Automatic Regularization Parameter Selections

Jodi Mead
Department of Mathematics
Boise State University

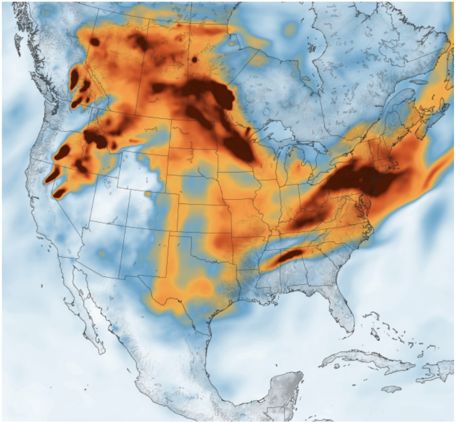
BOISE STATE UNIVERSITY

This work is supported by NSF-DMS-1043107

Outline

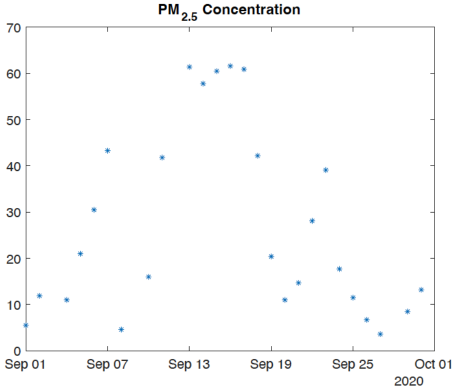
- The need for regularization (*Air quality data*)
- Automatic regularization parameter selection methods (*1D benchmarch problems*)
- χ^2 tests for regularization parameter selection
  - Linear problems (*Digital image reconstruction*)
  - Nonlinear problems (*Crosswell Tomography and Neural Networks*)
  - Multiple regularization parameters (*1D benchmark problem*)
  - Total variation regularization (*MRI image reconstruction*)

Wildfire Smoke, July 2021



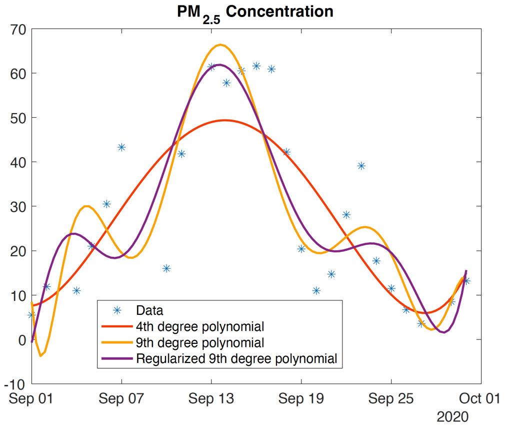
Courtesy of NASA Earth Observatory

Air Quality, Nampa, ID



Courtesy of United States Environmental Protection Agency (EPA)

Curve Fitting



The Need for Regularization

Regularization alleviates issues with

- Overfitting data
- Ill-conditioned problems
- Ill-posed problems

Least Squares

Fit data to polynomial

$$p(t) = c_n t^n + \ldots + c_1 t + c_0$$

where

$$\begin{aligned} \hat{\mathbf{c}} &= \arg\min_{\mathbf{c}} \left\{ \sum_{i=1}^{m} (d_i - p(t_i))^2 \right\} \\ &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{d} \end{aligned}$$

Tikhonov Regularization

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \left\{ \|\mathbf{d} - \mathbf{A}\mathbf{c}\|_2^2 + \alpha^2 \|\mathbf{L}\mathbf{c}\|_2^2 \right\}$$

gives $\hat{\mathbf{c}} = (\mathbf{A}^T\mathbf{A} + \alpha^2\mathbf{L}^T\mathbf{L})^{-1}\mathbf{A}^T\mathbf{d}$.

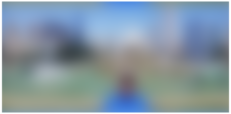
- $\mathbf{L}$ represents $\mathbf{I}$ or a derivative operator
- α is the regularization parameter

Regularization Parameter Selection

- α "small" $\Rightarrow$ data fitting
  - Not possible for ill-conditioned problems
- α "large" $\Rightarrow$ constrain the problem with $||\mathbf{Lc}||_2^2 \approx 0$
  - Derivative operator $\mathbf{L}$ gives smooth estimates
  - Can regularize with an initial estimate $||\mathbf{c} - \mathbf{c}_0||_2^2$

Digital Image Reconstruction

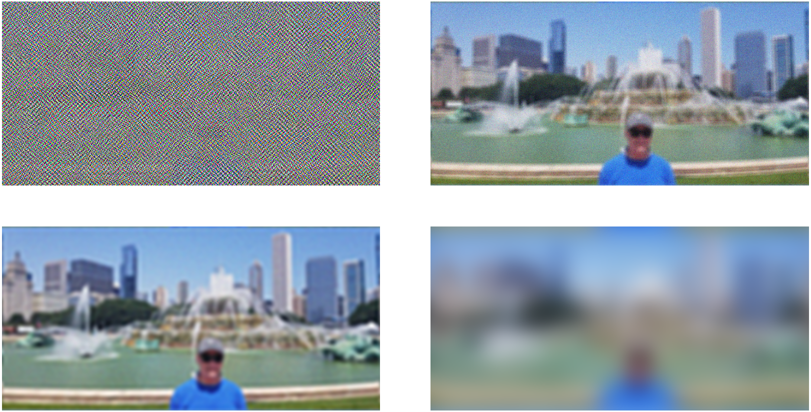
blurred, noisy   original

 $= \mathbf{A}\Big($  $\Big) + \mathbf{E}$

- $\mathbf{A}$ represents convolution of a point spread function
- $\mathbf{A}^{-1}$ does not exist or is ill-conditioned
- $\mathbf{E}$ - Additive noise.

Regularization Parameter Choices

small α



large α

Automatic Regularization Parameter Selection Methods

1. L-curve

   Plot $\|\mathbf{d} - \mathbf{A}\hat{\mathbf{c}}\|_2^2$ vs $\|\mathbf{L}\hat{\mathbf{c}}\|_2^2$ for a range of α, and choose α that minimizes both.

2. Generalized Cross Validation (GCV)

   Leave out data and choose α that minimizes prediction error in missing data.

3. Discrepancy principle

   Choose α so that $\|\mathbf{d} - \mathbf{A}\hat{\mathbf{c}}\|_2^2 < \delta$, δ represents data noise.

Maximum Likelihood Estimation (MLE)

$$\mathbf{d} = \mathbf{A}\mathbf{c} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$$

gives Likelihood function $L(\mathbf{c}|\mathbf{d}) \propto e^{-1/2\sigma^2\|\mathbf{d}-\mathbf{A}\mathbf{c}\|_2^2}$

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} \left\{ e^{-1/2\sigma^2\|\mathbf{d}-\mathbf{A}\mathbf{c}\|_2^2} \right\}$$

Residual properties

$$\mathbb{E}(\|\mathbf{d} - \mathbf{A}\mathbf{c}\|_2^2) = m\sigma^2$$

suggests finding roots

$$f(\alpha) = \|\mathbf{d} - \mathbf{A}\mathbf{c}(\alpha)\|_2^2 - m\sigma^2 = 0.$$

However,

$$\hat{f}(\alpha) = \|\mathbf{d} - \mathbf{A}\hat{\mathbf{c}}(\alpha)\|_2^2 - m\sigma^2$$

is biased so choosing α by solving $\hat{f}(\alpha) = 0$ leads to oversmoothing (Discrepancy principle).

Effective degrees of freedom in Tikhonov regularization

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \;\left\{\|\mathbf{d} - \mathbf{A}\mathbf{c}\|_2^2 + \alpha^2\|\mathbf{L}\mathbf{c}\|_2^2\right\}$$

gives ridge regression estimator

$$\hat{\mathbf{c}} = (\mathbf{A}^T\mathbf{A} + \alpha^2\mathbf{L}^T\mathbf{L})^{-1}\mathbf{A}^T\mathbf{d}.$$

Predictors are

$$\mathbf{A}\hat{\mathbf{c}} = \mathbf{N}(\alpha)\mathbf{d}, \quad \mathbf{N}(\alpha) = \mathbf{A}(\mathbf{A}^T\mathbf{A} + \alpha^2\mathbf{L}^T\mathbf{L})^{-1}\mathbf{A}^T$$

with*

$$\mathbb{E}(\|\mathbf{d} - \mathbf{A}\hat{\mathbf{c}}\|_2^2) = (m - \mathrm{tr}\mathbf{N}(\alpha))\sigma^2$$

*Hall et. al, 1987.

χ^2 test for Tikhonov regularization

$$\frac{\|\mathbf{d}-\mathbf{A}\hat{\mathbf{c}}\|_2^2}{\sigma^2} \sim \chi^2_{m-n}$$

Issue if $m \le n$, instead use regularized residual*

$$\frac{\|\mathbf{d}-\mathbf{A}\hat{\mathbf{c}}\|_2^2}{\sigma^2} + \alpha^2\|\mathbf{L}\hat{\mathbf{c}}\|_2^2 \sim \chi^2_{m-n+p}$$

$p =$rank$(\mathbf{L})$

True for large m, regardless of error distributions

*Mead 2008, 2013; Mead and Renaut 2009.

Comparison of Methods

Problem	noise	L-Curve	GCV	UPRE	χ^2
`shaw`	0.166	0.0357(0.008)	0.0344(0.013)	0.0161(0.008)	0.0120(0.004)
`shaw`	0.166	0.0354(0.008)	0.0342(0.013)	0.0162(0.008)	0.0125(0.004)
`phillips`	0.128	0.0379(0.011)	0.0268(0.012)	0.0298(0.011)	0.0225(0.006)
`phillips`	0.128	0.0379(0.011)	0.0283(0.013)	0.0297(0.011)	0.0229(0.006)
`ilaplace`	0.069	0.0367(0.008)	0.0244(0.014)	0.0194(0.010)	0.0169(0.007)
`ilaplace`	0.069	0.0373(0.009)	0.0217(0.012)	0.0198(0.011)	0.0172(0.008)

Mean and Standard Deviation of Risk with $n = 64$ *over* 500 *runs*

Regularization tools: A Matlab Toolbox, PC Hansen

Nonlinear Problems

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \ \left\{ \|\mathbf{d} - \mathbf{F}(\mathbf{c})\|_2^2 + ||\mathbf{L}\left(\mathbf{c} - \mathbf{c}_0\right)||_2^2 \right\}$$

Gauss-Newton optimization gives nonlinear χ^2 test at kth iterate:

$$\|\mathbf{P}_k^{-1/2}\left(\mathbf{r}_k + \mathbf{J}_k \triangle \mathbf{c}_k\right)\| \sim \chi_m^2$$

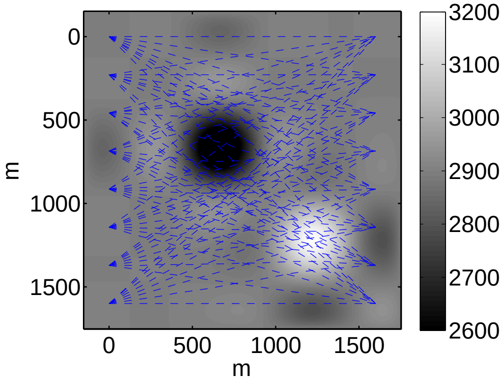
with $\quad \mathbf{r}_k = \mathbf{d} - \mathbf{F}(\mathbf{c}_k) + \mathbf{J}_k \mathbf{c}_k$

$$\triangle \mathbf{c}_k = \mathbf{c}_k - \mathbf{c}_0$$

$$\mathbf{P}_k = \mathbf{J}_k \mathbf{L} \mathbf{J}_k^T + \mathbf{I}$$

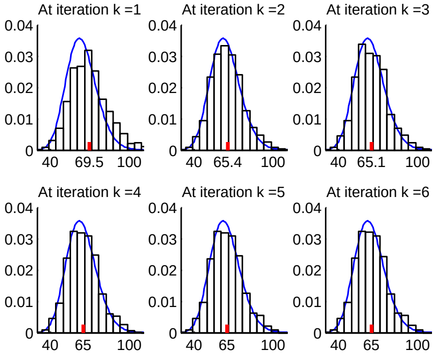
Mead and Hammerquist, 2013.

Nonlinear Cross-Well tomography

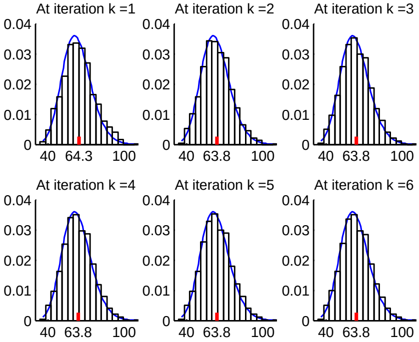


Numerical Validation of Nonlinear χ^2 tests

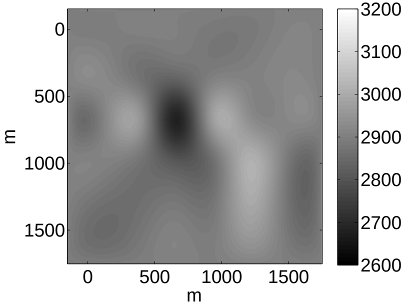
$$\left\|\mathbf{P}_k^{-1/2}\left(\mathbf{r}_k + \mathbf{J}_k \triangle \mathbf{c}_k\right)\right\| \sim \chi^2_{64}$$



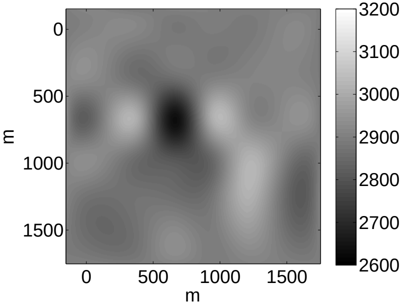
$$\left\|\mathbf{P}_k^{-1/2}\left(\mathbf{r}_k + \mathbf{J}_k \triangle \mathbf{c}_k\right)\right\| \sim \chi^2_{63}$$



Recovered velocity structure



Discrepancy Principle



χ^2 method

Neural Networks

Stack training feature-label pairs $(\mathbf{y}_i, \mathbf{c}_i)$, $i = 1, \ldots, s$:

$$\mathbf{Y}_0 \in \mathbb{R}^{s \times n}, \ \mathbf{C} \in \mathbb{R}^{s \times m}$$

Residual Neural Network (ResNet) forward propogation:

$$\mathbf{Y}_{j+1} = \mathbf{Y}_j + \sigma(\mathbf{Y}_j \mathbf{K}_j + b_j,) \quad j = 0, \ldots, N-1$$

with $\mathbf{C}^{pred} = \mathbf{h}(\mathbf{Y}_N \mathbf{W} + \mathbf{e}_s \boldsymbol{\mu}^T)$.

σ - activiation function, e.g. $\tanh(\mathbf{Y})$

$\mathbf{h}(\mathbf{X})$ - hypothesis function, e.g. $e^{\mathbf{X}}/(e^{X}\mathbf{e}_m)$

Learning Problem*

Inverse problem for the weights and bias

$$\min_{\mathbf{K}_j,\mathbf{W},b_j,\boldsymbol{\mu}} \left\{ \|\mathbf{C}^{pred} - \mathbf{C}\|_F^2 + \alpha_1 \sum_{j=1}^{N-1} \|\mathbf{K}_j - \mathbf{K}_{j-1}\|_F^2 + \alpha_2 \sum_{j=1}^{N-1} (b_j - b_{j-1})^2 \right\}$$

Magnitude of α_1 and α_2 control the extent to which:

- weights and bias are smoothly varying between layers
- overfitting occurs
- different weight values give the same classification

---

* Future work

Multiple Regularization Parameters

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \left\{ \|\mathbf{d} - \mathbf{A}\mathbf{c}\|_2^2 + \|\mathbf{W}\mathbf{L}\mathbf{c}\|_2^2 \right\}$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{W}_d \end{bmatrix}, \quad \mathbf{W}_i = \alpha_i \mathbf{I}_{m_i}$$

Multiple χ^2 Tests

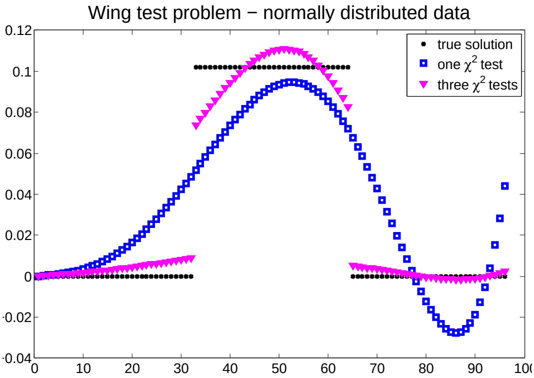
$$\|\mathbf{d} - \mathbf{A}\hat{\mathbf{c}}\|_2^2 + \|\mathbf{W}\mathbf{L}\hat{\mathbf{c}}\|_2^2 = k_1^2 + \ldots + k_m^2,$$

$k_i = (\mathbf{P}^{-1/2}\mathbf{r})_i$, gives d χ^2 Tests:

$$\begin{aligned} k_1^2 + \ldots + k_{m_1}^2 &= m_1 \\ k_{m_1+1}^2 + \ldots + k_{m_1+m_2}^2 &= m_2 \\ &\vdots \\ k_{m+1-m_d}^2 + \ldots + k_m^2 &= m_d \end{aligned}$$

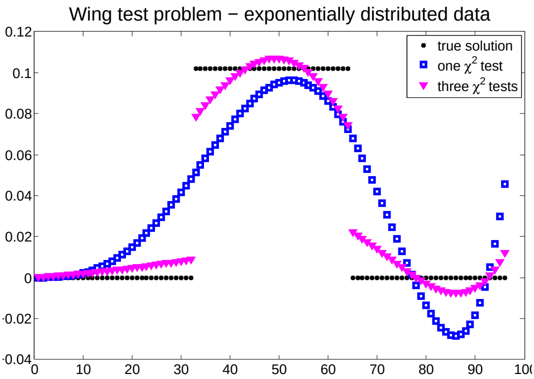
with $\sum_{i=1}^{d} m_i = m$.

Three Regularization Parameters, Normal Data



Regularization tools: A Matlab Toolbox, PC Hansen

Three Regularization Parameters, Exponential Data

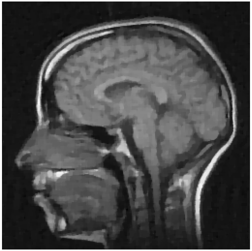


Regularization tools: A Matlab Toolbox, PC Hansen

Total Variation Regularization (TV)

$$\hat{\mathbf{c}} \in \arg\min_{\mathbf{c}} \left\{ \frac{\lambda}{2} \|\mathbf{d} - \mathbf{A}\mathbf{c}\|_2^2 + \|\mathbf{L}_1 \mathbf{c}\|_1 \right\}$$

$\lambda = 500$



$\lambda = 50000$

$\lambda = 500000$

Automatic TV regularization parameter selection

Approaches

- TV function viewed or approximated with a quadratic functional: **L-curve**, **Discrepancy principle**[*], **Unbiased Predictive Risk Estimator (UPRE)**[**], **Generalized Cross Validation (GCV)**[***].
- Predictive risk estimator: **Stein's unbiased risk estimate (SURE)**[†].

[*] Wen et. al, 2012; [**] Lin et. al, 2012; [***] Liao et. al, 2009

[†] Deledalle et. al, 2014

Degrees of Freedom

Tikhonov regularization defines smoothing matrix: $\mathbf{A}\hat{\mathbf{c}} = \mathbf{N}\mathbf{d}$

Nonlinear smoothers (e.g. TV) have:

$$\mathbf{A}\hat{\mathbf{c}} = \delta(\mathbf{d})$$

Degrees of freedom of δ are given by*

$$df(\mathbf{A}\hat{\mathbf{c}}) = \sum_{i=1}^{m} \mathsf{cov}(\hat{\mathbf{c}}_i, \mathbf{d}_i)/\sigma^2$$

e.g. $\mathrm{df}(\mathbf{A}\hat{\mathbf{c}}) = \mathrm{tr}(\mathbf{N})$

*Efron, 2004.

Degrees of Freedom for TV *,**

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \left\{ \frac{\lambda}{2} \|\mathbf{d} - \mathbf{A}\mathbf{c}\|_2^2 + \|\mathbf{L}_1 \mathbf{c}\|_1 \right\}$$

$$\begin{aligned} df(\mathbf{A}\hat{\mathbf{c}}) &= \sum_{i=1}^{m} \mathsf{cov}((\hat{\mathbf{c}})_i, \mathbf{d}_i)/\sigma^2 \\ &= \mathbb{E}[dim(\mathbf{A}(null(\mathbf{L}_{-\mathcal{A}})))], \quad \mathcal{A} = \{i : (\mathbf{D}\hat{\mathbf{x}}_{tv})_i \neq 0\} \end{aligned}$$

e.g. $nullity((\mathbf{L}_1)_{-\mathcal{A}}) = n \Rightarrow df(\mathbf{A}\hat{\mathbf{c}}) = n$

---

*Tibshirani 2012; **Dossal 2013.

χ^2 distribution for TV

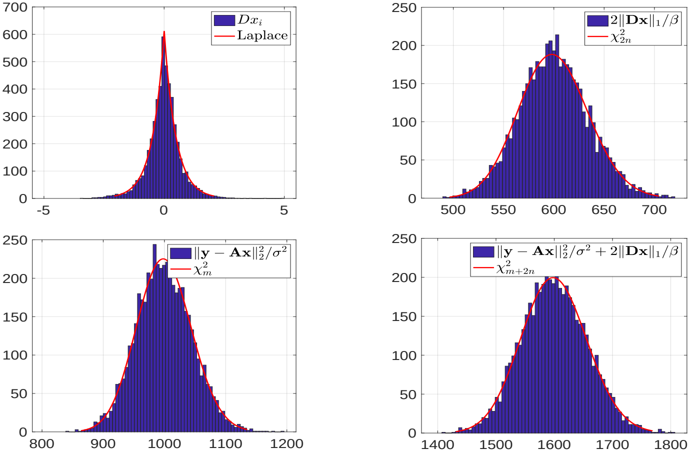
If $z_i \sim$ Laplace(θ, β) all independent, then

$$\sum_{i=1}^{n} \frac{2|z_i - \theta|}{\beta} \sim \chi^2_{2n}.$$

Since the TV functional is differentially Laplacian

$$\frac{\|\mathbf{d} - \mathbf{A}\mathbf{c}\|_2^2}{\sigma^2} + \frac{2\|\mathbf{L}_1\mathbf{c}\|_1}{\beta} \sim \chi^2_{m+2n}$$

Histograms illustrating χ^2 distribution for TV



χ^2 Test for TV

$$\frac{\|\mathbf{d}-\mathbf{A}\hat{\mathbf{c}}(\lambda)\|_2^2}{\sigma^2}+\frac{2\|\mathbf{L}_1\hat{\mathbf{c}}(\lambda)\|_1}{\beta}\sim\chi^2_{m-df(\mathbf{A}\hat{\mathbf{c}}(\lambda))+df(\mathbf{L}_1\hat{\mathbf{c}}(\lambda))}$$

Theorem

*Suppose that $(\mathbf{d}-\mathbf{A}\mathbf{c})_i\sim\mathcal{N}(0,\sigma)$ and $(\mathbf{L}_1\mathbf{c})_i\sim$ Laplace(θ,β) with $\mathbf{A}$ and $\mathbf{L}_1$ full rank, then

$$\frac{\|\mathbf{d}-\mathbf{A}\hat{\mathbf{c}}(\lambda)\|_2^2}{\sigma^2}+\frac{2\|\mathbf{L}_1\hat{\mathbf{c}}(\lambda)\|_1}{\beta}\sim\chi^2_m$$

or $\|\mathbf{d}-\mathbf{A}\hat{\mathbf{c}}(\lambda)\|_2^2+\frac{2}{\lambda}\|\mathbf{L}_1\hat{\mathbf{c}}(\lambda)\|_1\approx m\sigma^2$.

*Mead 2020

Numerical Tests - Evaluating Image Quality

MRI image filtered with a 15×15 uniform blur

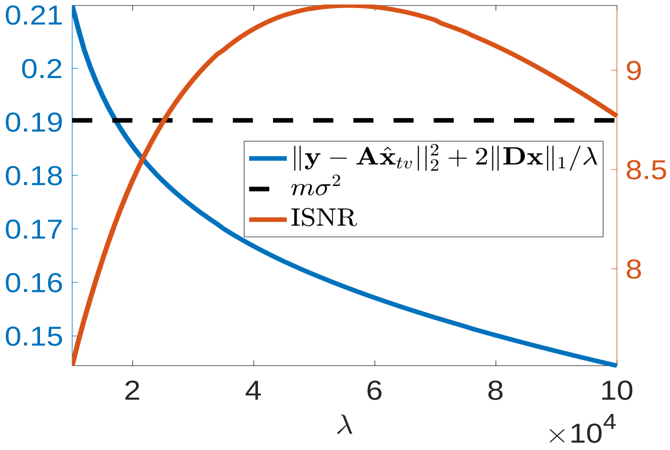
Input noise:

$$\textbf{BSNR} = 20 \log_{10} \frac{\|\mathbf{y} - \mathbf{A}\mathbf{c}\|_2}{m\sigma}$$

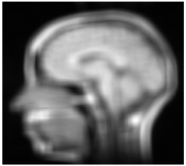
Recovered image quality:

$$\textbf{ISNR} = 20 \log_{10} \frac{\|\mathbf{d} - \mathbf{c}\|_2}{\|\hat{\mathbf{c}} - \mathbf{c}\|_2}$$

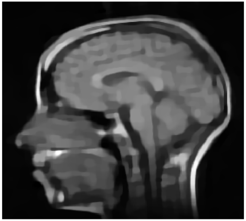
MRI BSNR = 40; χ^2 ISNR = 8.22; Max ISNR = 9.33



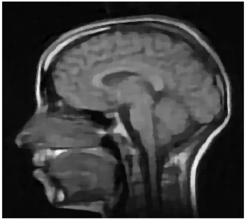
$\text{BSNR} = 40$



$\text{MAP ISNR} = 5.67$



χ^2 ISNR $= 8.22$



$\text{Maximum ISNR} = 9.30$

BSNR $= 30$

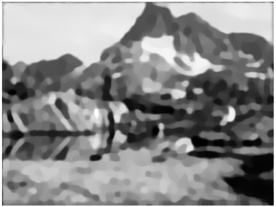
MAP ISNR $= 2.87$

χ^2 ISNR $= 5.36$

Maximum ISNR $= 5.64$

BSNR $= 20$

MAP ISNR $= 0.96$



χ^2 ISNR $= 2.10$

Maximum ISNR $= 2.28$

BSNR	MAP estimate	Discrepancy	χ^2 test	Maximum
Camerman $(m = n = 256)$				
40	5.0019	7.0914	7.1123	7.6329
30	2.8671	5.3398	5.3556	5.6426
20	1.8228	3.6031	3.6241	4.0441
MRI $(m = n = 256)$				
40	5.6696	8.1718	8.2201	9.2978
30	3.2113	5.9225	5.9510	6.5944
20	1.7017	3.8260	3.8474	4.5641
Mountain $(m = 480, n = 640)$				
40	2.8357	4.0904	4.0938	4.3440
30	1.6074	2.9915	2.9945	3.1432
20	0.9594	2.1004	2.1049	2.2803

Summary and Conclusions

- Regularization can prevent overfitting of data and make a problem well posed.
- The discrepancy principle is a simple method for automatically choosing a regularization parameter, but relies on inaccurate estimates of degrees of freedom.
- We have developed a framework for automatic and efficient selection of regularization parameters based on χ^2 properties, with theoretically justified degrees of freedom, and applied to
  - L2 or Tikhonov regularization, Ridge Regression
  - L1 or Total Variation regularization, LASSO
  - Nonlinear problems and varying regulation parameters
- The χ^2 method has been used for digital imaging problems, and problems in the geosciences. It has potential to improve the training of neural networks.