Empirical Risk Minimization as Parameter Choice Rule for General Linear Regularization Methods

Frank Werner[1]

Statistical Inverse Problems in Biophysics Group
Max Planck Institute for Biophysical Chemistry, Göttingen

and

Felix Bernstein Institute for Mathematical Statistics in the Biosciences
University of Göttingen

European Meeting of Statisticians 2017, Helsinki

[1]joint work with Housen Li

Ill-posed linear models

Model: Recover unknown f from n indirect noisy samples

$$Y = Tf + \sigma\xi \qquad \text{with } T \in \mathbb{R}^{n\times p}, \operatorname{rank}(T) = p, \xi \text{ standard Gaussian.}$$

Eigenvalues of T^*T: $\lambda_1 \geq \cdots \geq \lambda_p > 0$, assume

$$\lambda_k \asymp k^{-a} \qquad \text{with some } a > 1.$$

Normalized eigenvectors $e_1, ..., e_p$ $\rightsquigarrow$ Equivalent sequence model:

$$Y_k = \sqrt{\lambda_k} f_k + \sigma\xi_k, \qquad k = 1, \ldots, p,$$

where $Y_k := \langle \lambda_k^{-1/2} Te_k, Y\rangle$, $f_k = \langle f, e_k\rangle$, $\xi_k := \langle \lambda_k^{-1/2} Te_k, \xi\rangle \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

Linear regularization methods

Recall: least square estimator $\hat{f} := (T^*T)^{-1}T^*Y$.

Ill-posedness $\rightsquigarrow$ stable approximation $q_\alpha(\cdot)$ of $(\cdot)^{-1}$, that is,

$$\text{linear regularization methods:} \qquad \hat{f}_\alpha := q_\alpha(T^*T)T^*Y.$$

Definition

We call $q_\alpha : [0, \lambda_1] \to \mathbb{R}$ with $\alpha \in \mathcal{A} \subseteq \mathbb{R}_+$ an ordered filter if

(i) There exist $C_q', C_q'' > 0$ s.t. for every $\alpha \in \mathcal{A}$ and every $\lambda \in [0, \lambda_1]$

$$\alpha |q_\alpha(\lambda)| \leq C_q' \qquad \text{and} \qquad \lambda |q_\alpha(\lambda)| \leq C_q''.$$

(ii) $\alpha \mapsto (q_\alpha(\lambda_k))_{k=1}^p$ is strictly monotone and continuous.

Smoothness assumptions

We want to obtain minimax optimality over ellipsoids of the form

$$\mathcal{W} := \left\{ f \in \mathbb{R}^p : \sum_{k=1}^{p} w_k f_k^2 \leq 1 \right\} \qquad \text{with } w_k \asymp k^b.$$

But therefore, q_α must be able to take advantage of this!

Shorthand notation: $s_\alpha(\lambda) := \lambda q_\alpha(\lambda)$. Qualification condition

$$\sup_{\alpha \in \mathcal{A},\, \lambda \in [0, \lambda_1]} \alpha^{-v} \lambda^v |1 - s_\alpha(\lambda)| \leq C_v < \infty \qquad \text{for all } 0 < v \leq v_0.$$

The largest possible v_0 is called the polynomial qualification index.

Examples

Table: Summary of some ordered filters

Method	$q_\alpha(\lambda)$	C_q'	C_q''	v_0	Need SVD
Spectral cut-off	$\frac{1}{\lambda}\mathbf{1}_{[\alpha,\infty)}(\lambda)$	1	1	∞	Yes
Tikhonov	$\frac{1}{\lambda+\alpha}$	1	1	1	No
m-iterated Tikhonov	$\frac{(\lambda+\alpha)^m-\alpha^m}{\lambda(\lambda+\alpha)^m}$	m	1	m	No
Landweber ($\\|T\\| \leq 1$)	$\sum_{j=0}^{\lfloor\alpha\rfloor-1}(1-\lambda)^j$	1	1	∞	No
Showalter	$\frac{1-\exp\left(-\frac{\lambda}{\alpha}\right)}{\lambda}$	1	1	∞	No

A-priori parameter choice

Proposition (Bissantz et al. '07)

Let $\hat{f}_\alpha := q_\alpha(T^*T)T^*Y$ *with a filter* q_α, *and* $\alpha = \alpha_{\mathrm{or}} \asymp (\sigma^2)^{a/(a+b+1)}$.

- *If the qualification index* $v_0 \geq b/(2a)$, *then*

$$R(\alpha_{\mathrm{or}}, \mathcal{W}) := \sup_{f \in \mathcal{W}} \mathbb{E}\left[\|\hat{f}_{\alpha_{\mathrm{or}}} - f\|^2\right] \lesssim (\sigma^2)^{\frac{b}{a+b+1}}.$$

- *If further* $v_0 \geq b/(2a) + 1/2$, *then*

$$r(\alpha_{\mathrm{or}}, \mathcal{W}) := \sup_{f \in \mathcal{W}} \mathbb{E}\left[\|T\hat{f}_{\alpha_{\mathrm{or}}} - Tf\|^2\right] \lesssim (\sigma^2)^{\frac{a+b}{a+b+1}}.$$

Such rates are minimax optimal in order over $\mathcal{W}$.

Empirical prediction risk minimization

The optimality on the last slide relies on the smoothness of f (via α_{or}).

We consider the parameter choice rule $\hat{\alpha}$ given by

$$\hat{\alpha} := \underset{\alpha \in \mathcal{A}}{\operatorname{argmin}} \left[\| T \hat{f}_\alpha - Y \|^2 + 2\sigma^2 \mathrm{Trace}\left(s_\alpha \left(T^* T \right) \right) \right].$$

Intuition: minimize an unbiased estimator of the prediction risk

$$r(\alpha, f) := \mathbb{E}\left[\| T(\hat{f}_\alpha - f) \|^2 \right] = \sum_{k=1}^{p} \lambda_k (1 - s_\alpha(\lambda_k))^2 f_k^2 + \sigma^2 \sum_{k=1}^{p} s_\alpha(\lambda_k)^2,$$

since

$$\mathbb{E}\left[\| T \hat{f}_\alpha - Y \|^2 \right] = \underbrace{\sum_{k=1}^{p} \lambda_k (1 - s_\alpha(\lambda_k))^2 f_k^2 + \sigma^2 \sum_{k=1}^{p} s_\alpha(\lambda_k)^2}_{r(\alpha, f)} - \underbrace{2\sigma^2 \sum_{k=1}^{p} s_\alpha(\lambda_k)}_{2\sigma^2 \mathrm{Trace}(s_\alpha(T^* T))} + p\sigma^2.$$

Empirical prediction risk minimization (cont')

The $\hat{\alpha}$ was first introduced in (Mallows '73), thus a.k.a. Mallows C_L.

Practice: it is popular & attractive.

Theory: $\hat{\alpha}$ is order optimal w.r.t. prediction risk $r(\alpha, f)$ (Kneip '94).

- Unknown: Is $\hat{\alpha}$ also optimal for the risk $R(\alpha, f) := \mathbb{E}\left[\|\hat{f}_\alpha - f\|^2\right]$?
  - It is way more informative than $r(\alpha, f)$ due to the ill-posedness.
  - Spectral cut-off: this has recently been shown in (Chernousova & Golubev '14.)
- Our goal: Extend it to general linear regularization methods.
  - Why? Spectral cut-off relies on full SVD, thus impractical.

Order inequality

Assumption

(i) *As* $\alpha \searrow 0$, $s_\alpha(\alpha) \equiv \alpha q_\alpha(\alpha) \geq c_q > 0$.

(ii) *For* $\alpha \in \mathcal{A}$, *the function* $\lambda \mapsto s_\alpha(\lambda)$ *is non-decreasing.*

All mentioned regularization methods satisfy the assumption.

It requires proper parametrization. E.g. Tikhonov with re-parametrization $\alpha \mapsto \sqrt{\alpha}$, i.e. $q_\alpha(\lambda) = 1/(\sqrt{\alpha} + \lambda)$, still an ordered filter, but violates Ass. (i).

Theorem (Oracle inequality)

Let $r(\alpha_{\mathrm{or}}, f) := \min_{\alpha \in \mathcal{A}} r(\alpha, f)$. *Then for all* $f \in \mathcal{W}$

$$\mathbb{E}\left[\|\hat{f}_{\hat{\alpha}} - f\|^2\right] \lesssim r(\alpha_{\mathrm{or}}, f)^{\frac{b}{a+b}} + \sigma^{-2a} r(\alpha_{\mathrm{or}}, f)^{1+a} + \sigma^{1-2a} r(\alpha_{\mathrm{or}}, f)^{\frac{1+2a}{2}}.$$

Order optimality

$$\mathbb{E}\left[\|\hat{f}_{\hat{\alpha}} - f\|^2\right] \lesssim r(\alpha_{\mathrm{or}}, f)^{\frac{b}{a+b}} + \sigma^{-2a} r(\alpha_{\mathrm{or}}, f)^{1+a} + \sigma^{1-2a} r(\alpha_{\mathrm{or}}, f)^{\frac{1+2a}{2}}.$$

Recall:

$$r(\alpha_{\mathrm{or}}, f) \lesssim \sigma^{\frac{2(a+b)}{a+b+1}} \quad \text{if } v_0 \geq b/(2a) + 1/2$$

Thus, if $v_0 \geq b/(2a) + 1/2$,

$$\mathbb{E}\left[\|\hat{f}_{\hat{\alpha}} - f\|^2\right] \lesssim \sigma^{\frac{2b}{a+b+1}}. \quad \text{(order optimal)}$$

$v_0 \geq b/(2a) + 1/2$ means we need higher qualification (early saturation)

- Same price for the deterministic discrepancy principle and GCV, which also rely on the residual $\|T\hat{f}_\alpha - Y\|$.
- Better than Lepskiĭ ('90) principle, where one typically looses a log-factor.

Further results

Oracle inequality & optimality actually holds...

... in a more general setting $Y = Tf + \sigma\xi$ where

- T is an injective and compact operator between Hilbert spaces.,
- the Eigenvalues of T^*T decay in a general way,
- ξ is sub-Gaussian noise, and σ is unknown.

... under general smoothness assumptions:

- Source condition

$$f = \phi(T^*T)w \qquad \text{for some } \omega \text{ with } \|w\| \leq C.$$

- Qualification condition

$$\sup_{\lambda \in [0, \lambda_1]} \sqrt{\lambda}\phi(\lambda)|1 - s_\alpha(\lambda)| \lesssim \sqrt{\alpha}\phi(\alpha).$$

Experiment setting

Forward operator $T : \mathbf{L}^2([0,1]) \to \mathbf{L}^2([0,1])$

$$(Tf)(x) = \int_0^1 k(x,y) f(y) \,\mathrm{d}y, \qquad \text{with } k(x,y) = \min\{x(1-y), y(1-x)\}.$$

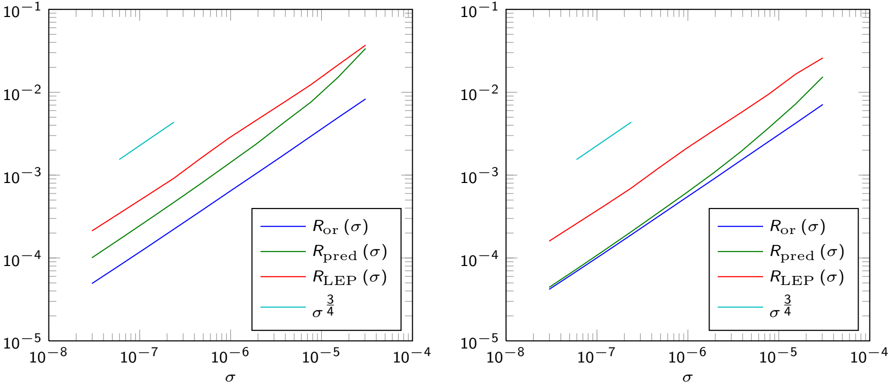
Obviously, $(Tf)'' = -f$, so the eigenvalues λ_k of T^*T satisfy $\lambda_k \asymp k^{-4}$

The unknown truth

$$f(x) = \begin{cases} x & \text{if } 0 \le x \le \frac{1}{2}, \\ 1-x & \text{if } \frac{1}{2} \le x \le 1. \end{cases}$$

Then $f_k = \frac{(-1)^k - 1}{4\pi^3 k^2}$ and the optimal rate is $\mathcal{O}\left(\sigma^{\frac{3}{4}-\varepsilon}\right)$ for any $\varepsilon > 0$.

Results



(a) Tikhonov regularization

(b) Showalter regularization

Figure: Average of $\|\hat{f} - f\|_2^2$ over 10^4 repetitions.

Conclusion

Theoretical explanations for the well-known parameter choice rule via empirical prediction risk minimization

Open questions

- Nonlinear problems;
- Different noise models;
- Exponentially ill-posed problems.

H. Li and F. Werner (2017).
Empirical risk minimization as parameter choice rule for general linear regularization methods.
arXiv: 1703.07809.