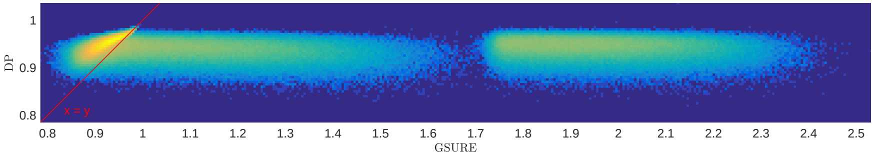
The Ill-Posedness Always Rings Twice

Risk Estimators for Choosing Regularization Parameters in Inverse Problems



Felix Lucka, University College London, f.lucka@ulc.ac.uk
joint with: Katharina Proksch, Christoph Brune, Nicolai Bissantz, Martin Burger, Holger Dette & Frank Wübbeling

Centre for Medical Image Computing

Cambridge & Heriot-Watt Workshop, Feb 20, 2017.

UCL CENTRE FOR INVERSE PROBLEMS

Discrete inverse problem:

$$y = Ax^* + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_m)$$

Variational regularization:

$$\hat{x}_\alpha(y) = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2}\|Ax - y\|_2^2 + \alpha R(x),$$

R convex such that the minimizer is unique for $\alpha > 0$.

Notation and Setting

Discrete inverse problem:

$$y = Ax^* + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_m)$$

Variational regularization:

$$\hat{x}_\alpha(y) = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2}\|Ax - y\|_2^2 + \alpha R(x),$$

R convex such that the minimizer is unique for $\alpha > 0$.

Every talk: "How did you choose α?"

An old problem

A problem as old as inverse problems / robust statistical inference.

- Ideal parameter choice:

$$\alpha^* := \underset{\alpha \geqslant 0}{\operatorname{argmin}} \; \|\hat{x}_\alpha(y) - x^*\|_2^2$$

! obviously not available (oracle solution)

- Many different approaches proposed. Focus here: Strategies that need accurate estimate of noise variance σ^2.

- Classical example: discrepancy principle:

$$find \quad \alpha \quad s.t. \quad \|A\hat{x}_\alpha(y) - y\|_2^2 = m\sigma^2.$$

✓ robust and easy-to-implement for many applications
! typically over-estimates α^*.

Stein's unbiased risk estimator (SURE) & GSURE

We want to minimize the quadratic risk function

$$\mathrm{R}_{\mathrm{SURE}}(\alpha) := \mathbb{E}\left[\|Ax^* - A\hat{x}_\alpha(y)\|_2^2\right],$$

but as $\mathrm{R}_{\mathrm{SURE}}$ depends on x^*, we replace it by an unbiased estimate:

$$\mathrm{SURE}(\alpha, y) := \|y - A\hat{x}_\alpha(y)\|_2^2 - m\sigma^2 + 2\sigma^2 \mathrm{df}_\alpha(y), \quad \mathrm{df}_\alpha(y) = \mathrm{tr}\left(\nabla_y \cdot A\hat{x}_\alpha(y)\right),$$

where unbiased means: $\mathbb{E}\left[\mathrm{SURE}(\alpha, y)\right] = \mathrm{R}_{\mathrm{SURE}}(\alpha)$

Risk in the domain, not in the image of the operator A:

$$\mathrm{R}_{\mathrm{GSURE}}(\alpha) := \mathbb{E}\left[\|\Pi(x^* - \hat{x}_\alpha(y))\|_2^2\right], \qquad \Pi := A^+ A$$

$$\mathrm{GSURE}(\alpha, y) := \|x_{\mathrm{ML}}(y) - \hat{x}_\alpha(y)\|_2^2 - \sigma^2 \mathrm{tr}\left((AA^*)^+\right) + 2\sigma^2 \mathrm{gdf}_\alpha(y)$$

$$\mathrm{gdf}_\alpha(y) := \mathrm{tr}((AA^*)^+ \nabla_y A\hat{x}_\alpha(y)), \qquad x_{\mathrm{ML}} = A^+ y = A^*(AA^*)^+ y,$$

GSURE seems more appropriate for ill-posed problems, since properties in data space do not tell much about the reconstruction quality!

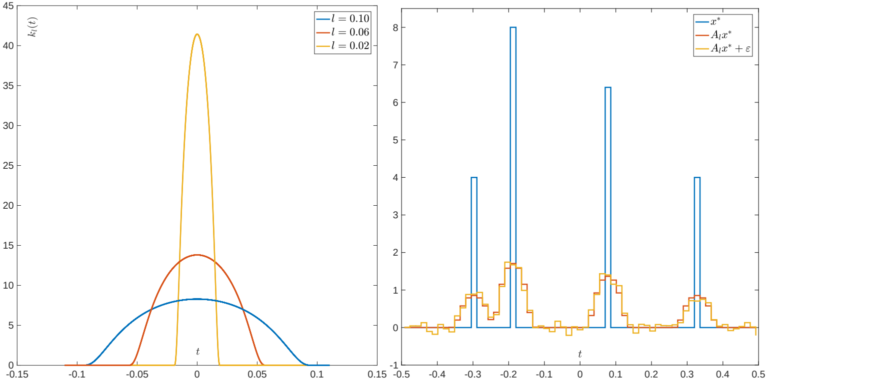
Recent works & a statistical perspective

Recently, a lot of work on risk estimators in imaging and inverse problems: *Blu, Chesneau, Deledalle, Dossal, Elad, Eldar, Fadili, Giryes, Kachour, Kocher, Luisier, Morel, Peyré, Ramani, Unser, Vaiter, Van De Ville, Wang*

Our interest is a statistical perspective:

- All parameter choice rules depend on data y and hence on random ε.
- Therefore, $\hat{\alpha}_{\mathsf{DP}}$, $\hat{\alpha}_{\mathsf{SURE}}$ and $\hat{\alpha}_{\mathsf{GSURE}}$ are random variables.
- Characteristics of their probability distributions?
- Distributions or error measures $dist(x^*, x_{\hat{\alpha}})$?

A simple example



$$y_\infty(s) = A_{\infty,l} x^*_\infty, \quad x^*_\infty(t) := \sum_{i=1}^{4} a_i \delta\left(b_i\right)$$

1D periodic convolution, kernel width l, mass-preserving discretization into ONB of piecewise constant functions

$$y_m = A_{m,l} x^*_m + \varepsilon_m, \qquad \varepsilon_m \sim \mathcal{N}(0, \sigma^2 I_m)$$

Quadratic regularization & singular system

Quadratic regularization leading to explicit, linear estimator:

$$\hat{x}_\alpha(y) = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2}\|Ax - y\|_2^2 + \frac{\alpha}{2}\|x\|_2^2 = (A^*A + \alpha I)^{-1} A^* y$$

Switch to singular system to analyse:

$$A = U \Sigma V^*, \qquad 1 = \gamma_1 \geq \ldots \geq \gamma_m > 0$$

$$y_i = \langle u_i, y \rangle, \qquad x_i^* = \langle v_i, x^* \rangle, \qquad \tilde{\epsilon}_i = \langle u_i, \epsilon \rangle$$

$$y = Ax + \varepsilon \Leftrightarrow y_i = \gamma_i x_i^* + \tilde{\varepsilon}_i, \qquad \tilde{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2)$$

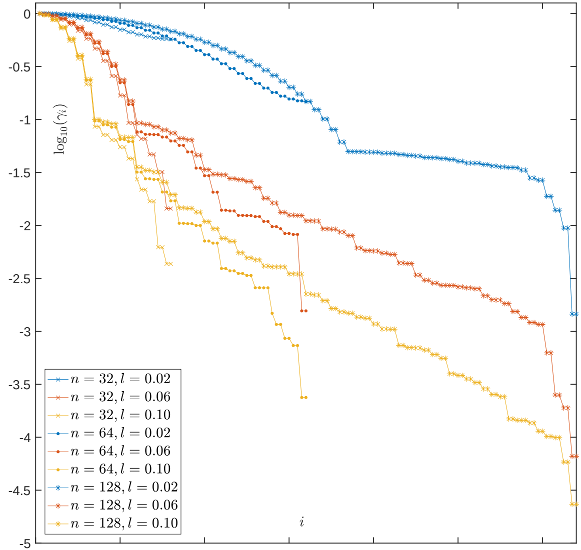
Risk representation for quadratic regularization

$$\begin{aligned}\mathrm{DP}(\alpha, y) &:= \|A\hat{x}_\alpha(y) - y\|_2^2 - m\sigma^2 \\ &= \sum_{i=1}^{m} \frac{\alpha^2}{(\gamma_i^2 + \alpha)^2} y_i^2 - m\sigma^2\end{aligned}$$

$$\begin{aligned}\mathrm{SURE}(\alpha, y) &= \|y - A\hat{x}_\alpha(y)\|_2^2 - m\sigma^2 + 2\sigma^2 \mathrm{df}_\alpha(y) \\ &= \sum_{i=1}^{m} \frac{\alpha^2}{(\gamma_i^2 + \alpha)^2} y_i^2 - m\sigma^2 + 2\sigma^2 \sum_{i=1}^{m} \frac{\gamma_i^2}{\gamma_i^2 + \alpha}\end{aligned}$$

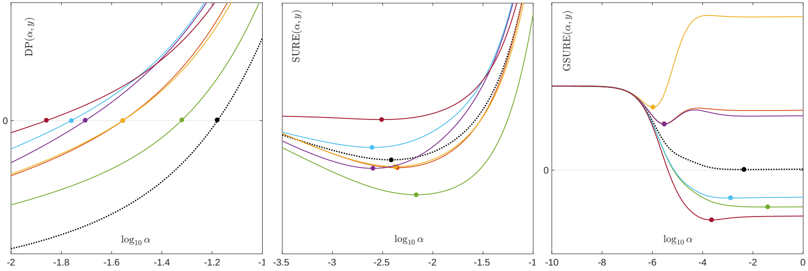
$$\begin{aligned}\mathrm{GSURE}(\alpha, y) &= \|x_{\mathrm{ML}}(y) - \hat{x}_\alpha(y)\|_2^2 - \sigma^2 \mathrm{tr}\left((AA^*)^+\right) + 2\sigma^2 \mathrm{gdf}_\alpha(y) \\ &= \sum_{i=1}^{r} \left(\frac{1}{\gamma_i} - \frac{\gamma_i}{\gamma_i^2 + \alpha}\right)^2 y_i^2 - \sigma^2 \sum_{i=1}^{r} \frac{1}{\gamma_i^2} + 2\sigma^2 \sum_{i=1}^{r} \frac{1}{\gamma_i^2 + \alpha}\end{aligned}$$

Singular values for simple example



Singular values γ_i of A_l for different choices of m and l.

Example of risk functions for simple example



(a) (b) (c)

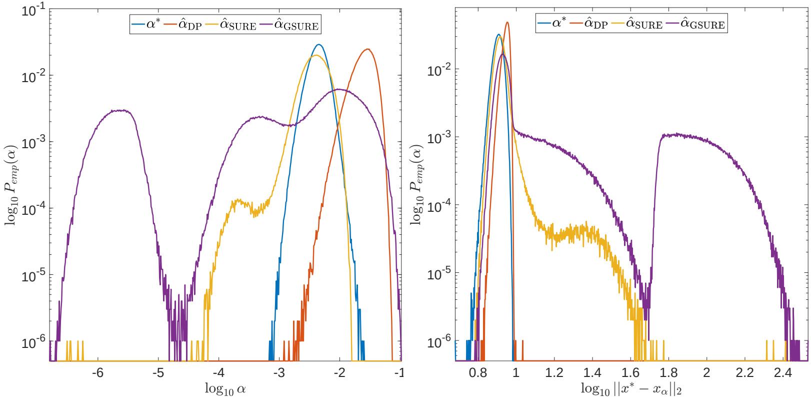
(a) $R_{\mathsf{DP}}(\alpha) = \|A\hat{x}_\alpha(Ax^*) - Ax^*\|_2^2 - m\sigma^2$ vs. 6 realizations of $\mathsf{DP}(\alpha, y) = \|A\hat{x}_\alpha(y) - y\|_2^2 - m\sigma^2$

(b) $R_{\mathsf{SURE}}(\alpha) = \mathbb{E}\left[\|Ax^* - A\hat{x}_\alpha(y)\|_2^2\right]$ vs. 6 realizations of $\mathsf{SURE}(\alpha, y) = \|y - A\hat{x}_\alpha(y)\|_2^2 - m\sigma^2 + 2\sigma^2 \mathrm{df}_\alpha(y)$.

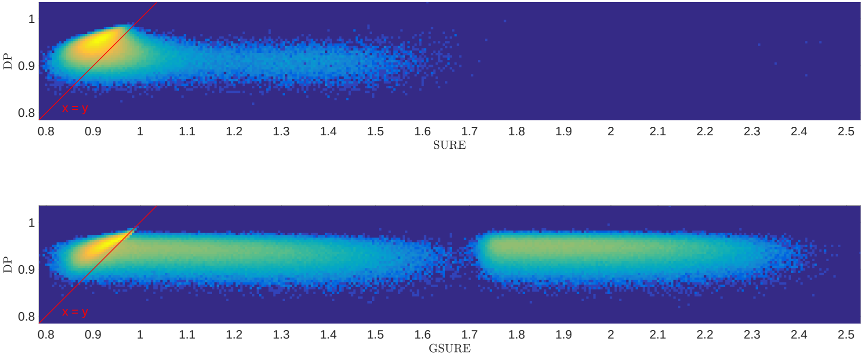
(c) $R_{\mathsf{GSURE}}(\alpha) = \mathbb{E}\left[\|\Pi(x^* - \hat{x}_\alpha(y))\|_2^2\right]$ vs. 6 realizations of $\mathsf{GSURE}(\alpha, y) = \|x_{\mathrm{ML}}(y) - \hat{x}_\alpha(y)\|_2^2 - \sigma^2 \mathrm{tr}\left((AA^*)^+\right) + 2\sigma^2 \mathrm{gdf}_\alpha(y)$

Empirical distributions

- fine logarithmical α-grid: $\log_{10}(\alpha_i)$ from -40 to 40, step size 0.01.
- $N_\varepsilon = 10^6$ samples of ε.
- $m = n = 64$, $l = 0.06$, $\sigma = 0.1$

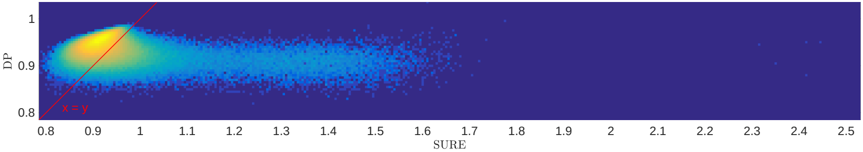


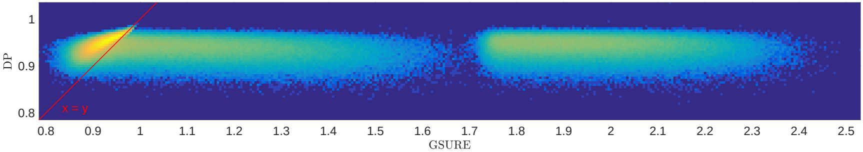
Empirical distributions (cont'd)



Joint empirical log-probabilities of $\log_{10} \|x^* - x_{\hat{\alpha}}\|_2$

Empirical distributions (cont'd)

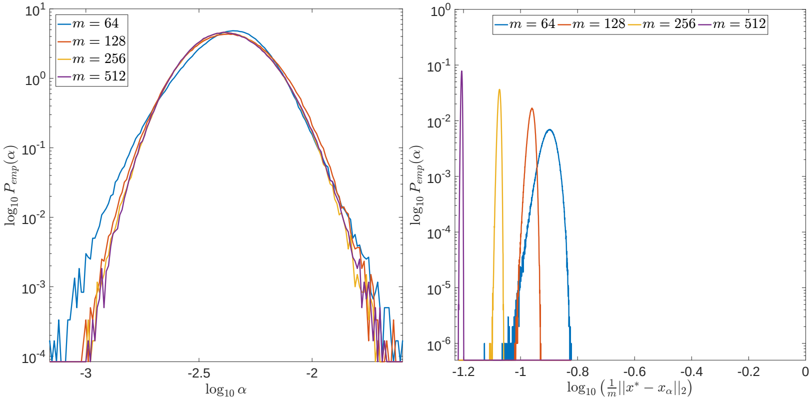


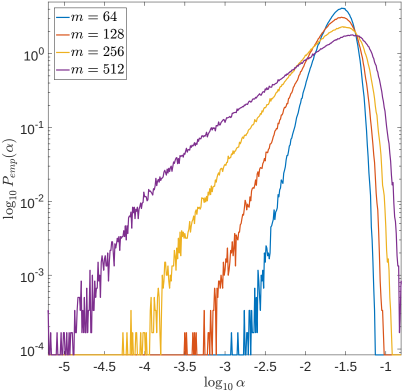


Joint empirical log-probabilities of $\log_{10} \|x^* - x_{\hat{\alpha}}\|_2$

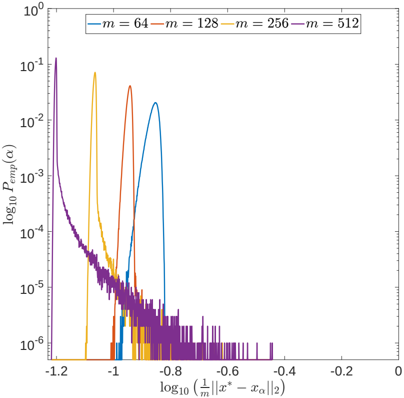
What's wrong? Let's do some more numerical studies...

Empirical probabilities for increasing m: Oracle



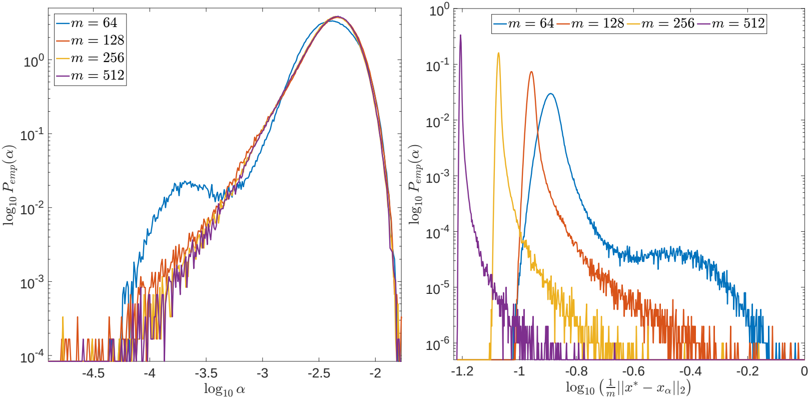


$m = 64$
$m = 128$
$m = 256$
$m = 512$
$\log_{10} P_{emp}(\alpha)$
$\log_{10} \alpha$

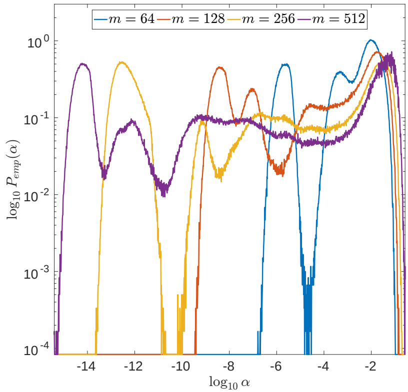


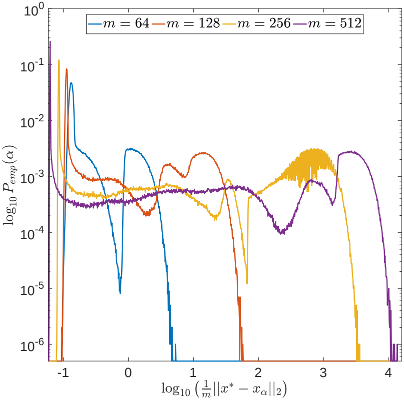
$m = 64$
$m = 128$
$m = 256$
$m = 512$
$\log_{10} P_{emp}(\alpha)$
$\log_{10} \left(\frac{1}{m}||x^* - x_\alpha||_2\right)$

Empirical probabilities for increasing m: SURE

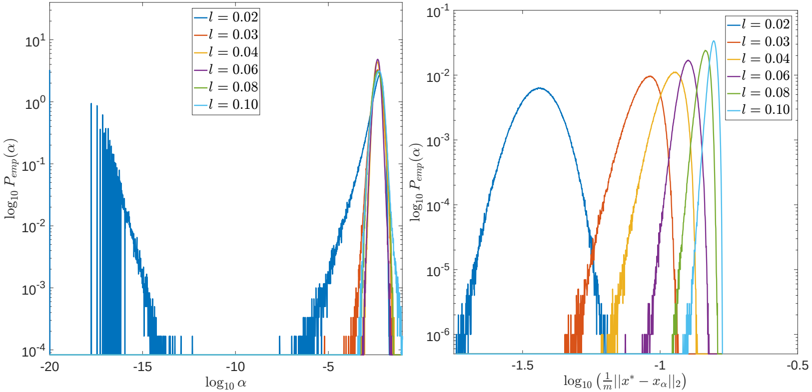


Empirical probabilities for increasing m: GSURE

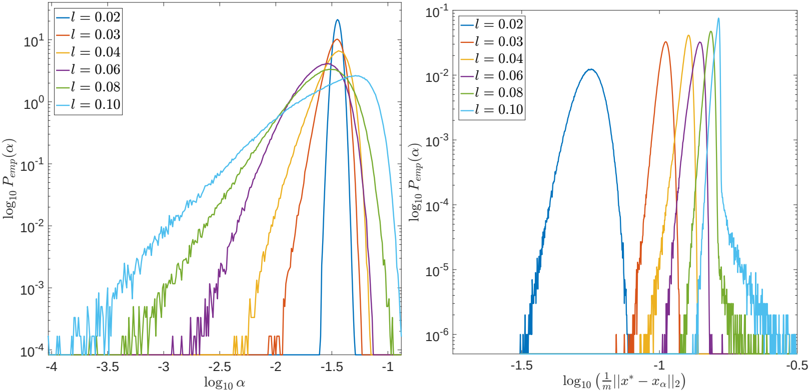




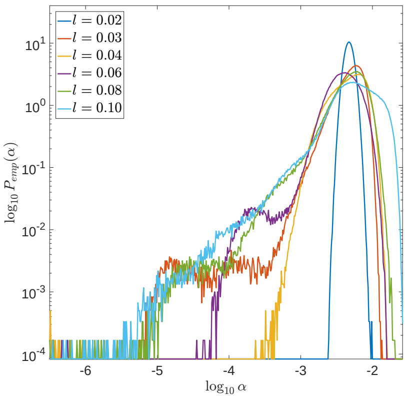
Empirical probabilities for increasing l: Oracle

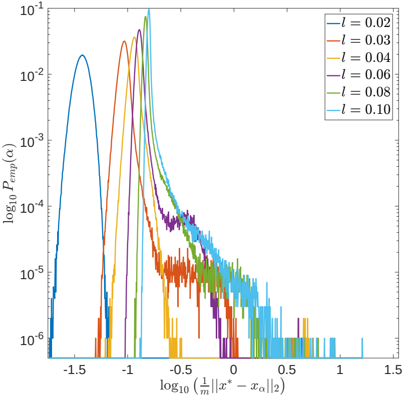


Empirical probabilities for increasing l: Discrepancy principle

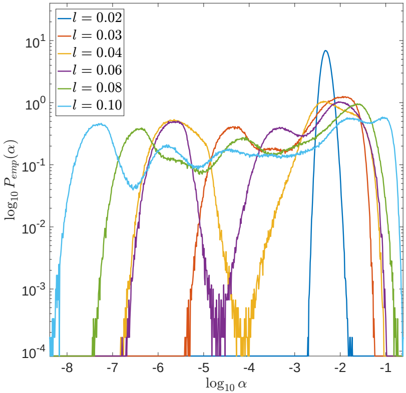


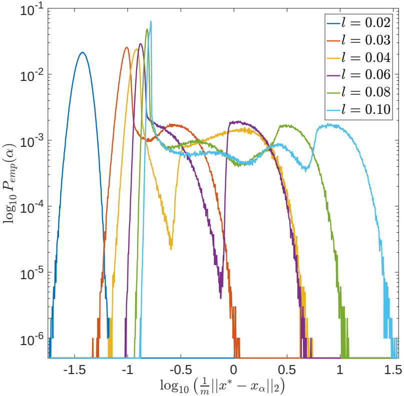
Empirical probabilities for increasing l: SURE





Empirical probabilities for increasing l: GSURE





Asymptotic analysis for quadratic regularization

UCL

Assume $A \in R^{m \times m}$, $1 = \gamma_1 \geq \ldots \geq \gamma_m > 0$ *and* $\|x^*\|_2^2 = O(m)$.
One can prove that for $m \longrightarrow \infty$*:*

$$\sup_{\alpha \in [0,\infty)} \left| \frac{1}{m} \left(\mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\alpha, y) \right) \right| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{m}} \right)$$

$$\sup_{\alpha \in [0,\infty)} \left| \frac{1}{m} \left(\mathrm{DP}(\alpha, y) - \mathbb{E}\left(\mathrm{DP}(\alpha, y) \right) \right) \right| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{m}} \right)$$

$$\sup_{\alpha \in [0,\infty)} \left| \frac{1}{m \, \mathrm{cond}(A_m)^2} \left(\mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha) \right) \right| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{m}} \right)$$

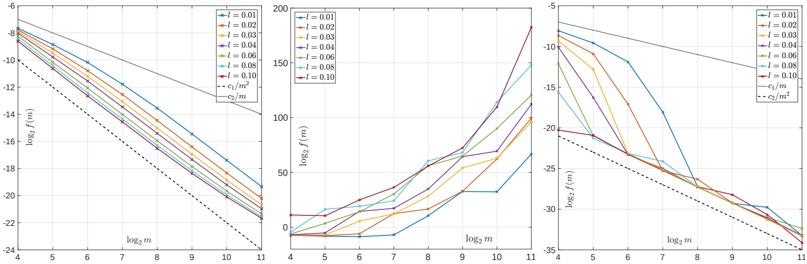
$$\mathbb{E}\left(\sup_{\alpha \in [0,\infty)} \left| \frac{1}{m} \left(\mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\alpha, y) \right) \right| \right)^2 = O\left(\frac{1}{m} \right)$$

$$\mathbb{E}\left(\sup_{\alpha \in [0,\infty)} \left| \frac{1}{m} \left(\mathrm{DP}(\alpha, y) - \mathbb{E}\left(\mathrm{DP}(\alpha, y) \right) \right) \right| \right)^2 = O\left(\frac{1}{m} \right)$$

$$\mathbb{E}\left(\sup_{\alpha \in [0,\infty)} \left| \frac{1}{m \, \mathrm{cond}(A_m)^2} \left(\mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha) \right) \right| \right)^2 = O\left(\frac{1}{m} \right)$$

Proof: Kolmogorov's maximal inequality & Doob's martingale inequality.

Numerical illustration of asymptotic theorems



(a) (b) (c)

(a) $\mathbb{E}\left(\sup_{\alpha\in[0,\infty)}\left|\frac{1}{m}\left(\mathrm{SURE}(\alpha,y)-\mathrm{R}_{\mathrm{SURE}}(\alpha,y)\right)\right|\right)^2$

(b) $\mathbb{E}\left(\sup_{\alpha\in[0,\infty)}\left|\frac{1}{m}\left(\mathrm{GSURE}(\alpha,y)-\mathrm{R}_{\mathrm{GSURE}}(\alpha)\right)\right|\right)^2$

(c) $\mathbb{E}\left(\sup_{\alpha\in[0,\infty)}\left|\frac{1}{m\ \mathrm{cond}(A_m)^2}\left(\mathrm{GSURE}(\alpha,y)-\mathrm{R}_{\mathrm{GSURE}}(\alpha)\right)\right|\right)^2$

Non-quadratic regularization

Sparsity-inducing regularization (LASSO):

$$\hat{x}_\alpha(y) = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|Ax - y\|_2^2 + \alpha \|x\|_1 \tag{1}$$

Let I be the support of $\hat{x}_\alpha(y)$, $|I| = k$, $P_I \in \mathbb{R}^{k \times n}$ projector onto I, A_I restriction of A to I. For our setting, we have that

$$\mathrm{df}_\alpha = \|\hat{x}_\alpha(y)\|_0 = k, \qquad \mathrm{gdf}_\alpha = \mathrm{tr}(\Pi P_I (A_I^* A_I)^{-1} P_I^*)$$

- **Deledalle, Vaiter, Peyré, Fadili, Dossal, 2012**. *Unbiased risk estimation for sparse analysis regularization*, IEEE ICIP.
- **Vaiter, Deledalle, Peyré, Fadili, Dossal, 2014**. *The Degrees of Freedom of Partly Smooth Regularizers*, arXiv:1404.5557.
- **Vaiter, Deledalle, Peyré, Dossal, Fadili, 2013**. *Local behavior of sparse analysis regularization: Applications to risk estimation*, Applied and Computational Harmonic Analysis 35(3).

Non-quadratic regularization

Sparsity-inducing regularization (LASSO):

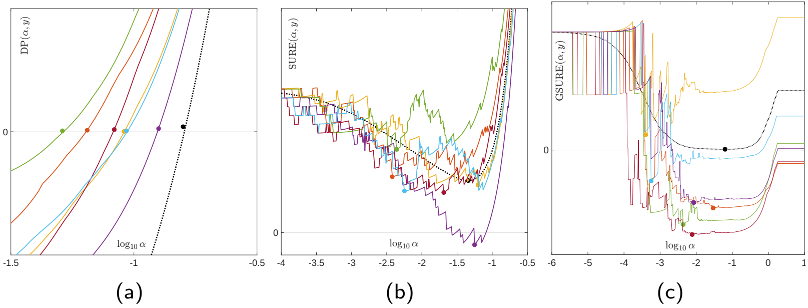
$$\hat{x}_\alpha(y) = \operatorname*{argmin}_{x \in \mathbb{R}^n} \ \frac{1}{2}\|Ax - y\|_2^2 + \alpha \|x\|_1 \tag{1}$$

Let I be the support of $\hat{x}_\alpha(y)$, $|I| = k$, $P_I \in \mathbb{R}^{k \times n}$ projector onto I, A_I restriction of A to I. For our setting, we have that

$$\mathrm{df}_\alpha = \|\hat{x}_\alpha(y)\|_0 = k, \qquad \mathrm{gdf}_\alpha = \mathrm{tr}(\Pi P_I (A_I^* A_I)^{-1} P_I^*)$$

- ! No theory, only numerical studies.
- ! Fast but accurate and consistent computation of $\hat{x}_\alpha(y)$ for α's ranging from 10^{-10} to 10^{10}.
- ✓ all-at-once implementation of ADMM solving (1) for all α simultaneously with $tol = 10^{-14}$ and 10^4 max iter.

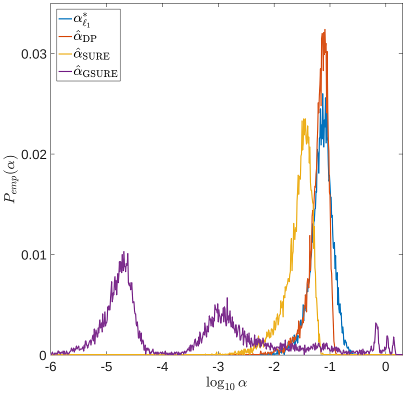
Example of risk functions for LASSO regularization

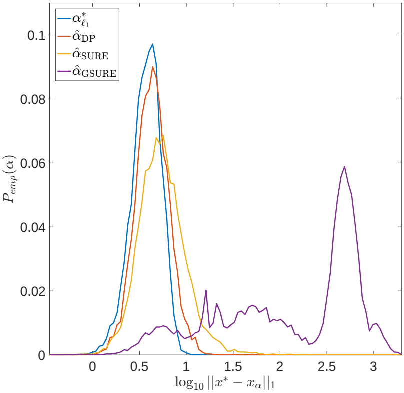


- (a) $R_{\mathrm{DP}}(\alpha) = \|A\hat{x}_\alpha(Ax^*) - Ax^*\|_2^2 - m\sigma^2$ vs. 6 realizations of $\mathrm{DP}(\alpha, y) = \|A\hat{x}_\alpha(y) - y\|_2^2 - m\sigma^2$
- (b) $R_{\mathrm{SURE}}(\alpha) = \mathbb{E}\left[\|Ax^* - A\hat{x}_\alpha(y)\|_2^2\right]$ vs. 6 realizations of $\mathrm{SURE}(\alpha, y) = \|y - A\hat{x}_\alpha(y)\|_2^2 - m\sigma^2 + 2\sigma^2 \mathrm{df}_\alpha(y)$.
- (c) $R_{\mathrm{GSURE}}(\alpha) = \mathbb{E}\left[\|\Pi(x^* - \hat{x}_\alpha(y))\|_2^2\right]$ vs. 6 realizations of $\mathrm{GSURE}(\alpha, y) = \|x_{\mathrm{ML}}(y) - \hat{x}_\alpha(y)\|_2^2 - \sigma^2 \mathrm{tr}\left((AA^*)^+\right) + 2\sigma^2 \mathrm{gdf}_\alpha(y)$

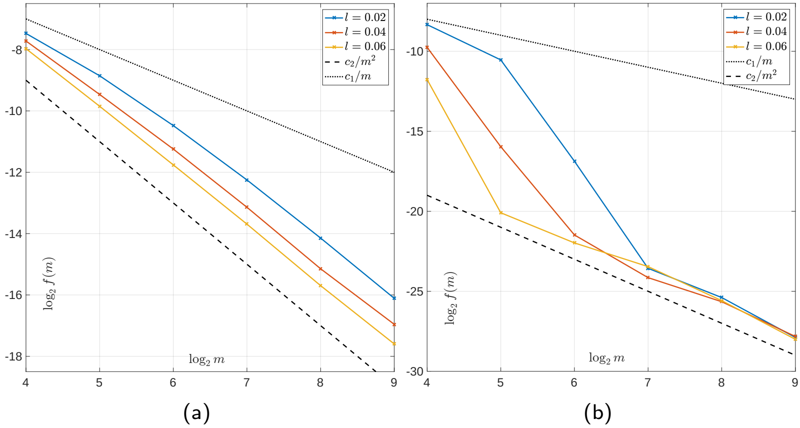
Empirical distributions for LASSO regularization

- fine logarithmical α-grid: $\log_{10}(\alpha_i)$ from -10 to 10, step size 0.01.
- $N_\varepsilon = 10^4$ samples of ε.
- $m = n = 64$, $l = 0.06$, $\sigma = 0.1$





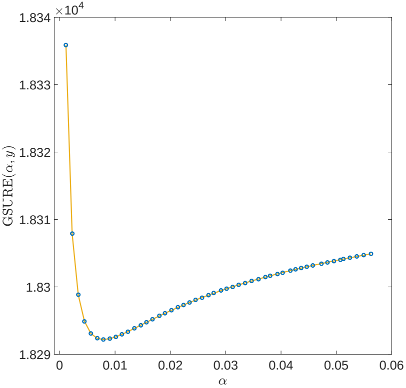
UCL



(a) $\mathbb{E}\left(\sup_{\alpha\in[0,\infty)}\left|\frac{1}{m}\left(\text{SURE}(\alpha,y)-\text{R}_{\text{SURE}}(\alpha,y)\right)\right|\right)^2$

(b) $\mathbb{E}\left(\sup_{\alpha\in[0,\infty)}\left|\frac{1}{m\ \text{cond}(A_m)^2}\left(\text{GSURE}(\alpha,y)-\text{R}_{\text{GSURE}}(\alpha)\right)\right|\right)^2$

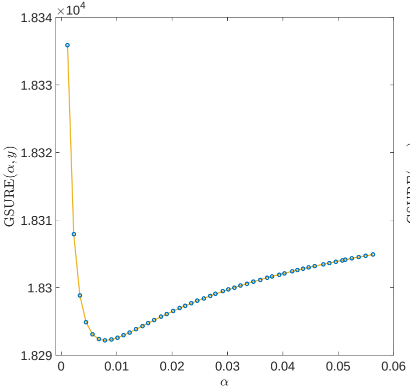
GSURE computed on a linear grid around "a reasonable value"...

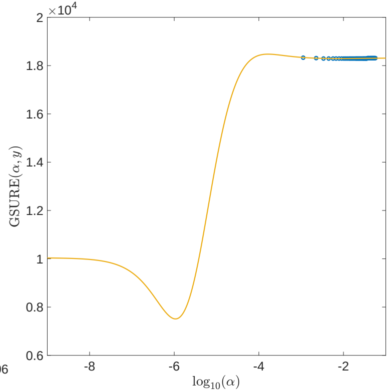


(quadratic regularization)

Why did no one notice this before?

GSURE computed on a linear grid around "a reasonable value"...



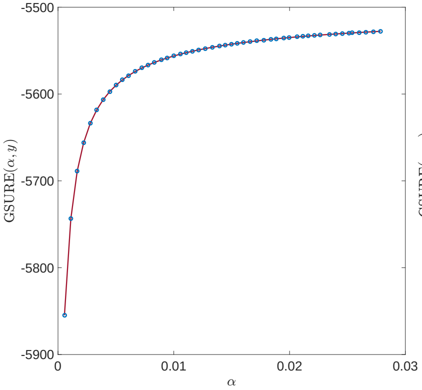


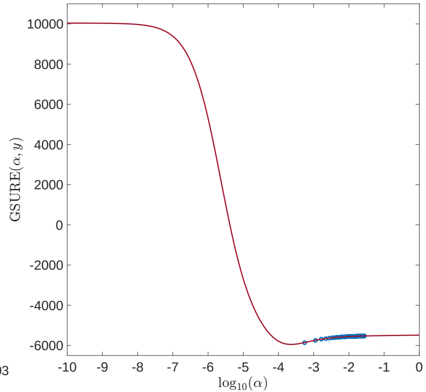
...and on a fine logarithmic grid.

(quadratic regularization)

Why did no one notice this before? (cont'd)

GSURE computed on a linear grid around "a reasonable value"...

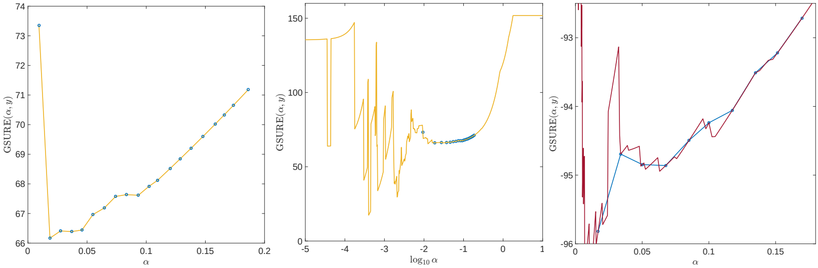




...and on a fine logarithmic grid.

(quadratic regularization)

Why did no one notice this before? (cont'd)

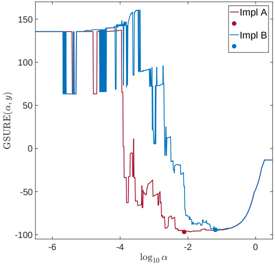


(LASSO regularization)

Why did no one notice this before? (cont'd)

In addition to fine logarithmic grids, you need an accurate solution.

- Solving large-scale problems with iterative solvers adds regularization.
- Often, scan over α is done with low accuracy only.



Many other works considered very mildly ill-posed problems (e.g., denoising) only, and only considered single noise realizations.

Conclusions & Outlook

- Unbiased risk estimators can be problematic for ill-posed problems.
- Asymptotic analysis suggests that GSURE is far off the real, reasonable risk function.
- In fact, risk estimation is an asymptotically ill-posed problem itself.
- Discrepancy principle was analysed in the same framework, and although often more conservative than SURE/GSURE, often more reliable.

- New risk estimators not based on Stein's method? Maybe not unbiased, i.e., regularized?
- LASSO: Asymptotic theory? Different GSURE risk more suitable (Bregman distances)?
- Non-Gaussian noise models?

L, Proksch, Brune, Bissantz, Burger, Dette & Wübbeling, 2017. *Risk Estimators for Choosing Regularization Parameters in Ill-Posed Problems - Properties and Limitations*, *submitted, arXiv:1701.04970.*