Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2018, Vol. 54, No. 2, 819–841
https://doi.org/10.1214/17-AIHP822


ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

# Degrees of freedom for piecewise Lipschitz estimators

Frederik Riis Mikkelsen and Niels Richard Hansen

*Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark.*
*E-mail: frm@math.ku.dk; Niels.R.Hansen@math.ku.dk*

Received 10 June 2016; revised 10 February 2017; accepted 22 February 2017

**Abstract.** A representation of the degrees of freedom akin to Stein's lemma is given for a class of estimators of a mean value parameter in $\mathbb{R}^n$. Contrary to previous results our representation holds for a range of discontinues estimators. It shows that even though the discontinuities form a Lebesgue null set, they cannot be ignored when computing degrees of freedom. Estimators with discontinuities arise naturally in regression if data driven variable selection is used. Two such examples, namely best subset selection and lasso-OLS, are considered in detail in this paper. For lasso-OLS the general representation leads to an estimate of the degrees of freedom based on the lasso solution path, which in turn can be used for estimating the risk of lasso-OLS. A similar estimate is proposed for best subset selection. The usefulness of the risk estimates for selecting the number of variables is demonstrated via simulations with a particular focus on lasso-OLS.

**Résumé.** Une représentation des degrés de liberté comparable au lemme de Stein est donnée pour une classe d'estimateurs du paramètre de la moyenne dans $\mathbb{R}^n$. Contrairement aux résultats précédents, notre représentation est valable pour une famille d'estimateurs discontinus. Cela montre que même si les discontinuités sont de mesure de Lebesgue zéro, elles ne peuvent pas être ignorées lors du calcul des degrés de liberté. Les estimateurs avec discontinuités apparaissent naturellement dans les modèles de régression si de la sélection de variables par les données est utilisée. Deux tels exemples, la sélection du meilleur sous-ensemble et le lasso-OLS, sont considérés en détail dans l'article. Pour le Lasso-OLS, la représentation générale mène à une estimation des degrés de liberté basée sur le chemin de la solution Lasso, qui à son tour peut être utilisée pour estimer le risque du lasso-OLS. Une estimation similaire est proposée pour la sélection du meilleur sous-ensemble. L'utilité des estimées de risque pour le choix du nombre de variables est démontrée par des simulations qui se concentrent en particulier sur lasso-OLS.

*MSC:* 62J05; 62J07

*Keywords:* Best subset selection; Lasso-OLS; Degrees of freedom; Stein's lemma

## 1. Introduction

Representations of the effective dimension of a statistical model have been studied extensively in many different frameworks. For classical model selection criteria such as AIC and Mallows's $C_p$ the dimension of the parameter space is used to adjust the empirical risk for its optimism so as to provide a fair model score across different dimensions. A number of extensions to models or methods without a well defined dimension exist, such as the trace of the smoother matrix for scatter plot smoothers, see e.g. [13], and the use of the divergence of a sufficiently differentiable estimator based on Stein's lemma as described in [5]. Stein's lemma was used by Zou et al. [28] and Tibshirani and Taylor [25] to demonstrate that for the lasso estimator in a linear regression model with Gaussian errors, the number of estimated non-zero parameters is an appropriate estimate of the effective dimension.

It is well known that neither Mallows's $C_p$ nor AIC or related information criteria correctly adjust for the optimism that results from selecting one model among a number of models of equal dimension. The usage of such methods for model selection without adequate adjustments was called "a quiet scandal in the statistical community" by Breiman

[1], who proposed a bootstrap based method for risk estimation as an alternative. Ye [27] defined the notion of generalized degrees of freedom for an estimator of the mean in a Gaussian model and showed how to use this number for risk estimation. The results by Ye apply to discontinuous estimators that involve model selection, but his proposal for computing the degrees of freedom was similarly to Breiman's based on refitting models to perturbed data.

If the estimator satisfies the differentiability requirements for Stein's lemma, Lemma 2 in [21], the divergence of the estimator w.r.t. the data is an unbiased estimate of the degrees of freedom in the generalized sense of [27]. This was used by Donoho and Johnstone [4], Meyer and Woodroofe [18], Zou et al. [28], Kato [14] and Tibshirani and Taylor [25] among others to derive formulas for the degrees of freedom of estimators that are Lipschitz continuous.

For estimators with discontinuities Stein's lemma generally breaks down and the divergence will not be an unbiased estimate of the degrees of freedom. Note that an estimator can be continuous or even differentiable almost everywhere – it can be a projection locally – and still be defined globally in such a way that it has non-ignorable discontinuities. This is, in particular, the case in regression when data adaptive variable selection is used to select among a number of projection estimators. Best subset selection is one central example, but variable selection procedures lead in general to non-ignorable discontinuities. A variable selection procedure effectively divides the sample space into a finite number of disjoint regions, with the estimator being a projection, say, on each region. The resulting estimator consisting of a selection step and a projection step will generally be discontinuous on the boundary between two regions.

Tibshirani [24] recently made headway with the computation of the degrees of freedom for some discontinuous estimators. Specifically, he considered a linear regression model with an orthogonal design and showed how to compute the degrees of freedom for hard thresholding, which for orthogonal designs is equivalent to the Lagrangian formulation of best subset selection. He also gave an extension of Stein's lemma to some discontinuous estimators, though it was not shown if this extension applies to subset selection estimators. Hansen and Sokol [12] gave a different generalization of Stein's lemma for all estimators that are metric projections onto a closed set. This generalization applies to subset selection and other estimators with non-convex constraints, but did not lead to a readily computable representation of the contribution to the degrees of freedom that are due to the discontinuities of the metric projection.

The first main contribution of this paper is the general Theorem 2.4, which is a version of Stein's lemma for estimators that are locally Lipschitz continuous on each of a finite number of open sets, whose union makes up Lebesgue almost all of $\mathbb{R}^n$. This is a broad class of estimators containing a number of regression estimators that include variable selection. Compared to existing results, Theorem 2.4 holds under verifiable conditions without putting restrictions on the design matrix such as orthogonality.

As a main example the lasso-OLS estimator in a linear regression setup is investigated in detail in Section 3. The lasso-OLS estimator consists of two steps: variable selection using lasso followed by ordinary least squares estimation using the selected variables. This estimator was referred to as the LARS-OLS hybrid in [6], and it is a limit case of the relaxed lasso as considered in [17]. We follow the terminology of [2], p. 34, and call it the lasso-OLS estimator.

The second main contribution of this paper is a derivation of a computable estimate of the degrees of freedom – and thus the risk – for lasso-OLS, which only involves the computation of a single lasso solution path and corresponding OLS estimators along the path. Simulation studies reported in Section 4 demonstrated that the resulting risk estimate leads to reliable model selection across a range of different designs and parameter settings, and that the risk estimate itself has smaller mean squared error than the computationally more demanding cross-validation estimate.

For the Lagrangian formulation of best subset selection it is also demonstrated that Theorem 2.4 holds, but the situation is more complicated than for lasso-OLS. However, it is possible to derive an approximation, which is exact for orthogonal designs, as shown in Section 5.

The proof of Theorem 2.4 and some auxiliary technical results are in the appendix.

## 2. A general representation of degrees of freedom

Throughout the paper we consider the multivariate Gaussian model $\mathcal{N}(\mu, \sigma^2 I)$ on $\mathbb{R}^n$ with $\mu$ the unknown parameter, and we let $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}^n$ denote an estimator of $\mu$. A typical application is to linear regression estimators of the form $X\hat{\beta}$ where $X$ denotes an $n \times p$ matrix and $\hat{\beta}$ denotes an estimator of the parameters in the linear regression model. When the estimator $\hat{\beta}$ sets some of the parameters to exactly zero we say that the estimator does variable selection. The lasso, [22], is an example of a globally Lipschitz continuous estimator that does variable selection, while best subset selection is a discontinuous estimator that does variable selection. The lasso-OLS – as studied

intensively in Section 3 – is another example of a discontinuous regression estimator that does variable selection. Though discontinuous regression estimators that do variable selection constitute the main motivation for the present paper, the general results are more conveniently formulated in terms of estimators of the mean $\mu$ without reference to the regression setup.

Letting $Y \sim \mathcal{N}(\mu, \sigma^2 I)$ the risk of the estimator is defined as

$$\mathrm{Risk}(\hat{\mu}) := E\left\|\mu - \hat{\mu}(Y)\right\|_2^2,$$

provided that $\hat{\mu}(Y)$ has finite second moment, which will thus be assumed throughout. The risk is a quantification of the error of $\hat{\mu}$, and tuning parameters are often chosen by minimising an estimate of the risk. Our main interest is to estimate the risk under the Gaussian model. The following definition introduces two notions of degrees of freedom that are useful when we want to estimate the risk. In the definition, $\psi(y; \mu, \sigma^2)$ denotes the density for the $\mathcal{N}(\mu, \sigma^2 I)$ distribution and $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^n$. The divergence operator is also needed. It is the differential operator defined as

$$\mathrm{div}(f) = \sum_{i=1}^{n} \partial_i f_i$$

for $f : \mathbb{R}^n \to \mathbb{R}^n$ Lebesgue almost everywhere differentiable and with $\partial_i$ denoting the partial derivative w.r.t. the $i$th coordinate.

**Definition 2.1.** For a measurable map $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\hat{\mu}(Y)$ has finite second moment the degrees of freedom of $\hat{\mu}$ is

$$\mathrm{df}(\hat{\mu}) := \sum_{i=1}^{n} \frac{\mathrm{cov}(Y_i, \hat{\mu}(Y)_i)}{\sigma^2} = \int \frac{\langle y - \mu, \hat{\mu}(y) \rangle}{\sigma^2} \psi\big(y; \mu, \sigma^2\big)\, dy. \tag{1}$$

If $\hat{\mu}$ is differentiable in Lebesgue almost all points and $\mathrm{div}(\hat{\mu})$ has finite first moment Stein's degrees of freedom of $\hat{\mu}$ is

$$\mathrm{df}_S(\hat{\mu}) := E\big(\mathrm{div}(\hat{\mu})(Y)\big). \tag{2}$$

A simple expansion of the risk yields

$$\mathrm{Risk} = E\left\|Y - \hat{\mu}(Y)\right\|_2^2 - n\sigma^2 + 2\sigma^2\, \mathrm{df}(\hat{\mu}). \tag{3}$$

Hence $\|Y - \hat{\mu}(Y)\|_2^2 - n\sigma^2 + 2\sigma^2 \widehat{\mathrm{df}}$ is an unbiased risk estimate if $\widehat{\mathrm{df}}$ is an unbiased estimate of $\mathrm{df}(\hat{\mu})$. In practice, $\sigma^2$ must be estimated as well and a bias of $\widehat{\mathrm{df}}$ can also be preferable if it reduces the variance. Hence exact unbiasedness of a risk estimate based on (3) is of secondary interest, but it is of interest to find adequate corrections of the squared error $\|Y - \hat{\mu}(Y)\|_2^2$ that can be used for model assessment and comparison.

If $\hat{\mu}$ is *almost differentiable* then $\mathrm{df}(\hat{\mu}) = \mathrm{df}_S(\hat{\mu})$ due to Stein's lemma (Lemma 2 in [21]), in which case $\mathrm{div}(\hat{\mu})(Y)$ is an unbiased estimate of $\mathrm{df}(\hat{\mu})$. However, most estimators with discontinuities are not almost differentiable, and for such estimators it is not clear if $\mathrm{div}(\hat{\mu})(Y)$ is a useful estimate of the degrees of freedom. Indeed, our main result, Theorem 2.4, provides a representation of $\mathrm{df}(\hat{\mu}) - \mathrm{df}_S(\hat{\mu})$, which is nonzero for a range of estimators. The result provides the theoretical basis for establishing more adequate estimates of the degrees of freedom and thus the risk. Furthermore, Theorem 3.2 provides a quite remarkable connection between $\mathrm{df}(\hat{\mu})$ and $\mathrm{df}_S(\hat{\mu})$ for the lasso-OLS estimator, which can be used to derive an estimate of $\mathrm{df}(\hat{\mu})$. This result is directly applicable in practice and provides fast and accurate risk estimation without the need for cross-validation, say.

Our main result is derived under the assumptions on the estimator as stated below. To fix notation we let $B(x, r)$ denote the closed ball in $\mathbb{R}^n$ of radius $r$ and center $x$. Additionally, we let $\mathcal{H}^{n-1}$ denote the $n-1$ dimensional Hausdorff measure – a generalisation of the surface measure of $n-1$ dimensional hypersurfaces in $\mathbb{R}^n$ (see e.g. [7] for details).

**Assumption 2.2.** The estimator $\hat{\mu}$ can be written as $\hat{\mu} = \sum_{i=1}^{N} 1_{U_i} \hat{\mu}_i$ for a collection of open and disjoint sets $\{U_i\}_{i=1}^{N}$ with $\bigcup_{i=1}^{N} \overline{U}_i = \mathbb{R}^n$. Additionally, for each $i = 1, \ldots, N$:

(a) The map $\hat{\mu}_i : \overline{U}_i \to \mathbb{R}^n$ is locally Lipschitz.
(b) The random variable $1_{U_i} \operatorname{div}(\hat{\mu}_i)(Y)$ has finite first moment and $\|\hat{\mu}_i\|$ is polynomially bounded on $U_i$.
(c) The function $r \mapsto \mathcal{H}^{n-1}(\partial U_i \cap B(0, r))$ is polynomially bounded.

**Remark 2.3.** The following points are worth noting:

(a) *Boundary values of the estimator.* Assumption 2.2(c) implies that the boundaries of the sets $U_i$ are Lebesgue null sets, and thus that $\mathbb{R}^n \setminus \bigcup_i U_i$ has Lebesgue measure zero. The estimator $\hat{\mu}$ is here defined to be zero on this null set, but with $Y$ having an absolutely continuous distribution its value on a null set is irrelevant. Note, however, that Assumption 2.2(a) ensures that $\hat{\mu}_i$ is uniquely defined on $\partial U_i$. In a concrete case there may be a natural way to define $\hat{\mu}$ on the common boundary between $U_i$ and $U_j$, say, but we make no abstract attempt to select between $\mu_i$ and $\mu_j$ on the boundary.
(b) *Degrees of freedom.* Assumption 2.2(a) implies by Rademacher's theorem (Theorems 3.1.6 and 3.1.7 in [9]) that $\operatorname{div}(\hat{\mu}_i)$ is defined Lebesgue a.e. Combining this with Assumption 2.2(b) we conclude that under Assumption 2.2 both $\operatorname{df}(\hat{\mu})$ and $\operatorname{df}_S(\hat{\mu})$ are well defined.
(c) *Existence of normal vectors.* Assumption 2.2(c) implies that the sets $U_i$ have locally finite perimeter (see Theorem 5.11.1 in [7]), thus a measure theoretic outer unit normal $\eta_i$ is defined on a subset of $\partial U_i$. In fact, by Lemma A.2 Assumption 2.2(c) only needs to hold for the reduced boundary $\partial^* U_i$ (see Definition 5.7 and Lemma 5.8.1 in [7]). Whenever $\partial U_i$ is smooth the measure theoretic unit normal coincides with the usual pointwise unit normal.

Estimators that involve data driven variable selection will generally fulfil Assumption 2.2 with each $U_i$ corresponding to a set of selected variables. Example 2.5 provides a thorough characterization of $U_i$ in the lasso-OLS setup. Moreover, a similar characterization of $U_i$ is given in Example 3.4 for a class of estimators defined via minimization of a penalized loss function.

The conditions in Assumption 2.2 are typically easy to verify, except perhaps the third condition, as it involves bounding Hausdorff measures. Section A.1 provides some results that can be helpful for verifying the third condition. For estimators satisfying Assumption 2.2 we have the following representation of the degrees of freedom.

**Theorem 2.4.** *If $\hat{\mu}$ satisfies Assumption 2.2 then*

$$\operatorname{df}(\hat{\mu}) = \operatorname{df}_S(\hat{\mu}) + \frac{1}{2} \sum_{i \neq j} \int_{\overline{U}_i \cap \overline{U}_j} \langle \hat{\mu}_j - \hat{\mu}_i, \eta_i \rangle \psi\big(\cdot; \mu, \sigma^2\big)\, d\mathcal{H}^{n-1}, \tag{4}$$

*where $\eta_i$ denotes the measure theoretic outer unit normal to $\partial U_i$.*

The proof is in Section A.2. The essential part is an application of a generalized version of Gauss–Green's formula combined with a dominated convergence argument. Note that though $\overline{U}_i \cap \overline{U}_j$ is a Lebesgue null set – on which $\hat{\mu}$ is defined to be zero – $\hat{\mu}_j$ and $\hat{\mu}_i$ are uniquely defined by Assumption 2.2(a) and generally non-zero and different, cf. also Remark 2.3(a).

If $\hat{\mu}$ satisfies Assumption 2.2 and is continuous then (4) reduces to $\operatorname{df}(\hat{\mu}) = \operatorname{df}_S(\hat{\mu})$, which is Stein's lemma for a class of locally Lipschitz continuous estimators. The boundary integrals therefore account for potential jumps of $\hat{\mu}$ across the boundary of any two adjacent regions $U_i$ and $U_j$. For two-step procedures consisting of a model selection step followed by a parameter estimation step, $\operatorname{df}_S$ generally only accounts for the contribution to the degrees of freedom by the estimation step, and the boundary integrals account for the contribution from the selection step.

The following example illustrates how to verify Assumption 2.2 for the lasso-OLS estimator, which is the estimator that will also be the main focus of the subsequent section.

**Example 2.5 (The lasso-OLS estimator).** Let $X$ be an $n \times p$-matrix. For any subset $A \subseteq \{1, \ldots, p\}$, $X_A$ denotes the matrix whose columns are those of $X$ indexed by $A$, and similarly, $\beta_A \in \mathbb{R}^{|A|}$ denotes $(\beta_i)_{i \in A}$ for $\beta \in \mathbb{R}^p$. We let

$$\mathcal{S} := \big\{ S = \operatorname{col}(X_A) \mid A \subseteq \{1, \ldots, p\} \big\}$$

denote the set of subspaces spanned by columns of $X$. The orthogonal projection onto a subspace $S \in \mathcal{S}$ is denoted by $\Pi_S$.

A *lasso estimator* $\hat{\mu}^\lambda_{\text{lasso}}(y)$ with tuning parameter $\lambda > 0$ is defined as $\hat{\mu}^\lambda_{\text{lasso}}(y) = X\hat{\beta}^\lambda$ where

$$\hat{\beta}^\lambda \in \arg\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

We do not make any assumptions on $X$, and therefore it may happen that multiple $\hat{\beta}^\lambda$-solutions exist. For a solution $\hat{\beta}^\lambda$, the support, $\operatorname{supp}(\hat{\beta}^\lambda) \subseteq \{1, \dots, p\}$, is called an *active set*. The lasso estimator $\hat{\mu}^\lambda_{\text{lasso}}(y) = X\hat{\beta}^\lambda$ belongs to the space $\operatorname{col}(X_A)$ for $A = \operatorname{supp}(\hat{\beta}^\lambda)$, and it follows by Lemma 7 in [25] that there exists a Lebesgue null set $N$, such that $\operatorname{col}(X_A)$ is invariant with respect to the choice of the active set of solutions for $y \notin N$. The map $\widehat{S}^\lambda : \mathbb{R}^n \setminus N \to \mathcal{S}$ returning $\operatorname{col}(X_A)$ when there is a solution $\hat{\beta}^\lambda$ with active set $A = \operatorname{supp}(\beta^\lambda)$ is therefore well defined. The *lasso-OLS* estimator $\hat{\mu}^\lambda_{\text{l-OLS}} := \Pi_{\widehat{S}^\lambda}$ is defined as the projection onto the space selected by the lasso, and is thus well-defined Lebesgue almost everywhere.

By defining the disjoint selection events

$$U^\lambda_S := \big(\widehat{S}^\lambda = S\big)$$

for each $S \in \mathcal{S}$, we immediately see from Lemma 6 in [25] that each selection event is open and that $\mathbb{R}^n = \bigcup_{S\in\mathcal{S}} \overline{U}^\lambda_S$. We can safely ignore any empty $U^\lambda_S$. From the proof of Lemma 6 in [25] we see that $\partial U^\lambda_S \subseteq (\bigcup_{T\in\mathcal{S}} U^\lambda_T)^c$ is a finite union of affine subspaces of dimensions $\leq n-1$, and $r \mapsto \mathcal{H}^{n-1}(\partial U^\lambda_S \cap B(0,r))$ is thus polynomially bounded. This follows by elementary considerations, but it is also a consequence of Lemma A.1. Consequently,

$$\hat{\mu}^\lambda_{\text{l-OLS}} = \sum_{S\in\mathcal{S}} 1_{U^\lambda_S} \Pi_S \quad \text{almost everywhere,}$$

and it satisfies all conditions in Assumption 2.2. Figure 1 provides an illustration of the partition of $\mathbb{R}^n$ for $n = p = 2$ for different choices of angles between the columns in $X$.

Note that since $\hat{\mu}^\lambda_{\text{l-OLS}} = \Pi_S$ on the open set $U^\lambda_S$, its divergence equals $\dim(S)$, hence Stein's degrees of freedom is

$$\operatorname{df}_S\big(\hat{\mu}^\lambda_{\text{l-OLS}}\big) = E\big(\dim\big(\widehat{S}^\lambda\big)\big).$$

From Lemma 3 in [23] it follows that $\dim(\widehat{S}^\lambda) = |\operatorname{supp}(\hat{\beta}^\lambda)|$ whenever the columns of $X$ are in general position, which is useful for practical computations.

The arguments above are based on results in [25], but see also [15] for related characterizations of the selection events for lasso.

## 3. Risk estimation for lasso-OLS

It is not obvious how the general formula in Theorem 2.4 for $\operatorname{df}(\hat{\mu})$ can be used for computing or estimating the degrees of freedom. The first term of (4), $\operatorname{df}_S(\hat{\mu})$, may be estimated by $\operatorname{div}(\hat{\mu})(Y)$, but the second term is more difficult. In this section we show how this second term can be related to the derivative of $\lambda \mapsto \operatorname{df}_S(\hat{\mu}^\lambda_{\text{l-OLS}})$ for lasso-OLS. First we recapitulate the computations in [24] of the degrees of freedom for lasso-OLS with $X$ orthogonal, which will reveal the general formula shown below.

**Example 3.1 (Continuation of Example 2.5).** Assume that $n = p$ and $X = I$. In this case it is well known that the lasso and the lasso-OLS estimators become the soft and hard thresholding estimators, respectively. That is,

$$\hat{\mu}^\lambda_{\text{lasso},i} = \begin{cases} Y_i - \lambda \operatorname{sign}(Y_i) & \text{if } |Y_i| > \lambda, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \hat{\mu}^\lambda_{\text{l-OLS},i} = \begin{cases} Y_i & \text{if } |Y_i| > \lambda, \\ 0 & \text{otherwise.} \end{cases}$$

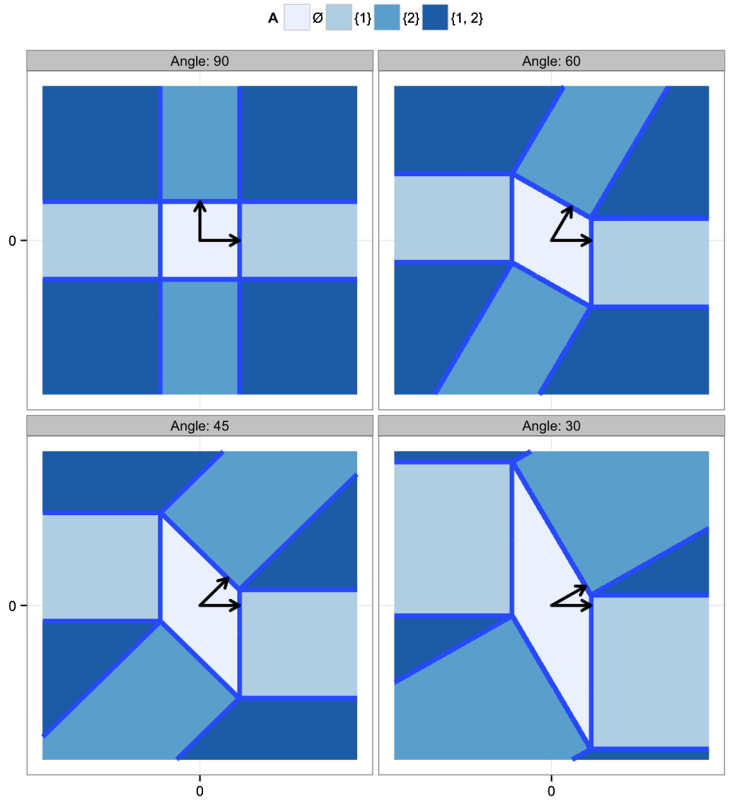


Fig. 1. Illustrations of the decomposition of $\mathbb{R}^2$ into the four sets $U^1_{\varnothing}$, $U^1_{\{1\}}$, $U^1_{\{2\}}$ and $U^1_{\{1,2\}}$ according to the lasso estimator with $\lambda = 1$. The set $U^1_{\varnothing}$ consists of the points shrunk to zero, the sets $U^1_{\{1\}}$ and $U^1_{\{2\}}$ to the points where either the second or the first coordinate, respectively, is shrunk to zero and $U^1_{\{1,2\}}$ to the set where none of the coordinates are shrunk to zero. The decomposition depends on the angle between the two columns in $X$.

We can write up closed form expressions for $\mathrm{df}(\hat{\mu}^{\lambda}_{\text{1-OLS}})$ and $\mathrm{df}_S(\hat{\mu}^{\lambda}_{\text{1-OLS}})$:

$$\begin{aligned}\mathrm{df}_S\big(\hat{\mu}^{\lambda}_{\text{1-OLS}}\big) &= \int \psi\big(y;\mu,\sigma^2\big)\sum_i 1_{(|y_i|>\lambda)}\,dy = \sum_i \int_{(|y_i|>\lambda)} \psi\big(y_i;\mu_i,\sigma^2\big)\,dy_i \\ &= \sum_i \Phi\left(\frac{-\lambda-\mu_i}{\sigma}\right) + \left(1-\Phi\left(\frac{\lambda-\mu_i}{\sigma}\right)\right),\end{aligned}$$

and as in [24]

$$\begin{aligned}\mathrm{df}\big(\hat{\mu}^{\lambda}_{\text{1-OLS}}\big) &= \sum_i \int_{\lambda}^{\infty} \psi\big(y_i;\mu_i,\sigma^2\big)\frac{y_i(y_i-\mu_i)}{\sigma^2}\,dy_i + \int_{-\infty}^{-\lambda} \psi\big(y_i;\mu_i,\sigma^2\big)\frac{y_i(y_i-\mu_i)}{\sigma^2}\,dy_i \\ &= \sum_i \big[-\psi\big(y_i;\mu_i,\sigma^2\big)y_i\big]_{\lambda}^{\infty} + \int_{\lambda}^{\infty} \psi\big(y_i;\mu_i,\sigma^2\big)\,dy_i \\ &\quad + \big[-\psi\big(y_i;\mu_i,\sigma^2\big)y_i\big]_{-\infty}^{-\lambda} + \int_{-\infty}^{-\lambda} \psi\big(y_i;\mu_i,\sigma^2\big)\,dy_i \\ &= \lambda\sum_i \big(\psi\big(\lambda;\mu_i,\sigma^2\big) + \psi\big(-\lambda;\mu_i,\sigma^2\big)\big) + \mathrm{df}_S\big(\hat{\mu}^{\lambda}_{\text{1-OLS}}\big).\end{aligned}$$

Letting $\partial_\lambda$ denote the differential operator with respect to $\lambda$ we observe that

$$\mathrm{df}\big(\hat{\mu}^{\lambda}_{\text{l-OLS}}\big) = \mathrm{df_S}\big(\hat{\mu}^{\lambda}_{\text{l-OLS}}\big) - \lambda \partial_\lambda \,\mathrm{df_S}\big(\hat{\mu}^{\lambda}_{\text{l-OLS}}\big), \tag{5}$$

which is a striking identity. This is because the formula for $\mathrm{df}(\hat{\mu}^{\lambda}_{\text{l-OLS}})$, though explicit, involves the unknown parameter $\mu$ and is not readily estimable. But we have the divergence estimator, $\sum_i 1_{(|y_i|>\lambda)}$, of $\mathrm{df_S}(\hat{\mu}^{\lambda}_{\text{l-OLS}})$, and if we from this can estimate its derivative as well, the formula above suggests how to estimate $\mathrm{df}(\hat{\mu}^{\lambda}_{\text{l-OLS}})$.

The remarkable fact that we will show is that (5) holds without the orthogonality assumption on $X$.

**Theorem 3.2.** *For the lasso-OLS estimator defined in Example 2.5 it holds that*

$$\mathrm{df}\big(\hat{\mu}^{\lambda}_{\text{l-OLS}}\big) = \mathrm{df_S}\big(\hat{\mu}^{\lambda}_{\text{l-OLS}}\big) - \lambda \partial_\lambda \,\mathrm{df_S}\big(\hat{\mu}^{\lambda}_{\text{l-OLS}}\big), \tag{6}$$

*where $\partial_\lambda$ denotes differentiation w.r.t. $\lambda$.*

Theorem 3.2 suggests that $\mathrm{df}(\hat{\mu}^{\lambda}_{\text{l-OLS}})$ can be estimated by differentiation of an estimate of $\mathrm{df_S}(\hat{\mu}^{\lambda}_{\text{l-OLS}})$. The divergence estimate of Stein's degrees of freedom is, however, not differentiable as a function of $\lambda$, and we need to somehow smooth it. To this end it is convenient to reparametrize the penalization in terms of $\delta = \log(\lambda)$, so that with

$$h(\delta) := \mathrm{df_S}\big(\hat{\mu}^{\exp(\delta)}_{\text{l-OLS}}\big),$$

then

$$\mathrm{df}\big(\hat{\mu}^{\exp(\delta)}_{\text{l-OLS}}\big) = h(\delta) - h'(\delta).$$

In simulations $h$ was found to be monotonically decreasing, and thus $h'$ to be negative, but we cannot prove that this is generally the case. The integral representation of $h'$ from Theorem 2.4 is not particularly helpful as the integrand can, in fact, be negative. Based on our computational observations – and to reduce variance of the resulting estimate – our proposal is based on the assumption that $h'$ is negative. It is effectively a kernel smoother that estimates the intensity of jumps for a monotone jump process.

We note that $\dim(\hat{S}^{\exp(\delta)})$ is an unbiased estimate of $h(\delta)$ and that the function $\delta \mapsto \dim(\hat{S}^{\exp(\delta)})$ is a step function. The problem of estimating the derivative, $h'$, of its mean is thus analogous to estimating the intensity for a jump process with one main difference; the step function can have jumps of negative as well as positive sign, though most jumps will be negative. Our proposed estimate ignores the positive excursions of the step function and is computed as follows:

- Compute the jump points, $\lambda_i$ and jump sizes, $\Delta_i := \inf_{\lambda<\lambda_i} \dim(\hat{S}^{\lambda}) - \dim(\hat{S}^{\lambda+})$, of the decreasing function $\lambda \mapsto \inf_{\lambda'<\lambda} \dim(\hat{S}^{\lambda'})$ for $i = 1, \ldots, M$.
- Apply a kernel density smoother to the points $\delta_i = \log(\lambda_i)$ for $i = 1, \ldots, M$ counted with the multiplicities $\Delta_i$. In the simulations presented in this paper an adaptive Gaussian kernel density smoother was used (see Section 10.4.3.2 in [11]).
- Rescale the density estimate by the total number of jumps, that is, by $\sum_{i=1}^{M} \Delta_i$.

As mentioned above, we can think of the proposed estimate of $h'$ as a non-parametric estimate of the intensity of the jumps for a monotonically decreasing jump process. Alternatively, we can think of it as smoothing the jumps by a sigmoidal function (the anti-derivative of the kernel) to obtain a smooth estimate of Stein's degrees of freedom, which can then be differentiated. Note that even if $\Delta_i$ may always be 1 in theory, the jumps are in practice computed on a grid and may thus be larger than 1, which the procedure accounts for. The estimate of $-\lambda\partial_\lambda \,\mathrm{df_S}(\hat{\mu}^{\lambda}_{\text{l-OLS}})$ resulting from the procedure above is denoted by $\widehat{\partial}$.

Using $\dim(\hat{S}^{\lambda}) + \widehat{\partial}$ as an estimate of degrees of freedom leads to the risk estimate

$$\widehat{\mathrm{Risk}}_{\mathrm{df}} := \big\|Y - \hat{\mu}^{\lambda}_{\text{l-OLS}}\big\|_2^2 - n\sigma^2 + 2\sigma^2\big(\dim\big(\hat{S}^{\lambda}\big) + \widehat{\partial}\big). \tag{7}$$

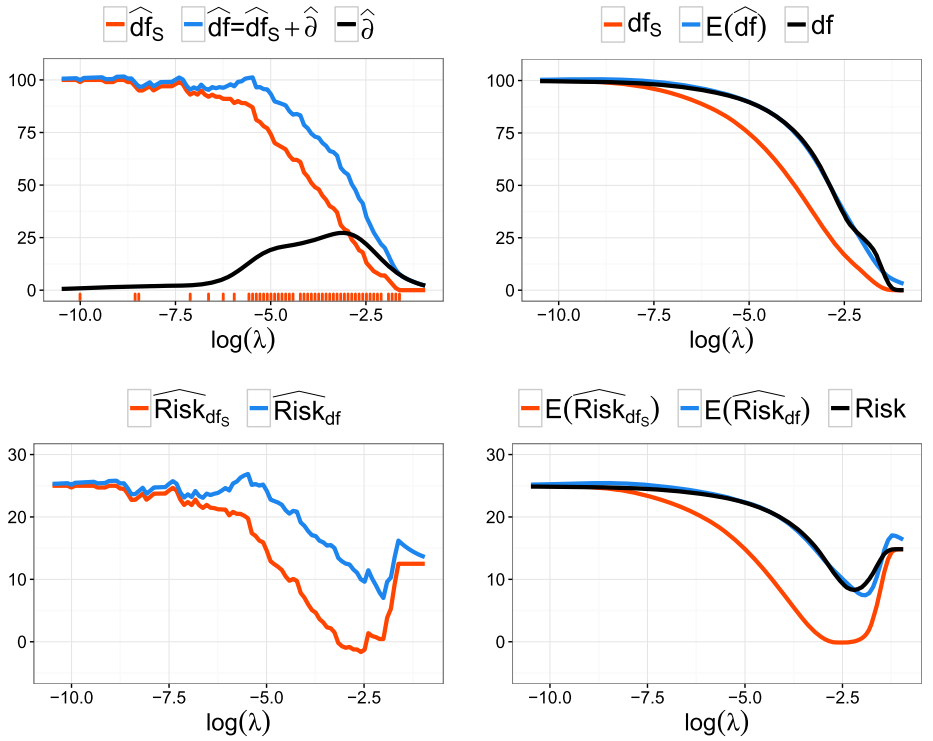


Fig. 2. Left: Realization of the estimates of degrees of freedom $\hat{\mathrm{df}}_S = \dim(\hat{S}^\lambda)$ and $\hat{\mathrm{df}} = \dim(\hat{S}^\lambda) + \widehat{\partial}$ as well as the correction term $\widehat{\partial}$ as a function of $\log(\lambda)$ (top) and corresponding estimates of the risk (bottom). Right: Similar to the left but mean values of the estimates obtained by averaging over 1000 samples along with the degrees of freedom $\mathrm{df} = \mathrm{df}(\hat{\mu}^\lambda_{\text{1-OLS}})$ obtained from the 1000 samples using the covariance definition (1). The design parameters were: $\sigma = 0.5$, $n = p = 100$, $\gamma = 1$, $\alpha = 0.1$ and the design type was (S) with constant correlation of $\rho = 0.1$ (see Section 4).

For an example of the above estimate see Figure 2, where $\widehat{\partial}$ and $\widehat{\mathrm{Risk}}_{\mathrm{df}}$ are applied to a single realization of $Y$ along with an average over 1000 replications.

To prove Theorem 3.2 we prove a more general intermediate result for estimators that are parametrized in a similar way by a tuning parameter. We use in the following $D$ to denote the differential operator w.r.t. $y$.

**Proposition 3.3.** *Let $q > 0$ and suppose that $\hat{\mu}^\lambda = \sum_i 1_{U_i^\lambda} \hat{\mu}_i$ where*

$$U_i^\lambda = \lambda^q U_i^1, \quad \text{for all } i = 1, \ldots, N. \tag{8}$$

*Assume that $\mathrm{div}(\hat{\mu}_i)$ is locally Lipschitz and both $\mathrm{div}(\hat{\mu}_i)$ and $D(\mathrm{div}(\hat{\mu}_i))$ are polynomially bounded for each $i = 1, \ldots, N$. If $\hat{\mu}^1$ satisfies Assumption 2.2 then*

$$-\frac{\lambda}{q}\partial_\lambda \mathrm{df}_S(\hat{\mu}^\lambda) = \frac{1}{2}\sum_{i\neq j}\int_{\overline{U}_i^\lambda \cap \overline{U}_j^\lambda} \big(\mathrm{div}(\hat{\mu}_j)(y) - \mathrm{div}(\hat{\mu}_i)(y)\big)\langle y, \eta_i\rangle \psi\big(y; \mu, \sigma^2\big)\, d\mathcal{H}^{n-1}(y). \tag{9}$$

**Proof.** First observe that $\partial U_i^\lambda \cap B(0,r) = \lambda^q(\partial U_i^1 \cap B(0, r/\lambda^q))$, hence if $\hat{\mu}^1$ satisfies Assumption 2.2 so does $\hat{\mu}^\lambda$ for all $\lambda$. Next, the change of variable formula yields

$$\begin{aligned}\mathrm{df}_S(\hat{\mu}^\lambda) &= \int \psi(y)\,\mathrm{div}\big(\hat{\mu}^\lambda\big)(y)\,dy = \sum_i \int_{U_i^\lambda} \psi(y)\,\mathrm{div}(\hat{\mu}_i)(y)\,dy \\ &= \sum_i \int_{U_i^1} \lambda^{qn}\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)\big(\lambda^q z\big)\,dz.\end{aligned}$$

Here $\psi = \psi(\cdot; \mu, \sigma^2)$ to ease notation.

The last integrand is differentiable w.r.t. $\lambda$ (for Lebesgue a.a. $z$) and its derivative is

$$\begin{aligned}&qn\lambda^{qn-1}\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)\big(\lambda^q z\big)+\lambda^{qn}\big\langle D\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)\big(\lambda^q z\big),q\lambda^{q-1}z\big\rangle\\&\quad=\frac{q}{\lambda}\lambda^{qn}\big(n\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)\big(\lambda^q z\big)+\big\langle D\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)\big(\lambda^q z\big),\lambda^q z\big\rangle\big),\end{aligned}$$

which is dominated in a neighbourhood of $\lambda$ by an integrable function due to the polynomial bounds. Hence, by the change of variable formula

$$\begin{aligned}\frac{\lambda}{q}\partial_\lambda\,\mathrm{df}_S\big(\hat{\mu}^\lambda\big)&=\sum_i\int_{U_i^1}\lambda^{qn}\big(n\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)\big(\lambda^q z\big)+\big\langle D\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)\big(\lambda^q z\big),\lambda^q z\big\rangle\big)\,dz\\&=\sum_i\int_{U_i^\lambda}n\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)(y)+\big\langle D\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)(y),y\big\rangle dy\\&=\sum_i\int_{U_i^\lambda}n\big(\psi\,\mathrm{div}(\hat{\mu}_i)\big)(y)+\big\langle\big(\psi D\,\mathrm{div}(\hat{\mu}_i)+\mathrm{div}(\hat{\mu}_i)D\psi\big)(y),y\big\rangle dy\\&=\sum_i\int_{U_i^\lambda}\psi(y)\,\mathrm{div}\big(y\,\mathrm{div}(\hat{\mu}_i)(y)\big)+\big\langle D\psi(y),y\,\mathrm{div}(\hat{\mu}_i)(y)\big\rangle dy.\end{aligned}$$

The last line is identified as $\mathrm{df}_S(\tilde{\mu}^\lambda)-\mathrm{df}(\tilde{\mu}^\lambda)$, where

$$\tilde{\mu}^\lambda(y):=\sum_i 1_{U_i^\lambda}(y)y\,\mathrm{div}(\hat{\mu}_i)(y).$$

Finally (9) follows by applying Theorem 2.4 to $\tilde{\mu}^\lambda$ (which also satisfies Assumption 2.2). □

**Example 3.4.** There are naturally occurring examples besides the lasso selection sets that satisfy (8). Consider still a linear regression setup with $X$ an $n\times p$-matrix. Let $\ell$ denote the penalized loss function

$$\ell(y,\beta,\lambda)=\frac{1}{2}\|y-X\beta\|_2^2+\lambda\,\mathrm{Pen}(\beta),$$

for some penalty function $\mathrm{Pen}:\mathbb{R}^p\to\mathbb{R}$ and define the sets

$$U_A^\lambda=\mathrm{int}\Big\{y\in\mathbb{R}^n\;\Big|\;\inf_{\beta:\mathrm{supp}(\beta)=A}\ell(y,\beta,\lambda)=\inf_\beta\ell(y,\beta,\lambda)\Big\},\tag{10}$$

for each $A\subseteq\{1,\dots,p\}$. Hence any $y\in U_A^\lambda$ has $A$ as an active set. If Pen is *positive homogeneous* of degree $k\in[0,2)$ then

$$\ell\big(\lambda^{\frac{1}{2-k}}y,\lambda^{\frac{1}{2-k}}\beta,\lambda\big)=\lambda^{\frac{2}{2-k}}\ell(y,\beta,1).$$

Hence $U_A^\lambda=\lambda^{\frac{1}{2-k}}U_A^1$ holds for all $A\subseteq\{1,\dots,p\}$ and $\lambda>0$. The (quasi) norms, $\mathrm{Pen}(\beta)=\|\beta\|_k^k$ for $k\in(0,2)$, and $\mathrm{Pen}(\beta)=\|\beta\|_0=|\,\mathrm{supp}(\beta)|$ are examples of positive homogeneous penalties. For these penalties only $k\in[0,1]$ will result in variable selection. With $\mathrm{Pen}(\cdot)=\|\cdot\|_1$ we see that for lasso the sets $U_S^\lambda$ in 2.5 satisfy (8) with $q=1$.

**Proof of Theorem 3.2.** Let $(U_S^\lambda)_{S\in\mathcal{S}}$ be defined as in Example 2.5, where it was also shown that Assumption 2.2 holds for the lasso-OLS estimator. Moreover, from Example 3.4 we see that $U_S^\lambda=\lambda U_S^1$ for all $\lambda>0$ and $S\in\mathcal{S}$. By Theorem 2.4 we know that the left hand side of (6) is

$$\begin{aligned}&\mathrm{df}\big(\hat{\mu}^\lambda_{\text{l-OLS}}\big)-\mathrm{df}_S\big(\hat{\mu}^\lambda_{\text{l-OLS}}\big)\\&\quad=\frac{1}{2}\sum_{S_1\neq S_2}\int_{\overline{U}^\lambda_{S_1}\cap\overline{U}^\lambda_{S_2}}\big\langle(\Pi_{S_2}-\Pi_{S_1})y,\eta_{S_1}(y)\big\rangle\psi(y)\,d\mathcal{H}^{n-1}(y).\end{aligned}\tag{11}$$

It will first be established that $\overline{U}^{\lambda}_{S_1} \cap \overline{U}^{\lambda}_{S_2}$ for $S_1 \neq S_2$ is a $\mathcal{H}^{n-1}$ null set unless $S_1$ and $S_2$ are nested and their dimensions differ by one.

By definition $\hat{\mu}^{\lambda}_{\text{lasso}} \in S$ on $U^{\lambda}_S$, and by continuity of $\hat{\mu}^{\lambda}_{\text{lasso}}$ (a consequence of Lemma 3 in [25]) we conclude that the same is true on $\overline{U}^{\lambda}_S$. Hence for $S_1, S_2 \in \mathcal{S}$

$$\hat{\mu}^{\lambda}_{\text{lasso}} \in S_1 \cap S_2 \quad \text{on } \overline{U}^{\lambda}_{S_1} \cap \overline{U}^{\lambda}_{S_2}. \tag{12}$$

For $A \subseteq \{1, \ldots, p\}$ and $s \in \{-1, 1\}^{|A|}$ we define the set

$$L_{A,s} := \{u \in \mathbb{R}^n \mid X_A^T u = \lambda s\}.$$

It now follows from the first order subgradient conditions for lasso that

$$y - \hat{\mu}^{\lambda}_{\text{lasso}} \in \bigcup_{\substack{A \subseteq \{1,\ldots,p\}: \\ \operatorname{col}(X_A) = S}} \bigcup_{s \in \{-1,1\}^{|A|}} L_{A,s} \tag{13}$$

for all $y \in U^{\lambda}_S$. Note that the dimension of the above set is $n - \dim(S)$. Since the set is closed and $\hat{\mu}^{\lambda}_{\text{lasso}}$ is continuous, (13) holds for $y \in \overline{U}^{\lambda}_S$ as well. We therefore conclude that

$$\begin{aligned} y - \hat{\mu}^{\lambda}_{\text{lasso}} &\in \Bigg( \bigcup_{\substack{A \subseteq \{1,\ldots,p\}: \\ \operatorname{col}(X_A) = S_1}} \bigcup_{s \in \{-1,1\}^{|A|}} L_{A,s} \Bigg) \cap \Bigg( \bigcup_{\substack{A \subseteq \{1,\ldots,p\}: \\ \operatorname{col}(X_A) = S_2}} \bigcup_{s \in \{-1,1\}^{|A|}} L_{A,s} \Bigg) \\ &\subseteq \bigcup_{\substack{A \subseteq \{1,\ldots,p\}: \\ \operatorname{col}(X_A) = S_1 + S_2}} \bigcup_{s \in \{-1,1\}^{|A|}} L_{A,s} \end{aligned} \tag{14}$$

for all $y \in \overline{U}^{\lambda}_{S_1} \cap \overline{U}^{\lambda}_{S_2}$ and $S_1, S_2 \in \mathcal{S}$.

From (12) and (14) we deduce that

$$\overline{U}^{\lambda}_{S_1} \cap \overline{U}^{\lambda}_{S_2} \subseteq S_1 \cap S_2 + \bigcup_{\substack{A \subseteq \{1,\ldots,p\}: \\ \operatorname{col}(X_A) = S_1 + S_2}} \bigcup_{s \in \{-1,1\}^{|A|}} L_{A,s} \tag{15}$$

for $S_1, S_2 \in \mathcal{S}$. Consequently, if $S_1 \neq S_2$ then $\mathcal{H}^{n-1}(\overline{U}^{\lambda}_{S_1} \cap \overline{U}^{\lambda}_{S_2}) = 0$, unless $S_1$ and $S_2$ are nested and their dimensions differ by 1.

We can therefore assume $S_1 \subseteq S_2$ and $\dim(S_2) = \dim(S_1) + 1$. Furthermore, $S_2 \ominus S_1 = (S_1 + S_2) \ominus (S_1 \cap S_2)$ is orthogonal to any of the faces $S_1 \cap S_2 + L_{A,s}$ in (15) and thus also orthogonal to $\overline{U}^{\lambda}_{S_1} \cap \overline{U}^{\lambda}_{S_2}$. This implies that $\eta_{S_1} = (\Pi_{S_2} - \Pi_{S_1})\eta_{S_1}$ and hence (11) becomes

$$\begin{aligned} &\operatorname{df}\big(\hat{\mu}^{\lambda}_{\text{l-OLS}}\big) - \operatorname{df}_S\big(\hat{\mu}^{\lambda}_{\text{l-OLS}}\big) \\ &\quad = \sum_{\substack{S_1 \subseteq S_2, \\ \dim(S_2) = \dim(S_1)+1}} \int_{\overline{U}^{\lambda}_{S_1} \cap \overline{U}^{\lambda}_{S_2}} \big\langle y, \eta_{S_1}(y) \big\rangle \psi(y)\, d\mathcal{H}^{n-1}(y) \\ &\quad = \sum_{\substack{S_1 \subseteq S_2, \\ \dim(S_2) = \dim(S_1)+1}} \int_{\overline{U}^{\lambda}_{S_1} \cap \overline{U}^{\lambda}_{S_2}} \underbrace{\big[\operatorname{div}(\Pi_{S_2}) - \operatorname{div}(\Pi_{S_1})\big]}_{=\dim(S_2) - \dim(S_1) = 1} \big\langle y, \eta_{S_1}(y) \big\rangle \psi(y)\, d\mathcal{H}^{n-1}(y) \\ &\quad = -\lambda \partial_\lambda \operatorname{df}_S\big(\hat{\mu}^{\lambda}_{\text{l-OLS}}\big) \end{aligned}$$

by Proposition 3.3. □

## 4. Simulation study

We report in this section the results from an extensive simulation study, whose purpose was to quantify how $\widehat{\mathrm{Risk}}_{\mathrm{df}}$ given by (7) performs as an estimate of the risk and in terms of selecting the penalty parameter $\lambda$. Its performance was compared to alternatives for risk estimation and tuning, and the resulting lasso-OLS estimator was compared to the lasso estimator. Throughout, the R package *glmnet*, [10], was used to compute the lasso solution path. This section is divided into subsections describing estimators and risk estimates, the design of the simulation study, and the results of the simulation study.

### 4.1. *Estimators and risk estimates*

The first alternative risk estimate for lasso-OLS is

$$\widehat{\mathrm{Risk}}_{\mathrm{df_S}} = \big\|Y - \hat{\mu}^{\lambda}_{\text{l-OLS}}\big\|_2^2 - n\sigma^2 + 2\sigma^2 \dim\big(\hat{S}^{\lambda}\big), \tag{16}$$

which does not adjust for the variable selection performed by lasso-OLS. The second alternative is $K$-fold cross-validation (denoted $\widehat{\mathrm{Risk}}_{\text{CV-K}}$) with $K = 5, 10$. This risk estimate is given by

$$\widehat{\mathrm{Risk}}_{\text{CV-K}} := \sum_{k=1}^{K} \big\|Y_k - X_k \hat{\beta}^{\lambda}_{\text{l-OLS}}(Y_{-k}, X_{-k})\big\|_2^2 - n\sigma^2, \tag{17}$$

where $Y_k$ and $X_k$ denote the entries of $Y$ and rows of $X$, respectively, corresponding to the $k$th fold, and similarly, $Y_{-k}$ and $X_{-k}$ denote the entries and rows not in the $k$th fold.

The lasso estimator was tuned by minimising the risk estimate

$$\widehat{\mathrm{Risk}}_{\text{lasso}} = \big\|Y - \hat{\mu}^{\lambda}_{\text{lasso}}\big\|_2^2 - n\sigma^2 + 2\sigma^2 \dim\big(\hat{S}^{\lambda}\big). \tag{18}$$

For tuning $\in \{\mathrm{df}, \mathrm{df_S}, \text{CV-5}, \text{CV-10}, \text{lasso}\}$ we let $\hat{\lambda}_{\text{tuning}}$ denote the value of $\lambda$ that minimises $\widehat{\mathrm{Risk}}_{\text{tuning}}$. The risk of the resulting estimator is denoted

$$\mathrm{Risk}(\text{tuning}) := E\big\|\mu - \hat{\mu}^{\hat{\lambda}_{\text{tuning}}}_{\text{l-OLS}}\big\|_2^2$$

for all but the lasso-tuning, whose risk instead is

$$\mathrm{Risk}(\text{lasso}) := E\big\|\mu - \hat{\mu}^{\hat{\lambda}_{\text{lasso}}}_{\text{lasso}}\big\|_2^2.$$

When the true mean is $\mu = X\beta$ with $\mathrm{supp}(\beta) = A$ we refer to $\Pi_A$ as the oracle-OLS estimator. This usage of the oracle terminology is in accordance with e.g. [8]. Its risk is

$$E\|\mu - \Pi_A Y\|_2^2 = \sigma^2 \operatorname{rank}(X_A).$$

The results from the simulation study are reported in terms of $\mathrm{Risk}(\text{tuning})/(\sigma^2 n)$ for each tuning method, which can then be compared to $\operatorname{rank}(X_A)/n$ – the fraction of nonzero parameters.

All simulations were carried out assuming either that $\sigma^2$ was known or using the following estimator of $\sigma^2$: first the lasso path $\lambda \mapsto \hat{\mu}_{\text{lasso}}(\lambda)$ was calculated, then $\hat{\lambda}$ was selected by minimising the generalized cross-validation criterion

$$\mathrm{gcv}(\lambda) = \frac{\|Y - \hat{\mu}^{\lambda}_{\text{lasso}}\|_2^2}{(1 - \frac{\dim(\hat{S}^{\lambda})}{n})^2},$$

and $\sigma^2$ was finally estimated as

$$\hat{\sigma}^2 = \frac{\|Y - \hat{\mu}^{\hat{\lambda}}_{\text{lasso}}\|_2^2}{n - \dim(\hat{S}^{\hat{\lambda}})}.$$

The main reason for choosing this estimator was computational efficiency, as the lasso path must be calculated for lasso-OLS anyway. Thus this variance estimate has virtually no extra computational costs. See also [20] for a comprehensive comparison of variance estimators.

### 4.2. *Simulation study design*

In the simulation study the mean was given as $X\beta$ with

$$\beta_i = \begin{cases} \gamma^{i-1} & \text{if } i \leq \lceil n\alpha \rceil, \\ 0 & \text{otherwise,} \end{cases}$$

for different choices of the dimension $n$, the $n \times p$ design matrix $X$ and the parameters $\gamma$ and $\alpha$.

Two simulation designs were implemented with parameters as follows:

| Parameter | Values for simulation study I | | | | | Values for simulation study II | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma$ | 0.5 | | | | | 0.1 | 0.2 | 0.5 | 1 | 2 |
| $\alpha$ | 0.1 | | | | | 0 | 0.05 | 0.1 | 0.3 | 0.5 |
| $n$ | 50 | 100 | 200 | 400 | 800 | 100 | 200 | | | |
| $p$ | 200 | 2000 | 20,000 | | | $n$ | | | | |
| $\gamma$ | 1 | | | | | 1 | 0.9 | | | |
| $X$ | S | | | | | O | S | E | | |
| $\rho$ | 0.1 | | | | | 0 | 0.1 | 0.4 | 0.7 | |

The parameter $\rho$ and the values of the design require some explanation. The three different design types are:

- Orthogonal (O), where $X = I$.
- Simulated (S), where the columns of $X$ are standard normally distributed with one of the following correlation structures:
  - Autoregressive setup: $\text{corr}(X_i, X_j) = \rho^{|i-j|}$ for all $i \neq j$.
  - Constant correlation setup: $\text{corr}(X_i, X_j) = \rho$ for all $i \neq j$.
- Empirical (E), where the rows and columns are randomly selected from the $240 \times 377$ matrix of microRNA expression values as used in the earlier study by [26].

The columns of the simulated and empirical designs were standardized to have norm one to obtain a comparable signal-to-noise ratio across the three designs.

The risk estimates were based on 1000 samples for each combination of the parameters, which were generated as follows. For each of the 1000 samples a design matrix $X$ was created/simulated and a single realization of $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ was drawn. For each sample the losses $\|\mu - \hat{\mu}^{\hat{\lambda}_{\text{lasso}}}_{\text{lasso}}\|_2^2$ and $\|\mu - \hat{\mu}^{\hat{\lambda}_{\text{tuning}}}_{\text{l-OLS}}\|_2^2$ for the different tuning methods were computed. The risks were estimated as the average of the losses over the 1000 samples.

In order to assess robustness to deviations from the Gaussian noise assumption, we replicated the second study design with two types of non-Gaussian noise: a $t$-distribution with 3 degrees of freedom, and a skew normal distribution with shape parameter 3. Location and scale parameters were set so that the noise distribution had mean 0 and variance $\sigma^2$.

### 4.3. *Results from study I*

We first report on the accuracy of the risk estimates. Figure 3 shows the risk estimates as a function of $\lambda$ for 50 samples along with a Monte Carlo estimate of the true risk. Cross-validation appears to give more variable estimates of the risk than $\widehat{\text{Risk}}_{\text{df}}$ across the entire range of $\lambda$-values. This is true even when the variance is estimated, though estimation of the variance does appear to degrade the performance of the risk estimates. We note that $\widehat{\text{Risk}}_{\text{df}}$ does not appear to be much more variable than $\widehat{\text{Risk}}_{\text{lasso}}$, though the former relies on the additional smoothed term for the estimation of degrees of freedom.

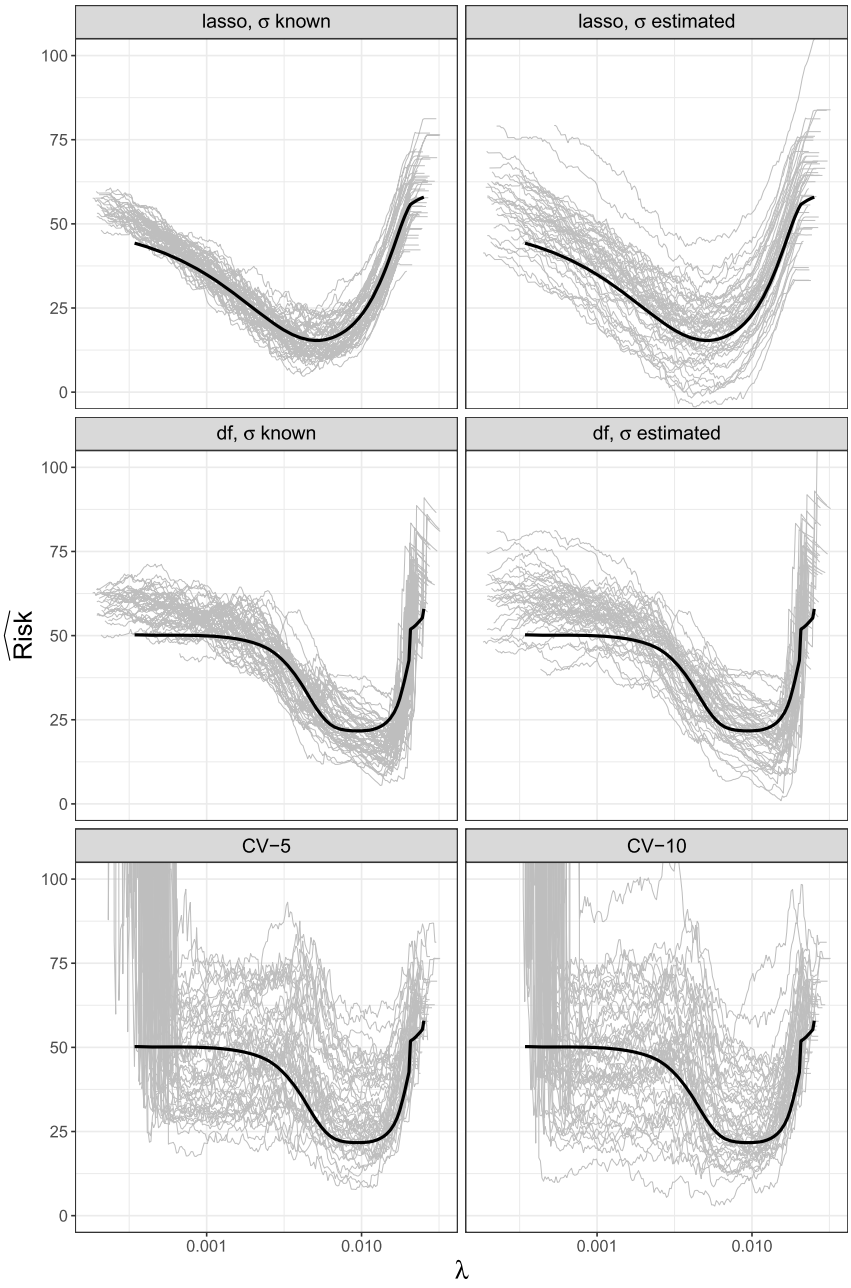


Fig. 3. Risk estimates $\widehat{\text{Risk}}_{\text{df}}$, $\widehat{\text{Risk}}_{\text{CV-5}}$, $\widehat{\text{Risk}}_{\text{CV-10}}$ and $\widehat{\text{Risk}}_{\text{lasso}}$ (gray lines) for 50 samples as a function of $\lambda$. The black lines are Monte Carlo estimates of the true risks. The design parameters were: $n = 200$, $p = 2000$, $\sigma = 0.5$, $\gamma = 1$, $\alpha = 0.1$, and the design type was (S) with a constant correlation of $\rho = 0.1$ (see Section 4.2).

Figure 4 shows mean squared errors (MSEs) for the risk estimates. The figure shows the integrated mean sequared error as well as the mean squared error in the optimal $\lambda$ (the $\lambda$ that minimizes risk as estimated from the Monte Carlo estimate of the risk based on 1000 replications). The cross-validation risk estimates generally have the largest MSEs, while $\widehat{\text{Risk}}_{\text{df}}$ has considerably smaller MSEs. This is true even when the variance is estimated except for $n = 50$ and $p = 2000, 20{,}000$. From this figure we see that $\widehat{\text{Risk}}_{\text{df}}$ does have a larger MSE than $\widehat{\text{Risk}}_{\text{lasso}}$. Moreover, for $n/p$ large the estimation of $\sigma$ does not affect the MSE of the risk estimates much.

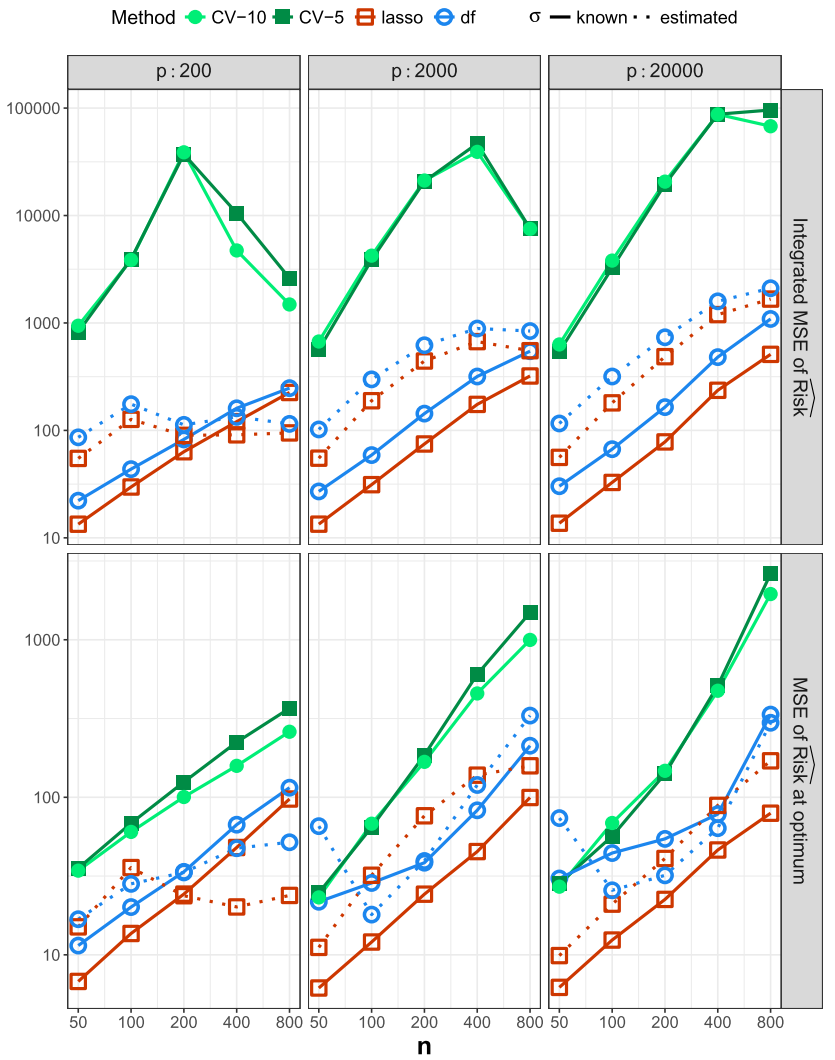


Fig. 4. Integrated mean squared error (top) and mean squared error at the optimal value of $\lambda$, $\hat{\lambda}$ (bottom) of the risk estimates $\widehat{\text{Risk}}_{\text{df}}$, $\widehat{\text{Risk}}_{\text{CV-5}}$, $\widehat{\text{Risk}}_{\text{CV-10}}$ and $\widehat{\text{Risk}}_{\text{lasso}}$. The integrated mean squared error was computed over the interval $[\hat{\lambda}/10, 10\hat{\lambda}]$ of $\log(\lambda)$-values. The design parameters were: $\sigma = 0.5$, $\gamma = 1$, $\alpha = 0.1$, and the design type was (S) with a constant correlation of $\rho = 0.1$ (see Section 4.2).

For this simulation study we also recorded the number of selected predictors as well as the computational time for evaluating and tuning the different estimators. The results can be found as Figure 1 in the supplementary material [19]. The lasso-OLS estimator selects fewer predictors than lasso, but when the variance is estimated, the number of selected predictors is increased – this is particularly so when $n/p$ is small. The lasso estimator using (18) for tuning is fastest, which is unsurprising as the computation of the lasso path is part of all estimators. Moreover, the lasso-OLS estimator using (7) for tuning is about a factor 4 faster than using 5-fold cross-validation for tuning and about a factor 8 faster than 10-fold cross-validation. Thus the added computation of the smoothed term to the estimate of degrees of freedom in (7) has an insignificant effect on the computation time.

### 4.4. *Results from study II*

Firstly, we discuss the comparison of the two tuning methods df and $\text{df}_S$ for the lasso-OLS estimator. The purpose of this comparison is to highlight the effect of correctly adjusting for the variable selection in the estimation of degrees of freedom via the term $\widehat{\partial}$. Secondly, we discuss the comparison of df to CV-5, CV-10 and lasso. The purpose of this

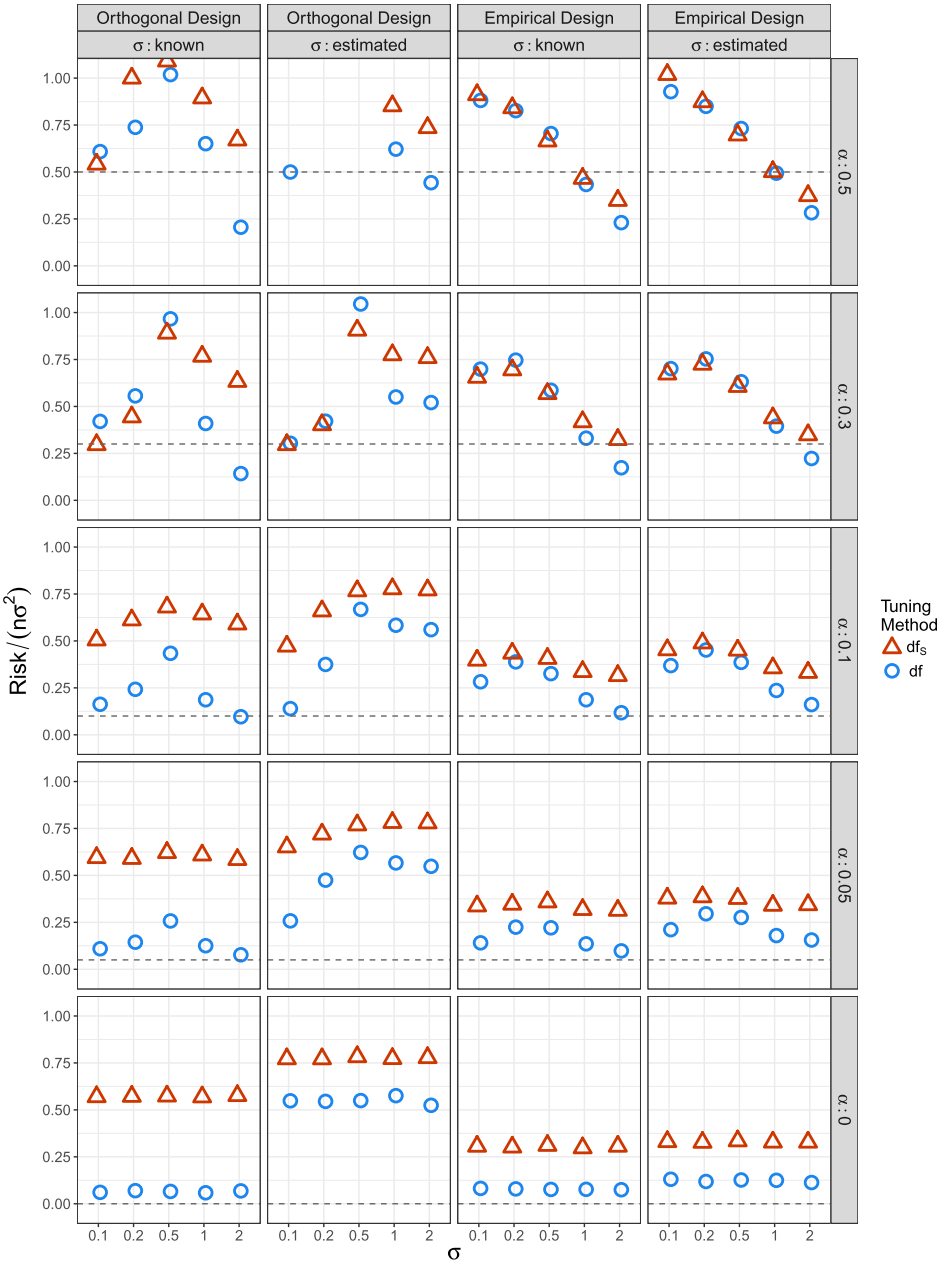


Fig. 5. Risk relative to $\sigma^2 n$ for the estimators $\hat{\mu}^{\hat{\lambda}_{\mathrm{df}_S}}_{\text{l-OLS}}$ and $\hat{\mu}^{\hat{\lambda}_{\mathrm{df}}}_{\text{l-OLS}}$ for orthogonal and empirical designs with $n = 100$ and $\gamma = 1$. The dashed line is $\lceil n\alpha \rceil / n \simeq \alpha$, the relative risk for the oracle-OLS estimator.

second comparison is two-fold. It provides a comparison of our proposed tuning method, df, to cross-validation based tuning, and it provides a comparison of lasso-OLS to lasso in terms of predictive performance.

Figure 5 shows the results for the two tuning methods df and $\mathrm{df}_S$ in the orthogonal and empirical designs with $\gamma = 1$ and $n = 100$. The results for all the other design parameters can be found in [19]. Tuning $\lambda$ by using $\dim(\hat{S}^\lambda) + \widehat{\partial}$ as an estimate of degrees of freedom is generally superior to using $\dim(\hat{S}^\lambda)$ and in the worst cases at least comparable. The differences are largest for the lowest signal-to-noise ratios. The benefit of using $\dim(\hat{S}^\lambda) + \widehat{\partial}$ generally increases with the dimension $n$, and it increases with decreasing signal-to-noise ratio. Furthermore, when the number of non-zero parameters is large and the signal-to-noise ratio is low (specifically, $\gamma = 0.9$, $\alpha$ large and $\sigma$ large), $\hat{\mu}^{\hat{\lambda}_{\mathrm{df}}}_{\text{l-OLS}}$ clearly

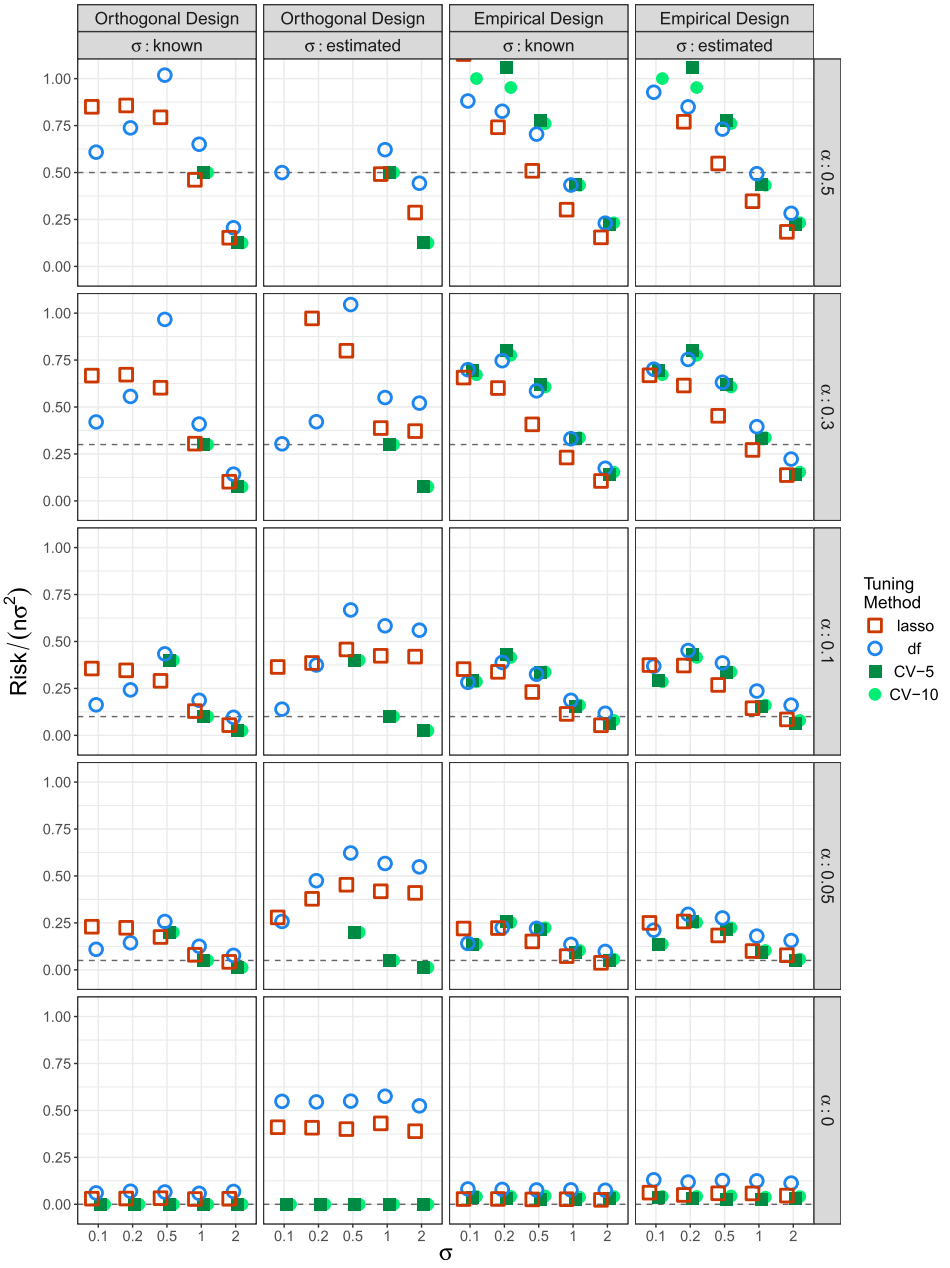


Fig. 6. Risk relative to $\sigma^2 n$ for the estimators $\hat{\mu}^{\hat{\lambda}_{\text{df}}}_{\text{l-OLS}}$, $\hat{\mu}^{\hat{\lambda}_{\text{CV-5}}}_{\text{l-OLS}}$, $\hat{\mu}^{\hat{\lambda}_{\text{CV-10}}}_{\text{l-OLS}}$ and $\hat{\mu}^{\hat{\lambda}_{\text{lasso}}}_{\text{lasso}}$ for orthogonal and empirical designs with $n = 100$ and $\gamma = 1$. The dashed line is $\lceil n\alpha \rceil / n \simeq \alpha$, the relative risk for the oracle-OLS estimator.

outperforms the oracle-OLS estimator, while $\hat{\mu}^{\hat{\lambda}_{\text{df}_S}}_{\text{l-OLS}}$ is comparable or worse than the oracle-OLS estimator. Neither of the estimators performs well for small variances and large signal-to-noise ratios. For the orthogonal design the estimation of the variance incurs a clear performance loss, which is not the case for the other designs. We ascribe this to the variance estimator being particularly poor for the orthogonal design.

Figure 6 shows the results for df, CV-5, CV-10 and lasso for the orthogonal and empirical designs with $\gamma = 1$ and $n = 100$. The results for the remaining design parameters are found in [19]. For the orthogonal design cross-validation is not an appropriate tuning method, since $\widehat{\text{Risk}}_{\text{CV-K}}$ is constant in $\lambda$. This relates to the fact that the folds cannot be considered replications of the same distribution. Consequently, for the orthogonal design, the tuning methods based

on degrees of freedom have clear advantages. On the other hand, the estimation of $\sigma$ has a quite large negative effect for precisely the orthogonal design.

When restricting attention to the non-orthogonal designs we observe that the tuning methods are quite comparable (see [19]). None of the tuning methods are generally superior or inferior to the others and their performance depends on both design type, signal-to-noise ratio and the signal decay parameter $\gamma$. The lasso estimator deviates most from the others, which is mainly due to this being a different estimator. It performs best at low signal-to-noise ratios, while lasso-OLS using either cross-validation of df tuning performs better at high signal-to-noise ratios ($\alpha$ large, $\sigma$ small and $\gamma = 1$). Cross-validation appears to perform best for highly correlated designs ($\rho$ large).

The results for the non-Gaussian error distributions are included in [19] as well. There are no major differences when compared to the Gaussian error distribution, with the most notable change being that lasso losses some of its performance for the $t$-distributed noise. The tuning based on df seems to be less affected. Still, all the tuning methods are generally comparable except for orthogonal designs. Since cross-validation does not rely on a Gaussian noise assumption, these results suggest that our proposed tuning method based on df is appropriate even in non-Gaussian settings.

## 5. Best subset selection

Example 3.4 demonstrates that (8) holds for other estimators than lasso-OLS, and Theorem 3.3 holds, in particular, for best subset selection in the Lagrangian formulation, which corresponds to $\operatorname{Pen}(\cdot) = \|\cdot\|_0$ in Example 3.4. Theorem 3.2 does, however, only partly extend to best subset selection. In this section we demonstrate that this may still provide a practically useful estimate of degrees of freedom.

The best subset selection estimator of $\mu$ with tuning parameter $\lambda > 0$, denoted by $\hat{\mu}^\lambda_{\mathrm{bs}}$, is

$$\hat{\mu}^\lambda_{\mathrm{bs}} = X\hat{\beta}^\lambda \quad \text{where } \hat{\beta}^\lambda = \arg\min_\beta \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_0.$$

It can be written on the form $\hat{\mu}^\lambda_{\mathrm{bs}} = \sum_{A\subseteq\{1,\dots,p\}} 1_{U^\lambda_A} \Pi_A$ (Lebesgue a.e.), where

$$U^\lambda_A := \left\{ y \in \mathbb{R}^n \;\middle|\; \lambda|A| - \frac{1}{2}\|\Pi_A y\|_2^2 < \min_{B\subseteq\{1,\dots,p\}\setminus A} \lambda|B| - \frac{1}{2}\|\Pi_B y\|_2^2 \right\}, \quad A \subseteq \{1,\dots,p\}. \tag{19}$$

It is straightforward to verify that $\hat{\mu}^\lambda_{\mathrm{bs}}$ fulfils Assumption 2.2 except 2.2(c), which follows by Lemma A.1 in the appendix. Hence Theorem 2.4 applies to $\hat{\mu}^\lambda_{\mathrm{bs}}$.

From (19) we note that the outer unit normal to $\partial U^\lambda_{A_1}$ on $\overline{U}^\lambda_{A_1} \cap \overline{U}^\lambda_{A_2}$ equals $(\Pi_{A_2} - \Pi_{A_1})y$ normalized to have norm 1. Theorem 2.4 yields

$$\begin{aligned}
\operatorname{df}(\hat{\mu}^\lambda_{\mathrm{bs}}) - \operatorname{df}_{\mathrm{S}}(\hat{\mu}^\lambda_{\mathrm{bs}}) &= \frac{1}{2}\sum_{A_1\neq A_2} \int_{\overline{U}^\lambda_{A_1}\cap\overline{U}^\lambda_{A_2}} \frac{\langle (\Pi_{A_2} - \Pi_{A_1})y, (\Pi_{A_2} - \Pi_{A_1})y\rangle}{\|(\Pi_{A_2} - \Pi_{A_1})y\|_2} \psi(y)\, d\mathcal{H}^{n-1}(y) \\
&= \frac{1}{2}\sum_{A_1\neq A_2} \int_{\overline{U}^\lambda_{A_1}\cap\overline{U}^\lambda_{A_2}} \|(\Pi_{A_2} - \Pi_{A_1})y\|_2 \psi(y)\, d\mathcal{H}^{n-1}(y),
\end{aligned}$$

which proves that $\mathrm{df} > \mathrm{df}_{\mathrm{S}}$ for best subsection selection. Moreover, Proposition 3.3 and Example 3.4 yields

$$-2\lambda\partial_\lambda \operatorname{df}_{\mathrm{S}}(\hat{\mu}^\lambda_{\mathrm{bs}}) = \frac{1}{2}\sum_{A_1\neq A_2} \int_{\overline{U}^\lambda_{A_1}\cap\overline{U}^\lambda_{A_2}} \psi(y) \frac{\langle y, (\Pi_{A_2} - \Pi_{A_1})y\rangle}{\|(\Pi_{A_2} - \Pi_{A_1})y\|_2} \big(|A_2| - |A_1|\big)\, d\mathcal{H}^{n-1}(y).$$

For $\operatorname{col}(X_{A_1}) \subseteq \operatorname{col}(X_{A_2})$ and $\operatorname{rank}(X_{A_2}) = \operatorname{rank}(X_{A_1}) + 1$, we see that the integrands in the two identities above coincide. Hence, if we define

$$\begin{aligned}
\mathcal{A}_1 &:= \big\{A_1, A_2 \subseteq \{1,\dots,p\} \mid \operatorname{col}(X_{A_1}) \subseteq \operatorname{col}(X_{A_2}) \text{ and } \operatorname{rank}(X_{A_2}) = \operatorname{rank}(X_{A_1}) + 1\big\} \quad \text{and} \\
\mathcal{A}_2 &:= \big\{A_1, A_2 \subseteq \{1,\dots,p\} \mid \operatorname{col}(X_{A_1}) \neq \operatorname{col}(X_{A_2}) \text{ and } (A_1, A_2) \notin \mathcal{A}_1, (A_2, A_1) \notin \mathcal{A}_1\big\},
\end{aligned}$$

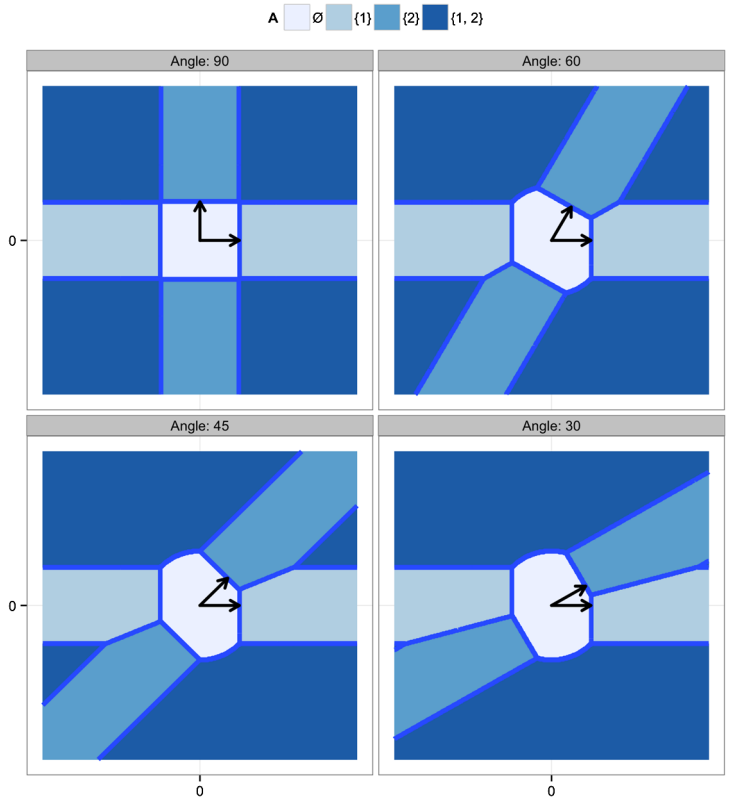


Fig. 7. Illustrations of the decomposition of $\mathbb{R}^2$ into the four sets $U^1_{\varnothing}$, $U^1_{\{1\}}$, $U^1_{\{2\}}$ and $U^1_{\{1,2\}}$ according to the best subset selection estimator in the Lagrangian formulation with $\lambda = 1$. The set $U^1_{\varnothing}$ consists of the points projected onto the 0-dimensional space $\{0\}$, the sets $U^1_{\{1\}}$, $U^1_{\{2\}}$ to the projections onto one of the two 1-dimensional subspaces and $U^1_{\{1,2\}}$ to the identity map. The decomposition depends on the angle between the two columns in $X$.

then

$$\mathrm{df}\big(\hat{\mu}^\lambda_{\mathrm{bs}}\big) - \mathrm{df}_{\mathrm{S}}\big(\hat{\mu}^\lambda_{\mathrm{bs}}\big) = -2\lambda\partial_\lambda \,\mathrm{df}_{\mathrm{S}}\big(\hat{\mu}^\lambda_{\mathrm{bs}}\big) + R,$$

where

$$R = \frac{1}{2}\sum_{(A_1,A_2)\in\mathcal{A}_2}\int_{\overline{U}^\lambda_{A_1}\cap\overline{U}^\lambda_{A_2}} \frac{\langle(\Pi_{A_2}-\Pi_{A_1})y, (\Pi_{A_2}-\Pi_{A_1}-(|A_2|-|A_1|)I_n)y\rangle}{\|(\Pi_{A_2}-\Pi_{A_1})y\|_2}\psi(y)\,d\mathcal{H}^{n-1}(y).$$

The usefulness of this hinges on $R$ being small. For $X$ orthogonal we have already demonstrated that $R = 0$ as $\hat{\mu}^\lambda_{\mathrm{bs}}$ then coincides with lasso-OLS, and in this case $\overline{U}^\lambda_{A_1} \cap \overline{U}^\lambda_{A_2}$ has Hausdorff measure zero for all $(A_1, A_2) \in \mathcal{A}_2$. For non-orthogonal $X$ this is no longer true, see Figure 7. For best subset selection there will generally be boundaries of non-zero Hausdorff measure between many more of the sets $\overline{U}^\lambda_A$ – boundaries that correspond to including or excluding more than one predictor at the time or replacing predictors. Compare this with lasso-OLS and Figure 1. However, by continuity in $X$ we have $R \to 0$ for $X$ tending to an orthogonal matrix, and we can expect $R$ to be small for matrices that are not too far from orthogonal matrices. Thus we expect

$$\mathrm{df}_{\mathrm{S}}\big(\hat{\mu}^\lambda_{\mathrm{bs}}\big) - 2\lambda\partial_\lambda \,\mathrm{df}_{\mathrm{S}}\big(\hat{\mu}^\lambda_{\mathrm{bs}}\big) \tag{20}$$

to be a useful approximation for $\mathrm{df}(\hat{\mu}^\lambda_{\mathrm{bs}})$ also for non-orthogonal $X$.

Using the same procedure for estimating the correction $-2\lambda\partial_\lambda \,\mathrm{df}_{\mathrm{S}}(\hat{\mu}^\lambda_{\mathrm{bs}})$ as outlined in Section 3 – using $2\widehat{\partial}$ instead of $\widehat{\partial}$ – we used simulations to investigate if (20) was actually a good approximation of $\mathrm{df}(\hat{\mu}^\lambda_{\mathrm{bs}})$. Figure 8 shows the

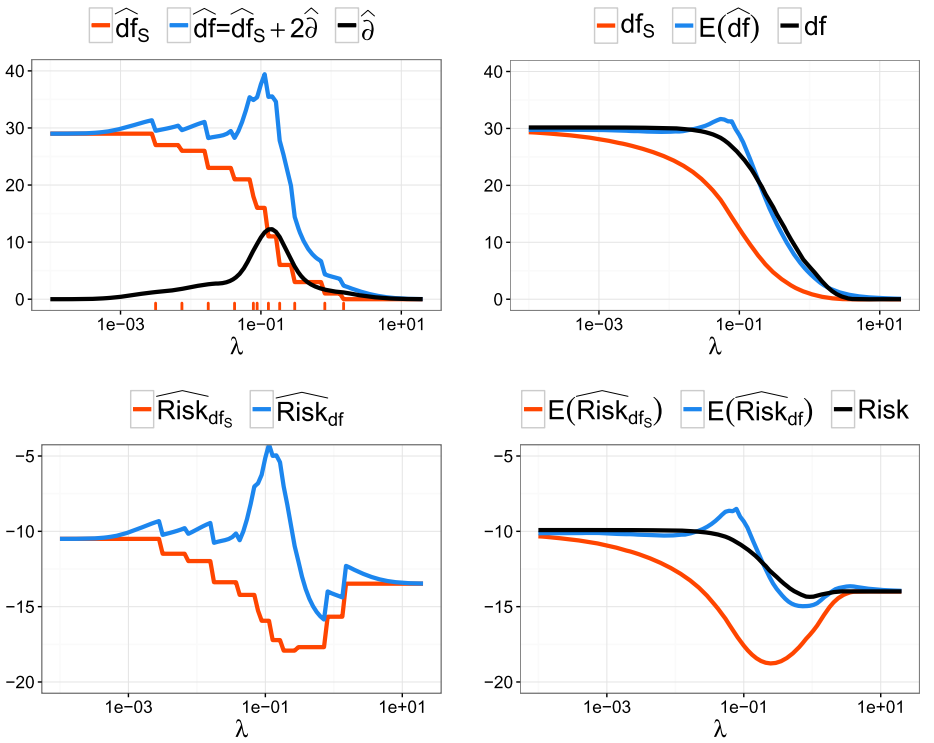


Fig. 8. Left: Realization of the estimates of degrees of freedom $\hat{\mathrm{df}}_S = \dim(\hat{S}^\lambda)$ and $\hat{\mathrm{df}} = \dim(\hat{S}^\lambda) + 2\widehat{\partial}$ as well as the correction term $\widehat{\partial}$ as a function of $\log(\lambda)$ for best subset selection (top) and corresponding estimates of the risk (bottom). Right: Similar to the left but mean values of the estimates obtained by averaging over 1000 samples along with the degrees of freedom $\mathrm{df} = \mathrm{df}(\hat{\mu}^\lambda_{\mathrm{bs}})$ obtained from the 1000 samples using the covariance definition (1). The design parameters were: $\sigma = 0.5$, $n = p = 30$, $\gamma = 1$, $\alpha = 0.1$ and the design type was (S) with constant correlation of $\rho = 0.1$ (see Section 4).

results using the same configurations as in Figure 2, except that $n$ was lowered to 30 due to computational constraints. The conclusion from this and other similar simulations (not shown) is that even with non-orthogonal designs, (20) is a practically useful approximation. That is, $-2\lambda\partial_\lambda \mathrm{df}_S(\hat{\mu}^\lambda_{\mathrm{bs}})$ accounts for the majority of the increase in the degrees of freedom due to variable selection.

## 6. Discussion

We have provided a new representation of degrees of freedom for a broad class of discontinuous, piecewise Lipschitz estimators. This representation provides us with a deeper insight into the effect of variable selection, among other things, on the effective dimension of the statistical model and the estimator used. We have demonstrated that for lasso-OLS it was, moreover, possible to derive a practically useful estimator of the degrees of freedom based on the general representation, and we also suggest that a similar estimator can be useful for best subset selection. The estimator was based on relating the derivative of $\lambda \mapsto \mathrm{df}_S(\hat{\mu}^\lambda)$ to the discontinuities of the estimator $\hat{\mu}^\lambda$ as expressed via the integral representation of $\mathrm{df}(\hat{\mu}^\lambda) - \mathrm{df}_S(\hat{\mu}^\lambda)$. This does, indeed, make some intuitive sense as the first expresses the mean jump of degrees of freedom per unit change of $\lambda$ and the other (in some sense) the mean discontinuity of degrees of freedom per unit change of $y$. Changing $\lambda$ for fixed $y$ or changing $y$ for fixed $\lambda$ are dual operations, and it is not surprising that we can relate the numbers.

A simulation study demonstrated that the risk of the lasso-OLS estimator can be estimated effectively by using our proposed estimate of degrees of freedom. Our proposal did not incur any substantial computational penalty, nor did it incur a substantial increase in the variance of the risk estimate. The simulation study also showed that lasso-OLS can be effectively tuned by minimising our proposed risk estimate, and that the resulting computations are faster

than using cross-validation. The resulting lasso-OLS estimator selects fewer predictors than lasso with a comparable predictive performance, but it is computationally more expensive.

If we were to generalize our results to other estimators that include a tuning parameter, we expect that it is only the derivative of the part of $\mathrm{df}_S(\hat{\mu}^\lambda)$ that corresponds to jumps that can be related to $\mathrm{df}(\hat{\mu}^\lambda) - \mathrm{df}_S(\hat{\mu}^\lambda)$. That is, in general, $\lambda \mapsto \mathrm{div}(\hat{\mu}^\lambda)$ will have jumps as well as smooth but non-constant pieces, and it is only the expectation of the jump part that we expect can be related to $\mathrm{df}(\hat{\mu}^\lambda) - \mathrm{df}_S(\hat{\mu}^\lambda)$. We believe that our suggested estimator of degrees of freedom may actually be generalizable to a number of discontinuous estimators involving variable selection as well as shrinkage. The requirement will be that the estimator has one or more tuning parameters and that it is computed on a grid or along a path of these. Then we can potentially estimate the derivative of the divergence of the estimator as a function of the tuning parameter(s). It is an ongoing research project to investigate this in detail.

For best subset selection we did not provide any bounds on the residual $R$ in the approximation of $\mathrm{df}(\hat{\mu}^\lambda) - \mathrm{df}_S(\hat{\mu}^\lambda)$. It would, indeed, be very interesting to investigate this approximation in more detail. It would, in particular, be interesting to understand if it in any way can be seen as a "first order approximation" and whether there are higher order terms worth including in some cases.

Finally, we have restricted attention to Gaussian noise in the theoretical derivations. Like Stein's classical lemma, Theorem 2.4 crucially relies on this assumption. Our simulation study demonstrated some robustness towards deviations from this assumption. However, extensions of Stein's lemma to non-Gaussian distributions do exist (see, e.g., [3]), but further investigations are required to determine if similar extensions can be made in the more general framework presented in this paper.

## 7. Supplementary material

The results from the entire simulation study as well as the R-code are available online http://doi.org/10.5281/zenodo.321847, [19].

## Appendix: Additional results and proofs

### A.1. *Semialgebraic sets*

Observe that for $A$ and $B$ subsets of $\mathbb{R}^n$ it holds that

$$\begin{aligned} &\partial A = \partial\left(A^c\right), \\ &\partial(A \cup B) \subseteq \partial A \cup \partial B, \\ &\partial(A \cap B) \subseteq \partial A \cup \partial B. \end{aligned} \tag{21}$$

Especially, the family of sets

$$\left\{E \in \mathcal{B}\left(\mathbb{R}^n\right) \mid r \mapsto \mathcal{H}^{n-1}\left(\partial E \cap B(0,r)\right) \text{ is polynomially bounded}\right\} \tag{22}$$

is stable under complement, finite union and finite intersection. This is a useful observation when we want to verify Assumption 2.2(c).

The following Lemma shows that *semialgebraic sets* belong to the family given by (22). A semialgebraic set is a finite union of finite intersections of sets of the form $(P = 0)$ and $(Q > 0)$, where $P$ and $Q$ are polynomials. A multivariate polynomial is of the form (using multi-index notation)

$$P(x) = \sum_{\alpha \in A} a_\alpha x^\alpha, \quad a_\alpha \in \mathbb{R} \text{ for each } \alpha \in A,$$

with $A \subseteq \mathbb{N}^n$ finite.

**Lemma A.1.** *If $E$ is semialgebraic then $r \mapsto \mathcal{H}^{n-1}(\partial E \cap B(0,r))$ is polynomially bounded.*

**Proof.** By the stability under finite set operations of the family given by (22) it suffices to show that $r \mapsto \mathcal{H}^{n-1}((P = 0) \cap B(0, r))$ is polynomially bounded for any nonzero polynomial $P$. But this follows from Corollary 1 in [16], which implies that

$$\mathcal{H}^{n-1}\big((P = 0) \cap B(0, r)\big) \leq \frac{\deg(P)\pi^{\frac{n+1}{2}}}{\Gamma(\frac{n}{2})} r^{n-1}$$

for any nonzero polynomial $P$ with $\deg(P) = \max_{a_\alpha \neq 0} |\alpha|$ denoting the degree of $P$. □

*A.2. Proof of Theorem 2.4*

The following Lemma characterizes the outer unit normal vectors $\eta_i$ for $i = 1, \ldots, N$.

**Lemma A.2.** *Under Assumption 2.2 the following holds*:

(a) $\eta_i = 0$ $\mathcal{H}^{n-1}$ *a.e. on* $\partial U_i \setminus \bigcup_{j \neq i} \overline{U}_j$ *for each* $i = 1, \ldots, N$.
(b) $\eta_i = -\eta_j$ $\mathcal{H}^{n-1}$ *a.e. on* $\partial U_i \cap \partial U_j$ *with* $i \neq j$.
(c) $\eta_i = 0$ $\mathcal{H}^{n-1}$ *a.e. on* $\partial U_i \cap \partial U_j \cap \partial U_k$ *with* $i, j, k$ *distinct.*

**Proof.** Firstly, note that the unit outer normal $\eta_i$ on $\partial U_i$ vanishes outside *the measure theoretic boundary* $\partial_* U_i$, see Definition 5.8 in [7]. Moreover, these two types of boundaries relates to *the reduced boundary* $\partial^* U_i$ (see Definition 5.7 in [7]) by the inclusions:

$$\partial^* U_i \subseteq \partial_* U_i \subseteq \partial U_i.$$

Furthermore, $\mathcal{H}^{n-1}(\partial_* U_i \setminus \partial^* U_i) = 0$ (see Lemma 5.8.1 in [7]). All in all, we see that the lemma holds if we can show the following claims:

$$\begin{aligned} &\partial^* U_i \subseteq \bigcup_{l \neq i} \overline{U}_l, \\ &\eta_i = -\eta_j \quad \text{on } \partial^* U_i \cap \partial^* U_j, \\ &\partial^* U_i \cap \partial^* U_j \cap \partial^* U_k = \varnothing \end{aligned} \tag{23}$$

holds for all $i, j, k$ distinct.

To prove the claims, define for each $i$ and $r > 0$ the sets

$$\begin{aligned} U_i^r(x) &= \big\{y \mid r(y - x) + x \in U_i\big\}, \\ H_i(x) &= \big\{y \mid \langle \eta_i, y - x\rangle \leq 0\big\}. \end{aligned}$$

Note that $\{U_i^r(x)\}_i$ are still disjoint. By Theorem 5.7.1 in [7]

$$1_{U_i^r(x)} \xrightarrow{r \to 0} 1_{H_i(x)} \quad \text{in } L^1_{\text{loc}}\big(\mathbb{R}^n\big) \text{ for all } x \in \partial^* U_i.$$

Therefore, if there existed $x \in \partial^* U_i \cap \partial^* U_j \cap \partial^* U_k$ for $i, j, k$ distinct, then

$$1_{U_i^r(x) \cup U_j^r(x) \cup U_k^r(x)} \xrightarrow{r \to 0} 1_{H_i(x)} + 1_{H_j(x)} + 1_{H_k(x)} \quad \text{in } L^1_{\text{loc}}\big(\mathbb{R}^n\big), \tag{24}$$

which is impossible as the right hand side is not Lebesgue a.e. an indicator. By the same argument one can deduce that $\eta_i = -\eta_j$ must hold for $x \in \partial^* U_i \cap \partial^* U_j$ and that any $x \in \partial^* U_i$ cannot belong to the open set $(\bigcup_{l \neq i} \overline{U}_l)^c$. □

**Proof of Theorem 2.4.** For $i = 1, \ldots, N$ Gauss–Green's formula (see Theorem 5.8.1 in [7] and Theorem 4.5.6 in [9]) gives that

$$\int_{U_i} \operatorname{div}(f)\, dm = \int_{\partial U_i} \langle f, \eta_i \rangle\, d\mathcal{H}^{n-1} \tag{25}$$

for all Lipschitz continuous vector fields $f$ with compact support. Here $\eta_i$ denotes the outer unit normal of $\partial U_i$, which is well defined and nonzero on a subset of $\partial U_i$ and zero everywhere else by definition.

Let $(g_r)_r$ be a sequence of smooth functions with

$$g_r(x) = \begin{cases} 1 & \text{if } x \in B(0, r), \\ 0 & \text{if } x \notin B(0, r+1), \end{cases}$$

and $(g_r)_r$ and $(Dg_r)_r$ uniformly bounded. Since $\hat{\mu}_i$ is Lipschitz continuous on $\overline{U}_i \cap B(0, r+1)$ Kirzbraun's theorem ensures that $\hat{\mu}_i$ has a Lipschitz extension, $\hat{\mu}_i^r : \mathbb{R}^n \to \mathbb{R}^n$. Then $f_r = g_r \psi \hat{\mu}_i^r$ is Lipschitz continuous with compact support and $g_r \hat{\mu}_i^r = g_r \hat{\mu}$ on $U_i$. Then (25) applied to $f_r$ yields

$$\int_{\partial U_i} g_r \psi \langle \hat{\mu}_i, \eta_i \rangle\, d\mathcal{H}^{n-1} = \int_{U_i} g_r \psi \operatorname{div}(\hat{\mu}_i)\, dm + \int_{U_i} \langle g_r D\psi + \psi\, Dg_r, \hat{\mu}_i \rangle\, dm.$$

Due to Assumption 2.2 all integrands above are dominated by integrable functions, and by letting $r \to \infty$ Lebesgue's Dominated Convergence Theorem yields

$$\int_{\partial U_i} \psi \langle \hat{\mu}_i, \eta_i \rangle\, d\mathcal{H}^{n-1} = \int_{U_i} \psi \operatorname{div}(\hat{\mu}_i)\, dm + \int_{U_i} \langle D\psi, \hat{\mu}_i \rangle\, dm.$$

By summing over $i$ we get

$$\operatorname{df}(\hat{\mu}) = \operatorname{df}_S(\hat{\mu}) - \sum_i \int_{\partial U_i} \psi \langle \hat{\mu}_i, \eta_i \rangle\, d\mathcal{H}^{n-1}. \tag{26}$$

By Lemma A.2 we see that

$$\begin{aligned} \operatorname{df}(\hat{\mu}) &= \operatorname{df}_S(\hat{\mu}) - \sum_{j \neq i} \int_{\partial U_i \cap \partial U_j} \psi \langle \hat{\mu}_i, \eta_i \rangle\, d\mathcal{H}^{n-1} \\ &= \operatorname{df}_S(\hat{\mu}) + \frac{1}{2} \sum_{j \neq i} \int_{\partial U_i \cap \partial U_j} \langle \hat{\mu}_j - \hat{\mu}_i, \eta_i \rangle \psi\, d\mathcal{H}^{n-1}. \end{aligned}$$

Since $\eta_i$ vanishes on $\partial U_i \cap \partial U_j \setminus (\overline{U}_i \cap \overline{U}_j)$ for $i \neq j$ we have proven (4). □

## References


[1] L. Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.* **87** (419) (1992) 738–754. MR1185196

[2] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg, 2011. MR2807761

[3] A. Dalalyan and A. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* **72** (2008) 39–61.

[4] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** (432) (1995) 1200–1224. MR1379464

[5] B. Efron. The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* **99** (467) (2004) 619–632. MR2090899

[6] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. Least angle regression. *Ann. Statist.* **32** (2) (2004) 407–499. With discussion, and a rejoinder by the authors. MR2060166

[7] L. Evans and R. Gariepy. *Measure Theory and Fine Properties of Functions*. *Studies in Advanced Mathematics*. Taylor & Francis, London, 1992. MR1158660

[8] J. Fan, L. Xue and H. Zou. Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** (3) (2014) 819–849. MR3210988

[9] H. Federer. *Geometric Measure Theory*. *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1969. MR0257325

[10] J. Friedman, T. Hastie and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** (1) (2010) 1–22.

[11] G. Givens and J. Hoeting. *Computational Statistics*. *Wiley Series in Computational Statistics*. John Wiley & Sons, Hoboken, 2012. MR3236433

[12] N. R. Hansen and A. Sokol. Degrees of freedom for nonlinear least squares estimation, 2014. Available at http://arxiv.org/abs/1402.2997.

[13] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. *Monographs on Statistics and Applied Probability*. **43**. Chapman and Hall Ltd., London, 1990. MR1082147

[14] K. Kato. On the degrees of freedom in shrinkage estimation. *J. Multivariate Anal.* **100** (7) (2009) 1338–1352. MR2514133

[15] J. D. Lee, D. L. Sun, Y. Sun and J. E. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** (3) (2016) 907–927. MR3485948

[16] T. Loi and P. Phien. Bounds of Hausdorff measures of tame sets. *Acta Math. Vietnam.* **39** (4) (2014) 637–647. MR3292588

[17] N. Meinshausen. Relaxed lasso. *Comput. Statist. Data Anal.* **52** (1) (2007) 374–393. MR2409990

[18] M. Meyer and M. Woodroofe. On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** (4) (2000) 1083–1104. MR1810920

[19] F. Mikkelsen and N. Hansen. Supplementary material for "Degrees of freedom for piecewise Lipschitz estimators", 2017. Available at http://doi.org/10.5281/zenodo.321847.

[20] S. Reid, R. Tibshirani and J. Friedman. A study of error variance estimation in lasso regression. *Statist. Sinica* **26** (1) (2016) 35–67. MR3468344

[21] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** (6) (1981) 1135–1151. MR0630098

[22] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., B* **58** (1) (1996) 267–288. MR1379242

[23] R. J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Stat.* **7** (2013) 1456–1490. MR3066375

[24] R. J. Tibshirani. Degrees of freedom and model search. *Statist. Sinica* **25** (3) (2015) 1265–1296. MR3410308

[25] R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *Ann. Statist.* **40** (2) (2012) 1198–1232. MR2985948

[26] M. Vincent, K. Perell, F. Nielsen, G. Daugaard and N. Hansen. Modeling tissue contamination to improve molecular identification of the primary tumor site of metastases. *Bioinformatics* **30** (10) (2014) 1417–1423.

[27] J. Ye. On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** (441) (1998) 120–131. MR1614596

[28] H. Zou, T. Hastie and R. Tibshirani. On the degrees of freedom of the lasso. *Ann. Statist.* **35** (5) (2007) 2173–2192. MR2363967