Supplementary Material for “Kernels over Sets of Finite Sets using RKHS Embeddings, with Application to Bayesian (Combinatorial) Optimization”

Unless indicated otherwise, referenced equation labels are to be understood with respect to equations of the main article.

A Elements of literature review

Before reviewing some foundational machine learning papers dealing with kernels on sets of (sub)sets and related objects, let us start by some preliminary remarks on how an elementary class of positive definite kernels can be constructed in the context of measure spaces and why these kernels are not necessarily ideal for the prediction and optimization objectives we have in mind. Consider here a set $\mathcal{X}$ equipped with a sigma-algebra $\mathcal{A}$ and a measure μ, making it a measure space $(\mathcal{X}, \mathcal{A}, \mu)$. Then it comes without much effort that the mapping k defined by

$$k : (S, S') \in \mathcal{A}^2 \to \mu(S \cap S') \in [0, \infty)$$

constitutes a positive definite kernel. Indeed, taking arbitrary $n \geq 1$, $a_1, \dots, a_n \in \mathbb{R}$, $S_1, \dots, S_n \in \mathcal{A}$ and recalling that $\mu(S \cap S') = \int_{\mathcal{X}} \mathbf{1}_S(\mathbf{u})\mathbf{1}_{S'}(\mathbf{u})\mathrm{d}\mu(\mathbf{u})$, we do have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(S_i, S_j) = \int_{\mathcal{X}} \left(\sum_{i=1}^{n} a_i \mathbf{1}_{S_i}(\mathbf{u})\right)^2 \mathrm{d}\mu(\mathbf{u}) \geq 0$$

In the particular case where $\mathcal{X}$ is finite, $\mathcal{A}$ is the associated power set $\mathcal{P}(\mathcal{X})$, and μ is the counting measure, we find that

$$k(S, S') = \#(S \cap S') = \sum_{\mathbf{x} \in S}\sum_{\mathbf{x}' \in S} \frac{1}{2}\delta_{\mathbf{x},\mathbf{x}'},$$

a kernel that does account for the position of points only to the extent that it counts the number of points simultaneously in both sets (without any account for the closeness of non-coinciding points). Such a kernel is referred to as *default kernel on sets* in (Gärtner et al., 2004, Example 4.2), where it appears as a particular case of an abstract construction denoted *default kernel for basic terms* (Definition 4.1, p. 213) and that is also applied for instance to multisets (Example 4.3 of the same page). For the case of the default kernel on sets, the authors comment following Example 4.2 that *“the intuition here is that using the matching kernel for the elements of the set corresponds to computing the cardinality of the intersection of the two sets. Alternatively, this computation can be seen as the inner product of the bit-vectors representing the two sets”*.

Yet another important class of kernels for structured data, notably put to the fore by Gärtner et al. (2004) yet by pointing out high associated computational costs, is the class of *convolution kernels* dating back to Haussler (1999). Convolution kernels can accommodate a variety of so-called “composite structures” by relying on their respective “parts”. They are constructed based on prescribed kernels between vectors of parts by instantiating and summing them with respect to all vectors of parts generating the considered compositive structures (Theorem 1 in Haussler (1999)). The proof of the latter theorem turns out to be based on the following Lemma that focuses on composite structures writing as finite subsets of a base set (say $\mathcal{X}$, to stick to the notation of the present paper):

Proposition A.1 (Lemma 1 of Haussler (1999))**.** *Let k be a kernel on $\mathcal{X} \times \mathcal{X}$ and for all finite, nonempty $A, B \subseteq \mathcal{X}$ define $k'(A, B) = \sum_{x \in A, y \in B} k(x, y)$. Then k' is a kernel on the product of the set of all finite, nonempty, subsets of $\mathcal{X}$ with itself.*

Let us remark that this construction is none other than what we refer to as the double sum kernels throughout the paper, notably at the heart of (Kim et al., 2019).

In contrast, the approach employed in (Kondor and Jebara, 2003) to create classes of kernels between sets consists in viewing these sets as samples from multivariate Gaussian distributions and then defining their baseline kernel in terms of Bhattacharyya affinity between those distributions. The resulting approach is then further enriched or "kernelized" thanks to the introduction of a second kernel defined between elementary vectors. In Cuturi et al. (2005), the focus is on kernels on measures characterized by the fact that the value of the kernel between two measures is a function of their sum, and the proposed constructions rely on common quantities defined on measures such as entropy or generalized variance. Quoting the article, *"the considered kernels can be used to derive kernels on structured objects, such as images and texts, by representing these objects as sets of components, such as pixels or words, or more generally as measures on the space of components"*. Here again, given an other kernel on the space of components itself, the approach is further extended using the "kernel trick".

Christmann and Steinwart (2010) investigate universal kernels on non-standard input spaces. They consider in particular a kernel on the set of probability measures obtained by chaining a radial Gaussian kernel and the RKHS distance between embedded distributions, coinciding in the case of uniform distributions over finite sets with our proposed class of *Deep Embedding* kernels. They show that in case of a compact base space and with probability measures endowed with the topology of weak convergence, the kernels of interest are universal. The reader is also referred to (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Sriperumbudur et al., 2011; Muandet et al., 2017) and references therein for more background results on RKHS embeddings of probability measures. Besides this, RKHS embeddings are also at the heart of the thesis Sutherland (2016), focusing on "Scalable, Flexible and Active Learning on Distributions". Kernel distribution embeddings have been recently further studied in Simon-Gabriel and Schölkopf (2018) from a functional analysis perspective, resulting in a proof that for kernels, being *universal*, *characteristic*, and *strictly positive definite* (where the definitions are slightly extended) are essentially equivalent. The latter paper gives furthermore a complete characterization of kernels whose associated Maximum Mean Discrepancy distance metrizes weak convergence, and it is shown in turn that kernel mean embeddings can be extended from probability measures to Schwartz distributions.

B Proofs of theoretical results

Proposition 1. *Let $\mathcal{X}$ be a set, $k_{\mathcal{X}}$ be a positive definite kernel on $\mathcal{X}$ with associated reproducing kernel Hilbert space $\mathcal{H}_{k_{\mathcal{X}}}$, and $\mathcal{S}_{fin}(\mathcal{X})$ be the set of non-empty finite subsets of $\mathcal{X}$. Let $\mathcal{E} : S \in \mathcal{S}_{fin}(\mathcal{X}) \mapsto \mathcal{H}_{k_{\mathcal{X}}}$, $k_0 : \mathcal{S}_{fin}(\mathcal{X}) \times \mathcal{S}_{fin}(\mathcal{X}) \mapsto \mathbb{R}$, $d_{\mathcal{E}} : \mathcal{S}_{fin}(\mathcal{X}) \times \mathcal{S}_{fin}(\mathcal{X}) \mapsto [0, \infty)$ be defined by Equations 1,2,3, respectively. Then,*

a) *$k_0(S, S') = \langle \mathcal{E}(S), \mathcal{E}(S') \rangle_{\mathcal{H}_{k_{\mathcal{X}}}}$ for any $S, S' \in \mathcal{S}_{fin}(\mathcal{X})$, and k_0 is positive definite on $\mathcal{S}_{fin}(\mathcal{X})$ while $d_{\mathcal{E}}$ is a pseudometric on $\mathcal{S}_{fin}(\mathcal{X})$.*

Let us furthermore introduce for $n \geq 2$ the sets

$$A_n = \Bigg\{ \Bigg(\overbrace{\frac{1}{n_1}, \ldots, \frac{1}{n_1}}^{(n_1 - \ell)\ times}, \overbrace{\frac{n_2 - n_1}{n_1 n_2}, \ldots, \frac{n_2 - n_1}{n_1 n_2}}^{\ell\ times}, \overbrace{\frac{-1}{n_2}, \ldots, \frac{-1}{n_2}}^{(n_2 - \ell)\ times} \Bigg),$$

$$n_1, n_2 \geq 1, \ell \geq 0 : n_1 + n_2 + \ell = n \Bigg\} \subset \mathbb{R}^n \ (n \geq 2).$$

b) *Then, the following assertions are equivalent:*

i) *$k_{\mathcal{X}}$ satisfies $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) > 0$ for all $n \geq 2$, pairwise distinct $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, and $(a_1, \ldots, a_n) \in A_n$.*

ii) *$\mathcal{E}$ is injective.*

iii) *$d_{\mathcal{E}}$ is a metric on $\mathcal{S}_{fin}(\mathcal{X})$.*

In particular, if $k_{\mathcal{X}}$ is strictly positive definite on $\mathcal{X}$, then all three conditions above are fulfilled.

Proof of Prop. 1. **a)** $k(S,S') = \langle \mathcal{E}(S), \mathcal{E}(S') \rangle_{\mathcal{H}_{k_{\mathcal{X}}}}$ $(S,S' \in \mathcal{S}_{\text{fin}}(\mathcal{X}))$ follows directly from scalar product bilinearity and $\langle k_{\mathcal{X}}(\mathbf{x},\cdot), k_{\mathcal{X}}(\mathbf{x}',\cdot) \rangle_{\mathcal{H}_{k_{\mathcal{X}}}} = k_{\mathcal{X}}(\mathbf{x},\mathbf{x}')$ $(\mathbf{x},\mathbf{x}' \in \mathcal{X})$, by reproducing property. Positive definiteness is then inherited from the scalar product as, for any $n \geq 1$, $a_1,\dots,a_n \in \mathbb{R}$ and $S_1,\dots,S_n \in \mathcal{S}_{\text{fin}}(\mathcal{X})$, $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(S_i,S_j) = ||\sum_{i=1}^n a_i \mathcal{E}(S_i)||^2_{\mathcal{H}_{k_{\mathcal{X}}}} \geq 0$. Similarly, the non-negativity, symmetry, and triangle inequality for $d_{\mathcal{E}}$ are inherited from the metric $||\cdot||_{\mathcal{H}_{k_{\mathcal{X}}}}$, making the former a pseudometric on $\mathcal{S}_{\text{fin}}(\mathcal{X})$. **b)** First, **ii)** $\Leftrightarrow$ **iii)** as $d_{\mathcal{E}}(S,S') = ||\mathcal{E}(S) - \mathcal{E}(S')||_{\mathcal{H}_{k_{\mathcal{X}}}}$ and **ii)** means that for $S \neq S'$ $\mathcal{E}(S) \neq \mathcal{E}(S')$, or equivalently $||\mathcal{E}(S) - \mathcal{E}(S')||_{\mathcal{H}_{k_{\mathcal{X}}}} \neq 0$ for $S \neq S'$, which is exactly what is needed for the pseudo-metric $d_{\mathcal{E}}$ to qualify as a metric on $\mathcal{S}_{\text{fin}}(\mathcal{X})$. **i)** $\Rightarrow$ **ii)**: Let $S = \{\mathbf{y}_1,\dots,\mathbf{y}_{n_1}\}$ and $S' = \{\mathbf{y}_1,\dots,\mathbf{y}_{n_2}\}$ be distinct elements of $\mathcal{S}_{\text{fin}}(\mathcal{X})$. Let us denote by $\ell \geq 0$ $(\ell \leq n_1 + n_2)$ the number of elements in $S \cap S'$ and denote $n = n_1 + n_2 - \ell$ and by $\mathbf{x}_1,\dots,\mathbf{x}_n$ the elements of $S \cup S'$ ordered so as to have as first $n_1 - \ell$ elements those of $S \backslash S'$, then the ℓ elements from $S \cap S'$, and finally those of $S' \backslash S$ (the orders within those three categories being arbitrary). Denote further here $\mathbf{X}_n = (\mathbf{x}_1,\dots,\mathbf{x}_n)$. Then,

$$\mathcal{E}(S) - \mathcal{E}(S') = \frac{1}{n_1} \sum_{i=1}^{n_1-\ell} k_{\mathcal{X}}(\mathbf{x}_i,\cdot)$$
$$+ \left(\frac{1}{n_1} - \frac{1}{n_2} \right) \sum_{i=n_1-\ell+1}^{n_1} k_{\mathcal{X}}(\mathbf{x}_i,\cdot) + \frac{1}{n_2} \sum_{i=n_1+1}^{n} k_{\mathcal{X}}(\mathbf{x}_i,\cdot),$$

whereof, putting $a_i = \frac{1}{n_1}$ $(1 \leq i \leq n_1 - \ell)$, $a_i = \frac{1}{n_1} - \frac{1}{n_2}$ $(n_1 - \ell + 1 \leq i \leq n_1)$, $a_i = \frac{1}{n_2}$ $(n_1 + 1 \leq i \leq n)$, and noting $k_{\mathcal{X}}(\mathbf{X}_n) = (k_{\mathcal{X}}(\mathbf{x}_i,\mathbf{x}_j))_{i,j \in \{1,\dots,n\}}$, we have

$$||\mathcal{E}(S) - \mathcal{E}(S')||_{H_{k_{\mathcal{X}}}} = \sqrt{\mathbf{a}' k_{\mathcal{X}}(\mathbf{X}_n)\mathbf{a}} > 0$$

where $\mathbf{a} = (a_1,\dots,a_n) \in A_n$ and the positivity follows from **i)**, implying that $\mathcal{E}(S) \neq \mathcal{E}(S')$ indeed. Assuming now that **ii)** holds and considering elements $\mathbf{x}_1,\dots,\mathbf{x}_n \in \mathcal{X}$ and $\mathbf{a} = (a_1,\dots,a_n) \in A_n$ such as in **i)** (with ℓ, n_1, n_2 following from $\mathbf{a}$), we define this time $S = \{\mathbf{x}_1,\dots,\mathbf{x}_{n_1+\ell}\}$ and $S' = \{\mathbf{x}_{n_1+1},\dots,\mathbf{x}_n\}$ and conclude that **i)** holds by pointing out that $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k_{\mathcal{X}}(\mathbf{x}_i,\mathbf{x}_j) = ||\mathcal{E}(S) - \mathcal{E}(S')||_{H_{k_{\mathcal{X}}}} > 0$, where $S \neq S'$ follows from the assumption of pairwise distinct $\mathbf{x}_i$'s. □

Proposition 2 (Non-strict positive definiteness of double sum kernels)**.** *Let us keep the notation of Proposition 1 and denote furthermore in the case of a finite set $\mathcal{X}$ with cardinality $c \geq 1$ and elements $\mathbf{X}_c = (\mathbf{x}_1,\dots,\mathbf{x}_c)$ by $u : S \in \mathcal{S}_{fin}(\mathcal{X}) \to u(S) = \frac{1}{\#S}(\mathbf{1}_{\mathbf{x}_i \in S})_{1 \leq i \leq c} \in \mathbb{R}^c$ the mapping returning for any nonempty subset of $\mathcal{X}$ a vector with components $\frac{1}{\#S}$ or 0 depending whether $\mathbf{x}_i \in S$ or not. Then we have:*

a) *For $\mathcal{X}$ finite, for any $S,S' \in \mathcal{S}_{fin}(\mathcal{X})$,*

$$k_0(S,S') = u(S)^T k_{\mathcal{X}}(\mathbf{X}_c) u(S').$$

Consequently, for $q \geq 1$ and $\mathbf{S} = (S_1,\dots,S_q) \in \mathcal{S}^q$, the covariance matrix $k_0(\mathbf{S})$ associated with $k_{\mathcal{X}}$ and $\mathbf{S}$ can be compactly written as

$$k_0(\mathbf{S}) = U(\mathbf{S})^T k_{\mathcal{X}}(\mathbf{X}_c) U(\mathbf{S}),$$

with the notation $U(\mathbf{S}) = [u(S_1),\dots,u(S_q)]$.

b) *For arbitrary $\mathcal{X}$, the two following assertions are mutually exclusive*

i) *$\#\mathcal{X} = 1$ and $k_{\mathcal{X}}$ is non-zero.*

ii) *k_0 is not strictly positive definite on $\mathcal{S}_{fin}(\mathcal{X})$.*

Proof of Prop. 2. **a)** Putting $k_{\mathcal{X}}(\mathbf{X}_c) = (k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j))_{i,j\in\{1,\dots,c\}}$ and $u(S) = \frac{1}{\#S}(\mathbf{1}_{\mathbf{x}_i \in S})_{1\leq i\leq c}$ in the right hand side directly delivers that

$$u(S)^T k_{\mathcal{X}}(\mathbf{X}_c) u(S) = \sum_{i=1}^{c}\sum_{j=1}^{c} \mathbf{1}_{\mathbf{x}_i\in S}\mathbf{1}_{\mathbf{x}_j\in S'} \frac{k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j)}{\#S\#S'},$$

which coincides indeed with Eq. 2's $k_0(S, S')$. Eq. 6 then simply follows as a Gram matrix associated with the bilinear form defined by Eq. 5. **b)** That **i)** $\Rightarrow$ **ii)** follows from the fact that if $\mathcal{X} = \{\mathbf{x}\}$ has cardinality 1 and $k_{\mathcal{X}}$ is strictly positive definite on $\mathcal{X}$, then $\mathcal{S}_{\text{fin}}(\mathcal{X})$ consists of the single element $\{\mathbf{x}\}$, and $k(\{\mathbf{x}\}, \{\mathbf{x}\}) = k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}) > 0$ whereof k is strictly positive definite on $\mathcal{S}_{\text{fin}}(\mathcal{X})$. To prove that **ii)** $\Rightarrow$ **i)**, let us now consider the case where $\mathcal{X}$'s cardinality is at least 2 (finite or not). From this assumption, it is possible to choose two distinct elements in $\mathbf{x}_A, \mathbf{x}_B \in \mathcal{X}$; let us denote here $\mathbf{X} = \{\mathbf{x}_A, \mathbf{x}_B\}$, and set $S_1 = \{\mathbf{x}_A\}$, $S_2 = \{\mathbf{x}_B\}$, $S_3 = \{\mathbf{x}_A, \mathbf{x}_B\}$, and $\mathbf{S} = (S_1, S_2, S_3)$. Following the same route as for Eq. 6, we then get

$$k_0(\mathbf{S}) = U(\mathbf{S})^T k_{\mathcal{X}}(\mathbf{X}) U(\mathbf{S}) = M(\mathbf{S})^T M(\mathbf{S}),$$

with $M(\mathbf{S}) = k_{\mathcal{X}}(\mathbf{X})^{\frac{1}{2}} U(\mathbf{S})$. Hence $\text{rank}(k_0(\mathbf{S})) \leq \text{rank}(k_{\mathcal{X}}(\mathbf{X})^{\frac{1}{2}}) = \text{rank}(k_{\mathcal{X}}(\mathbf{X})) \leq 2$ and so the 3×3 matrix $\text{rank}(k(\mathbf{S}))$ is non-invertible, proving indeed that k is not strictly positive definite on $\mathcal{S}_{\text{fin}}(\mathcal{X})$. □

Remark B.1. *The first equation of point **a)** highlights the fact that even if $k_{\mathcal{X}}(\mathbf{X})$ is a positive definite matrix (in particular, assuming that $k_{\mathcal{X}}$ is strictly p.d. on $\mathcal{X}$), the matrix $k_0(\mathbf{S})$ will actually be systematically singular for $q > c$. It turns out to also possibly happen in situations where $q \leq c$, as is for instance the case with $c = 5, q = 4$, and* $U(\mathbf{S}) \propto \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}$.

Proposition 3 ((Strict) positive definiteness of k_{DE})**.** *Let us consider here again the notation of Proposition 1 and consider furthermore the class of kernels $k_{DE} : (S, S') \in \mathcal{S}_{fin}(\mathcal{X}) \to k_H \circ d_{\mathcal{E}}(S, S')$ of Eq. 4, where $k_H : [0, \infty) \to \mathbb{R}$ is chosen such that $(h, h') \in \mathcal{H}^2 \to k_H(||h - h'||_{\mathcal{H}})$ is positive definite for any Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$. Then,*

a) *k_{DE} is positive definite on $\mathcal{S}_{fin}(\mathcal{X})$.*

b) *If furthermore $k_{\mathcal{X}}$ satisfies **i)** of condition **b)** in Proposition 1, and $k_H : [0, \infty) \to \mathbb{R}$ is chosen such that $(h, h') \in \mathcal{H}^2 \to k_H(||h - h'||_{\mathcal{H}})$ is strictly positive definite for any Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$, then k_{DE} is strictly positive definite on $\mathcal{S}_{fin}(\mathcal{X})$.*

Proof of Prop. 3. Both points essentially rely on the fact that $d_{\mathcal{E}}(S, S') = ||\mathcal{E}(S) - \mathcal{E}(S')||_{\mathcal{H}_{k_{\mathcal{X}}}}$ and that, as Reproducing Kernel Hilbert Space, $\mathcal{H}_{k_{\mathcal{X}}}$ is in the first place a Hilbert space. Indeed, writing $k_{\text{DE}}(S, S') = k_{\text{H}}(||\mathcal{E}(S) - \mathcal{E}(S')||_{\mathcal{H}_{k_{\mathcal{X}}}})$, we then directly obtain **a)** by composition of the positive definite kernel $(h, h') \in \mathcal{H}^2 \to k_{\text{H}}(||h - h'||_{\mathcal{H}_{k_{\mathcal{X}}}})$ with the mapping $\mathcal{E} : \mathcal{S}_{\text{fin}}(\mathcal{X}) \mapsto \mathcal{H}_{k_{\mathcal{X}}}$. As for **b)**, assuming furthermore k_{H} to be strictly positive definite on any Hilbert space and **i)** of condition **b)** in Proposition 1 to hold, then the strict positive definiteness of k_{DE} follows from the one of k_{H} and the injectivity of $\mathcal{E}$ ensured by Proposition 1. □

Proposition 4. *Let $r_{\mathcal{X}}$ be an isotropic positive definite kernel on $\mathcal{X} = [0, 1]^d$ assumed to be monotonically decreasing to 0 with respect to the Euclidean distance between elements of $\mathcal{X}$, with range parameter $\theta_{\mathcal{X}} > 0$. Then the $d_{\mathcal{E}_{r_{\mathcal{X}}}}$-diameter of $\mathcal{S}_p(\mathcal{X})$ ($p \geq 1$), i.e. $\sup_{S,S'\in\mathcal{S}_p} d_{\mathcal{E}_{r_{\mathcal{X}}}}(S, S')$, is reached with arguments $\{\mathbf{0}_d, \dots, \mathbf{0}_d\}$ and $\{\mathbf{1}_d, \dots, \mathbf{1}_d\}$, where $\mathbf{0}_d = (0, \dots, 0), \mathbf{1}_d = (1, \dots, 1) \in \mathcal{X}$. Furthermore, the supremum of this diameter with respect to $\theta_{\mathcal{X}} \in (0, +\infty)$ is given by $\sqrt{2}$.*

Proof of Prop. 4. Let us consider two sets $S = \{\mathbf{x}_1, \dots, \mathbf{x}_p\}, S' = \{\mathbf{x}'_1, \dots, \mathbf{x}'_p\} \in \mathcal{S}_p$. Then, from the fact that a correlation kernel is upper-bounded by 1, we get

$$\begin{aligned}
d^2_{\mathcal{E}_{r_{\mathcal{X}}}}(S,S') &= \frac{1}{p^2}\left(\sum_{i=1}^{p}\sum_{j=1}^{p} r_{\mathcal{X}}(\mathbf{x}_i,\mathbf{x}_j) + \sum_{i=1}^{p}\sum_{j=1}^{p} r_{\mathcal{X}}(\mathbf{x}'_i,\mathbf{x}'_j)\right. \\
&\quad \left. -2\sum_{i=1}^{p}\sum_{j=1}^{p} r_{\mathcal{X}}(\mathbf{x}_i,\mathbf{x}'_j)\right) \\
&\leq \frac{1}{p^2}\left(2p^2 - 2\sum_{i=1}^{p}\sum_{j=1}^{p} r_{\mathcal{X}}(\mathbf{x}_i,\mathbf{x}'_j)\right) \\
&\leq \frac{1}{p^2}\left(2p^2 - 2\sum_{i=1}^{p}\sum_{j=1}^{p} r_{\mathcal{X}}(\mathbf{0}_d,\mathbf{1}_d)\right),
\end{aligned}$$

where the last inequality follows from the assumed monotonicity of $r_{\mathcal{X}}$ with respect to the Euclidean distance between elements of $\mathcal{X}$ and the fact that the maximal distance between two points of $\mathcal{X}$, i.e. the Euclidean diameter of $[0,1]^d$, is precisely attained for $\mathbf{x} = \mathbf{0}_d$ and $\mathbf{x}' = \mathbf{1}_d$. Finally, by assumption again, $r_{\mathcal{X}}(\mathbf{0}_d, \mathbf{1}_d)$ is monotonically decreasing to 0 when $\theta_{\mathcal{X}}$ decreases to 0, and so the upper bound of $d^2_{\mathcal{E}_{r_{\mathcal{X}}}}$ tends to $\frac{1}{p^2}\left(2p^2 - 0\right) = 2$, showing that upper bound of the $d_{\mathcal{E}_{r_{\mathcal{X}}}}$-diameter of $\mathcal{S}_p$ with respect to $\theta_{\mathcal{X}} \in (0, +\infty)$ is $\sqrt{2}$ indeed, independently of the dimension. □

C Complements on the methodology

C.1 Maximum likelihood estimation for GPs with Deep Embedding kernel

In the numerical experiments, we make predictions under a stationary GP model which assumes a constant unknown trend (following the route of Ordinary Kriging prediction such as exposed in (Roustant et al., 2012)). When both $k_{\mathcal{X}}$ and k_{H} are assumed to be Gaussian kernels (still with the parametrization mentioned in (Roustant et al., 2012)), the introduced Deep Embedding kernel takes the form

$$\begin{aligned}
k_{DE}(S,S') &= k_{\mathrm{H}} \circ d_{\mathcal{E}}(S,S') \\
&= \sigma_H^2 r_H \circ d_{\mathcal{E}}(S,S') \qquad \text{(C.1.1)}\\
&= \sigma_H^2 \exp\left(-\frac{1}{2}\frac{d^2_{\mathcal{E}}(S,S')}{\theta_H^2}\right), \qquad \text{(C.1.2)}
\end{aligned}$$

where

$$\begin{aligned}
d_{\mathcal{E}}(S,S') = &\left(\frac{1}{\#S\#S}\sum_{\mathbf{x}_1,\mathbf{x}_2\in S}\exp\left(-\frac{1}{2}\frac{\|\mathbf{x}_1-\mathbf{x}_2\|^2}{\theta_{\mathcal{X}}^2}\right) + \frac{1}{\#S'\#S'}\sum_{\mathbf{x}'_1,\mathbf{x}'_2\in S'}\exp\left(-\frac{1}{2}\frac{\|\mathbf{x}'_1-\mathbf{x}'_2\|^2}{\theta_{\mathcal{X}}^2}\right)\right. \\
&\left. - \frac{2}{\#S\#S'}\sum_{\mathbf{x}\in S,\mathbf{x}'\in S'}\exp\left(-\frac{1}{2}\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{\theta_{\mathcal{X}}^2}\right)\right)^{\frac{1}{2}}. \qquad \text{(C.1.3)}
\end{aligned}$$

The three hyperparameters are determined by Maximum Likelihood Estimation (MLE). The expression of k_{DE} as a function of r_H in Equation C.1.1 allows us to use the concentrated log-likelihood, optimized with respect to θ_H and $\theta_{\mathcal{X}}$ via genetic algorithm with derivatives (Mebane Jr et al., 2011). This can be done in a similar manner to the method given in Appendix A of Roustant et al. (2012). Assuming positive values for the hyperparameters, the derivatives of $r_H(\cdot,\cdot)$ with respect to the two hyperparameters θ_H and $\theta_{\mathcal{X}}$ exist and are respectively given by:

$$\frac{\partial r_H(S,S')}{\partial \theta_H} = \exp\left(-\frac{1}{2}\frac{d_{\mathcal{E}}(S,S')^2}{\theta_H^2}\right)\left(\frac{d_{\mathcal{E}}(S,S')^2}{\theta_H^3}\right) = r_H(S,S')\left(\frac{d_{\mathcal{E}}(S,S')^2}{\theta_H^3}\right), \qquad \text{(C.1.4)}$$

and

$$\frac{\partial r_H(S,S')}{\partial \theta_{\mathcal{X}}} = -\frac{1}{2\theta_H^2} \exp\left(-\frac{1}{2}\frac{d_{\mathcal{E}}(S,S')^2}{\theta_H^2}\right) \frac{\partial d_{\mathcal{E}}(S,S')^2}{\partial \theta_{\mathcal{X}}} = -\frac{1}{2\theta_H^2} r_H(S,S') \frac{\partial d_{\mathcal{E}}(S,S')^2}{\partial \theta_{\mathcal{X}}}, \tag{C.1.5}$$

where

$$\begin{aligned} \frac{\partial d_{\mathcal{E}}(S,S')^2}{\partial \theta_{\mathcal{X}}} &= \frac{1}{\#S^2} \sum_{\mathbf{x}_1,\mathbf{x}_2 \in S} \exp\left(-\frac{1}{2}\frac{\|\mathbf{x}_1-\mathbf{x}_2\|^2}{\theta_{\mathcal{X}}^2}\right)\left(\frac{\|\mathbf{x}_1-\mathbf{x}_2\|^2}{\theta_{\mathcal{X}}^3}\right) \\ &+ \frac{1}{\#S'^2} \sum_{\mathbf{x}'_1,\mathbf{x}'_2 \in S'} \exp\left(-\frac{1}{2}\frac{\|\mathbf{x}'_1-\mathbf{x}'_2\|^2}{\theta_{\mathcal{X}}^2}\right)\left(\frac{\|\mathbf{x}'_1-\mathbf{x}'_2\|^2}{\theta_{\mathcal{X}}^3}\right) \\ &- \frac{2}{\#S\#S'} \sum_{\mathbf{x}\in S,\mathbf{x}' \in S'} \exp\left(-\frac{1}{2}\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{\theta_{\mathcal{X}}^2}\right)\left(\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{\theta_{\mathcal{X}}^3}\right). \end{aligned} \tag{C.1.6}$$

C.2 Condition number and jitter for matrix inversion

The condition number of an $n \times n$ positive definite matrix $\mathbf{R}$ under the 2-norm is defined by

$$\kappa(\mathbf{R}) = \|\mathbf{R}\|_2 \left\|\mathbf{R}^{-1}\right\|_2 = \frac{\lambda_n}{\lambda_1}, \tag{C.2.7}$$

where λ_n and λ_1 are the largest and smallest positive eigenvalues of $\mathbf{R}$, respectively. A matrix is said to be ill-conditioned when its condition number is larger than some prescribed threshold.

Given an ill-conditioned matrix, one can perturb the matrix by adding a small "jitter" δ to diagonal in order to decrease its condition number:

$$\mathbf{R}_\delta = \mathbf{R} + \delta \mathbf{I}, \tag{C.2.8}$$

where $\mathbf{I}$ denotes the identity matrix with appropriate dimension. The eigenvalues of the perturbed matrix $\mathbf{R}_\delta$ become $\lambda_i + \delta$, $i = 1, 2, 3, ..., n$ where λ_i is the ith smallest eigenvalue of the original matrix $\mathbf{R}$.

In Gaussian Process modelling, it is not rare that the inversion of ill-conditioned covariance/correlation matrices constitutes a bottleneck, motivating to introduce a positive jitter δ; yet, finding an appropriate value for such a δ is no straightforward task and too small a value might not fix the issue of near singularity while too big a value could cause over-regularization and result in a poor surrogate of the inverse. One approach is to consider the jitter as a model hyperparameter and estimate it, e.g., by MLE. However, implementing this method may end up introducing positive jitter values even the matrix itself is well-conditioned. Also, things can be challenging from the computational point of view when δ takes a variety of values in the course of hyperparameter optimization.

Ranjan et al. (2011) proposed an alternative way by finding a lower bound of the jitter that can overcome the ill-condition issue while minimizing the over-smoothing. As proven in (Ranjan et al., 2011), the condition number $\kappa(\mathbf{R}_\delta)$, setting a jitter level to

$$\delta\,(a) = \frac{\lambda_n\,(\kappa(\mathbf{R}) - \exp(a))}{\kappa(\mathbf{R})(\exp(a) - 1)}, \tag{C.2.9}$$

will ensure that the condition number of $\mathbf{R}_\delta$ remains below a prescribed value $\exp(a)$.

D Complementary experimental results

D.1 DS kernel +jitter for contaminant source localization test cases

Due to conditioning issues in combinatorial problems, the double sum kernel is not readily applicable for the contaminant source localization test case. We hence apply the described jitter trick in the case of GP prediction with DS kernel on this test case. In particular, to find an appropriately small jitter, we vary the value of "a" $= 1, 2, 3, ..., 7$ in Equation C.2.9, and compare both prediction and optimization performances of the modified DS kernel when the corresponding bound values for the jitter are used.

In the numerical experiments, once the jitter δ is set, the correlation matrix $\mathbf{R}_\delta = \mathbf{R} + \delta$ is used in all computations. This includes not only the computation of predictive mean and variance, but also the log-likelihood as well as its partial derivatives with respect to hyperparameters.

D.1.1 Prediction performance

Table D.1.1 gives Q^2 values for GP models with the proposed DE kernel against DS ones with multiple values of "a" on the four considered scenarios for the contamination test case (refer to Table 1 in the main article).

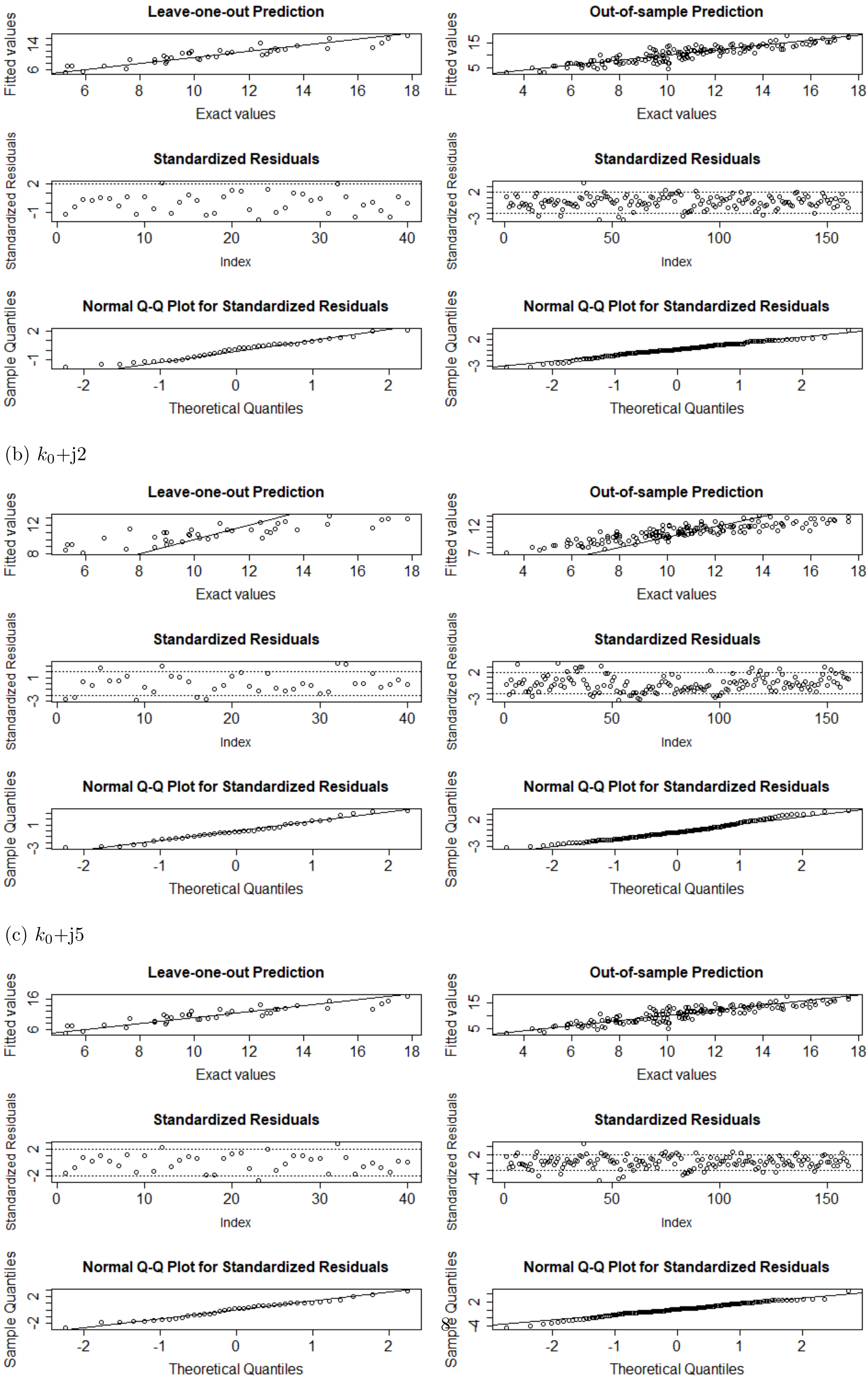
We can see from the table that small values of "a", e.g. $a = 1$ and 2, which corresponds to larger jitter levels, yield higher prediction errors. Here in fact, the DE kernel outperforms the DS kernels on all cases.

Table D.1.1: Q^2 values for GP predictions on contamination test cases with DE versus DS kernels (k_{DE} versus k_0+j)

Q^2	Ratio	k_{DE}	$k_0 + j1$	$k_0 + j2$	$k_0 + j3$	$k_0 + j4$	$k_0 + j5$	$k_0 + j6$	$k_0 + j7$
Src A, Geo 1	20:80	0.7607	0.3177	0.5756	0.7117	0.7501	0.7437	0.7109	0.6568
	50:50	0.9133	0.3557	0.6506	0.7970	0.8391	0.8445	0.8438	0.8424
	80:20	0.9352	0.4060	0.6930	0.8326	0.8728	0.8804	0.8815	0.8818
Src A, Geo 2	20:80	0.7239	0.2393	0.4884	0.6399	0.7013	0.7130	0.7025	0.6584
	50:50	0.8855	0.3557	0.6430	0.8001	0.8449	0.8485	0.8476	0.8460
	80:20	0.9240	0.3352	0.6514	0.8206	0.8673	0.8729	0.8724	0.8719
Src B, Geo 1	20:80	0.7977	0.2946	0.5457	0.7087	0.7775	0.7901	0.7720	0.7354
	50:50	0.9190	0.3302	0.6450	0.8152	0.8668	0.8746	0.8749	0.8743
	80:20	0.9447	0.3878	0.6847	0.8369	0.8818	0.8904	0.8916	0.8918
Src B, Geo 2	20:80	0.8486	0.2930	0.5672	0.7434	0.8182	0.8389	0.8398	0.8338
	50:50	0.9151	0.3904	0.6916	0.8465	0.8880	0.8944	0.8946	0.8941
	80:20	0.9439	0.4922	0.7543	0.8862	0.9207	0.9252	0.9259	0.9258

Figures D.1.1-D.1.8 show residual analyses for both leave-one-out and out-sample validation errors over four contaminant test cases. Here, we present only results for k_0+j2 and k_0+j5 (corresponding to the case when "a"= 2 and "a"= 5, respectively) to give a compact yet representative illustration of compared performances against the DE kernel.

As one can see, assigning an inappropriate "a" value can lead to very poor predictive results ($a = 2$). The fact that using the exposed approach with jitter heavily relies on the value of "a" confers a relative robustness advantage to strictly positive definite DE kernels as no jitter is needed. This comes of course at the price of an additional hyperparameter to be fitted, yet with an estimation that can be more conveniently conducted together with the estimation of the other hyperparameters.

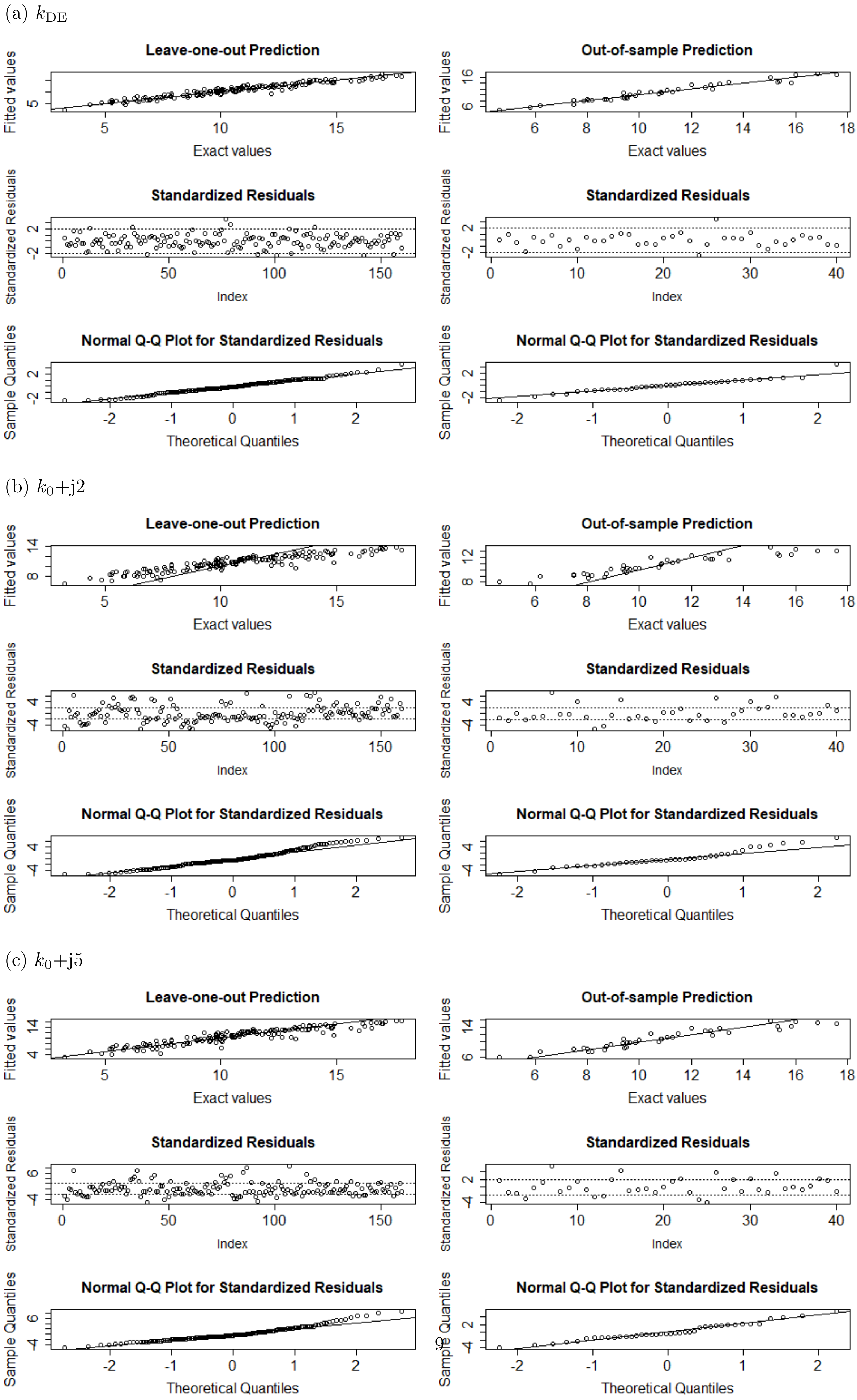


(a) k_{DE}

(b) k_0+j2

(c) k_0+j5

Figure D.1.1: Residual analysis on contamination test case **(Src A, Geo 1) with (20:80)**, (a) k_{DE}, (b) k_0+j2 and (c) k_0+j5

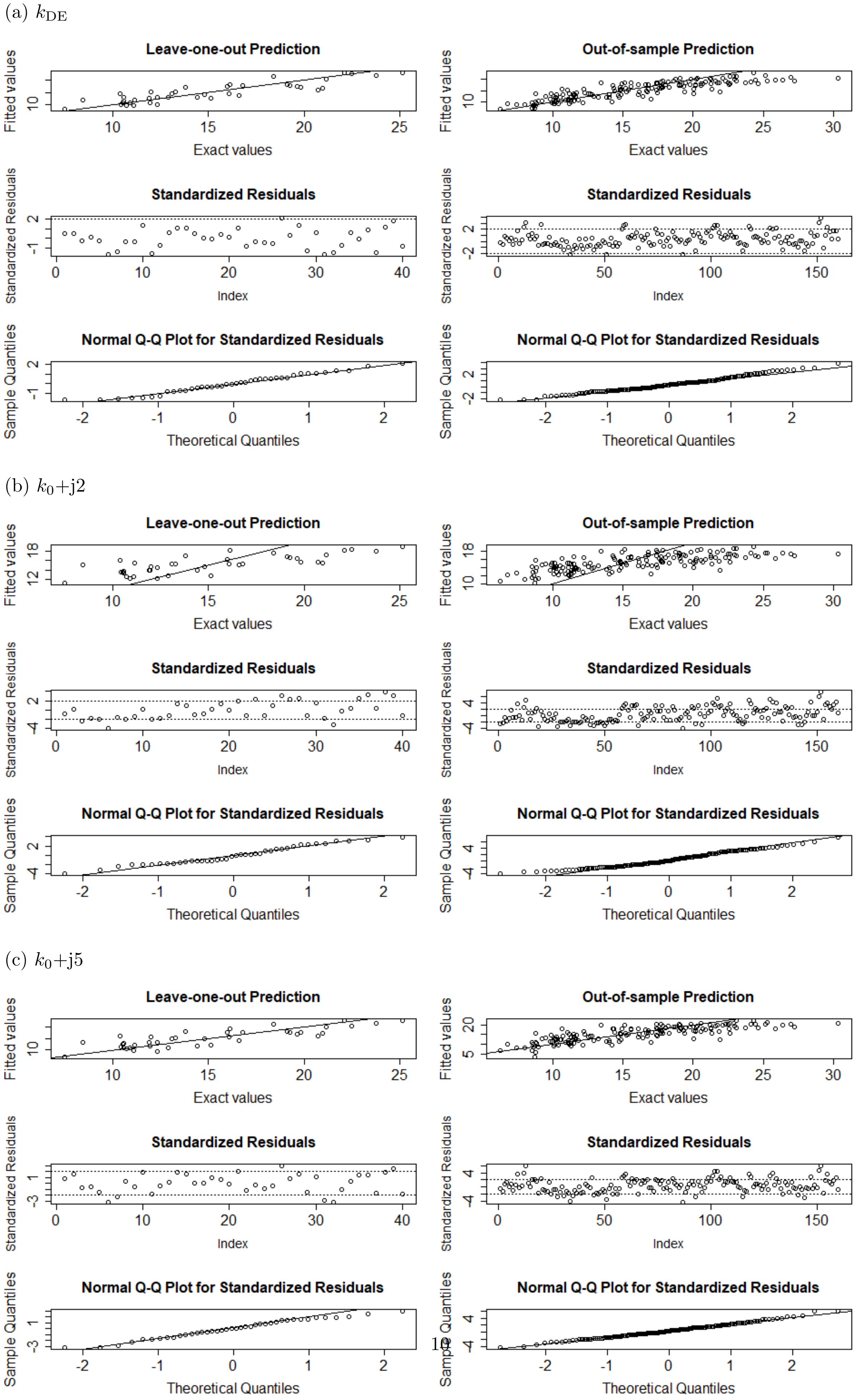


(a) k_{DE}

(b) k_0+j2

(c) k_0+j5

Figure D.1.2: Residual analysis on contamination test case **(Src A, Geo 1) with (80:20)**, (a) k_{DE}, (b) k_0+j2 and (c) k_0+j5

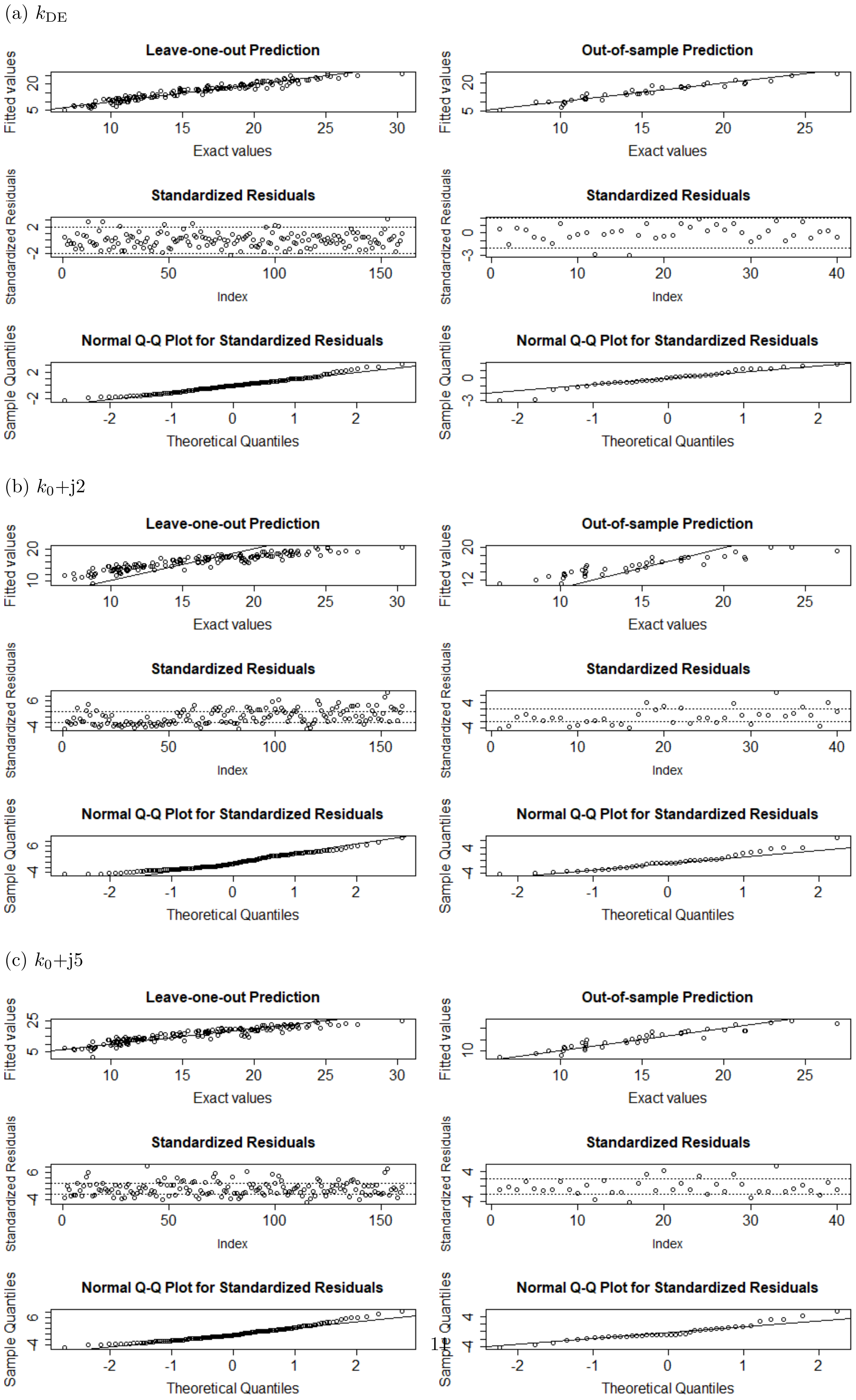


(a) k_{DE}

(b) k_0+j2

(c) k_0+j5

Figure D.1.3: Residual analysis on contamination test case **(Src A, Geo 2) with (20:80)**, (a) k_{DE}, (b) k_0+j2 and (c) k_0+j5

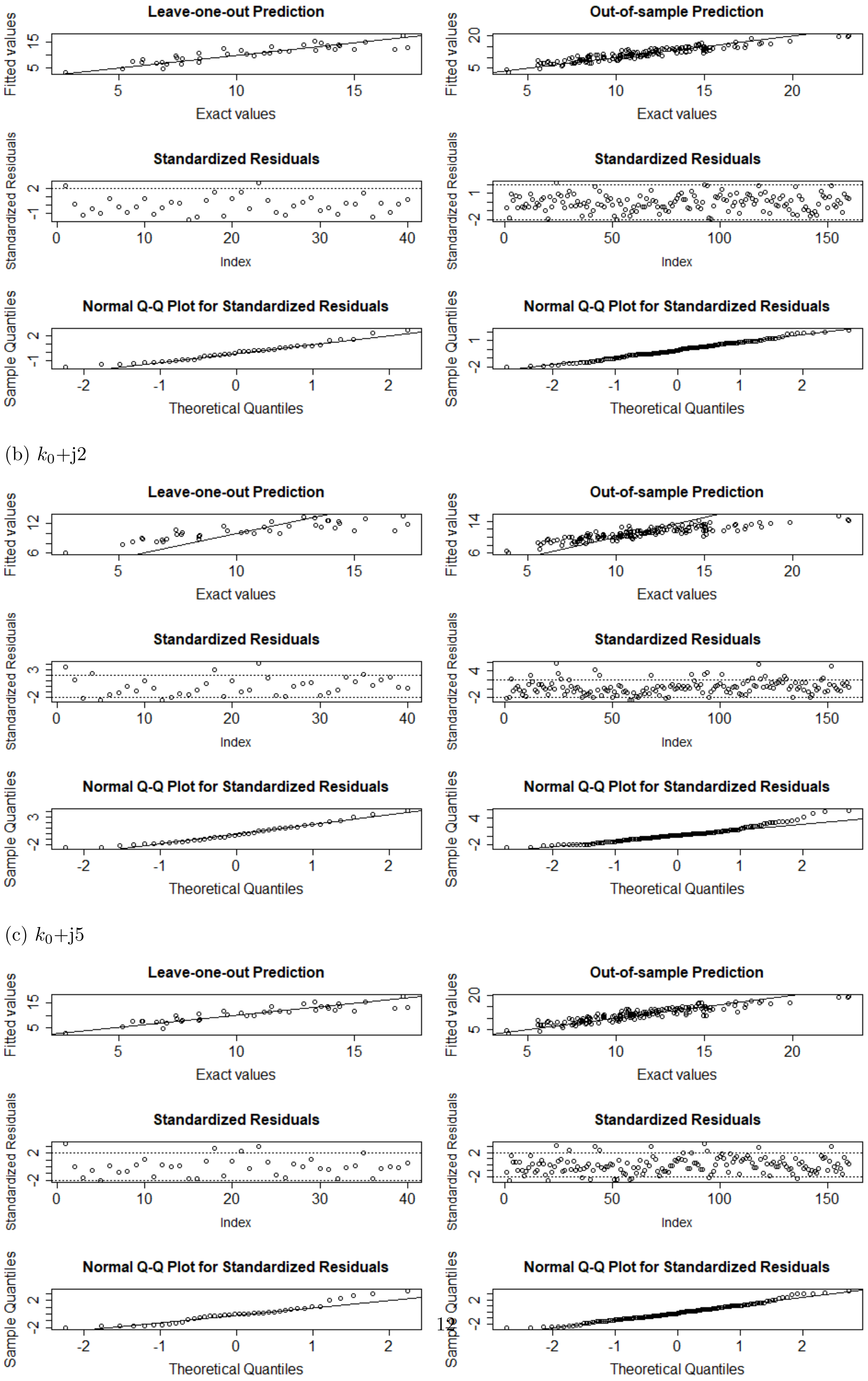


(a) k_{DE}

(b) k_0+j2

(c) k_0+j5

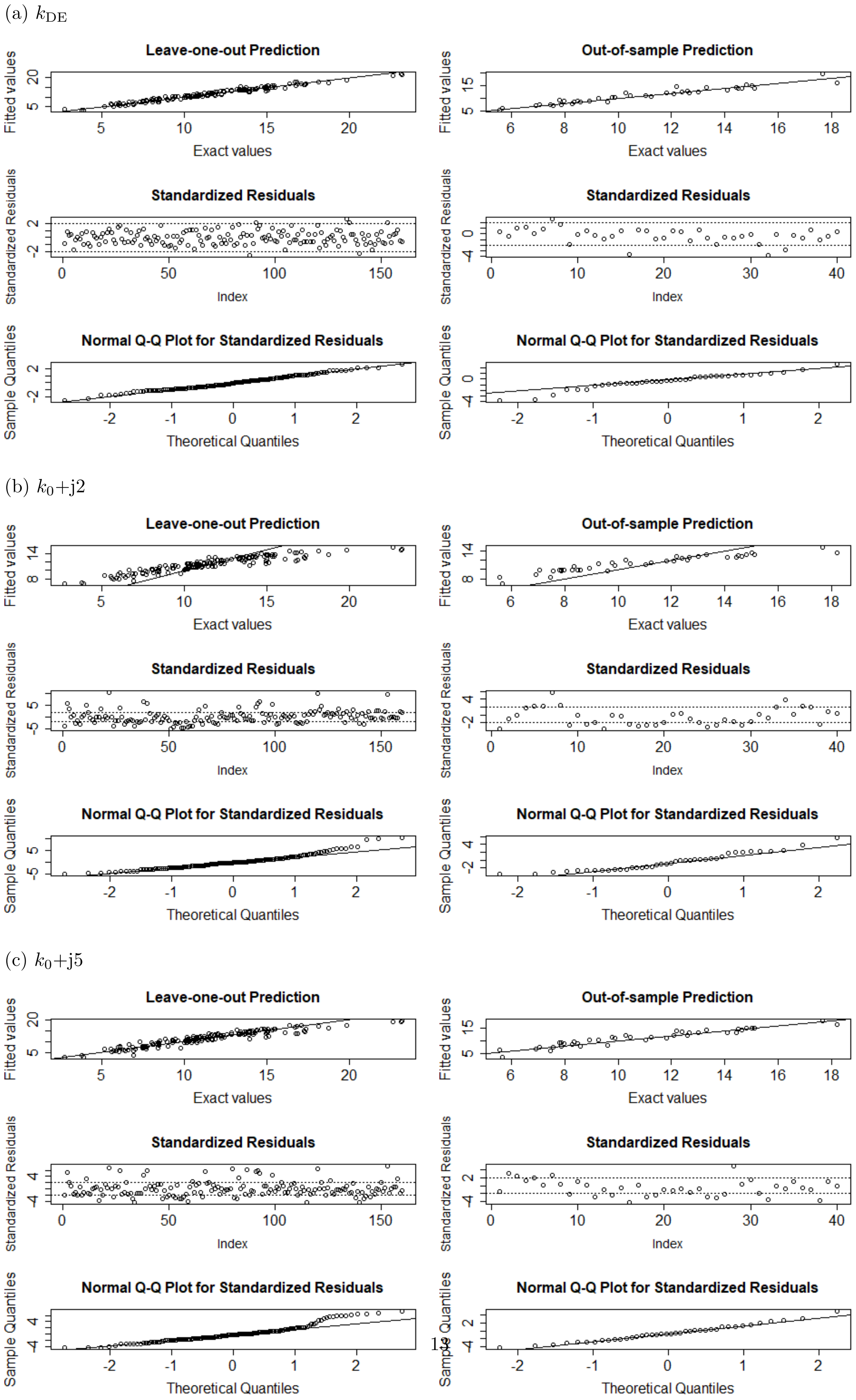
Figure D.1.4: Residual analysis on contamination test case **(Src A, Geo 2) with (80:20)**, (a) k_{DE}, (b) k_0+j2 and (c) k_0+j5


(a) k_{DE}

(b) k_0+j2

(c) k_0+j5

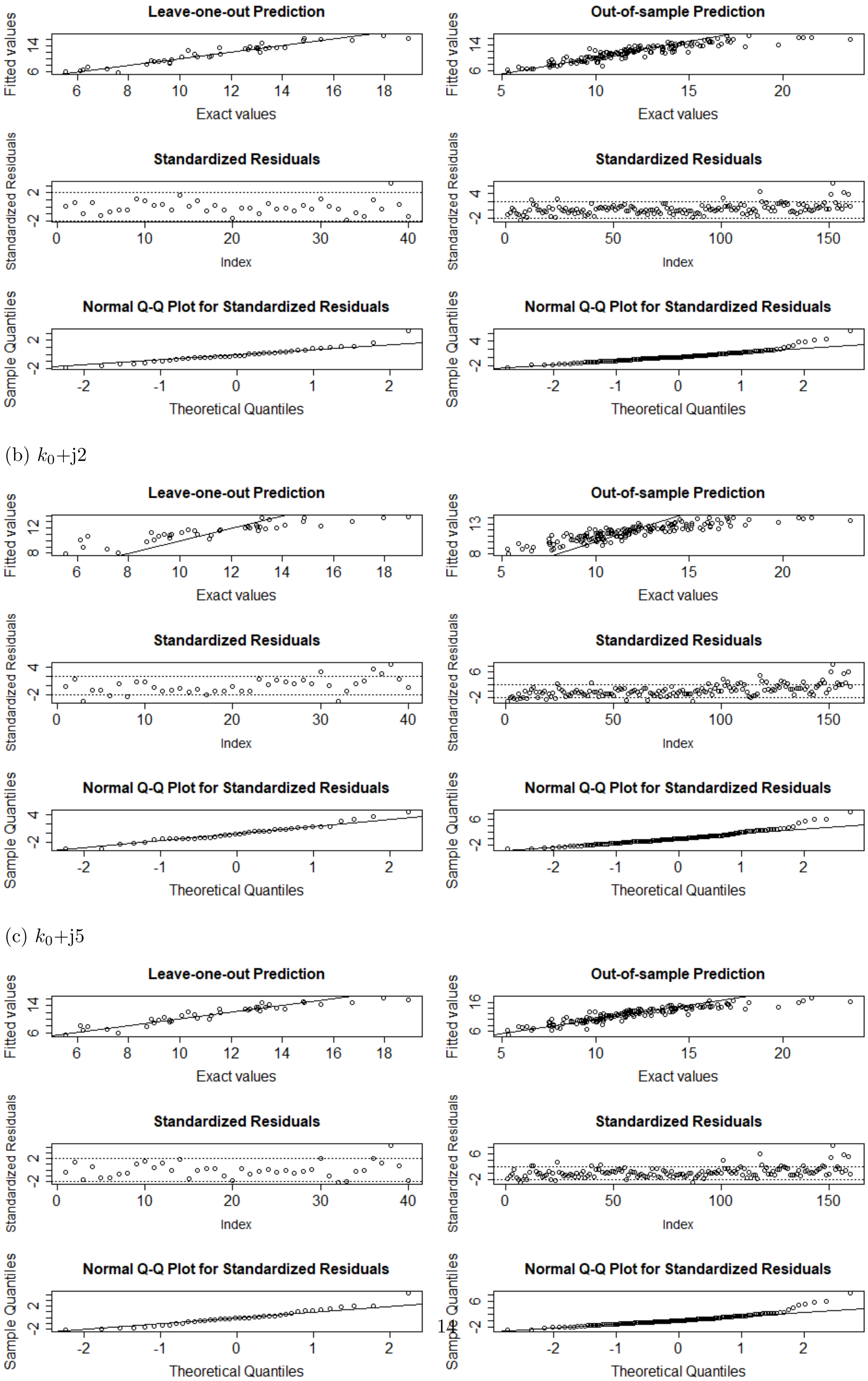
Figure D.1.5: Residual analysis on contamination test case **(Src B, Geo 1) with (20:80)**, (a) k_{DE}, (b) k_0+j2 and (c) k_0+j5


(a) k_{DE}

(b) k_0+j2

(c) k_0+j5

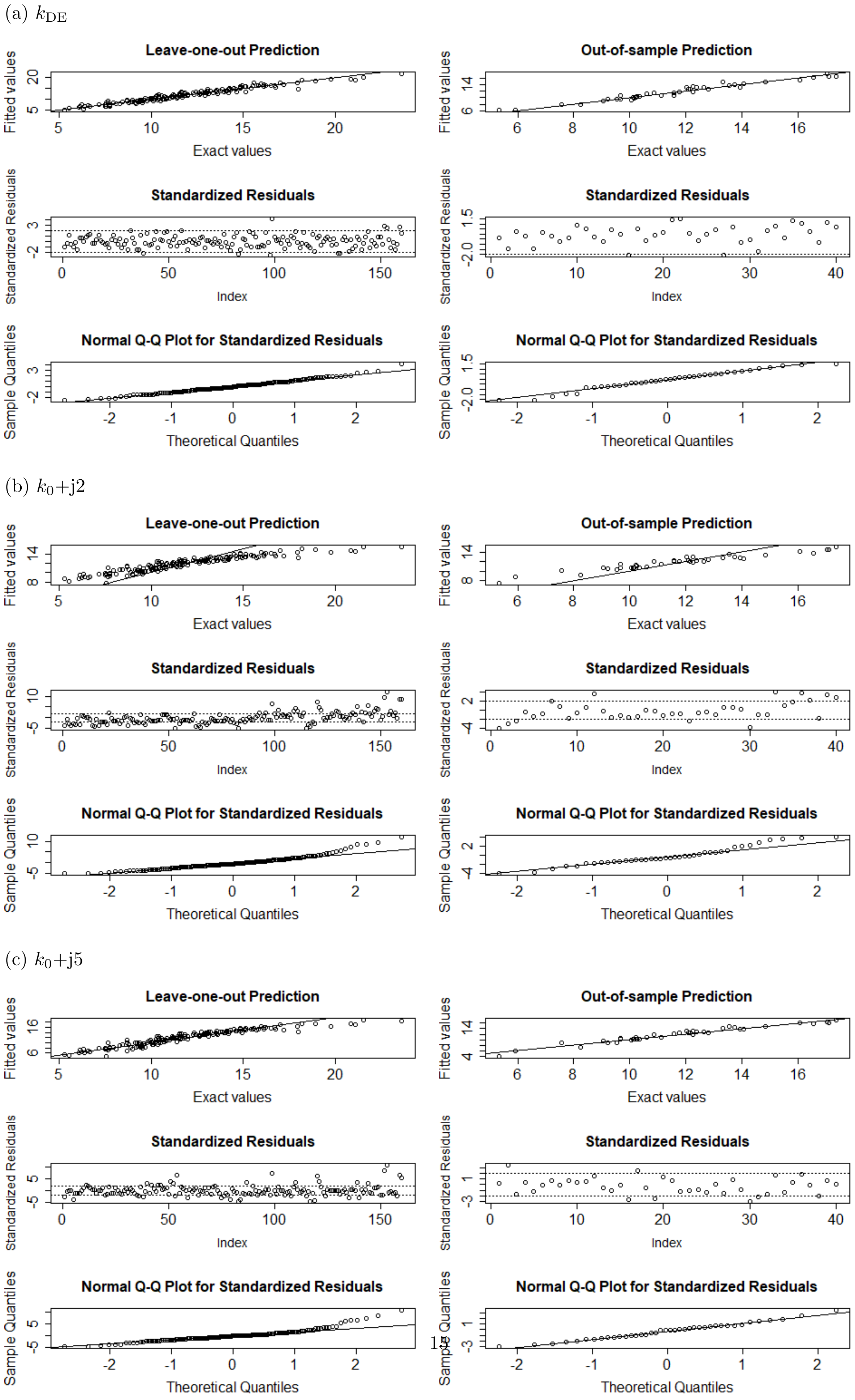
Figure D.1.6: Residual analysis on contamination test case **(Src B, Geo 1) with (80:20)**, (a) k_{DE}, (b) k_0+j2 and (c) k_0+j5


(a) k_{DE}

(b) k_0+j2

(c) k_0+j5

Figure D.1.7: Residual analysis on contamination test case **(Src B, Geo 2) with (20:80)**, (a) k_{DE}, (b) k_0+j2 and (c) k_0+j5


(a) k_{DE}

(b) k_0+j2

(c) k_0+j5

Figure D.1.8: Residual analysis on contamination test case **(Src B, Geo 2) with (80:20)**, (a) k_{DE}, (b) k_0+j2 and (c) k_0+j5

D.1.2 Optimization performance

In line with Section 3.3 of the main article, in this section, we present complete results of (1) the number of trials such that the minimum is found by EI with k_{DE} and $k_0 + j$ in Table D.1.2; (2) the progress curves in terms of the median value of current best response in Figure D.1.9; and (3) the 95th percentile of current best response in Figure D.1.10.

Table D.1.2: Number of trials (out of 100) such that minimum is found by EI algorithms with DE and DS kernels (k_{DE} versus k_0+j) on four contamination problems

Problem	**EI-k_{DE}**	**EI-$k_0 + j1$**	**EI-$k_0 + j2$**	**EI-$k_0 + j3$**	**EI-$k_0 + j4$**
(a) Src A, Geo 1	100	17	63	87	95
(b) Src A, Geo 2	66	15	36	46	52
(c) Src B, Geo 1	100	26	59	77	95
(d) Src B, Geo 2	78	42	64	76	81
Problem	**EI-$k_0 + j5$**	**EI-$k_0 + j6$**	**EI-$k_0 + j7$**	**RANDOM**	
(a) Src A, Geo 1	98	96	97	0	
(b) Src A, Geo 2	46	47	44	0	
(c) Src B, Geo 1	96	96	95	0	
(d) Src B, Geo 2	82	82	81	0	

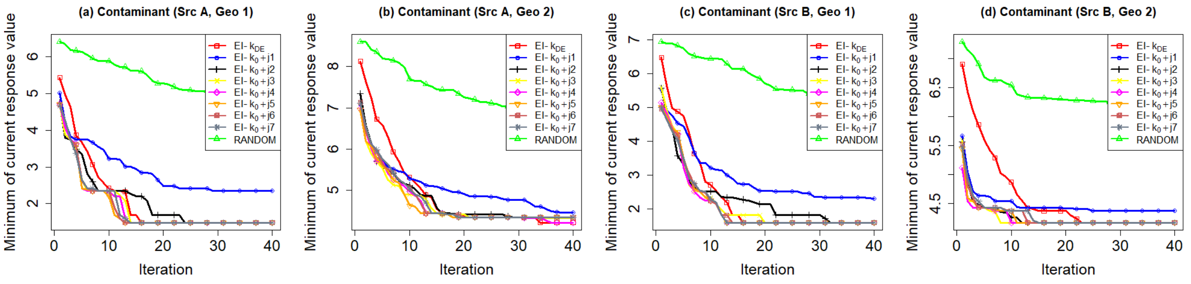


Figure D.1.9: The median of current best response over 40 iterations on four contamination test cases

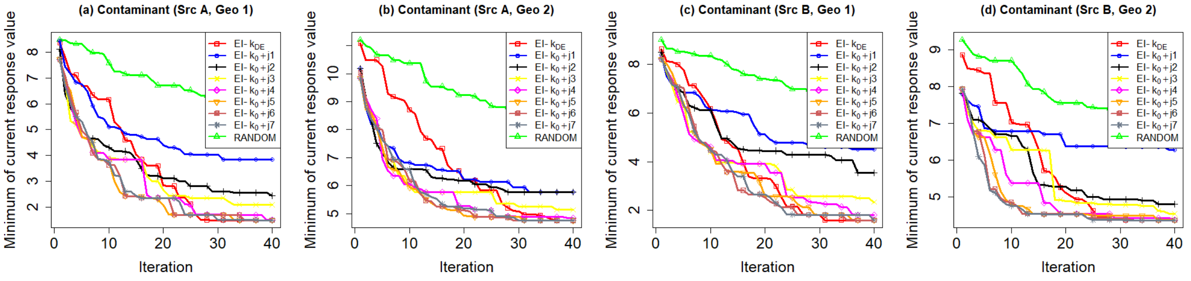
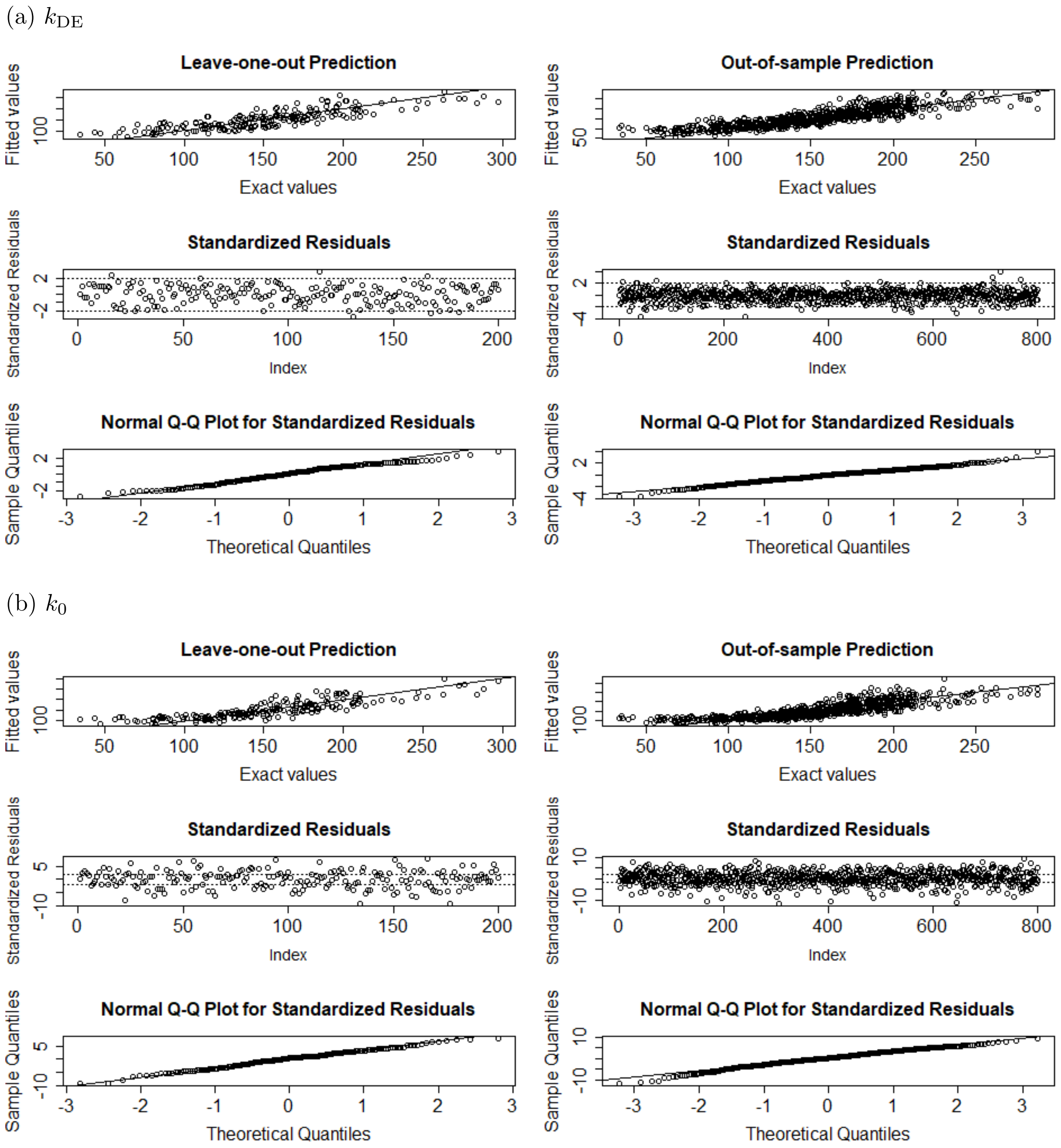


Figure D.1.10: The 95th percentile of current best response over 40 iterations on four contamination test cases

Table D.1.2 indicates that with the DE kernel, EI could locate the true minimum for more replications than that with the DS kernels (at all jitter levels) for all problems, except for Source B, Geology 2. The progress curves of median and 95th percentile values suggest that regardless of the jitter level added, EI-$k_0 + j$ method decreases the function value quickly at the beginning of the course when the kernel is still very well conditioned. With more points in the observation sets, jitter cannot be avoided as the kernel becomes ill-conditioned. When this happens, the performance of $k_0 + j$ heavily depends on the jitter levels, as the progress curve starts to flatten out. Notice how the EI-k_{DE} curve crosses the EI-$k_0 + j$ one in the 95th percentile plots. Because the model accuracy as well as optimization performance of the DS kernel relies on the jitter levels, this makes the approach less robust than the DE kernel.

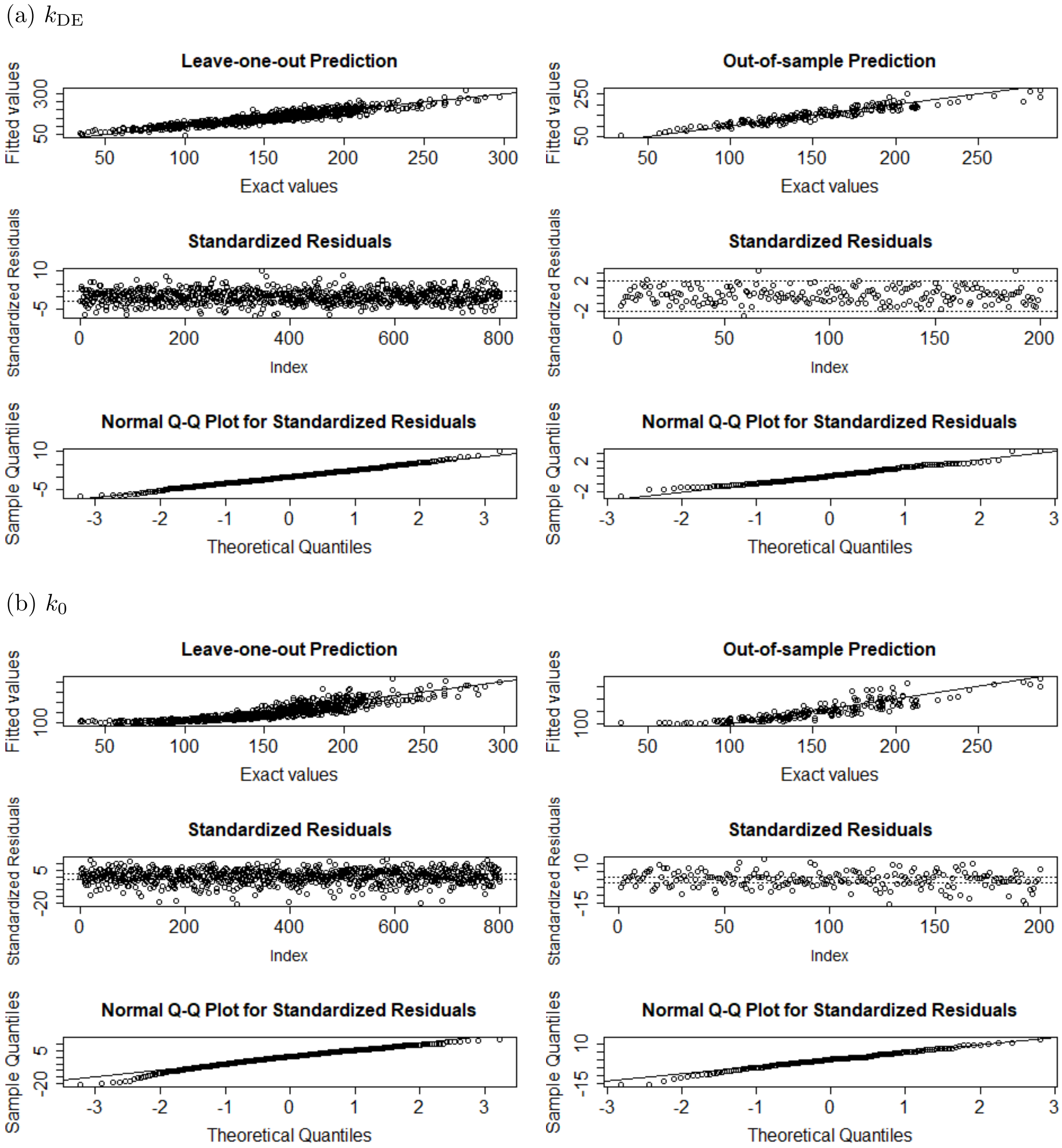
D.2 Complementary residual analyses for the synthetic and Castem test cases



(a) k_{DE}

(b) k_0

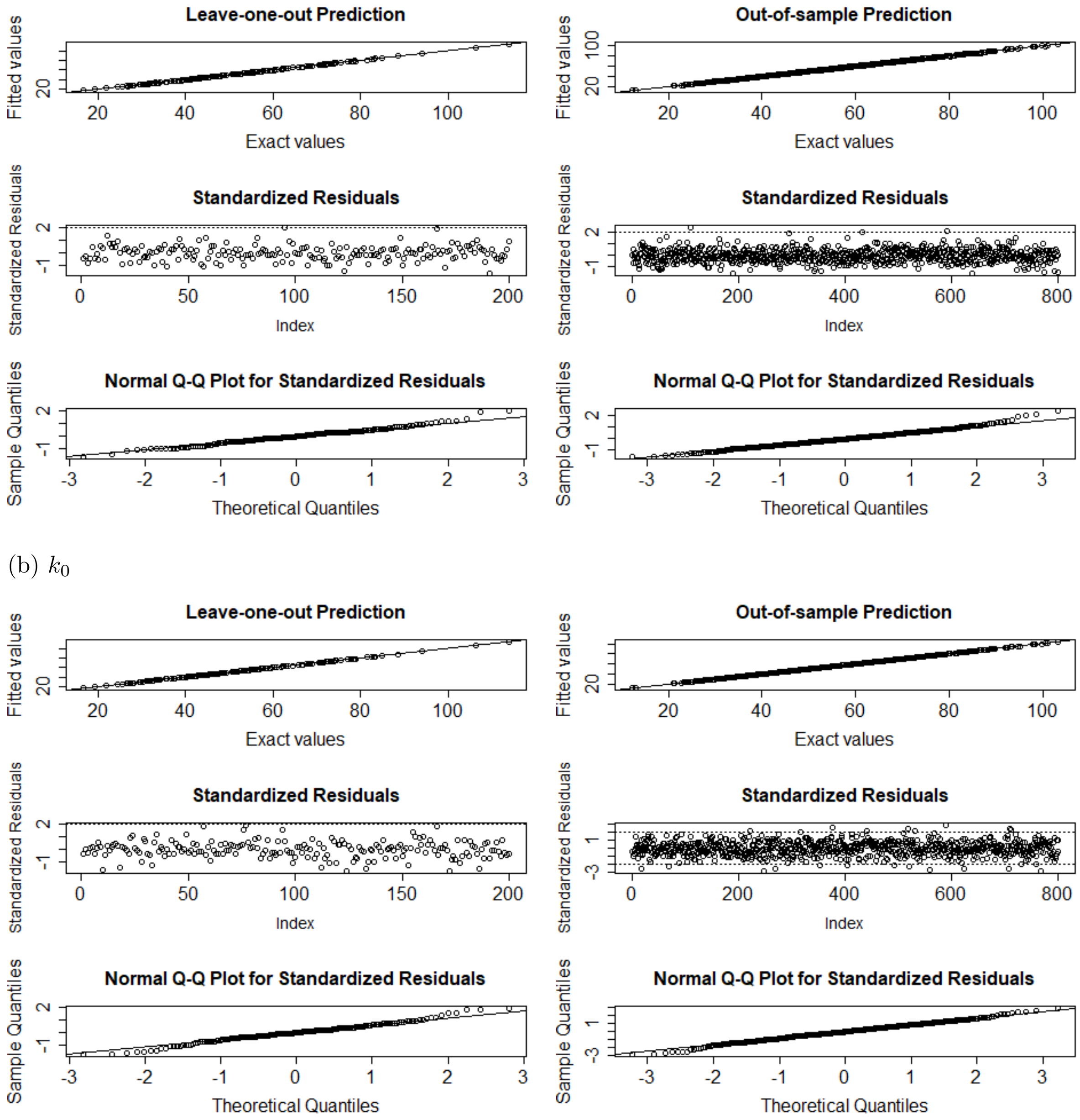
Figure D.2.11: Residual analysis on **MAX with (20:80)**, (a) k_{DE} and (b) k_0



(a) k_{DE}

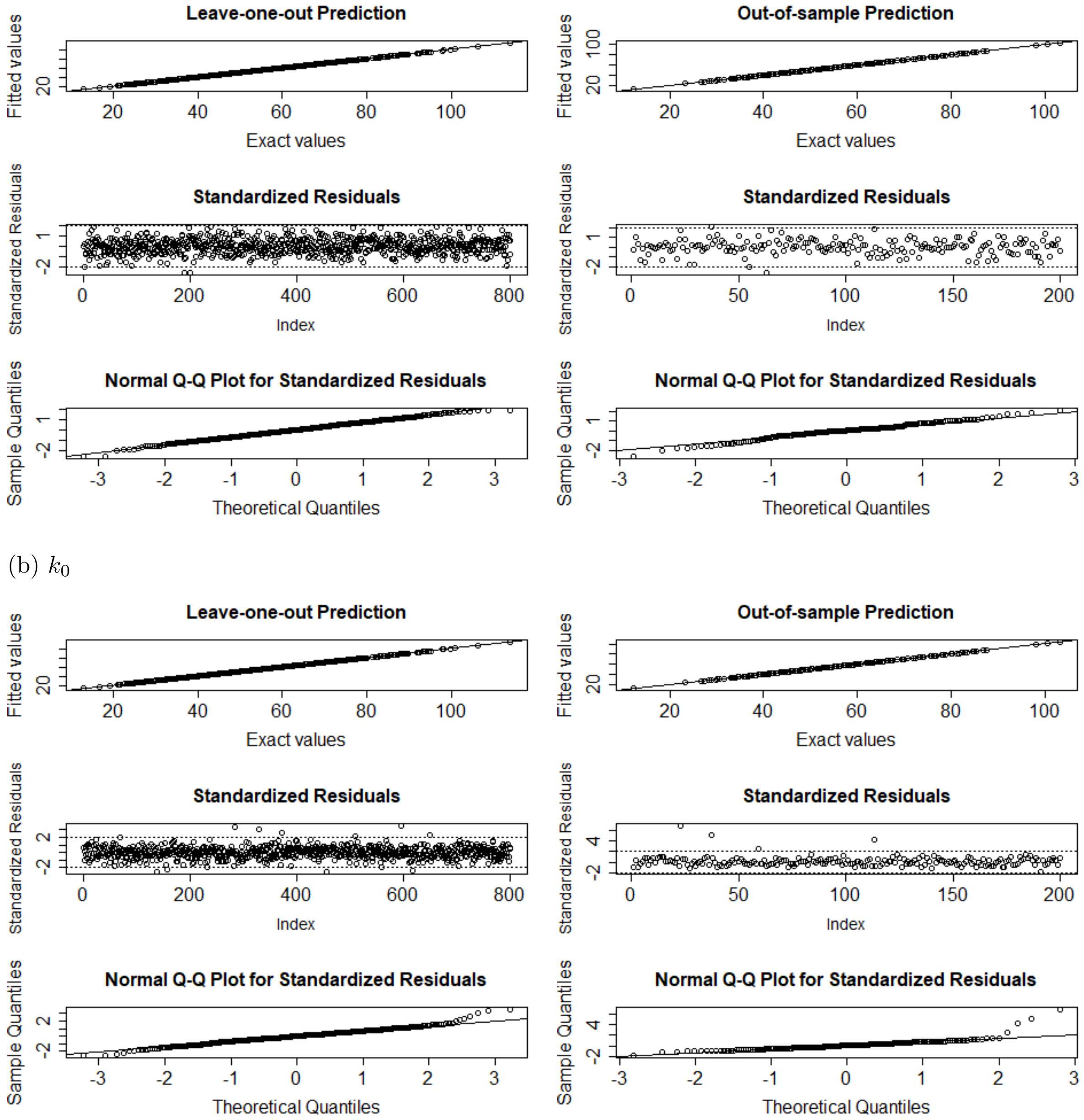
(b) k_0

Figure D.2.12: Residual analysis on **MAX with (80:20)**, (a) k_{DE} and (b) k_0


(a) k_{DE}

(b) k_0

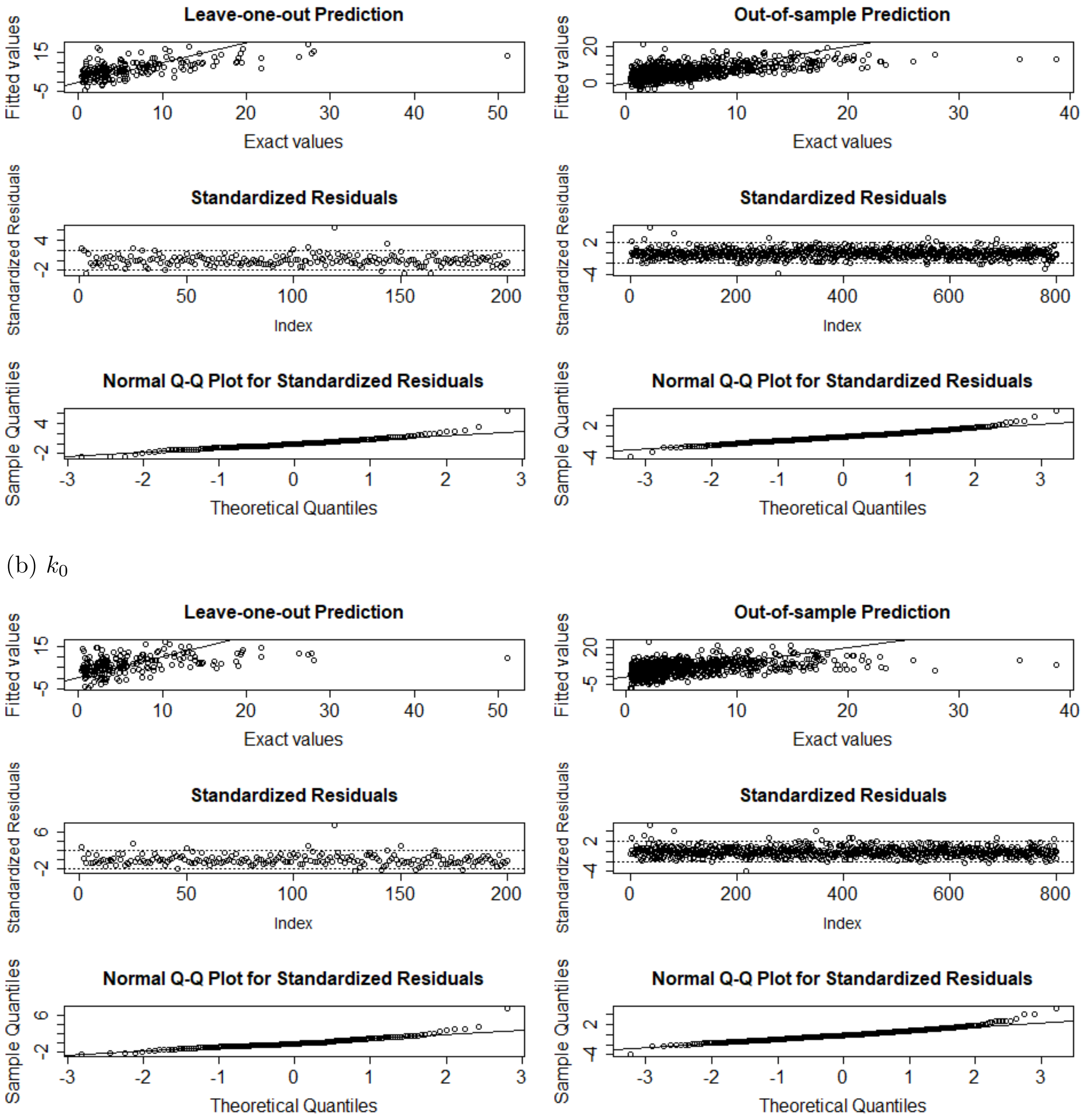
Figure D.2.13: Residual analysis on **MEAN with (20:80)**, (a) k_{DE} and (b) k_0



(a) k_{DE}

(b) k_0

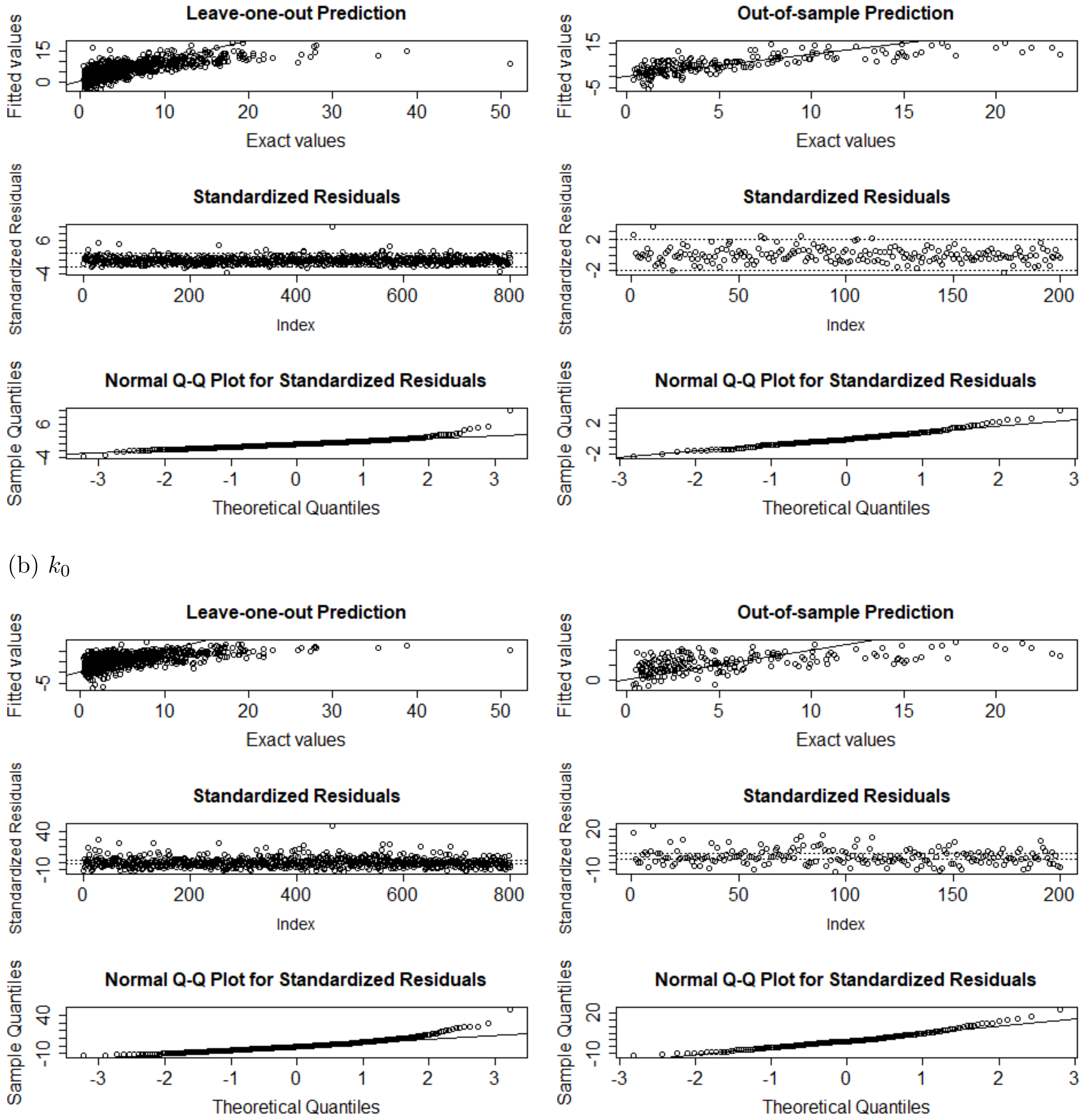
Figure D.2.14: Residual analysis on **MEAN with (80:20)**, (a) k_{DE} and (b) k_0



(a) k_{DE}

(b) k_0

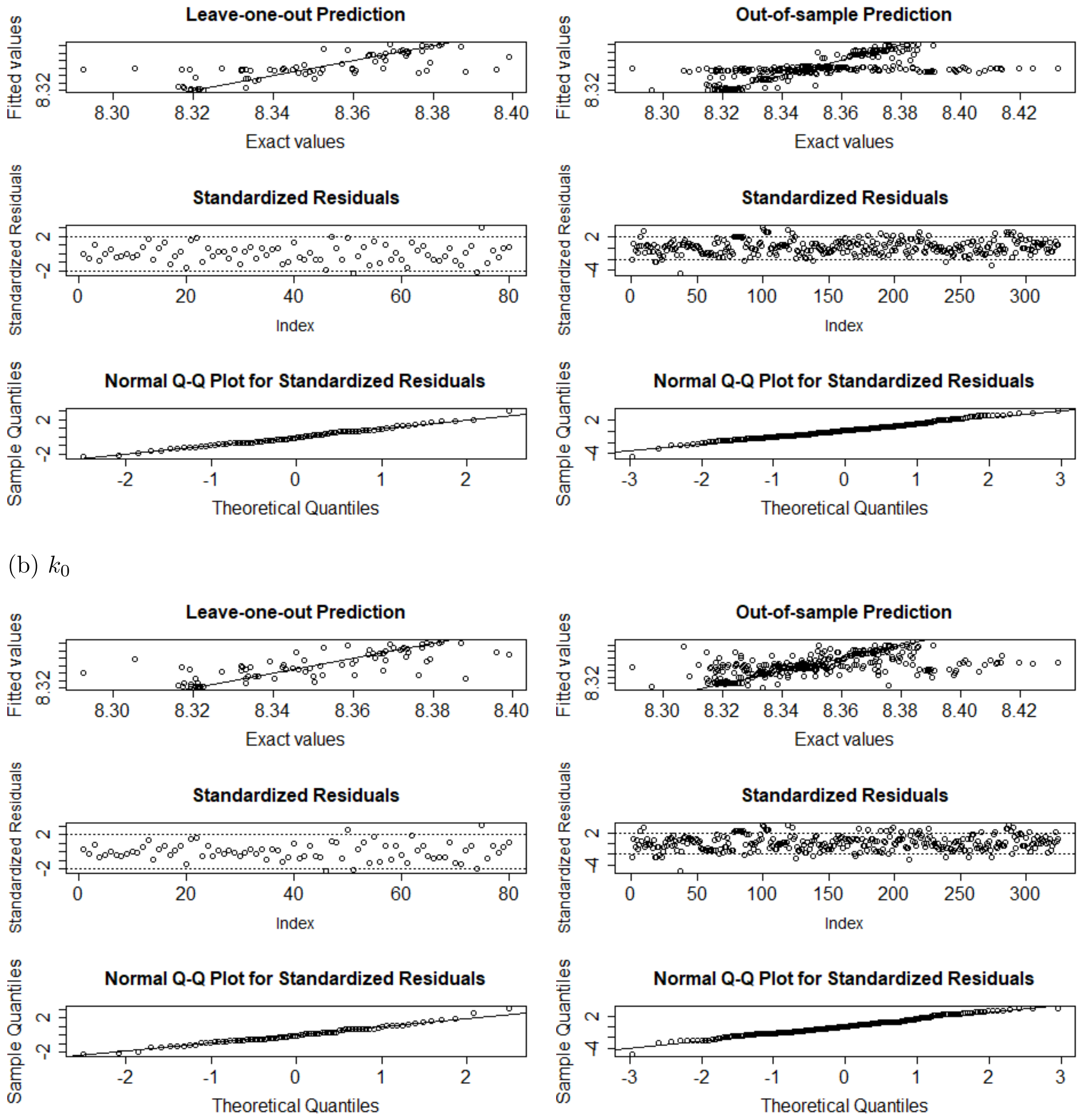
Figure D.2.15: Residual analysis on **MIN with (20:80)**, (a) k_{DE} and (b) k_0



(a) k_{DE}

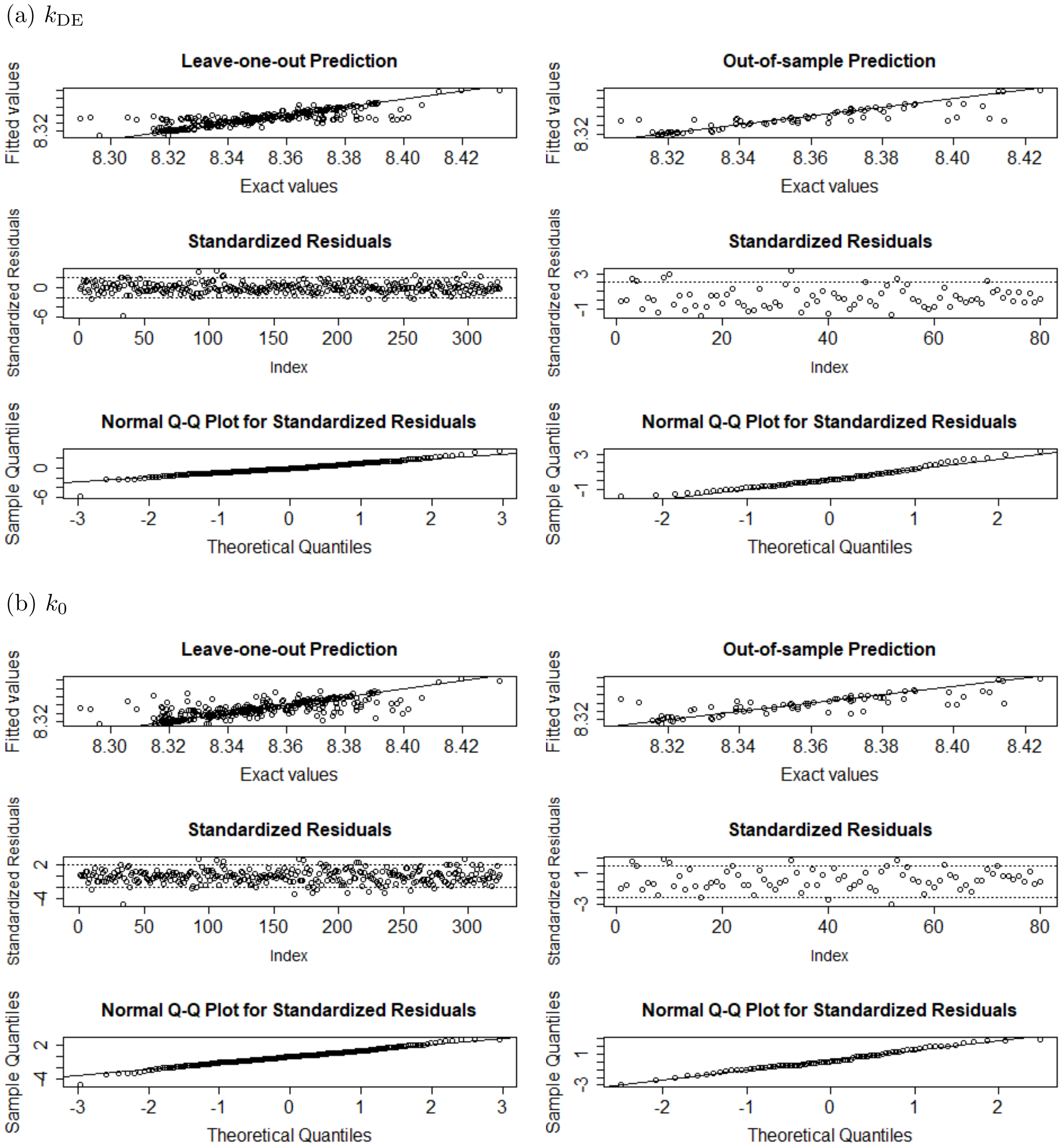
(b) k_0

Figure D.2.16: Residual analysis on **MIN with (80:20)**, (a) k_{DE} and (b) k_0


(a) k_{DE}

(b) k_0

Figure D.2.17: Residual analysis on **CASTEM with (20:80)**, (a) k_{DE} and (b) k_0



(a) k_{DE}

(b) k_0

Figure D.2.18: Residual analysis on **CASTEM with (80:20)**, (a) k_{DE} and (b) k_0

References

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics.* Kluwer Academic Publishers.

Christmann, A. and Steinwart, I. (2010). Universal kernels on non-standard input spaces. In *Advances in neural information processing systems*, pages 406–414.

Cuturi, M., Fukumizu, K., and Vert, J. (2005). Semigroup kernels on measures. *Journal of Machine Learning Research*, 6:1169–1198.

Gärtner, T., Lloyd, J., and Flach, P. A. (2004). Kernels and distances for structured data. *Machine Learning*, 57.

Haussler, D. (1999). Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz, Department of Computer Science.

Kim, J., McCourt, M., You, T., Kim, S., and Choi, S. (2019). Bayesian optimization over sets. In *6th ICML Workshop on Automated Machine Learning.* arXiv:1905.09780.

Kondor, R. and Jebara, T. (2003). A kernel between sets of vectors. In *Proceedings of the Twentieth International Conference on Machine Learning.*

Mebane Jr, W. R., Sekhon, J. S., et al. (2011). Genetic optimization using derivatives: the rgenoud package for r. *Journal of Statistical Software*, 42(11):1–26.

Muandet, K., Fukumizu, K., and B., S. (2017). Kernel mean embedding of distributions : A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141.

Ranjan, P., Haynes, R., and Karsten, R. (2011). A computationally stable approach to gaussian process interpolation of deterministic computer simulation data. *Technometrics*, 53(4):366–378.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012). Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018). Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 1.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In *Algorithmic Learning Theory: 18th International Conference*, page 13–31. Springer.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, (12):2389–2410.

Sutherland, D. (2016). *Scalable, Flexible and Active Learning on Distributions.* PhD thesis.