Regularization in Linear Regression Problems

MATS VIBERG

Chalmers University of Technology

Department of Signals and Systems

SE-412 96 Göteborg, Sweden

Email: viberg@s2.chalmers.se

WIP/Beats Workshop 2004, Visby.

Abstract and Outline

This presentation gives a brief review of regularized least squares. Some motivations and interpretations are given, and methods for selecting the regularization parameter are summarized. Outline:

- Linear regression using least-squares
- Regularized (Tikhonov) least-squares
- Selection of the regularization parameter
- Least-squares with unknown noise color
- Weighted least-squares with regularization
- Hyper-parameter estimation

Linear Regression Problem

Measured signal $\mathbf{y}$ $(N \times 1)$ contains signal and noise:

$$\mathbf{y} = \mathbf{x} + \mathbf{e}$$

Signal part modeled by basis function expansion:

$$\mathbf{x} = \sum_{k=1}^{n} \mathbf{a}_k s_k = [\mathbf{a}_1, \ldots, \mathbf{a}_n]\mathbf{s} = \mathbf{A}\mathbf{s}\,,\ n \leq N.$$

We assume noise $\mathbf{e}$ is zero-mean white, $E[\mathbf{e}\mathbf{e}^*] = \sigma_e^2 \mathbf{I}$ (colored noise considered later).

Given $\mathbf{y}$ we wish to:

- Determine signal amplitudes $\mathbf{s}$ - detection/classification
- Reconstruct $\mathbf{x}$ - filtering/smoothing/prediction

Signal Processing Application

Wireless channel prediction using sinusoidal modeling:

$$y(t) = \sum_{k=1}^{n} s_k e^{j\omega_k t} + e(t)$$

The vector of observations for $t = 0, 1, \ldots, N-1$ is

$$\mathbf{y} = [y(0), \ldots, y(N-1)]^T = \sum_{k=1}^{n} \mathbf{a}(\omega_k) s_k + \mathbf{e} = \mathbf{A}\mathbf{s} + \mathbf{e}$$

where $\mathbf{a}(\omega) = [1, e^{j\omega}, \ldots, e^{j(N-1)\omega}]^T$ is the DFT vector.

Assume frequencies ω_k known (or accurately estimated). The task is to estimate amplitudes $\mathbf{s}$ and then predict future values of $y(t)$!

Difficulty: ω_ks tend to be closely spaced $\Rightarrow$ $\mathbf{A}$ is ill-conditioned!

Linear Regression

Least-squares solution:

$$\hat{\mathbf{s}}_{LS} = \arg\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{A}\mathbf{s}\|^2$$

Leads to

$$\begin{aligned} \hat{\mathbf{s}}_{LS} &= (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{y} = \mathbf{A}^\dagger\mathbf{y} \\ \hat{\mathbf{x}}_{LS} &= \mathbf{A}\hat{\mathbf{s}}_{LS} = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{y} = \mathbf{\Pi}_{\mathbf{A}}\mathbf{y} \end{aligned}$$

Motivations:

- $\hat{\mathbf{s}}_{LS}$ is the BLUE (Best Linear Unbiased Estimator)
- If $\mathbf{e}$ is Gaussian distributed, $\hat{\mathbf{s}}_{LS}$ is also ML

So what is the problem?

Least-Squares Performance

Amplitude estimation performance

Insert $\mathbf{y}$ into $\hat{\mathbf{s}}_{LS}$:

$$\hat{\mathbf{s}}_{LS} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*(\mathbf{A}\mathbf{s} + \mathbf{e})$$

which gives

$$MSE_{LS,s} = E[(\hat{\mathbf{s}}_{LS} - \mathbf{s})(\hat{\mathbf{s}}_{LS} - \mathbf{s})^*] = \sigma_e^2(\mathbf{A}^*\mathbf{A})^{-1}$$

Potential trouble if $\mathbf{A}$ ill-conditioned; $\mathbf{A}^*\mathbf{A}$ nearly singular!

Let the SVD of $\mathbf{A}$ be $\mathbf{A} = \sum_{k=1}^{n} \mathbf{u}_k \sigma_k \mathbf{v}_k^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. Then,

$$MSE_{LS,s} = \sigma_e^2 \mathbf{V}\mathbf{\Sigma}^{-2}\mathbf{V}^*$$

Least-Squares Performance

Signal estimation (prediction) performance

The reconstructed signal is

$$\hat{\mathbf{x}}_{LS} = \mathbf{A}\hat{\mathbf{s}}_{LS} = \mathbf{x} + \mathbf{\Pi_A e}$$

so

$$MSE_{LS,x} = E[(\hat{\mathbf{x}}_{LS} - \mathbf{x})(\hat{\mathbf{x}}_{LS} - \mathbf{x})^*] = \sigma_e^2 \mathbf{\Pi_A}$$

We can define an average signal estimation error:

$$\overline{MSE}_{LS,x} = \frac{1}{N} E[\|\hat{\mathbf{x}}_{LS} - \mathbf{x}\|^2] = \frac{1}{N}\sigma_e^2 \mathrm{Tr}\{\mathbf{\Pi_A}\} = \sigma_e^2 \frac{n}{N}$$

Works fine as long as $n \ll N$, independent of **A** and **s**!

Regularization

Regularization is a way to avoid ill-conditioning, both for numerical and statistical reasons!

Motivation 1: Ill-conditioning leads to large $\|\mathbf{s}\|$. Add a penalty term:

$$\hat{\mathbf{s}}_{Reg} = \arg\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{A}\mathbf{s}\|^2 + \lambda \|\mathbf{s}\|^2$$

where $\lambda > 0$ is the *regularization parameter*. The solution is

$$\hat{\mathbf{s}}_{Reg} = (\mathbf{A}^*\mathbf{A} + \lambda \mathbf{I})^{-1}\mathbf{A}^*\mathbf{y} = \mathbf{R}_{\lambda}^{-1}\mathbf{A}^*\mathbf{y}$$

Motivation 2: Model $\mathbf{s}$ as zero-mean random with $E[\mathbf{s}\mathbf{s}^*] = \sigma_s^2\mathbf{I}$. Then, LMMSE (Linear Minimum Mean Square Error Estimate) of $\mathbf{s}$ is

$$\hat{\mathbf{s}}_{Reg} = \mathbf{R}_{\lambda}^{-1}\mathbf{A}^*\mathbf{y}$$

with $\lambda = \sigma_e^2/\sigma_s^2 = SNR^{-1}$.

Regularized LS Performance

Amplitude estimation performance

We first compute the average performance, assuming $\mathbf{s}$ is random:

$$\hat{\mathbf{s}}_{Reg} = \mathbf{R}_\lambda^{-1}\mathbf{A}^*(\mathbf{A}\mathbf{s} + \mathbf{e}) = \mathbf{s} - \lambda\mathbf{R}_\lambda^{-1}\mathbf{s} + \mathbf{R}_\lambda^{-1}\mathbf{A}^*\mathbf{e}$$

Thus, with $\lambda = \sigma_e^2/\sigma_s^2$ we have

$$\begin{aligned} MSE_{Reg,s} &= \lambda^2\sigma_s^2\mathbf{R}_\lambda^{-2} + \sigma_e^2\mathbf{R}_\lambda^{-1}\mathbf{A}^*\mathbf{A}\mathbf{R}_\lambda^{-1} \\ &= \sigma_e^2\mathbf{R}_\lambda^{-1} \end{aligned}$$

In terms of the SVD of $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$:

$$MSE_{Reg,s} = \sigma_e^2\mathbf{V}(\Sigma^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^*$$

Regularized LS Performance

Signal estimation (prediction) performance

$$MSE_{Reg,x} = \sigma_e^2 \mathbf{A}\mathbf{R}_\lambda^{-1}\mathbf{A}^*$$

The average signal reconstruction error is

$$\overline{MSE}_{Reg,x} = \frac{1}{N}\sigma_e^2 \mathrm{Tr}\{\mathbf{A}\mathbf{R}_\lambda^{-1}\mathbf{A}^*\}$$

In terms of the singular values $\{\sigma_k\}_{k=1}^n$ this becomes

$$\overline{MSE}_{Reg,x} = \frac{\sigma_e^2}{N}\sum_{k=1}^{n}\frac{\sigma_k^2}{\sigma_k^2+\lambda} = \overline{MSE}_{LS,x} - \frac{\sigma_e^2}{N}\sum_{k=1}^{n}\frac{\lambda}{\sigma_k^2+\lambda}$$

Interpretation: If $\alpha\%$ of the singular values obey $\sigma_k^2 \ll \lambda = \sigma_e^2/\sigma_s^2$, then the MSE is reduced by (at least) $\alpha\%$!

Performance for a Fixed Amplitude

What can we say for a fixed **s**? LS performance independent of **s**! For regularized LS:

$$MSE_{Reg,s|s} = \lambda^2 \mathbf{R}_\lambda^{-1} \mathbf{s}\mathbf{s}^* \mathbf{R}_\lambda^{-1} + \sigma^2 \mathbf{R}_\lambda^{-1} \mathbf{A}^* \mathbf{A} \mathbf{R}_\lambda^{-1}$$

Easy to show that

$$\sigma^2 \mathbf{R}_\lambda^{-1} \mathbf{A}^* \mathbf{A} \mathbf{R}_\lambda^{-1} \leq \sigma^2 (\mathbf{A}^* \mathbf{A})^{-1}$$

and for a fixed value of λ we find

- For small $\|\mathbf{s}\|$: $MSE_{Reg,s|s} < MSE_{LS,s|s}$
- For large $\|\mathbf{s}\|$: $MSE_{Reg,s|s} > MSE_{LS,s|s}$

Optimality

We conclude that no linear estimator is uniformly optimal!

Easy to see that

$$\left.\frac{\partial MSE_{Reg,s|s}}{\partial \lambda}\right|_{\lambda=0} < 0$$

so regularization is always better if λ is "small enough".

From Stein's classical result ($\mathbf{A} = \mathbf{I}$, $n = N$) we know that LS is not *admissible*, there exist other estimators that are uniformly (for all $\mathbf{s}$) better!

Choice of Regularization Parameter

It seems like a good idea to determine λ from data! A direct MSE optimization would be:

1. Use a "reasonable" λ and obtain preliminary estimates $\hat{\sigma}^2$ and $\hat{\mathbf{s}}$
2. Choose "optimal" λ by minimizing $MSE_{Reg,s|s}$ or $MSE_{Reg,x|s}$, evaluated at $\hat{\sigma}$ and $\hat{\mathbf{s}}$
3. Compute improved $\hat{\mathbf{s}}$ and $\hat{\mathbf{x}}$

Unfortunately, it does not work; λ will tend to 0!

Popular methods that work:

- Cross-validation techniques
- Hyper-parameter estimation (ML or Bayesian)

Cross-Validation Techniques

The Jack-knife (leave-one-out): compute $\hat{\hat{x}}_k(\lambda)$ using $y_l,\ l \neq k$. Determine λ by minimizing

$$CV(\lambda) = \sum_{k=1}^{N} |y_k - \hat{\hat{x}}_k(\lambda)|^2$$

Variation: Leave K out, $K > 1$

Generalized Cross-Validation [Golub and Heath, 1979]: select λ to minimize

$$GCV(\lambda) = \frac{\|\mathbf{y} - \hat{\mathbf{x}}_{Reg}(\lambda)\|}{\mathrm{Tr}\{\mathbf{I} - \mathbf{A}\mathbf{R}_\lambda^{-1}\mathbf{A}^*\}}$$

This can be interpreted as leave-one-out applied to transformed data!

Hyper-Parameter Estimation

An interesting idea is to *estimate* λ using e.g. ML. This is possible only if the *marginalized likelihood* w.r.t. **s** is used. Galatsanos and Katsaggelos (1992) used the Gaussian prior:

$$\mathbf{s} \in \mathcal{N}(0, \lambda/\sigma^2 \mathbf{I})$$

The likelihood function for σ^2 and λ is now

$$f_y(\mathbf{y}; \sigma^2, \lambda) = \int_{\mathbf{s}} f_{y|s}(\mathbf{y}|\mathbf{s}, \sigma^2, \lambda) f_s(\mathbf{s}; \sigma^2, \lambda) d\mathbf{s}$$

and the maximizing arguments yield $\hat{\sigma}^2$ and $\hat{\lambda}$.

Fortunately, the integral can be solved in closed form. It is also possible to find $\hat{\sigma}^2$ in terms of λ, but $\hat{\lambda}$ requires a scalar search!

Hyper-Parameter Estimation

The resulting estimates are

$$\hat{\sigma}_e^2 = \frac{1}{N}\mathbf{y}^*(\mathbf{I} - \mathbf{A}\mathbf{R}_\lambda^{-1}\mathbf{A}^*)\mathbf{y}$$

and

$$\hat{\lambda} = \arg\min_{\lambda} ML(\lambda)$$

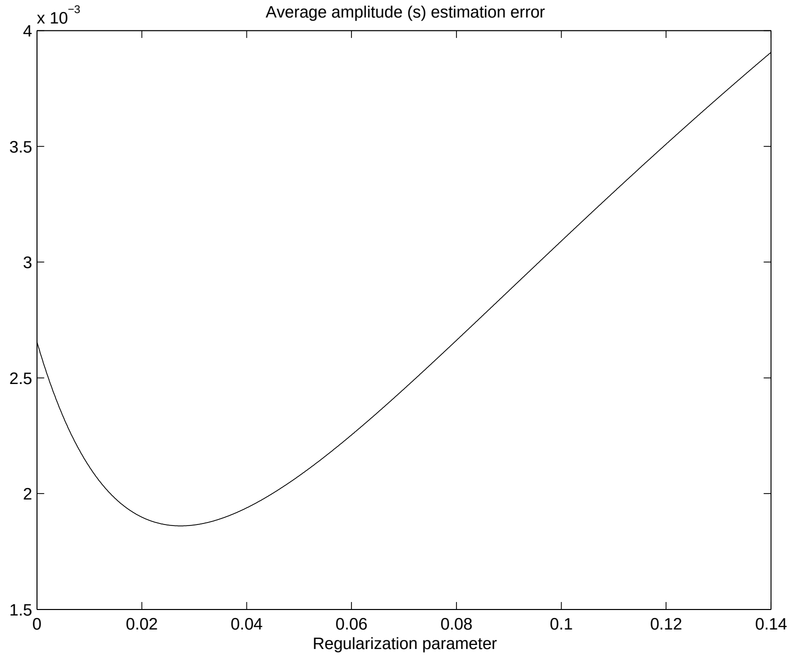
where

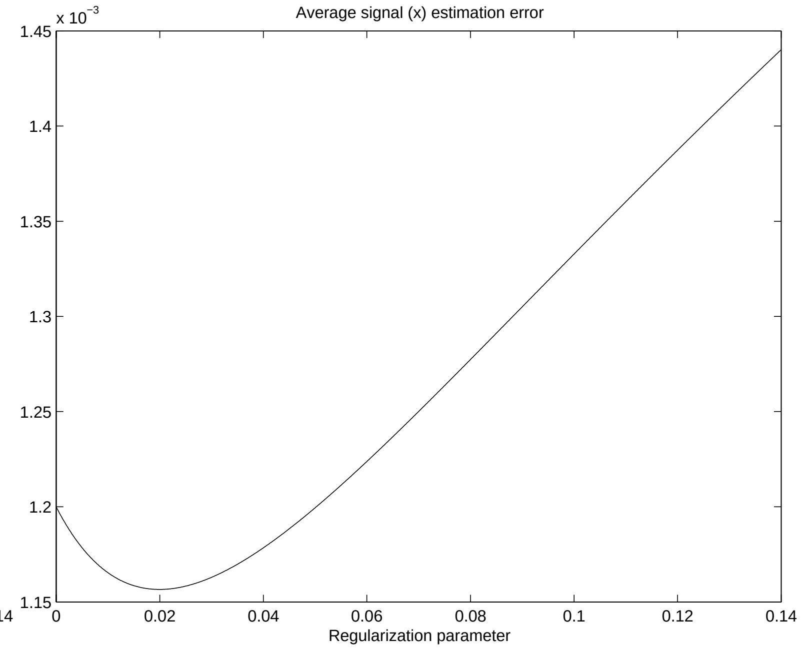
$$ML(\lambda) = \log|\mathbf{y}^*(\mathbf{I} - \mathbf{A}\mathbf{R}_\lambda^{-1}\mathbf{A}^*)\mathbf{y}| - \frac{1}{N}\log|\mathbf{I} - \mathbf{A}\mathbf{R}_\lambda^{-1}\mathbf{A}^*|$$

(Another possibility: assign prior on λ and marginalize w.r.t. λ instead)

Example: Mean Square Error Performance

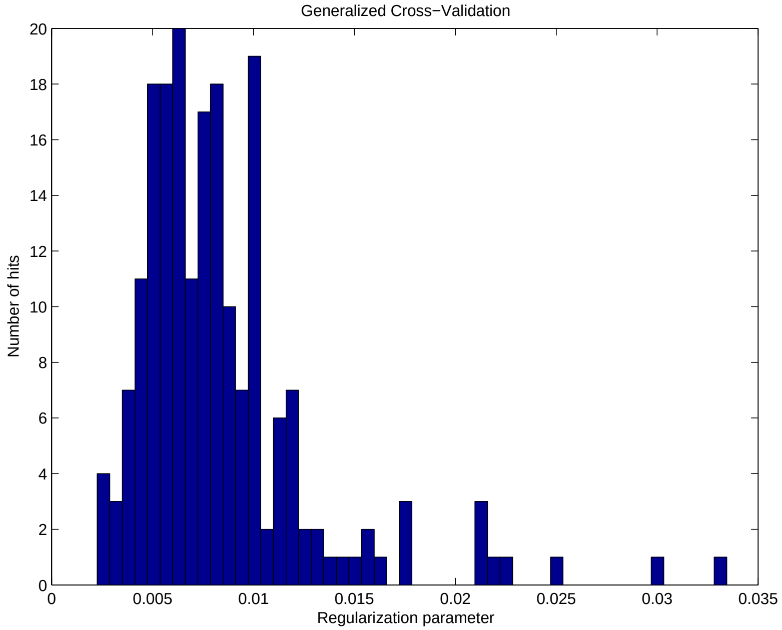
$E\|\hat{\mathbf{s}}(\lambda) - \mathbf{s}\|^2/N$ and $E\|\hat{\mathbf{x}}(\lambda) - \mathbf{x}\|^2/N$ vs λ for "representative" scenario:

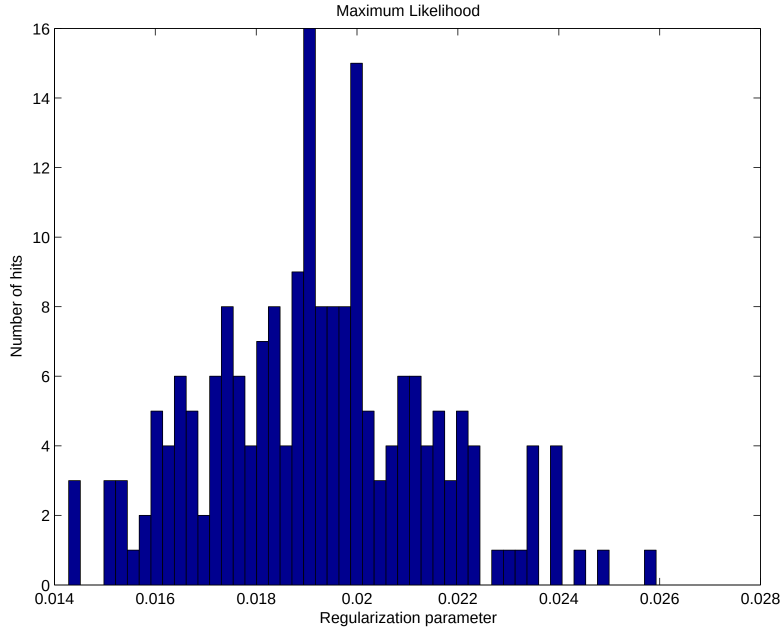




Example: Histograms

Histograms for $\hat{\lambda}_{GCV}$ (left) and $\hat{\lambda}_{ML}$ (right)





Example: "Conclusions"

In this example:

- Both ML and GCV yield $\hat{\lambda}$ that reduce the MSE over LS
- ML "estimates" have lower variance, and are clustered around the value that miminizes $\overline{MSE}_{Reg,x|s}$
- GCV tends to chose λ "too small"

Unknown Noise Color

A related problem ($n = 1$ for simplicity):

$$\mathbf{y} = \mathbf{x} + \mathbf{e} = \mathbf{a}\, s + \mathbf{e}$$

where $E[\mathbf{e}\mathbf{e}^*] = \mathbf{Q}$ is unknown. Noise color estimated from training data:

$$\mathbf{Z} = [\mathbf{z}(0), \mathbf{z}(1), \ldots, \mathbf{z}(M-1)]$$

with $E[\mathbf{z}(k)\mathbf{z}^*(l)] = \mathbf{Q}\, \delta_{k,l}$.

Weighted Least-Squares (WLS) with Certainty Equivalence (CE):

$$\begin{aligned} \hat{\mathbf{Q}} &= \frac{1}{M}\mathbf{Z}\mathbf{Z}^* \\ \hat{s} &= (\mathbf{a}^*\hat{\mathbf{Q}}^{-1}\mathbf{a})^{-1}\mathbf{a}^*\hat{\mathbf{Q}}^{-1}\mathbf{y} \end{aligned}$$

Poor performance if M "too small", and even impossible if $M < N$!

Signal Processing Application

Space-Time Adaptive Processing (STAP) in radar: $\mathbf{x}$ contains backscattered signal from moving target, received at K antennas during L pulses:

$$\mathbf{a} = \mathbf{a}(\theta) \otimes \mathbf{a}(\omega) \quad (N \times 1)\,, N = KL$$

Here, θ is the Direction-of-Arrival and ω the target Doppler frequency.

Major noise source: ground clutter – highly structured space-time color!

Noise color estimated using secondary data at other carrier frequencies and/or range bins. Usually not enough data!

Maximum Likelihood Estimation

If $\mathbf{e}$ and $\mathbf{z}(k)$ are $\mathcal{N}(0, \mathbf{Q})$, the joint MLE is identical to WLS-CE:

$$\begin{aligned} \hat{\mathbf{Q}} &= \frac{1}{M}\mathbf{Z}\mathbf{Z}^* \\ \hat{s} &= (\mathbf{a}^*\hat{\mathbf{Q}}^{-1}\mathbf{a})^{-1}\mathbf{a}^*\hat{\mathbf{Q}}^{-1}\mathbf{y} \end{aligned}$$

"Bayesian" likelihood: assign prior $\mathbf{Q}^{-1} \in f_Q(\mathbf{Q}^{-1})$ and marginalize:

$$f_y(\mathbf{y}; s) = \int_{\mathbf{Q}} f_{y|Q}(\mathbf{y}|s, \mathbf{Q}^{-1}) f_Q(\mathbf{Q}^{-1}) d\mathbf{Q}^{-1}$$

Then, the MLE of the signal amplitude is

$$\hat{s}_{ML} = \arg\max_s f_y(\mathbf{y}; s)$$

Choice of Prior Distribution

Non-informative priors add as little information as possible and are parameterization invariant!

Jeffrey's prior

$$f_Q(\mathbf{Q}^{-1}) \propto |\mathbf{FIM}|^{1/2} \propto |\mathbf{Q}^{-1}|^{-N}$$

(**FIM** is the Fisher Information Matrix)

Bad luck: using Jeffrey's prior also leads to WLS-CE:

$$\hat{s} = (\mathbf{a}^* \hat{\mathbf{Q}}^{-1} \mathbf{a})^{-1} \mathbf{a}^* \hat{\mathbf{Q}}^{-1} \mathbf{y}$$

Reference prior is more "non-informative" than Jeffrey's, but does not allow an explicit solution (MCMC sampling)!

Regularization Prior

Regularization has been found to work well in practice. Use WLS with

$$\hat{\mathbf{Q}}_\lambda = \frac{1}{M}\mathbf{Z}\mathbf{Z}^* + \lambda\mathbf{I}$$

The problem is to choose λ!

The above is like saying we have extra training data with sample covariance $\lambda\mathbf{I}$. We can as well move this to the prior for $\mathbf{Q}^{-1}$:

$$f_Q(\mathbf{Q}, \lambda) \propto |\mathbf{Q}^{-1}|^{-K} \mathrm{etr}\{-\mathbf{Q}^{-1}\lambda\}$$

Interpreting this as a *regularization prior* we can determine λ by hyper-parameter estimation!

Hyperparameter Estimation

Using the regularization prior results in

$$\hat{s}(\lambda) = (\mathbf{a}^* \hat{\mathbf{Q}}_\lambda^{-1} \mathbf{a})^{-1} \mathbf{a}^* \hat{\mathbf{Q}}_\lambda^{-1} \mathbf{y}$$
$$\hat{\mathbf{Q}}_\lambda = \frac{1}{M} \mathbf{Z}\mathbf{Z}^* + \lambda \mathbf{I}$$

and we can estimate the hyper-parameter by

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} f_y(\mathbf{y}; \hat{s}(\lambda), \lambda)$$

where $f_y(\mathbf{y}; \hat{s}(\lambda), \lambda)$ is given in closed form.

(It is also possible to assign prior $f_\lambda(\lambda)$ and integrate again!)

Concluding Remarks

- Linear regression looks simple, but fortunately we can make it quite complicated!
- Regularization is very useful to deal both with numerical problems and sensitivity to noise and model imperfections
- The regularization parameter λ can be set from prior info, or from data
- Data-driven methods for selecting λ generally work well
- Our example favors ML over GCV but no generality is claimed
- Regularized WLS can be interpreted using a prior distribution of the noise covariance
- Regularization prior allows selecting λ by hyper-parameter estimation!