# Regularization Parameter Estimation: Stabilization of LSQR Algorithms by Iterative Reweighting

Rosemary Renaut[1] Saeed Vatankhah[2]

1: School of Mathematical and Statistical Sciences, Arizona State University

2: University of Tehran

SIAM Applied Linear Algebra, October 2015

# Outline

- Background: Tikhonov Regularization for Ill-Posed Problems
  - Standard Approaches to Estimate Regularization Problem
- Parameter estimation on the projected problem
  - UPRE is a good estimator [RVA15]
  - Identifying the weight parameter in the GCV [CNO08]
- Numerical Illustrations
  - Numerical Illustrations $1d$ Underdetermined Cases
- Identifying the optimal Subspace
  - Appearance of Noise in the Subspace [HPS09]
  - Minimization of the GCV for the truncated SVD [CKO15]
- Simulations: Two dimensional Examples
- Iteratively Reweighted Regularization [LK83]
- Conclusions

## Background:Tikhonov Regularization for Ill-Posed Problems

Ill-Posed Equations in the presence of noise

$$A\mathbf{x} \approx \mathbf{b} \quad A \in \mathbb{R}^{m \times n}$$

$$\mathbf{b} = \mathbf{b}_{\text{true}} + \boldsymbol{\eta}, \quad \text{noise} \quad \boldsymbol{\eta} \sim \mathbb{N}(0, C_{\eta})$$

Tikhonov Regularization:

$$\mathbf{x}(\lambda) = \underset{\mathbf{x} \in \mathbb{R}^n}{\operatorname{argmin}} \{\|A\mathbf{x} - \mathbf{b}\|^2_{W_\eta} + \lambda^2 \|L(\mathbf{x} - \mathbf{x}_0)\|^2_2\}$$

Mapping $L$ defines basis for $\mathbf{x}$

Prior $\mathbf{x}_0$

Weighting $W_\eta = C_\eta^{-1}$, $\|\mathbf{y}\|_{W_\eta} = \mathbf{y}^T W_\eta \mathbf{y}$. Whitens noise in $\mathbf{b}$.

**Requires automatic estimation of $\lambda$**

## Regularization Parameter Estimation using the SVD: Examples ($m = n$)

For $L$ invertible, and SVD $W_\eta^{1/2} A L^{-1} = U \Sigma V^T$, $\Sigma = \mathrm{diag}(\sigma_i)$.
Find $\lambda^{\mathrm{opt}}$ from e.g.

Unbiased Predictive Risk Minimize functional

$$U(\lambda) = \sum_{i=1}^{n} \left( \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^2 (\mathbf{u}_i^T \mathbf{b})^2 + 2 \sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2}$$

Morozov Discrepancy Principle Given parameter $\nu$, solve

$$M(\lambda) = \sum_{i=1}^{n} \left( \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^2 (\mathbf{u}_i^T \mathbf{b})^2 - \nu n = 0$$

GCV : Minimize rational function

$$G(\lambda) = \left( \sum_{i=1}^{n} \left( \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^2 (\mathbf{u}_i^T \mathbf{b})^2 \right) \left( \sum_{i=1}^{n} \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^{-2}$$

**Not practical for large scale problems**

## Regularization of the LSQR solution: Questions

(i) Determine optimal $t$ The choice of the subspace impacts the regularizing properties of the iteration: For large $t$ noise due to numerical precision and data error enters the projected space.

(ii) Determine optimal $\zeta_t$ How do regularization parameter techniques translate to the projected problem?

(iii) Relation optimal $\zeta_t$ and optimal $\lambda$ Given $t$ how well does optimal $\zeta_t$ for projected space yield optimal $\lambda$ for full space, or when is this the case?

## Needed Properties and Definitions:

**Interlace Properties** Singular values, $\gamma_i$, of $B_t$, $\sigma_i$ of $A$, interlace

$$\sigma_1 \geq \gamma_1 \geq \sigma_2 \cdots \geq \gamma_t \geq \sigma_{t+1} \geq 0.$$

**Residuals** Full, $\mathbf{r}^{\text{full}}(\mathbf{x}_t)$, and projected, $\mathbf{r}^{\text{proj}}(\mathbf{w}_t)$,

$$\begin{aligned}\mathbf{r}^{\text{full}}(\mathbf{x}_t) &= A\mathbf{x}_t - \mathbf{b} = AG_t\mathbf{w}_t - \beta_1 H_{t+1}\mathbf{e}_1^{(t+1)} \\ &= H_{t+1}(B_t\mathbf{w}_t - \beta_1\mathbf{e}_1^{(t+1)}) = H_{t+1}\mathbf{r}^{\text{proj}}(\mathbf{w}_t).\end{aligned}$$

**Pseudoinverse** Use $A^\dagger(\lambda) = (A^TA + \lambda^2 I)^{-1}A^T$ for pseudo inverse of $[A; \lambda I]$, then

$$\begin{aligned}\mathbf{w}_t(\zeta_t) &= \beta_1({B_t}^T B_t + {\zeta_t}^2 I_t)^{-1}{B_t}^T\mathbf{e}_1^{(t+1)} = \beta_1 B_t^\dagger(\zeta_t)\mathbf{e}_1^{(t+1)} \\ &= ({G_t}^T A^T A G_t + {\zeta_t}^2 I_t)^{-1}{G_t}^T A^T\mathbf{b} = (AG_t)^\dagger(\zeta_t)\mathbf{b}.\end{aligned}$$

**Influence** $A(\lambda) = AA^\dagger(\lambda)$ for the influence matrix, likewise $(AG_t)(\zeta_t) = AG_t(AG_t)^\dagger(\zeta_t)$.

## Calculating Unbiased Predictive Risk using $\mathbf{w}_t(\lambda)$[RVA15]

Full problem

$$\lambda^{\text{opt}} = \underset{\lambda}{\operatorname{argmin}}\{\|\mathbf{r}^{\text{full}}(\mathbf{x}(\lambda))\|_2^2 + 2\,\operatorname{Tr}(A(\lambda)) - m\} = \underset{\lambda}{\operatorname{argmin}}\{U^{\text{full}}(\lambda)\}.$$

Using the projected solution for parameter $\lambda$ and

$$\operatorname{Tr}\left((AG_t)(\lambda)\right) = \operatorname{Tr}\left(B_t(\lambda)\right)$$

$$\begin{aligned} U^{\text{full}}(\lambda) &= \|\left((AG_t)(\lambda) - I_m\right)\mathbf{b}\|_2^2 + 2\,\operatorname{Tr}\left((AG_t)(\lambda)\right) - m \\ &= \|\beta_1(B_t(\lambda) - I_{t+1})\mathbf{e}_1^{t+1}\|_2^2 + 2\,\operatorname{Tr}(B_t(\lambda)) - m \end{aligned}$$

**$\lambda^{\text{opt}}$ for $U^{\text{full}}(\lambda)$ can be estimated given projected SVD**

# Deriving UPRE for the projected problem

**Is $\lambda^{\text{opt}}$ relevant to ${\zeta_t}^{\text{opt}}$ for the projected problem?**

Noise in the right hand side For $\mathbf{b} = \mathbf{b}^{\text{true}} + \boldsymbol{\eta}$, $\boldsymbol{\eta} \sim \mathbb{N}(0, I_m)$

$$\beta_1 \mathbf{e}_1^{t+1} = H_{t+1}^T \mathbf{b} = H_{t+1}^T \mathbf{b}^{\text{true}} + H_{t+1}^T \boldsymbol{\eta}.$$

Noise in projected right hand side $\beta_1 \mathbf{e}_1^{t+1}$, satisfies

$$H_{t+1}^T \boldsymbol{\eta} \sim \mathbb{N}(0, I_{t+1})$$

Immediately

$$\begin{aligned} U^{\text{proj}}(\zeta_t) &= \|\beta_1 (B_t(\zeta_t) - I_{t+1}) \mathbf{e}_1^{(t+1)}\|_2^2 + 2\,\text{Tr}(B_t(\zeta_t)) - (t+1) \\ &= U^{\text{full}}(\zeta_t) + m - (t+1). \end{aligned}$$

**Minimizer of $U^{\text{proj}}(\zeta_t)$ is minimizer of $U^{\text{full}}(\zeta_t)$**

## ${\zeta_t}^{\text{opt}}$ calculated for projected problem may not yield $\lambda^{\text{opt}}$ on full problem

**${\zeta_t}^{\text{opt}}$ depends on $t$, $\lambda^{\text{opt}}$ depends on $m^* =: \min(m, n)$**

Trace Relations By linearity and cycling.

$$\mathrm{Tr}(A(\lambda)) = \mathrm{Tr}(A(A^TA + \lambda^2 I_n)^{-1}A^T) = n - \lambda^2 \mathrm{Tr}((A^TA + \lambda^2 I_n)^{-1})$$

$$= m^* - \lambda^2 \sum_{i=1}^{m^*} (\sigma_i^2 + \lambda^2)^{-1}.$$

Immediately $\mathrm{Tr}(B_t(\zeta_t)) = t - {\zeta_t}^2 \sum_{i=1}^{t} (\gamma_i^2 + {\zeta_t}^2)^{-1}$.

Interlacing For $\sigma_i \approx \gamma_i$, $1 \leq i \leq t$, $\sigma_i^2/(\sigma_i^2 + \lambda^2) \approx 0$, $i > t$,

$$\mathrm{Tr}(A(\lambda)) = t - \lambda^2 \sum_{i=1}^{t} (\sigma_i^2 + \lambda^2)^{-1} + (m^* - t) - \lambda^2 \sum_{i=t+1}^{m^*} (\sigma_i^2 + \lambda^2)^{-1}$$

$$\approx \mathrm{Tr}(B_t(\lambda)) + (m^* - t) - \lambda^2 \sum_{i=t+1}^{m^*} (\sigma_i^2 + \lambda^2)^{-1} \approx \mathrm{Tr}(B_t(\lambda)).$$

**If $t$ approx numerical rank $A$, ${\zeta_t}^{\text{opt}} \approx \lambda^{\text{opt}}$ for $\mathcal{K}_t(A^TA, A^T\mathrm{b})$**

# Other Estimation Techniques for the Projected Problem

GCV: [CNO08] *weighted* GCV is introduced for $\omega > 0$.

$$G^{\text{proj}}(\zeta_t, \omega) = \frac{\|\mathbf{r}^{\text{proj}}(\mathbf{w}_t(\zeta_t))\|_2^2}{(\text{Tr}(\omega B_t(\zeta_t) - I_{t+1}))^2}, \quad G(\lambda) = G^{\text{proj}}(\lambda, 1).$$

Optimal Analysing as for UPRE: $\omega = \frac{t+1}{m} < 1$.

Discrepancy Principle Seek $\lambda$ such that $\|\mathbf{r}^{\text{full}}(\mathbf{x}(\lambda))\|_2^2 = \delta \approx m$.

To avoid over smoothing: $\delta = \upsilon m$, $\upsilon > 1$

Discrepancy for the Projected Problem Seek $\zeta_t$ such that

$$\|\mathbf{r}^{\text{proj}}(\mathbf{w}_t(\zeta_t))\|_2^2 \approx \delta^{\text{proj}} = \upsilon(t+1).$$

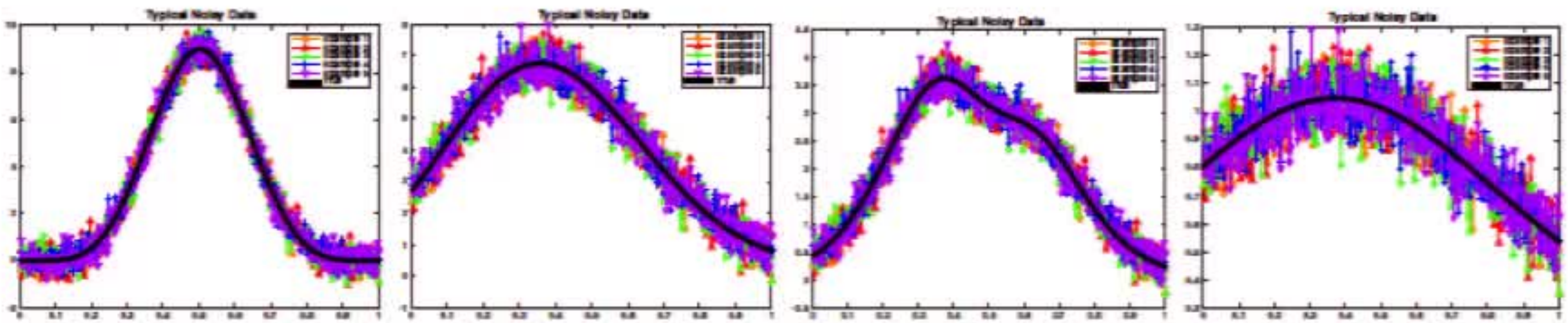
**We do not obtain in these cases $\zeta_t^{\text{opt}} \approx \lambda^{\text{opt}}$**

## Numerical Illustrations $1d$ **Underdetermined** Cases

Regularization Tools `phillips`: Picard condition not satisfied, `shaw`: severely ill-posed, and `gravity` $d = .75$ severely ill-posed, $d = .25$ less severe.

Noise levels SNR approx $-10\log 10(\eta\sqrt{m})$ for noise level $\eta$.

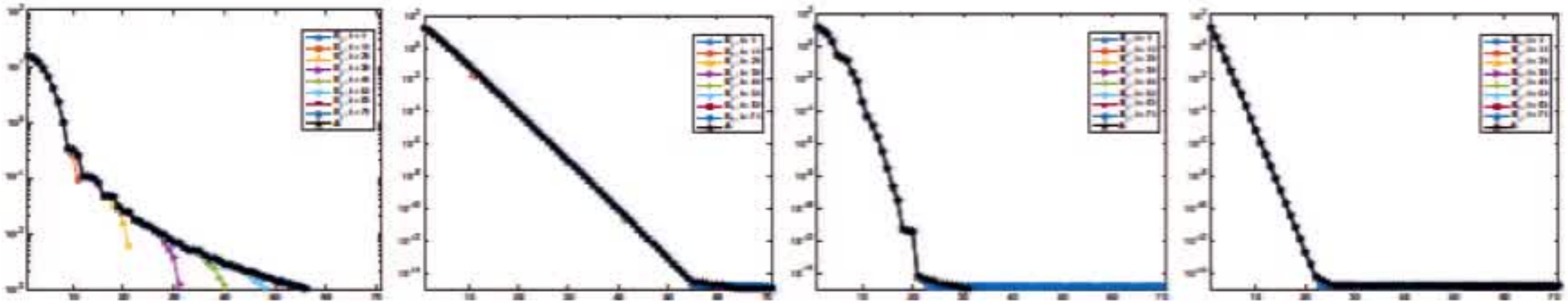
Underdetermined $m = 152$ and $n = 304$. $50\%$ undersampling.



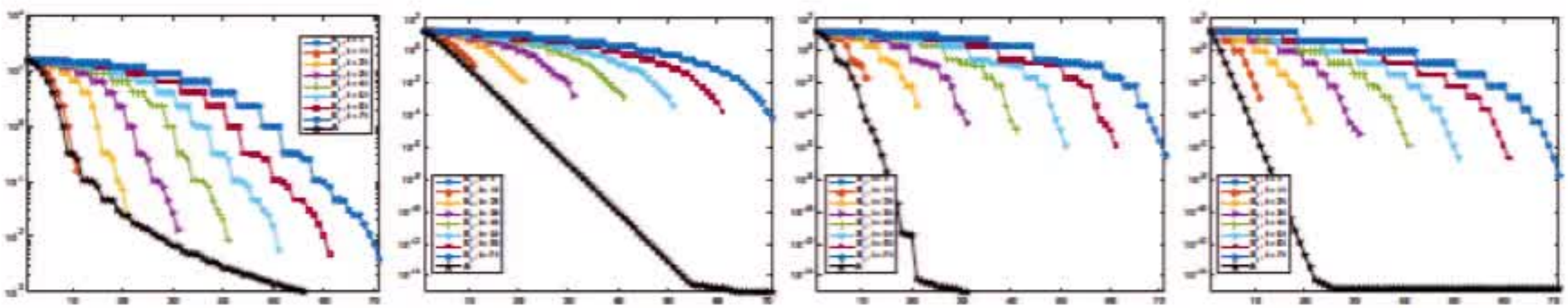
(a) `phillips`, $\eta = .005$ (b) `gravity`, $d = .25$, $\eta = .005$ (c) `shaw`, $\eta = .005$ (d) `gravity`, $d = .75$, $\eta = .005$

Figure: Illustrative test data high noise $\eta = .005$ for 5 sample right hand side data. Exact data are solid lines

# Significance of Reorthogonalization: Clustering of singular values



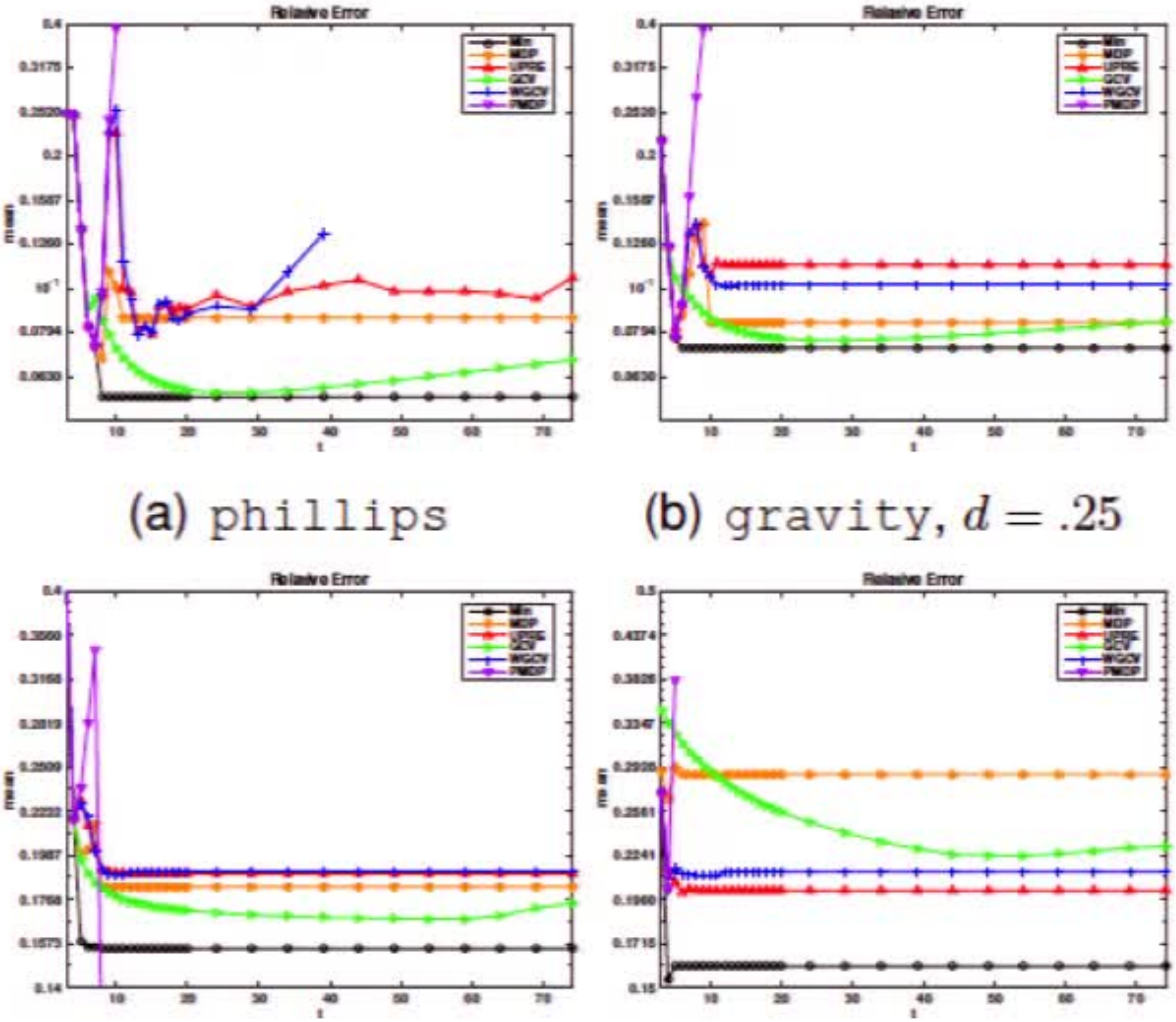
(a) `phillips`: R (b) `gravity`: R $d = .25$ (c) `shaw`: R (d) `gravity`: R $d = .75$



(e) `phillips`: NR (f) `gravity`: NR $d = .25$ (g) `shaw`: NR (h) `gravity`: NR $d = .75$

Figure: Singular values against index for $B_t$, increasing $t$ compared to $A$. With and without reorthogonalization (R) and (NR) in 2(a)-2(b) and 2(e)-2(f), resp.. Notice clustering of spectral values without high accuracy reorthogonalization.

## Average Relative Error over 50 samples: Reorthogonalized




(a) `phillips` (b) `gravity`, $d = .25$

(c) `shaw` (d) `gravity`, $d = .75$

Figure:
Minimum
MDP UPRE
GCV WGCV
PMDP

**UPRE and WGCV may outperform GCV for small $t$**

## Observations

- For the most severe case: `gravity` $d = .75$ UPRE and WGCV yield optimal results
- PMDP gives good results small $t$ but blows up.
- For these examples optimal solutions live on a small projected space.
- Small scale demonstrates the theoretical analysis.
- WGCV and UPRE perform similarly and stabilize with respect to $t$.

## Identifying optimal subspace size $t$

Noise revealing function: [HPS09] suppose $\theta_j$ and $\beta_j$ on diagonal and sub diagonal of $B_t$

$$\rho(t) = \prod_{j=1}^{t} (\theta_j / \beta_{j+1})$$

Optimal $t$ is given by (for user determined $t^{\min}$)

$$t^{\text{opt}-\rho} = \min\{\underset{t > t^{\min}}{\text{argmax}}(\rho(t))\} + \text{step}$$

step$= 2$ is to assure that noise has entered the entries in $\rho(t)$ and hence the basis.

$t^{\min}$ is chosen based on examination of $\rho(t)$.

**Only useful if discrete Picard condition holds [HPS09].**

## Identifying optimal subspace size $t$:

Minimization of the GCV for the truncated SVD of $B_{t^*}$ [CKO15]

Projected subspace size is defined to be $t^*$

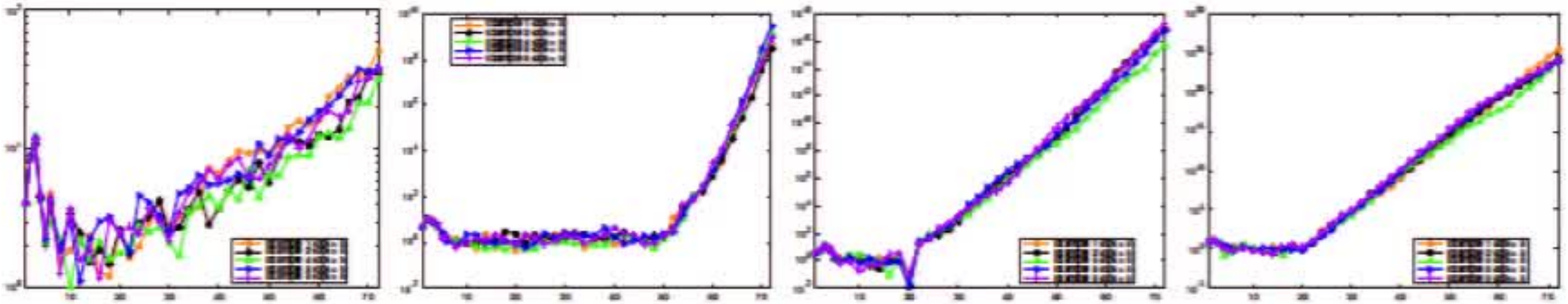
$$\mathcal{G}(t, t^*) = \frac{t^*}{(t^* - t)^2} \sum_{t+1}^{t^*} |\mathbf{u}_i^T \mathbf{b}|^2.$$

Optimal $t$ is given by

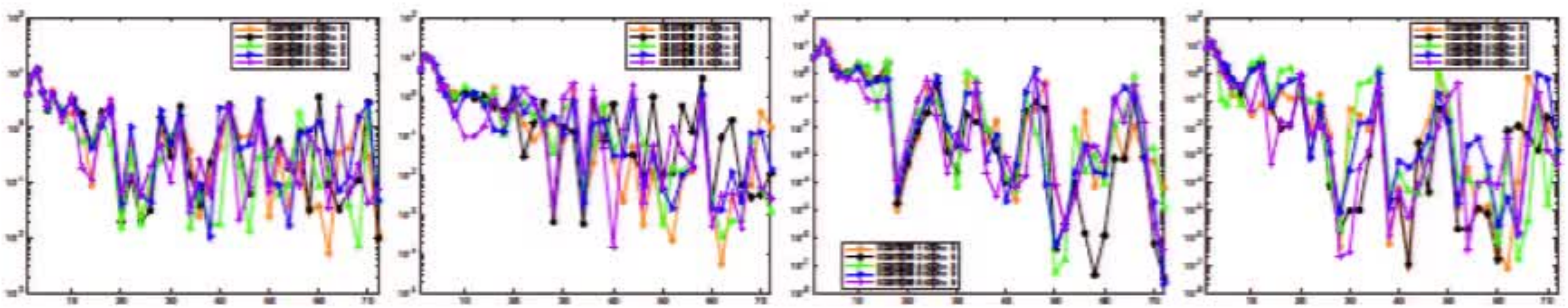
$$t^{\mathrm{opt}-\mathcal{G}} = \operatorname*{argmin}_{t} \mathcal{G}(t, t^*)$$

**Does not require Picard condition, but $t^{\mathrm{opt}-\mathcal{G}}$ depends on $t^*$**

## Significance of reorthogonalization: Estimating $t$



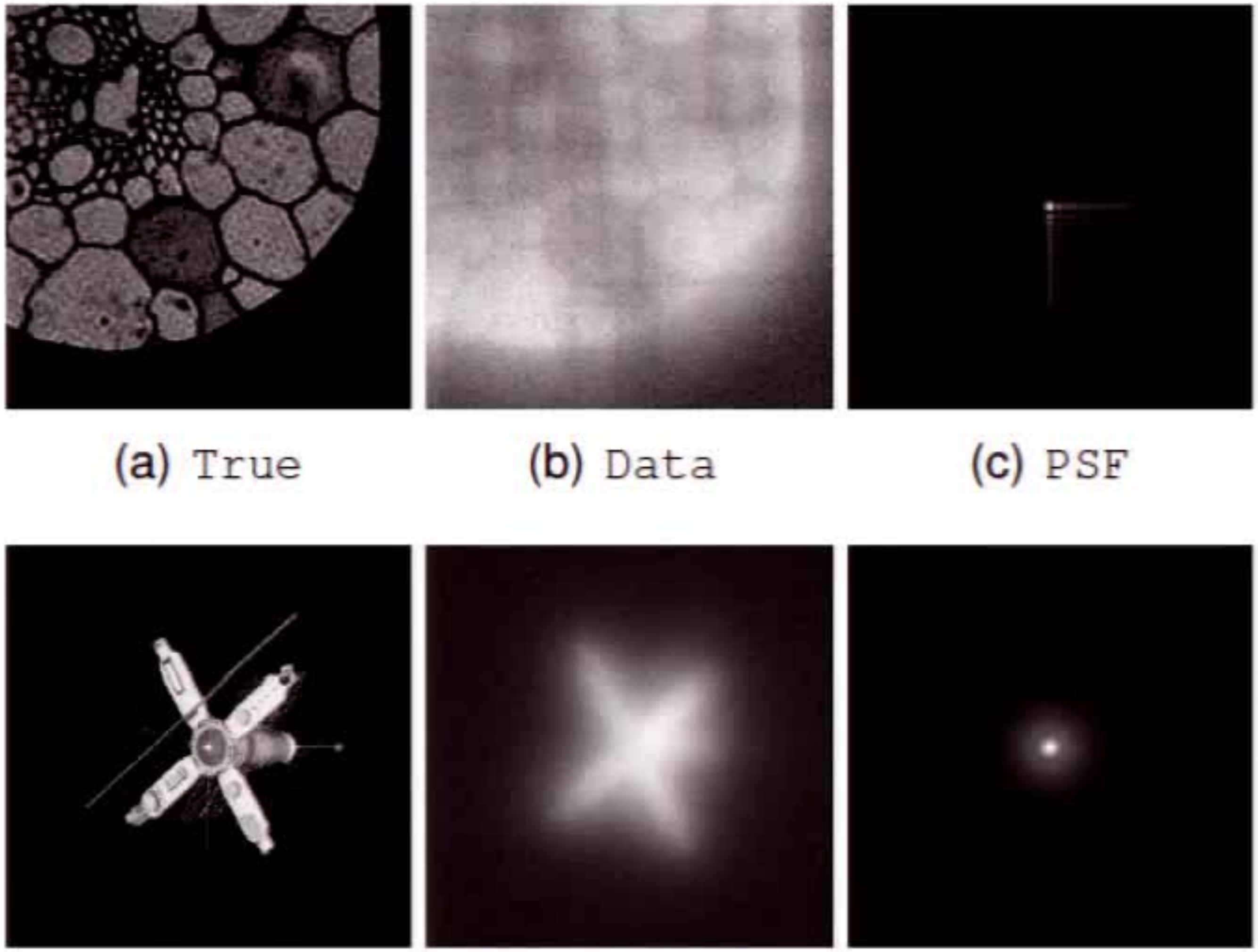
(a) `phillips` R (b) `gravity` R $d = .25$ (c) `shaw` R (d) `gravity` R $d = .75$



(e) `phillips` NR (f) `gravity` NR $d = .25$ (g) `shaw` NR (h) `gravity` NR $d = .75$

Figure: Noise revealing function $\rho(t)$.

## Two dimensional image deblurring [NPP04] Problem size $256 \times 256$

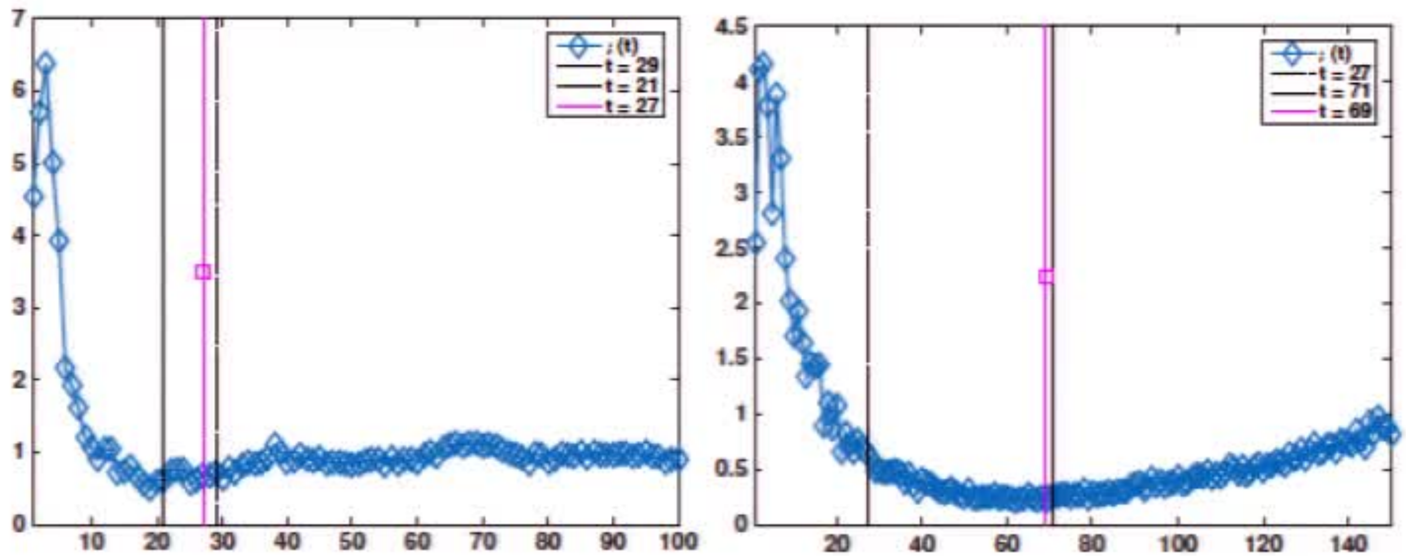


(a) True (b) Data (c) PSF

(d) True (e) Data (f) PSF

Figure: Data for grain and satellite images with blur by the given point spread function and noise level $10\%$.

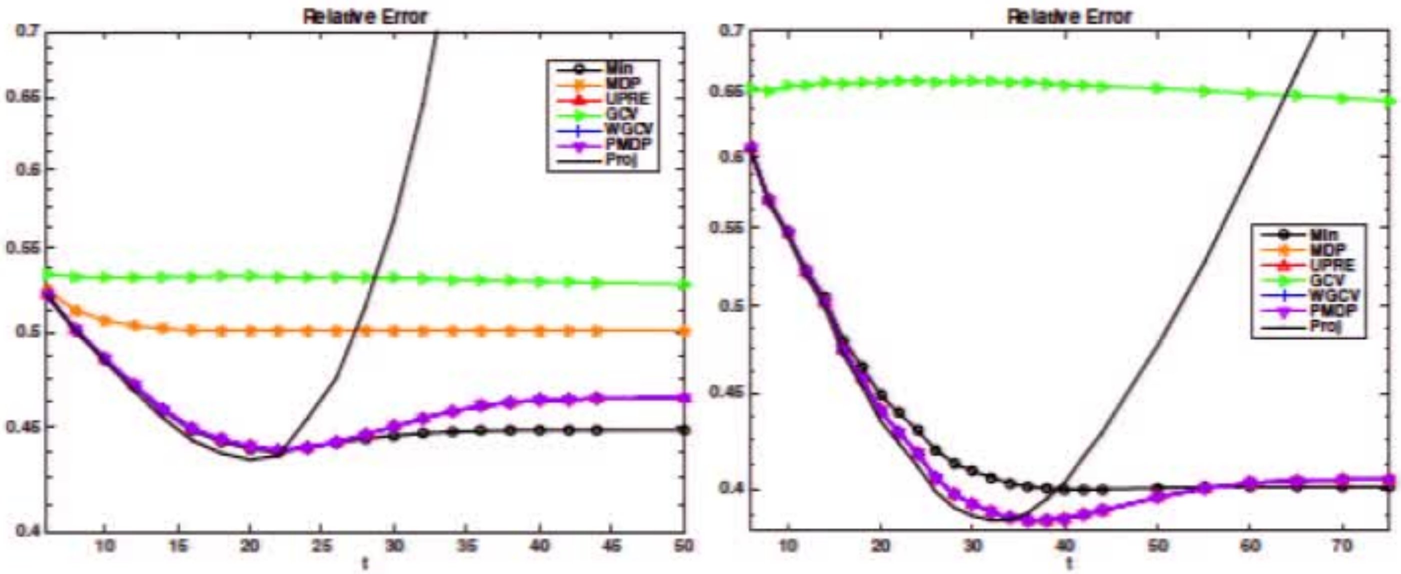
## Noise Revealing Function $\rho(t)$: comparing $t^{\text{opt}-\rho}$, $t^{\text{opt}-\mathcal{G}}$, $t^{\text{opt}-\min}$




(a) `grain` (b) `satellite`

Figure: $\rho(t)$ using $t^{\min} = 25$. Dashed-dot $t^{\text{opt}-\rho}$, magenta $t^{\text{opt}-\mathcal{G}}$ and black $t^{\text{opt}-\min}$, location of minimum for $\rho(t)$ plus step.

# Evaluating Image Quality : Relative error

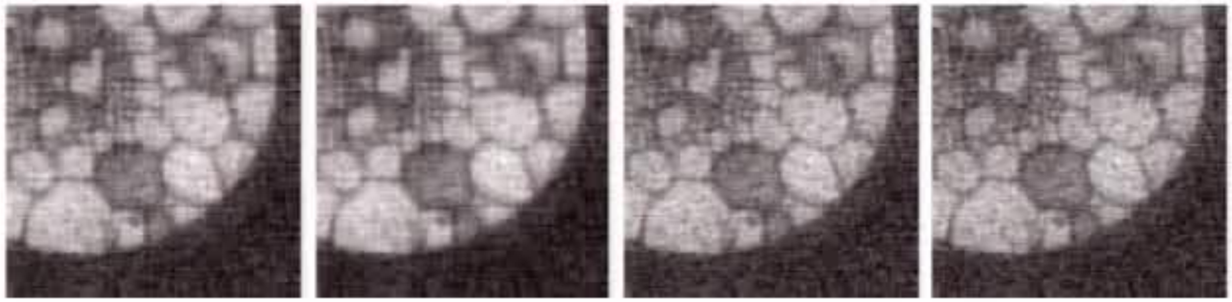



(a) `grain` RE (b) `satellite` RE

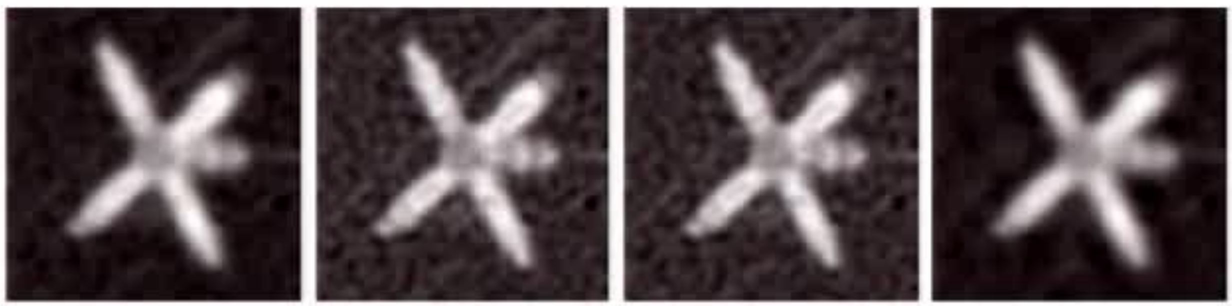
Figure: Relative error (RE) with increasing $t$. Solid line in each case is solution with projection and without regularization.

**UPRE, WGCV and PMDP outperform GCV**

## Solutions for different $t^{\text{opt}}$: (MIN, $t^{\text{opt}-\min}$, $t^{\text{opt}-\mathcal{G}}$, $t^{\text{opt}-\rho}$) Noise level $10\%$



(a) 22 (b) 21 (c) 27 (d) 29



(e) 42 (f) 71 (g) 69 (h) 27

Figure: UPRE to find $\zeta$ and comparing to solutions obtained for $t^{\text{opt}-\rho}$, $t^{\text{opt}-\min}$ and $t^{\text{opt}-\mathcal{G}}$ as compared to solution with minimum error, MIN, (a) and (e).

**Solutions inadequate**

## Iteratively Reweighted Regularization [LK83]

Minimum Support Stabilizer Regularization operator $L^{(k)}$.

$$(L^{(k)})_{ii} = ((\mathbf{x}_i^{(k-1)} - \mathbf{x}_i^{(k-2)})^2 + \beta^2)^{-1/2} \quad \beta > 0$$

Parameter $\beta$ ensures $L^{(k)}$ invertible

Invertibility use $(L^{(k)})^{-1}$ as right preconditioner for $A$

$$(L^{(k)})_{ii}^{-1} = ((\mathbf{x}_i^{(k-1)} - \mathbf{x}_i^{(k-2)})^2 + \beta^2)^{1/2} \quad \beta > 0$$

Initialization $L^{(0)} = I$, $\mathbf{x}^{(0)} = \mathbf{x}_0$. (might be $0$)

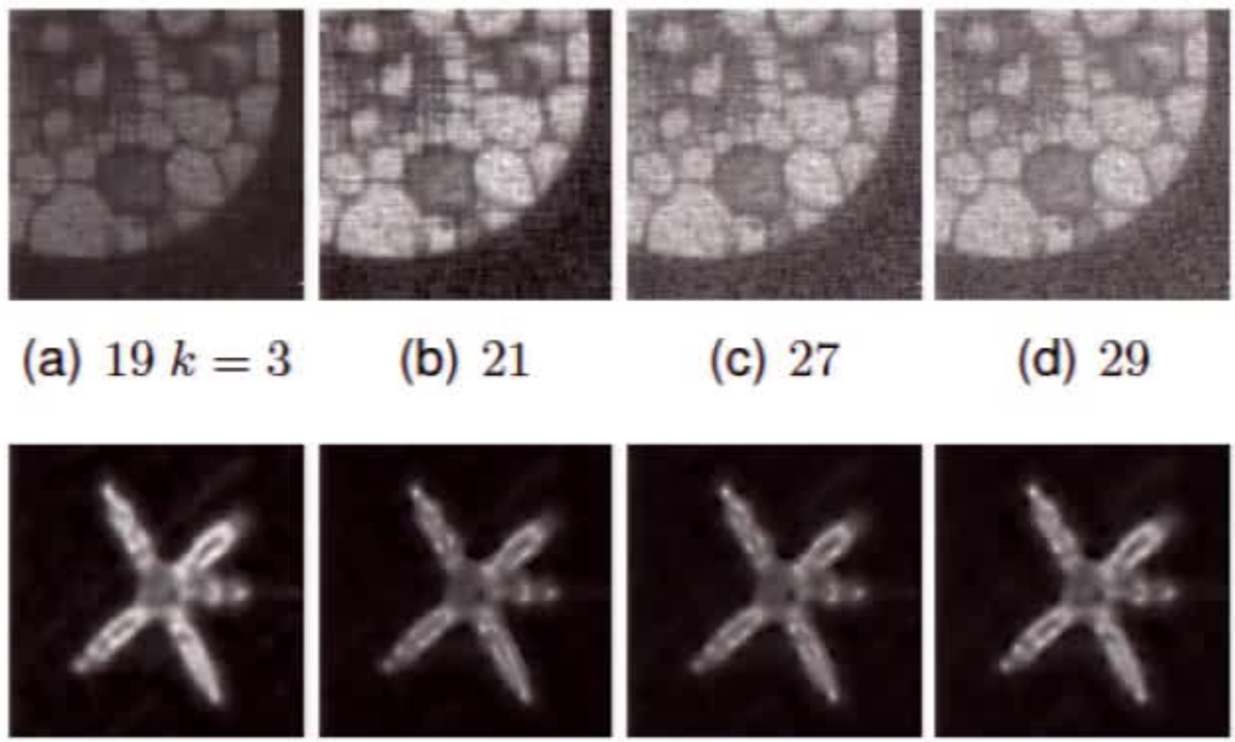
Reduced System When $\beta = 0$ and $\mathbf{x}_i^{(k-1)} = \mathbf{x}_i^{(k-2)}$ remove column $i$, matrix is $\hat{A}$.

Update Equation Solve $\hat{A}\hat{\mathbf{y}} \approx \mathbf{r} = \mathbf{b} - A\mathbf{x}^{(k-1)}$. With correct indexing set $\mathbf{y}_i = \hat{\mathbf{y}}_i$ if updated, else $\mathbf{y}_i = 0$.

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \mathbf{y}$$

**Cost of $L^{(k)}$ is minimal**

## Solutions $t^{\mathrm{opt}}$ after two steps IRR: (MIN, $t^{\mathrm{opt-min}}$, $t^{\mathrm{opt-}\mathcal{G}}$, $t^{\mathrm{opt-}\rho}$)



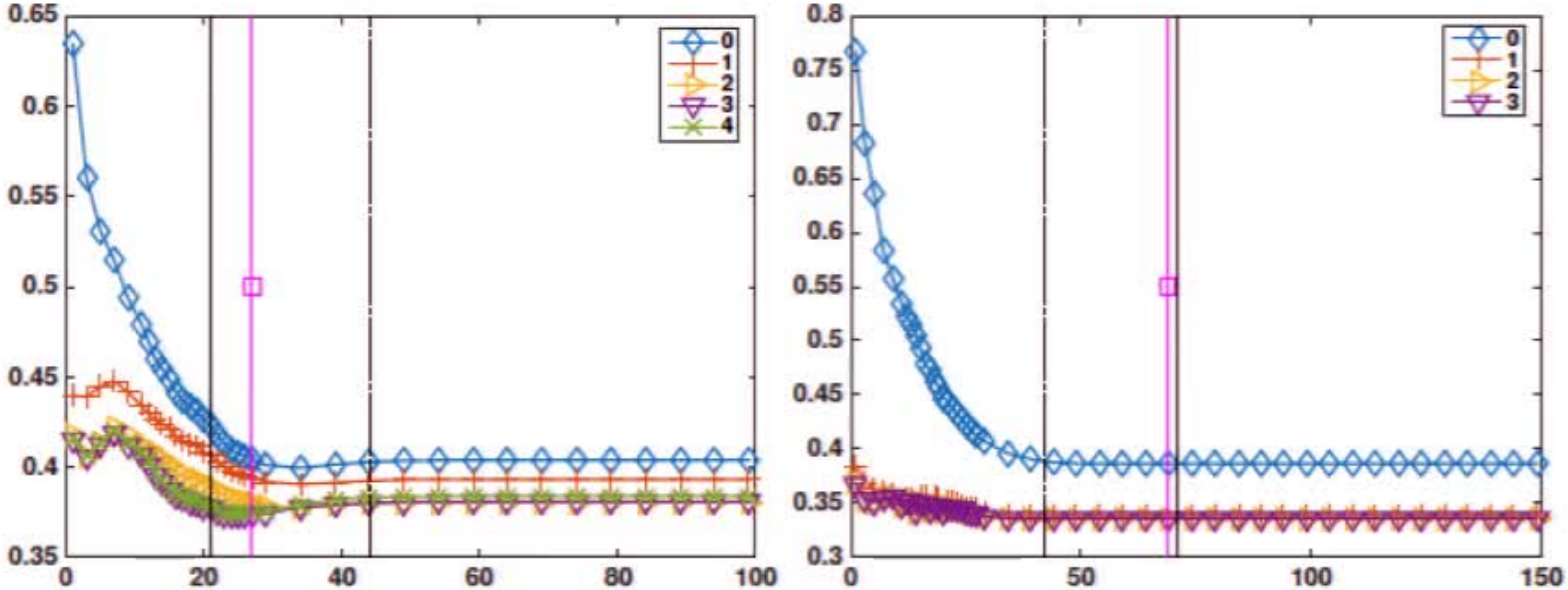
(a) $19$ $k = 3$ (b) $21$ (c) $27$ (d) $29$

(e) $35$ (f) $71$ (g) $69$ (h) $27$

Figure: IRR $k = 2$ Grain $k = 2$ MIN solution is at $t^{\mathrm{opt-min}}$, show $k = 3$.

**Solutions are stabilized less dependent on $t$**

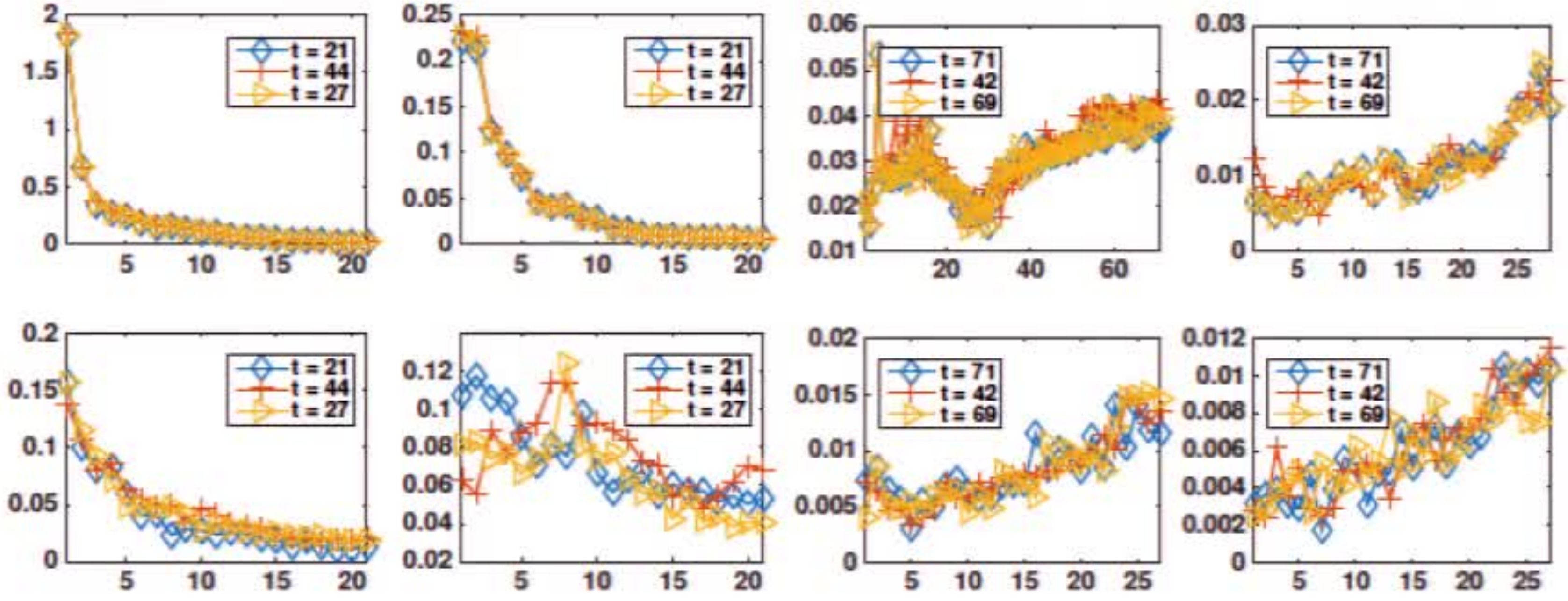
# Relative error with $k$: $5\%$ error




(a) Grain RE (b) Satellite

Figure: Relative errors decrease initially with $k$ and then increase. Dashed-dot $t^{\mathrm{opt}-\rho}$, magenta $t^{\mathrm{opt}-\mathcal{G}}$, black $t^{\mathrm{opt}-\mathrm{min}}$.

# Noise revealing function $\rho(t)$ with $k$ $5\%$ error




(a) Grain $\rho(t)$ (b) Satellite

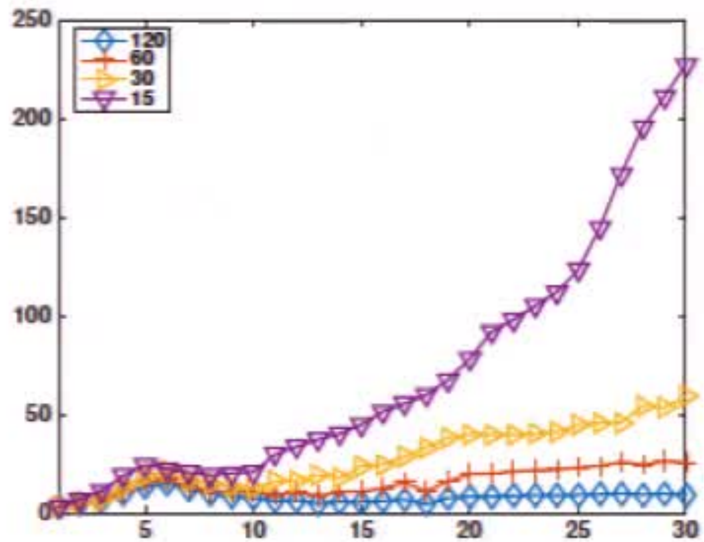
Figure: Determining $t^{\text{opt}}$ with $k$ for $5\%$ noise using $\rho(t)$. Stopping Critera : Grain $k = 4$ noise enters, use $k = 2$. Satellite $k = 3$ noise enters, use $k = 1$.

## Sparse tomographic reconstruction: walnut [HHK+15, HKK+13]

**Projection Problem** Resolution of data $164 \times 120$.

**Downsampling** $50\%$, $25\%$, eg $m = 164 \times 60$, $m = 164 \times 30$

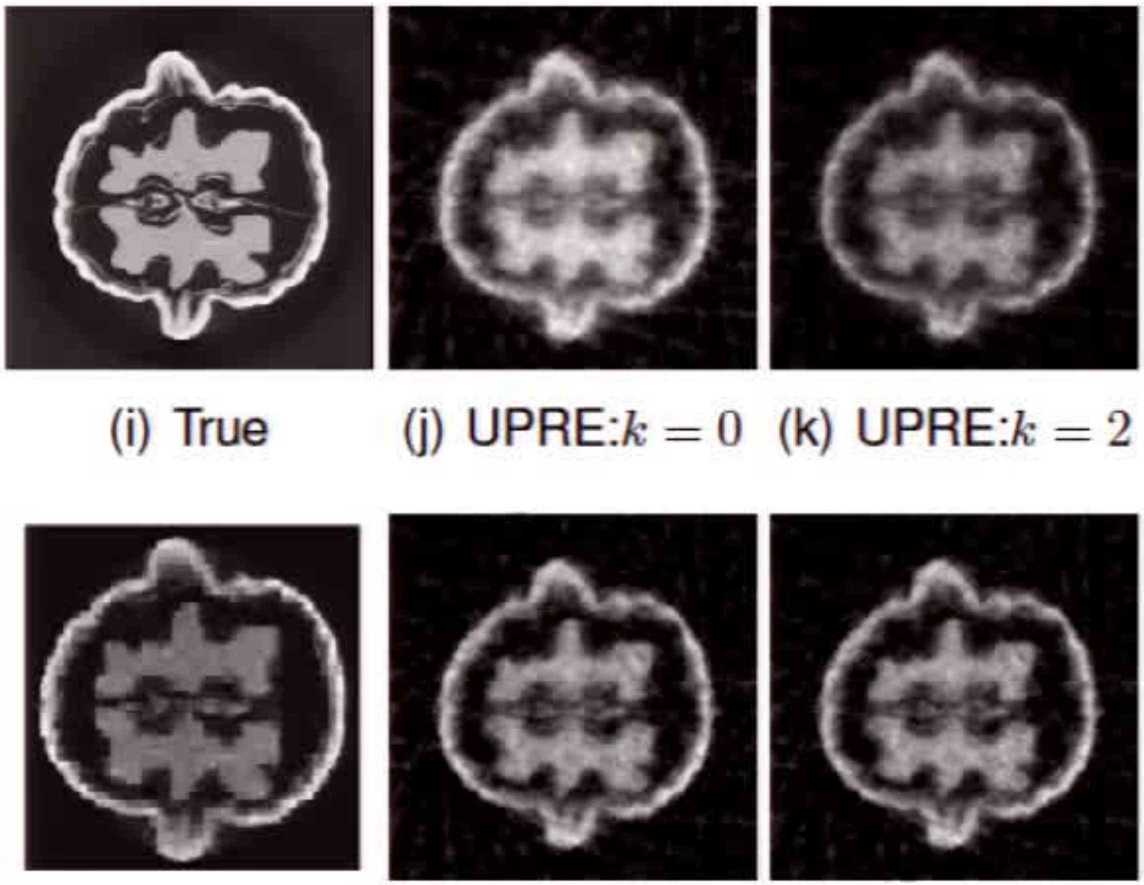
**Resolution** Full problem is $164 \times 164$




(a) $\rho(t)$ for increasing sparsity

$\rho(t)$ quite consistent for small $t$ and $m$.

## Solutions at $t^{\mathrm{opt}-\rho} = 8$



(i) True (j) UPRE:$k = 0$ (k) UPRE:$k = 2$

(l) Comparison (m) GCV:$k = 0$ (n) GCV:$k = 2$

$t^{\min} = 5$, $t^{\mathrm{opt}-\rho} = 8$, sampling at $12^\circ$ intervals, 30 projections. Comparison from [HKK+13], sparsity with prior, and resolution $256 \times 256$

**Stablized Projection no characteristic TV *blocky structure***

## Conclusions

UPRE/WGCV regularization parameter estimation explained for projected problem.

$\zeta_t^{\mathrm{opt}}$, $\lambda^{\mathrm{opt}}$ related across levels

Underdetermined problems are also solved.

Iteratively Reweighted Regularization stabilizes the projected solution

Sensitivity to choice of $t^{\mathrm{opt}}$ reduced by IRR

$t^{\mathrm{opt}}$ can be estimated using $\rho(t)$, use $t^{\mathrm{opt-min}}$ as independent of other parameters

Future

(i) extend to more realistic large scale problems

(ii) embed in TV solvers

(iii) alternative iterative methods/randomized approaches