Stat Comput (2010) 20: 177–189
DOI 10.1007/s11222-009-9124-0

Least squares and shrinkage estimation under bimonotonicity constraints

Rudolf Beran · Lutz Dümbgen

Received: 14 September 2008 / Accepted: 30 March 2009 / Published online: 24 April 2009

Abstract In this paper we describe active set type algorithms for minimization of a smooth function under general order constraints, an important case being functions on the set of bimonotone $r \times s$ matrices. These algorithms can be used, for instance, to estimate a bimonotone regression function via least squares or (a smooth approximation of) least absolute deviations. Another application is shrinkage estimation in image denoising or, more generally, regression problems with two ordinal factors after representing the data in a suitable basis which is indexed by pairs $(i, j) \in \{1, \dots, r\} \times \{1, \dots, s\}$. Various numerical examples illustrate our methods.

Keywords Active set algorithm · Dynamic programming · Estimated risk · Pool-adjacent-violators algorithm · Regularization

1 Introduction

Monotonicity and other qualitative constraints play an important role in contemporary nonparametric statistics. One reason for this success is that such constraints are often plausible or even justified theoretically, within an appropriate mathematical formulation of the application. Moreover, by imposing shape constraints one can often avoid more traditional smoothness assumptions which typically lead to procedures requiring the choice of some tuning parameter. A good starting point for statistical inference under qualitative constraints is the monograph by Robertson et al. (1988).

Estimation under order constraints leads often to the following optimization problem: For some dimension $p \geq 2$ let $Q : \mathbb{R}^p \to \mathbb{R}$ be a given functional. For instance,

$$Q(\boldsymbol{\theta}) = \sum_{u=1}^{p} w_u (Z_u - \theta_u)^2 \tag{1}$$

with a certain weight vector $\boldsymbol{w} \in (0, \infty)^p$ and a given data vector $\boldsymbol{Z} \in \mathbb{R}^p$. In general we assume that Q is continuously differentiable, strictly convex and coercive, i.e.

$$Q(\boldsymbol{\theta}) \to \infty \quad \text{as } \|\boldsymbol{\theta}\| \to \infty,$$

where $\|\cdot\|$ is some norm on $\mathbb{R}^p$. The goal is to minimize Q over the following subset $\mathbb{K}$ of $\mathbb{R}^p$: Let $\mathcal{C}$ be a given collection of pairs (u, v) of different indices $u, v \in \{1, 2, \dots, p\}$, and define

$$\mathbb{K} = \mathbb{K}(\mathcal{C}) = \left\{\boldsymbol{\theta} \in \mathbb{R}^p : \theta_u \leq \theta_v \text{ for all } (u, v) \in \mathcal{C}\right\}.$$

This defines a closed convex cone in $\mathbb{R}^p$ containing all constant vectors.

For instance, if $\mathcal{C}$ consists of $(1, 2), (2, 3), \dots, (p-1, p)$, then $\mathbb{K}$ is the cone of all vectors $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_p$. Minimizing (1) over all such vectors is a standard problem and can be solved in $O(p)$ steps via the pool-adjacent-violators algorithm (PAVA). The latter was introduced in a special setting by Ayer et al. (1955) and extended later by numerous authors, see Robertson et al. (1988) and Best and Chakravarti (1990).

As soon as $Q(\cdot)$ is not of type (1) or $\mathcal{C}$ differs from the aforementioned standard example, the minimization of $Q(\cdot)$ over $\mathbb{K}$ becomes more involved. Here is another example for

R. Beran
University of California, Davis, CA, USA

L. Dümbgen (✉)
University of Bern, Bern, Switzerland
e-mail: duembgen@stat.unibe.ch

$\mathbb{K}$ and $\mathcal{C}$ which is of primary interest in the present paper: Let $p = rs$ with integers $r, s \geq 2$, and identify $\mathbb{R}^p$ with the set $\mathbb{R}^{r\times s}$ of all matrices with r rows and s columns. Further let $\mathbb{K}_{r,s}$ be the set of all matrices $\boldsymbol{\theta} \in \mathbb{R}^{r\times s}$ such that

$$\theta_{i,j} \leq \theta_{i+1,j} \quad \text{whenever } i < r \quad \text{and}$$
$$\theta_{i,j} \leq \theta_{i,j+1} \quad \text{whenever } j < s.$$

This corresponds to the set $\mathcal{C}_{r,s}$ of all pairs $((i,j),(k,\ell))$ with $i,k \in \{1,\dots,r\}$ and $j,\ell \in \{1,\dots,s\}$ such that either $(k,\ell) = (i+1,j)$ or $(k,\ell) = (i,j+1)$. Hence there are $\#\mathcal{C} = 2rs - r - s$ constraints.

Minimizing the special functional (1), i.e. $Q(\boldsymbol{\theta}) = \sum_{i,j} w_{ij}(Z_{ij} - \theta_{ij})^2$, over the bimonotone cone $\mathbb{K}_{r,s}$ is a well recognized problem with various proposed solutions, see, for instance, Spouge et al. (2003), Burdakow et al. (2004), and the references cited therein. However, all these algorithms exploit the special structure of $\mathbb{K}_{r,s}$ or (1). For general functionals $Q(\cdot)$, e.g. quadratic functions with positive definite but non-diagonal Hessian matrix, different approaches are needed.

The remainder of this paper is organized as follows. In Sect. 2 we describe the *bimonotone regression* problem and argue that the special structure (1) is sometimes too restrictive even in that context. In Sect. 3 we derive possible algorithms for the general optimization problem described above. These algorithms involve a discrete optimization step which gives rise to a dynamic program in case of $\mathbb{K} = \mathbb{K}_{r,s}$. For a general introduction to dynamic programming see Cormen et al. (1990). Other ingredients are active methods as described by, for instance, Fletcher (1987), Best and Chakravarti (1990) or Dümbgen et al. (2007), sometimes combined with the ordinary PAVA in a particular fashion. It will be shown that all these algorithms find the exact solution in finitely many steps, at least when $Q(\cdot)$ is an arbitrary quadratic and strictly convex function. Finally, in Sect. 4 we adapt our procedure to image denoising via *bimonotone shrinkage* of generalized Fourier coefficients. The statistical method in this section was already indicated in Beran and Dümbgen (1998) but has not been implemented yet, for lack of an efficient computational algorithm.

2 Least squares estimation of bimonotone regression functions

Suppose that one observes (x^1, y^1, Z^1), (x^2, y^2, Z^2), …, (x^n, y^n, Z^n) with real components x^t, y^t and Z^t. The points (x^t, y^t) are regarded as fixed points, which is always possible by conditioning, while

$$Z^t = \mu(x^t, y^t) + \varepsilon^t$$

for an unknown regression function $\mu : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and independent random errors $\varepsilon^1, \varepsilon^2, \dots, \varepsilon^n$ with mean zero. In some applications it is plausible to assume μ to be bimonotone increasing, i.e. non-decreasing in both arguments. Then it would be desirable to estimate μ under that constraint only. One possibility would be to minimize

$$\sum_{t=1}^{n} (Z^t - \mu(x^t, y^t))^2$$

over all bimonotone functions μ. The resulting minimizer $\hat{\mu}$ is uniquely defined on the finite set of all design points (x^t, y^t), $1 \leq t \leq n$.

For a more detailed discussion, suppose that we want to estimate μ on a finite rectangular grid

$$\big\{(x_{(i)}, y_{(j)}) : 1 \leq i \leq r, 1 \leq j \leq s\big\},$$

where $x_{(1)} < x_{(2)} < \cdots < x_{(r)}$ and $y_{(1)} < y_{(2)} < \cdots < y_{(s)}$ contain at least the different elements of $\{x^1, x^2, \dots, x^n\}$ and $\{y^1, y^2, \dots, y^n\}$, respectively, but maybe additional points as well. For $1 \leq i \leq r$ and $1 \leq j \leq s$ let w_{ij} be the number of all $t \in \{1,\dots,n\}$ such that $(x^t, y^t) = (x_{(i)}, y_{(j)})$, and let Z_{ij} be the average of Z^t over these indices t. Then $\sum_{t=1}^{n}(Z^t - \mu(x^t, y^t))^2$ equals

$$Q(\boldsymbol{\theta}) = \sum_{i,j} w_{ij}(Z_{ij} - \theta_{ij})^2,$$

where $\boldsymbol{\theta} = (\theta_{ij})_{i,j}$ stands for the matrix $(\mu(x_{(i)}, y_{(j)}))_{i,j} \in \mathbb{K}_{r,s}$.

Setting 1: *Complete layout* Suppose that $w_{ij} > 0$ for all $(i,j) \in \{1,\dots,r\} \times \{1,\dots,s\}$. Then the resulting optimization problem is precisely the one described in the introduction.

Setting 2a: *Incomplete layout and simple interpolation/extrapolation* Suppose that the set $\mathcal{U}$ of all index pairs (i,j) with $w_{ij} > 0$ differs from $\{1,\dots,r\} \times \{1,\dots,s\}$. Then

$$Q(\boldsymbol{\theta}) = \sum_{u \in \mathcal{U}} w_u(Z_u - \theta_u)^2$$

fails to be coercive. Nevertheless it can be minimized over $\mathbb{K}_{r,s}$ with the algorithms described later. Let $\check{\boldsymbol{\theta}}$ be such a minimizer. Since it is uniquely defined on $\mathcal{U}$ only, we propose to replace it with $\hat{\boldsymbol{\theta}} = 2^{-1}(\underline{\boldsymbol{\theta}} + \overline{\boldsymbol{\theta}})$, where

$$\underline{\theta}_{ij} = \max\Big(\big\{\check{\theta}_{i'j'} : (i',j') \in \mathcal{U}, i' \leq i, j' \leq j\big\} \cup \{\check{\theta}_{\min}\}\Big),$$
$$\overline{\theta}_{ij} = \min\Big(\big\{\check{\theta}_{i'j'} : (i',j') \in \mathcal{U}, i \leq i', j \leq j'\big\} \cup \{\check{\theta}_{\max}\}\Big),$$

and $\check{\theta}_{\min}$ and $\check{\theta}_{\max}$ denote the minimum and maximum, respectively, of $\{\check{\theta}_u : u \in \mathcal{U}\}$. Note that $\underline{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\theta}}$ belong to

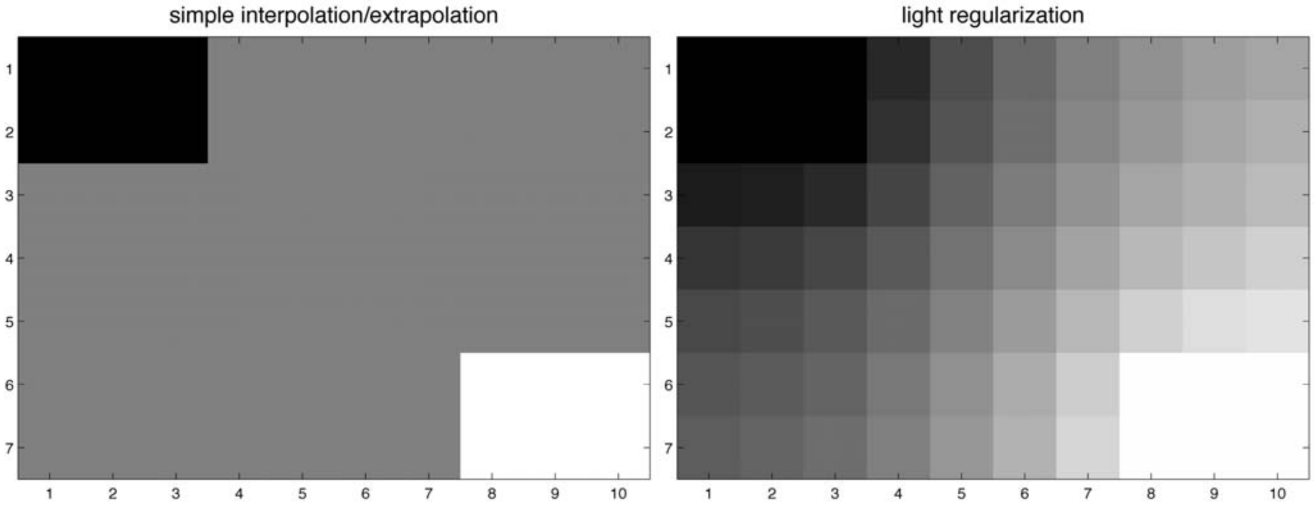

Fig. 1 Simple interpolation/extrapolation versus light regularization

$\mathbb{K}_{r,s}$ and are extremal in the sense that any matrix $\boldsymbol{\theta} \in \mathbb{K}_{r,s} \cap [\check{\theta}_{\min}, \check{\theta}_{\max}]^{r\times s}$ with $\theta_u = \check{\theta}_u$ for all $u \in \mathcal{U}$ satisfies necessarily $\underline{\theta}_{ij} \le \theta_{ij} \le \overline{\theta}_{ij}$ for all (i, j).

Setting 2b: *Incomplete layout and light regularization* Instead of restricting attention to the index set $\mathcal{U}$, one can estimate the full matrix $(\mu(x_{(i)}, y_{(j)}))_{i,j} \in \mathbb{R}^{r\times s}$ by minimizing a suitably penalized sum of squares,

$$Q(\boldsymbol{\theta}) = \sum_{u\in\mathcal{U}} w_u(Z_u - \theta_u)^2 + \lambda P(\boldsymbol{\theta}),$$

over $\mathbb{K}_{r,s}$ for some small parameter $\lambda > 0$. Here $P(\cdot)$ is a convex quadratic function on $\mathbb{R}^{r\times s}$ such that $Q(\cdot)$ is strictly convex. One possibility would be Tychonov regularisation with $P(\boldsymbol{\theta}) = \sum_{i,j}(\theta_{ij} - \theta_o)^2$ and a certain reference value θ_o, for instance, $\theta_o = \sum_{i,j} w_{ij} Z_{ij} / \sum_{i,j} w_{ij}$. In our particular setting we prefer the penalty

$$P(\boldsymbol{\theta}) = \sum_{((i,j),(k,\ell))\in\mathcal{C}_{r,s}} (\theta_{k\ell} - \theta_{ij})^2, \tag{2}$$

because it yields smoother interpolations than the recipe for Setting 2a or the Tychonov penalty. One can easily show that the resulting quadratic function Q is strictly convex but with non-diagonal Hessian matrix. Thus it fulfills our general requirements but is not of type (1).

Note that adding a penalty term such as (2) could be worthwhile even in case of a complete layout if the underlying function μ is assumed to be smooth. But this leads to the nontrivial task of choosing $\lambda > 0$ appropriately. Here we use the penalty term mainly for smooth interpolation/extrapolation with λ just large enough to ensure a well-conditioned Hessian matrix. We refer to this as “light regularization”, and the exact value of λ is essentially irrelevant.

Example 2.1 To illustrate the difference between simple interpolation/extrapolation and light regularization with penalty (2) we consider just two observations, $(x^1, y^1, Z^1) = (2, 3, 0)$ and $(x^2, y^2, Z^2) = (6, 8, 1)$, and let $r = 7$, $s = 10$ with $x_{(i)} = i$ and $y_{(j)} = j$. Thus $w_{ij} = 0$ except for $w_{2,3} = w_{6,8} = 1$, while $Z_{2,3} = 0$ and $Z_{6,8} = 1$. Any minimizer $\check{\boldsymbol{\theta}}$ of $\sum_{u\in\mathcal{U}} w_u(Z_u - \theta_u)^2$ over $\mathbb{K}_{7,10}$ satisfies $\check{\theta}_{2,3} = 0$ and $\check{\theta}_{6,8} = 1$, so the recipe for Setting 2a yields

$$\hat{\theta}_{ij} = \begin{cases} 0, & \text{if } i \le 2,\ j \le 3, \\ 1, & \text{if } i \ge 6,\ j \ge 8, \\ 0.5, & \text{else.} \end{cases}$$

The left panel of Fig. 1 shows the latter fit $\hat{\boldsymbol{\theta}}$, while the right panel shows the regularized fit based on (2) with $\lambda = 10^{-4}$. In these and most subsequent pictures we use a gray scale from black = 0 to white = 1.

Example 2.2 (Binary regression) We generated a random matrix $\boldsymbol{Z} \in \{0, 1\}^{r\times s}$ with $r = 70$ rows, $s = 100$ columns and independent components Z_{ij}, where

$$\begin{aligned}\Pr(Z_{ij} = 1) &= \theta_{ij} \\ &= \frac{x_{(i)} + y_{(j)}}{4} + \frac{1\{y_{(j)} \ge 1/2 + \cos(\pi x_{(i)})/4\}}{2}\end{aligned}$$

with $x_{(i)} = (i - 0.5)/r$ and $y_{(j)} = (j - 0.5)/s$. Thereafter we removed randomly all but 700 of the 7000 components Z_{ij}. The resulting data are depicted in the upper left panel of Fig. 2, where missing values are depicted grey, while the upper right panel shows the true signal $\boldsymbol{\theta}$. The lower panels depict the least squares estimator with simple interpolation/extrapolation (left) and light regularization based on (2) with $\lambda = 10^{-4}$ (right). Note that both estimators are very

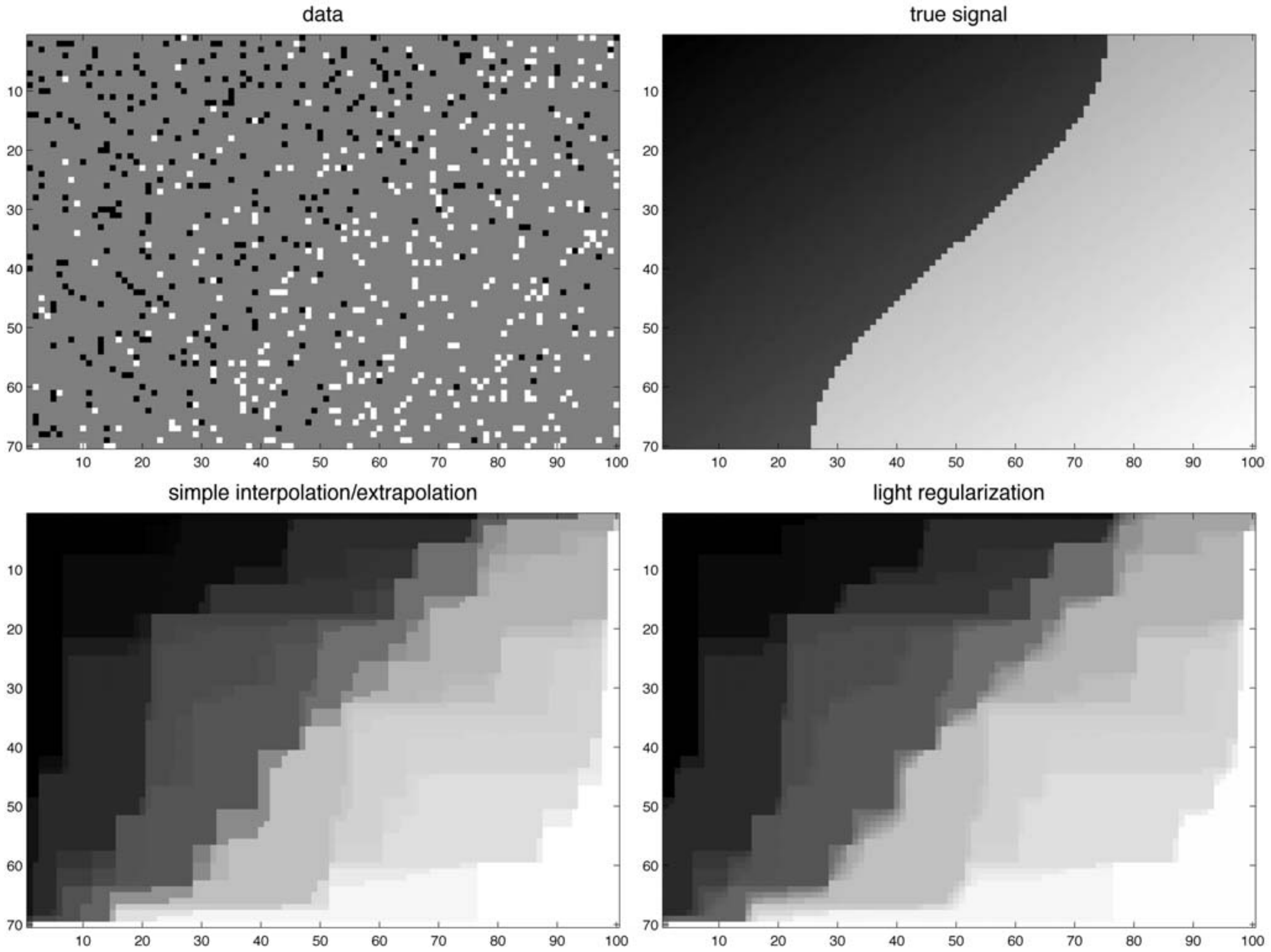

Fig. 2 Binary regression with incomplete layout

similar. Due to the small value of λ, the main differences occur in regions without data points.

The quality of an estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ may be quantified by the average absolute deviation,

$$\mathrm{AAD} = \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}|\hat{\theta}_{ij} - \theta_{ij}|.$$

For the estimator with simple interpolation/extrapolation, AAD turned out to be 7.5607×10^{-2}, the estimator based on light regularization performed slightly better with $\mathrm{AAD} = 7.4039 \times 10^{-2}$.

3 The general algorithmic problem

We return to the general framework introduced in the beginning with a continuously differentiable, strictly convex and coercive functional $Q : \mathbb{R}^p \to \mathbb{R}$ and a closed convex cone $\mathbb{K} = \mathbb{K}(\mathcal{C}) \in \mathbb{R}^p$ determined by a collection $\mathcal{C}$ of inequality constraints.

Before starting with explicit algorithms, let us characterize the point

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{K}} Q(\boldsymbol{\theta}).$$

It is well-known from convex analysis that a point $\boldsymbol{\theta} \in \mathbb{K}$ coincides with $\boldsymbol{\theta}$ if, and only if,

$$\nabla Q(\boldsymbol{\theta})^\top \boldsymbol{\theta} = 0 \le \nabla Q(\boldsymbol{\theta})^\top \boldsymbol{\eta} \quad \text{for all } \boldsymbol{\eta} \in \mathbb{K}, \tag{3}$$

where $\nabla Q(\boldsymbol{\theta})$ denotes the gradient of Q at $\boldsymbol{\theta}$. This characterization involves infinitely many inequalities, but it can be replaced with a criterion involving only finitely many constraints.

3.1 Extremal directions of $\mathbb{K}$

Note that $\mathbb{K}$ contains all constant vectors $c\mathbf{1}$, $c \in \mathbb{R}$, where $\mathbf{1} = \mathbf{1}_p = (1)_{i=1}^{p}$. It can be represented as follows:

Lemma 3.1 *Define*

$\mathcal{E} = \mathbb{K} \cap \{0,1\}^p.$

Then any vector $\boldsymbol{x} \in \mathbb{K}$ *may be represented as*

$$\boldsymbol{x} \;=\; \min(\boldsymbol{x})\mathbf{1} + \sum_{\boldsymbol{e}\in\mathcal{E}} \lambda_{\boldsymbol{e}}\boldsymbol{e}$$

with coefficients $\lambda_{\boldsymbol{e}} \geq 0$ *such that* $\sum_{\boldsymbol{e}\in\mathcal{E}} \lambda_{\boldsymbol{e}} = \max(\boldsymbol{x}) - \min(\boldsymbol{x})$.

Here $\min(\boldsymbol{x})$ and $\max(\boldsymbol{x})$ denote the minimum and maximum, respectively, of the components of $\boldsymbol{x}$.

Modified characterization of $\hat{\boldsymbol{\theta}}$ By means of Lemma 3.1 one can easily verify that (3) is equivalent to the following condition:

$$\nabla Q(\boldsymbol{\theta})^\top \boldsymbol{\theta} = 0 \leq \nabla Q(\boldsymbol{\theta})^\top \boldsymbol{e} \quad \text{for all } \boldsymbol{e} \in \mathcal{E} \cup \{-\mathbf{1}\}. \tag{4}$$

Thus we have to check only finitely many constraints. Note, however, that the cardinality of $\mathcal{E}$ may be substantially larger than the dimension p, so that checking (4) is far from trivial.

Application to $\mathbb{K}_{r,s}$ Applying Lemma 3.1 to the cone $\mathbb{K}_{r,s} \subset \mathbb{R}^{r\times s}$ yields the following representation: With

$$\mathcal{E}_{r,s} \;=\; \mathbb{K}_{r,s} \cap \{0,1\}^{r\times s}$$

any matrix $\boldsymbol{x} \in \mathbb{K}$ may be written as

$$\boldsymbol{x} \;=\; a_o \mathbf{1}_{r\times s} + \sum_{\boldsymbol{e}\in\mathcal{E}_{r,s}} \lambda_{\boldsymbol{e}}\boldsymbol{e}$$

with coefficients $a_o \in \mathbb{R}$ and $\lambda_{\boldsymbol{e}} \geq 0$, $\boldsymbol{e} \in \mathcal{E}_{r,s}$.

There is a one-to-one correspondence between the set $\mathcal{E}_{r,s}$ and the set of all vectors $\tilde{\boldsymbol{e}} \in \{1, 2, \dots, r+s\}^r$ with components $\tilde{e}_1 < \tilde{e}_2 < \cdots < \tilde{e}_r$ via the mapping

$$\boldsymbol{e} \mapsto \left(i + \sum_{j=1}^{s} e_{ij} \right)_{i=1}^{r}.$$

Since such a vector $\tilde{\boldsymbol{e}}$ corresponds to a subset of $\{1, 2, \dots, r+s\}$ with r elements, we end up with

$$\#\mathcal{E}_{r,s} = \binom{r+s}{r} = \binom{r+s}{s}.$$

Hence the cardinality of $\mathcal{E}_{r,s}$ grows exponentially in $\min(r, s)$. Nevertheless, minimizing a linear functional over $\mathcal{E}_{r,s}$ is possible in $O(rs)$ steps, as explained in the next section.

Proof of Lemma 3.1 For $\boldsymbol{x} \in \mathbb{K}$ let $a_0 < a_1 < \cdots < a_m$ be the different elements of $\{x_1, x_2, \dots, x_p\}$, i.e. $a_0 = \min(\boldsymbol{x})$ and $a_m = \max(\boldsymbol{x})$. Then

$$\boldsymbol{x} \;=\; a_0\mathbf{1} + \sum_{i=1}^{m} (a_i - a_{i-1}) \big(1\{x_t \geq a_i\}\big)_{t=1}^{p}.$$

Obviously, these weights $a_i - a_{i-1}$ are nonnegative and sum to $\max(\boldsymbol{x}) - \min(\boldsymbol{x})$. Furthermore, one can easily deduce from $\boldsymbol{x} \in \mathbb{K}$ that $(1\{x_t \geq a\})_{t=1}^{p}$ belongs to $\mathcal{E}$ for any real threshold a. □

3.2 A dynamic program for $\boldsymbol{\mathcal{E}_{r,s}}$

For some matrix $\boldsymbol{a} \in \mathbb{R}^{r\times s}$ let $L : \mathbb{R}^{r\times s} \to \mathbb{R}$ be given by

$$L(\boldsymbol{x}) = \sum_{i=1}^{r}\sum_{j=1}^{s} a_{ij}x_{ij}.$$

The minimum of $L(\cdot)$ over the finite set $\mathcal{E}_{r,s}$ may be obtained by means of the following recursion: For $1 \leq k \leq r$ and $1 \leq \ell \leq s$ define

$$H(k,\ell) = \min\left\{ \sum_{i=k}^{r}\sum_{j=1}^{s} a_{ij}e_{ij} : \boldsymbol{e} \in \mathcal{E}_{r,s}, e_{k\ell} = 1 \right\},$$

$$H(k, s+1) = \min\left\{ \sum_{i=k}^{r}\sum_{j=1}^{s} a_{ij}e_{ij} : \boldsymbol{e} \in \mathcal{E}_{r,s} \right\}.$$

Then

$$\min_{\boldsymbol{e}\in\mathcal{E}_{r,s}} L(\boldsymbol{e}) = H(1, s+1),$$

and

$$H(k,1) = \sum_{i=k}^{r}\sum_{j=1}^{s} a_{ij},$$

$$H(k,\ell+1) = \min\left(H(k,\ell), \sum_{j=\ell+1}^{s} a_{ij} + H(k+1,\ell+1) \right)$$

where we use the conventions that $H(k+1,\cdot) = 0$ and $\sum_{j=s+1}^{s} \cdot = 0$. In the recursion formula for $H(k,\ell+1)$, the term $\sum_{j=\ell+1}^{s} a_{ij} + H(k+1,\ell+1)$ is the minimum of $L_k(\boldsymbol{e}) = \sum_{i=k}^{r}\sum_{j=1}^{s} a_{ij}e_{ij}$ over all matrices $\boldsymbol{e} \in \mathcal{E}_{r,s}$ with $e_{k\ell} = 0$ and $e_{k,\ell+1} = 1$ (if $\ell < s$), while $H(k,\ell)$ is the minimum of $L_k(\boldsymbol{e})$ over all $\boldsymbol{e} \in \mathcal{E}_{k,s}$ with $e_{k\ell} = 1$.

Table 1 provides pseudocode for an algorithm that determines a minimizer of $L(\cdot)$ over $\mathcal{E}_{r,s}$.

3.3 Active set type algorithms

Throughout this exposition we assume that minimization of Q over an affine linear subspace of $\mathbb{R}^p$ is feasible. This is certainly the case if Q is a quadratic functional. If Q is twice

Table 1 Minimizing a linear functional over $\mathcal{E}_{r,s}$

```
Algorithm $e \leftarrow$ DynamicProgram($a$)
$b \leftarrow (\sum_{j=\ell}^{s} a_{k,j})_{k\le r,\ell\le s+1}$
$H \leftarrow (0)_{k\le r+1,\ell\le s+1}$
for $k \leftarrow r$ downto $1$ do
  $H_{k,1} \leftarrow H_{k+1,1} + b_{k,1}$
  for $\ell \leftarrow 1$ to $s$ do
    $H_{k,\ell+1} \leftarrow \min(H_{k,\ell}, b_{k,\ell+1} + H_{k+1,\ell+1})$
  end for
end for
$e \leftarrow (0)_{k\le r,\ell\le s}$
$k \leftarrow 1$, $\ell \leftarrow s$
while $k \le r$ and $\ell \ge 1$ do
  if $H_{k,\ell+1} = H_{k,\ell}$ then
    $(e_{i,\ell})_{i=k}^{r} \leftarrow (1)_{i=k}^{r}$
    $\ell \leftarrow \ell - 1$
  else
    $k \leftarrow k + 1$
  end if
end while.
```

continuously differentiable with positive definite Hessian matrix everywhere, this minimization problem can be solved with arbitrarily high accuracy by a Newton type algorithm.

All algorithms described in this paper alternate between two basic procedures which are described next. In both procedures $\boldsymbol{\theta} \in \mathbb{K}$ is replaced with a vector $\boldsymbol{\theta}_{\text{new}} \in \mathbb{K}$ such that $Q(\boldsymbol{\theta}_{\text{new}}) < Q(\boldsymbol{\theta})$ unless $\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta}$.

Basic procedure 1: *Checking optimality of $\boldsymbol{\theta} \in \mathbb{K}$* Suppose that $\boldsymbol{\theta} \in \mathbb{K}$ satisfies already the following two equations:

$$\nabla Q(\boldsymbol{\theta})^\top \boldsymbol{\theta} = 0 = \nabla Q(\boldsymbol{\theta})^\top \mathbf{1}. \tag{5}$$

According to (3), this vector is already the solution $\hat{\boldsymbol{\theta}}$ if, and only if, $\nabla Q(\boldsymbol{\theta})^\top \boldsymbol{e} \ge 0$ for all $\boldsymbol{e} \in \mathcal{E}$. Thus we determine

$$\Delta \in \underset{\boldsymbol{e}\in\mathcal{E}}{\text{argmin}}\ \nabla Q(\boldsymbol{\theta})^\top \boldsymbol{e}$$

and do the following: If $\nabla Q(\boldsymbol{\theta})^\top \Delta \ge 0$, we know that $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and stop the algorithm. Otherwise we determine

$$t_o = \underset{t\in\mathbb{R}}{\text{argmin}}\ Q(\boldsymbol{\theta} + t\Delta) > 0$$

and replace $\boldsymbol{\theta}$ with

$$\boldsymbol{\theta}_{\text{new}} := \boldsymbol{\theta} + t_o\Delta.$$

This vector $\boldsymbol{\theta}_{\text{new}}$ lies in the cone $\mathbb{K}$, too, and satisfies the inequality $Q(\boldsymbol{\theta}_{\text{new}}) < Q(\boldsymbol{\theta})$. Then we proceed with basic procedure 2.

Basic procedure 2: *Replacing $\boldsymbol{\theta} \in \mathbb{K}$ with a "locally optimal" point $\boldsymbol{\theta}_{\text{new}} \in \mathbb{K}$* The general idea of basic procedure 2 is to find a point $\boldsymbol{\theta}_{\text{new}} \in \mathbb{K}$ such that

$$\boldsymbol{\theta}_{\text{new}} = \underset{\boldsymbol{x}\in\mathbb{V}}{\text{argmin}}\ Q(\boldsymbol{x}) \tag{6}$$

for some $\mathbb{V}$ in a finite family $\mathcal{V}$ of linear subspaces of $\mathbb{R}^p$. Typically these subspaces $\mathbb{V}$ are obtained by replacing some inequality constraints from $\mathcal{C}$ with equality constraints and ignoring the remaining ones. This approach is described below as basic procedure 2a. But we shall see that it is potentially useful to modify this strategy; see basic procedures 2b and 2c.

Basic procedure 2a: *The classical active set approach* For $\boldsymbol{\theta} \in \mathbb{K}$ define

$$\mathbb{V}(\boldsymbol{\theta}) = \{\boldsymbol{x} \in \mathbb{R}^p : x_u = x_v \text{ for all } (u, v) \in \mathcal{C} \text{ with } \theta_u = \theta_v\}.$$

This is a linear subspace of $\mathbb{R}^p$ containing $\mathbf{1}$ and $\boldsymbol{\theta}$ which is determined by those constraints from $\mathcal{C}$ which are "active" in $\boldsymbol{\theta}$. It has the additional property that for any vector $\boldsymbol{x} \in \mathbb{V}(\boldsymbol{\theta})$,

$$\lambda(\boldsymbol{\theta}, \boldsymbol{x}) = \max\{t \in [0, 1] : (1 - t)\boldsymbol{\theta} + t\boldsymbol{x} \in \mathbb{K}\} > 0.$$

Precisely, $\lambda(\boldsymbol{\theta}, \boldsymbol{x}) = 1$ if $\boldsymbol{x} \in \mathbb{K}$, and otherwise,

$$\lambda(\boldsymbol{\theta}, \boldsymbol{x}) = \min_{(u,v)\in\mathcal{C}:x_u>x_v} \frac{\theta_v - \theta_u}{\theta_v - \theta_u - x_v + x_u}.$$

The key step in basic procedure 2a is to determine $\boldsymbol{x}_o = \text{argmin}_{\boldsymbol{x}\in\mathbb{V}(\boldsymbol{\theta})} Q(\boldsymbol{x})$ and $\lambda(\boldsymbol{\theta}, \boldsymbol{x}_o)$. If $\boldsymbol{x}_o \in \mathbb{K}$, which is equivalent to $\lambda(\boldsymbol{\theta}, \boldsymbol{x}_o) = 1$, we are done and return $\boldsymbol{\theta}_{\text{new}} = \boldsymbol{x}_o$. This vector satisfies (6) with $\mathbb{V} = \mathbb{V}(\boldsymbol{\theta})$ and $\mathbb{V} = \mathbb{V}(\boldsymbol{\theta}_{\text{new}})$. The latter fact follows simply from $\mathbb{V}(\boldsymbol{\theta}_{\text{new}}) \subset \mathbb{V}(\boldsymbol{\theta})$. If $\boldsymbol{x}_o \notin \mathbb{K}$, we repeat this key step with $\boldsymbol{\theta}_{\text{new}} = (1 - \lambda(\boldsymbol{\theta}, \boldsymbol{x}_o)\boldsymbol{\theta} + \lambda(\boldsymbol{\theta}, \boldsymbol{x}_o)\boldsymbol{x}_o$ in place of $\boldsymbol{\theta}$.

In both cases the key step yields a vector $\boldsymbol{\theta}_{\text{new}}$ satisfying $Q(\boldsymbol{\theta}_{\text{new}}) < Q(\boldsymbol{\theta})$, unless $\boldsymbol{x}_o = \boldsymbol{\theta}$. Moreover, if $\boldsymbol{x}_o \notin \mathbb{K}$, then the vector space $\mathbb{V}(\boldsymbol{\theta}_{\text{new}})$ is contained in $\mathbb{V}(\boldsymbol{\theta})$ with strictly smaller dimension, because at least one additional constraint from $\mathcal{C}$ becomes active. Hence after finitely many repetitions of the key step, we end up with a vector $\boldsymbol{\theta}_{\text{new}}$ satisfying (6) with $\mathbb{V} = \mathbb{V}(\boldsymbol{\theta}_{\text{new}})$. Table 2 provides pseudocode for basic procedure 2a.

Basic procedure 2b: *Working with complete orders* The determination and handling of the subspace $\mathbb{V}(\boldsymbol{\theta})$ in basic procedure 2a may be rather involved, in particular, when the set $\mathcal{C}$ consists of more than p constraints. One possibility to avoid this is to replace $\mathbb{V}(\theta)$ and $\mathbb{K}$ in the key step with the following subspace $\mathbb{V}^*(\boldsymbol{\theta})$ and cone $\mathbb{K}^*(\boldsymbol{\theta})$, respectively:

Table 2 Basic procedure 2a

```
Algorithm $\boldsymbol{\theta}_{\text{new}} \leftarrow$ BasicProcedure2a($\boldsymbol{\theta}$)
$\boldsymbol{\theta}_{\text{new}} \leftarrow \boldsymbol{\theta}$
$\boldsymbol{x}_o \leftarrow \operatorname{argmin}_{\boldsymbol{x}\in\mathbb{V}(\boldsymbol{\theta}_{\text{new}})} Q(\boldsymbol{x})$
$\lambda \leftarrow \lambda(\boldsymbol{\theta}_{\text{new}}, \boldsymbol{x}_o)$
while $\lambda < 1$ do
    $\boldsymbol{\theta}_{\text{new}} \leftarrow (1-\lambda)\boldsymbol{\theta}_{\text{new}} + \lambda\boldsymbol{x}_o$
    $\boldsymbol{x}_o \leftarrow \operatorname{argmin}_{\boldsymbol{x}\in\mathbb{V}(\boldsymbol{\theta}_{\text{new}})} Q(\boldsymbol{x})$
    $\lambda \leftarrow \lambda(\boldsymbol{\theta}_{\text{new}}, \boldsymbol{x}_o)$
end while
$\boldsymbol{\theta}_{\text{new}} \leftarrow \boldsymbol{x}_o$
```

$$\begin{aligned}
\mathbb{V}^*(\boldsymbol{\theta}) &= \big\{\boldsymbol{x} \in \mathbb{R}^p : \text{for all } u, v \in \{1, \ldots, p\}, \\
&\qquad x_u = x_v \text{ if } \theta_u = \theta_v\big\}, \\
\mathbb{K}^*(\boldsymbol{\theta}) &= \big\{\boldsymbol{x} \in \mathbb{R}^p : \text{for all } u, v \in \{1, \ldots, p\}, \\
&\qquad x_u \le x_v \text{ if } \theta_u \le \theta_v\big\}.
\end{aligned}$$

Note that $\mathbf{1}, \boldsymbol{\theta} \in \mathbb{K}^*(\boldsymbol{\theta}) \subset \mathbb{V}^*(\boldsymbol{\theta})$, and one easily verifies that $\mathbb{K}^*(\boldsymbol{\theta}) \subset \mathbb{K}$ if $\boldsymbol{\theta} \in \mathbb{K}$. Basic procedure 2b works precisely like basic procedure 2a, but with $\mathbb{V}^*(\cdot)$ in place of $\mathbb{V}(\cdot)$, and $\lambda(\boldsymbol{\theta}, \boldsymbol{x})$ is replaced with

$$\lambda^*(\boldsymbol{\theta}, \boldsymbol{x}) = \max\big\{t \in [0,1] : (1-t)\boldsymbol{\theta} + t\boldsymbol{x} \in \mathbb{K}^*(\boldsymbol{\theta})\big\}.$$

Then basic procedure 2b yields a vector $\boldsymbol{\theta}_{\text{new}}$ satisfying (6) with $\mathbb{V} = \mathbb{V}^*(\boldsymbol{\theta}_{\text{new}})$.

When implementing this procedure, it is useful to determine a permutation $\sigma(\cdot)$ of $\{1, \ldots, p\}$ such that $\theta_{\sigma(1)} \le \theta_{\sigma(2)} \le \cdots \le \theta_{\sigma(p)}$. Let $1 \le i_1 < i_2 < \cdots < i_q = p$ denote those indices i such that $\theta_{\sigma(i)} < \theta_{\sigma(i+1)}$ if $i < p$. Then, with $i_0 = 0$,

$$\begin{aligned}
\mathbb{V}^*(\boldsymbol{\theta}) &= \big\{\boldsymbol{x} \in \mathbb{R}^p : \text{for } 1 \le \ell \le q, \\
&\qquad x_{\sigma(i)} \text{ is constant in } i \in \{i_{\ell-1}+1, \ldots, i_\ell\}\big\}, \\
\mathbb{K}^*(\boldsymbol{\theta}) &= \big\{\boldsymbol{x} \in \mathbb{V}^*(\boldsymbol{\theta}) : \text{for } 1 \le \ell < q,\ x_{\sigma(i_\ell)} \le x_{\sigma(i_{\ell+1})}\big\},
\end{aligned}$$

and

$$\begin{aligned}
&\lambda^*(\boldsymbol{\theta}, \boldsymbol{x}) \\
&\quad = \min_{2 \le \ell \le p : x_{\sigma(i_{\ell-1})} > x_{\sigma(i_\ell)}} \frac{\theta_{\sigma(i_\ell)} - \theta_{\sigma(i_{\ell-1})}}{\theta_{\sigma(i_\ell)} - \theta_{\sigma(i_{\ell-1})} - x_{\sigma(i_\ell)} + x_{\sigma(i_{\ell-1})}}.
\end{aligned}$$

Basic procedure 2c: *A shortcut via the PAVA* In the special case of $Q(\boldsymbol{\theta})$ being the weighted least squares functional in (1), one can determine

$$\boldsymbol{\theta}_{\text{new}} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{K}^*(\boldsymbol{\theta})} Q(\boldsymbol{x})$$

directly by means of the PAVA with a suitable modification for the equality constraints defining $\mathbb{V}^*(\boldsymbol{\theta})$.

3.4 The whole algorithm and its validity

All subspaces $\mathbb{V}(\boldsymbol{\theta})$ and $\mathbb{V}^*(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{K}$, correspond to partitions of $\{1, 2, \ldots, p\}$ into index sets. Namely, the linear subspace corresponding to such a partition consists of all vectors $\boldsymbol{x} \in \mathbb{R}^p$ with the property that $x_u = x_v$ for arbitrary indices u, v belonging to the same set from the partition. Thus the subspaces used in basic procedures 2a–b belong to a finite family $\mathcal{V}$ of linear subspaces of $\mathbb{R}^p$ all containing $\mathbf{1}$.

We may start the algorithm with initial point

$$\boldsymbol{\theta}^{(0)} = \Big(\operatorname*{argmin}_{t\in\mathbb{R}} Q(t\mathbf{1})\Big) \cdot \mathbf{1}.$$

Now suppose that $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(k)} \in \mathbb{K}$ have been chosen such that

$$\boldsymbol{\theta}^{(\ell)} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{V}^{(\ell)}} Q(\boldsymbol{x}) \quad \text{for } 1 \le \ell \le k$$

with linear spaces $\mathbb{V}^{(0)}, \ldots, \mathbb{V}^{(k)} \in \mathcal{V}$. Then $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ satisfies (5), and we may apply basic procedure 1 to check whether $\boldsymbol{\theta}^{(k)} = \hat{\boldsymbol{\theta}}$. If not, we may also apply a variant of basic procedure 2 to get $\boldsymbol{\theta}^{(k+1)} \in \mathbb{K}$ minimizing Q on a linear subspace $\mathbb{V}^{(k+1)} \in \mathcal{V}$, where $Q(\boldsymbol{\theta}^{(k+1)}) < Q(\boldsymbol{\theta}^{(k)})$. Since $\mathcal{V}$ is finite, we will obtain $\hat{\boldsymbol{\theta}}$ after finitely many steps.

Similar arguments show that our algorithm based on basic procedure 2c reaches an optimum after finitely many steps, too.

Final remark on coercivity As mentioned for Setting 2a, the algorithm above may be applicable even in situations when the functional Q fails to be coercive. In fact, we only need to assume that Q attains a minimum, possibly non-unique, over any linear space $\mathbb{V}(\boldsymbol{\theta})$, $\mathbb{V}^*(\boldsymbol{\theta})$ or any cone $\mathbb{K}^*(\boldsymbol{\theta})$, and we have to able to compute it. In Setting 2a, one can verify this easily.

4 Shrinkage estimation

We consider a regression setting as in Sect. 2, this time with Gaussian errors $\varepsilon^t \sim \mathcal{N}(0, \sigma^2)$. As before, the regression function $\mu : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is reduced to a matrix

$$\boldsymbol{M} = \big(\mu(x_{(i)}, y_{(j)})\big)_{i,j} \in \mathbb{R}^{r\times s}$$

for given design points $x_{(1)} < x_{(2)} < \cdots < x_{(r)}$ and $y_{(1)} < y_{(2)} < \cdots < y_{(s)}$. This matrix is no longer assumed to be bimonotone, but the latter constraint will play a role in our estimation method.

4.1 Transforming the signal

At first we represent the signal $\boldsymbol{M}$ with respect to a certain basis of $\mathbb{R}^{r\times s}$. To this end let $\boldsymbol{U} = [\boldsymbol{u}_1\, \boldsymbol{u}_2 \cdots \boldsymbol{u}_r]$ and $\boldsymbol{V} = [\boldsymbol{v}_1\, \boldsymbol{v}_2 \cdots \boldsymbol{v}_s]$ be orthonormal matrices in $\mathbb{R}^{r\times r}$ and $\mathbb{R}^{s\times s}$, respectively, to be specified later. Then we write

$$\boldsymbol{M} = \boldsymbol{U}\tilde{\boldsymbol{M}}\boldsymbol{V}^\top = \sum_{i,j} \tilde{M}_{ij}\, \boldsymbol{u}_i \boldsymbol{v}_j^\top$$

$$\text{with } \tilde{\boldsymbol{M}} = \boldsymbol{U}^\top \boldsymbol{M}\boldsymbol{V} = \big(\boldsymbol{u}_i^\top \boldsymbol{M}\boldsymbol{v}_j\big)_{i,j}.$$

Thus $\tilde{\boldsymbol{M}}$ contains the coefficients of $\boldsymbol{M}$ with respect to the new basis matrices $\boldsymbol{u}_i \boldsymbol{v}_j^\top \in \mathbb{R}^{r\times s}$. The purpose of such a transformation is to obtain a transformed signal $\tilde{\boldsymbol{M}}$ with many coefficients being equal or at least close to zero.

One particular construction of such basis matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ is via discrete smoothing splines: For given degrees k, $\ell \geq 1$, consider *annihilators*

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1,k+1} & & & 0 \\ & a_{22} & \cdots & a_{2,k+2} & & \\ & & \ddots & & \ddots & \\ 0 & & & a_{r-k,r-k} & \cdots & a_{r-k,r} \end{bmatrix}$$
$$\in \mathbb{R}^{(r-k)\times r},$$

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & \cdots & b_{1,\ell+1} & & & 0 \\ & b_{22} & \cdots & b_{2,\ell+2} & & \\ & & \ddots & & \ddots & \\ 0 & & & b_{s-\ell,s-\ell} & \cdots & b_{s-\ell,s} \end{bmatrix}$$
$$\in \mathbb{R}^{(s-\ell)\times s},$$

with unit row vectors such that

$$\boldsymbol{A}\big(x_{(i)}^e\big)_{i=1}^r = \boldsymbol{0} \quad \text{for } e = 0, \ldots, k-1,$$
$$\boldsymbol{B}\big(y_{(j)}^e\big)_{j=1}^s = \boldsymbol{0} \quad \text{for } e = 0, \ldots, \ell-1.$$

An important special case is $k = \ell = 1$. Here

$$\boldsymbol{A} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 & & & 0 \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ 0 & & & 1 & -1 \end{bmatrix} \quad \text{and}$$

$$\boldsymbol{B} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 & & & 0 \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ 0 & & & 1 & -1 \end{bmatrix}$$

satisfy the equations $\boldsymbol{A}\boldsymbol{1}_r = \boldsymbol{0}$ and $\boldsymbol{B}\boldsymbol{1}_s = \boldsymbol{0}$.

Next we determine singular value decompositions of $\boldsymbol{A}$ and $\boldsymbol{B}$, namely,

$$\boldsymbol{A} = \tilde{\boldsymbol{U}} \cdot \big[\boldsymbol{0}_{(r-k)\times k}\ \underbrace{\mathrm{diag}(a_1, \ldots, a_{r-k})}_{0 \leq a_1 \leq \cdots \leq a_{r-k}}\big] \cdot \boldsymbol{U}^\top,$$

$$\boldsymbol{B} = \tilde{\boldsymbol{V}} \cdot \big[\boldsymbol{0}_{(s-\ell)\times \ell}\ \underbrace{\mathrm{diag}(b_1, \ldots, b_{s-\ell})}_{0 \leq b_1 \leq \cdots \leq b_{s-\ell}}\big] \cdot \boldsymbol{V}^\top$$

with column-orthonormal matrices $\tilde{\boldsymbol{U}}$, $\boldsymbol{U} = [\boldsymbol{u}_1\, \boldsymbol{u}_2 \cdots \boldsymbol{u}_r]$, $\tilde{\boldsymbol{V}}$ and $\boldsymbol{V} = [\boldsymbol{v}_1\, \boldsymbol{v}_2 \cdots \boldsymbol{v}_s]$. The vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_\ell$ correspond to the space of polynomials of order at most k and ℓ, respectively. In particular, we always choose $\boldsymbol{u}_1 = r^{-1/2}\boldsymbol{1}_r$ and $\boldsymbol{v}_1 = s^{-1/2}\boldsymbol{1}_s$. Then

$$\begin{aligned} \boldsymbol{M} &= \tilde{M}_{11}\, \boldsymbol{u}_1 \boldsymbol{v}_1^\top \quad \text{(constant part)} \\ &\quad + \sum_{i=2}^{r} \tilde{M}_{i1}\, \boldsymbol{u}_i \boldsymbol{v}_1^\top + \sum_{j=2}^{s} \tilde{M}_{1j}\, \boldsymbol{u}_1 \boldsymbol{v}_j^\top \quad \text{(additive part)} \\ &\quad + \sum_{i,j\geq 2} \tilde{M}_{ij}\, \boldsymbol{u}_i \boldsymbol{v}_j^\top \quad \text{(interactions).} \end{aligned}$$

One may also write

$$\boldsymbol{M} = \boldsymbol{U} \begin{array}{|c|c|} \hline \text{polynomial part} & \text{half-polyn. interactions} \\ k \times \ell & k \times (s-\ell) \\ \hline \text{half-polyn. interactions} & \text{non-polyn. interactions} \\ (r-k) \times \ell & (r-k) \times (s-\ell) \\ \hline \end{array} \boldsymbol{V}^\top.$$

For moderately smooth functions μ we expect $|\tilde{M}_{ij}|$ to have a decreasing trend in $i > k$ and in $j > \ell$. This motivates a class of shrinkage estimators which we describe next.

4.2 Shrinkage estimation in the simple balanced case

In the case of $n = p = rs$ observations such that each grid point $(x_{(i)}, y_{(j)})$ is contained in $\{(x^1, y^1), \ldots, (x^n, y^n)\}$, our input data may be written as a matrix

$$\boldsymbol{Z} = \boldsymbol{M} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon} \in \mathbb{R}^{r\times s}$ having independent components $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Reexpressing such data with respect to the discrete spline basis leads to $\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{M}} + \tilde{\boldsymbol{\varepsilon}}$ with $\tilde{\boldsymbol{Z}} := \boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{V}$ and $\tilde{\boldsymbol{\varepsilon}} := \boldsymbol{U}^\top \boldsymbol{\varepsilon}\boldsymbol{V}$. Note that the raw data $\boldsymbol{Z}$ is the maximum likelihood estimator of $\boldsymbol{M}$. To benefit from the bias-variance trade-off, we consider component-wise shrinkage of the coefficient matrix $\tilde{\boldsymbol{Z}}$: For $\boldsymbol{\gamma} \in [0,1]^{r\times s}$ we consider the candidate estimator

$$\hat{\boldsymbol{M}}^{(\boldsymbol{\gamma})} = \boldsymbol{U}\, (\gamma_{ij}\tilde{Z}_{ij})_{i,j}\, \boldsymbol{V}^\top. \tag{7}$$

Eventually we will choose a shrinkage matrix $\hat{\boldsymbol{\gamma}}$ depending on the data and compute the shrinkage estimator

$$\hat{\boldsymbol{M}} = \hat{\boldsymbol{M}}^{(\hat{\boldsymbol{\gamma}})}. \tag{8}$$

Let $\|\boldsymbol{A}\|_F$ denote the Frobenius norm of a matrix $\boldsymbol{A}$, i.e. $\|\boldsymbol{A}\|_F^2 = \sum_{i,j} A_{ij}^2 = \operatorname{trace}(\boldsymbol{A}^\top \boldsymbol{A})$. As a measure of risk of the estimator (7), we consider

$$\begin{aligned} R(\boldsymbol{\gamma}, \boldsymbol{M}) &= \mathrm{E}\,\big\|\hat{\boldsymbol{M}}^{(\gamma)} - \boldsymbol{M}\big\|_F^2 \\ &= \sum_{i,j} \big((1-\gamma_{ij})^2 \tilde{M}_{ij}^2 + \sigma^2 \gamma_{ij}^2\big) \\ &= \sum_{i,j} (\tilde{M}_{ij}^2 + \sigma^2)\Big(\gamma_{ij} - \frac{\tilde{M}_{ij}^2}{\tilde{M}_{ij}^2 + \sigma^2}\Big)^2 \\ &\quad + \sum_{i,j} \frac{\tilde{M}_{ij}^2 \sigma^2}{\tilde{M}_{ij}^2 + \sigma^2}. \end{aligned}$$

Here we used the fact that the transformed error matrix $\tilde{\boldsymbol{\varepsilon}}$ has the same distribution as $\boldsymbol{\varepsilon}$. An estimator of this risk is given by

$$\begin{aligned} \hat{R}(\boldsymbol{\gamma}) &= \sum_{i,j} \big(\hat{\sigma}^2 \gamma_{ij}^2 + (1-\gamma_{ij})^2 (\tilde{Z}_{ij}^2 - \hat{\sigma}^2)\big) \\ &= \sum_{i,j} \tilde{Z}_{ij}^2 \big(\gamma_{ij} - (1 - \hat{\sigma}^2/\tilde{Z}_{ij}^2)\big)^2 \\ &\quad + \sum_{i,j} \hat{\sigma}^2 (1 - \hat{\sigma}^2/\tilde{Z}_{ij}^2), \end{aligned}$$

where $\hat{\sigma}$ is a certain estimator of σ, e.g. based on high frequency components of $\tilde{\boldsymbol{Z}}$, see later.

Thus optimal shrinkage factors would be given by $\check{\gamma}_{ij} = \tilde{M}_{ij}^2/(\tilde{M}_{ij}^2 + \sigma^2)$, but these depend on the unknown signal $\boldsymbol{M}$. Naive estimators would be $\hat{\gamma}_{ij} = (1 - \hat{\sigma}^2/\tilde{Z}_{ij}^2)^+$. The resulting estimator's performance is rather poor, but it improves substantially if $\hat{\boldsymbol{\gamma}}$ in (8) is given by

$$\hat{\gamma}_{ij} = \max\Big(1 - \frac{\tau \log(p) \hat{\sigma}^2}{\tilde{Z}_{ij}^2}, 0\Big) \tag{9}$$

with τ close to 2; cf. Donoho and Johnstone (1994).

An alternative strategy, utilized for instance by Beran and Dümbgen (1998), is to restrict $\boldsymbol{\gamma}$ to a certain convex set of shrinkage matrices serving as a caricature of the optimal $\boldsymbol{\gamma}$. The previous considerations suggest to restrict $-\boldsymbol{\gamma}$ to be contained in $\mathbb{K}_{r,s}^{(k,\ell)}$, the set of all matrices $\boldsymbol{\theta} \in \mathbb{R}^{r\times s}$ such that

- $\theta_{1,j} = \theta_{2,j} = \cdots = \theta_{k,j}$ is non-decreasing in $j > \ell$,
- $\theta_{i,1} = \theta_{i,2} = \cdots = \theta_{i,\ell}$ is non-decreasing in $i > k$,
- $(\theta_{ij})_{i>k, j>\ell}$ belongs to $\mathbb{K}_{r-k,s-\ell}$.

The set of all such shrinkage matrices $\boldsymbol{\gamma}$ is denoted by $\mathbb{G}_{r,s}^{(k,\ell)} = (-\mathbb{K}_{r,s}^{(k,\ell)}) \cap [0,1]^{r\times s}$. Thus we propose to use the shrinkage matrix

$$\hat{\boldsymbol{\gamma}} \;=\; \operatorname*{argmin}_{\boldsymbol{\gamma} \in \mathbb{G}_{r,s}^{(k,\ell)}} \hat{R}(\boldsymbol{\gamma}). \tag{10}$$

In the present setting one can show (cf. Beran and Dümbgen 1998) that

$$\check{\boldsymbol{\gamma}} = \operatorname*{argmin}_{\boldsymbol{\gamma} \in \mathbb{G}_{r,s}^{(k,\ell)}} R(\boldsymbol{\gamma}, \boldsymbol{M}) = \Big(\frac{\check{\eta}_{ij}}{\check{\eta}_{ij} + \sigma^2}\Big)_{i,j}$$

$$\text{with } \check{\boldsymbol{\eta}} = -\operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{K}_{r,s}^{(k,\ell)}} \sum_{i,j} \big(-(\tilde{M}_{ij}^2 + \sigma^2) - \theta_{ij}\big)^2.$$

Similarly,

$$\hat{\boldsymbol{\gamma}} = \operatorname*{argmin}_{\boldsymbol{\gamma} \in \mathbb{G}_{r,s}^{(k,\ell)}} \hat{R}(\boldsymbol{\gamma}) = \big((1 - \hat{\sigma}^2/\hat{\eta}_{ij})^+\big)_{i,j}$$

$$\text{with } \hat{\boldsymbol{\eta}} = -\operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{K}_{r,s}^{(k,\ell)}} \sum_{i,j} (-\tilde{Z}_{ij}^2 - \theta_{ij})^2.$$

This allows one to experiment with different values for $\hat{\sigma}$ with little effort.

Estimation of the noise level Two particular estimators are given by

$$\begin{aligned} \hat{\sigma}_{1,\kappa} &= \Big(\frac{\sum_{i/r + j/s \ge \kappa} \tilde{Z}_{ij}^2}{\#\{(i,j) : i/r + j/s \ge \kappa\}}\Big)^{1/2} \quad \text{or} \\ \hat{\sigma}_{2,\kappa} &= \frac{\operatorname{Median}\big(|\tilde{Z}_{ij}| : i/r + j/s \ge \kappa\big)}{\Phi^{-1}(3/4)} \end{aligned} \tag{11}$$

for a certain number $\kappa \in (0,2)$, where Φ^{-1} denotes the standard Gaussian quantile function. The idea is that for $i \gg 1$ and $j \gg 1$, the components $\tilde{Z}_{ij}$ are essentially equal to the noise variables $\tilde{\varepsilon}_{ij} \sim \mathcal{N}(0, \sigma^2)$. Otherwise both estimators tend to overestimate σ.

As to the choice of κ, we propose to choose it via visual inspection of the graphs of $\kappa \mapsto \hat{\sigma}_{1,\kappa}$ and $\kappa \mapsto \hat{\sigma}_{2,\kappa}$. Typically these functions are almost constant and close to σ on a large subinterval of $(0,2)$, non-increasing to the left of that interval, and show random fluctuations to the right. As we shall illustrate later, the quality of the shrinkage estimator is rather robust with respect to the estimator $\hat{\sigma}$. In particular, overestimating σ slightly is typically harmless or even beneficial.

Consistency We now augment the foregoing discussion with consistency results that follow from more general considerations in Beran and Dümbgen (1998). First of all, for large p, the normalized quadratic loss $p^{-1}\|\hat{\boldsymbol{M}}^{(\boldsymbol{\gamma})} - \boldsymbol{M}\|_F^2$ of a candidate estimator is close to its normalized risk $p^{-1}R(\boldsymbol{\gamma}, \boldsymbol{M})$, uniformly over $\boldsymbol{\gamma} \in \mathbb{G}_{r,s}^{(k,\ell)}$. Precisely,

$$\begin{aligned} &\mathrm{E} \sup_{\boldsymbol{\gamma} \in \mathbb{G}_{r,s}^{(k,\ell)}} \big|p^{-1}\|\hat{\boldsymbol{M}}^{(\boldsymbol{\gamma})} - \boldsymbol{M}\|_F^2 - p^{-1}R(\boldsymbol{\gamma}, \boldsymbol{M})\big| \\ &\quad \le C\, \frac{\sigma^2 + \sigma p^{-1/2}\|\boldsymbol{M}\|_F}{\max(r,s)^{1/2}} \end{aligned}$$

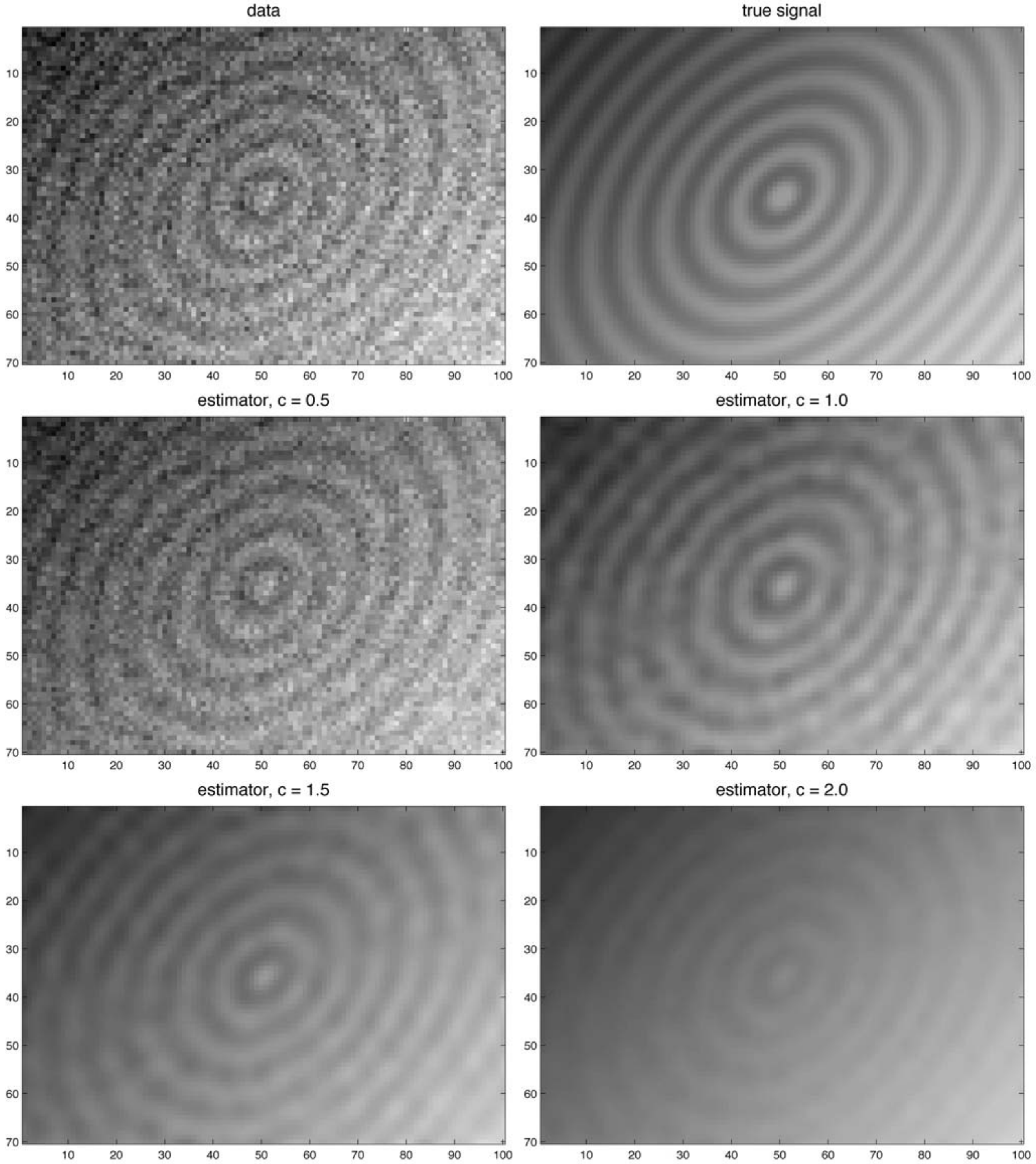

Fig. 3 Shrinkage estimation: data and true signal (*1st row*), estimators with $\hat{\sigma} \leftarrow c\hat{\sigma}$ for $c = 0.5, 1.0, 1.5, 2.0$ (*2nd and 3rd row*)

with C denoting a generic universal constant. Moreover, if the variance estimator $\hat{\sigma}^2$ is L_1-consistent, the normalized estimated risk $p^{-1}\hat{R}(\boldsymbol{\gamma})$ differs little from the normalized true risk $p^{-1}R(\boldsymbol{\gamma}, \boldsymbol{M})$, uniformly in $\boldsymbol{\gamma} \in \mathbb{G}_{r,s}^{(k,\ell)}$. Namely,

$$\mathrm{E} \sup_{\boldsymbol{\gamma} \in \mathbb{G}_{r,s}^{(k,\ell)}} \left| p^{-1}\hat{R}(\boldsymbol{\gamma}) - p^{-1}R(\boldsymbol{\gamma}, \boldsymbol{M}) \right| \le C\,\frac{\sigma^2 + \sigma p^{-1/2}\|\boldsymbol{M}\|_F}{\max(r,s)^{1/2}} + C\,\mathrm{E}\left|\hat{\sigma}^2 - \sigma^2\right|.$$

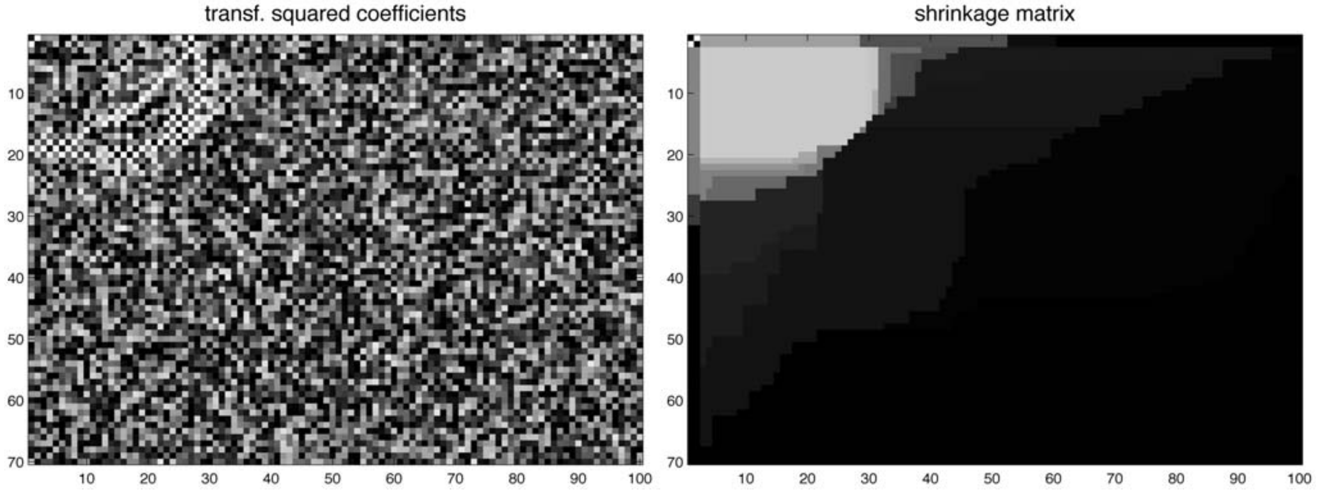

Fig. 4 Shrinkage estimation: Transformed squared coefficients $\tilde{Z}_{ij}^2/(1+\tilde{Z}_{ij}^2)$ (*left*) and bimonotone shrinkage matrix $\hat{\boldsymbol{\gamma}}$ (*right*)

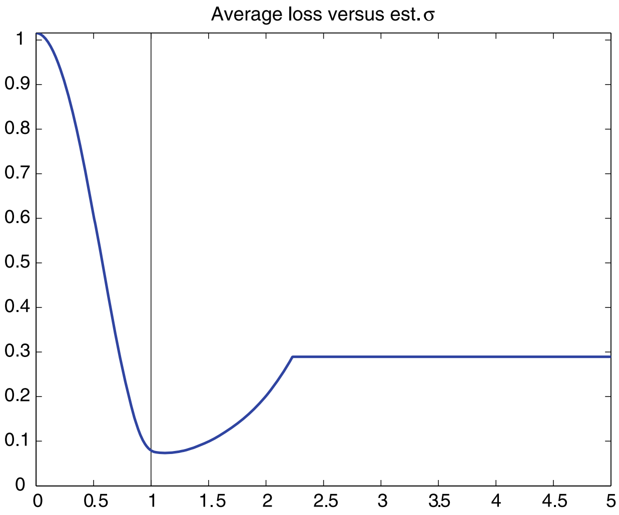

Fig. 5 Shrinkage estimation: Average quadratic loss as a function of $\hat{\sigma}$

In particular, the shrinkage matrix $\hat{\boldsymbol{\gamma}}$ in (10) and the corresponding estimator $\hat{\boldsymbol{M}} = \hat{\boldsymbol{M}}^{(\hat{\boldsymbol{\gamma}})}$ satisfy the inequalities

$$\left.\begin{array}{l} \mathrm{E}\,|p^{-1}\hat{R}(\hat{\boldsymbol{\gamma}}) - p^{-1}R_{\min}(\boldsymbol{M})| \\ \mathrm{E}\,|p^{-1}\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|_F^2 - p^{-1}R_{\min}(\boldsymbol{M})| \end{array}\right\}$$

$$\le C\,\frac{\sigma^2 + \sigma p^{-1/2}\|\boldsymbol{M}\|_F}{\max(r,s)^{1/2}} + C\,\mathrm{E}\,\big|\hat{\sigma}^2 - \sigma^2\big|,$$

where $R_{\min}(\boldsymbol{M})$ denotes the minimum of $R(\boldsymbol{\gamma}, \boldsymbol{M})$ over all $\boldsymbol{\gamma} \in \mathbb{G}_{r,s}^{(k,\ell)}$.

Example 4.1 We generated a random matrix $\boldsymbol{Z} \in \mathbb{R}^{r\times s}$ with $r = 60$ rows, $s = 100$ columns and independent components $Z_{ij} \sim \mathcal{N}(\mu(x_{(i)}, y_{(j)}), 1)$, where $x_{(i)} = (i-0.5)/r$, $y_{(j)} = (j-0.5)/s$, and

$$\mu(x,y) = 2\tau(x,y)^{-0.25}\sin(\tau(x,y)) + 0.05(x+y),$$

$$\tau(x,y) = \sqrt{3x^2 + 2xy + 3y^2} + 1.$$

We smoothed this data matrix $\boldsymbol{Z}$ as described above with annihilators of order $k = \ell = 2$. The estimators $\hat{\sigma}_{1,\kappa}$ and $\hat{\sigma}_{2,\kappa}$ turned out to be almost constant and slightly smaller than 1.0 on $(0.5, 0.65)$, so we chose $\hat{\sigma} = 1$. The first row of Fig. 3 shows gray scale images of the raw data $\boldsymbol{Z}$ (left) and the true signal $\boldsymbol{M}$ (right). The second and third row depict the matrix $\hat{\boldsymbol{M}}$ for different values of $\hat{\sigma}$. Precisely, to show the effect of varying the estimated noise level, we replaced $\hat{\sigma}$ with $c\hat{\sigma}$, where $c = 0.5$ (undersmoothing), $c = 1.0$ (original estimator), $c = 1.5$ (oversmoothing) and $c = 2.0$ (heavy oversmoothing). In these pictures the gray scale ranges from -7 (black) to 7 (white).

Figure 4 depicts the transformed squared coefficients $\tilde{Z}_{ij}^2/(1+\tilde{Z}_{ij}^2)$ (left panel) and the bimonotone shrinkage matrix $\hat{\boldsymbol{\gamma}}$ (right panel).

Figure 5 shows the average squared loss $p^{-1}\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|_F^2$ as a function of $\hat{\sigma}$. The emerging pattern is very stable over all simulations we looked at. This plot and Fig. 4 show that there is a rather large range of values for $\hat{\sigma}$ leading to estimators of similar quality. Overestimation of $\hat{\sigma}$ is less severe than underestimation and sometimes even beneficial.

Since this is just one simulation, we also conducted a simulation study. We generated 5000 such data matrices $\boldsymbol{Z}$. Each time we estimated the noise level via $\hat{\sigma} = \hat{\sigma}_{1,1}$. Then we computed the shrinkage estimators $\hat{\boldsymbol{M}}$ in (8), where the shrinkage matrices $\hat{\boldsymbol{\gamma}}$ were given by (10) and by (9) with τ running through a fine grid of points in $(0, 2]$. It turned out that $\tau = 0.60$ yielded optimal performance, although this value depends certainly on the underlying signal and

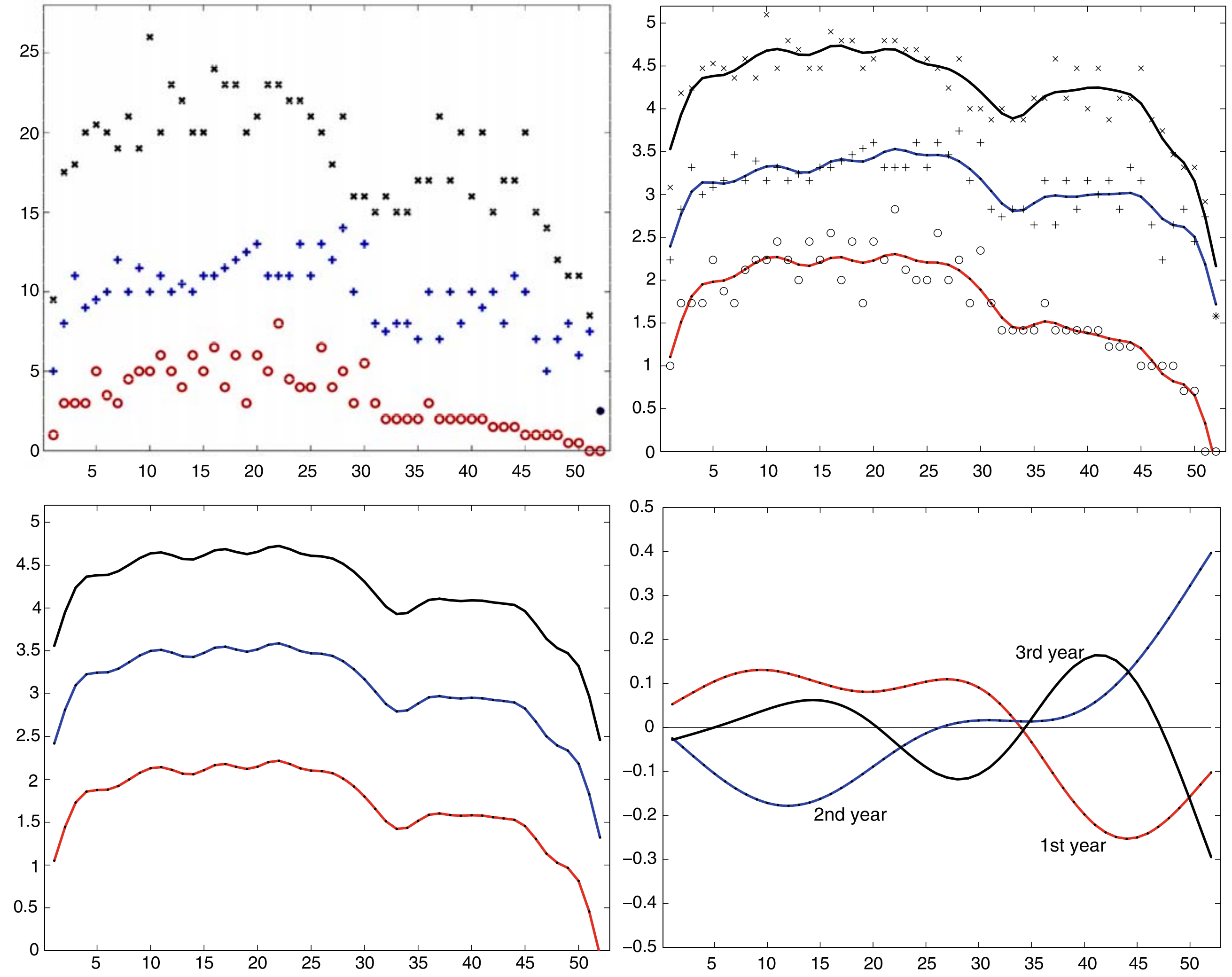

Fig. 6 Raw vineyard data (*top left*), transformed data and fitted values (*top right*), additive part (*bottom left*) and interactions (*bottom right*)

Table 3 Estimated risks of different estimators in Example 4.1

Bimonotone	Componentwise thresholding (9) with				
shrinkage (10)	$\tau = 0.5$	$\tau = 0.6$	$\tau = 1.0$	$\tau = 1.5$	$\tau = 2.0$
0.0790 (0.0044)	0.0922 (0.0050)	0.0888 (0.0051)	0.1044 (0.0061)	0.1342 (0.0073)	0.1619 (0.0082)

noise level. Table 3 provides Monte Carlo estimates of the corresponding risk, i.e. the expectation of the normalized quadratic loss $p^{-1}\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|_F^2$. The values into parentheses are the estimated standard deviations of the latter loss. This table shows that bimonotone shrinkage yields better results than componentwise (soft) thresholding.

4.3 Viticultural case study

In this case study, row i of the data matrix $\boldsymbol{Y} \in \mathbb{R}^{52\times 3}$ reports the grape yields harvested in 3 successive years from a vineyard near Lake Erie that has 52 rows of vines. The data is taken from Chatterjee et al. (1995). The grape yields, measured in lugs of grapes harvested from each vineyard-row, are plotted in the upper left panel of Fig. 6, using a different plotting character for each of the three years. The analysis seeks to bring out patterns in the vineyard-row yields that persist across years. Year and vineyard-row are both ordinal covariates. The covariate vineyard-row summarizes location-dependent effects that may be due to soil fertility and microclimate. The covariate year summarizes time-varying effects that may be due to rainfall pattern, temperatures, and viticultural practices.

A preliminary data analysis based on running means and variance estimates from triplets $(Y_{i,j}, Y_{i+1,j}, Y_{i+2,j})$, $1 \le i \le 50$, revealed that a square-root transformation yields a data matrix $\boldsymbol{Z} \in \mathbb{R}^{52\times 3}$ which may be viewed as a two-way layout in which both the row and column numbers are ordinal covariates, the measurement errors are independent with mean zero and common unknown variance σ^2 and unknown mean matrix $\boldsymbol{M} = \mathrm{E}\,\boldsymbol{Z}$.

Now we applied the orthonormal transformation into spline bases with $x_{(i)} = i$ and $y_{(j)} = j$, where $k = 2$ and $\ell = 1$. In particular, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are proportional to $\mathbf{1}_{52}$ and $(i - 26.5)_{i=1}^{52}$, respectively. Similarly, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$ are proportional to $\mathbf{1}_3$, $(-1, 0, 1)^\top$ and $(1, -2, 1)^\top$, respectively. The graphs of $\kappa \mapsto \hat{\sigma}_{1,\kappa}$ and $\kappa \mapsto \hat{\sigma}_{2,\kappa}$ revealed that $\hat{\sigma} = 0.25$ is a plausible estimator for σ. The resulting fitted matrix $\hat{\boldsymbol{M}}$ is shown in the upper right panel of Fig. 6, adding linear interpolation between adjacent elements to bring out their trend. In addition the transformed data Z_{ij} are superimposed as single points.

The estimated mean grape yields reveal shared patterns across the three years. Large dips in estimated mean grape yields occur in the outermost rows of the vineyard and near row 33. These point to possible geographical variations in growing conditions, such as harsher climate at the vineyard edges or changes in soil fertility.

It is also interesting to split the fit $\hat{\boldsymbol{M}}$ into an additive part (including constant) and interactions,

$$\begin{aligned}\hat{\boldsymbol{M}}_{\mathrm{add}} &= \hat{\gamma}_{11}\tilde{Z}_{11}\,\boldsymbol{u}_1\boldsymbol{v}_1^\top \\ &\quad + \sum_{i=2}^{r}\hat{\gamma}_{i1}\tilde{Z}_{i1}\,\boldsymbol{u}_i\boldsymbol{v}_1^\top + \sum_{j=2}^{s}\hat{\gamma}_{1j}\tilde{Z}_{1j}\,\boldsymbol{u}_1\boldsymbol{v}_j^\top, \\ \hat{\boldsymbol{M}}_{\mathrm{inter}} &= \sum_{i=2}^{r}\sum_{j=2}^{s}\hat{\gamma}_{ij}\tilde{Z}_{ij}\,\boldsymbol{u}_i\boldsymbol{v}_j^\top.\end{aligned}$$

The lower panels of Fig. 6 depict these parts separately. The plot of the additive part emphasizes the pattern across rows just described and the (nonlinear) increase across years. The interactions reveal that a simple additive model does not seem appropriate for these data.

Acknowledgements This work was supported by the Swiss National Science Foundation. We are grateful to two reviewers for constructive comments.

References

Ayer, M., Brunk, H.D., Reid, W.T., Silverman, E.: An empirical distribution function for sampling with incomplete information. Ann. Math. Stat. **26**, 641–647 (1955)

Beran, R., Dümbgen, L.: Modulation of estimators and confidence sets. Ann. Stat. **26**, 1826–1856 (1998)

Best, M.J., Chakravarti, N.: Active set algorithms for isotonic regression; a unifying framework. Math. Program. **47**, 425–439 (1990)

Burdakow, O., Grimwall, A., Hussian, M.: A generalised PAV algorithm for monotone regression in several variables. In: Antoch, J. (ed.) COMPSTAT 2004—Proceedings in Computational Statistics, 16th Symposium Held in Prague, Czech Republic, pp. 761–767. Physica-Verlag, Heidelberg (2004)

Chatterjee, S., Handcock, M.S., Simonoff, J.S.: A Casebook for a First Course in Statistics and Data Analysis. Wiley, New York (1995)

Cormen, T.H., Leiserson, C.E., Rivest, R.L.: Introduction to Algorithms. MIT Press, Cambridge (1990)

Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**, 425–455 (1994)

Dümbgen, L., Hüsler, A., Rufibach, K.: Active set and EM algorithms for log–concave densities based on complete and censored data. Technical Report 61, IMSV, University of Bern (2007). arXiv:0707.4643

Fletcher, R.: Practical Methods of Optimization, 2nd edn. Wiley, New York (1987)

Robertson, T., Wright, F.T., Dykstra, R.L.: Order Restricted Statistical Inference. Wiley, New York (1988)

Spouge, J., Wan, H., Wilbur, W.J.: Least squares isotonic regression in two dimensions. J. Optim. Theory Appl. **117**, 585–605 (2003)