Bayesian hierarchical models: convexity, sparsity and model reduction

Daniela Calvetti
Case Western Reserve University

based on joint work with Monica Pragliola, Erkki Somersalo and Alexander Strang

Supported in part by NSF grants DMS-1522334.

Reconstruction Methods for Inverse Problems
Banff, June 27, 2019

Problem statement

Want to reconstruct $x \in \mathbb{R}^n$ from few indirect, noisy observations.
In the case of a linear observation model

$$b = \mathsf{A}x + e, \quad \mathsf{A} \in \mathbb{R}^{m \times n}, \quad m \ll n.$$

Assume that

- additive Gaussian noise e; where $E \sim \mathcal{N}(0, \mathsf{I}_m)$
- x is believed to be sparse, i.e.,

$$\|x\|_0 \ll n.$$

- or to admit a sparse representation

$$x = \mathsf{L}z, \quad \|z\|_0 \ll n.$$

Sparsity Considerations

Sparsity means a signal with a *sparse representation*

- The sparse vector in that case contains the coefficients of a suitable representation, for example
- Wavelet basis
- Fourier basis
- First order differencing matrix for piecewise constant signals in terms of their increments

The conditionally Gaussian random variable is the presumably sparse coefficient vector.

Sparsity promotion via hierarchical model

- Conditionally Gaussian prior for sparse object

$$X \sim \mathcal{N}(0, \mathsf{D}_\theta), \quad \mathsf{D}_\theta = \mathrm{diag}(\theta_1, \ldots, \theta_n),$$

$$\pi_{x|\theta}(x \mid \theta) = \frac{1}{(2\pi)^{n/2}\sqrt{\theta_1 \cdots \theta_n}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n} \frac{x_j^2}{\theta_j}\right).$$

- Mutually independent unknown prior variances $\theta_j > 0$ follow generalized gamma distributions,

$$\Theta_j \sim \mathrm{GenGamma}(r, \vartheta_j, \beta), \quad \pi_{\Theta_j}(\theta_j) = \frac{1}{\Gamma(\beta)\vartheta_j}\left(\frac{\theta_j}{\vartheta_j}\right)^{r\beta-1} \exp\left(-\frac{\theta_j}{\vartheta_j}\right)^r.$$

- Posterior density

$$\pi_{X,\Theta|B}(x, \theta) \propto \exp\left(-\frac{1}{2}\|b - \mathsf{A}x\|^2 - \frac{1}{2}\sum_{j=1}^{n}\frac{x_j^2}{\theta_j} + \eta\sum_{j=1}^{n}\log\frac{\theta_j}{\vartheta_j} - \sum_{j=1}^{n}\left(\frac{\theta_j}{\vartheta_j}\right)^r\right)$$

where $\eta = r\beta - 3/2 > 0$.

Iterated Alternating Sequential (IAS) algorithm

To compute x_{MAP} we minimize the Gibbs energy

$$\mathscr{E}(x;\theta) = \overbrace{\frac{1}{2}\|b - \mathsf{A}x\|^2 + \sum_{j=1}^{n} \frac{x_j^2}{2\theta_j}}^{(a)} \underbrace{- \sum_{j=1}^{n} \left(\eta \log \frac{\theta_j}{\vartheta_j} - \left(\frac{\theta_j}{\vartheta_j}\right)^r \right)}_{(\mathscr{P}(x,\theta))} \qquad (1)$$

Given the initial value $\theta^0 = \vartheta$, $x^0 = 0$, and $k = 0$, iterate until convergence:

(a) Update $x^k \to x^{k+1}$ by minimizing $\mathscr{E}(x \mid \theta^k)$;

(b) Update $\theta^k \to \theta^{k+1}$ by minimizing $\mathscr{E}(\theta \mid x^{k+1})$.

IAS algorithm for Generalized Gamma hyperpriors

1. Given θ, $x_{k+1} = \operatorname{argmin}\left\{\|b - \mathsf{A}x\|^2 + \|\mathsf{D}_\theta^{-1/2} x\|^2\right\}$ solves

$$\begin{bmatrix} \mathsf{A} \\ \mathsf{D}_\theta^{-1/2} \end{bmatrix} x = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

   in the least squares sense.

2. The update of θ is componentwise. From the first order optimality condition θ_j must satisfy

$$-\frac{1}{2}\frac{x_j^2}{\theta_j^2} - \left(r\beta - \frac{2}{3}\right)\frac{1}{\theta_j} + r\frac{\theta_j^{r-1}}{\vartheta_j^r} = 0, \; x_j = x_j^{t+1}.$$

Convexity and Convergence: $r = 1$

For the gamma hyperprior ($r = 1$):

- The Gibbs energy functional $\mathscr{E}$ is strictly convex and has a unique minimizer
- In exact arithmetic, the IAS algorithm converges to the global minimizer
- For $\eta > 0$ small, the Gibbs energy (1) is approximately equal to the penalized least squares functional with a weighted ℓ_1-penalty.

.

Theorem

For a gamma hyperprior, the exact IAS algorithm converges to the unique minimizer $(\widehat{x}, \widehat{\theta})$ of the Gibbs energy functional. Moreover, the minimizer $(\widehat{x}, \widehat{\theta})$ satisfies the fixed point condition

$$\widehat{x} = \mathrm{argmin}\left\{\mathscr{E}\left(x \mid F(x)\right)\right\}, \quad \widehat{\theta} = F(\widehat{x}),$$

where F is the map with jth component f_j. [1].

[1] Calvetti D, Pascarella A, Pitolli F, Somersalo E, Vantaggi B (2015) A hierarchical Krylov–Bayes iterative inverse solver for MEG with physiological preconditioning. Inverse Problems 31:125005

Scale parameter and sparsity: $r = 1$

Under the assumptions of our hierarchical Bayesian model we have shown that

- The exact IAS iteration converges to the global minimizer of the functional

$$\mathscr{L}_\eta(x) = \mathscr{E}(x, f(x))$$

and, for small $\eta > 0$

- $$\mathscr{L}_\eta(x) = \mathscr{L}_1(x) + \underbrace{\eta g(x, \eta)}_{\to 0 \text{ as } \eta \to 0},$$

where

$$\mathscr{L}_1(x) = \frac{1}{2}\|b - \mathsf{A}x\|^2 + \sqrt{2}\sum_{j=1}^{n} \frac{|x_j|}{\sqrt{\vartheta_j}}.$$

and the sum extends only over the support of x,

$$S = \mathrm{supp}(x) = \{j \mid x_j \neq 0\}.$$

ℓ_2 Stable Recovery: $r = 1$

$$\underbrace{x_\eta = \text{argmin}\left\{\mathscr{L}_\eta(x)\right\}}_{=IAS\ \ solution} \qquad \underbrace{x_1 = \text{argmin}\left\{\mathscr{L}_1(x)\right\}}_{=\ell_1 penalized\ \ solution}.$$

1. The size of $x_\eta - x_1$ depends continuously on η. Thus η controls the sparsity of the solution.
2. If A is of the kind for which the ℓ_1-magic works and the data come from a sparse vector[2], then x_η is close to the underlying sparse solution.
3. The scale parameters ϑ_j play the role of sensitivity weights in inverse problems: Data components may have different sensitivity to different components x_j.

[2] Candes E, Romberg JK and Tao T(2006): Stable Signal Recovery from Incomplete and Inaccurate Measurements, Comm Pure Appl Math LIX: 1207–1223

Sparsity and exchangeability

Assume the underlying signal x is sparse $\mathrm{supp}(x) = \mathsf{S} \subset \{1, 2, \ldots, n\}$ and b_0 is the noiseless measurement. Define

$$\mathrm{SNR}_{\mathsf{S}} = \frac{E\left\{\|b_0\|^2 \mid \mathrm{supp}(x) = \mathsf{S}\right\}}{E\left\{\|e\|^2\right\}}, \; e \sim \mathcal{N}(0, \Sigma).$$

Lemma

With our assumptions about X and the noise

$$\mathrm{SNR}_{\mathsf{S}} = \frac{\sum_{j \in \mathsf{I}} \nu(r, \beta) \vartheta_j \|\mathsf{A} e_j\|^2}{\mathrm{tr}\,(\Sigma)} + 1, \; \nu(r, \beta) = \frac{\Gamma(\beta + 1) r}{\Gamma(\beta r)}.$$

Proof.

$$E\left\{\|b_0\|^2\right\} = \mathrm{Tr} E\left\{b_0 b_0^{\mathrm{T}}\right\} = \mathrm{Tr} E\left\{\mathsf{A} x x^{\mathrm{T}} \mathsf{A}^{\mathrm{T}}\right\} = \mathrm{Tr}\big(\mathsf{A} E\left\{x x^{\mathrm{T}}\right\} \mathsf{A}^{\mathrm{T}}\big),$$

and from the generalized gamma hyperprior

$$E\left\{x x^{\mathrm{T}}\right\} = E_\theta\left\{E\left\{x x^{\mathrm{T}} \mid \theta\right\}\right\} = E\big(\mathrm{diag}(\theta)\big) = \mathrm{diag}\big(\nu(r, \beta)\vartheta\big).$$

Scale parameter and sensitivity scaling, in Bayesian way

How should ϑ be chosen?

Theorem

Given an estimate $\overline{\mathrm{SNR}}$ *of SNR, if*

$$P\big(\|x\|_0 = k\big) = p_k, \quad p_0 = p_n = 0, \quad \sum_{k=1}^{n} p_k = 1$$

and if

$$\mathrm{SNR}_S = \mathrm{SNR}_{S'}, \quad \forall\, S, S' : \mathrm{card}(S) = \mathrm{card}(S'),$$

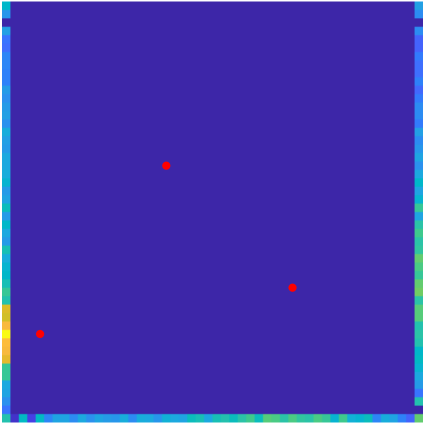
then

$$\vartheta_j = \frac{C}{\|Ae_j\|^2}, \quad C = \frac{(\overline{\mathrm{SNR}} - 1)\,\mathrm{Tr}(\Sigma)}{\nu(r,\beta)} \sum_{j=1}^{n} \frac{p_k}{k}$$

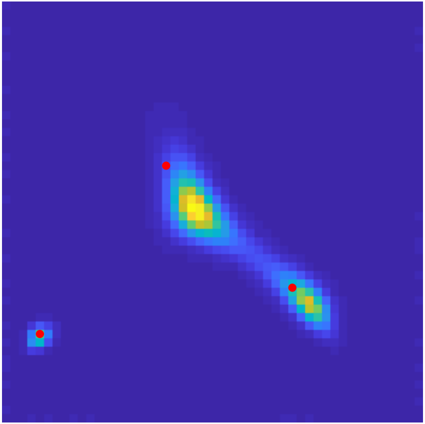
In the literature $\|Ae_j\|$ is the *sensitivity* of the data to jth component of x.

Sensitivity can make a difference.

Without sensitivity



With sensitivity



Sparsity and quadratic convergence: $r = 1$

For the gamma hyperprior, as η goes to zero the sequence of IAS minimizers remains bounded.

Lemma

There is a constant $B > 0$ such that

$$\|x_\eta\| \leq B,$$

for all η, $0 \leq \eta \leq \frac{1}{2}$.

Theorem

If the matrix A *is such that the minimizer*

$$x_1 = \operatorname{argmin}\{F_1(x)\}$$

of the ℓ_1-penalized functional F_1 is unique, then, as $\eta \to 0+$, the minimizers x_η converge to the minimizer x_1.

Intermezzo: Sparse or compressible?

- **Sparsity**
  If A is a matrix such that the ℓ_1 regularized solution x_1 is sparse, then the solution of the IAS algorithm with $\eta > 0$ small can be made arbitrarily small outside the support of x_1.
- **Compressibility**
  If the components of x_1 are smaller than a threshold outside a set $S \subset \{1, 2, \ldots, n\}$, the same is true for the IAS solution x_η with a slightly larger threshold when $\eta > 0$ is small enough.
- **Bayesian Sparsity is Compressibility**
  The Bayesian target reconstruction of a sparse signal is a compressible signal.

Convergence of IAS for $r = 1$

Theorem

In the IAS algorithm, the updates of x converge at least $\widehat{\theta}$-linearly, that is, linearly in the Mahalanobis norm

$$\|x\|_{\widehat{\theta}}^2 = x^{\mathsf{T}} \mathsf{D}_{\widehat{\theta}}^{-1} x$$

evaluated at the MAP estimate. Moreover, if $\mathrm{supp}(\widehat{x}) \subsetneq \{1, 2, \ldots, n\}$, the convergence of θ in the complement of the support is quadratic[3].

[3] D. Calvetti, E.Somersalo and A. Strang. Hierachical Bayesian models and sparsity: ℓ_2 -magic. Inverse Problems 35: 035003.

Generalized gamma hyperpriors and IAS[4]

For the family of generalized gamma hyperpriors for sparse recovery we want to investigate the

- Convexity - or lack thereof - of Gibbs functional
- Form and behavior of θ update
- Type of regularization effect on components
- Similarity with classical regularization functionals
- Role of r and shape parameter.

Non-dimensionalization:

- WLOG we assume that $\vartheta_j = 1$ or, equivalently,
- scale x_j by $\sqrt{\vartheta_j}$ and θ_j by ϑ_j.

[4] D. Calvetti, M.Pragliola, E. Somersalo and A. Strang. Sparse reconstructions from few noisy data via hierarchical Bayesian models with generalized gamma hyperpriors: convergence, convexity and performance. Manuscript.

The θ update as a function of r

For generalized gamma hyperpriors, the function $\theta_{k+1} = f(x_{k+1})$ is the unique solution of the IVP:

$$\frac{d}{dx} f(x) = \frac{2x f(x)}{2r^2 f(x)^{r+1} + x^2}, \quad f(0) = \left(\frac{\eta}{r}\right)^{\frac{1}{r}}, \; x > 0$$

and $f(x) = f(-x)$. Moreover, f is

- Monotonically increasing and unbounded above
- Asymptotically, when $|x|$ is small

$$f(x) \propto \left(\frac{\eta}{r}\right)^{\frac{1}{r}} + \frac{1}{2\eta r} x^2$$

- Asymptotically, when $|x|$ is large

$$\begin{aligned} f(x) &\propto |x|^p, \; p = \frac{2}{r+1} & r > 0 \\ f(x) &\propto x^2 & r < 0, \end{aligned}$$

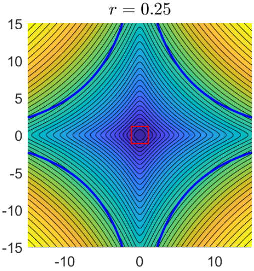
with growth linear for $r = 1$, less than linear $r > 1$, quadratic $r < 1$.

Effective local penalty functional

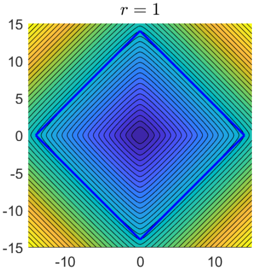
- Shape parameter determines initial value $f(0)$
- Shape parameter does not affect variance of large $|x|$

$$\mathscr{P}_j(x_j \mid \theta_j) = \frac{x_j^2}{2\theta_j} - \eta \log \theta_j + (\theta_j)^r)$$

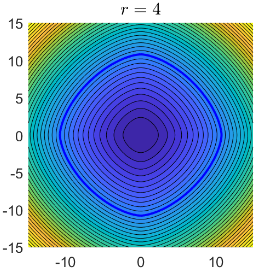
- For small $|x_j|$: $\mathscr{P}_j(x)$ is quadratic in $|x|$;
- For large $|x_j|$: $\mathscr{P}_j(x)$ is proportional to
  - $|x_j|^p, p = \frac{2r}{1+r}, \; r > 0$
  - $\log |x_j|, \; r < 0.$
- When $r = 2, \; p = 4/3$.
- When $r = 1$ $p = 1$, thus ℓ_1-like penalty.
- When $0 < r < 1, \; p < 1$ and the penalty strongly enforces sparsity.


$r = 0.25$
15
10
5
0
-5
-10
-15
-10
0
10

p = 2/5


$r = 1$
15
10
5
0
-5
-10
-15
-10
0
10

p = 1


$r = 4$
15
10
5
0
-5
-10
-15
-10
0
10

p = 8/5

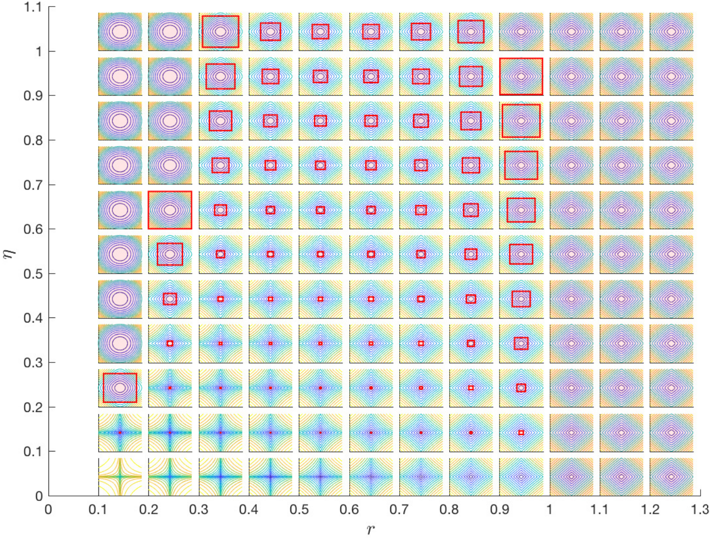
Convexity, Sparsity and Penalization

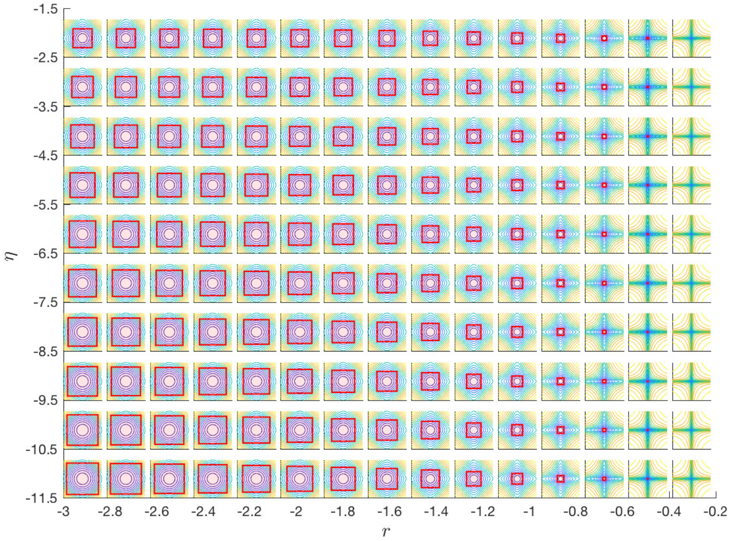
The Gibbs functional $\mathscr{E}(x,\theta)$ is convex

- for all x, θ if $r \geq 1$ and $\eta > 0$
- for all x, $\theta < \overline{\theta} = \left(\frac{\eta}{r(1-r)}\right)$ if $r < 1$

Convexity region:

- Let $\overline{x} = f^{-1}(\overline{\theta})$. The convexity region is all $x: \ \|x\|_\infty < \overline{x}$.
- The radius of the convexity region $\overline{x}$ increases monotonically with η
- η is proportional to the radius of the convexity region centered at origin


1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
η
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
r


-1.5
-2.5
-3.5
-4.5
-5.5
-6.5
-7.5
-8.5
-9.5
-10.5
-11.5
η
-3
-2.8
-2.6
-2.4
-2.2
-2
-1.8
-1.6
-1.4
-1.2
-1
-0.8
-0.6
-0.4
-0.2
r

Support of the signal: the meaning of θ

In light of the Bayesian set up:

- The entries of x with large variance are more likely to contain large values
- The prior variance of x_j is θ_j
- The entries of θ above a threshold identify the support of the signal
- The more sparsity promoting the hyperprior, the more θ greedy the IAS

At each IAS iteration, the system learns the support of the signal and uses it to improve the reconstruction.

IAS with bound constraints

The IAS method can be modified to include bounds on the entries of the solution.

- Assume we believe
$$0 < x_j < H$$
- Define
$$G(x) = \begin{cases} 0, & \text{when } 0 < x \leq H, \\ \infty & \text{otherwise,} \end{cases}$$
- Write posterior density with the bound constraints as
$$\pi(x, \theta \mid b) \propto \exp\left(-\mathscr{E}(x,\theta) - G(x)\right) = \exp\left(-\mathscr{E}_G(x,\theta)\right).$$

Moreau-Yoshida envelope and box contraints

- Consider the Moreau-Yoshida envelope

$$\Phi_G^\lambda(x,\theta) = \mathscr{E}(x,\theta) + G^\lambda(x),$$

where

$$G^\lambda(x) = \min_{u\in\mathbb{R}^n}\left\{G(u) + \frac{1}{2\lambda}\|x-u\|^2\right\},\ \lambda > 0.$$

- The Moreau-Yoshida envelope is differentiable and

$$\nabla_x\Phi_G^\lambda(x,\theta) = \nabla_x\mathscr{E}(x,\theta) + \frac{1}{\lambda}(x - \mathrm{prox}_G^\lambda(x)),$$

where the proximal operator is

$$\begin{aligned}\mathrm{prox}_G^\lambda(x) &= \mathrm{argmin}_{u\in\mathbb{R}^n}\left\{G(u) + \frac{1}{2\lambda}\|x-u\|^2\right\}\\ &= \left\{\begin{array}{ll} x, & \text{if } G(x) = 0,\\ Pz, & \text{if } G(x) = \infty. \end{array}\right.,\end{aligned}$$

and P is the orthogonal projector on the feasible set $[0,H]^n$.

What is the Moreau-Yoshida envelope doing for us?

It has been shown that

- as $\lambda \to 0+$,
- the posterior distribution in terms of the Moreau-Yoshida envelope
- converges to the posterior distribution + positivity constraint.

IAS with bound constraint

- The inclusion of the bounds does not change ∇_θ,
- The IAS algorithm can be extended for bound constrained problems
- Replace the least squares minimization by the sequential procedure:
- Given the current θ^t:
  (a) Find $x = x^*$ solving $\nabla_x \mathscr{E}(x, \theta^t) = 0$ in the least squares sense,
  (b) Update $x^{t+1} = \mathrm{prox}_G^\lambda(x^*)$ by projecting x^* onto the feasible set.

Approximate IAS and reduced model

In the case where $\mathrm{A} \in \mathbb{R}^{m\times n}$, $m < n$ at each IAS step, instead of solving

$$\begin{bmatrix} \mathrm{A} \\ \mathrm{D}_\theta^{-1/2} \end{bmatrix} x = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

solve approximately

$$\mathrm{A}\mathrm{D}_\theta^{1/2} w = b, \quad x = \mathrm{D}_\theta^{1/2} w$$

with the CGLS methods equipped with stopping rule.

- Each CGLS iteration requires only 1 matvec with A and one with A′
- If θ_j is small, the corresponding column of $\mathrm{A}\mathrm{D}_\theta^{1/2}$ is almost deflated
- Equivalently, the corresponding solution entry is made smaller
- The more sparsity promoting the prior, the fewer the large θ_j

Three computed examples

- Example 1: Deconvolution of one dimensional staircase signal blurred with Airy kernel[5]. Exact and CGLS-AS
- Example 2: Reconstruction of two dimensional nearly black object recovery from blurred, noisy data (Gaussian blur). Exact and CGLS-IAS
- Example 3: Limited angle computed tomography problem. CGLS-IAS only.

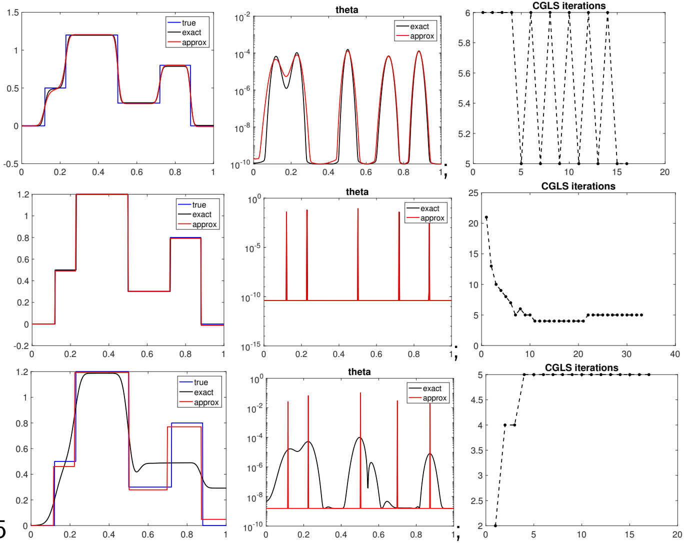
[5] $\frac{J(k|x|)}{k|x|}$

$r = 1$, $r = 0.5$ and $r = -0.5$ with 1% noise

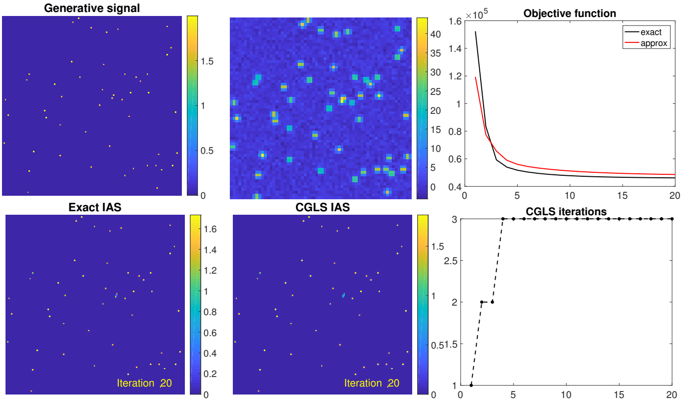
$r = 1$

$r = 0.5$

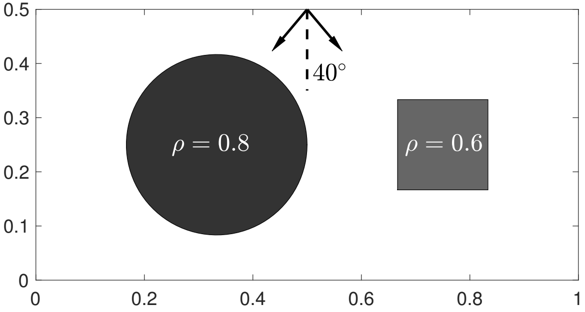
$r = -0.5$

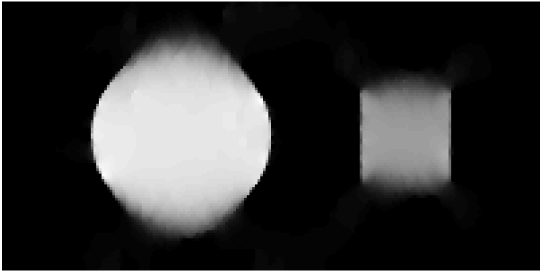


Starry night: $r = 1$

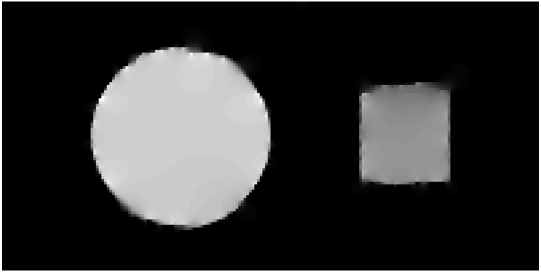


Limited angle tomography





$r = 1$



$r = 0.5$

Horizontal and vertical profiles, and CGLS steps