Available online at www.sciencedirect.com
SCIENCE DIRECT®
Statistics & Probability Letters 70 (2004) 87–94
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Shrinkage estimation for convex polyhedral cones

Anna Amirdjanova*, Michael Woodroofe

Department of Statistics, The University of Michigan, 455 West Hall, 550 E University, Ann Arbor, MI 48109-1092, USA

Abstract

Estimation of a multivariate normal mean is considered when the latter is known to belong to a convex polyhedron. It is shown that shrinking the maximum likelihood estimator towards an appropriate target can reduce mean squared error. The proof combines an unbiased estimator of a risk difference with some geometrical considerations. When applied to the monotone regression problem, the main result shows that shrinking the maximum likelihood estimator towards modifications that have been suggested to alleviate the spiking problem can reduce mean squared error.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Degrees of freedom; Maximum likelihood estimator; Mean squared error; Primal–dual bases; Projections; Stein's Identity; Target estimator

1. Introduction

Let $y = (y_1, \ldots, y_n)'$ denote a normally distributed random vector with unknown mean θ and covariance matrix $\sigma^2 I_n$ and suppose that θ is known to belong to a closed convex polyhedron $\Omega \subseteq \mathbb{R}^n$. For such a model, the maximum likelihood estimator, $\tilde{\theta}$ say, of θ is the projection of y onto Ω, an operation that is described in some detail in the next section. The main result is that shrinking $\tilde{\theta}$ toward an appropriate target estimator $\tilde{\theta}^{\circ}$ reduces the mean squared error.

The model arises in isotonic regression, in particular, where Ω is the set of non-decreasing vectors, $\Omega = \{\omega \in \mathbb{R}^n : -\infty < \omega_1 \leqslant \cdots \leqslant \omega_n < \infty\}$. For this model the maximum likelihood estimators $\tilde{\theta}_1$ and $\tilde{\theta}_n$ are biased, highly variable, and asymptotically inconsistent. Suggestions

*Corresponding author.
E-mail address: anutka@umich.edu (A. Amirdjanova).

0167-7152/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2004.08.007

for alleviating this problem include simply replacing the first and last few $\tilde{\theta}_k$'s by their average. Modifications of this nature date from Wegman (1970), for the related problem of density estimation and can improve asymptotic performance. See, for example, Woodroofe and Sun (1993) and Wu et al. (2001). As an application of the main result, it is shown that shrinking the maximum likelihood estimator towards such a modification improves finite sample performance.

The main result is presented in Section 4. Section 2 contains preliminary material on the structure of $\tilde{\theta}$, and Section 3 on the structure of $\tilde{\theta}^{\circ}$. Related results have been obtained by Sengupta and Sen (1991) and Ouassou and Strawderman (2002). The main result here allows a more flexible form of shrinkage, a flexibility that is needed for the application. Bock (1982) and Wright (1978) made early contributions.

2. Preliminaries

Let $\Omega \subseteq \mathbb{R}^n$ be a convex polyhedral cone. Specifically, suppose that there are linearly independent $\gamma_1, \ldots, \gamma_m \in \mathbb{R}^n$ for which

$$\Omega = \{z \in \mathbb{R}^n : \langle \gamma_j, z \rangle \geqslant 0, \quad j = 1, \ldots, m\}. \tag{1}$$

Thus $m \leqslant n$. Given $y \in \mathbb{R}^n$, let $\tilde{\theta} = \tilde{\theta}(y)$ be the projection of y onto Ω, $\tilde{\theta} = \tilde{\theta}(y) = \Pi_{\Omega}(y)$, so that $\|y - \tilde{\theta}\| = \inf_{\theta \in \Omega} \|y - \theta\|$. Then

$$\langle y - \tilde{\theta}, \theta \rangle \leqslant 0, \tag{2}$$

for all $\theta \in \Omega$. In fact, (2) holds for every $\theta \in \Omega$ for which $\tilde{\theta} + \varepsilon\theta \in \Omega$ for some $\varepsilon > 0$, so that $\langle y - \tilde{\theta}, \tilde{\theta} \rangle = 0$.

Let L be the orthogonal complement of span $\{\gamma_1, \ldots, \gamma_m\}$; let $\Gamma = [\gamma_1, \ldots, \gamma_m]$ $(n \times m)$; write

$$\Gamma(\Gamma'\Gamma)^{-1} = [\delta_1, \ldots, \delta_m];$$

and let $\delta_{m+1}, \ldots, \delta_n$ be an orthonormal basis for L. Then $\delta_1, \ldots, \delta_n$ is a (not necessarily orthonormal) basis for $\mathbb{R}^n$, and $\langle \gamma_i, \delta_j \rangle = 1$ or 0 accordingly as $i = j$ or $i \neq j$, since $\Gamma'[\delta_1, \ldots, \delta_m] = I_m$ and $\Gamma'[\delta_{m+1}, \ldots, \delta_n] = 0$. So, if $z \in \mathbb{R}^n$ is written as $z = \sum_{i=1}^n c_i \delta_i$, then $c_j = \langle \gamma_j, z \rangle$ for $j = 1, \ldots, m$, and $z \in \Omega$ iff $c_1, \ldots, c_m \geqslant 0$. The sets $\{\gamma_1, \ldots, \gamma_m\}$ and $\{\delta_1, \ldots, \delta_m\}$ have been called *primal and dual bases* (for span$\{\gamma_1, \ldots, \gamma_m\}$) in Fraser and Massam (1989) and Meyer (1999).

Example (*Monotone regression*). In the monotone regression problem, Ω is of form (1) with $m = n - 1$, $\gamma_1 = (-1, 1, 0, \ldots, 0)', \ldots,$ and $\gamma_{n-1} = (0, \ldots, 0, -1, 1)'$. After some routine calculations,

$$\delta_j = \frac{j}{n} 1^n - 1^j,$$

for $j \leqslant n - 1$, where 1^j is a list of j 1's followed by $(n - j)$ 0's for $j = 1, \ldots, n$. In this example, L is the linear subspace spanned by 1^n.

Recall that $\tilde{\theta} = \tilde{\theta}(y)$ is the projection of y onto Ω, and write

$$\tilde{\theta} = \sum_{j=1}^{n} \tilde{c}_j \delta_j, \tag{3}$$

where $\tilde{c}_j = \tilde{c}_j(y) = \langle \gamma_j, \tilde{\theta} \rangle \geqslant 0$ for $j = 1, \ldots, m$. For subsets $a \subseteq \{1, \ldots, m\}$, let $a^c = \{1, \ldots, m\} - a$ and

$$B_a = \{y \in \mathbb{R}^n : \tilde{c}_j > 0 \text{ for } j \in a \ \& \ \tilde{c}_j = 0 \text{ for } j \in a^c\}.$$

From Meyer and Woodroofe (2000), the sets B_a partition $\mathbb{R}^n$, and their boundaries are of Lebesgue measure zero. Moreover, if $y \in B_a$, then $\tilde{\theta}$ is the projection of y onto $\text{span}\{\delta_j,\ j \in a \cup \{m+1, \ldots, n\}\}$, and $y - \tilde{\theta}$ is the projection of y onto $\text{span}\{\gamma_j : j \in a^c\}$. If $a = \{k_1, \ldots, k_r\}$, where $0 = k_0 < k_1 < k_2 < \cdots < k_r < k_{r+1} = (m+1) \wedge n$, let

$$X_a = [\delta_{k_1}, \ldots, \delta_{k_r}]\ (n \times r). \tag{4}$$

Then the projection operator onto $\text{span}\{\delta_j : j \in a\}$ is $P_a = X_a(X'_a X_a)^{-1} X'_a$; and $\tilde{\theta}(y) = P_a y + \pi_L y$ for $y \in B_a$, where $\pi_L y$ is the projection of y onto L.

3. The target estimator

Now let ℓ be an integer for which $1 \leqslant \ell \leqslant m$ and that is fixed for the remainder of the paper. Define the target estimator

$$\tilde{\theta}^\circ = \sum_{j=\ell+1}^{n} \tilde{c}_j \delta_j. \tag{5}$$

Equivalently, for $y \in B_a$, let $a = \{k_1, \ldots, k_r\}$ be as in (4) and let s denote the unique integer for which $k_s \leqslant \ell < k_{s+1}$. Then

$$\tilde{\theta}^\circ = \sum_{j=s+1}^{r} \tilde{c}_{k_j} \delta_{k_j} + \sum_{j=m+1}^{n} \tilde{c}_j \delta_j,$$

where an empty sum is to be interpreted as zero. Thus,

$$\tilde{\theta} - \tilde{\theta}^\circ = \sum_{j=1}^{s} \tilde{c}_{k_j} \delta_{k_j}, \quad \text{and} \quad \langle y - \tilde{\theta}, \tilde{\theta} - \tilde{\theta}^\circ \rangle = 0, \tag{6}$$

since $\langle \gamma_i, \delta_j \rangle = 0$ for $i \in a^c$ and $j \in a \cup \{m+1, \ldots, n\}$. In (5) $\ell = \ell_n$ does not depend on y; s depends on a, but this dependence is suppressed in the notation to reduce the number of higher-order subscripts. If $\ell = m$, then $\tilde{\theta}^\circ = \pi_L y$ is just the projection of y onto L.

Example (*Monotone regression*). After some algebra,

$$\tilde{\theta}_j^\circ = \tilde{\omega} + \begin{cases} \tilde{\theta}_{\ell+1} & \text{if } j \leqslant \ell+1, \\ \tilde{\theta}_j & \text{if otherwise,} \end{cases} \quad \text{where } \tilde{\omega} = \frac{1}{n} \sum_{i=1}^{\ell} (\tilde{\theta}_i - \tilde{\theta}_{\ell+1}).$$

Thus, $\tilde{\theta}_j^\circ$ is constant for $j \leqslant \ell + 1$, as in Wu et al. (2001). By relabeling $\gamma_1, \ldots, \gamma_{n-1}$, it is possible to construct target estimators $\tilde{\theta}^\circ$ for which $\tilde{\theta}_j^\circ$ is constant for j over other subintervals of $1, \ldots, n$, as in Wright (1982) . It is known that such grouping improves asymptotic performance, under mild conditions. The main results show that shrinking $\tilde{\theta}$ towards an estimator of the form $\tilde{\theta}^\circ$ can reduce mean squared error in moderate samples.

It is convenient to represent $\tilde{\theta} - \tilde{\theta}^{\circ}$ in matrix notation. For $y \in B_a$, let $c^* = [\tilde{c}_{k_1}, \ldots, \tilde{c}_{k_r}]'$, where $k_1, \ldots, k_r$ are as in (4); that is, $c^* = (X'_a X_a)^{-1} X'_a y$. Next, let

$$M_a = \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix}.$$

Then $M_a c^* = [\tilde{c}_{k_1}, \ldots, \tilde{c}_{k_s}, 0 \cdots, 0]'$, $X_a M_a c^* = [\delta_{k_1}, \ldots, \delta_{k_r}][\tilde{c}_{k_1}, \cdots, \tilde{c}_{k_s}, 0, \ldots, 0]' = \tilde{\theta} - \tilde{\theta}^{\circ}$, and, therefore, $\tilde{\theta} - \tilde{\theta}^{\circ} = Q_a y$, where $Q_a = X_a M_a (X'_a X_a)^{-1} X'_a$.

Two properties of Q_a are relevant.

Lemma 1. *For any $a \subseteq \{1, \ldots, m\}$, $Q_a^2 = Q_a$ and $\mathrm{tr}(Q_a) = s$.*

Proof. This follows from

$$Q_a^2 = X_a M_a (X'_a X_a)^{-1} X'_a X_a M_a (X'_a X_a)^{-1} X'_a = X_a M_a^2 (X'_a X_a)^{-1} X'_a = Q_a,$$

and $\mathrm{tr}(Q_a) = \mathrm{tr}(M_a (X'_a X_a)^{-1} X'_a X_a) = \mathrm{tr}(M_a) = s$. □

An alternative statement of the second assertion is $\mathrm{tr}(Q_a) = \sum_{i=1}^m H[\langle \gamma_i, \tilde{\theta} - \tilde{\theta}^{\circ} \rangle]$, where H is the Heaveside function $H(x) = 0$ for $x \leqslant 0$ and $H(x) = 1$ for $x > 0$.

From Meyer and Woodroofe (2000), $\tilde{\theta}$ is Lipschitz continuous and, therefore, almost differentiable in the sense of Stein (1981). The same is then true of $\tilde{\theta}^{\circ}$. Let $D = \mathrm{div}(\tilde{\theta} - \tilde{\theta}^{\circ})$, the divergence of the vector field $\tilde{\theta} - \tilde{\theta}^{\circ}$; that is,

$$D(y) = \sum_{i=1}^n \frac{\partial}{\partial y_i} (\tilde{\theta}_i - \tilde{\theta}_i^{\circ})(y),$$

where the partial derivatives are understood in the sense of Stein (1981). Then, D is constant on the interior of each B_a,

$$D(y) = \mathrm{tr}(Q_a) = \sum_{i=1}^m H[\langle \gamma_i, \tilde{\theta} - \tilde{\theta}^{\circ} \rangle], \tag{7}$$

but discontinuous on $\mathbb{R}^n$.

Let ψ be the beta(2, 2) distribution function, $\psi(x) = 0$, $3x^2 - 2x^3$, or 1 for $x \leqslant 0$, $0 < x \leqslant 1$, or $x > 1$, so that $\psi'(x) = 6x(1-x)$ for $0 \leqslant x \leqslant 1$, and observe that $x\psi'(x) \geqslant 0$ for all $x \in \mathbb{R}$. Next, given an $\varepsilon > 0$, let

$$\Psi(z) = \sum_{i=1}^m \psi(\langle \gamma_i, z \rangle) \quad \text{and} \quad D_\varepsilon(y) = \Psi\left(\frac{\tilde{\theta} - \tilde{\theta}^{\circ}}{\varepsilon}\right), \tag{8}$$

for $y, z \in \mathbb{R}^n$.

Lemma 2. *$0 \leqslant D_\varepsilon(y) \leqslant D(y)$ for $\varepsilon > 0$ and $\lim_{\varepsilon \downarrow 0} D_\varepsilon(y) = D(y)$ for all $y \in \mathbb{R}^n$. Moreover, $\langle \nabla D_\varepsilon(y), Q_a y \rangle \geqslant 0$ on B_a for each $a \subseteq \{1, \ldots, m\}$ and $\varepsilon > 0$.*

Proof. The first two assertions are clear from (7) and (8). For the third assertion, first observe that

$$\frac{\partial}{\partial z_j} \Psi(z) = \sum_{i=1}^m \frac{\partial}{\partial z_j} \psi(\langle \gamma_i, z \rangle) = \sum_{i=1}^m \gamma_{ij} \psi'(\langle \gamma_i, z \rangle).$$

So,

$$\langle \nabla \Psi(z), z\rangle = \sum_{j=1}^{n}\sum_{i=1}^{m} \gamma_{ij}\psi'(\langle \gamma_i, z\rangle)z_j = \sum_{i=1}^{m} \psi'(\langle \gamma_i, z\rangle)\langle \gamma_i, z\rangle \geqslant 0,$$

for all $z \in \mathbb{R}^n$. On the interior of any B_a, $D_\varepsilon(y) = \Psi(Q_a y/\varepsilon)$, $\nabla D_\varepsilon(y) = (Q_a'/\varepsilon)\nabla\Psi(Q_a y/\varepsilon)$ and, therefore,

$$\langle \nabla D_\varepsilon(y), Q_a y\rangle = \frac{1}{\varepsilon}\left\langle Q_a' \nabla\Psi\left(\frac{Q_a y}{\varepsilon}\right), Q_a y\right\rangle = \frac{1}{\varepsilon}\left\langle \nabla\Psi\left(\frac{Q_a y}{\varepsilon}\right), Q_a^2 y\right\rangle = \left\langle \nabla\Psi\left(\frac{Q_a y}{\varepsilon}\right), \frac{Q_a y}{\varepsilon}\right\rangle \geqslant 0,$$

since $Q_a^2 = Q_a$. $\square$

4. Shrinkage estimation

Known σ^2. Suppose first that $\sigma^2 = 1$ is known, so that $y \sim \text{Normal}(\theta, I_n)$ for some unknown $\theta \in \Omega$. Let g be an absolutely continuous, non-decreasing function on $[0,\infty)$ for which $0 \leqslant g \leqslant 2$; and let

$$\hat{\theta}_\varepsilon = \tilde{\theta}^\circ + \left[1 - (D_\varepsilon - 2)_+ \frac{g(\|\tilde{\theta} - \tilde{\theta}^\circ\|^2)}{\|\tilde{\theta} - \tilde{\theta}^\circ\|^2}\right](\tilde{\theta} - \tilde{\theta}^\circ),$$

where $(D_\varepsilon - 2)_+ := \max\{0, D_\varepsilon - 2\}$.

Alternatively,

$$\hat{\theta}_\varepsilon(y) = \tilde{\theta}(y) - h_\varepsilon(y), \tag{9}$$

where

$$h_\varepsilon(y) = (D_\varepsilon(y) - 2)_+ \frac{g(\|\tilde{\theta} - \tilde{\theta}^\circ\|^2)}{\|\tilde{\theta} - \tilde{\theta}^\circ\|^2}(\tilde{\theta} - \tilde{\theta}^\circ) = (D_\varepsilon(y) - 2)_+ \frac{g(\|Q_a y\|^2)}{\|Q_a y\|^2} Q_a y,$$

and

$$\|h_\varepsilon(y)\|^2 = [(D_\varepsilon(y) - 2)_+{}^2]\frac{g^2(\|Q_a y\|^2)}{\|Q_a y\|^2},$$

on B_a for each a. When $\varepsilon = 0$, $\hat{\theta}_\varepsilon$ and h_ε will be denoted simply by $\hat{\theta}$ and h. A role of g here is to ensure that $\hat{\theta}_\varepsilon \in \Omega$. For example, if $\ell \geqslant 3$ and $g(w) = \min\{2, w/(\ell - 2)\}$, then

$$\left[1 - (D_\varepsilon - 2)_+ \frac{g(\|\tilde{\theta} - \tilde{\theta}^\circ\|^2)}{\|\tilde{\theta} - \tilde{\theta}^\circ\|^2}\right] \geqslant 0,$$

and, therefore, $\hat{\theta}_\varepsilon \in \Omega$ (since $0 \leqslant D_\varepsilon \leqslant D \leqslant \ell$ and $\tilde{\theta}, \tilde{\theta}^\circ \in \Omega$).

Theorem 1. *Let $\ell \geqslant 3$ be an integer and let g be a non-decreasing, absolutely continuous function for which $0 \leqslant g \leqslant 2$, and $0 < g < 2$ on a set of positive Lebesgue measure. Then $E_\theta\|\hat{\theta}_\varepsilon - \theta\|^2 < E_\theta\|\tilde{\theta} - \theta\|^2$ for all $\theta \in \Omega$ for each $\varepsilon \geqslant 0$.*

Proof. Suppose first that $\varepsilon > 0$. Then,

$$E_\theta\|\hat{\theta}_\varepsilon - \theta\|^2 = E_\theta\|(\tilde{\theta} - \theta) - h_\varepsilon(y)\|^2 = E_\theta\|\tilde{\theta} - \theta\|^2 - E_\theta[2\langle\tilde{\theta} - \theta, h_\varepsilon(y)\rangle - \|h_\varepsilon(y)\|^2],$$

from (9). Here

$$\langle\tilde{\theta} - \theta, h_\varepsilon(y)\rangle = (D_\varepsilon - 2)_+ \frac{g(\|\tilde{\theta} - \tilde{\theta}^\circ\|^2)}{\|\tilde{\theta} - \tilde{\theta}^\circ\|^2}\langle\tilde{\theta} - \theta, \tilde{\theta} - \tilde{\theta}^\circ\rangle$$

and

$$\langle\tilde{\theta} - \theta, \tilde{\theta} - \tilde{\theta}^\circ\rangle = \langle y - \theta, \tilde{\theta} - \tilde{\theta}^\circ\rangle + \langle\tilde{\theta} - y, \tilde{\theta} - \tilde{\theta}^\circ\rangle = \langle y - \theta, \tilde{\theta} - \tilde{\theta}^\circ\rangle,$$

since $\langle\tilde{\theta} - y, \tilde{\theta} - \tilde{\theta}^\circ\rangle = 0$, by (6). So, $\langle\tilde{\theta} - \theta, h_\varepsilon(y)\rangle = \langle y - \theta, h_\varepsilon(y)\rangle$, and

$$E_\theta[\langle\tilde{\theta} - \theta, h_\varepsilon(y)\rangle] = E_\theta[\langle y - \theta, h_\varepsilon(y)\rangle] = E_\theta[\mathrm{div}(h_\varepsilon)],$$

by Stein's Identity, Stein (1981). Collecting terms,

$$\Delta_\varepsilon(\theta) := E_\theta\|\hat{\theta}_\varepsilon - \theta\|^2 - E_\theta\|\tilde{\theta} - \theta\|^2 = -E_\theta[2\mathrm{div}(h_\varepsilon)(y) - \|h_\varepsilon(y)\|^2]. \tag{10}$$

On any B_a, the ith component of h_ε is

$$h_{\varepsilon,i}(y) = (D_\varepsilon(y) - 2)_+ \frac{g(\|Q_a y\|^2)}{\|Q_a y\|^2}(Q_a y)_i,$$

and $(Q_a y)_j = \sum_{k=1}^n q_{jk}^a y_k$, where the jth element of $Q_a y$ and the (j,k)th element of Q_a have been denoted by $(Q_a y)_j$ and q_{jk}^a. So, $\partial(Q_a y)_j/\partial y_i = q_{ji}^a$ and

$$\frac{\partial}{\partial y_i}\|Q_a y\|^2 = 2\sum_{j=1}^n q_{ji}^a (Q_a y)_j,$$

on the interior of any B_a and, therefore, almost everywhere. Of course, $h_{\varepsilon,i}(y) = 0$ if $D_\varepsilon(y) \leqslant 2$. On $B_a^o \cap \{D_\varepsilon > 2\}$,

$$\begin{aligned}\frac{\partial}{\partial y_i}h_{\varepsilon,i}(y) = {} & \left[\frac{\partial}{\partial y_i}D_\varepsilon(y)\right]\frac{g(\|Q_a y\|^2)}{\|Q_a y\|^2}(Q_a y)_i + 2(D_\varepsilon(y) - 2)_+\left[\frac{g'(\|Q_a y\|^2)}{\|Q_a y\|^2} - \frac{g(\|Q_a y\|^2)}{\|Q_a y\|^4}\right] \\ & \times\left[\sum_{j=1}^n q_{ji}^a (Q_a y)_j\right](Q_a y)_i + (D_\varepsilon(y) - 2)_+ \frac{g(\|Q_a y\|^2)}{\|Q_a y\|^2} q_{ii}^a.\end{aligned}$$

So,

$$\begin{aligned}\mathrm{div}h_\varepsilon(y) = {} & \langle\nabla D_\varepsilon(y), Q_a y\rangle\frac{g(\|Q_a y\|^2)}{\|Q_a y\|^2} + 2(D_\varepsilon(y) - 2)_+\left[\frac{g'(\|Q_a y\|^2)}{\|Q_a y\|^2} - \frac{g(\|Q_a y\|^2)}{\|Q_a y\|^4}\right] \\ & \times\left[\sum_{i=1}^n\sum_{j=1}^n q_{ji}^a (Q_a y)_j (Q_a y)_i\right] + (D_\varepsilon(y) - 2)_+ \frac{g(\|Q_a y\|^2)}{\|Q_a y\|^2} D(y).\end{aligned}$$

The double sum here is

$$\sum_{j=1}^n\sum_{i=1}^n q_{ji}^a (Q_a y)_j (Q_a y)_i = \sum_{j=1}^n (Q_a y)_j (Q_a^2 y)_j = \|Q_a y\|^2,$$

since $Q_a^2 = Q_a$. Moreover, $g' \geqslant 0$ since g is non-decreasing, and $\langle \nabla D_\varepsilon(y), Q_a y\rangle \geqslant 0$ by Lemma 2. So,

$$\begin{aligned}\operatorname{div} h_\varepsilon(y) &\geqslant -2(D_\varepsilon(y)-2))_+ \frac{g(\|Q_a y\|^2)}{\|Q_a y\|^2} + (D_\varepsilon(y)-2)_+ \frac{g(\|Q_a y\|^2)}{\|Q_a y\|^2} D(y)\\ &\geqslant \left[(D_\varepsilon(y)-2)_+{}^2\right] \frac{g(\|Q_a y\|^2)}{\|Q_a y\|^2},\end{aligned}$$

since $D \geqslant D_\varepsilon > 2$. It follows that

$$-2\operatorname{div} h_\varepsilon(y) + \|h_\varepsilon(y)\|^2 \leqslant -\left[(D_\varepsilon(y)-2)_+{}^2\right] \frac{g(\|Q_a y\|^2)[2-g(\|Q_a y\|^2)]}{\|Q_a y\|^2},$$

which is non-positive everywhere and negative on a set of positive Lebesgue measure. So, recalling (10)

$$\Delta_\varepsilon(\theta) \leqslant -E_\theta\left\{\left[(D_\varepsilon(y)-2)_+{}^2\right] \frac{g(\|Q_a y\|^2)[2-g(\|Q_a y\|^2)]}{\|Q_a y\|^2}\right\} < 0, \tag{11}$$

for any $\theta \in \Omega$. This establishes the theorem for $\varepsilon > 0$. The case $\varepsilon = 0$, then follows from Fatou's Lemma. □

Remarks. (1) The conditions that $g \leqslant 2$ and $g < 2$ on a set of positive Lebesgue measure were used only in the final step. Thus, the first inequality in (11) is valid, provided only that g is bounded.

(2) The proof shows that $E_\theta\|\hat{\theta}_\varepsilon - \theta\|^2 < E_\theta\|\tilde{\theta} - \theta\|^2$ for all $\theta \in \mathbb{R}^n$.

Unknown σ^2. The theorem is easily extended to the case where $y \sim$ Normal $[\theta, \sigma^2 I_n]$ for some unknown $\theta \in \Omega$ and $\sigma^2 > 0$, provided that there is an independent estimator S^2 of σ^2 for which $S^2 \sim \sigma^2\chi_r^2/r$ for some (known) $r \geqslant 1$. For this case, let

$$\hat{\theta}_\varepsilon = \tilde{\theta}^\circ + \left[1 - \frac{(D_\varepsilon - 2)_+{}^2}{Z} g(Z)\right](\tilde{\theta} - \tilde{\theta}^\circ), \quad \text{where } Z = \frac{\|\tilde{\theta} - \tilde{\theta}^\circ\|^2}{S^2}.$$

Theorem 2. *Let $\ell \geqslant 3$ be an integer; and let g be a non-decreasing, absolutely continuous function for which $g(x)/x$ is non-increasing and $0 \leqslant g < 2r/(r+2)$ with strict inequality on a set of positive Lebesgue measure. Then $E_{\theta,\sigma^2}\|\hat{\theta}_\varepsilon - \theta\|^2 < E_{\theta,\sigma^2}\|\tilde{\theta} - \theta\|^2$ for all $\theta \in \Omega$ and $\sigma^2 > 0$ for each $\varepsilon \geqslant 0$.*

Proof. By scale equivariance, it suffices to establish the inequality when $\sigma^2 = 1$. For this case, write E_θ for $E_{\theta,1}$ and define Δ_ε by (10), but with the new definition of $\hat{\theta}_\varepsilon$. Then conditioning on S^2 in the proof of Theorem 1 leads to

$$\Delta_\varepsilon(\theta) \leqslant -E_\theta\left\{\frac{(D_\varepsilon - 2)_+{}^2}{\|\tilde{\theta} - \tilde{\theta}^\circ\|^2} S^2 g(Z)[2 - S^2 g(Z)]\right\}$$

for any $\varepsilon > 0$. Remark 1 is relevant here. Next, conditioning on y,

$$E_\theta\left\{\frac{(D_\varepsilon - 2)_+^{\,2}}{\|\tilde{\theta} - \tilde{\theta}^\circ\|^2} S^2 g(Z)[2 - S^2 g(Z)]\middle| y\right\} = \frac{(D_\varepsilon - 2)_+^{\,2}}{\|\tilde{\theta} - \tilde{\theta}^\circ\|^2} E[2f_y(W) - \frac{1}{r} W f_y(W)^2],$$

where $W \sim \chi^2_{r+2}$ and $f_y(w) = g(t^2/w)$ with $t^2 = r\|\tilde{\theta} - \tilde{\theta}^\circ\|^2$. By assumption, $f_y(w)$ is non-increasing in w and $wf_y(w)$ is non-decreasing. So, $E[Wf_y(W)^2] \leqslant E[Wf_y(W)]E[f_y(W)]$ and, therefore,

$$\Delta_\varepsilon(\theta) \leqslant -E_\theta\left\{\frac{(D_\varepsilon - 2)_+^{\,2}}{\|\tilde{\theta} - \tilde{\theta}^\circ\|^2} E[f_y(W)|y]\left[2 - \frac{1}{r} E[Wf_y(W)]\right]\right\}.$$

The integrand on the right is non-negative, because

$$E[Wf_y(W)] \leqslant \frac{2r}{r+2} E(W) = 2r,$$

and positive with positive probability. This establishes the theorem for $\varepsilon > 0$, and the case $\varepsilon = 0$ follows from Fatou's Lemma. □

Acknowledgements

This research was supported in part by the National Security Agency, US Army Research Office and Rackham Interdisciplinary Grant.

References

Bock, M.E., 1982. Employing vague inequality information in the estimation of normal mean vectors (estimators that shrink to closed convex polyhedra). Statistical Decision Theory and Related Topics, III, vol. 1 (West Lafayette, Ind., 1981). Academic Press, New York, pp. 169–193.

Fraser, D., Massam, H., 1989. A mixed primal–dual bases algorithm for regression under inequality constraints. application to concave regression. Scand. J. Statist. 16 (1), 65–74.

Meyer, M., 1999. An extension of the mixed primal–dual bases algorithm to the case of more constraints than dimensions. J. Statist. Plann. Inference 81 (1), 13–31.

Meyer, M., Woodroofe, M., 2000. On the degrees of freedom in shape-restricted regression. Ann. Statist. 28 (4), 1083–1104.

Ouassou, I., Strawderman, W.E., 2002. Estimation of a parameter vector restricted to a cone. Statist. and Probab. Lett. 56, 121–129.

Sengupta, D., Sen, P.K., 1991. Shrinkage estimation in a restricted parameter space. Sankhya, Ser. A 53, 389–411.

Stein, C., 1981. Estimation of the mean of a multivariate normal distribution. Ann. Statist. 9 (6), 1135–1151.

Wegman, E., 1970. Maximum likelihood estimation of a unimodal density function. Ann. Math. Statist. 41, 457–471.

Woodroofe, M., Sun, J., 1993. A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing. Statist. Sinica 3 (2), 501–515.

Wright, F.T., 1978. Estimating strictly increasing regression functions. J. Amer. Statist. Assoc. 73 (363), 636–639.

Wright, F.T., 1982. Monotone regression estimates for grouped observations. Ann. Statist. 10, 278–286.

Wu, W.B., Woodroofe, M., Mentz, G., 2001. Isotonic regression: another look at the change point problem. Biometrika 88 (3), 793–804.