Structured nonlinear variable selection - supplement

Magda Gregorová **Alexandros Kalousis**
Geneva School of Business Administration, HES-SO, Switzerland
University of Geneva, Switzerland

Stéphane Marchand-Maillet
University of Geneva
Switzerland

1 CODE AND REPLICATION FILES

The implementation of our NVSD algorithm and the replication files for the experiments presented in the main text of our paper are available publicly at the Bitbucket repository `https://bitbucket.org/dmmlgeneva/nvsd_uai2018/`.

2 PROOFS OF PROPOSITIONS FROM THE MAIN TEXT

Proof of Proposition 1. We may decompose any function $f \in \mathcal{F}$ as $f = f_{\|} + f_{\perp}$, where $f_{\|}$ lies in the span of the kernel sections $k_{\mathbf{x}^i}$ and its partial derivatives $[\partial_a k_{\mathbf{x}^i}]$ centred at the n training points, and $f_{\perp}$ lies in its orthogonal complement.

The 1st term $\widehat{\mathcal{L}}(f)$ depends on the function f only through its evaluations at the training points $f(\mathbf{x}^i), i \in \mathbb{N}_n$. For each training point $\mathbf{x}^i$ we have

$$f(\mathbf{x}^i) = \langle f, k_{\mathbf{x}^i} \rangle_{\mathcal{F}} = \langle f_{\|} + f_{\perp}, k_{\mathbf{x}^i} \rangle_{\mathcal{F}} = \langle f_{\|}, k_{\mathbf{x}^i} \rangle_{\mathcal{F}} \quad ,$$

where the last equality is the result of the orthogonality of the complement $\langle f_{\perp}, k_{\mathbf{x}^i} \rangle_{\mathcal{F}} = 0$. By this the term $\widehat{\mathcal{L}}(f)$ is independent of $f_{\perp}$.

The 2nd term $\widehat{\mathcal{R}}(f)$ depends on the function f only through the evaluations of its partial derivatives at the training points $\partial_a f(\mathbf{x}^i), i \in \mathbb{N}_i, a \in \mathbb{N}_d$. For each training point $\mathbf{x}^i$ and dimension a we have

$$\partial_a f(\mathbf{x}^i) = \langle f, [\partial_a k_{\mathbf{x}^i}] \rangle_{\mathcal{F}} = \langle f_{\|}, [\partial_a k_{\mathbf{x}^i}] \rangle_{\mathcal{F}} \quad ,$$

by the orthogonality of the complement $\langle f_{\perp}, [\partial_a k_{\mathbf{x}^i}] \rangle_{\mathcal{F}} = 0$. By this the term $\widehat{\mathcal{R}}(f)$ is independent of $f_{\perp}$ for the empirical versions of all three considered regularizers $\mathcal{R}^L, \mathcal{R}^{GL}, \mathcal{R}^{EN}$. For the 3rd term we have $||f||^2_{\mathcal{F}} = ||f_{\|} + f_{\perp}||^2_{\mathcal{F}} = ||f_{\|}||^2_{\mathcal{F}} + ||f_{\perp}||^2_{\mathcal{F}}$ because $\langle f_{\|}, f_{\perp} \rangle_{\mathcal{F}} = 0$. Trivially, this is minimised when $f_{\perp} = 0$. $\square$

Proof of Proposition 2. Using the matrices and vector introduced in section 4.1 and proposition 1 we have

$$f(\mathbf{x}^i) = \sum_{j=1}^{n} \alpha_j K_{ji} + \sum_{j=1}^{n} \sum_{a=1}^{d} \beta_{aj} \tilde{D}^a_{ij}$$

$$\partial_a f(\mathbf{x}^i) = \sum_{j=1}^{n} \alpha_j \tilde{D}^a_{ij} + \sum_{j=1}^{n} \sum_{c=1}^{d} \beta_{cj} L^{ca}_{ji}$$

For the 1st term $\widehat{\mathcal{L}}(f)$ we have

$$\begin{aligned}
\widehat{\mathcal{L}}(f) &= \sum_{i=1}^{n}\left(y^i - f(\mathbf{x}^i)\right)^2 = \sum_{i=1}^{n}\left(y^i - \sum_{j=1}^{n}\alpha_j K_{ji} - \sum_{j=1}^{n}\sum_{a=1}^{d}\beta_{aj}\tilde{D}^a_{ij}\right)^2 \\
&= \sum_{i=1}^{n}\Big((y^i)^2 - 2y^i\sum_{j=1}^{n}\alpha_j K_{ji} - 2y^i\sum_{j=1}^{n}\sum_{a=1}^{d}\beta_{aj}\tilde{D}^a_{ij} + \sum_{j,l}^{n}\alpha_j\alpha_l K_{ji}K_{l,i} + 2\sum_{j,l}^{n}\sum_{a=1}^{d}\beta_{aj}\alpha_l\tilde{D}^a_{ij}K_{l,i} \\
&+ \sum_{j,l}^{n}\sum_{a,b}^{d}\beta_{aj}\beta_{bl}\tilde{D}^a_{ij}\tilde{D}^b_{i,l}\Big) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{K}\mathbf{a} - 2\sum_{a}^{d}\mathbf{y}^T\tilde{\mathbf{D}}^a\mathbf{B}^T_{a,:} + \boldsymbol{\alpha}^T\mathbf{K}\mathbf{K}\boldsymbol{\alpha} + 2\sum_{a}^{d}\boldsymbol{\alpha}^T\mathbf{K}\tilde{\mathbf{D}}^a\mathbf{B}^T_{a,:} + \sum_{a,b}^{d}\mathbf{B}_{a,:}\mathbf{D}^a\tilde{\mathbf{D}}^b\mathbf{B}^T_{b,:} \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{K}\mathbf{a} - 2\mathbf{y}^T\mathbf{D}^T\boldsymbol{\beta} + \boldsymbol{\alpha}^T\mathbf{K}\mathbf{K}\boldsymbol{\alpha} + 2\boldsymbol{\alpha}^T\mathbf{K}\mathbf{D}^T\boldsymbol{\beta} + \sum_{a,b}^{d}\boldsymbol{\beta}^T\mathbf{D}\mathbf{D}^T\boldsymbol{\beta} \\
&= ||\mathbf{y} - \mathbf{K}\boldsymbol{\alpha} - \mathbf{D}^T\boldsymbol{\beta}||_2^2 \ ,
\end{aligned}$$

where $\mathbf{B}$ is the $d \times n$ matrix with the β coefficients $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B}^T)$

For the 2nd term we have

$$\begin{aligned}
\widehat{\mathcal{R}}^L(f) &= \sum_{a=1}^{d}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\partial_a f(\mathbf{x}^i)\right)^2} = \sum_{a=1}^{d}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\alpha_j\tilde{D}^a_{ji} + \sum_{j=1}^{n}\sum_{c=1}^{d}\beta_{cj}L^{ca}_{ji}\Big)^2\right]^{0.5} \\
&= \sum_{a=1}^{d}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j,l}^{n}\alpha_j\alpha_l\tilde{D}^a_{ji}\tilde{D}^a_{l,i} + 2\sum_{j,l}^{n}\sum_{c=1}^{d}\alpha_j\beta_{cl}\tilde{D}^a_{ji}L^{ca}_{l,i} + \sum_{j,l}^{n}\sum_{c,r}^{d}\beta_{cj}\beta_{rl}L^{ca}_{ji}L^{ra}_{l,i}\Big)\right]^{0.5} \\
&= \sum_{a=1}^{d}\frac{1}{\sqrt{n}}\left[\boldsymbol{\alpha}^T\tilde{\mathbf{D}}^a\mathbf{D}^a\boldsymbol{\alpha} + 2\sum_{c=1}^{d}\alpha^T\tilde{\mathbf{D}}^a\mathbf{L}^{ac}\mathbf{B}^T_{c:} + \sum_{c,r}^{d}\mathbf{B}_{c:}\mathbf{L}^{ca}\mathbf{L}^{ar}\mathbf{B}^T_{r:}\right]^{0.5} \\
&= \sum_{a=1}^{d}\frac{1}{\sqrt{n}}\left[\boldsymbol{\alpha}^T\tilde{\mathbf{D}}^a\mathbf{D}^a\boldsymbol{\alpha} + 2\alpha^T\tilde{\mathbf{D}}^a\mathbf{L}^a\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{L}^{aT}\mathbf{L}^a\boldsymbol{\beta}\right]^{0.5} = \sum_{a=1}^{d}\frac{1}{\sqrt{n}}||\mathbf{D}^a\boldsymbol{\alpha} + \mathbf{L}^a\boldsymbol{\beta}||_2
\end{aligned}$$

$\widehat{\mathcal{R}}^{GL}(f)$ and $\widehat{\mathcal{R}}^{EN}(f)$ follow in analogy.

For the 3rd term we have

$$
\begin{aligned}
||f||_{\mathcal{F}}^2 &= ||\sum_{j=1}^{n} \alpha_j k_{\mathbf{x}^j} + \sum_{j=1}^{n}\sum_{a=1}^{d} \beta_{aj} [\partial_a k_{\mathbf{x}^j}] ||_{\mathcal{F}}^2 \\
&= \langle \sum_{j=1}^{n} \alpha_j k_{\mathbf{x}^j}, \sum_{i=1}^{n} \alpha_i k_{\mathbf{x}^i} \rangle_{\mathcal{F}} + 2 \langle \sum_{j=1}^{n} \alpha_j k_{\mathbf{x}^j}, \sum_{i=1}^{n}\sum_{a=1}^{d} \beta_{ai} [\partial_a k_{\mathbf{x}^i}] \rangle_{\mathcal{F}} \\
&+ \langle \sum_{j=1}^{n}\sum_{a=1}^{d} \beta_{aj} [\partial_a k_{\mathbf{x}^j}], \sum_{i=1}^{n}\sum_{c=1}^{d} \beta_{ci} [\partial_c k_{\mathbf{x}^i}] \rangle_{\mathcal{F}} \\
&= \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + 2 \sum_{ij}^{n} \sum_{a}^{d} \alpha_j \beta_{ai} \, \partial_a k_{\mathbf{x}^j} (\mathbf{x}^i) + \sum_{ij}^{n} \sum_{ac}^{d} \beta_{aj} \beta_{ci} \frac{\partial^2}{\partial x_a^j \partial x_c^i} k(\mathbf{x}^j, \mathbf{x}^i) \\
&= \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + 2 \sum_{ij}^{n} \sum_{a}^{d} \alpha_j \beta_{ai} \tilde{D}_{ji}^a + \sum_{ij}^{n} \sum_{ac}^{d} \beta_{aj} \beta_{ci} L_{ji}^{ac} \\
&= \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + 2 \sum_{a}^{d} \boldsymbol{\alpha}^T \tilde{\mathbf{D}}^a \mathbf{B}_{a:}^T + \sum_{ac}^{d} \mathbf{B}_{:j} \mathbf{L}^{ac} \mathbf{B}_{c:}^T \\
&= \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + 2 \boldsymbol{\alpha}^T \mathbf{D}^T \boldsymbol{\beta} + \sum_{a}^{d} \mathbf{B}_{a:} \mathbf{L}^a \boldsymbol{\beta} \\
&= \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + 2 \boldsymbol{\alpha}^T \mathbf{D}^T \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{L} \boldsymbol{\beta}
\end{aligned}
$$

□

Proof of Proposition 4. The proximal problem in step $S2$ for $\mathcal{R}^L$ for a single partition $\boldsymbol{\varphi}_a$ is

$$
\mathcal{R}^L : \, \boldsymbol{\varphi}_a^{(k+1)} = \underset{\boldsymbol{\varphi}_a}{\operatorname{argmin}} \, \frac{\tau}{\sqrt{n}} ||\boldsymbol{\varphi}_a||_2 + \frac{\rho}{2} ||\mathbf{Z}^a \boldsymbol{\omega}^{(k+1)} - \boldsymbol{\varphi}_a + \boldsymbol{\lambda}_a^{(k)}||_2^2
$$

This convex problem is non-differentiable at the point $\boldsymbol{\varphi} = \mathbf{0}$. It is, however, sub-differentiable with the optimality condition for the minimizing $\boldsymbol{\varphi}^*$

$$
\mathbf{0} \in \partial \, \frac{\tau}{\sqrt{n}} ||\boldsymbol{\varphi}_a^*||_2 - \rho \left(\mathbf{Z}^a \boldsymbol{\omega}^{(k+1)} - \boldsymbol{\varphi}_a + \boldsymbol{\lambda}_a^{(k)} \right) \ ,
$$

where for any function $f : \mathbb{R}^d \to \mathbb{R}$, $\partial f(\mathbf{x}) \subset \mathbb{R}^d$ is the sub-differential of f at x defined as

$$
\partial f(\mathbf{x}) = \{ \mathbf{g} \, | \, f(\mathbf{z}) \geq f(\mathbf{x}) + \mathbf{g}^T (\mathbf{z} - \mathbf{x}) \} \ .
$$

For notational simplicity, in what follows we introduce the variable $\mathbf{v} = \mathbf{Z}^a \boldsymbol{\omega}^{(k+1)} + \boldsymbol{\lambda}_a^{(k)}$, and we drop the sub-/superscripts of the partitions a and the iterations k.

Part A For all points other than $\boldsymbol{\varphi}^* = \mathbf{0}$ the optimality condition reduces to

$$
\mathbf{0} = \frac{\tau}{\sqrt{n}} \frac{\boldsymbol{\varphi}^*}{||\boldsymbol{\varphi}^*||_2} - \rho \left(\mathbf{v} - \boldsymbol{\varphi}^* \right) \ ,
$$

From which we get

$$
\begin{aligned}
\left(\frac{\tau}{\rho \sqrt{n} ||\boldsymbol{\varphi}^*||_2} + 1 \right) \boldsymbol{\varphi}^* &= \mathbf{v} \\
\left(\frac{\tau}{\rho \sqrt{n} ||\boldsymbol{\varphi}^*||_2} + 1 \right) ||\boldsymbol{\varphi}^*||_2 &= ||\mathbf{v}||_2 \\
||\boldsymbol{\varphi}^*||_2 &= ||\mathbf{v}||_2 - \frac{\tau}{\rho \sqrt{n}} \ .
\end{aligned}
$$

We use this result in the optimality condition

$$\begin{aligned}
\mathbf{0} &= \frac{\tau}{\sqrt{n}} \frac{\boldsymbol{\varphi}^*}{||\mathbf{v}||_2 - \frac{\tau}{\rho\sqrt{n}}} - \rho\,(\mathbf{v} - \boldsymbol{\varphi}^*) \\
\frac{\tau}{\sqrt{n}}\,\boldsymbol{\varphi}^* &= \rho\,(\mathbf{v} - \boldsymbol{\varphi}^*)(||\mathbf{v}||_2 - \frac{\tau}{\rho\sqrt{n}}) \\
\frac{\tau}{\sqrt{n}}\,\boldsymbol{\varphi}^* &= (\rho||\mathbf{v}||_2 - \frac{\tau}{\sqrt{n}})\mathbf{v} - \rho\,||\mathbf{v}||_2\,\boldsymbol{\varphi}^* + \frac{\tau}{\sqrt{n}}\,\boldsymbol{\varphi}^* \\
\boldsymbol{\varphi}^* &= \left(1 - \frac{\tau}{\rho\sqrt{n}||\mathbf{v}||_2}\right)\mathbf{v}
\end{aligned}$$

Part B For the point $\boldsymbol{\varphi}^* = \mathbf{0}$ we have $\partial||\boldsymbol{\varphi}^*||_2 = \{\mathbf{g} \,|\, ||\mathbf{g}||_2 \leq 1\}$ (from the definition of sub-differential and the Cauchy-Schwarz inequality).

From the optimality condition

$$\begin{aligned}
\mathbf{0} &= \frac{\tau}{\sqrt{n}}\,\mathbf{g} - \rho\,\mathbf{v} \qquad\qquad (\boldsymbol{\varphi}^* = \mathbf{0}) \\
\rho\,\mathbf{v} &= \frac{\tau}{\sqrt{n}}\,\mathbf{g} \\
\rho\,||\mathbf{v}||_2 &= \frac{\tau}{\sqrt{n}}\,||\mathbf{g}||_2 \\
||\mathbf{v}||_2 &\leq \frac{\tau}{\rho\sqrt{n}} \qquad\qquad (||\mathbf{g}||_2 \leq 1)
\end{aligned}$$

Putting the results from part A and B together we obtain the final result

$$\boldsymbol{\varphi}^* = \left(1 - \frac{\tau}{\rho\sqrt{n}||\mathbf{v}||_2}\right)_+ \mathbf{v}$$

The proofs for $\mathcal{R}^{GL}$ and $\mathcal{R}^{EN}$ follow similarly. $\square$

3 Examples of kernel partial derivatives

We list here the 1st and 2nd order partial derivatives which form the elements of the derivative matrices $\mathbf{D}$ and $\mathbf{L}$ introduced in section 4.1 for some common kernel functions k.

Linear kernel
Kernel gram matrix

$$K_{i,j} = k(\mathbf{x}^i, \mathbf{x}^j) = \langle \mathbf{x}^i, \mathbf{x}^j \rangle$$

1st order partial-derivative matrix

$$D^a_{i,j} = \frac{\partial k(\mathbf{s}, \mathbf{x}^j)}{\partial s_a}|_{\mathbf{s}=\mathbf{x}^i} = x^j_a$$

2nd order partial-derivative matrix

$$L^{ab}_{i,j} = \frac{\partial^2 k(\mathbf{s}, \mathbf{r})}{\partial s_a \partial r_b}|_{\substack{\mathbf{s}=\mathbf{x}^i \\ \mathbf{r}=\mathbf{x}^j}} = \begin{cases} 0 & \text{if } a \neq b \\ 1 & \text{if } a = b \end{cases}$$

Polynomial of order $p > 1$
Kernel gram matrix

$$K_{i,j} = (\langle \mathbf{x}^i, \mathbf{x}^j \rangle + c)^p$$

1st order partial-derivative matrix

$$D^a_{i,j} = p\,(\langle \mathbf{x}^i, \mathbf{x}^j \rangle + c)^{p-1}\; x^j_a$$

2nd order partial-derivative matrix

$$L^{ab}_{i,j} = \begin{cases} p(p-1)\,(\langle \mathbf{x}^i, \mathbf{x}^j \rangle + c)^{p-2}\; x^i_b x^j_a & \text{if } a \neq b \\ p(p-1)\,(\langle \mathbf{x}^i, \mathbf{x}^j \rangle + c)^{p-2}\; x^i_a x^j_a + p\,(\langle \mathbf{x}^i, \mathbf{x}^j \rangle + c)^{p-1} & \\ & \text{if } a = b \end{cases}$$

Gaussian kernel
Kernel gram matrix

$$K_{i,j} = \exp\left(-\frac{||\mathbf{x}^i - \mathbf{x}^j||_2^2}{2\sigma^2}\right)$$

1st order partial-derivative matrix

$$D^a_{i,j} = \exp\left(-\frac{||\mathbf{x}^i - \mathbf{x}^j||_2^2}{2\sigma^2}\right)\frac{x^j_a - x^i_a}{\sigma^2}$$

2nd order partial-derivative matrix

$$L^{ab}_{i,j} = \begin{cases} \exp\left(-\frac{||\mathbf{x}^i - \mathbf{x}^j||_2^2}{2\sigma^2}\right)\frac{(x^j_a - x^i_a)(x^i_b - x^j_b)}{\sigma^4} & \text{if } a \neq b \\ \exp\left(-\frac{||\mathbf{x}^i - \mathbf{x}^j||_2^2}{2\sigma^2}\right)\frac{(x^i_a - x^j_a)^2 - \sigma^2}{-\sigma^4} & \text{if } a = b \end{cases}$$