Model selection for ill-posed inverse problems

Jean-Michel Loubes
(joint work with Ana Karina Fermin, Carenne Ludeña)

We are interested in recovering an unobservable signal x_0 based on observations

$$y(t_i) = F(x_0)(t_i) + \varepsilon_i, \tag{1}$$

where $F : X \to Y$ is a regular functional, with X, Y Hilbert spaces and $t_i, i = 1, \ldots, n$ is a fixed observation scheme. $x_0 : \mathbb{R} \to \mathbb{R}$ is the unknown function to be recovered from the data $y(t_i)$, $i = 1, \ldots, n$. The regularity condition over the unknown parameter of interest is expressed through the assumption $x_0 \in X$. We assume that the observations $y(t_i) \in \mathbb{R}$ and that the observation noise ε_i are i.i.d. realizations of a certain random variable ε. Throughout the paper, we shall denote $\mathbf{y} = (y(t_i))_{i=1}^n$. We assume F is Fréchet differentiable and ill posed in the sense that our noise corrupted observations might lead to large deviations when trying to estimate x_0. In a deterministic framework, the statistical model is formulated as the problem of approximating the solution of $F(x) = y$, when y is not known, and is only available through an approximation y^δ,

$$\|y - y^\delta\| \leq \delta.$$

It is important to remark that whereas in this case consistency of the estimators depends on the approximation parameter δ, which depends on the number of observations n.

In the linear case, the best L^2 approximation of x_0 is $x^+ = F^+ y$, where F^+ is the Moore-Penrose (generalized) inverse of F. We will say the problem is ill-posed if F^+ is unbounded. This might entail, and is generally the case, that $F^+(y^\delta)$ is not close to x^+. Hence, the inverse operator needs to be, in some sense, regularized. In the nonlinear case, by ill-posedness we will always mean that the solutions do not depend continuously on the data. If F is a compact operator, local injectivity around x^+ a solution of $F(x^+) = y$, implies ill-posedness of the problem.

Regularization methods replace an ill-posed problem by a family of well-posed problems. Their solution, called regularized solutions, are used as approximations of the desired solution of the inverse problem. These methods always involve some parameter measuring the closeness of the regularized and the original (unregularized) inverse problem. Rules (and algorithms) for the choice of these regularization parameters as well as convergence properties of the regularized solutions are central points in the theory of these methods, since they allow to find the right balance between stability and accuracy.

Our goal in this article is to develop algorithms providing estimators that achieve optimal rates of convergence when the smoothness of the true solution is not known a priori. We will assume operator F can be linear or non linear, but in the latter case we will assume it satisfies a local linear invariance condition.

For this we consider penalized M-estimators minimizing quantities of the form

$$\hat{x}_n = \arg\min_{\alpha_n \in \Theta} \arg\min_{x \in \mathcal{X}} \left(\gamma(y - F(x)(t)) + \alpha_n \mathrm{pen}(x, \mathcal{X}) + pen(\alpha_n)\right), \tag{2}$$

where $\mathcal{X}$ is a specific set, $\gamma(.)$ is a loss-function, $\mathrm{pen}(.,.)$ is a penalty over x and/or $\mathcal{X}$, and $\alpha_n \in \Theta$ is a decreasing sequence all of which will be defined precisely later. The idea of penalized M-estimators is to find an estimator close enough to the data, close in the sense defined by γ and with a regularity property induced by the choice of the penalty pen. Adaptivity means here, that the construction of the estimator does not require knowing beforehand the regularity of the function of interest to be recovered x_0. In the inverse problems literature this is known as a posteriori methods. But, we do assume that the inverse operator is known as well as some assumptions, such as its degree of ill-posedness.

When F is linear, the statistical problem has been extensively studied, although in general efficient adaptive regularization-parameter choice is still under active research. Two main kinds of estimators have been considered. First regularized estimators such as Tikhonov type estimators, then non linear thresholded estimators. We refer to [5], [6], [2] or [1] for more references.

1. Definitions and notations

We introduce certain standard assumptions on the observation noise

AN moment condition for the errors: ε is a centered random variable satisfying the moment condition $\mathbf{E}(|\varepsilon|^p/\sigma^p) \leq p!/2$ and $\mathbf{E}(\varepsilon^2) = \sigma^2$.

The smoothness of the function is expressed through the following assumption.

SC source condition: There exists $0 < \nu \leq 1/2$ such that

$$x_0 \in Range((T^*T)^\nu) = \mathcal{R}((T^*T)^\nu).$$

We assume the regularity of the problem is defined by that of $F'(x_0)$. In the linear case we will write $F = T$. This linear operator acts with a degree of ill-posedness defined by an index p. This is generally expressed by the fact that T maps L^2 into some Sobolev space H_p, or by assuming that T acts along a Hilbert scale H_s.

IP ill posedness of the operator: There exists $p > 0$ such that $F'(x_0)(H_s) = H_{s+p}$.

Now consider approximation subsets, namely $\forall m \in \mathcal{M}_n$, $Y_m \subset Y$ with $d_m = \dim(Y_m)$. We start out with a big enough subspace Y_{m_0} and in order to deal with the ill posedness of $(T^*\Pi^n_{Y_m})^+$ use Tikhonov-like regularization methods. Consider the corresponding sets in space X, defined by $X_m = (\Pi^n_{Y_m}T)^+X$. To begin assume that m_0 is such that

$$\|(I - \Pi_{X_{m_0}})x_0\| \leq \inf_m [\|(I - \Pi_{X_m})x_0\| + \sqrt{\frac{d_m}{n}}\frac{1}{\gamma_m}].$$

This quantity can be chosen so as not to depend on the unknown regularity of the solution x_0. Under assumption **SC** the above inequality is satisfied if the

dimension of the set is such that

$$d_{m_0}^{2\nu p} \geq n^{\frac{2\nu p}{4\nu p+2p+1}}.$$

Thus it is enough to choose m_0 such that $d_{m_0} \geq n^{1/(2p+1)}$.

2. Estimation procedure

Next consider for a given $k \in \mathcal{K}$ a sequence of (diagonal) matrixes $A_k(n) \in \mathbb{M}_{m_0 \times m_0}(\mathbb{R})$ as and define the following penalized estimator:

$$\hat{x}_{A_k(n)} = \arg\min_{x \in X_m} \left[\|\Pi_{Y_{m_0}}(y - F(x))\|_n^2 + \|A_k(n)x\|^2 \right]. \tag{3}$$

The behaviour of this general estimator depends on the choice of the smoothing matrixes $A_k(n)$. For particular choices of the penalty, this estimator can be seen as a Tikhonov regularized estimator or a projection estimator. We would like to choose $A_k(n)$, among all the $A_k(n)$, $k \in \mathcal{K}$ based on the data in such a way that optimal rates are maintained. This choice must also not depend on a priori regularity assumptions.

First we note that, for each fixed $A_k(n)$, the expression (3) can be written in the following way:

$$\hat{x}_{A_k(n)} = \arg\min_{x \in X_{m_0}} \|(T^t \Pi_{Y_{m_0}} T + A_k(n)^t A_k(n))^{-1} T^t \Pi_{Y_{m_0}}(y - T(x))\|^2. \tag{4}$$

In practice the second one is more complicated (the matrix to inverse might be big), but it is simpler to deal with in order to show our results concerning the selection of A_k. With this notation set

$$R_{A_k(n)} = (T^t \Pi_{Y_{m_0}} T + A_k(n)^t A_k(n) I_{m_0})^{-1} T^t \Pi_{Y_{m_0}}.$$

Now set $\gamma(x, \alpha_k) = \|R_{A_k}(y - F(x))\|^2$ and

$$\text{pen}(A_k) = r\sigma^2(1 + L_k)[Tr(R_{A_k}^t R_{A_k}) + \rho^2(R_{A_k})],$$

with $r > 2$. We choose $\hat{x}_{A_{\hat{k}}}$ such that

$$\hat{x}_{A_{\hat{k}}} = \arg\min_{k \in \mathcal{K}, x \in X_{m_0}} \left(\gamma(x, A_k(n)) + \text{pen}(A_k(n))\right).$$

Let $x_{A_k} = R_{A_k} T x_0$. Set

$$\Sigma(d) = \sum_k 2 \left[\sqrt{\frac{dTr(R_{A_k}^t R_{A_k})}{\rho^2(R_{A_k})}} + 1 \right] \frac{d}{\rho^2(nR_{A_k})} e^{-\sqrt{dL_k[Tr(R_{A_k}^t R_{A_k}) + \rho^2(R_{A_k})]/\rho^2(R_{A_k})}},$$

for d. We have the following result.

Theorem. There exists a constant C which depends on r and on T, such that the following inequality holds true

$$\mathbf{E}\|\hat{x}_{A_{\hat{k}}} - x_0\|^2 \leq 2\|(I - \Pi_{X_{m_0}})x_0\|^2 + C \inf_{k \in \mathcal{K}} \left[\|x_{A_k} - x_0\|^2 + 2\text{pen}(A_k) \right] + \frac{\Sigma(d)}{n}. \tag{5}$$

Hence, the estimator is optimal in the sense that the adaptive estimator achieves the best rate of convergence among all the regularized estimators.

Proof. In the linear case $F = T$, for any x_{A_k} and any $k \in \mathcal{N}$

$$\|R_{A_{\hat{k}}}(y - T\hat{x}_{A_{\hat{k}}})\|^2 + pen(A_{\hat{k}}) \leq \|R_{A_k}(y - Tx_{A_k})\|^2 + pen(A_k)$$

and

$$\|R_{A_k}(y - Tx_{A_k})\|^2 = \|R_{A_k}T(x_0 - x_{A_k})\|^2 + 2\langle R_{A_k}T(x_0 - x_{A_k}), R_{A_k}\varepsilon\rangle + \|R_{A_k}\varepsilon\|^2$$

Thus, following standard arguments we have

$$\begin{aligned}
&\|R_{A_{\hat{k}}}T(x_0 - \hat{x}_{A_{\hat{k}}})\|^2 \\
\leq &\|R_{A_k}T(x_0 - x_{A_k})\|^2 - 2 < R_{A_{\hat{k}}}T(x_0 - \hat{x}_{A_{\hat{k}}}), R_{A_{\hat{k}}}\varepsilon > \\
&+ 2 < R_{A_k}T(x_0 - x_{A_k}), R_{A_k}\varepsilon > - \|R_{A_{\hat{k}}}\varepsilon\|^2 + \|R_{A_k}\varepsilon\|^2 + \mathrm{pen}(A_k) - \mathrm{pen}(A_{\hat{k}}).
\end{aligned}$$

Let $0 < \kappa < 1$. Since $2ab \leq \kappa a^2 + \frac{1}{\kappa}b^2$, for any a, b we have for any k and $x_{A_k} \in X_m$

$$\begin{aligned}
&(1-\kappa)\|R_{A_{\hat{k}}}T(x_0 - \hat{x}_{A_{\hat{k}}})\|^2 \\
&\leq (1+\kappa)\|R_{A_k}T(x_0 - x_{A_k})\|^2 + (2 + \frac{1}{\kappa})\mathrm{pen}(A_k) + 2\sup_k\{(1 + \frac{1}{\kappa})\|R_{A_k}\varepsilon\|^2 - \mathrm{pen}(A_k)\},
\end{aligned}$$

On the other hand, using that is $1 \leq R_{A_k}T \leq C$, we have that for any $x_{A_k} \in X_{m_0}$ and any $k \in \mathbb{N}$,

$$\begin{aligned}
&(1-\kappa)\|x_0 - \hat{x}_{A_{\hat{k}}}\|^2 \leq C(1-\kappa)\|x_0 - x_{A_k}\|^2 \\
&+ (2 + \frac{1}{\kappa})\,\mathrm{pen}(A_k) + 2\,C_1 \sup_k \{\|R_{A_k}\varepsilon\|^2 - (1 + \frac{1}{\kappa})^{-1}\mathrm{pen}(A_k)\}.
\end{aligned}$$

As above, the proof then follows directly from the concentration of the supremum of the empirical process under the linear application as defined by the regularization family, see [4]. The choice of κ will depend on r in the penalty. □

Particular choises of penalty enable to handle nonlinear operatores, see [4] and iterative methods, see [3].

References

[1] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov, *Oracle inequalities for inverse problems*, Ann. Statist., **30(3)** (2002),843–874.

[2] Laurent Cavalier and Alexandre Tsybakov, *Sharp adaptation for inverse problems with random noise*, Probab. Theory Related Fields, **123(3)** (2002),323–354.

[3] A-K. Fermin, C. Ludeña, *Iterative methods for inverse problems*, working paper (2005).

[4] J-M. Loubes, C. Ludenña, *Complexity regularization for general inverse problems*, Prépublications de l'Université Paris Sud, 2005.

[5] Finbarr O'Sullivan, *A statistical perspective on ill-posed inverse problems*, Statist. Sci., **1(4)**, (1986), 502–527.

[6] Bernard A. Mair and Frits H. Ruymgaart, *Statistical inverse estimation in Hilbert scales*, SIAM J. Appl. Math., **56(5)** (1996),1424–1444.