Regularization Methods

An Applied Mathematician's Perspective

Curt Vogel

Montana State University

Department of Mathematical Sciences

Bozeman, MT 59717-2400

Outline

- Well (and Ill-) Posedness
- Regularization
  - Optimization approach (Tikhonov)
    - Bayesian connection (MAP)
  - Filtering approach
  - Iterative approach
- Regularization Parameter Selection
- Applied Math Wish List

Focus on linear problems.

Well-Posedness

Definition due to Hadamard, 1915: Given mapping $A : X \to Y$, equation

$$A\mathbf{x} = \mathbf{y}$$

is well-posed provided

- (Existence) For each $\mathbf{y} \in Y$, $\exists \mathbf{x} \in X$ such that $A\mathbf{x} = \mathbf{y}$;
- (Uniqueness) $A\mathbf{x}_1 = A\mathbf{x}_2 \Rightarrow \mathbf{x}_1 = \mathbf{x}_2$; and
- (Stability) A^{-1} is continuous.

Equation is ill-posed if it is not well-posed.

Linear, Finite-Dimensional Case

$A : \mathbb{R}^n \to \mathbb{R}^n$ ($n \times n$ matrix).

$$A\mathbf{x} = \mathbf{y} \text{ well-posed} \iff \begin{cases} A^{-1} \text{ exists} \\ \det A \neq 0 \\ A\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{x} = \mathbf{0} \\ \quad\vdots \end{cases}$$

Existence imposed by considering least squares solutions

$$\mathbf{x}_{\mathrm{LS}} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} ||A\mathbf{x} - \mathbf{y}||^2.$$

Uniqueness imposed by taking the min norm least squares solution

$$\mathbf{x}_{\mathrm{LSMN}} = \arg\min\{||\mathbf{x}_{\mathrm{LS}}||\} = A^{\dagger}\mathbf{y}.$$

Infinite-Dimensional Example

(Compact) diagonal operator on (Hilbert) space ℓ^2

$$\mathbf{x} = (x_1, x_2, \ldots, x_n, \ldots) \in \ell^2 \quad \Longleftrightarrow \quad \sum_{i=1}^{\infty} x_i^2 < \infty.$$

Define $A : \ell^2 \to \ell^2$ by

$$A\mathbf{x} = \left(x_1, \frac{x_2}{2}, \ldots, \frac{x_n}{n}, \ldots \right).$$

Formal (unbounded) inverse is

$$A^{-1}\mathbf{y} = (y_1, 2y_2, \ldots, ny_n, \ldots),$$

so we have uniqueness (and existence of solutions for certain $\mathbf{y}$).

Don't have stability!

Take

$$\mathbf{y}_n = (0, \ldots, 0, \underbrace{1/\sqrt{n}}_{nth}, 0, \ldots)$$

Then $\mathbf{y}_n \to \mathbf{0}$, but

$$||A^{-1}\mathbf{y}_n|| = \sqrt{n} \to \infty.$$

Also don't have existence of solns to $A\mathbf{x} = \mathbf{y}$ for all $\mathbf{y} \in Y$.
E.g., $\mathbf{y} = (1, 1/2, 1/3, \ldots) = A(1, 1, 1, \ldots)$, but $(1, 1, 1, \ldots) \notin \ell^2$.

Does this matter?

- Example was contrived.
- Practical computations are discrete, finite dimensional.
- Can replace (finite dimensional) A^{-1} by pseudo-inverse $A^{\dagger}$.

But ...

- Discrete problems approximate underlying infinite dimensional problems (Discrete problems become increasingly ill-conditioned as they become more accurate).
- In Inverse Problems applications A is often compact, and it acts like the diagonal operator in the above example (Compact operators can be diagonalized using the SVD; diagonal entries decay to zero).

Regularization

Remedy for ill-posedness (or ill-conditioning, in discrete case).

Informal Definition: "Imposes stability on an ill-posed problem in a manner that yields accurate approximate solutions, often by incorporating prior information".

More Formal Definition: Parametric family of "approximate inverse operators" $R_\alpha : Y \to X$ with the following property. If $\mathbf{y}_n = A\mathbf{x}_{\text{true}} + \eta_n$, and $\eta_n \to \mathbf{0}$, we can pick parameters α_n such that

$$\mathbf{x}_{\alpha_n} \stackrel{\text{def}}{=} R_{\alpha_n}\mathbf{y}_n \to \mathbf{x}_{\text{true}}.$$

Tikhonov Regularization

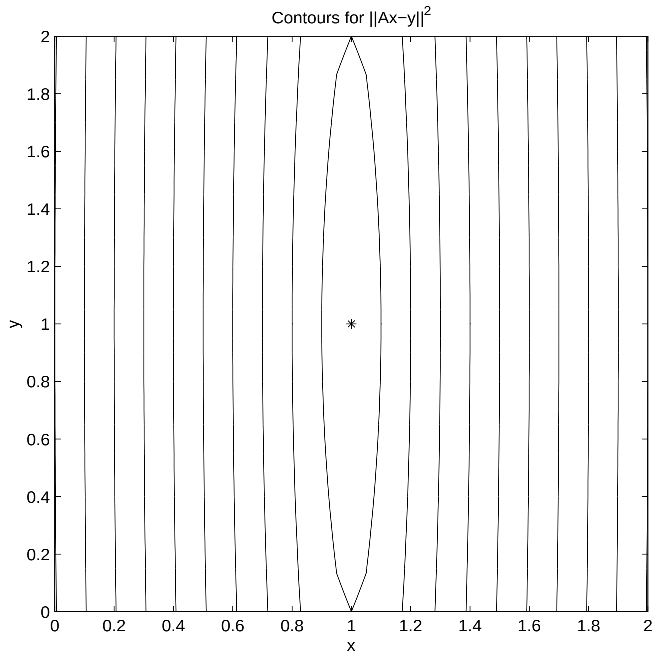
Math Interpretation. In simplest case, assume X, Y are Hilbert spaces. To obtain regularized soln to $A\mathbf{x} = \mathbf{y}$, choose $\mathbf{x}$ to fit data $\mathbf{y}$ in least-squares sense, but penalize solutions of large norm. Solve minimization problem

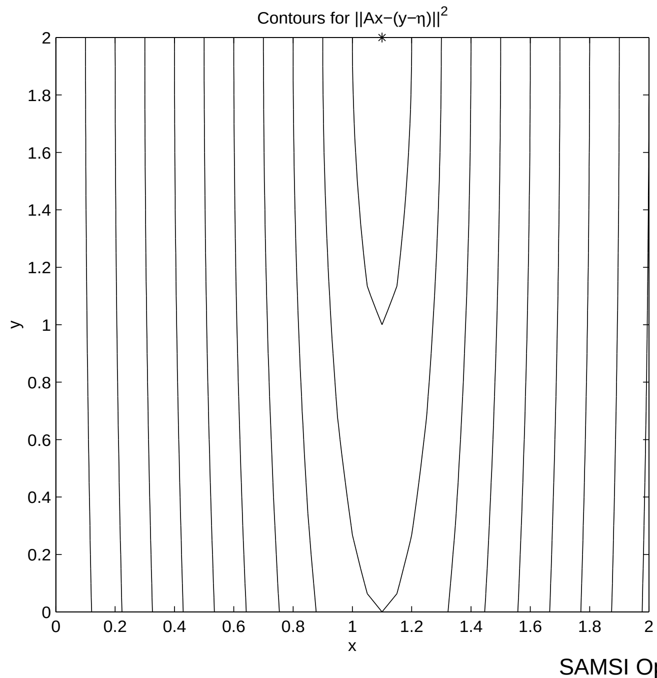
$$\begin{aligned} \mathbf{x}_\alpha &= \arg\min_{\mathbf{x}\in X} ||A\mathbf{x} - \mathbf{y}||_Y^2 + \alpha||\mathbf{x}||_X^2 \\ &= \underbrace{(A^*A + \alpha I)^{-1} A^*}_{R_\alpha} \mathbf{y}. \end{aligned}$$

$\alpha > 0$ is called the regularization parameter.

Geometry of Linear Least Squares

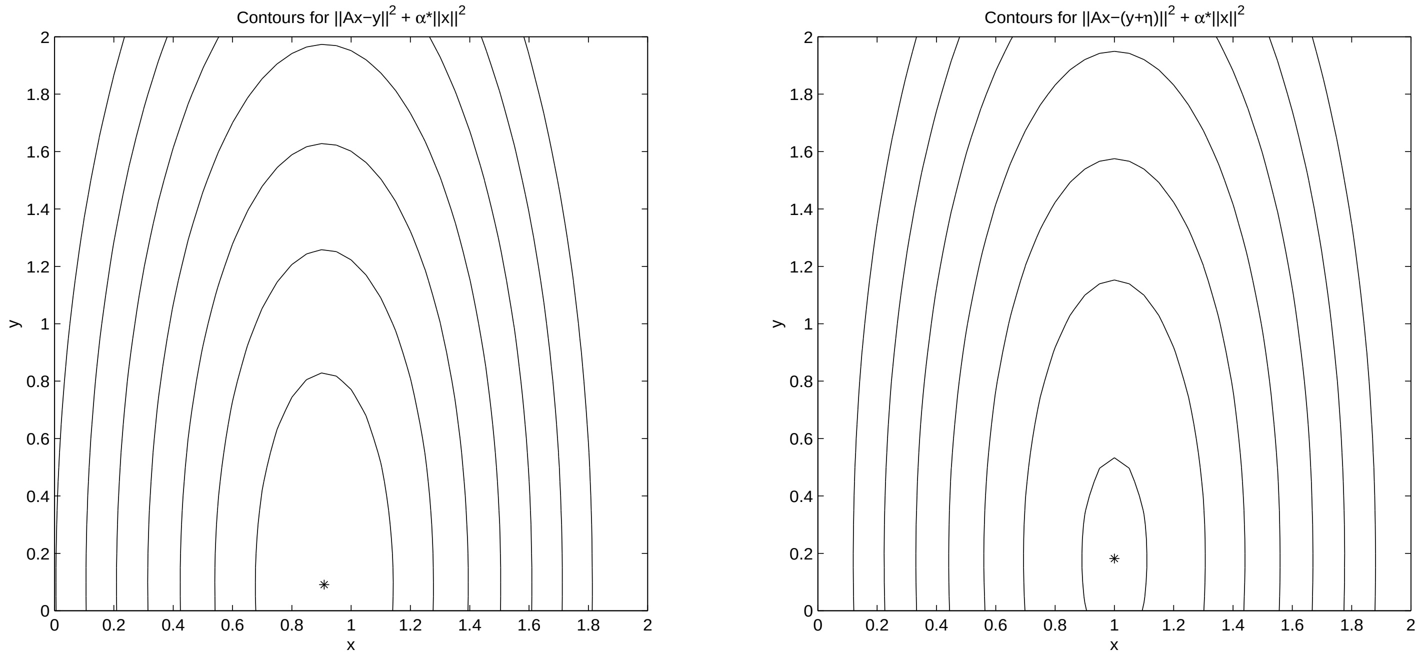
$$J(\mathbf{x}) = \left\| \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & .1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\mathbf{x}} - \left(\underbrace{\begin{bmatrix} 1 \\ .1 \end{bmatrix}}_{\mathbf{y}} + \underbrace{\begin{bmatrix} .1 \\ .1 \end{bmatrix}}_{\eta} \right) \right\|^2$$





Geometry of Tikhonov Regularization

$$J(\mathbf{x}) = ||A\mathbf{x} - (\mathbf{y} + \eta)||^2 + \underbrace{0.1}_{\alpha}\, ||\mathbf{x}||^2$$



Bayesian Interpretation (MAP)

Bayes Law: Assume X, Y are jointly distributed continuous random variables.

$$\underbrace{\pi(\mathbf{x}|\mathbf{y})}_{posterior\ pdf} = \underbrace{\pi(\mathbf{y}|\mathbf{x})}_{conditional\ pdf} \underbrace{\pi(\mathbf{x})}_{prior} \Big/ \underbrace{\pi(\mathbf{y})}_{indep\ of\ \mathbf{x}}$$

Maximum a posteriori (MAP) extimator is max of posterior pdf. Equivalently, minimize w.r.t. $\mathbf{x}$

$$-\log \pi(\mathbf{x}|\mathbf{y}) = -\underbrace{\log \pi(\mathbf{y}|\mathbf{x})}_{log\ likelihood} - \underbrace{\log \pi(\mathbf{x})}_{log\ \ prior}$$

First term on rhs is “fit-to-data” term; second is “regularization” term.

Illustrative Example

If $X \sim$ Normal$(\mathbf{0}, \sigma_x^2 I)$, then prior is

$$\pi(\mathbf{x}) = \frac{1}{(2\pi\sigma_x^2)^{n/2}} \exp\left[-||\mathbf{x}||^2/2\sigma_x^2\right]$$

If $Y = AX + \eta$ and $\eta \sim$ Normal$(\mathbf{0}, \sigma_\eta^2 I)$, conditional pdf is

$$\pi(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi\sigma_\eta^2)^{n/2}} \exp\left[-||A\mathbf{x} - \mathbf{y}||^2/2\sigma_\eta^2\right]$$

Tikhonov cost functional is

$$J(\mathbf{x}) = ||A\mathbf{x} - \mathbf{y}||^2 + \alpha||\mathbf{x}||^2, \quad \alpha = \frac{\sigma_\eta^2}{\sigma_x^2} = \text{SNR}^{-2}.$$

Singular Value Decomposition

Important tool for analysis and computation. Gives bi-orthogonal diagonalization of linear operator,

$$A = USV^*.$$

In $n \times n$ matrix case, $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$, $S = \text{diag}(s_1, \ldots, s_n)$, and $V = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ with

$$s_1 \geq s_2 \geq \ldots \geq s_n \geq 0,$$

$$A\mathbf{v}_i = s_i \mathbf{u}_i, \qquad A^*\mathbf{u}_i = s_i \mathbf{v}_i,$$

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{ij}, \;\; \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{ij} \quad \Rightarrow \quad U^*U = I, \; V^*V = I$$

Tikhonov Filtering

In the case of Tikhonov regularization, using the SVD $A = USV^*$ (and assuming $n \times n$ matrix with $s_i > 0$ for simplicity),

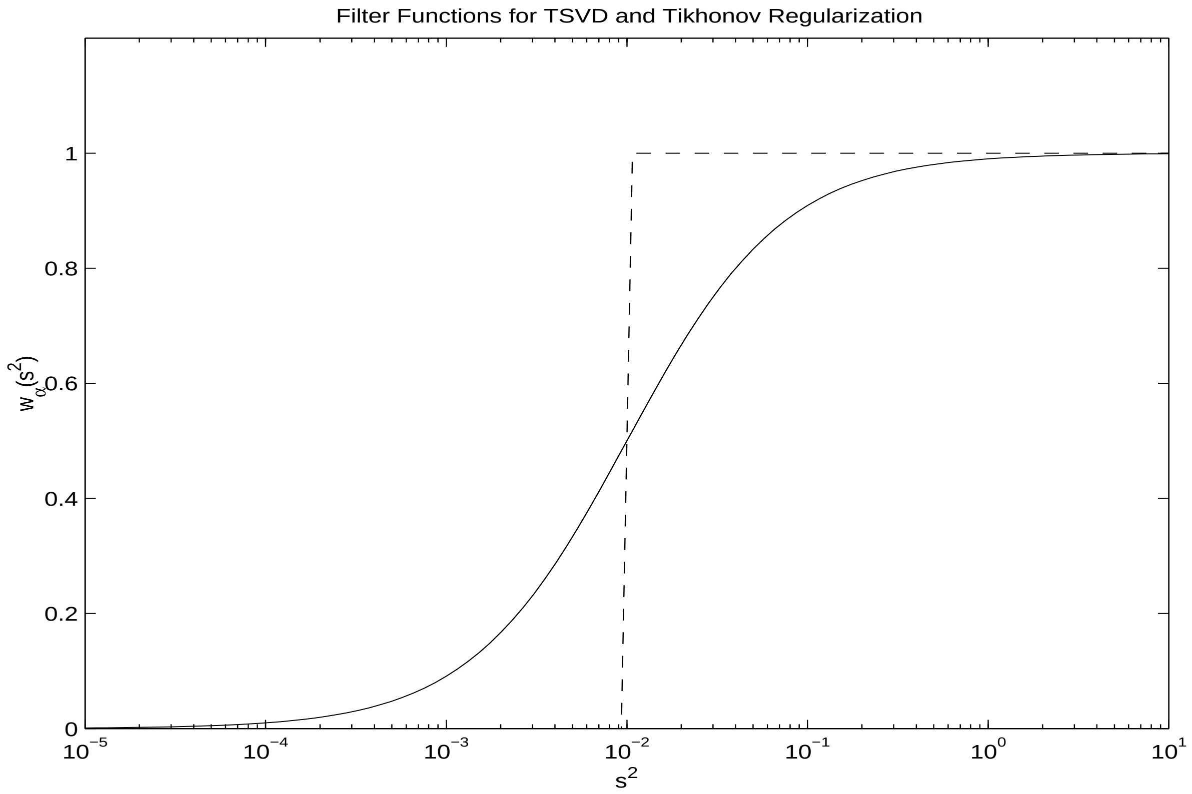
$$\begin{aligned} R_\alpha &= (A^*A + \alpha I)^{-1} A^* \\ &= (VS^*U^*USV^* + \alpha VIV^*)^{-1} VS^*U^* \\ &= V\,(S^*S + \alpha I)^{-1} S^*\,U^* \\ &= V\,\text{diag}(\underbrace{\frac{s_i^2}{s_i^2+\alpha}}_{w_\alpha(s_i^2)}\frac{1}{s_i})\,U^* \end{aligned}$$

If $\alpha \to 0$, then $w_\alpha(s_i^2) \to 1$, so

$$R_\alpha \to V\,\text{diag}(1/s_i)U^* \stackrel{\text{def}}{=} A^\dagger \quad \text{as} \quad \alpha \to 0.$$

Tikhonov Filtering, Continued

Plot of Tikhonov filter function $w_\alpha^{\mathrm{Tikh}}(s^2) = \frac{s^2}{s^2+\alpha}$ shows that Tikhonov regularization filters out singular components that are small (relative to α) while retaining components that are large.



Truncated SVD (TSVD) Regularization

TSVD filtering function is

$$w_\alpha^{\mathrm{TSVD}}(s_i^2) = \begin{cases} 0, & s_i^2 \le \alpha, \\ 1, & s_i^2 > \alpha. \end{cases}$$

Has “sharp cut-off” behavior instead of “smooth roll-off behavior” of Tikhonov filter.

Iterative Regularization

Certain iterative methods, e.g., steepest descent, conjugate gradients, and Richardson-Lucy (EM), have regularizing effects with the regularization parameter equal to the number of iterations. These are useful in applications, like 3-D imaging, with many unknowns.

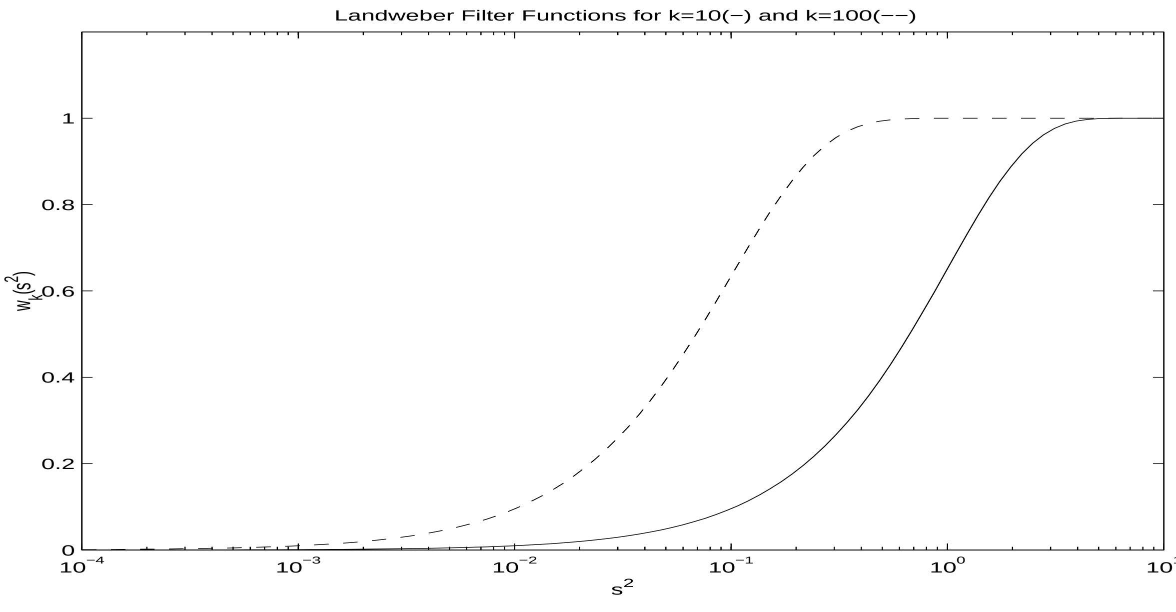
An example is Landweber iteration, a variant of steepest descent. Minimize the least squares fit-to-data functional

$$J(\mathbf{x}) = \frac{1}{2}||A\mathbf{x} - \mathbf{y}||^2$$

using gradient descent iteration, initial guess $\mathbf{x}^0 = \mathbf{0}$, and fixed step length parameter $0 < \tau < 1/||A||^2$.

Landweber Iteration

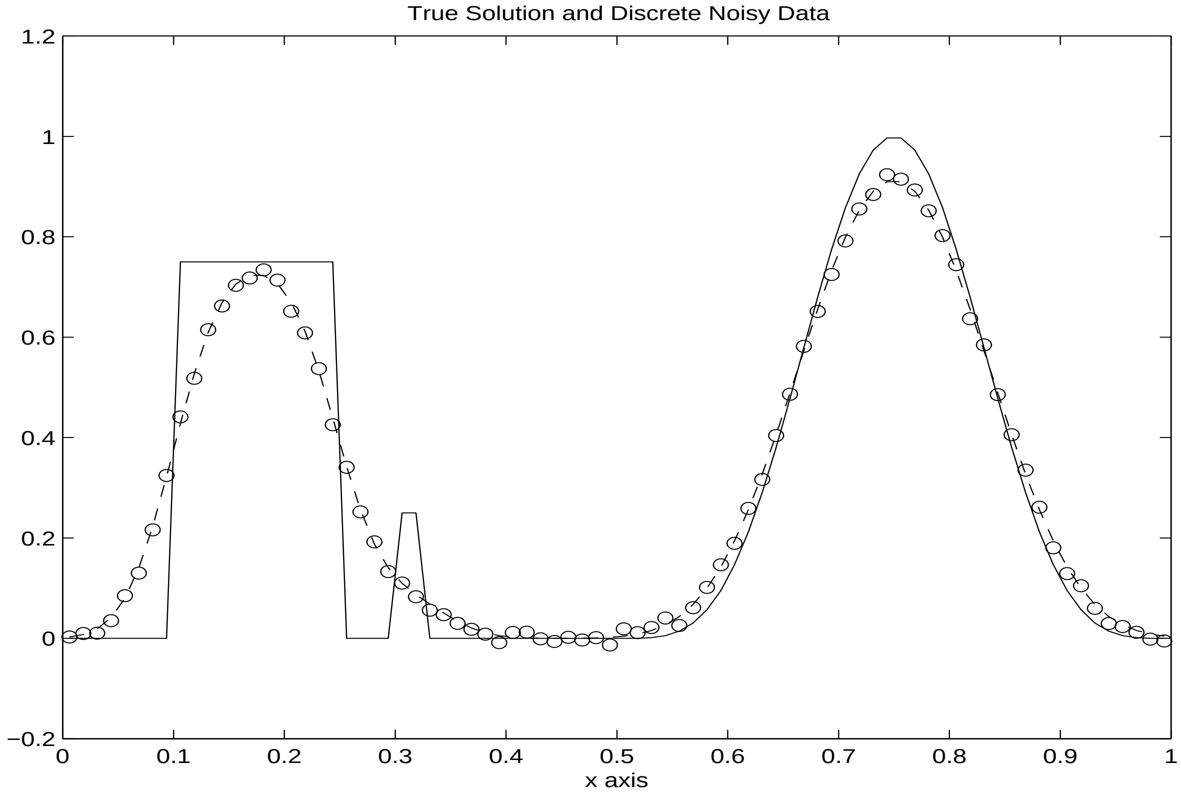
$$
\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{x}^k - \tau \,\text{grad}\, J(\mathbf{x}^k), \quad k = 0, 1, 2, \ldots \\
&= \mathbf{x}^k - \tau A^*(A\mathbf{x}^k - \mathbf{y}) \\
&= (I - \tau A^* A)\mathbf{x}^k + \tau A^* \mathbf{y} \\
&= V \,\text{diag}(\underbrace{1 - (1 - \tau s_i^2)^k}_{Landweber\, filter\, fn})\, U^* \mathbf{y}.
\end{aligned}
$$



Effect of Regularization Parameter

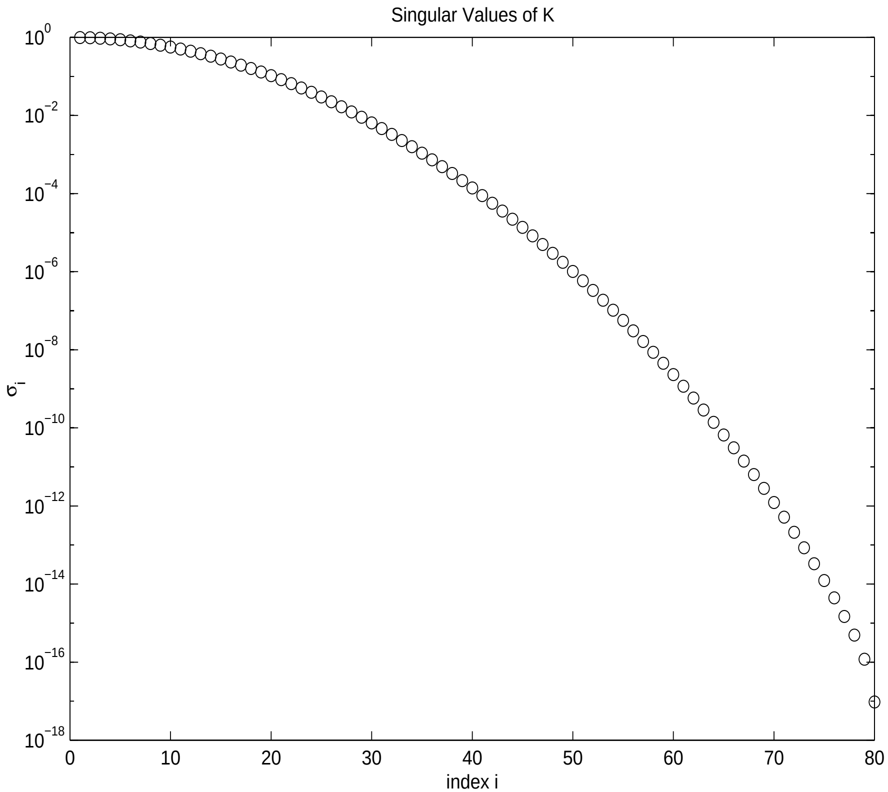
Illustrative Example: 1-D deconvolution with Gaussian kernel $a(t) = C\exp(-t^2/2\gamma^2)$ and discrete data

$$d_i = \int_0^1 a(s_i - t)\, x_{\text{true}}(t)\, dt + \text{noise}, \quad i = 1, \dots, n.$$

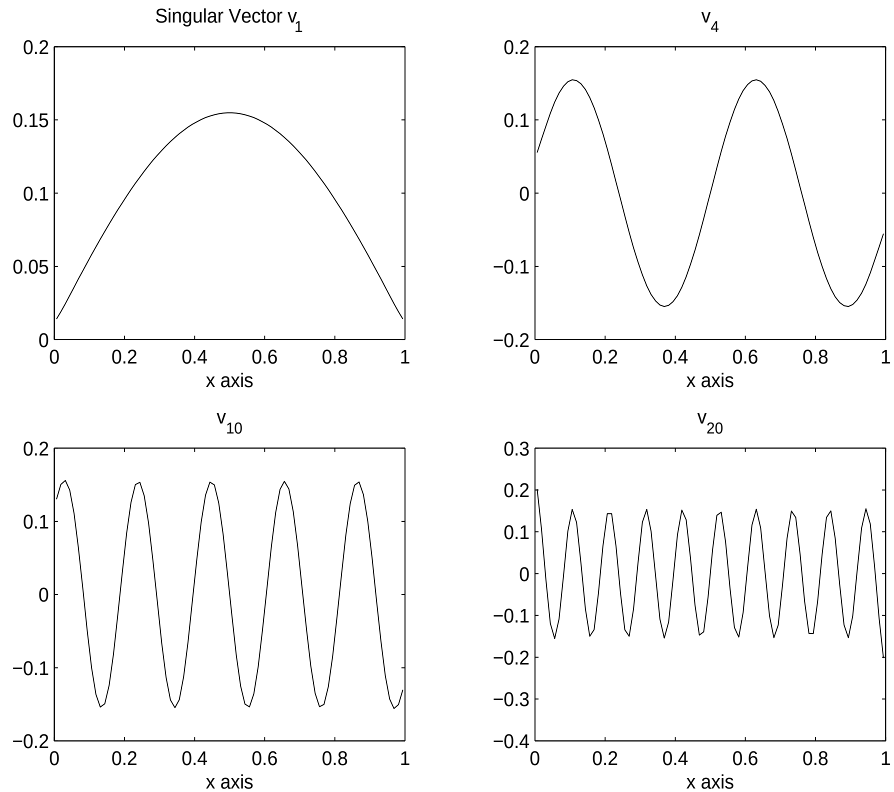


Graphical Representation of SVD

Singular Values

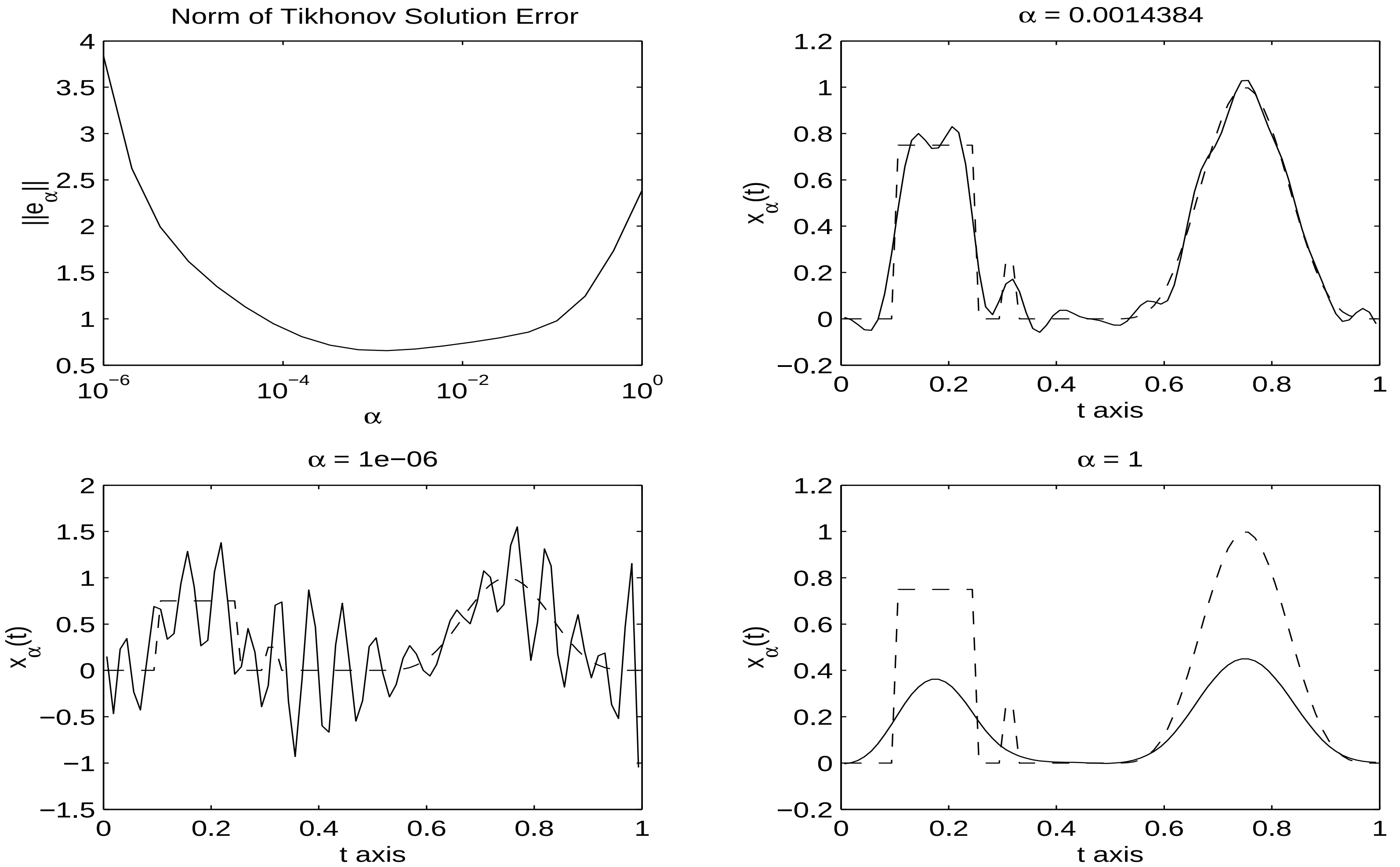


Singular Vectors



Tikhonov Solutions vs α

Tikhonov regularized solution is $x_\alpha = (A^*A + \alpha I)^{-1} A^* \mathbf{d}$.
Solution error is $e_\alpha = x_\alpha - x_{\text{true}}$.



Error Indicators

Linear, Additive Noise Data Model:

$$\mathbf{d} = A\mathbf{x}_{\text{true}} + \eta$$

Regularized Solution:

$$\mathbf{x}_\alpha = R_\alpha \mathbf{d} = V \,\text{diag}\big(w_\alpha(s_i^2)\,/s_i\big)\, U^* \mathbf{d}$$

Solution Error:

$$\mathbf{e}_\alpha \stackrel{\text{def}}{=} \mathbf{x}_\alpha - \mathbf{x}_{\text{true}} = \underbrace{(R_\alpha A - I)\mathbf{x}_{\text{true}}}_{\text{“}bias\text{''}} + \underbrace{R_\alpha \eta}_{\text{“}variance\text{''}}$$

Predictive Error:

$$\mathbf{p}_\alpha \stackrel{\text{def}}{=} A\mathbf{x}_\alpha - A\mathbf{x}_{\text{true}} = A\mathbf{e}_\alpha = (AR_\alpha - I)A\mathbf{x}_{\text{true}} + AR_\alpha \eta$$

Unbiased Predictive Risk Estimator

The Influence Matrix is

$$B(\alpha) \stackrel{\text{def}}{=} AR_\alpha,$$

so we can write the predictive error as

$$\mathbf{p}_\alpha = (B(\alpha) - I)A\mathbf{x}_{\text{true}} + B(\alpha)\eta$$

Residual is

$$\mathbf{r}_\alpha \stackrel{\text{def}}{=} A\mathbf{x}_\alpha - \mathbf{d} = (B(\alpha) - I)A\mathbf{x}_{\text{true}} + (B(\alpha) \underbrace{-I}_{\textit{new term}})\eta$$

Let $\mathcal{E}$ denote expected value operator. Assume $\mathbf{x}_{\text{true}}$ is deterministic (or independent of η), assume $\mathcal{E}(\eta) = \mathbf{0}$, and note that $B(\alpha)$ is symmetric. Then ...

UPRE, Continued

$$\mathcal{E}||\mathbf{r}_\alpha||^2 = \underbrace{||(B(\alpha)-I)A\mathbf{x}_{\text{true}}||^2 + \mathcal{E}[\eta^* B(\alpha)^2\eta]}_{\mathcal{E}||\mathbf{p}_\alpha||^2} - 2\,\mathcal{E}[\eta^* B(\alpha)\eta] + \mathcal{E}||\eta||^2.$$

So up to const $\mathcal{E}||\eta||^2$, an unbiased estimator for $||\mathbf{p}_\alpha||^2$ is

$$U(\alpha) \stackrel{\text{def}}{=} ||\mathbf{r}_\alpha||^2 + 2\mathcal{E}[\eta^* B(\alpha)\eta] = ||\mathbf{r}_\alpha||^2 + 2\sigma_\eta^2\,\text{trace}\,B(\alpha)$$

Last equality follows if

$$\mathcal{E}[\eta_i\eta_j] = \begin{cases} \sigma_\eta^2, & i=j, \\ 0, & i \neq j \end{cases}$$

Comments about UPRE

UPRE regularization parameter selection method, also known as Mallow's C_L method, is to pick α to mimimize $U(\alpha)$.

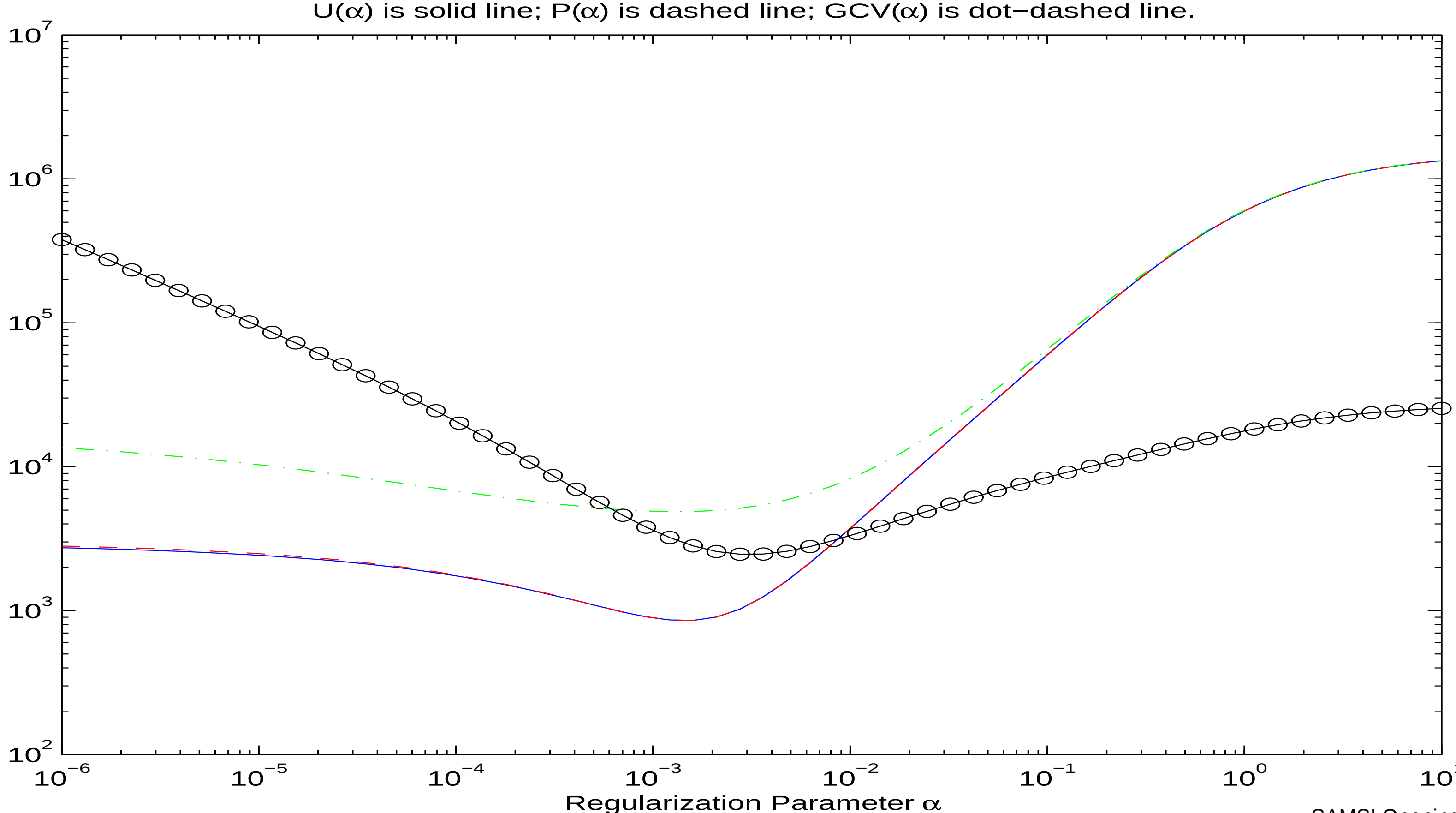
- Predictive error norm $||\mathbf{p}_\alpha||$ and solution error norm $||\mathbf{e}_\alpha||$ need not have the same minimizer, but the mins are often quite close.
- There is a variant of UPRE, called generalized cross validation (GCV), which requires minimization of

$$V(\alpha) \stackrel{\text{def}}{=} \frac{||\mathbf{r}_\alpha||^2}{[\text{trace}(I - B(\alpha))]^2}.$$

This does not require prior knowledge of variance σ_η^2.

Illustrative Example of Indicators

2-D image reconstruction problem, noise $\eta \sim \mathrm{N}(\mathbf{0}, \sigma_\eta^2 I)$, Tikhonov regularization. o-o indicates soln error norm; — — indicates GCV; – indicates $U(\alpha)$; and — — indicates predictive error norm.



Mathematical Summary

- There exists a well-developed mathematical theory of regularization.
- There are a number of different approaches to regularization.
  - optimization-based (equivalent to MAP)
  - filtering-based
  - iteration-based
- There are robust schemes for choosing regularization parameters.

These techniques often work well in practical applications.

But Things Can Get Ugly ...

Astronomical Imaging Application. Light intensity

$$I(p,q) = \int\int \underbrace{a(p-p', q-q')}_{PSF} \underbrace{x(p',q')}_{object} \, dp \, dq.$$

This is measured by a ccd array (digital camera), giving data

$$d_i = I(p_{i1}, q_{i2}) + \text{“noise”}.$$

For high contrast imaging (dim object near very bright object), accurate modeling of noise is critical.

With ordinary (and even weighted) least squares, dim object is missed.

Model for Data from CCD Array

$$d_i = c_i(\mathbf{x}) + b_i + \eta_i, \quad i = 1, \ldots, n$$

- Photon count for “signal”

$$c_i(\mathbf{x}) \sim \text{Poisson}(\lambda_i), \quad \lambda_i = I(p_{i1}, q_{i2}) \approx [A\mathbf{x}]_i.$$

- Background photon count

$$b_i \sim \text{Poisson}(b), \quad b \text{ fixed, known.}$$

- Instrument “read noise”

$$\eta_i \sim \text{N}(0, \sigma^2), \quad \sigma^2 \text{ fixed, known.}$$

Imaging Example, Continued

- Log likelihood ($\propto \log \pi(\mathbf{d}|A\mathbf{x})$) is messy

$$L(A\mathbf{x};\mathbf{d}) = -\sum_{i=1}^{n} \log \sum_{j=0}^{\infty} \frac{e^{-[Ax]_i - b}([Ax]_i + b)^j}{j!} \; e^{-(d_i - [Ax]_i - b)^2/\sigma^2}$$

- Light source (object) intensity is nonnegative. Constraint $x(t) \geq 0$.
- With "pixel" discretization, dimension is very large, e.g., size$(\mathbf{x})$ = size$(\mathbf{d}) = 256^2$ or more.
- Problem is ill-posed. Need regularization (prior), e.g.,

$$\alpha ||\mathbf{x}||^2, \quad \alpha > 0.$$

- Regularization parameter (strength of prior) is unknown.

Applied Mathematician's Wish List

- Optimization-based regularization methods (Tikhonov, MAP) require soln of minimization problems. Need fast, robust, large-scale, nonlinear constrained numerical optimization techniques.
- When the parameter-to-observation map is nonlinear, regularization functionals may be non-convex. Need optimization methods which yield the global minimizer (not just a local min) and are fast, robust, ...
- Need indicators of reliability (e.g., confidence intervals) for regularized solutions.
- Need good priors.
- Need fast, robust schemes for choosing regularization parameters.

Challenge for the Statistics Community

- Can MCMC techniques provide fast, robust alternatives to optimization-based regularization methods?

Relevant Reference: J. Kaipio, et al, "Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography", Inverse Problems, vol 16 (2000), pp. 1487-1522.

Relevant Caveat: There is no free lunch.