

Table 2: Mean together with first and ninth deciles (within parentheses) of explained variance, R^2 ; bandwidth selection time in milliseconds, t ; and selected bandwidth, σ , for the U.K. rainfall data. The Jacobian method performs significantly better and significantly faster than the competing methods. While unseeded MML selects a too large bandwidth, hampering performance, Jacobian seeded MML performs identically to the Jacobian method, but at a much higher computational time. The p-value of 0.00024 corresponds to the Jacobian method performing better in all 12 experiments.

Method	R^2	t [ms]	σ
Jacobian	0.981 (0.967, 0.994)	0.108 (0.100, 0.119)	0.214 (0.209, 0.219)
GCV	0.969 (0.953, 0.988) $p_{\text{wil}} = 0.00024$	25400 (20800, 28900) $p_{\text{wil}} = 0.00024$	0.296 (0.131, 0.352)
MML	0.554 (0.408, 0.678) $p_{\text{wil}} = 0.00024$	29100 (24300, 33700) $p_{\text{wil}} = 0.00024$	6.56 (6.47, 6.65)
Seeded MML	0.981 (0.967, 0.994) $p_{\text{wil}} = 1$	210000 (176000, 240000) $p_{\text{wil}} = 0.00024$	0.214 (0.209, 0.219)
Silverman	0.971 (0.957, 0.990) $p_{\text{wil}} = 0.00024$	0.168 (0.16, 0.18) $p_{\text{wil}} = 0.00024$	0.306 (0.304, 0.309)

A ADDITIONAL EXPERIMENTS

In this section, we provide additional experiments to those of the main manuscript.

In Figure 7, we compare the approximate and true Jacobian norms as a function of the bandwidth on the temperature and synthetic data sets. The approximate norm captures the structure of the true norm quite well. In the absence of regularization, the minima of the two functions approximately agree. When regularization is added the selected bandwidth, σ_0 , is close to the elbow of both the approximate and true norms.

In Figures 8, we vary the regularization strength, λ for fixed sample size. For the synthetic data, 1000 test observations were generated, while the real data was randomly split into training and testing data. For each data set, n was chosen to a value where the different methods performed approximately equally well in the experiments with varying sample size (Figure 6 in the main manuscript). Thus the splits were 25/15, 100/148, and 50/1000 for the 2D temperature, 1D temperature, and synthetic data, respectively. In all cases 1000 random splits were used to estimate the variance of R^2 , i.e. the proportion of the variation in the data that is explained by $\hat{f}(x^*)$, the selected bandwidth σ , and bandwidth selection computation time in milliseconds, t . The experiments were run on a cluster with Intel Xeon Gold 6130 CPUs. It is again confirmed that the Jacobian method, in addition to being much faster than GCV and MML, is much more stable in terms of bandwidth selection. For the Cauchy distributed data, the median version of the Jacobian method was used; this method requires slightly more time than the standard Jacobian method.

We note that when n is large compared to λ for the 1D temperature data, and when λ is large for the Cauchy distributed data, MML performs very badly, which can be attributed to a local minimum of the likelihood function. This leads us to introduce Jacobian seeded MML, where we instead of Brent’s method, which does not use a seed, use the Nelder-Mead method (Nelder & Mead, 1965) and seed it with the Jacobian bandwidth. In Figure 9, we revisit the jackknife resampling of the 1D temperature data, this time including the Jacobian seeded MML method, which performs much better than standard MML. In Table 2 we extend Table 1 in the main manuscript with Jacobian seeded MML. Using the Jacobian seed, MML performs identically to the Jacobian method in terms of R^2 and σ , albeit more than a million times slower.

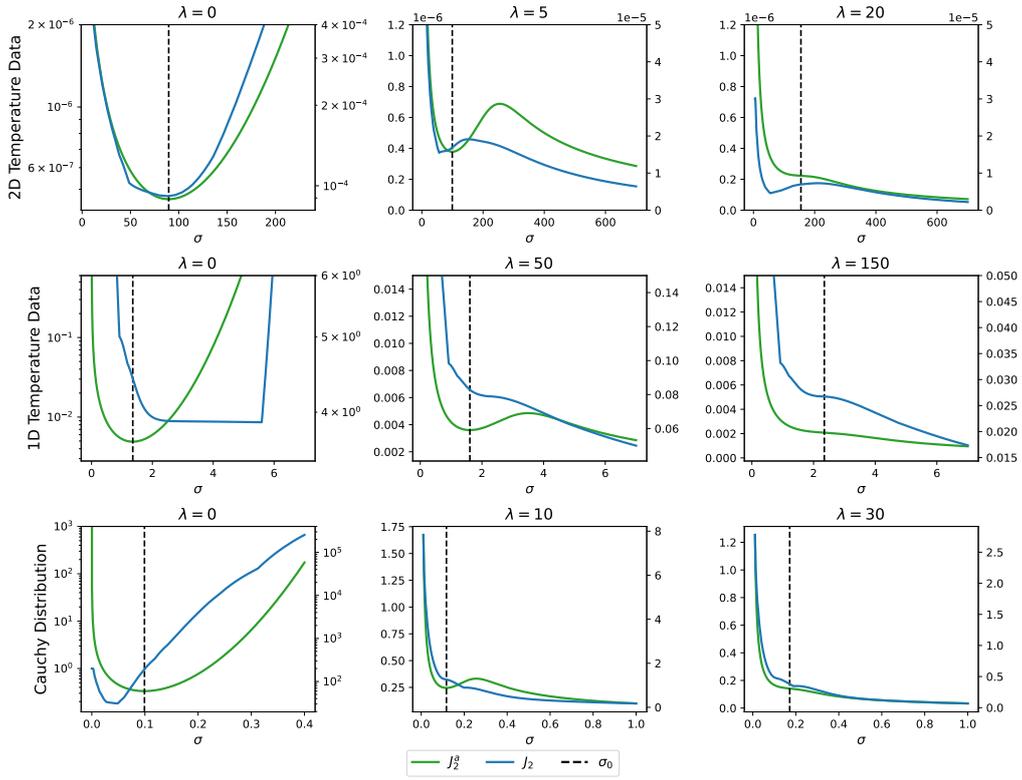


Figure 7: Comparison of the approximate (green, left y-axis) and true (blue, right y-axis) Jacobian norms as a function of bandwidth and regularization. The approximate Jacobian norm captures the structure of the true Jacobian norm quite well, especially for the 2D temperature data, where for $\lambda = 0$, the minima of the two functions agree very well. For $\lambda > 0$, the selected bandwidth, σ_0 , is close to the elbow of both the approximate and true norms. In the rightmost panel, $\lambda > 2ne^{-3/2}$, which means that the approximate Jacobian norm has no local minimum and σ_0 is selected as if $\lambda = 2ne^{-3/2}$.

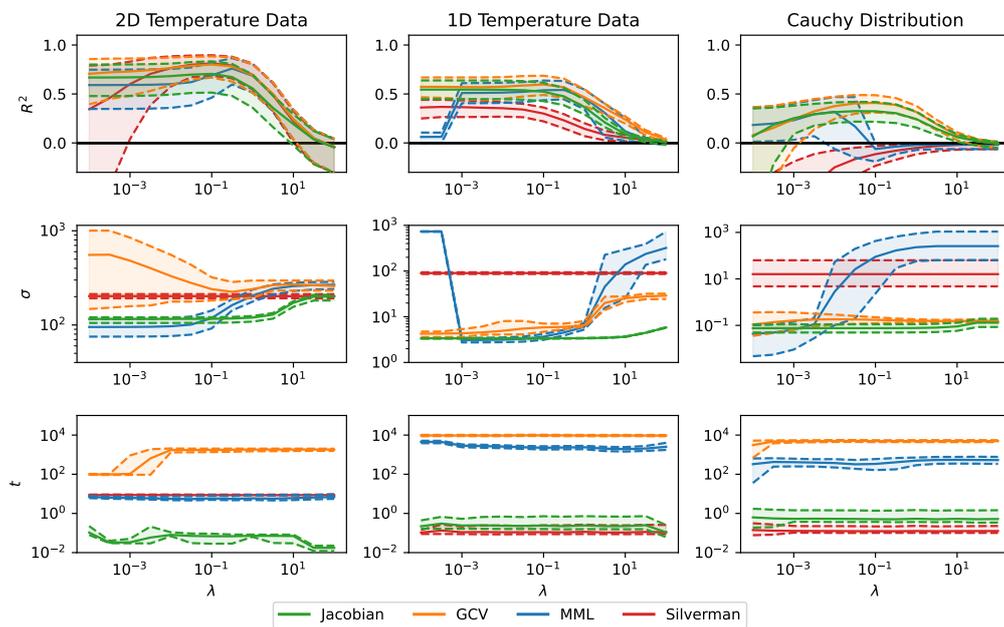


Figure 8: Mean together with first and ninth deciles for explained variance, R^2 ; selected bandwidth, σ ; and computation time in milliseconds t , for different regularization strengths, using the four bandwidth selection methods. The Jacobian and Silverman’s methods are several orders of magnitude quicker than the two other methods. They are also much more stable in terms of bandwidth selection. In terms of prediction, the Jacobian method generally performs better than, or on par with, the competing methods. For both the 1D temperature data and the Cauchy data, MML gets stuck in a local minimum. For the Cauchy data, the, slightly slower, median version of the Jacobian method was used.

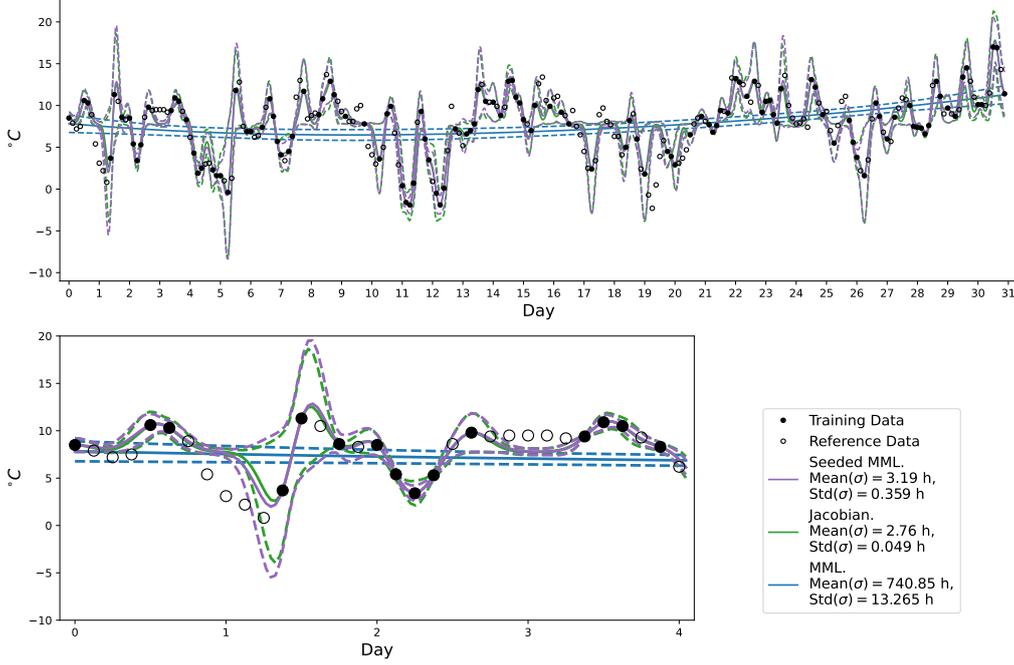


Figure 9: Means and standard deviations of KRR predictions from jackknife resampling on the 1D temperature data. The lower bottom plot shows a zoom-in on the first 4 days. Jacobian seeded MML performs much better than standard MML, and similarly to the Jacobian method.

B PROOFS

In this section, we provide the proofs of the propositions in the main manuscript.

Proof of Proposition 1.

Denote $d := \frac{2l_{\max}}{((n-1)^{1/p}-1)\pi}$. Then

$$J_2^a(\sigma, l, n, p, \lambda) = J_2^a(\sigma, n, d, \lambda) = \frac{1}{\sigma \left(n \exp\left(-\left(\frac{\sigma}{d}\right)^2\right) + \lambda \right)},$$

from which we obtain

$$\lim_{\sigma \rightarrow 0^+} J_2^a(\sigma) = +\infty$$

and

$$\lim_{\sigma \rightarrow +\infty} J_2^a(\sigma) = \begin{cases} +\infty & \text{if } \lambda = 0 \\ 0 & \text{if } \lambda > 0. \end{cases}$$

We now identify stationary points by setting the derivative to 0.

$$\begin{aligned} \frac{\partial J_2^a(\sigma)}{\partial \sigma} &= -\frac{\exp\left(\left(\frac{\sigma}{d}\right)^2 - \frac{1}{2}\right) \left(n + \lambda \exp\left(\left(\frac{\sigma}{d}\right)^2\right) - 2n \left(\frac{\sigma}{d}\right)^2 \right)}{\left(\sigma \left(n + \lambda \exp\left(\left(\frac{\sigma}{d}\right)^2\right) \right) \right)^2}. \\ n + \lambda e^{\left(\frac{\sigma}{d}\right)^2} - 2n \left(\frac{\sigma}{d}\right)^2 = 0 &\iff -\frac{\lambda\sqrt{e}}{2n} = e^{\frac{1}{2}-\left(\frac{\sigma}{d}\right)^2} \left(\frac{1}{2} - \left(\frac{\sigma}{d}\right)^2 \right) \\ \iff \left(\frac{1}{2} - \left(\frac{\sigma}{d}\right)^2 \right) &= W\left(-\frac{\lambda\sqrt{e}}{2n}\right) \implies \sigma = \frac{d}{\sqrt{2}} \sqrt{1 - 2W\left(-\frac{\lambda\sqrt{e}}{2n}\right)}, \end{aligned}$$

where W denotes the Lambert W function. Since this function has real outputs only if its argument is greater than $-e^{-1}$, in order to obtain stationary points we need

$$-\frac{\lambda\sqrt{e}}{2n} \geq -e^{-1} \iff \lambda \leq 2ne^{-3/2}$$

which gives us the two stationary points

$$\sigma_0 = \frac{\sqrt{2}}{\pi} \frac{l_{\max}}{(n-1)^{1/p} - 1} \sqrt{1 - 2W_0\left(-\frac{\lambda\sqrt{e}}{2n}\right)}$$

and

$$\sigma_{-1} = \frac{\sqrt{2}}{\pi} \frac{l_{\max}}{(n-1)^{1/p} - 1} \sqrt{1 - 2W_{-1}\left(-\frac{\lambda\sqrt{e}}{2n}\right)}.$$

$W_{-1}(x) < W_0(x)$ for $x \in (-e^{-1}, 0)$, which implies that $\sigma_0 < \sigma_{-1}$. Combined with the limits above, this implies that, when existing, σ_0 is a local minimum and σ_{-1} is a local maximum.

Finally, for $\lambda = 0$, $W_0(0) = 0$ and $\lim_{\lambda \rightarrow 0} W_{-1}\left(-\frac{\lambda\sqrt{e}}{2n}\right) = -\infty$, which means that in the absence of λ , $\sigma_0 = \frac{\sqrt{2}}{\pi} \frac{l_{\max}}{(n-1)^{1/p} - 1}$ and $\sigma_{-1} = +\infty$. \square

Proof of Proposition 2.

We first note that for $\mathbf{d}_i = \mathbf{x}^* - \mathbf{x}_i$, $\frac{\partial \hat{f}(\mathbf{x}^*)}{\partial \mathbf{d}_i} = \frac{\partial \hat{f}(\mathbf{x}^*)}{\partial \mathbf{x}^*}$:

$$\begin{aligned} \frac{\partial \mathbf{d}_i}{\partial \mathbf{x}^*} &= \frac{\partial (\mathbf{x}^* - \mathbf{x}_i)}{\partial \mathbf{x}^*} = \frac{\partial \mathbf{x}^*}{\partial \mathbf{x}^*} - \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}^*} = \mathbf{I} - \mathbf{0} = \mathbf{I} \\ \frac{\partial \hat{f}(\mathbf{x}^*)}{\partial \mathbf{x}^*} &= \frac{\partial \hat{f}(\mathbf{x}^*)}{\partial \mathbf{d}_i} \cdot \frac{\partial \mathbf{d}_i}{\partial \mathbf{x}^*} = \frac{\partial \hat{f}(\mathbf{x}^*)}{\partial \mathbf{d}_i} \cdot \mathbf{I} = \frac{\partial \hat{f}(\mathbf{x}^*)}{\partial \mathbf{d}_i}. \end{aligned}$$

Now,

$$\begin{aligned} \left\| \frac{\partial \hat{f}(\mathbf{x}^*)}{\partial \mathbf{d}_i} \right\|_2 &= \left\| \frac{\partial \hat{f}(\mathbf{x}^*)}{\partial \mathbf{x}^*} \right\|_2 = \left\| \frac{\partial}{\partial \mathbf{x}^*} \left(\mathbf{k}(\mathbf{x}^*, \mathbf{X})^\top \cdot (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I})^{-1} \cdot \mathbf{y} \right) \right\|_2 \\ &= \left\| \frac{\partial \mathbf{k}(\mathbf{x}^*, \mathbf{X})^\top}{\partial \mathbf{x}^*} \cdot (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I})^{-1} \mathbf{y} \right\|_2 \\ &\leq \left\| \frac{\partial \mathbf{k}(\mathbf{x}^*, \mathbf{X})^\top}{\partial \mathbf{x}^*} \right\|_2 \cdot \left\| (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I})^{-1} \right\|_2 \cdot \|\mathbf{y}\|_2 \\ &\leq \sqrt{n} \cdot \left\| \frac{\partial \mathbf{k}(\mathbf{x}^*, \mathbf{X})^\top}{\partial \mathbf{x}^*} \right\|_1 \cdot \left\| (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I})^{-1} \right\|_2 \cdot \|\mathbf{y}\|_2 \\ &= \sqrt{n} \cdot \left\| \left[\frac{\partial k(\mathbf{x}^*, \mathbf{x}_1)}{\partial \mathbf{x}^*} \quad \frac{\partial k(\mathbf{x}^*, \mathbf{x}_2)}{\partial \mathbf{x}^*} \quad \dots \quad \frac{\partial k(\mathbf{x}^*, \mathbf{x}_n)}{\partial \mathbf{x}^*} \right] \right\|_1 \cdot \left\| (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I})^{-1} \right\|_2 \cdot \|\mathbf{y}\|_2 \\ &= \sqrt{n} \cdot \max_{\mathbf{x}_i \in \mathbf{X}} \left\| \frac{\partial k(\mathbf{x}^*, \mathbf{x}_i)}{\partial \mathbf{x}^*} \right\|_1 \cdot \left\| (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I})^{-1} \right\|_2 \cdot \|\mathbf{y}\|_2 \\ &\stackrel{(a)}{=} \sqrt{n} \cdot \max_{\mathbf{x}_i \in \mathbf{X}} \left\| \frac{\partial k(\mathbf{x}_i + \mathbf{d}_i, \mathbf{x}_i)}{\partial \mathbf{d}_i} \right\|_1 \cdot \left\| (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I})^{-1} \right\|_2 \cdot \|\mathbf{y}\|_2, \end{aligned}$$

where in (a), we used the chain rule together with $\frac{\partial \mathbf{d}_i}{\partial \mathbf{x}^*} = \mathbf{I}$. \square

Proof of Proposition 3.

In spherical coordinates,

$$\left\| \frac{\partial k_G(\mathbf{d}_i, \sigma)}{\partial \mathbf{d}_i} \right\|_1 = \left| \frac{\partial k_G(\mathbf{d}_i, \sigma)}{\partial d_i} \right| + \sum_{j=2}^p \left| \frac{1}{d_i} \frac{\partial k_G(\mathbf{d}_i, \sigma)}{\partial \theta_j} \right|,$$

where the sum is over the angular coordinates. Since the Gaussian kernel is rotationally invariant, this sum is 0 and

$$\left\| \frac{\partial k_G(\mathbf{d}_i, \sigma)}{\partial \mathbf{d}_i} \right\|_1 = \left| \frac{\partial}{\partial d_i} \exp\left(-\frac{d_i^2}{2\sigma^2}\right) \right| = \frac{d_i}{\sigma^2} \exp\left(-\frac{d_i^2}{2\sigma^2}\right).$$

To find the d_i that maximizes the derivative, we look where the second derivative is zero.

$$\frac{\partial}{\partial d_i} \left| \frac{\partial k_G(d_i, \sigma)}{\partial d_i} \right| = \left(\left(\frac{d_i}{\sigma^2} \right)^2 - \frac{1}{\sigma^2} \right) \exp\left(-\frac{d_i^2}{2\sigma^2}\right).$$

Setting the second derivative to zero amounts to

$$\left(\frac{d_i}{\sigma^2} \right)^2 = \frac{1}{\sigma^2} \iff d_i^2 = \sigma^2 \implies d_i = \sigma.$$

Plugging this into the first derivative we obtain $\frac{1}{\sigma} \exp\left(-\frac{1}{2}\right)$, which is greater than

$$\left| \frac{\partial k_G(0, \sigma)}{\partial d_i} \right| = \left| \frac{\partial k_G(\infty, \sigma)}{\partial d_i} \right| = 0,$$

and consequently

$$\max_{\mathbf{d}_i} \left\| \frac{\partial k_G(\mathbf{d}_i, \sigma)}{\partial \mathbf{d}_i} \right\|_1 = \max_{d_i} \left| \frac{\partial k_G(d_i, \sigma)}{\partial d_i} \right| = \frac{1}{\sigma\sqrt{e}}.$$

□

Proof of Proposition 4.

To alleviate notation, from now on we do not explicitly state that \mathbf{K} depends on \mathbf{X} . We first note that $\left\| (\mathbf{K} + \lambda \mathbf{I})^{-1} \right\|_2 = \frac{1}{s_n(\mathbf{K} + \lambda \mathbf{I})}$, where s_n denotes the smallest singular value of \mathbf{K} . Since \mathbf{K} is symmetric and positive semi-definite, it is diagonalizable as $\mathbf{K} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{U}^\top$, while $\lambda \mathbf{I} = \lambda \mathbf{U} \mathbf{U}^\top$, which means that $\mathbf{K} + \lambda \mathbf{I} = \mathbf{U} (\boldsymbol{\Sigma} + \lambda \mathbf{I}) \mathbf{U}^\top$, i.e. the singular values of $\mathbf{K} + \lambda \mathbf{I}$ are the singular values of \mathbf{K} , shifted by λ .

According to Bermanis et al. (2013), for $\mathbf{x} \in \mathbb{R}^p$, where each x_i is restricted to an interval of length l_i , $i = 1, 2, \dots, p$, for a Gaussian kernel matrix $\mathbf{K} \in \mathbb{R}^{m \times n}$, with singular values s_1, \dots, s_n , the number of singular values larger than $\delta \cdot s_1$ for some $\delta > 0$, $R_\delta(\mathbf{K})$, is bounded according to

$$R_\delta(\mathbf{K}) := \#\left\{j : \frac{s_j(\mathbf{K})}{s_1(\mathbf{K})} \geq \delta\right\} \leq \prod_{i=1}^d \left(\frac{2 l_i}{\pi \sigma} \sqrt{\log(1/\delta)} + 1 \right) \leq \left(\frac{2 l_{\max}}{\pi \sigma} \sqrt{\log(1/\delta)} + 1 \right)^p.$$

Solving for δ , we obtain

$$\begin{aligned} R_\delta(\mathbf{K}) \leq \left(\frac{2 l_{\max}}{\pi \sigma} \sqrt{\log(1/\delta)} + 1 \right)^p &\iff (R_\delta(\mathbf{K})^{1/p} - 1) \frac{\pi \sigma}{2 l_{\max}} \leq \sqrt{\log(1/\delta)} \\ \iff \delta \leq \exp\left(-\left(\frac{(R_\delta(\mathbf{K})^{1/p} - 1)\pi \sigma}{2 l_{\max}}\right)^2\right). \end{aligned}$$

Now, if $R_\delta(\mathbf{K}) = n$, then all singular values (including s_n) are larger than or equal to $\delta \cdot s_1$. If $R_\delta(\mathbf{K}) = n - 1$, then all but one (namely s_n) of the singular values are larger than or equal to $\delta \cdot s_1$. So for $R_\delta(\mathbf{K}) = n - 1$, $s_n < \delta \cdot s_1$, which implies

$$s_n < s_1 \delta \leq s_1 \exp\left(-\left(\frac{((n-1)^{1/p} - 1)\pi \sigma}{2 l_{\max}}\right)^2\right) \leq n \exp\left(-\left(\frac{((n-1)^{1/p} - 1)\pi \sigma}{2 l_{\max}}\right)^2\right),$$

where we used $\sigma_1(\mathbf{K}) \leq n \cdot \|\mathbf{K}\|_{\max} = n \cdot 1$.

Thus

$$\left\| (\mathbf{K} + \lambda \mathbf{I})^{-1} \right\|_2 = \frac{1}{s_n + \lambda} \geq \frac{1}{n \exp\left(-\left(\frac{((n-1)^{1/p} - 1)\pi \sigma}{2 l_{\max}}\right)^2\right) + \lambda}.$$

□