Towards optimal regularization by incorporating prior knowledge

Toon van Waterschoot and Marc Moonen
Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
toon.vanwaterschoot@esat.kuleuven.ac.be

1 Regularization in least squares parameter estimation

Consider the linear estimation problem

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} u(1) & u(0) & \dots & u(1-n_F) \\ u(2) & u(1) & \dots & u(2-n_F) \\ \vdots & \vdots & \ddots & \vdots \\ u(N) & u(N-1) & \dots & u(N-n_F) \end{bmatrix} \cdot \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n_F} \end{bmatrix} + \begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(N) \end{bmatrix},$$

or in matrix notation

$$\mathbf{y} = \mathbf{U}\mathbf{f} + \mathbf{v}. \tag{1}$$

An estimator for $\mathbf{f}$ may be obtained by minimizing the least squares (LS) criterion

$$\min_{\hat{\mathbf{f}}} V_{LS}(\hat{\mathbf{f}}) = \min_{\hat{\mathbf{f}}} (\mathbf{y} - \mathbf{U}\hat{\mathbf{f}})^T (\mathbf{y} - \mathbf{U}\hat{\mathbf{f}}) \tag{2}$$

which results in the well-known LS estimator

$$\boxed{\hat{\mathbf{f}}_{\mathbf{LS}} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{y}.} \tag{3}$$

When the input signal $u(t)$ is not (or hardly) persistently exciting, the matrix $\mathbf{U}^T\mathbf{U}$ may be ill-conditioned or even singular. A common solution is to apply Tikhonov regularization by adding a scaled identity matrix to $\mathbf{U}^T\mathbf{U}$:

$$\boxed{\hat{\mathbf{f}}_{\mathbf{RgLS}} = (\mathbf{U}^T\mathbf{U} + \alpha\mathbf{I})^{-1}\mathbf{U}^T\mathbf{y}.} \tag{4}$$

This is in fact the solution to a ridge regression problem as formulated by the criterion

$$\min_{\hat{\mathbf{f}}} V_{RR}(\hat{\mathbf{f}}) = \min_{\hat{\mathbf{f}}} \big((\mathbf{y} - \mathbf{U}\hat{\mathbf{f}})^T (\mathbf{y} - \mathbf{U}\hat{\mathbf{f}}) + \alpha \|\hat{\mathbf{f}}\|_2^2 \big). \tag{5}$$

We address the question whether it is convenient to replace the scaled unity matrix $\alpha\mathbf{I}$ by a non-identity, possibly non-diagonal regularization matrix $\mathbf{P}$:

$$\boxed{\hat{\mathbf{f}}_{\mathbf{RgLS}} = (\mathbf{U}^T\mathbf{U} + \mathbf{P})^{-1}\mathbf{U}^T\mathbf{y},} \tag{6}$$

and if it is, what is the optimal choice for $\mathbf{P}$?

2 Linear mimimum mean square error estimation

Let us derive an expression for the minimum mean square error (MMSE) estimator of $\mathbf{f}$:

$$\min_{\hat{\mathbf{f}}} V_{MMSE}(\hat{\mathbf{f}}) = \min_{\hat{\mathbf{f}}} E(\hat{\mathbf{f}} - \mathbf{f})^T (\hat{\mathbf{f}} - \mathbf{f}) \tag{7}$$

under the following assumptions:

- The estimator is a linear function of the data in $\mathbf{y}$:

$$\hat{\mathbf{f}}_{\mathbf{MMSE}} = \mathbf{Z}^T\mathbf{y}. \tag{8}$$

- The measurement noise in $\mathbf{v}$ is drawn from a stationary white noise process with zero mean and variance σ_v^2:

$$\boldsymbol{\mu}_{\mathbf{v}} \triangleq E\mathbf{v} = \mathbf{0}, \tag{9}$$
$$\mathbf{R}_{\mathbf{v}} \triangleq \mathrm{cov}(\mathbf{v}) = E\mathbf{v}\mathbf{v}^T = \sigma_v^2\mathbf{I}. \tag{10}$$

- The true parameter vector $\mathbf{f}$ is considered as a random variable on which some prior knowledge may be available. More specifically, let the prior probability density function (PDF) $p(\mathbf{f})$ be characterized by its first and second order moments:

$$\boldsymbol{\mu}_{\mathbf{f}} \triangleq E\mathbf{f}, \tag{11}$$
$$\mathbf{R}_{\mathbf{f}} \triangleq \mathrm{cov}(\mathbf{f}) = E(\mathbf{f} - E\mathbf{f})(\mathbf{f} - E\mathbf{f})^T. \tag{12}$$

Then the linear MMSE estimator can be obtained as the mean of the posterior PDF $p(\mathbf{f}|\mathbf{y})$ after the data have been recorded:

$$\hat{\mathbf{f}}_{\mathbf{MMSE}} = E(\mathbf{f}|\mathbf{y}) = \boldsymbol{\mu} + (\mathbf{U}^T\mathbf{R}_{\mathbf{v}}{}^{-1}\mathbf{U} + \mathbf{R}_{\mathbf{f}}{}^{-1})^{-1}\mathbf{U}^T\mathbf{R}_{\mathbf{v}}{}^{-1}(\mathbf{y} - \mathbf{U}\boldsymbol{\mu}).$$

We will construct the prior knowledge on $\mathbf{f}$ in such a way that $\boldsymbol{\mu}_{\mathbf{f}} = \mathbf{0}$. Then, also using the white noise assumption in (10), the expression for $\hat{\mathbf{f}}_{\mathbf{MMSE}}$ can be rewritten as

$$\boxed{\hat{\mathbf{f}}_{\mathbf{MMSE}} = (\mathbf{U}^T\mathbf{U} + \sigma_v^2\mathbf{R}_{\mathbf{f}}{}^{-1})^{-1}\mathbf{U}^T\mathbf{y}.} \tag{13}$$

From this point of view, applying Tikhonov regularization as in (4) is equivalent to assuming that the true parameter vector $\mathbf{f}$ is drawn from a stationary white noise process. In many applications however, more information on the true parameter will be available and an appropriate non-identity covariance matrix $\mathbf{R}_{\mathbf{f}}$ may be constructed.

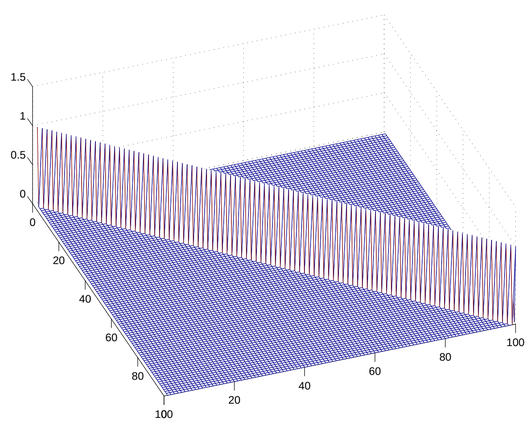
3 An example application: Acoustic echo cancellation

Consider an acoustic echo cancellation (AEC) application where:

- the loudspeaker signal $u(t)$ is an audio signal (typically speech),
- the near-end noise signal $v(t)$ is assumed to be drawn from a stationary white noise process,
- the acoustic impulse response (AIR) $F(q) = f_0 + f_1 q^{-1} + \dots + f_{n_F} q^{-n_F}$, with q the time shift operator, is assumed to be stationary, and
- the microphone signal is made up of the echo signal $x(t) = F(q)u(t)$ and the near-end noise signal: $y(t) = x(t) + v(t)$.

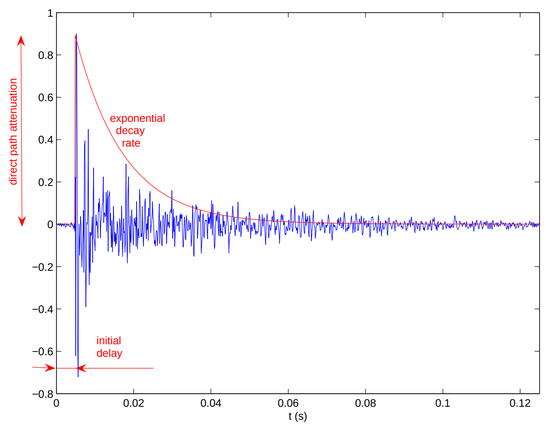
Typically the loudspeaker signal exhibits a tonal spectrum and the AIR is of high order. Therefore the lack of persistent excitation may indeed be a problem in the AEC application.

If an LS estimator with Tikhonov regularization is applied in the AEC problem, the regularization matrix $\mathbf{P} = \alpha\mathbf{I}$ (with the regularization parameter typically chosen as $\alpha = \sigma_v^2$) looks like this:

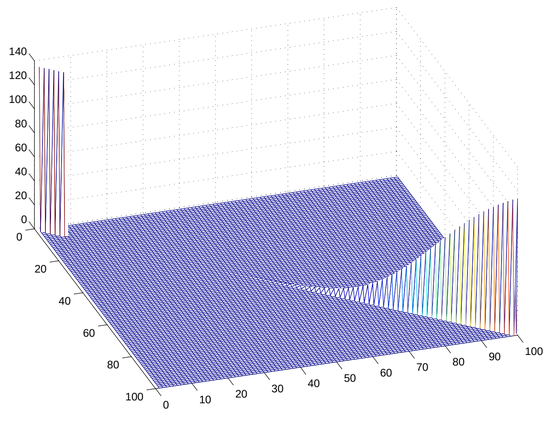


An acoustic impulse response has a very typical form, which may be characterized by three parameters:

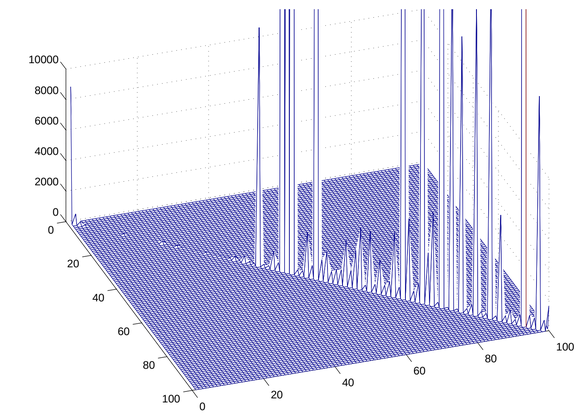
- the initial delay, which corresponds to the time needed by the loudspeaker sound wave to reach the microphone through a direct path (i.e. without reflections),
- the direct path attenuation, which determines the peak response in the AIR, and
- the exponential decay rate, which models the reverberant tail of the AIR.



These three parameters may be calculated from the acoustic setup (distance between loudspeaker and microphone, acoustic absorption of the walls, room volume, etc.). Hence they can be considered as prior knowledge. If these three parameters are taken into account, a diagonal regularization matrix may be constructed that looks like this:



An idealized case, which is interesting as a reference method, occurs when the true AIR is known. In this case a diagonal regularization matrix may be constructed with the diagonal elements equal to the inverse square values of the true parameter vector coefficients:



4 Simulation results

Four different least squares estimators were compared for the AEC application: the LS estimator $\hat{\mathbf{f}}_{\mathbf{LS}}$ and the regularized LS estimator $\hat{\mathbf{f}}_{\mathbf{RgLS}}$ with Tikhonov regularization ($\mathbf{P} = \sigma_v^2\mathbf{I}$), regularization based on the three AIR parameters described above ($\mathbf{P} = \sigma_v^2\mathbf{R}_{\mathbf{f},synth}^{-1}$), and on the true AIR ($\mathbf{P} = \sigma_v^2\mathbf{R}_{\mathbf{f},true}^{-1}$).

For each estimator, 100 simulation runs were performed with a different near-end noise realization.

Two types of loudspeaker signals were used: a sum of $n_F - 1$ sinusoids with random frequencies, uniformly distributed in the interval between DC and the Nyquist frequency, and a male speech signal.

Two types of acoustic impulse responses were used: a hearing aid impulse response with $n_F + 1 = 100$ coefficients, and a room impulse response with $n_F + 1 = 1000$ coefficients.

