Derivation of the Validation MSE and its Derivative w.r.t. λ

"All models are wrong, but some are useful." — G. E. P. Box

1 Problem setup and notation

Let

- $N \in \mathbb{N}$ be the length of the time series.
- $\boldsymbol{Z} \in \mathbb{R}^N$ be the observed data vector:

$$\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_N \end{bmatrix}.$$

- $d \in \mathbb{N}$ be the order of differencing used in the penalty.
- $K \in \mathbb{R}^{(N-d)\times N}$ be the differencing matrix of order d.
- $\lambda > 0$ be the smoothing (penalty) parameter.

We consider the penalized least squares problem

$$\min_{\boldsymbol{t}\in\mathbb{R}^N} J(\boldsymbol{t};\lambda) := \big\|\boldsymbol{Z} - \boldsymbol{t}\big\|_2^2 + \lambda\big\|K\boldsymbol{t}\big\|_2^2. \tag{1}$$

The goal is:

- For each $\lambda > 0$, compute the trend estimator

$$\widehat{\boldsymbol{t}}(\lambda) = \arg\min_{\boldsymbol{t}} J(\boldsymbol{t};\lambda).$$

- On a separate validation set, define the mean squared error as a function of λ,

$$f(\lambda) := \mathrm{MSE}_{\mathrm{val}}(\lambda),$$

and derive an explicit expression for the derivative $\dfrac{d}{d\lambda} f(\lambda)$.

2 Closed form of the penalized estimator

We solve (1) explicitly.

First note that

$$\|\boldsymbol{Z}-\boldsymbol{t}\|_2^2 = (\boldsymbol{Z}-\boldsymbol{t})^\top(\boldsymbol{Z}-\boldsymbol{t}), \qquad \|K\boldsymbol{t}\|_2^2 = (K\boldsymbol{t})^\top(K\boldsymbol{t}).$$

Thus

$$\begin{aligned} J(\boldsymbol{t};\lambda) &= (\boldsymbol{Z}-\boldsymbol{t})^\top(\boldsymbol{Z}-\boldsymbol{t}) + \lambda(K\boldsymbol{t})^\top(K\boldsymbol{t}) \\ &= (\boldsymbol{Z}-\boldsymbol{t})^\top(\boldsymbol{Z}-\boldsymbol{t}) + \lambda\boldsymbol{t}^\top K^\top K\boldsymbol{t}. \end{aligned}$$

Expand the first quadratic term:

$$(\boldsymbol{Z}-\boldsymbol{t})^\top(\boldsymbol{Z}-\boldsymbol{t}) = \boldsymbol{Z}^\top\boldsymbol{Z} - 2\boldsymbol{Z}^\top\boldsymbol{t} + \boldsymbol{t}^\top\boldsymbol{t}.$$

Hence

$$J(\boldsymbol{t};\lambda) = \boldsymbol{Z}^\top\boldsymbol{Z} - 2\boldsymbol{Z}^\top\boldsymbol{t} + \boldsymbol{t}^\top\boldsymbol{t} + \lambda\boldsymbol{t}^\top K^\top K\boldsymbol{t}.$$

Group the terms that depend on $\boldsymbol{t}$:

$$\begin{aligned} J(\boldsymbol{t};\lambda) &= \boldsymbol{Z}^\top\boldsymbol{Z} + \left(\boldsymbol{t}^\top\boldsymbol{t} + \lambda\boldsymbol{t}^\top K^\top K\boldsymbol{t} - 2\boldsymbol{Z}^\top\boldsymbol{t}\right) \\ &= \boldsymbol{Z}^\top\boldsymbol{Z} + \left(\boldsymbol{t}^\top(I_N + \lambda K^\top K)\boldsymbol{t} - 2\boldsymbol{Z}^\top\boldsymbol{t}\right), \end{aligned}$$

where I_N is the $N \times N$ identity matrix.

Define the symmetric positive definite matrix

$$A(\lambda) := I_N + \lambda K^\top K \in \mathbb{R}^{N\times N}.$$

Then

$$J(\boldsymbol{t};\lambda) = \boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{t}^\top A(\lambda)\boldsymbol{t} - 2\boldsymbol{Z}^\top\boldsymbol{t}.$$

To find the minimizer, compute the gradient of J with respect to $\boldsymbol{t}$ and set it equal to $\mathbf{0}$:

$$\nabla_{\boldsymbol{t}} J(\boldsymbol{t};\lambda) = \mathbf{0}.$$

Recall:

- If A is symmetric, then $\nabla_{\boldsymbol{t}}(\boldsymbol{t}^\top A\boldsymbol{t}) = 2A\boldsymbol{t}$.
- $\nabla_{\boldsymbol{t}}(\boldsymbol{Z}^\top\boldsymbol{t}) = \boldsymbol{Z}$.

Apply this:

$$\begin{aligned}\nabla_{\boldsymbol{t}} J(\boldsymbol{t};\lambda) &= \nabla_{\boldsymbol{t}}\left[\boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{t}^\top A(\lambda)\boldsymbol{t} - 2\boldsymbol{Z}^\top \boldsymbol{t}\right] \\ &= 0 + 2A(\lambda)\boldsymbol{t} - 2\boldsymbol{Z}.\end{aligned}$$

Set gradient equal to $\mathbf{0}$:

$$\begin{aligned}\nabla_{\boldsymbol{t}} J(\boldsymbol{t};\lambda) = \mathbf{0} &\iff 2A(\lambda)\boldsymbol{t} - 2\boldsymbol{Z} = \mathbf{0} \\ &\iff 2A(\lambda)\boldsymbol{t} = 2\boldsymbol{Z} \\ &\iff A(\lambda)\boldsymbol{t} = \boldsymbol{Z}.\end{aligned}$$

Thus the minimizer solves the linear system

$$A(\lambda)\widehat{\boldsymbol{t}}(\lambda) = \boldsymbol{Z}.$$

Since $A(\lambda)$ is symmetric positive definite, it is invertible, so

$$\widehat{\boldsymbol{t}}(\lambda) = A(\lambda)^{-1}\boldsymbol{Z}. \tag{2}$$

Define the *smoothing matrix*

$$S(\lambda) := A(\lambda)^{-1} = \left(I_N + \lambda K^\top K\right)^{-1}.$$

Then

$$\widehat{\boldsymbol{t}}(\lambda) = S(\lambda)\boldsymbol{Z}. \tag{3}$$

3 Validation set and validation MSE

Let $\mathcal{V} \subset \{1, 2, \dots, N\}$ denote the index set of validation observations, and let

$$n_{\text{val}} := |\mathcal{V}|$$

be the number of validation points.

Define the *selection matrix* $R \in \mathbb{R}^{n_{\text{val}} \times N}$ such that

$$\boldsymbol{Z}_{\text{val}} := R\boldsymbol{Z} \in \mathbb{R}^{n_{\text{val}}}$$

is the vector of validation observations. Concretely, each row of R is a standard basis vector selecting the corresponding index in $\mathcal{V}$.

Similarly, the fitted trend on the validation points is

$$\widehat{\boldsymbol{t}}_{\text{val}}(\lambda) := R\widehat{\boldsymbol{t}}(\lambda) = RS(\lambda)\boldsymbol{Z}.$$

Define the validation residual vector

$$\boldsymbol{r}(\lambda) := \boldsymbol{Z}_{\text{val}} - \widehat{\boldsymbol{t}}_{\text{val}}(\lambda) = R\boldsymbol{Z} - RS(\lambda)\boldsymbol{Z}. \tag{4}$$

The validation mean squared error (MSE) as a function of λ is

$$f(\lambda) := \text{MSE}_{\text{val}}(\lambda) := \frac{1}{n_{\text{val}}} \left\| \boldsymbol{r}(\lambda) \right\|_2^2 = \frac{1}{n_{\text{val}}} \boldsymbol{r}(\lambda)^\top \boldsymbol{r}(\lambda). \tag{5}$$

Our objective is now:

$$\text{derive } \frac{d}{d\lambda} f(\lambda).$$

4 Derivative of the smoothing matrix $S(\lambda)$

Recall

$$S(\lambda) = A(\lambda)^{-1}, \qquad A(\lambda) = I_N + \lambda K^\top K.$$

First compute the derivative of $A(\lambda)$:

$$\frac{d}{d\lambda} A(\lambda) = \frac{d}{d\lambda}\big(I_N + \lambda K^\top K\big) = \mathbf{0} + 1 \cdot K^\top K = K^\top K.$$

Now use the well-known formula for the derivative of an inverse:

$$\frac{d}{d\lambda} A(\lambda)^{-1} = -A(\lambda)^{-1} \left(\frac{d}{d\lambda} A(\lambda) \right) A(\lambda)^{-1}.$$

Apply this with $A(\lambda)$ above:

$$\begin{aligned} \frac{d}{d\lambda} S(\lambda) &= \frac{d}{d\lambda} A(\lambda)^{-1} \\ &= -A(\lambda)^{-1}\big(K^\top K\big)A(\lambda)^{-1} \\ &= -S(\lambda) K^\top K S(\lambda). \end{aligned}$$

Thus

$$S'(\lambda) := \frac{d}{d\lambda} S(\lambda) = -S(\lambda) K^\top K S(\lambda). \tag{6}$$

5 Derivative of the validation residual $\boldsymbol{r}(\lambda)$

From (4), we have

$$\boldsymbol{r}(\lambda) = R\boldsymbol{Z} - RS(\lambda)\boldsymbol{Z}.$$

Differentiate with respect to λ:

$$\begin{aligned}\frac{d}{d\lambda}\boldsymbol{r}(\lambda) &= \frac{d}{d\lambda}\big(R\boldsymbol{Z} - RS(\lambda)\boldsymbol{Z}\big)\\ &= \frac{d}{d\lambda}(R\boldsymbol{Z}) - \frac{d}{d\lambda}(RS(\lambda)\boldsymbol{Z}).\end{aligned}$$

Note:

- R does not depend on λ.
- $\boldsymbol{Z}$ does not depend on λ.

Therefore

$$\frac{d}{d\lambda}(R\boldsymbol{Z}) = R \cdot \mathbf{0} = \mathbf{0}.$$

For the second term, treat R as constant and apply the chain rule:

$$\frac{d}{d\lambda}\big(RS(\lambda)\boldsymbol{Z}\big) = R\frac{d}{d\lambda}\big(S(\lambda)\boldsymbol{Z}\big).$$

Since $\boldsymbol{Z}$ is constant, we differentiate $S(\lambda)$:

$$\frac{d}{d\lambda}\big(S(\lambda)\boldsymbol{Z}\big) = S'(\lambda)\boldsymbol{Z}.$$

Thus

$$\frac{d}{d\lambda}\big(RS(\lambda)\boldsymbol{Z}\big) = RS'(\lambda)\boldsymbol{Z}.$$

Combining both pieces:

$$\begin{aligned}\frac{d}{d\lambda}\boldsymbol{r}(\lambda) &= \mathbf{0} - RS'(\lambda)\boldsymbol{Z}\\ &= -RS'(\lambda)\boldsymbol{Z}.\end{aligned}$$

Now substitute $S'(\lambda)$ from (6):

$$S'(\lambda) = -S(\lambda)K^\top KS(\lambda).$$

Therefore

$$\begin{aligned}\frac{d}{d\lambda}\boldsymbol{r}(\lambda) &= -R\left(-S(\lambda)K^\top KS(\lambda)\right)\boldsymbol{Z} \\ &= RS(\lambda)K^\top KS(\lambda)\boldsymbol{Z}.\end{aligned}$$

So we have

$$\boldsymbol{r}'(\lambda) := \frac{d}{d\lambda}\boldsymbol{r}(\lambda) = RS(\lambda)K^\top KS(\lambda)\boldsymbol{Z}. \tag{7}$$

6 Derivative of the validation MSE $f(\lambda)$

Recall the definition from (5):

$$f(\lambda) = \frac{1}{n_{\text{val}}}\boldsymbol{r}(\lambda)^\top \boldsymbol{r}(\lambda).$$

Let us write explicitly:

$$f(\lambda) = \frac{1}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}} r_i(\lambda)^2, \tag{8}$$

where $r_i(\lambda)$ is the i-th component of the vector $\boldsymbol{r}(\lambda)$.

6.1 Step-by-step scalar derivative

Differentiate (8) term by term:

$$\begin{aligned}\frac{d}{d\lambda}f(\lambda) &= \frac{d}{d\lambda}\left(\frac{1}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}} r_i(\lambda)^2\right) \\ &= \frac{1}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}}\frac{d}{d\lambda}\big(r_i(\lambda)^2\big).\end{aligned}$$

For each i, apply the chain rule:

$$\frac{d}{d\lambda}\big(r_i(\lambda)^2\big) = 2r_i(\lambda)\cdot r_i'(\lambda).$$

Thus

$$\begin{aligned}\frac{d}{d\lambda}f(\lambda) &= \frac{1}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}} 2r_i(\lambda)r_i'(\lambda) \\ &= \frac{2}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}} r_i(\lambda)r_i'(\lambda).\end{aligned}$$

Now rewrite the sum as an inner product of vectors:

$$\sum_{i=1}^{n_{\text{val}}} r_i(\lambda) r_i'(\lambda) = \boldsymbol{r}(\lambda)^\top \boldsymbol{r}'(\lambda).$$

Therefore

$$f'(\lambda) := \frac{d}{d\lambda} f(\lambda) = \frac{2}{n_{\text{val}}} \boldsymbol{r}(\lambda)^\top \boldsymbol{r}'(\lambda). \tag{9}$$

6.2 Substituting $\boldsymbol{r}(\lambda)$ and $\boldsymbol{r}'(\lambda)$

From (4):

$$\boldsymbol{r}(\lambda) = R\boldsymbol{Z} - RS(\lambda)\boldsymbol{Z} = R\big(I_N - S(\lambda)\big)\boldsymbol{Z}.$$

From (7):

$$\boldsymbol{r}'(\lambda) = RS(\lambda)K^\top KS(\lambda)\boldsymbol{Z}.$$

Substitute these into (9):

$$f'(\lambda) = \frac{2}{n_{\text{val}}} \Big(R\big(I_N - S(\lambda)\big)\boldsymbol{Z}\Big)^\top \Big(RS(\lambda)K^\top KS(\lambda)\boldsymbol{Z}\Big).$$

Note that we cannot simplify much further in general, because R may not be square or symmetric. This is already a compact expression.

Hence the final expression for the derivative is

$$f'(\lambda) = \frac{2}{n_{\text{val}}} \left[R\big(I_N - S(\lambda)\big)\boldsymbol{Z}\right]^\top \left[RS(\lambda)K^\top KS(\lambda)\boldsymbol{Z}\right]. \tag{10}$$

7 Special case: validation on all points

If the validation set is the entire sample, then $R = I_N$ and $n_{\text{val}} = N$.

In this case,

$$\boldsymbol{r}(\lambda) = (I_N - S(\lambda))\boldsymbol{Z},$$

and

$$\boldsymbol{r}'(\lambda) = S(\lambda)K^\top KS(\lambda)\boldsymbol{Z}.$$

Thus

$$f(\lambda) = \frac{1}{N}\big\|(I_N - S(\lambda))\boldsymbol{Z}\big\|_2^2,$$
$$f'(\lambda) = \frac{2}{N}\left[(I_N - S(\lambda))\boldsymbol{Z}\right]^\top \left[S(\lambda)K^\top KS(\lambda)\boldsymbol{Z}\right].$$

8 On solving $f'(\lambda) = 0$

To find the optimal $\lambda^\star$ that minimizes the validation MSE, one would like to solve

$$f'(\lambda^\star) = 0.$$

However, even if we diagonalize

$$K^\top K = U\Lambda U^\top, \quad \Lambda = \mathrm{diag}(\delta_1, \ldots, \delta_N),$$

and write $S(\lambda) = (I + \lambda K^\top K)^{-1} = U(I + \lambda\Lambda)^{-1}U^\top$, the expression for $f(\lambda)$ becomes a sum of rational functions in λ of the form

$$\frac{(\text{polynomial in } \lambda)}{\prod_j (1 + \lambda\delta_j)^2},$$

and $f'(\lambda) = 0$ turns into a high-degree rational equation in λ with no general closed-form solution.

For this reason, in practice $\lambda^\star$ is found by one-dimensional numerical optimization (e.g., golden-section search, Brent's method, or Newton's method using $f'(\lambda)$ from (10)).

$$\boxed{\frac{d}{d\lambda}\,\mathrm{MSE}_{\mathrm{val}}(\lambda) = \frac{2}{n_{\mathrm{val}}}\left[R(I_N - S(\lambda))\boldsymbol{Z}\right]^\top \left[RS(\lambda)K^\top KS(\lambda)\boldsymbol{Z}\right], \quad S(\lambda) = (I_N + \lambda K^\top K)^{-1}.}$$