $V = (v_{i,j})_{i,j=1}^{n,J} \in \mathbb{R}^{n \times J}$ is a matrix, we define the ℓ^{q_1,q_2}-norms by

$$\|V\|_{q_1,q_2} = \left(\sum_{i=1}^n \left(\sum_{j=1}^J |v_{i,j}|^{q_1} \right)^{q_2/q_1} \right)^{1/q_2}.$$

- Again by convention the $\ell^{2,0}$-'norm' is defined as $\|V\|_{2,0} = \left| \left\{ i : \sum_{j=1}^J |v_{i,j}|^2 \neq 0 \right\} \right|$.

Definitions and notations surrounding joint sparsity

Definition

A vector $v \in \mathbb{R}^n$ is s-sparse for some $1 \leq s \leq N$ if

$$\|v\|_0 = |\text{supp}(v)| \leq s.$$

A collections of vectors $v_1, \ldots, v_J \in \mathbb{R}^n$ is s-joint sparse if

$$\|V\|_{2,0} = \left| \bigcup_{j=1}^{J} \text{supp}\,(v_j) \right| \leq s,$$

where $V = [v_1 | \cdots | v_J]$.

Weighted ℓ_p regularization informed by joint sparsity

in collaboration with Theresa Scarnati, supported by AFOSR #FA9550-18-1-0316.

1. x is a signal with corresponding sparse domain vector g.
2. Acquire measurements $y^j = A^j x + e^j, \quad j = 1, \ldots, J$.
3. Calculate J sparse vector approximations of g from measurements:

$$\tilde{g}^j = B^j y^j, \quad j = 1, \ldots, J.$$

4. Determine spatially varying matrix $W = \text{diag}\,(w_1, w_2, ..., w_N)$. The diagonal entries of W are (roughly) inversely proportional to the support of g, which we learn from $\{\tilde{g}^j\}_{j=1}^J$.
5. The weighted ℓ_p regularization problem is

$$x^* = \underset{q \in \mathbb{R}^N}{\arg\min} \left\{ \frac{1}{p} ||W\mathcal{L}q||_p^p + \frac{1}{2} ||Aq - y||_2^2 \right\}.$$

6. $\mathcal{L}$ is a sparsifying transform operator with $\mathcal{L}x \approx g$.

Weighted ℓ_p regularization informed by joint sparsity

$$x^* = \arg\min_{q \in \mathbb{R}^N} \left\{ \frac{1}{p} ||WLq||_p^p + \frac{1}{2} ||Aq - y||_2^2 \right\}.$$

One dimensional example: Measurements $y^j = F^j x + e^j$ are Fourier coefficients of a ramp function in different frequency bands. Determine $\tilde{g}^j$ *directly* from the measurements without approximating reconstruction.

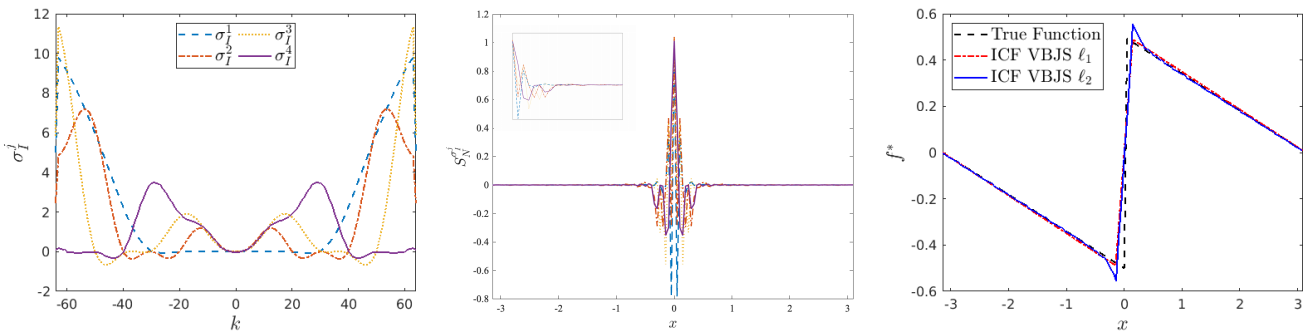

Figure: (left) different Fourier bands of data to obtain y^j. (middle) $\{\tilde{g}^j\}_{j=1}^4$. (right) weighted ℓ_p reconstruction.

We will return later to how to effectively construct W using the concept of joint sparsity.

Motivation for using Bayesian framework

- Many techniques have been introduced to improve the robustness and efficacy of compressive sensing algorithms in both single and multiple measurement cases.
- However, since the recovery is always limited to point estimate solutions, it is not possible to quantify the uncertainty.
- This disadvantage is non-trivial because in practice it is often crucial to know how reliable the recovery is, especially in the case of noisy or limited data availability within in each measurement vector.
- Even within the same application, compressive sensing algorithms inevitably require hand tuning of the regularization parameter λ, with small changes often resulting in drastically different recovery outcomes.

The Bayesian approach for solving inverse problems

- Let $\mathbf{X}, \mathbf{Y}, \mathbf{E}$ be independent random variables sampled from a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.
- We consider the linear forward model of the form

$$\mathbf{Y} = A\mathbf{X} + \mathbf{E}.$$

  $\mathbf{X}$ represents the underlying, unknown we seek to recover, $\mathbf{Y}$ corresponds to the distribution of the observable data, $A \in \mathbb{C}^{n \times n}$ is a known forward operator, and $\mathbf{E} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$ models the noise.
- A particular observation of is given by $y = Ax + e$.
- As we already discussed, compressive sensing algorithms are used to obtain a point estimate solution. However, to assess the uncertainty of the our recovery, it is often more valuable to explore the entire posterior probability density corresponding to the distribution of $\boldsymbol{X}$.

The Bayesian approach for solving inverse problems

Let $x, y, e \in \Omega$ and $\lambda \in [0,1]$ be realizations of the random variables $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{E}$ and $\boldsymbol{\lambda}$ respectively. Bayes' theorem is given by

$$f_{\mathbf{X},\boldsymbol{\lambda}|\mathbf{Y}}(x,\lambda|y) = \frac{f_{\mathbf{Y}|\mathbf{X}}(y|x) f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda) f_{\boldsymbol{\lambda}}(\lambda)}{f_{\mathbf{Y}}(y)} \propto f_{\mathbf{Y}|\mathbf{X}}(y|x) f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda) f_{\boldsymbol{\lambda}}(\lambda).$$

- The likelihood (for $\mathbf{E} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$) is

$$f_{\mathbf{Y}|\mathbf{X}}(y|x) = f_{\mathbf{E}}(y - Ax) \propto \exp\{-\frac{1}{2\sigma^2}||y - Ax||_2^2\}.$$

- The prior is $\tilde{f}_{\mathbf{X},\boldsymbol{\lambda}}(x,\lambda) = f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda) f_{\boldsymbol{\lambda}}(\lambda)$, where $f_{\boldsymbol{\lambda}}(\lambda)$ is the hyper-prior.
- The evidence $f_{\mathbf{Y}}(y)$ is often computationally intractable so we use the un-normalized version (standard in practice). It is nonzero because otherwise the observation has a probability of zero.

Empirical Bayesian Inference using Joint Sparsity

- For now we do not focus on the likelihood. In the future we may want to want to weigh the data fidelity terms differently depending on a variety of factors.

Empirical Bayesian Inference using Joint Sparsity

- For now we do not focus on the likelihood. In the future we may want to want to weigh the data fidelity terms differently depending on a variety of factors.
- The focus of the current work is to construct an appropriate prior, that is informed by the support locations in the sparse domain.

Empirical Bayesian Inference using Joint Sparsity

- For now we do not focus on the likelihood. In the future we may want to want to weigh the data fidelity terms differently depending on a variety of factors.
- The focus of the current work is to construct an appropriate prior, that is informed by the support locations in the sparse domain.
- Ideally, this will allow us to (spatially) separate regions based on extractable features of the underlying image (signal).

Empirical Bayesian Inference using Joint Sparsity

- For now we do not focus on the likelihood. In the future we may want to want to weigh the data fidelity terms differently depending on a variety of factors.
- The focus of the current work is to construct an appropriate prior, that is informed by the support locations in the sparse domain.
- Ideally, this will allow us to (spatially) separate regions based on extractable features of the underlying image (signal).
- For example, our results indicate that using the *support informed* sparse prior yields less uncertainty in homogeneous regions than typical Laplace priors do. Moreover, one can gain insight into regions that may require further interrogation.

Estimating the hyper-prior $f_{\boldsymbol{\lambda}}(\lambda)$.

We have $f_{\mathbf{X},\boldsymbol{\lambda}|\mathbf{Y}}(x,\lambda|y) \propto f_{\mathbf{Y}|\mathbf{X}}(y|x) f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda) f_{\boldsymbol{\lambda}}(\lambda)$.

The prior is $\tilde{f}_{\mathbf{X},\boldsymbol{\lambda}}(x,\lambda) = f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda) f_{\boldsymbol{\lambda}}(\lambda)$. We will begin with the hyper-prior $f_{\boldsymbol{\lambda}}(\lambda)$:

- It is non-trivial in practice to determine $f_{\boldsymbol{\lambda}}(\lambda)$.
- Within the empirical Bayesian analysis framework, we can approximate the hyper-prior as:

$$f_{\boldsymbol{\lambda}}(\lambda) \approx f_{\boldsymbol{\lambda}|\mathbf{Y}}(\lambda|y) = \frac{f_{\mathbf{Y}|\boldsymbol{\lambda}}(y|\lambda)\tilde{f}_{\boldsymbol{\lambda}}(\lambda)}{f(y)} \propto f_{\mathbf{Y}|\boldsymbol{\lambda}}(y|\lambda)\tilde{f}_{\boldsymbol{\lambda}}(\lambda),$$

  where $\tilde{f}_{\boldsymbol{\lambda}}(\lambda)$ is the prior on the hyper-prior.
- By assuming $\lambda \sim U[0,1]$ so that $\tilde{f}_{\boldsymbol{\lambda}}(\lambda) = 1$, we have

$$f_{\boldsymbol{\lambda}}(\lambda) \propto f_{\mathbf{Y}|\boldsymbol{\lambda}}(y|\lambda),$$

- Its mode can be approximated using the MAP estimate as

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\{f_{\mathbf{Y}|\boldsymbol{\lambda}}(y|\lambda)\} = \underset{\lambda}{\operatorname{argmax}} \int_{\mathbb{R}^n} f_{\mathbf{Y}|\mathbf{X}}(y|x) f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda)\,dx.$$

Estimating the hyper-prior density $f_{\boldsymbol{\lambda}}(\lambda)$.

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\{f_{\mathbf{Y}|\boldsymbol{\lambda}}(y|\lambda)\} = \underset{\lambda}{\operatorname{argmax}} \int_{\mathbb{R}^n} f_{\mathbf{Y}|\mathbf{X}}(y|x) f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda) dx.$$

- Integration with respect to x over the space $\mathbb{R}^n$ is not computationally feasible.
- We instead employ K-fold cross-validation to approximate $\hat{\lambda}$ given the data $y_j \in \Omega$ for $j = 1, \ldots, J$.
- The general process for K-fold cross-validation is to first calculate M MAP estimates of the unknown from a subset of $M < J$ observable measurements using sample candidates of the hyper-prior $\hat{\lambda}$.
- These are then used to generate M *training vectors*, which are in turn compared to the remaining $J - M$ observable measurements, or *testing vectors*.
- The whole process is then repeated for K independent trials, and $\hat{\lambda}$ is chosen to minimize the mean square error of $y = Ax$ between all training and testing vectors.

K fold cross validation to estimate $\hat{\boldsymbol{\lambda}}$.

Input: Set of multiple measurements (MMVs) $\mathbf{y} = [y_1, y_2, y_3, \ldots, y_J]$.
Output: Point Estimate of the hyper-prior $\hat{\lambda}$.
FOR $k = 1$ to K:

1. Randomly partition MMVs into M training vectors and $J - M$ testing vectors.
   FOR $i = 1$ to M:
   1. Sample candidate hyper-prior from $\tilde{\lambda}_{i,k} \sim U[0, 1]$.
   2. Calculate MAP estimate of unknown as $\hat{x}_{i,k} = \underset{x}{\operatorname{argmax}}\{f_{\mathbf{X},\boldsymbol{\lambda}|\mathbf{Y}}(x, \tilde{\lambda}_{i,k}|y_i)\}$.
   3. Compute training data according to $\hat{y}_{i,k} = A\hat{x}_{i,k}$.
   4. Evaluate the training data by calculating

$$E_{i,k} = \frac{1}{J - M} \sum_{l=M+1}^{J} \mathrm{MSE}(\hat{y}_{i,k}, y_l),$$

   where $\{y_l\}_{l=M+1}^{J}$ are the partitioned $J - M$ testing vectors.

Choose $(i^*, k^*) = \arg\min_{i,k} E_{i,k}$ and set $\hat{\lambda} = \tilde{\lambda}_{i^*,k^*}$.

K fold cross validation to estimate $\hat{\boldsymbol{\lambda}}$.

Remarks

- The hyper-prior $\hat{\lambda}$ recovered in the K fold cross validation process approximates the solution such that $f_{\boldsymbol{\lambda}}(\lambda) \approx \delta_{\hat{\lambda}}(\lambda)$.
- Intuitively, we are simply concentrating the density on the value $\hat{\lambda}$, which is our estimate of the mode of the hyper-prior distribution.
- The hyper-prior $\hat{\lambda}$ is consistent with how one ideally chooses the regularization parameter in compressive sensing approximations, since it should be chosen to offset the variance of the noise. That is, the K-fold cross validation method provides an empirical approach to calculating this variance.

A support informed sparsity prior

- The main goal is to devise a support informed sparsity prior $\tilde{f}_{\mathbf{X},\boldsymbol{\lambda}}(x,\lambda) = f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda) f_{\boldsymbol{\lambda}}(\lambda)$ for the posterior.
- We have the approximation $f_{\boldsymbol{\lambda}}(\lambda) \approx \delta_{\hat{\lambda}}(\lambda)$. Hence $\tilde{f}_{\mathbf{X},\boldsymbol{\lambda}}(x,\lambda) \approx f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\hat{\lambda})$.
- Therefore we need to determine $f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\hat{\lambda})$.
- We make two assumptions:
  1. The prior enforces the sparsity of some transformation of $\mathbf{X}$.
  2. There are multiple observations y of $\mathbf{Y}$ for each realization x of $\mathbf{X}$. (It is actually possible to relax this assumption.)

A support informed sparsity prior

- Sparsity is often enforced through the Laplace prior:

$$f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\hat{\lambda}) \propto \exp\{-\hat{\lambda}||\mathcal{L}x||_1\},$$

  where $\mathcal{L}$ is the transform operator used to approximate the sparse domain.
- In compressive sensing algorithms, a MAP estimate is constructed using this prior.
- Regardless of how the hyper-parameter $\hat{\lambda}$ is chosen, the Laplace prior does not actually provide the most information about what is known about the underlying signal.
- Specifically, it does not take into account the locations of the support in the sparse domain.
- The idea now is to use joint sparsity to obtain this information.

Variance based joint sparsity

Recall the definition:

Definition

A vector $v \in \mathbb{R}^n$ is s-sparse for some $1 \leq s \leq N$ if

$$\|v\|_0 = |\text{supp}(v)| \leq s.$$

A collections of vectors $v_1, \ldots, v_J \in \mathbb{R}^n$ is s-joint sparse if

$$\|V\|_{2,0} = \left| \bigcup_{j=1}^{J} \text{supp}\,(v_j) \right| \leq s,$$

where $V = [v_1 | \cdots | v_J]$.

The (spatial) entries of the sample variance $\nu \in \mathbb{R}^n$ across the rows of V is given by

$$\nu_i = \frac{1}{J} \sum_{j=1}^{J} V_{i,j}^2 - \left(\frac{1}{J} \sum_{j=1}^{J} V_{i,j} \right)^2, \quad i = 1, \ldots, n.$$

Variance based joint sparsity

The (spatial) entries of the sample variance $\nu \in \mathbb{R}^n$ across the rows of V is given by

$$\nu_i = \frac{1}{J}\sum_{j=1}^{J} V_{i,j}^2 - \left(\frac{1}{J}\sum_{j=1}^{J} V_{i,j}\right)^2, \quad i = 1, \ldots, n.$$

- ν_i small suggests that the corresponding sparse vector component is close to zero.
- ν_i large suggests that the corresponding sparse vector component is *not* close to zero, i.e. it is a place of *support*.
- In the weighted ℓ_p regularization (compressive sensing approach), we construct W so that $w_{ii} \propto \frac{1}{\nu_i}$.
- The sparse prior is heavily enforced in sparse regions. (It also allows a separation of scales in the underlying solution.)

Variance based joint sparsity

The (spatial) entries of the sample variance $\nu \in \mathbb{R}^n$ across the rows of V is given by

$$\nu_i = \frac{1}{J}\sum_{j=1}^{J} V_{i,j}^2 - \left(\frac{1}{J}\sum_{j=1}^{J} V_{i,j}\right)^2, \quad i = 1, \ldots, n.$$

- To adapt this concept for empirical Bayesian inference, we start with $w_{ii} \propto \frac{1}{\nu_i}$.
- But instead of using varying weight values, we generate a binary mask for a threshold $\tau \approx \frac{1}{n}$:

$$m_i = \begin{cases} 1, & w_i \geq \tau \\ 0, & w_i < \tau. \end{cases}$$

- This allows us to still apply the hyper-prior $\hat{\lambda}$ as determined by the K fold cross validation.

Variance based joint sparsity mask construction

Consider the following signal for $0 \leq s \leq 1$:

$$x(s) = \begin{cases} 40, & 0.1 \leq s \leq 0.25 \\ 10, & 0.35 \leq s \leq 0.325 \\ 20\sqrt{2\pi}e^{-\left(\frac{s-0.75}{0.05}\right)^2}, & s > 0.5 \\ 0, & \text{otherwise.} \end{cases}$$

Variance based joint sparsity mask construction

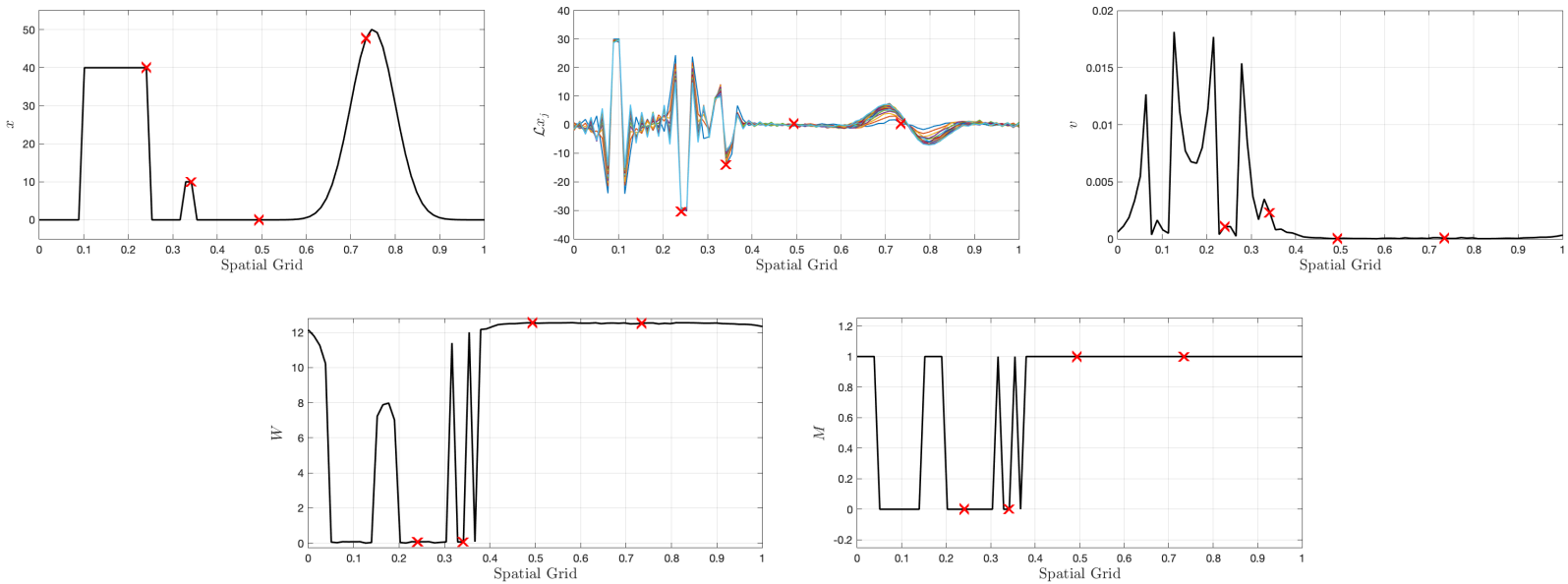

Figure: (top-left) True function. (top-middle) Jointly sparse vectors generated from indirect, noisy measurements $y_j, j = 1, \ldots, 10$, of x. (top-right) Spatial variance ν across joint sparsity vectors. (bottom-left) $w_{ii} \propto \frac{1}{\nu_i}$. (bottom-right) Mask generated for use in the support informed sparsity prior. The red 'x's indicate locations where we calculate evaluation metrics.

A support informed sparsity prior

Input: Set of measurement vectors, $\vec{y} = [y_1, y_2, \ldots, y_J]$. Threshold $\tau \approx \frac{1}{n}$
Output: Support informed sparsity prior.

1. Approximate sparse domain vectors $\check{v}_j, j = 1, \ldots, J$, from the given measurements.
2. Estimate the hyper-parameter $\hat{\lambda}$ using K fold cross validation.
3. Compute the variance ν of $\check{v}_j, j = 1, \ldots, J$.
4. Construct weighting W with diagonal $\mathbf{w} = w_1, \ldots, w_n$ using $w_i \propto \frac{1}{\nu_i}$
5. Construct the binary mask M using M according to chosen threshold τ.
6. Define the support informed sparsity prior density as

$$\tilde{f}_{\mathbf{X}|\hat{\boldsymbol{\lambda}},M}(x|\hat{\lambda}, M) = C \exp\{-\hat{\lambda}||M\mathcal{L}x||_1\}.$$

Defining the posterior for the model $\mathbf{Y} = A\mathbf{X} + \mathbf{E}$.

- Bayes' theorem: $f_{\mathbf{X},\boldsymbol{\lambda}|\mathbf{Y}}(x,\lambda|y) = \propto f_{\mathbf{Y}|\mathbf{X}}(y|x) f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda) f_{\boldsymbol{\lambda}}(\lambda)$.
- The likelihood (for $\mathbf{E} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$) is

$$f_{\mathbf{Y}|\mathbf{X}}(y|x) = f_{\mathbf{E}}(y - Ax) \propto \exp\{-\frac{1}{2\sigma^2}||y - Ax||_2^2\}.$$

- The prior is $\tilde{f}_{\mathbf{X},\boldsymbol{\lambda}}(x,\lambda) = f_{\mathbf{X}|\boldsymbol{\lambda}}(x|\lambda) f_{\boldsymbol{\lambda}}(\lambda)$, where $f_{\boldsymbol{\lambda}}(\lambda)$ is the hyper-prior.
- We consider two posterior estimates, both using the hyperparameter $\hat{\lambda}$ calculated by K fold cross validation.

$$\hat{f}_{\mathbf{X},\boldsymbol{\lambda}|\mathbf{Y}}(x,\hat{\lambda}|y) = C_L \exp\{-\hat{\lambda}||\mathcal{L}x||_1 - \frac{1}{2\sigma^2}||y - Ax||_2^2\}, \qquad C_L > 0,$$

$$\hat{f}_{\mathbf{X},\boldsymbol{\lambda}|\mathbf{Y}}(x,\hat{\lambda}|y) = C_M \exp\{-\hat{\lambda}||M\mathcal{L}x||_1 - \frac{1}{2\sigma^2}||y - Ax||_2^2\}, \qquad C_M > 0.$$

The first uses the Laplace prior and the second uses the support informed sparsity prior constructed using *variance ased joint sparsity*. C_L and C_M are not explicitly known but are not needed for calculation.

Evaluating the posterior

- Point estimation via MAP estimate:

$$x^j_{MAP} \equiv \underset{x\in\mathbb{R}^n}{\text{argmax}} \left\{ \log \left[\hat{f}_{\mathbf{X},\boldsymbol{\lambda}|\mathbf{Y}}(x, \hat{\lambda}|y_j) \right] \right\} \quad j = 1, \ldots, J.$$

A corresponding convex optimization problem is then easily derived as

$$\begin{aligned} x^j_{MAP} &= \underset{x\in\mathbb{R}^n}{\text{argmax}} \left\{ \log(C) - \left(\hat{\lambda}||\mathcal{T}x||_1 + \frac{1}{2\sigma^2}||y_j - Ax||_2^2 \right) \right\} \\ &= \underset{x\in\mathbb{R}^n}{\arg\min} \left\{ \hat{\lambda}||\mathcal{T}x||_1 + \frac{1}{2\sigma^2}||y_j - Ax||_2^2 \right\}, \quad j = 1, \ldots, J, \end{aligned}$$

where $\mathcal{T} = \mathcal{L}$ or $\mathcal{ML}$ and $C = C_L$ or C_M according to the chosen posterior.

Evaluating the posterior

- Recover samples from the posterior using Metropolis Hasting (MH) Markov Chain Monte Carlo (MCMC)
  - We choose the proposal distribution to be Gaussian: $x^{cand} \sim \mathcal{N}(x^{k-1}, \sigma_p)$
  - The acceptance ratio (for the symmetric proposal) is then

$$\alpha(x^{cand}|x^{k-1}) = \min\left\{1, \frac{\hat{f}_{\mathbf{X},\boldsymbol{\lambda}|\mathbf{Y}}(x^{cand}, \hat{\lambda}|y)}{\hat{f}_{\mathbf{X},\boldsymbol{\lambda}|\mathbf{Y}}(x^{k-1}, \hat{\lambda}|y)}\right\}.$$

  - We used burn in rate of $25,000$ and total length of chain $50,000$ which we found in other comparable cases.

Numerical Experiments: Credibility intervals

For these experiments, $\mathbf{E} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$ with $\sigma = 6$.

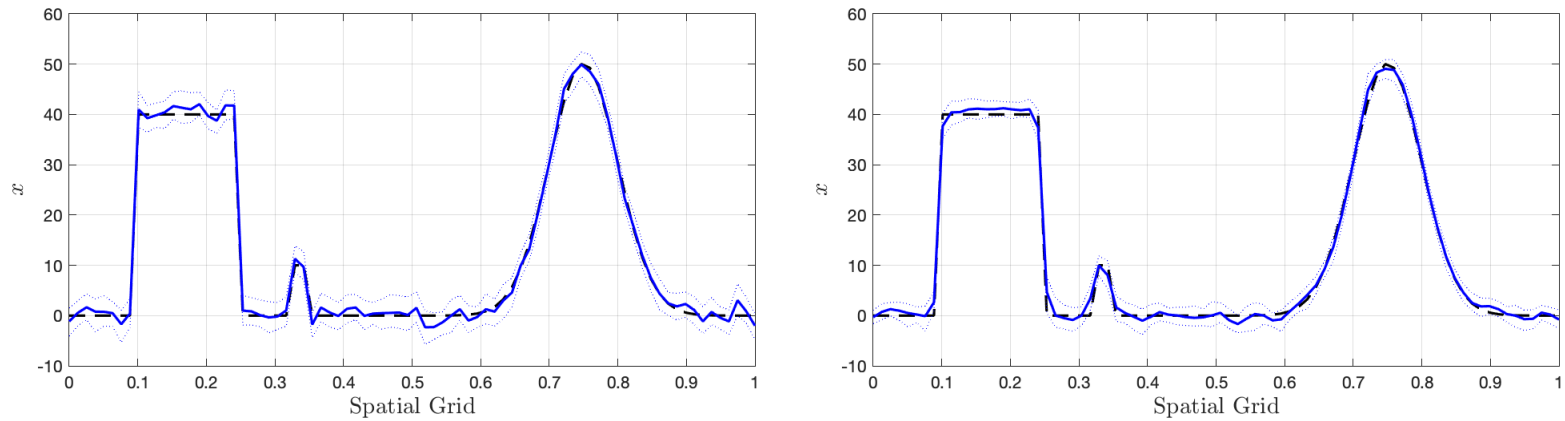

Figure: Approximate reconstructions (mean of the Markov chain) displayed along with credibility intervals $C_{0.05}$ when using (left) the Laplace prior and (right) the proposed support informed sparsity prior.

Numerical Experiments: Acceptance Ratio

The support informed sparse prior yields an acceptance ratio near .5.

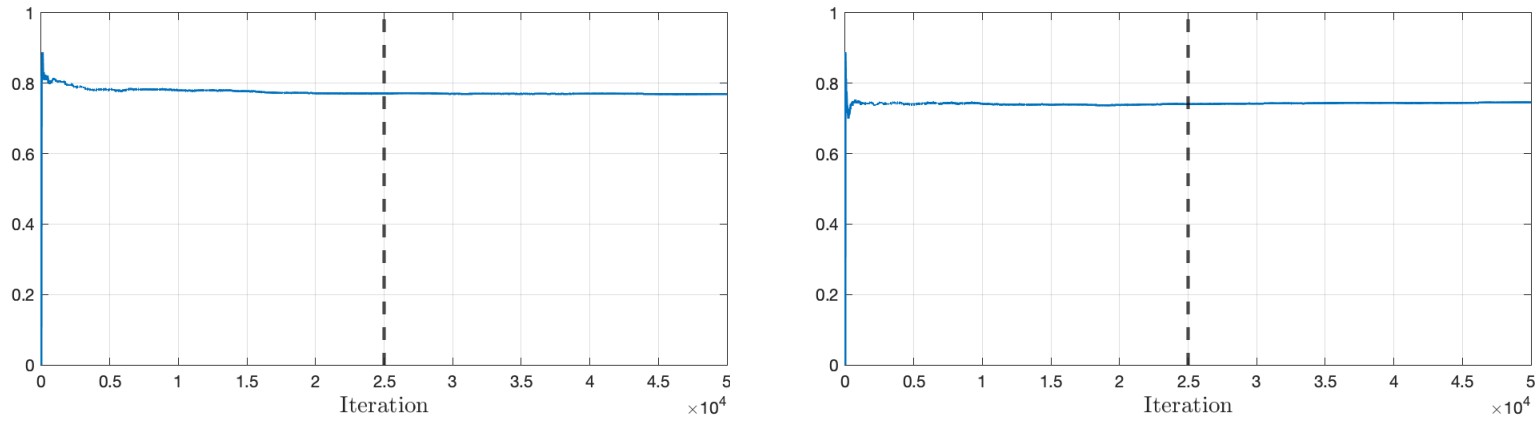

Figure: Acceptance ratio plots resulting from sampling the posterior defined using (left) Laplace prior and (right) support informed sparse prior.

Numerical Experiments: Correlograms

Results at different points in the domain.

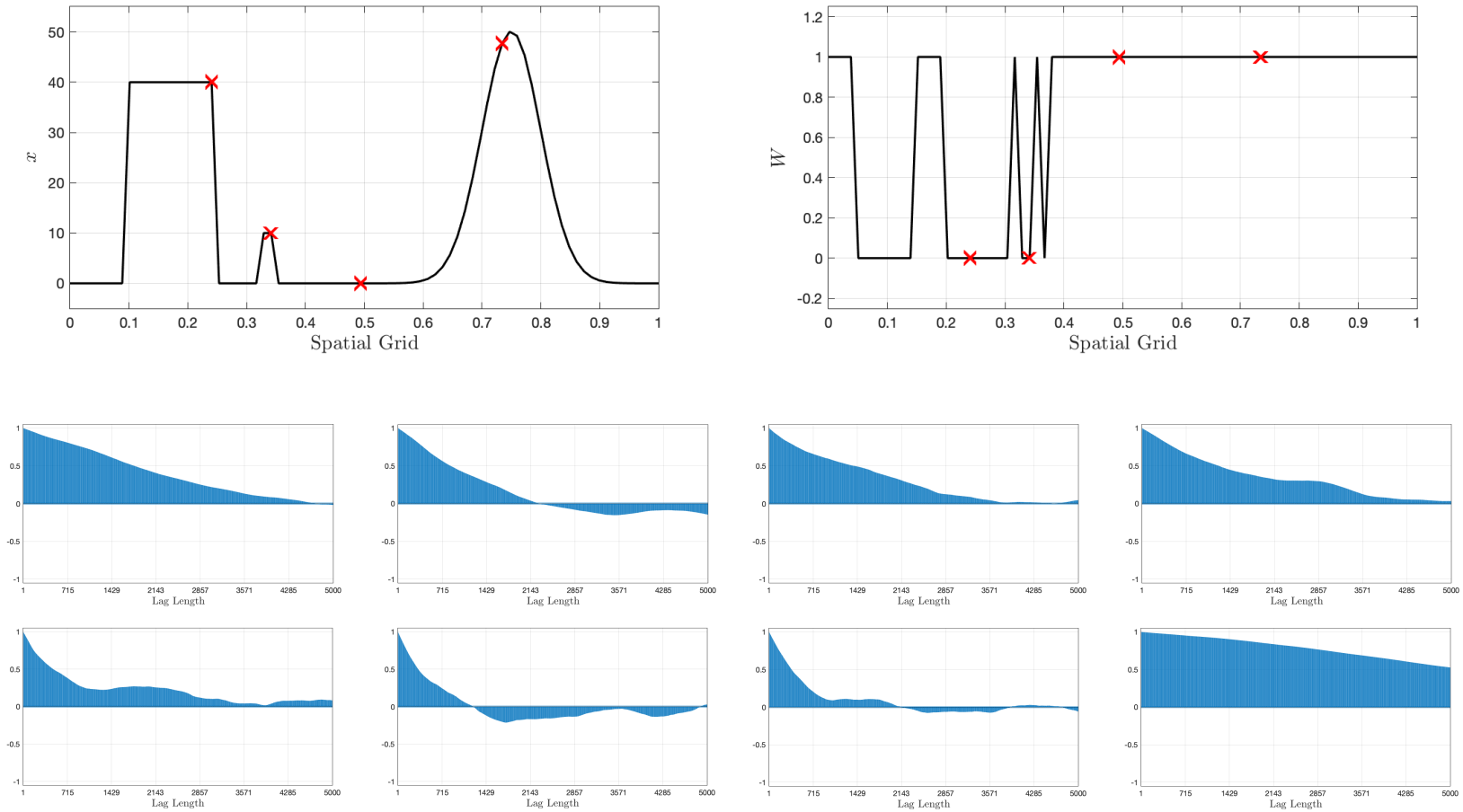

Figure: The corellograms display the autocorrelation function (ACF) calculated for various lag lengths at the points in the domain indicated by x in the above graphs. Samples are taken from the posterior distributions defined using the (top) Laplace prior and (bottom) support informed sparsity prior.

Numerical Experiments: Trace plots

Results at different points in the domain.

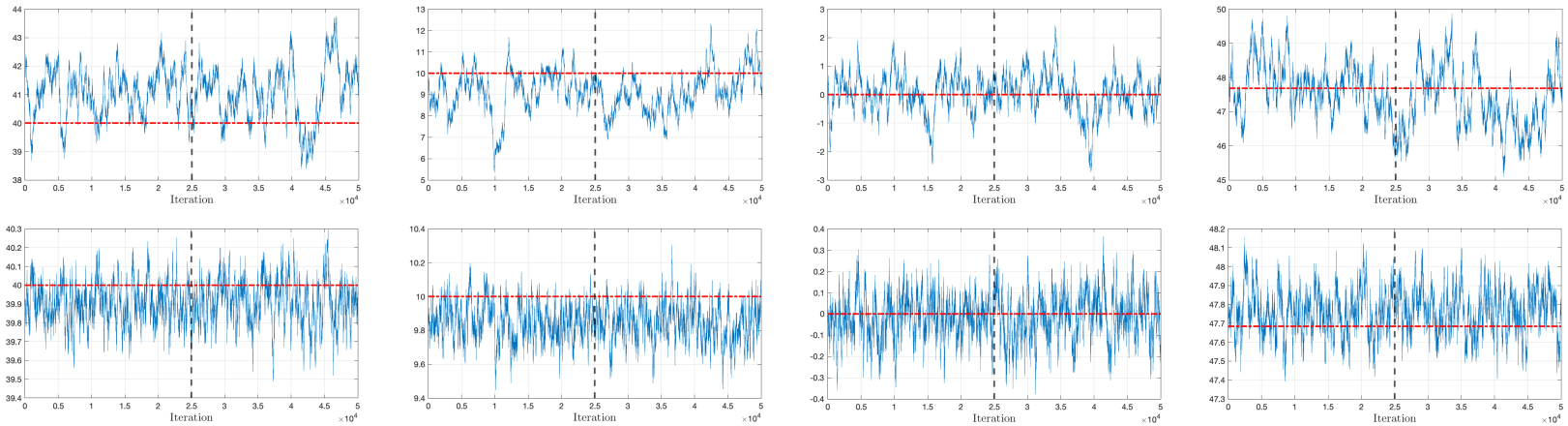

Figure: Trace plots calculated at the points described by x in above graphs. We display the samples found using the posterior with (top) the Laplace prior and (bottom) the support informed sparsity prior.

Numerical Experiments: Relative and absolute errors

We calculate the relative and absolute errors for increasing levels of noise.

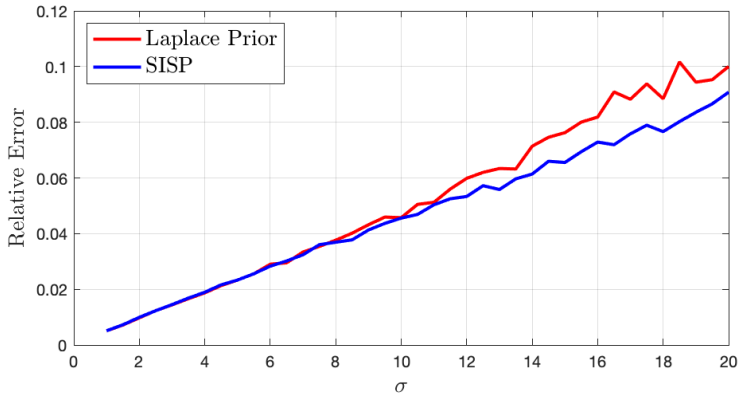

Figure: The relative reconstruction error calculated across the entire spatial domain given noisy measurements.

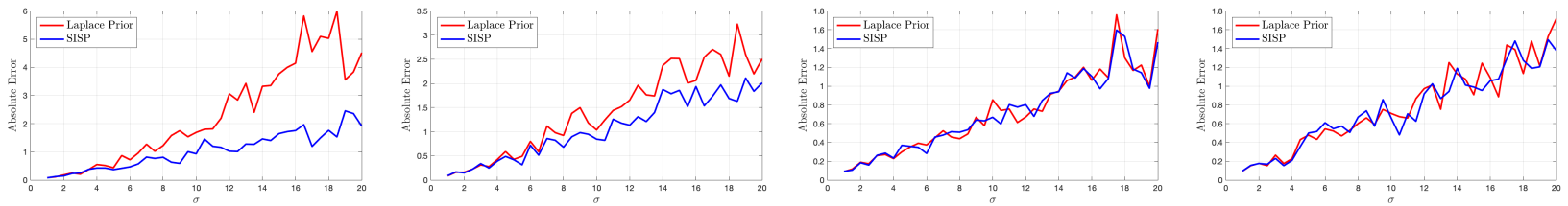

Figure: Absolute error at the x values.

Thoughts moving forward

- We compared our technique to the one described in the sequence of papers by Bardsley et. al. (In fact that is where we got our model).
- There is also some relationship to SBL methods which we can discuss.
- We have some two dimensional results, but we are still working some things out.

Thoughts moving forward

- There are clearly some benefits in using a support informed sparse prior.
- However, the support locations are not the only features that matter, and we have to be careful how we characterize the prior as a result.
- On the other hand, it is also apparent that we can combine methodology (as is) to give us more insight into what is going on in different regions.
- In these examples we used sparsely sampled Fourier data, but we can use different sources of data, and combine different sources of data, provided that we have a good way of determining the support (or some other feature).