Andrej Pazman

Elimination of less informative design points in regression models with a known or parametrized covariance function

Paper

Original Citation:

Pazman, Andrej
(2005)
Elimination of less informative design points in regression models with a known or parametrized covariance function.


This version is available at: https://epub.wu.ac.at/856/
Available in ePubWU: June 2005

ePubWU, the institutional repository of the WU Vienna University of Economics and Business, is provided by the University Library and the IT-Services. The aim is to enable open access to the scholarly output of the WU.
Elimination of Less Informative Design Points in Regression Models with a Known or Parametrized Covariance Function

Andrej Pázman

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series
Report 18
June 2005

http://statistik.wu-wien.ac.at/
ELIMINATION OF LESS INFORMATIVE DESIGN POINTS
IN REGRESSION MODELS WITH A KNOWN
OR PARAMETRIZED COVARIANCE FUNCTION

Andrej Pázman, Comenius University, Bratislava
pazman@fmph.uniba.sk

Abstract

We consider a regression model $E[y(x)] = \eta(\theta, x)$ where $x$ is a design point taken from a finite design space $X$. The covariance of observations is $\text{Cov}[y(x), y(x^*]) = C(x, x^*, \beta)$. Here $\theta, \beta$ are unknown vector parameters. The quality of the ML estimators of $\theta$ and $\beta$ is measured by optimality criteria applied on the Fisher information matrix taken at a fixed $\theta, \beta$ (=local optimality). In this paper we give formulae to identify the design points which have little influence on this quality. We also propose a simple algorithm which is deleting such points and leads to a better (not necessarily optimum) design.

1. Introduction

We consider a regression model of the form

$$y(x_i) = \eta(\theta, x_i) + \epsilon_i$$

with the points $x_1, ..., x_N$ (=the design) taken from a finite set $X$ (= the design space), with the vector parameter $\theta = (\theta_1, ..., \theta_p)^T$ being unknown and taken from the parameter space $\Theta$, and with $\eta(.)$ a known function. The model is supposed to be without systematic errors (i.e. $E(\epsilon_i) = 0$), its covariance structure is supposed to be related to the positions $x_1, ..., x_N$ of the observations,

$$\text{Cov}(y(x_i), y(x_j)) = C(x_i, x_j, \beta)$$

The function $C(.)$ is supposed to be known, with another unknown vector parameter $\beta = (\beta_1, ..., \beta_q)^T \in B$.

Notation. Any set $A \subset X$ corresponds to a design, with observations in each point of $A$ and without replications. Let us denote by $N_A$ the number of points in $A$. For any $A, B \subset X$ we denote by $C(A, B, \beta)$ the $N_A \times N_B$ matrix with entries $\{C(A, B, \beta)\}_{x,z} = C(x, z, \beta) ; x \in A, z \in B$, and $C(A, \beta) = C(A, A, \beta)$. We suppose here that the matrix $C(X, \beta)$ is
nonsingular, consequently \( C(A, \beta) \) is nonsingular for any \( A \subset X \). Further, we use the notation
\[
G(X, \beta) = C^{-1}(X, \beta)
\]
and similarly we define the \( N_A \times N_B \) matrix \( \{ G(A, B, \beta) \}_{x,z} = \{ G(X, \beta) \}_{x,A,z,B} \), and the \( N_A \times N_A \) matrix \( G(A, \beta) = G(A, A, \beta) \), which are submatrices of \( G(X, \beta) \). So we must distinguish between \( G(A, \beta) \), which is a submatrix of \( G(X, \beta) = C^{-1}(X, \beta) \), and \( C^{-1}(A, \beta) \), which is the inverse of \( C(A, \beta) \), a submatrix of \( C(X, \beta) \).

**The information matrix.** For normally distributed errors of observations having small variances, the mean square error matrix of the maximum likelihood estimator of \( (\theta, \beta) \) can be approximated by \( [M(A, \theta, \beta)]^{-1} \), even for small numbers of observations (cf. Pázmán (2004)). Here \( M(A, \theta, \beta) \) is the Fisher information matrix of the design \( A \)
\[
M(A, \theta, \beta) = \begin{pmatrix}
M_I(A, \theta, \beta) & 0 \\
0 & M_{II}(A, \beta)
\end{pmatrix}
\]  
with
\[
M_I(A, \theta, \beta) = \sum_{x,z \in A} f_\theta(x) \left\{ C^{-1}(A, \beta) \right\}_{x,z} f_\theta^T(z)
\]
\[
f_\theta(x) = \frac{\partial \eta(x, \theta)}{\partial \theta}
\]
and
\[
\{ M_{II}(A, \beta) \}_{ij} = \frac{1}{2} \text{tr} \left\{ C^{-1}(A, \beta) \frac{\partial C(A, \beta)}{\partial \beta_i} C^{-1}(A, \beta) \frac{\partial C(A, \beta)}{\partial \beta_j} \right\}
\]
In the present paper we shall express in a convenient form the differences \( M(\theta, \beta, X) - M(\theta, \beta, A) \), \( M_I(\theta, \beta, X) - M_I(\theta, \beta, A) \), \( M_{II}(\theta, \beta, X) - M_{II}(\theta, \beta, A) \) as well as the difference \( \Phi[M(\theta, \beta, X)] - \Phi[M(\theta, \beta, A)] \) for differentiable optimality criteria \( \Phi \). In particular, we are interested in the case that \( X - A = \{x_0\} \), a one-point set.

*Since in the following exposition the values of \( \theta \) and \( \beta \) are fixed, we omit to write them along the whole paper. So we write simply \( C(A) \), \( \frac{\partial C(A)}{\partial \beta} \), \( G(A) \), \( M_I(A) \), \( M_{II}(A) \), \( f(x) \), etc.*

2. **The case of estimation of \( \theta \).**
In this section the Fisher information matrix of the design $A$ is equal to $M_I(A)$.

**Lemma 1.** The following equalities hold

$$C(\mathcal{X}, A) C^{-1}(A) C(A, \mathcal{X})$$

$$= C(\mathcal{X}) - \begin{pmatrix} 0 & C(A) - C(A^c, A) C^{-1}(A) C(A, A^c) \\ 0 & 0 \end{pmatrix}$$

$$= C(\mathcal{X}) - \begin{pmatrix} 0 & 0 \\ 0 & G^{-1}(A^c) \end{pmatrix}$$

where $A^c = \mathcal{X} - A$.

**Proof.** We decompose the matrix $C(\mathcal{X})$ into blocks

$$C(\mathcal{X}) = \begin{pmatrix} C(A) & C(A, A^c) \\ C(A^c, A) & C(A^c) \end{pmatrix}$$

and verify by a direct multiplication that

$$C(\mathcal{X}, A) C^{-1}(A) C(A, \mathcal{X})$$

$$= C(\mathcal{X}) \begin{pmatrix} C^{-1}(A) & 0 \\ 0 & 0 \end{pmatrix} C(\mathcal{X})$$

$$= \begin{pmatrix} C(A) & C(A, A^c) \\ C(A^c, A) & C(A^c, A) C^{-1}(A) C(A, A^c) \end{pmatrix}$$

$$= C(\mathcal{X}) - \begin{pmatrix} 0 & 0 \\ 0 & C(A^c) - C(A^c, A) C^{-1}(A) C(A, A^c) \end{pmatrix}$$

This proves the first equality. The second follows from

$$G^{-1}(A^c) = C(A^c) - C(A^c, A) C^{-1}(A) C(A, A^c)$$

(cf. theorem 8.5.11 in Harville (1997)).

**Lemma 2.** If there is a set $D \subset \mathcal{X}$ and a mapping $a: u \in D \rightarrow a(u) \in \mathbb{R}^p$ such that for every $x \in \mathcal{X}$

$$f(x) = \sum_{z \in D} C(x, z) a(z)$$

(4)
then
\[ M_I (A) = M_I (X) \]
for every \( A \supset D \).

(Cf. Näther (1985), where also examples are given, or Theorem 1 below which gives an alternative proof.) Consequently, when the aim is to estimate \( \theta \), we can reject from the design space all design points which are belonging to \( D^c \). They are non-informative, i.e. bring no information about \( \theta \), whatever is the optimality criterion that we consider.

Let us define for every \( u \in X \)
\[ a (u) = \sum_{x \in X} \{ G (X) \} \sum_{x} f (x) \]  
which corresponds to
\[ f (x) = \sum_{u \in X} C (x, u) a (u) \]
We can write
\[ M_I (X) = \sum_{u,v \in X} a (u) C (u, v) a^T (v) \]  
We see that the assumption (4) is equivalent to the assumption
\[ a (u) = 0 \text{ for every } u \in D^c \]

**Lemma 3.** For any differentiable convex optimality criterion \( \Phi (M) \) the gradient \( \nabla \Phi (M) \) with components
\[ \{ \nabla \Phi (M) \}_{ij} = \frac{\partial \Phi (M)}{\partial \{ M \}_{ij}} \]
is a negative semidefinite matrix.

For a proof cf. Müller & Pázman (2003). This allows to define a norm (or a pseudonorm) of \( a (u) \)
\[ \| a (u) \|_{\Phi}^2 = -a^T (u) [\nabla \Phi (M)]_{M_I (X)} a (u) \]  
For example, for the criterion of D-optimality when \( \Phi (M) = -\ln \det (M) \), and \( \nabla \Phi (M) = -M^{-1} \), we have \( \| a (u) \|_{\Phi}^2 = a^T M^{-1} a \). The norm appeared in certain algorithms for computing \( \Phi \)-optimal or approximate \( \Phi \)-optimal...
designs. This indicates the conjecture that the vector $a(u)$ measures the importance of the point $u \in X$ in a design.

Here we want to support this idea by further arguments. We can write, using Lemma 1

$$M_I(A) = \sum_{x,z \in A} \left[ \sum_{u \in X} a(u) C(u, x) \right] \left[ C^{-1}(A) \right]_{x,z} \left[ \sum_{v \in A} C(z, v) a^T(v) \right]$$

$$= \sum_{u \in A} a(u) \left[ C(X, A) C^{-1}(A) C(A, X) \right]_{u,v} a^T(v)$$

$$= \sum_{u,v \in A} a(u) C(u, v) a^T(v) - \sum_{u \in A, v \in A^c} a(u) \left\{ G^{-1}(A^c) \right\}_{u,v} a^T(v)$$

$$= M_I(X) - \sum_{u,v \in A^c} a(u) \left\{ C^{cond}(A^c) \right\}_{uv} a^T(v)$$

Notice also that when $A^c = \{x_o\}$ is a one point set, then $G(A^c) = \{G(X)\}_{x_o,x_o}$, hence $[G(A^c)]^{-1} = 1/\{G(X)\}_{x_o,x_o}$.

So we proved the following statement.

**Theorem 1.** For every design $A \subset X$ we have

$$M_I(X) - M_I(A) = \sum_{u \in A^c, v \in A^c} a(u) \left\{ G^{-1}(A^c) \right\}_{u,v} a^T(v)$$

which expresses in terms of $a(u)$ the loss of information when using the design $A$. This difference is always nonnegative definite. In the particular case that $A^c = \{x_o\}$ we have the simple expression

$$M_I(X) - M_I(A) - \sum_{u \in A^c, v \in A^c} a(u) \left\{ C^{cond}(A^c) \right\}_{uv} a^T(v)$$

where

$$C^{cond}(A^c) = G^{-1}(A^c) = C(A^c) - C(A^c, A) [C(A)]^{-1} C(A, A^c)$$

Notice that Lemma 2 is a corollary of Theorem 1, i.e.

$$\{ \forall u \in A^c \ a(u) = 0 \} \Rightarrow M_I(X) = M_I(A)$$

We can also write instead of (8)

$$M_I(X) - M_I(A) = \sum_{u,v \in A^c} a(u) \left\{ C^{cond}(A^c) \right\}_{uv} a^T(v)$$

Notice that Lemma 2 is a corollary of Theorem 1, i.e.

$$\{ \forall u \in A^c \ a(u) = 0 \} \Rightarrow M_I(X) = M_I(A)$$

We can also write instead of (8)

$$M_I(X) - M_I(A) = \sum_{u,v \in A^c} a(u) \left\{ C^{cond}(A^c) \right\}_{uv} a^T(v)$$

where

$$C^{cond}(A^c) = G^{-1}(A^c) = C(A^c) - C(A^c, A) [C(A)]^{-1} C(A, A^c)$$

5
is the conditional covariance matrix of the vector of observations \((y (x) : x \in A^c)\) given the complementary subvector \((y (x) : x \in A)\). We see that the difference \(M_I (\mathcal{X}) - M_I (A)\) is small when the conditional covariance \(C_{\text{cond}} (A^c)\) is small (i.e. when knowing the observations \(y (x)\) at points \(x \in A\) we can very precisely predict the values of \(y (x)\) at \(x \in A^c\), so observations at \(x \in A^c\) are almost unnecessary), or when the vectors \(a (u); u \in A^c\) are small. In both cases we can reject observations in points of \(A^c\).

To consider the influence of \(a (u)\) on the value of an optimality criterion \(\Phi\) we write

\[
\Phi \left[ M_I (A) \right] = \Phi \left[ M_I (\mathcal{X}) - V_{\text{cond}} (A^c) \right]
\]

where

\[
V_{\text{cond}} (A^c) = \sum_{u \in A^c, v \in A^c} a (u) C_{\text{cond}} (A^c) a^T (v)
\]

is the conditional variance matrix of the (normal) random vector

\[
\sum_{x \in A^c} y (x) a (x)
\]

given the values of the observations \(y (x) ; x \in A\).

If \(V_{\text{cond}} (A^c)\) is small, we can use the linear Taylor expansion of \(\Phi (M)\) at \(M = M_I (X)\). So we obtain from (8)

**Corollary of Theorem 1.**

\[
0 \leq \Phi \left[ M_I (A) \right] - \Phi \left[ M_I (\mathcal{X}) \right] \doteq tr \left\{ V_{\text{cond}} (A^c) \nabla \Phi (M)_{M=M_I (\mathcal{X})} \right\}
\]

\[
= \sum_{u, v \in A^c} < a (u), a (v)>_\Phi \left\{ G^{-1} (A^c) \right\}_{uv}
\]

where

\[
< a (u), a (v)>_\Phi = -a^T (u) \left[ \nabla \Phi (M)_{M=M_I (\mathcal{X})} \right] a (v)
\]

is the inner product (or pseudoproduct) corresponding to the norm \(\| a (u) \|_\Phi\).

In particular when \(A^c = \{x_o\}\) we have

\[
\Phi \left[ M_I (\mathcal{X} - \{x_o\}) \right] \doteq \Phi \left[ M_I (\mathcal{X}) \right] + \frac{\| a (x_o) \|_\Phi^2}{\{G(\mathcal{X})\}_{x_o, x_o}} \tag{10}
\]
It is easy to find \( x_o \in \mathcal{X} \) which minimizes these expressions. This can be used to construct the following very simple algorithm for finding a design \( A \) having a smaller value of \( \Phi (M_I (A)) \):

a) Compute the matrix \( G (\mathcal{X}) = C^{-1} (\mathcal{X}) \), the vectors \( a(x) \) according to (5) and \( \|a (x)\|_\Phi^2 \) according to (7) for every \( x \in \mathcal{X} \).

b) Exclude from \( \mathcal{X} \) the point \( x_o \), which is approximately the less informative under the design \( \mathcal{X} \), i.e. which minimizes

\[
\frac{\|a (x)\|_\Phi^2}{\{G (\mathcal{X})\}_{x,x}}
\]

c) Denote \( \mathcal{X} - \{x_o\} \) by \( \mathcal{X} \) and go back to point a), or stop if the number of points in \( \mathcal{X} - \{x_o\} \) is equal to \( N = \) the required size of the design.

3. The case of estimation of \( \beta \)

In this section we shall consider the difference \( M_{II} (\mathcal{X}) - M_{II} (A) \).

**Theorem 2.** For any design \( A \subset \mathcal{X} \) we have

\[
\{M_{II} (\mathcal{X})\}_{ij} - \{M_{II} (A)\}_{ij} = \text{tr} \left\{ \frac{\partial C (\mathcal{X})}{\partial \beta_i} \Psi_A (\mathcal{X}) \frac{\partial C (\mathcal{X})}{\partial \beta_j} \left[ G (\mathcal{X}) - \frac{1}{2} \Psi_A (\mathcal{X}) \right] \right\}
\]

(11)

where

\[
\Psi_A (\mathcal{X}) = G (\mathcal{X}, A^c) G^{-1} (A^c) G (A^c, \mathcal{X})
\]

In the particular case that \( A^c = \{x_o\} \) we have

\[
\{M_{II} (\mathcal{X})\}_{ij} - \{M_{II} (\mathcal{X} - \{x_o\})\}_{ij} = \frac{1}{\{G (\mathcal{X})\}_{x_o,x_o}} \left[ G (\{x_o\}, \mathcal{X}) \frac{\partial C (\mathcal{X})}{\partial \beta_i} G (\mathcal{X}, \{x_o\}) \frac{\partial C (\mathcal{X})}{\partial \beta_j} G (\mathcal{X}, \{x_o\}) \right]
\]

\[
- \frac{1}{2 \{G (\mathcal{X})\}_{x_o,x_o}^2} \left[ G (\{x_o\}, \mathcal{X}) \frac{\partial C (\mathcal{X})}{\partial \beta_i} G (\mathcal{X}, \{x_o\}) \right] \times \left[ G (\{x_o\}, \mathcal{X}) \frac{\partial C (\mathcal{X})}{\partial \beta_j} G (\mathcal{X}, \{x_o\}) \right]
\]

(12)
**Proof.**

Putting into Lemma 1 \( G(X) \) instead of \( C(X) \) and \( A^c \) instead of \( A \) we obtain

\[
\Psi_A(X) = G(X) - \begin{pmatrix} C^{-1}(A) & 0 \\ 0 & 0 \end{pmatrix}
\]

Hence

\[
G(X) - \frac{1}{2} \Psi_A(X) = \frac{1}{2} \begin{pmatrix} C^{-1}(A) & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{2} G(X)
\]

Thus the right-hand side of (11) is equal to

\[
\frac{1}{2} \text{tr} \left\{ \frac{\partial}{\partial \beta_i} \left[ G(X) - \begin{pmatrix} C^{-1}(A) & 0 \\ 0 & 0 \end{pmatrix} \right] \frac{\partial}{\partial \beta_j} \left[ G(X) + \begin{pmatrix} C^{-1}(A) & 0 \\ 0 & 0 \end{pmatrix} \right] \right\}
\]

\[
= \frac{1}{2} \text{tr} \left\{ \frac{\partial C(X)}{\partial \beta_i} C^{-1}(X) \frac{\partial C(X)}{\partial \beta_j} C^{-1}(X) \right\}
\]

\[
- \frac{1}{2} \text{tr} \left\{ \frac{\partial C(A)}{\partial \beta_i} C^{-1}(A) \frac{\partial C(A)}{\partial \beta_j} C^{-1}(A) \right\}
\]

\[
= \{M_{II}(X)\}_{ij} - \{M_{II}(A)\}_{ij}
\]

**Remark** It can useful to formulate these results in a form which is symmetric to the statements in Theorem 1, if we introduce for each \( u, v \in X \) the vector

\[
b(u, v) = - \frac{\partial \{G(X)\}_{u,v}}{\partial \beta} = \left\{ C^{-1}(X) \frac{\partial C(X)}{\partial \beta} C^{-1}(X) \right\}_{u,v}
\]

which is a counterpart of the vector \( a(u) \) defined in (5). We see that

\[
M_{II}(X) = \frac{1}{2} \sum_{u,v,u^*,v^* \in X} b(v, u) C(u, v^*) b^T(v^*, u^*) C(u^*, v)
\]

(compare with (6) in Section 2). The reformulation of Theorem 2 can be done for every \( A^c \). We present here just the case when \( A^c \) is a one-point set.

**Corollary 1.** We have

\[
M_{II}(X) - M_{II}(X - \{x_o\}) = \frac{1}{\{G(X)\}_{x_o,x_o}} \sum_{u,v \in X} b(x_o, u) C(u, v) b^T(v, x_o)
\]

\[
- \frac{1}{2} \left[ \{G(X)\}_{x_o,x_o} \right]^{-1} b(x_o, x_o) b^T(x_o, x_o)
\]
Now we shall consider the changes of the criterion when deleting one point \( \{x_o\} \). By the linear Taylor expansion we have
\[
\Phi [M_{II}(A)] = \Phi [M_{II}(X)] + tr \left\{ \nabla \Phi (M)_{M=M_{II}(X)} [M_{II}(A) - M_{II}(X)] \right\}
\]

Using corollary of Theorem 2 we obtain
\[
\Phi [M_{II}(A)] = \Phi [M_{II}(X)] + \frac{1}{2} \left\{ \left\{ G(X) \right\}_{x_o,x_o} \right\} b^T (x_o, x_o) \left[ \nabla \Phi (M)_{M=M_{II}(X)} \right] b (x_o, x_o)
\]
\[
- \frac{1}{\left\{ G(X) \right\}_{x_o,x_o}} \sum_{u,v \in X} C(u,v) b^T (v, x_o) \left[ \nabla \Phi (M)_{M=M_{II}(X)} \right] b (x_o, u)
\]

So, we have the following corollary of Theorem 2

**Corollary 2.** When the rejected point \( x_o \) is not very informative, we can write
\[
\Phi [M_{II}(X- \{x_o\})] = \Phi [M_{II}(X)] - \frac{1}{2} \left\{ \left\{ G(X) \right\}_{x_o,x_o} \right\} b^T (x_o, x_o) \left\| b (x_o, x_o) \right\|^2 \Phi \\
+ \frac{1}{\left\{ G(X) \right\}_{x_o,x_o}} \sum_{u,v \in X} C(u,v) < b(x_o,v), b(x_o,u) > \Phi
\]

where the inner product and the norm are \(<c,d>_{\Phi} = -c^T \nabla \Phi [M]_{M=M_{II}(X)} d, \\left\| c \right\|^2_{\Phi} = <c,c>_{\Phi} \).

4. **The case when both \( \theta \) and \( \beta \) are estimated**

From the partition of the information matrix given in (2) it follows that
\[
M(X) - M(A) = \begin{pmatrix} M_{I}(X) - M_{I}(A) & 0 \\ 0 & M_{II}(X) - M_{II}(A) \end{pmatrix}
\]

where \( M_{I}(X) - M_{I}(A) \) and \( M_{II}(X) - M_{II}(A) \) are given in Theorems 1 and 2.

Further from (2) it also follows that
\[
\log \det [M(A)] = \log \det [M_{I}(A)] + \log \det [M_{II}(A)]
\]
\[
tr \left\{ M^{-1}(A) \right\} = tr \left\{ [M_{I}(A)]^{-1} + tr \left\{ [M_{II}(A)]^{-1} \right\} \right\}
\]
Hence the criteria of D- and A-optimaliy have the property
\[ \Phi[M(A)] = \Phi[M_I(A)] + \Phi[M_{II}(A)] \]
for any design \( A \subset X \). So
\[
\Phi[M(X - \{x_o\})] - \Phi[M(X)] = \Phi[M_I(X - \{x_o\})] - \Phi[M_I(X)] + \Phi[M_{II}(X - \{x_o\})] - \Phi[M_{II}(X)]
\]
This leads to the following

**Algorithm for improving designs**

a) Compute the matrices
\[
G(X) = C^{-1}(X), \nabla \Phi[M_{M=M_f[X]}], \nabla \Phi[M_{M=M_{II}[X]}],
\]
the vectors \( a(x) \) according to (5) and their squared norms \( \|a(x)\|_\Phi^2 \), the vectors \( b(x, u) \) according to (13), and their inner products \( \langle b(x, v), b(x, u) \rangle_\Phi \) for every \( x, u, v \in X \).
b) Exclude from \( X \) the point \( x_o \), which is approximately the less informative (with respect to the criterion \( \Phi \), and under the design \( X \)), i.e. which minimizes over \( x \in X \) the expression
\[
\frac{\|a(x)\|_\Phi^2}{\{G(X)\}_{x,x}} + \frac{\sum_{u, v \in X} C(u, v) \langle b(x, v), b(x, u) \rangle_\Phi}{\{G(X)\}_{x,x}} - \frac{\|b(x, x)\|_\Phi^2}{2 \{\{G(X)\}_{x,x}\}^2}
\]
c) Stop if the number of points in \( X - \{x_o\} \) is equal to \( N = \) the required size of the design. Otherwise, denote \( X - \{x_o\} \) by \( X \), and go back to point a).

**References**

