On the degrees of freedom in shrinkage estimation

Kengo Kato
Graduate School of Economics, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
kato_ken@hkg.odn.ne.jp

October, 2007

Abstract

We study the degrees of freedom in shrinkage estimation of the regression coefficients. Generalizing the idea of the Lasso, we consider the problem of estimating the coefficients by the projection of the ordinary least squares estimator onto a closed convex set. Then an unbiased estimator of the degrees of freedom is derived in terms of geometric quantities under a smoothness condition on the boundary of the closed convex set. The result presented in this paper is applicable to estimation with a wide class of constraints. As an application, we obtain a C_p-type criterion and AIC for selecting the tuning parameter.

Keywords: AIC, degrees of freedom, fused Lasso, group Lasso, Lasso, Mallows' C_p, second fundamental form, shrinkage estimation, Stein's lemma, tubal coordinates.

Running title: Degrees of freedom in shrinkage estimation

1 Introduction

In recent years, much attention has been paid for shrinkage methods in estimating coefficients of a linear model. Compared with the ordinary least squares (OLS), shrinkage methods often improve the prediction accuracy. In addition, if the constraint region towards which the estimator is shrunk has edges or corners, some coefficients can be set to exactly zero.

To be precise, suppose $y = (y_1, \ldots, y_n)'$ is the response vector and $x_j = (x_{1j}, \ldots, x_{nj})'$, $j = 1, \ldots, p$ are p linearly independent predictors. Let $X = [x_1 \cdots x_p]$ be the design matrix. We consider a linear model

$$y = X\beta + \epsilon, \tag{1.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)'$ is the coefficient vector and $\epsilon \sim N_n(0, \sigma^2 I_n)$. Without loss of generality, we assume that the predictors are centered so that the intercept is not included in the above linear model.

A canonical example of shrinkage methods is the Lasso (Tibshirani [10]). Let $\|\cdot\|_2$ be the ordinary Euclidean norm: $\|z\|_2 = (z'z)^{\frac{1}{2}}$ for $z \in \mathbb{R}^n$. The Lasso estimate is defined as the solution of the following problem:

$$\min_{\beta} \|y - X\beta\|_2^2 \ \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \tag{1.2}$$

or equivalently

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1.3}$$

where t and λ are non-negative tuning parameters. The Lasso shrinks the coefficients towards zero as t decreases or λ increases. An important feature of the Lasso is that, depending on the tuning parameter, some coefficients are set exactly equal to zero. It should be noted that although (1.2) and (1.3) are equivalent as *minimization problems*, the solutions of these two problems are different as *estimators* since the correspondence between t and λ generally depends on the data.

As explained in Efron [2], the *degrees of freedom* plays an important role in selecting the optimal tuning parameter. The degrees of freedom reflects the model complexity controlled by the shrinkage and corresponds to the penalty term of model selection criteria such as Mallows' C_p (Mallows [4]) and Akaike's information criterion (AIC, Akaike [1]). Recently, Zou et al. [15] show that, with parametrization (1.3), the number of non-zero coefficients is an unbiased estimator of the degrees of the freedom of the Lasso. Their derivation, however, requires the local explicit form of the Lasso estimator and can not be applied to estimation with a more general restriction.

The Lasso can be viewed as the projection of the OLS estimator onto the diamond shaped region. For $u, v \in \mathbb{R}^p$, we denote

$$\langle u, v \rangle = u'Vv, \tag{1.4}$$

where $V = X'X$ and let $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. Then the Lasso problem (1.2) is rewritten as

$$\min_{\beta} \|\beta - \hat{\beta}^\circ\| \ \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \tag{1.5}$$

where $\hat{\beta}^\circ$ is the OLS estimator of β. The natural generalization of the minimization problem (1.5) is

$$\min_{\beta} \|\beta - \hat{\beta}^\circ\| \ \text{ subject to } \beta \in K, \tag{1.6}$$

with a closed convex set $K \subset \mathbb{R}^p$. The solution $\hat{\beta}_K$ to the problem (1.6) is given by the projection of $\hat{\beta}^\circ$ onto K. Since K is closed and convex, $\hat{\beta}_K$ is uniquely defined. The problem of selecting the optimal tuning parameter is viewed as the problem of selecting the optimal constraint region K among a given collection of closed convex sets. The class of estimation methods considered here includes the Lasso, the fused Lasso (Tibshirani et al. [11]), and the group Lasso (Yuan and Lin [12]).

Here we present illustrative examples of the constraint regions of the Lasso and the group Lasso. The left of the figures below corresponds to the Lasso constraint $|\beta_1| + |\beta_2| + |\beta_3| \leq 1$. The right one corresponds to the group Lasso constraint $(\beta_1^2 + \beta_2^2)^{\frac{1}{2}} + |\beta_3| \leq 1$.

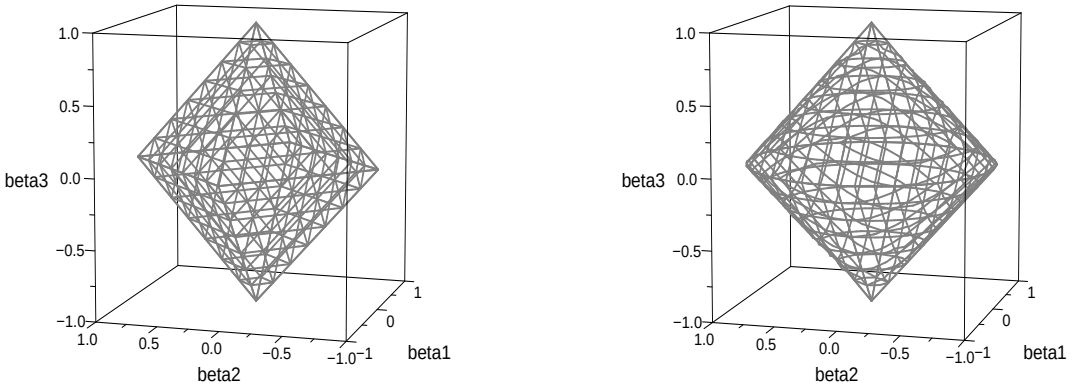

Fig. The constraint regions of the Lasso (left) and the group Lasso (right).

In this paper, we study the degrees of freedom of the fit $\hat{\mu}_K = X\hat{\beta}_K$. From Stein's lemma (Stein [8]), an unbiased estimator of the degrees of freedom is given by the divergence of $\hat{\mu}_K$ with respect to y, which coincides with the divergence of $\hat{\beta}_K$ with respect to $\hat{\beta}^\circ$. However, in general, the estimator $\hat{\beta}_K$ can not be expressed in an explicit form. Thus it is often impossible to directly calculate the divergence.

To overcome this difficulty, we use the idea of the "tubal coordinates" (Weyl [14]). From an approach similar to that of Kuriki and Takemura [3], we derive the divergence of the projection onto K in terms of geometric quantities under a regularity condition on the boundary ∂K of K. Hence we obtain an unbiased estimator of the degrees of freedom of $\hat{\mu}_K$. As an application, a C_p-type statistic and AIC for $\hat{\mu}_K$ are also derived.

The organization of this paper is as follows. In Section 2, we briefly review Stein's unbiased risk theory. In Section 3, we first prepare notations of geometry of a piecewise smooth boundary of a closed convex set and derive a divergence formula for the projection onto K from the differential geometric approach. An unbiased estimator of the degrees of freedom of $\hat{\mu}_K$ is provided in Section 3.2. The result presented in this paper seems to be fairly general. In Section 4, we exemplify our method to obtain unbiased estimators of the degrees of freedom for the Lasso and its variants. Section 5 is devoted to some concluding remarks.

2 Unbiased estimation of the prediction risk

In this section, according to Efron [2], we first introduce Stein's unbiased risk estimation theory. The precise definition of the degrees of freedom is given. Then we explain the strategy to derive an unbiased estimator of the degrees of freedom for the estimator defined by the solution to the minimization problem (1.6).

Given a fit $\hat{\mu} = \hat{\mu}(y) = X\hat{\beta}$ where $\hat{\beta}$ is an estimator of β, we focus on the accuracy of $\hat{\mu}$ to predict future data. Suppose y^{new} is a new response vector generated from the same distribution as y. We shall consider to estimate the prediction risk $E(\|y^{new} - \hat{\mu}\|_2^2)/n$.

Define $\mu = X\beta$. Partitioning $(y_i^{new} - \hat{\mu}_i)^2$ as

$$(y_i^{new} - \hat{\mu}_i)^2 = (y_i^{new} - \mu_i)^2 + 2(y_i^{new} - \mu_i)(\mu_i - \hat{\mu}_i) + (\mu_i - \hat{\mu}_i)^2 \tag{2.1}$$

and substituting

$$(\mu_i - \hat{\mu}_i)^2 = (y_i - \hat{\mu}_i)^2 - (y_i - \mu_i)^2 + 2(y_i - \mu_i)(\hat{\mu}_i - \mu_i)$$

into (2.1), we obtain

$$(y_i^{new} - \hat{\mu}_i)^2 = (y_i - \hat{\mu}_i)^2 + 2(y_i - \mu_i)(\hat{\mu}_i - \mu_i) + (y_i^{new} - \mu_i)^2 - (y_i - \mu_i)^2 + 2(y_i^{new} - \mu_i)(\mu_i - \hat{\mu}_i). \tag{2.2}$$

Taking expectation of both sides of the equation (2.2), we obtain the decomposition

$$E(\|y^{new} - \hat{\mu}\|_2^2) = E(\|y - \hat{\mu}\|_2^2) + 2df(\hat{\mu})\sigma^2,$$

where

$$df(\hat{\mu}) = \sum_{i=1}^{n} \mathrm{cov}(\hat{\mu}_i, y_i)/\sigma^2 \tag{2.3}$$

is called the *degrees of freedom* of the fit $\hat{\mu}$.

When $\hat{\mu}$ is given by a linear function of y, i.e., $\hat{\mu} = Sy$ with some matrix S being independent of y, the degrees of freedom is $df(\hat{\mu}) = \mathrm{tr}\, S$, which is a known constant. However, in general it is necessary to estimate $df(\hat{\mu})$. We employ Stein's lemma to accomplish the task.

Lemma 2.1 (Stein's lemma). *Suppose $\hat{\mu}_i : \mathbb{R}^n \to \mathbb{R}$ is absolutely continuous in i-th coordinate for $i = 1, \ldots, n$. If $E(|\partial\hat{\mu}_i/\partial y_i|) < \infty$ for each i, then*

$$\sum_{i=1}^{n} \mathrm{cov}(\hat{\mu}_i, y_i)/\sigma^2 = E(\mathrm{div}\, \hat{\mu}),$$

where $\mathrm{div}\, \hat{\mu} = \sum_{i=1}^{n} \partial\hat{\mu}_i/\partial y_i$.

Therefore an unbiased estimator of the degrees of freedom is given by

$$\widehat{df}(\hat{\mu}) = \mathrm{div}\, \hat{\mu}, \tag{2.4}$$

and we can define a C_p-type criterion by

$$C_p(\hat{\mu}) = \frac{\|y - \hat{\mu}\|_2^2}{n} + \frac{2\widehat{df}(\hat{\mu})}{n}\sigma^2$$

which is an unbiased estimator of the prediction risk.

Let $\hat{\beta}_K$ be the estimator defined as the solution to the problem (1.6) with a closed convex set K. We verify the absolute continuity of $\hat{\mu}_K$ with $\hat{\mu}_K = X\hat{\beta}_K$.

Lemma 2.2. *For every i, $\hat{\mu}_{K,i}$ is absolutely continuous in each coordinate and $\partial\hat{\mu}_{K,i}/\partial y = (\partial\hat{\mu}_{K,i}/\partial y_1, \ldots, \partial\hat{\mu}_{K,i}/\partial y_n)'$ is essentially bounded.*

Proof. Since $\hat{\beta}_K$ is the projection of $\hat{\beta}^\circ$ onto K, $\hat{\beta}_K$ is Lipschitz continuous in $\hat{\beta}^\circ$ (see Theorem 2.4.2 of Webster [13]). Therefore $\hat{\mu}_K$ is shown to be Lipschitz continuous in y and so is each $\hat{\mu}_{K,i}$. The absolute continuity and the essential boundedness follow directly from the Lipschitz continuity. □

Note that if $\hat{\beta}_K$ is differentiable in $\hat{\beta}^\circ$, the divergence $\operatorname{div} \hat{\mu}_K$ is same as the divergence of $\hat{\beta}_K$ with respect to $\hat{\beta}^\circ$. This can be verified by the chain rule:

$$\begin{aligned}\operatorname{div} \hat{\mu}_K &= \operatorname{tr} X \left(\frac{\partial \hat{\beta}_K}{\partial \hat{\beta}^\circ}\right)\left(\frac{\partial \hat{\beta}^\circ}{\partial y}\right) \\ &= \operatorname{tr} X \left(\frac{\partial \hat{\beta}_K}{\partial \hat{\beta}^\circ}\right)(X'X)^{-1}X' \\ &= \operatorname{tr}\left(\frac{\partial \hat{\beta}_K}{\partial \hat{\beta}^\circ}\right),\end{aligned}$$

where $\partial\hat{\beta}_K/\partial\hat{\beta}^\circ$ is the matrix whose (i,k)-th component is $\partial\hat{\beta}_{K,i}/\partial\hat{\beta}^\circ_k$ and $\partial\hat{\beta}^\circ/\partial y$ is the matrix whose (k,j)-th component is $\partial\hat{\beta}^\circ_k/\partial y_j$. Therefore we only need to calculate the divergence of $\hat{\beta}_K$ with respect to $\hat{\beta}^\circ$ in order to derive an unbiased estimator of the degree of the freedom $df(\hat{\mu}_K)$.

For the normal linear model (1.1), $\hat{\beta}^\circ$ is a complete sufficient statistic for β when σ^2 is known and $(\hat{\beta}^\circ, y'y)$ is a complete sufficient statistic for (β, σ^2) when σ^2 is unknown. In either case, $\widehat{df}(\hat{\mu}_K) = \operatorname{tr}(\partial\hat{\beta}_K/\partial\hat{\beta}^\circ)$ is shown to be the unique uniformly minimum variance unbiased estimator of the degrees of freedom $df(\hat{\mu}_K)$ since $\widehat{df}(\hat{\mu}_K)$ is a function of $\hat{\beta}^\circ$. Thus, in terms of estimating the degrees of freedom, the analytical estimator $\widehat{df}(\hat{\mu}_K)$ is more efficient than cross-validation and related nonparametric methods.

3 Main results

In this section, we first derive a divergence formula for the projection onto K under a smoothness condition on the boundary ∂K. As noted in the previous section, it enables us to obtain an unbiased estimator of the degrees of freedom for the shrinkage estimator projected on K. The result presented here is an extension of that of Meyer and Woodroofe [5], which treats the case where K is a convex polyhedral cone.

3.1 Divergence formula

Let $K \subset \mathbb{R}^p$ be a closed convex set. For $x \in \mathbb{R}^p$, x_K denotes the orthogonal projection of x onto K in terms of $\langle\cdot,\cdot\rangle$:

$$\|x - x_K\| = \min_{z\in K} \|x - z\|.$$

Recall that the inner product $\langle\cdot,\cdot\rangle$ is defined by (1.4).

Since K is closed and convex, x_K is uniquely defined. Our main aim is to evaluate the divergence of the projection onto K defined as $f(x) = (f_1(x), \ldots, f_p(x))' = x_K$. Note that f is Lipschitz continuous (see the proof of Lemma 2.2).

Let ∂K be boundary of K. For $s \in \partial K$, the normal cone of K at s is defined by

$$N(K, s) = \{z - s \mid z_K = s\}.$$

Depending on the dimension of the normal cone $N(K, s)$, we have a disjoint partition of the boundary ∂K as

$$\partial K = D_1 \cup \cdots \cup D_p,$$

where

$$D_m = \{s \in \partial K \mid \dim N(K, s) = m\}.$$

Define

$$E_m = \{x \in \mathbb{R}^p \setminus K \mid x_K \in D_m\}.$$

Then we have a disjoint partition of $\mathbb{R}^p \setminus K$ as

$$\mathbb{R}^p \setminus K = E_1 \cup \cdots \cup E_p.$$

We put a condition on smoothness of ∂K as in Kuriki and Takemura [3]. E_m° denotes the interior of E_m.

Assumption 3.1. *D_m is a $(p-m)$-dimensional C^2-manifold consisting of a finite number of relatively open connected components. Furthermore the Lebesgue measure of $E_m \setminus E_m^\circ$ is zero.*

Remark 3.1. In Kuriki and Takemura [3], they call ∂K "piecewise smooth" if ∂K meets Assumption 3.1.

Let $T_s D_m$ be the tangent space of D_m at s and $T_s^\perp D_m$ be the orthogonal complement of $T_s D_m$ in terms of $\langle \cdot, \cdot \rangle$: $T_s^\perp D_m = \{v \in \mathbb{R}^p \mid \langle v, z \rangle = 0,\ \forall z \in T_s D_m\}$. Clearly, $T_s^\perp D_m$ is the affine hull of $N(K, s)$. Following Milnor [6], the normal bundle of D_m is defined as

$$N_m = \{(s, v) \mid s \in D_m, v \in T_s^\perp D_m\}.$$

It is not difficult to show that N_m is a p-dimensional C^1-manifold imbedded in $\mathbb{R}^{2p}$. Let us define $\varphi : N_m \to \mathbb{R}^p$ as $\varphi(s, v) = s + v$. Notice that φ is a C^1-mapping.

Then we show the following basic fact.

Lemma 3.1. *For each fixed $\bar{x} \in E_m^\circ$, there exist an ϵ-ball $B_\epsilon = \{x \in \mathbb{R}^p \mid \|x - \bar{x}\|_2 < \epsilon\} \subset E_m^\circ$ around $\bar{x}$ with sufficiently small $\epsilon > 0$ and an open neighborhood W of $(\bar{x}_K, \bar{x} - \bar{x}_K)$ in N_m such that $\varphi|_W : W \to B_\epsilon$ is a diffeomorphism and $(\varphi|_W)^{-1}(x) = (x_K, x - x_K)$ for $x \in B_\epsilon$. Especially, f is continuously differentiable on E_m°.*

Proof. See Appendix A.1. □

To calculate the divergence of f in an explicit form, we introduce the "tubal coordinates" on E_m°. Let $\theta = (\theta^1, \ldots, \theta^{p-m})$ be a C^2-local coordinate system on D_m and write $s \in D_m$ as $s(\theta) = s(\theta^1, \ldots, \theta^{p-m})$. The tangent space $T_{s(\theta)} D_m$ at $s(\theta)$ is spanned by

$$\left\{ b_a(\theta) = \frac{\partial s}{\partial \theta^a}(\theta), a = 1, \ldots, p - m \right\}.$$

Let $\{n_\alpha(\theta), \alpha = 1, \ldots, m\}$ be an orthonormal basis of $T_s^\perp D_m$ in terms of $\langle \cdot, \cdot \rangle$. Since $\{b_a(\theta)\}$ are C^1-mappings in θ, we can choose $\{n_\alpha(\theta)\}$ so as to be of class C^1 as well. Hence we know that

$$(\theta, \tau) \mapsto (s(\theta), \sum_{\alpha=1}^{m} \tau^\alpha n_\alpha(\theta)),$$

with $\tau = (\tau^1, \ldots, \tau^m) \in \mathbb{R}^m$, gives a C^1-local parametrization of N_m.

From Lemma 3.1, taking

$$(\theta, \tau) \mapsto \varphi(\theta, \tau) = (s(\theta) + \sum_{\alpha=1}^{m} \tau^\alpha n_\alpha(\theta)) \tag{3.1}$$

as a C^1-local parametrization of E_m°, we can express f in the local coordinates (θ, τ) as $f(\theta, \tau) = s(\theta)$. Thus the Jacobian matrix of f with respect to x at $x = \varphi(\theta, \tau)$ is given by

$$\big[b_1(\theta) \cdots b_{p-m}(\theta) \ \underbrace{0 \cdots 0}_{m}\big] (J\varphi)^{-1}_{(\theta,\tau)}, \tag{3.2}$$

where $(J\varphi)_{(\theta,\tau)}$ is the Jacobian matrix of φ with respect to (θ, τ). Especially, the divergence of f with respect to x at $x = \varphi(\theta, \tau)$ is given by the trace of the Jacobian matrix (3.2).

To state our main result, we prepare some notations used in differential geometry: the "first fundamental form" and the "second fundamental form". The first fundamental form of D_m associated with the coordinate system $\theta = (\theta^1, \ldots, \theta^{p-m})$ is the symmetric matrix

$$G(\theta) = (g_{ab}(\theta))_{1 \le a,b \le p-m}$$

with

$$g_{ab}(\theta) = \langle b_a(\theta), b_b(\theta) \rangle.$$

The second fundamental form of D_m in the normal direction $n_\alpha(\theta)$ is defined as

$$H_\alpha(\theta) = (h_{ab\alpha}(\theta))_{1 \le a,b \le p-m}$$

with

$$h_{ab\alpha}(\theta) = \langle n_\alpha(\theta), \frac{\partial^2 s}{\partial \theta^a \partial \theta^b}(\theta) \rangle.$$

For $x = \varphi(\theta, \tau)$, we define

$$\begin{aligned} H(\theta, \tau) &= -\left(\langle x - x_K, \frac{\partial^2 s}{\partial \theta^a \partial \theta^b}(\theta) \rangle \right)_{1 \le a,b \le p-m} \\ &= -\sum_{\alpha=1}^{m} \tau^\alpha H_\alpha(\theta), \end{aligned} \tag{3.3}$$

which is a positive semi-definite matrix. See Appendix A.2.

Lemma 3.2. *The divergence* $\operatorname{div} f(x) = \sum_{j=1}^{p} \partial f_j(x)/\partial x_j$ *of* f *at* $x \in E_m^\circ$ *is given by*

$$\operatorname{div} f(x) = \sum_{a=1}^{p-m} \frac{1}{1+\kappa_a(x)},$$

where $\kappa_a(x) = \kappa_a(\theta, \tau)$, $a = 1, \dots p-m$ *are the eigenvalues satisfying the equation*

$$|H(\theta, \tau) - \kappa G(\theta)| = 0. \tag{3.4}$$

Proof. We need to evaluate the Jacobian matrix (3.2). In the following calculation, we abbreviate arguments like $b_a = b_a(\theta)$. Since the elements of the Jacobian matrix $J\varphi = [\partial\varphi/\partial\theta^1 \cdots \partial\varphi/\partial\theta^{p-m}\ \partial\varphi/\partial\tau^1 \cdots \partial\varphi/\partial\tau^m]$ is given by

$$\frac{\partial\varphi}{\partial\theta^a} = b_a + \sum_{\alpha=1}^{m} \tau^\alpha \frac{\partial n_\alpha}{\partial\theta^a},$$
$$\frac{\partial\varphi}{\partial\tau^\beta} = n_\beta,$$

we have

$$\begin{aligned}
&(J\varphi)'V\,[b_1 \cdots b_{p-m}\ n_1 \cdots n_m] \\
&= \begin{bmatrix} \left(g_{ab} + \sum_{\alpha=1}^{m} \tau^\alpha \langle \frac{\partial n_\alpha}{\partial\theta^a}, b_b\rangle\right)_{1\le a,b\le p-m} & \left(\sum_{\alpha=1}^{m} \tau^\alpha \langle \frac{\partial n_\alpha}{\partial\theta^a}, n_\beta\rangle\right)_{1\le a\le p-m, 1\le\beta\le m} \\ 0 & I_m \end{bmatrix}.
\end{aligned} \tag{3.5}$$

Differentiating both sides of $\langle n_\alpha, b_b\rangle = 0$ with respect to θ^a, we obtain

$$0 = \frac{\partial}{\partial\theta^a}\langle n_\alpha, b_b\rangle = \langle \frac{\partial n_\alpha}{\partial\theta^a}, b_b\rangle + \langle n_\alpha, \frac{\partial^2 s}{\partial\theta^a\partial\theta^b}\rangle,$$

and hence

$$\langle \frac{\partial n_\alpha}{\partial\theta^a}, b_b\rangle = -\langle n_\alpha, \frac{\partial^2 s}{\partial\theta^a\partial\theta^b}\rangle.$$

Thus the right hand side of (3.5) is written as

$$\begin{aligned}
&\begin{bmatrix} \left(g_{ab} - \sum_{\alpha=1}^{m} \tau^\alpha \langle n_\alpha, \frac{\partial^2 s}{\partial\theta^a\partial\theta^b}\rangle\right)_{1\le a,b\le p-m} & \left(\sum_{\alpha=1}^{m} \tau^\alpha \langle \frac{\partial n_\alpha}{\partial\theta^a}, n_\beta\rangle\right)_{1\le a\le p-m, 1\le\beta\le m} \\ 0 & I_m \end{bmatrix} \\
&= \begin{bmatrix} A_{11} & A_{12} \\ 0 & I_m \end{bmatrix},
\end{aligned}$$

where $(p-m)\times(p-m)$ matrix A_{11} and $(p-m)\times m$ matrix A_{12} are given by

$$\begin{aligned}
A_{11} &= \left(g_{ab} - \sum_{\alpha=1}^{m} \tau^\alpha \langle n_\alpha, \frac{\partial^2 s}{\partial\theta^a\partial\theta^b}\rangle\right)_{1\le a,b\le p-m} \\
&= G(\theta) + H(\theta, \tau), \\
A_{12} &= \left(\sum_{\alpha=1}^{m} \tau^\alpha \langle \frac{\partial n_\alpha}{\partial\theta^a}, n_\beta\rangle\right)_{1\le a\le p-m, 1\le\beta\le m}.
\end{aligned}$$

Therefore we obtain

$$(J\varphi)^{-1} = \begin{bmatrix} A_{11} & 0 \\ A_{12}' & I_m \end{bmatrix}^{-1} [b_1 \cdots b_{p-m}\ n_1 \cdots n_m]' V.$$

The Jacobian matrix (3.2) is given by

$$\begin{aligned} [B \quad 0] \begin{bmatrix} A_{11} & 0 \\ A_{12}' & I_m \end{bmatrix}^{-1} \begin{bmatrix} B' \\ N' \end{bmatrix} V &= [B \quad 0] \begin{bmatrix} A_{11}^{-1} & 0 \\ A_{12}' A_{11}^{-1} & I_m \end{bmatrix} \begin{bmatrix} B' \\ N' \end{bmatrix} V \\ &= BA_{11}^{-1}B'V \\ &= B(G+H)^{-1}B'V \\ &= B(B'VB+H)^{-1}B'V, \end{aligned} \tag{3.6}$$

where $G = G(\theta)$, $H = H(\theta, \tau)$, $B = [b_1 \cdots b_{p-m}]$, $N = [n_1 \cdots n_m]$. Let $\kappa_1(\theta, \tau), \ldots, \kappa_{p-m}(\theta, \tau)$ be the eigenvalues of $H(\theta, \tau)$ with respect to $G(\theta)$, i.e., solutions of the equation (3.4).

Then, the divergence is written as

$$\begin{aligned} \operatorname{tr} B(G+H)^{-1}B'V &= \operatorname{tr}(G+H)^{-1}G \\ &= \sum_{a=1}^{p-m} \frac{1}{1+\kappa_a}. \end{aligned}$$

Therefore the proof is completed. □

Remark 3.2. The local coordinates (θ, τ) given in (3.1) is called the "tubal coordinates", which is used in Weyl [14] to derive formulas for the volume of tubes.

Remark 3.3. When K is a convex polyhedron, it holds that $B(\theta) \equiv B$ (constant matrix) and $H(\theta, \tau) \equiv 0$. In this case, the Jacobian matrix (3.6) reduces to the constant projection matrix.

Remark 3.4. In Kuriki and Takemura [3], the "average codimension" $d(x)$ is defined as

$$\begin{aligned} d(x) &= m + \operatorname{tr}(I_{p-m} + HG^{-1})^{-1} HG^{-1} \\ &= m + \sum_{a=1}^{p-m} \frac{\kappa_a}{1+\kappa_a} \\ &= p - \sum_{a=1}^{p-m} \frac{1}{1+\kappa_a} \end{aligned}$$

for $x \in E_m$. Hence we have the relation $\operatorname{div} f(x) = p - d(x), a.e.$

3.2 Degrees of freedom

Using Lemma 3.2, we can derive an unbiased estimator of the degrees of freedom $df(\hat{\mu}_K)$. We assume that K is a closed convex set satisfying Assumption 3.1.

For $\hat{\beta}^\circ \in E_m$, identifying $x = \hat{\beta}^\circ$ and $x_K = \hat{\beta}_K$, let $\kappa_{m,1}(\hat{\beta}^\circ), \ldots, \kappa_{m,p-m}(\hat{\beta}^\circ)$ be the eigenvalues satisfying (3.4). Formally we define $E_0 = K$ and $\kappa_{0,a}(\hat{\beta}^\circ) \equiv 0$, $a = 1, \ldots, p$. Then, we obtain the following theorem. Note that $\hat{\beta}^\circ \in E_m$ is equivalent to $\hat{\beta}^\circ \notin K$ and $\hat{\beta}_K \in D_m$ for $m \geq 1$.

Theorem 3.1. *Suppose K is a closed convex set satisfying Assumption 3.1. Then,*

$$\widehat{df}(\hat{\mu}_K) = \sum_{m=0}^{p} \sum_{a=1}^{p-m} \frac{1}{1+\kappa_{m,a}(\hat{\beta}^\circ)} I(\hat{\beta}^\circ \in E_m) \tag{3.7}$$

gives an unbiased estimator of the degrees of freedom $df(\hat{\mu}_K)$. Here, $I(\cdot)$ is an indicator function.

Hence, a C_p-type criterion for $\hat{\mu}_K$ is given by

$$C_p(\hat{\mu}_K) = \frac{\|y-\hat{\mu}_K\|_2^2}{n} + \frac{2\widehat{df}(\hat{\mu}_K)}{n}\sigma^2,$$

which is an unbiased estimator of the prediction risk $E[\|y^{new}-\hat{\mu}_K\|_2^2]/n$. Equivalently, we can define AIC for $\hat{\mu}_K$ as

$$\mathrm{AIC}(\hat{\mu}_K) = \frac{\|y-\hat{\mu}_K\|_2^2}{n\sigma^2} + \frac{2\widehat{df}(\hat{\mu}_K)}{n}.$$

When σ^2 is unknown, it is replaced by an unbiased estimate.

In our setting (1.6), K plays a role of a tuning parameter. Practically, we choose the optimal K which minimizes $C_p(\hat{\mu}_K)$ or $\mathrm{AIC}(\hat{\mu}_K)$ among a given collection $\mathcal{K}$ of closed convex sets satisfying Assumption 3.1. For instance, $\mathcal{K} = \{\{\beta \in \mathbb{R}^p \mid \sum_{j=1}^p |\beta_j| \le t\} \mid t > 0\}$ in the Lasso case.

The usefulness of Theorem 3.1 is that it does not required to know the functional form of $\hat{\beta}_K$ in calculation of (3.7). Once we know the numerical values of $\hat{\beta}^\circ$ and $\hat{\beta}_K$, we can calculate the value of (3.7) through the geometric quantities such as the first fundamental form and the second fundamental form. Especially, if K is a convex polyhedron, all $\kappa_{m,a}$'s turn out to be zero. Therefore, (3.7) is simply expressed as

$$\widehat{df}(\hat{\mu}_K) = p - \sum_{m=1}^{p} m I(\hat{\beta}^\circ \in E_m), \tag{3.8}$$

which coincides with the dimension of the face which contains $\hat{\beta}_K$ as a relatively interior point when $\hat{\beta}^\circ \notin K$.

4 Examples

In this section, we provide unbiased estimators of the degrees of freedom for the Lasso, the fused Lasso, and the group Lasso. Our result is also applicable to order restricted inference. The degrees of freedom in order restricted inference is studied in Meyer and Woodroofe [5] in the case where K is a convex polyhedral cone.

4.1 Lasso

For the Lasso, the constraint region is given by

$$K = \{\beta \in \mathbb{R}^p \mid \sum_{j=1}^{p} |\beta_j| \leq t\}.$$

We denote the Lasso estimator by $\hat{\beta}(t)$ rather than $\hat{\beta}_K$. Since K is a convex polyhedron, an unbiased estimator $\widehat{df}(t)$ of the degrees of freedom of $\hat{\mu}(t) = X\hat{\beta}(t)$ is given by (3.8). In this case, if $\hat{\beta}(t) \in D_m$, then the number of zeros in $\hat{\beta}(t)$ is equal to $m - 1$. Therefore we obtain the expression

$$\widehat{df}(t) = \begin{cases} \#\{j \mid \hat{\beta}(t)_j \neq 0\} - 1 & \text{if } \sum_{j=1}^{p} |\hat{\beta}_j^{\circ}| > t, \\ p & \text{if } \sum_{j=1}^{p} |\hat{\beta}_j^{\circ}| \leq t. \end{cases}$$

A similar result is presented in Zou et al. [15], although their parametrization is not same as ours.

4.2 Fused Lasso

The fused Lasso (Tibshirani et al. [11]) is the shrinkage method with the constraint region

$$K = \{\beta \in \mathbb{R}^p \mid \sum_{j=1}^{p} |\beta_j| \leq t_1, \ \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq t_2\}.$$

We assume $t_1 \neq t_2$. Let $\hat{\beta}(t)$ be the fused Lasso estimator with $t = (t_1, t_2)$. Since K is a convex polyhedron, an unbiased estimator $\widehat{df}(t)$ of the degrees of freedom of $\hat{\mu}(t) = X\hat{\beta}(t)$ is given by (3.8).

Define

$$K_1 = \{\beta \in \mathbb{R}^p \mid \sum_{j=1}^{p} |\beta_j| \leq t_1\}$$

and

$$K_2 = \{\beta \in \mathbb{R}^p \mid \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq t_2\}.$$

Corresponding to the 2^p different possible signs for p components of β, K_1 is expressed as the solution set of 2^p linear inequalities:

$$K_1 = \{\beta \in \mathbb{R}^p \mid a_i'\beta \leq t_1, i = 1, \ldots, 2^p\}.$$

Similarly, K_2 is expressed as the solution set of 2^{p-1} linear inequalities:

$$K_2 = \{\beta \in \mathbb{R}^p \mid b_i'\beta \leq t_2, i = 1, \ldots, 2^{p-1}\}.$$

For instance, if $p = 3$, $a_1 = (1, 1, 1)'$, $a_2 = (-1, 1, 1)'$, $a_3 = (1, -1, 1)'$, $a_4 = (1, 1, -1)'$, $a_5 = (-1, -1, 1)'$, $a_6 = (-1, 1, -1)'$, $a_7 = (1, -1, -1)'$, $a_8 = (-1, -1, -1)'$ and $b_1 = (-1, 0, 1)'$, $b_2 = (1, -2, 1)'$, $b_3 = (-1, 2, -1)'$, $b_4 = (1, 0, -1)'$.

Each open face of the polytope $K = K_1 \cap K_2$ is of the form

$$\{\beta \in \mathbb{R}^p \mid a_i'\beta = t_1, i \in \mathcal{I}_1, b_i'\beta = t_2, i \in \mathcal{I}_2, \\ a_j'\beta < t_1, j \in \{1, \ldots, 2^p\} \backslash \mathcal{I}_1, b_j'\beta < t_2, j \in \{1, \ldots, 2^{p-1}\} \backslash \mathcal{I}_2\}, \tag{4.1}$$

where $\mathcal{I}_1 \subset \{1, \ldots, 2^p\}$ and $\mathcal{I}_2 \subset \{1, \ldots, 2^{p-1}\}$. Suppose a nonempty open face F of K is given by (4.1) where the matrix whose column vectors are a_i, $i \in \mathcal{I}_1$ and b_i, $i \in \mathcal{I}_2$ is of rank m. Then the dimension of F is $p - m$.

From these observations, we know that the unbiased estimator $\widehat{df}(t)$ of $df(\hat{\mu}(t))$ is given by

$$\widehat{df}(t) = \begin{cases} p - m_1(t) & \text{if } \hat{\beta}(t) \in \partial K_1 \cap K_2^\circ \text{ and } \hat{\beta}^\circ \notin K, \\ p - m_2(t) & \text{if } \hat{\beta}(t) \in K_1^\circ \cap \partial K_2 \text{ and } \hat{\beta}^\circ \notin K, \\ p - m_3(t) & \text{if } \hat{\beta}(t) \in \partial K_1 \cap \partial K_2 \text{ and } \hat{\beta}^\circ \notin K, \\ p & \text{if } \hat{\beta}^\circ \in K, \end{cases}$$

where

$$m_1(t) = \#\{j \mid \hat{\beta}(t)_j = 0\} + 1,$$
$$m_2(t) = \#\{j \geq 2 \mid \hat{\beta}(t)_j - \hat{\beta}(t)_{j-1} = 0\} + 1,$$
$$m_3(t) = \#\{j \mid \hat{\beta}(t)_j = 0\} + \#\{j \geq 2 \mid \hat{\beta}(t)_j - \hat{\beta}(t)_{j-1} = 0, \hat{\beta}(t)_{j-1}, \hat{\beta}(t)_j \neq 0\} + 2.$$

Remark 4.1. In Tibshirani et al. [11], with the penalization formulation, they propose

$$p - \#\{j \mid \hat{\beta}_j = 0\} - \#\{j \geq 2 \mid \hat{\beta}_j - \hat{\beta}_{j-1} = 0, \hat{\beta}_j, \hat{\beta}_{j-1} \neq 0\}$$

as an estimator of the degrees of freedom for the fused Lasso, where $\hat{\beta}$ is the fused Lasso estimator. They, however, do not present a mathematical proof for unbiasedness of this estimator.

4.3 Group Lasso

The group Lasso is proposed in Yuan and Lin [12]. The constraint region of the group Lasso is

$$K = \{\beta \in \mathbb{R}^p \mid \sum_{j=1}^{J} (\beta_{[j]}' V_j \beta_{[j]})^{1/2} \leq t\},$$

where β is partitioned as $\beta = (\beta_{[1]}', \ldots, \beta_{[J]}')'$ with $\beta_{[j]}$ being a $p_j \times 1$ vector, and V_j is a $p_j \times p_j$ symmetric positive definite matrix. In the subsequent calculation, we assume that X is orthonormal, i.e., $X'X = I_p$ and hence $V = I_p$.

For $x \in \mathbb{R}^p$, let $x_{[1]} = (x_1, \ldots, x_q)'$. We first treat the case

$$K = \{x \in \mathbb{R}^p \mid \|x_{[1]}\|_2 + |x_{q+1}| + \cdots + |x_p| \leq t\}, \tag{4.2}$$

where $\|x_{[1]}\|_2 = (\sum_{j=1}^{q} x_j^2)^{\frac{1}{2}}$. We focus on the following surface area:

$$\begin{aligned} M =& \{x \in \mathbb{R}^p \mid \|x_{[1]}\|_2 + x_{q+1} + \cdots + x_{q+r} = t, \\ & \|x_{[1]}\|_2 > 0, x_{q+1} > 0, \ldots, x_{q+r} > 0, x_{q+r+1} = \cdots = x_p = 0\}. \end{aligned}$$

The set M is a $(q+r-1)$-dimensional smooth manifold. To introduce a local coordinate system on M, we transform $x_{[1]}$ into polar coordinates (Takemura [9]) as

$$x_{[1]} = \theta_q u(\theta_1, \ldots, \theta_{q-1}),$$

with

$$u(\theta_1, \ldots, \theta_{q-1}) = \begin{pmatrix} \cos\theta_1 \\ \sin\theta_1 \cos\theta_2 \\ \vdots \\ \sin\theta_1 \sin\theta_2 \cdots \cos\theta_{q-1} \\ \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{q-1} \end{pmatrix},$$

where $0 \le \theta_i \le \pi$,$i = 1, \ldots, q-2$, $0 \le \theta_{q-1} < 2\pi$, and $0 < \theta_q < t$. Then the rest of the variables $x_{q+1}, \ldots, x_{q+r}$ must satisfy

$$x_{q+1} + \cdots + x_{q+r} = t - \theta_q.$$

Let $e_i \in \mathbb{R}^p$ be the vector of which only i-th component is 1 and all other components are zero. Take $b_{q+j} = e_{q+1+j} - e_{q+1}$, $j = 1, \ldots, r-1$. Then $x \in M$ is expressed as

$$\begin{aligned} x = x(\theta_1, \ldots, \theta_{q+r-1}) &= \begin{pmatrix} x_{[1]} \\ x_{q+1} \\ \vdots \\ x_{q+r} \\ 0_{p-q-r} \end{pmatrix} \\ &= \theta_q \begin{pmatrix} u(\theta_1, \ldots, \theta_{q-1}) \\ 0_{p-q} \end{pmatrix} + (t - \theta_q)(e_{q+1} + \theta_{q+1} b_{q+1} + \cdots + \theta_{q+r-1} b_{q+r-1}), \end{aligned}$$

where 0_i is $i \times 1$ zero vector, and $\theta_{q+1}, \ldots, \theta_{q+r-1}$ satisfy $\theta_{q+j} > 0$, $j = 1, \ldots, r-1$ and $\sum_{j=1}^{r-1} \theta_{q+j} < 1$.

The partial derivative of $u(\theta_1, \ldots, \theta_{q-1})$ with respect to θ_1 is given by

$$\begin{aligned} \frac{\partial u}{\partial \theta_1}(\theta_1, \ldots, \theta_{q-1}) &= \begin{pmatrix} -\sin\theta_1 \\ \cos\theta_1 \sin\theta_2 \\ \vdots \\ \cos\theta_1 \sin\theta_2 \cdots \sin\theta_{q-2} \cos\theta_{q-1} \\ \cos\theta_1 \sin\theta_2 \cdots \sin\theta_{q-2} \sin\theta_{q-1} \end{pmatrix} \\ &\equiv v(\theta_1, \ldots, \theta_{q-1}). \end{aligned}$$

Define $v(\theta_i, \ldots, \theta_{q-1})$ for $i \geq 2$ in the similar manner. Then, we have

$$\frac{\partial u}{\partial \theta_i}(\theta_1, \ldots, \theta_{q-1}) = \sin\theta_1 \cdots \sin\theta_{i-1} \begin{pmatrix} 0_{i-1} \\ v(\theta_i, \ldots, \theta_{q-1}) \end{pmatrix}.$$

Thus the tangent space T_xM at x is spanned by the following $(q+r-1)$ linearly independent vectors:

$$\begin{aligned}
\frac{\partial x}{\partial \theta_i} &= \theta_q \sin\theta_1 \cdots \sin\theta_{i-1} \begin{pmatrix} 0_{i-1} \\ v(\theta_i, \ldots, \theta_{q-1}) \\ 0_{p-q} \end{pmatrix}, \quad i = 1, \ldots, q-1, \\
\frac{\partial x}{\partial \theta_q} &= \begin{pmatrix} u(\theta_1, \ldots, \theta_{q-1}) \\ 0_{p-q} \end{pmatrix} - (e_{q+1} + \theta_{q+1} b_{q+1} + \cdots + \theta_{q+r-1} b_{q+r-1}), \\
\frac{\partial x}{\partial \theta_{q+j}} &= (t - \theta_q) b_{q+j}, \quad j = 1, \ldots, r-1.
\end{aligned}$$

It is easy to see that the orthonormal system

$$\{n_1, \ldots, n_{p-q-r+1}\},$$

with

$$n_1 = \frac{1}{\sqrt{r+1}} \begin{pmatrix} u(\theta_1, \ldots, \theta_{q-1}) \\ 1_r \\ 0_{p-q-r} \end{pmatrix}$$

and $n_2 = e_{q+r+1}, \ldots, n_{p-q-r+1} = e_p$, gives a basis of $T_x^{\perp}M$. Here, $1_r = (\underbrace{1, \ldots, 1}_{r})'$. To calculate the second fundamental forms, we evaluate the second partial derivatives of x, which are summarized as follows:

$$\begin{aligned}
\frac{\partial^2 x}{\partial \theta_i^2} &= -\theta_q \sin\theta_1 \cdots \sin\theta_{i-1} \begin{pmatrix} 0_{i-1} \\ u(\theta_i, \ldots, \theta_{q-1}) \\ 0_{p-q} \end{pmatrix}, \quad i = 1, \ldots, q-1, \\
\frac{\partial^2 x}{\partial \theta_i \partial \theta_j} &= \theta_q \sin\theta_1 \cdots \cos\theta_i \cdots \sin\theta_{j-1} \begin{pmatrix} 0_{j-1} \\ v(\theta_j, \ldots, \theta_{q-1}) \\ 0_{p-q} \end{pmatrix}, \quad 1 \leq i < j \leq q-1, \\
\frac{\partial^2 x}{\partial \theta_i \partial \theta_q} &= \sin\theta_1 \cdots \sin\theta_{i-1} \begin{pmatrix} 0_{i-1} \\ v(\theta_i, \ldots, \theta_{q-1}) \\ 0_{p-q} \end{pmatrix}, \quad i = 1, \ldots, q-1, \\
\frac{\partial^2 x}{\partial \theta_q \partial \theta_{q+j}} &= -b_{q+j}, \quad j = 1, \ldots, r-1, \\
\frac{\partial^2 x}{\partial \theta_i \partial \theta_{q+j}} &= \frac{\partial^2 x}{\partial \theta_q^2} = \frac{\partial^2 x}{\partial \theta_{q+j}^2} = 0, \quad i = 1, \ldots, q-1, \ j = 1, \ldots, r-1.
\end{aligned}$$

Here, $u(\theta_i, \ldots, \theta_{q-1})$, $i \geq 2$ are defined in the similar way as $u(\theta_1, \ldots, \theta_{q-1})$.

Therefore the second fundamental forms are calculated as follows:

$$\begin{aligned} H_1 &= \left(\langle n_1, \frac{\partial^2 x}{\partial\theta_i \partial\theta_j} \rangle \right)_{1\le i,j\le q+r-1} \\ &= -\frac{1}{\sqrt{r+1}} \left(\begin{array}{ccc|c} h_1 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & h_{q-1} & 0 \\ \hline 0 & \dots & 0 & 0 \end{array} \right), \end{aligned}$$

with $h_i = \theta_q \sin^2\theta_1 \cdots \sin^2\theta_{i-1}$, $i = 1, \dots, q-1$, and

$$\begin{aligned} H_k &= \left(\langle n_k, \frac{\partial^2 x}{\partial\theta_i \partial\theta_j} \rangle \right)_{1\le i,j\le q+r-1} \\ &= 0, \end{aligned}$$

for $k = 2, \dots, p-q-r+1$.

Since the first fundamental form is given in the form

$$G = \left(\begin{array}{ccc|c} \theta_q h_1 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \theta_q h_{q-1} & 0 \\ \hline 0 & \dots & 0 & G_{22} \end{array} \right),$$

the eigenvalues of $H = -\sum_{k=1}^{p-q-r+1} \tau_k H_k = -\tau_1 H_1$ with respect to G are given by

$$\kappa_1 = \cdots = \kappa_{q-1} = \frac{\tau_1'}{\theta_q}, \ \kappa_q = \cdots = \kappa_{q+r-1} = 0,$$

where $\tau_1' = \tau_1/\sqrt{r+1}$.

Returning to the original problem, let $\hat\beta(t)$ be the resulting estimator with the constraint region (4.2). When $\hat\beta(t) \in M$, θ_q and τ_1' correspond to $\theta_q = \|\hat\beta(t)_{[1]}\|_2$ and $\tau_1' = \|\hat\beta^\circ_{[1]} - \hat\beta(t)_{[1]}\|_2$. Since $\hat\beta(t)'_{[1]}(\hat\beta^\circ_{[1]} - \hat\beta(t)_{[1]}) = (\theta_q u(\theta_1, \dots, \theta_{q-1}))'(\tau_1' u(\theta_1, \dots, \theta_{q-1})) = \theta_q \tau_1' = \|\hat\beta(t)_{[1]}\|_2 \|\hat\beta^\circ_{[1]} - \hat\beta(t)_{[1]}\|_2$, we have

$$\|\hat\beta(t)_{[1]}\|_2 + \|\hat\beta^\circ_{[1]} - \hat\beta(t)_{[1]}\|_2 = \|\hat\beta^\circ_{[1]}\|_2.$$

Thus an unbiased estimator $\widehat{df}(t)$ of $df(\hat\mu(t))$, where $\hat\mu(t) = X\hat\beta(t)$, is given by

$$\begin{aligned} \widehat{df}(t) &= r + (q-1)\frac{1}{1 + \|\hat\beta^\circ_{[1]} - \hat\beta(t)_{[1]}\|_2 / \|\hat\beta(t)_{[1]}\|_2} \\ &= r + (q-1)\frac{\|\hat\beta(t)_{[1]}\|_2}{\|\hat\beta^\circ_{[1]}\|_2}, \end{aligned}$$

when $\hat\beta(t) \in M$ and $\hat\beta^\circ \notin K$. A similar calculation shows that entire $\widehat{df}(t)$ is given by

$$\widehat{df}(t) = \begin{cases} I(\|\hat\beta(t)_{[1]}\|_2 > 0) \left(1 + (q-1)\frac{\|\hat\beta(t)_{[1]}\|_2}{\|\hat\beta^\circ_{[1]}\|_2} \right) + \sum_{j=1}^{p-q} I(|\hat\beta(t)_{q+j}| > 0) - 1 & \text{if } \hat\beta^\circ \notin K, \\ p & \text{if } \hat\beta^\circ \in K. \end{cases}$$

Since $\|\hat{\beta}^{\circ}_{[1]}\|_2 > 0$ with probability 1, we also have $\widetilde{df}(t)$ as an unbiased estimator of $df(\hat{\mu}(t))$, where

$$\widetilde{df}(t) = \begin{cases} I(\|\hat{\beta}(t)_{[1]}\|_2 > 0) + \sum_{j=1}^{p-q} I(|\hat{\beta}(t)_{q+j}| > 0) + (q-1)\frac{\|\hat{\beta}(t)_{[1]}\|_2}{\|\hat{\beta}^{\circ}_{[1]}\|_2} - 1 & \text{if } \hat{\beta}^{\circ} \notin K, \\ p & \text{if } \hat{\beta}^{\circ} \in K. \end{cases}$$

Next, for $x \in \mathbb{R}^p$, we write $x = (x'_{[1]}, \ldots, x'_{[J]})'$ as a partition of x where $x_{[j]}$ is a $p_j \times 1$ vector. Define $\|x\|_{[j]} = (x'_{[j]}x_{[j]})^{\frac{1}{2}}$ for $x \in \mathbb{R}^p$. We consider the group Lasso estimation with the constraint region

$$K = \{\beta \in \mathbb{R}^p \mid \sum_{j=1}^{J} \|\beta\|_{[j]} \leq t\}.$$

Let $\hat{\beta}(t)$ be the resulting estimator. Assuming that X is orthonormal, an unbiased estimator of the degrees of freedom $df(\hat{\mu}(t))$ with $\hat{\mu}(t) = X\hat{\beta}(t)$ is given by

$$\widetilde{df}(t) = \begin{cases} \sum_{j=1}^{J} I(\|\hat{\beta}\|_{[j]} > 0) + \sum_{j=1}^{J}(p_j - 1)\frac{\|\hat{\beta}(t)\|_{[j]}}{\|\hat{\beta}^{\circ}\|_{[j]}} - 1 & \text{if } \hat{\beta}^{\circ} \notin K, \\ p & \text{if } \hat{\beta}^{\circ} \in K. \end{cases}$$

The proof is similar as above and thus omitted.

Remark 4.2. Even when X is not orthonormal, we can calculate the estimator (3.7) of the degrees of freedom numerically.

5 Concluding remarks

In this paper, we have derived an unbiased estimator of the degrees of freedom for the shrinkage estimator towards a closed convex set with piecewise smooth boundary. Setting the estimation problem to (1.6), we can treat selection criteria for the tuning parameter in recently proposed estimation methods such as the Lasso, the fused Lasso, and the group Lasso in unified sense.

It seems to be necessary to study some optimal properties of C_p or AIC in selecting the tuning parameter. For the traditional variable selection problem in linear model, there are lots of literature on properties of model selection criteria (for example, Shao [7]). This topic remains in the future research.

Acknowledgment

The author would like to thank Professor Tatsuya Kubokawa for his encouragement and helpful suggestions.

A Appendix

A.1 Proof of Lemma 3.1

Let $\bar{x} \in E_m^\circ$ be an arbitrary fixed vector. From Remark A.1 below, $(\bar{x}_K, \bar{x} - \bar{x}_K)$ is a regular point of φ. Thus the inverse function theorem implies that there exists an open neighborhood $U \cap N_m$ of $(\bar{x}_K, \bar{x} - \bar{x}_K)$ in N_m such that $\varphi|_{U \cap N_m} : U \cap N_m \to \varphi(U \cap N_m)$ is a diffeomorphism. Here U is an open set in $\mathbb{R}^{2p}$ containing $(\bar{x}_K, \bar{x} - \bar{x}_K)$. Let $L > 0$ be the Lipschitz constant of f.

Let us define

$$B_\epsilon = \{x \in \mathbb{R}^p \mid \|x - \bar{x}\|_2 < \epsilon\}$$

and

$$\begin{aligned} Q_\epsilon = \{z \in \mathbb{R}^{2p} \mid & |z_i - \bar{x}_i| < L\epsilon, 1 \le i \le p, \\ & |z_{p+j} - (\bar{x}_j - \bar{x}_{K,j})| < (1+L)\epsilon, 1 \le j \le p\} \end{aligned}$$

with $\epsilon > 0$ small enough to have

$$B_\epsilon \subset \varphi(U \cap N_m), \ Q_\epsilon \subset U.$$

Since f is Lipschitz continuous with Lipschitz constant L, it holds that for $x \in B_\epsilon$,

$$\|x_K - \bar{x}_K\|_2 < L\epsilon,$$

and

$$\begin{aligned} \|(x - x_K) - (\bar{x} - \bar{x}_K)\|_2 &\le \|x - \bar{x}\|_2 + \|x_K - \bar{x}_K\|_2 \\ &< (1+L)\epsilon. \end{aligned}$$

Therefore we have $(x_K, x - x_K) \in Q_\epsilon \cap N_m \subset U \cap N_m$ for $x \in B_\epsilon$. Define

$$W = (\varphi|_{U \cap N_m})^{-1}(B_\epsilon) \subset Q_\epsilon \cap N_m.$$

Note that W is an open set in N_m. Then it is seen that the diffeomorphism $\varphi|_W : W \to B_\epsilon$ corresponds to the mapping $(x_K, x - x_K) \mapsto x$. □

A.2 Positive semi-definiteness of the matrix (3.3)

We follow the notations used in Section 3.1. Let $s_0 \in D_m$ and $v_0 \in N(K, s_0)$ be arbitrary fixed vectors. We take a C^2-local coordinate system $(\theta^1, \ldots, \theta^{p-m})$ of D_m around s_0 such that $s_0 = s(0, \ldots, 0)$. Then we shall show the following fact:

Lemma A.1. *The matrix*

$$-\left(\langle v_0, \frac{\partial^2 s}{\partial\theta^a \partial\theta^b}(0)\rangle\right)_{1 \le a,b \le p-m} \tag{A.1}$$

is positive semi-definite.

Proof. Define

$$L(\theta) = -\langle v_0, s(\theta) - s_0 \rangle.$$

in an appropriate neighborhood of 0. From Theorem 2.4.1 of Webster [13], it follows that $L(\theta) \geq 0$ for all θ in the neighborhood and $L(0) = 0$. Hence $\theta = 0$ is the minimizer of $L(\theta)$. Noting that

$$\frac{\partial^2 L}{\partial \theta^a \partial \theta^b}(0) = -\langle v_0, \frac{\partial^2 s}{\partial \theta^a \partial \theta^b}(0) \rangle,$$

the second order necessary condition for the minimizer ensures that the matrix (A.1) is indeed positive semi-definite. □

Remark A.1. From this lemma and Assertion 6.4 of Milnor [6] (or our calculation in the proof of Lemma 3.2), it can be proved that $(x_K, x - x_K)$ with $x \in E_m$ is a regular point of φ.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum liklihood principle. *Second International Symposium on Information Theory* 267-281.

[2] Efron, B. (2004). The estimation of prediction error: covariance penalties and cross validation. *J. Amer. Statist. Assoc.* **99**, 619-632.

[3] Kuriki, S. and Takemura, A. (2000). Shrinkage estimation towards a closed convex set with a smooth boundary. *J. Multivariate Anal.* **75**, 79-111.

[4] Mallows, C. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

[5] Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28**, 1083-1104.

[6] Milnor, J. (1963). *Morse Theory*. Ann. Math. Stud. **51**, Princeton Univ. Press, Princeton.

[7] Shao, J. (1997). An asymtotic theory for linear model selection (with discussion). *Statist. Sinica* **7**, 221-242.

[8] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135-1151.

[9] Takemura, A. (1991). *Foundation of Multivariate Statistical Inference* (in Japanese). Kyoritsu Shuppan, Tokyo.

[10] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267-288.

[11] Tibshirani, R., Saunders, M., Rosset, S., Zu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 91-108.

[12] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 49-67.

[13] Webster, R. (1994). *Convexity*. Oxford Univ. Press, Oxford.

[14] Weyl, H. (1939). On the volume of tubes. *Amer. J. Math.* **61**, 461-472.

[15] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the Lasso. *Ann. Statist.* to appear.