We give a general result on the effective degrees of freedom for nonlinear least squares estimation, which relates the degrees of freedom to the divergence of the estimator. The result implies that Stein’s unbiased risk estimate (SURE) is biased if the least squares estimator is not sufficiently differentiable, and it gives an exact representation of the bias. In the light of the general result we treat $\ell_1$-constrained nonlinear least squares estimation, and present an application to model selection for systems of linear ODE models.

1. Introduction. The ability to estimate, or bound, the risk of an estimator is a central statistical problem. Such risk estimates, obtained e.g. by cross-validation, are in widespread use for model selection – in particular for high-dimensional statistical modeling. Less computationally demanding alternatives to refitting methods, based on theoretical considerations, are abundant. We mention the information criteria (AIC, BIC, TIC, etc.), generalized cross-validation (GCV), and Stein’s unbiased risk estimate (SURE), see e.g. Burnham and Anderson (2002), Claeskens and Hjort (2008) and Efron (2004). SURE assumes a Gaussian error model, but is particularly interesting in the high-dimensional case, because it is based on non-asymptotic arguments and applies to nonlinear estimators, see Efron (2004) for a thorough treatment of SURE.

For the computation of SURE it is necessary to compute an estimate of the effective degrees of freedom for the estimator. To this end, sufficient differentiability of the estimator is required, and this was, for instance, established in Meyer and Woodroofe (2000) in the context of shape restricted regression. In Meyer and Woodroofe (2000) the mean in a multivariate Gaussian model is estimated by projection onto a closed convex set. The related case of $\ell_1$-penalized estimators has been studied thoroughly, and it has been established in generality that for $\ell_1$-penalized linear least squares regression, the number of nonzero parameters is an unbiased estimate of the effective degrees of freedom, see Efron et al. (2004), Zou, Hastie and Tibshirani (2007) and Tibshirani and Taylor (2012).

We have a particular interest in using $\ell_1$-regularized estimators in the context of high-dimensional dynamic models, e.g. estimation of parameters
in a $d$-dimensional ODE. This will often amount to $\ell_1$-constrained or $\ell_1$-penalized nonlinear least squares estimation. To apply SURE in this case, several questions arise. First, is the estimator sufficiently differentiable, and if not, what are the consequences? Second, how do we in practice compute the effective degrees of freedom or an estimate thereof? The present paper provides some answers to these questions.

We consider the setup where $Y \sim \mathcal{N}(\xi, \sigma^2 I_n)$, $K \subseteq \mathbb{R}^n$ is a nonempty closed set,

$$\text{pr}(y) \in \arg \min_{x \in K} \|y - x\|^2_2$$

is a point that minimizes the Euclidean distance from $y$ to $K$ and $\text{pr}(Y)$ is the estimator of $\xi$. The map $\text{pr} : \mathbb{R}^n \to \mathbb{R}^n$ is known as the metric projection onto $K$. Though it may not be uniquely defined everywhere, it is, in fact, Lebesgue almost everywhere unique. For the purpose of this introduction we assume that a (Borel measurable) selection has been made on the Lebesgue null set where the metric projection is not unique. With

$$\text{Risk} = E\|\xi - \text{pr}(Y)\|_2^2$$

denoting the risk of the estimator, it is well known that

(1) $$\text{Risk} = E\|Y - \text{pr}(Y)\|_2^2 - n\sigma^2 + 2\sigma^2 df$$

where

$$df = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{cov}(Y_i, \text{pr}_i(Y)).$$

See e.g. Tibshirani and Taylor (2012), Efron (2004) and Ye (1998).

It turns out that the metric projection is Lebesgue almost everywhere differentiable, see Section 2, and we can therefore introduce the Stein degrees of freedom as

$$df_S = E(\nabla \cdot \text{pr}(Y))$$

with $\nabla \cdot \text{pr} = \sum_{i=1}^{n} \partial_i \text{pr}_i$ denoting the divergence of $\text{pr}$. It follows from Lemma 2 (Stein’s lemma) in Stein (1981) that if $\text{pr}$ is almost differentiable then

$$df = df_S.$$

However, differentiability Lebesgue almost everywhere does not imply almost differentiability, and our main result, Theorem 2 in Section 2, gives that in general

$$df - df_S \geq 0.$$
This implies the lower bound

\[ \text{Risk} \geq E||Y - \text{pr}(Y)||^2 - n\sigma^2 + 2\sigma^2\text{df}_S. \]

Theorem 2 gives a characterization of \( \text{df} - \text{df}_S \), whose size is closely related to the distance from \( \xi \) to points where the metric projection is non-differentiable, and the “magnitude” of the non-differentiability – see also the discussion in Section 6. The Stein unbiased risk estimate (SURE)

\[ \hat{\text{Risk}} = ||Y - \text{pr}(Y)||^2 - n\sigma^2 + 2\sigma^2\nabla \cdot \text{pr}(Y) \]

can thus be biased in general – systematically underestimating the true risk. In Section 4 we study an example where the bias turns out to be undetectable, and where SURE still works well for risk estimation and model selection. In such cases the challenge is the actual computation of the divergence \( \nabla \cdot \text{pr} \).

To give results on the computation of \( \nabla \cdot \text{pr} \) we consider the case where \( K \) is a parametrized set, that is, there is a map \( \zeta : \mathbb{R}^p \rightarrow \mathbb{R}^n \), and

\[ K = \zeta(\Theta) \]

for \( \Theta \subseteq \mathbb{R}^p \) a closed set. This setup includes most linear and nonlinear regression models. Moreover, by taking parameter sets of the form

\[ \Theta = \{\beta \in \mathbb{R}^p \mid J(\beta) \leq s\} \]

for \( s \geq 0 \) and some function \( J : \mathbb{R}^p \rightarrow [0, \infty) \), the setup includes many regularization methods, see Figure 1. In Section 3 we give two results on the computation of \( \nabla \cdot \text{pr} \). First, if \( \text{pr}(Y) = \zeta(\hat{\beta}) \) for \( \hat{\beta} \in \Theta^o \) the divergence is computable if \( \zeta \) is a local \( C^2 \) diffeomorphism around \( \hat{\beta} \), and we provide an explicit formula. Similar results can be found in the mathematical literature.
on differentiability of metric projections. We show that the resulting estimate of degrees of freedom in this case coincides with the plug-in estimate of the effective number of parameters used in Takeuchi’s information criterion (TIC). Second, we present a result for the case where

$$\Theta = \left\{ \beta \in \mathbb{R}^p \left| \sum_{k=1}^{p} \omega_k |\beta_k| \leq s \right. \right\},$$

which gives an estimate of the degrees of freedom for nonlinear least squares regression with a weighted $\ell_1$-constraint.

The results are illustrated in Section 4 by an application to model selection for dynamical systems modeled using linear ODEs. In this example, the mean is given in terms of matrix exponentiation. We consider the $\ell_1$-constrained least squares estimator of the parameter matrix in a $d$-dimensional linear ODE using SURE, as given by (3), for model selection.

2. Degrees of freedom for the metric projection. In this section we present the main general results on differentiability of the metric projection, and how the divergence is related to the degrees of freedom. This gives a characterization of the bias of SURE in cases where the metric projection does not satisfy a sufficiently strong differentiability condition. The proofs are given in Section 5.

**Definition 1.** With $D \subseteq \mathbb{R}^n$ we say that a function $f : D \to \mathbb{R}^n$ is differentiable in $y \in D$ in the extended sense if there is a neighborhood $N$ of $y$ such that $D^c \cap N$ is a Lebesgue null set and

$$f(x) = f(y) + A(x - y) + o(||x - y||_2)$$

for $x \in D \cap N$ and a matrix $A$.

If $f$ is differentiable in $y$ in the extended sense the matrix $A$, depending on $y$, is necessarily unique by denseness of $D \cap N$ in $N$. We define the partial derivatives – and thus the divergence – of $f$ in $y$ in terms of $A$ by

$$\partial_j f_i(y) = A_{ij}$$

for $i, j = 1, \ldots, n$. Note that the partial derivatives of $f$ in $y$ need not exist in the classical sense if $f$ is differentiable in $y$ in the extended sense, but if they do, they coincide with $A_{ij}$. 
Theorem 1. There exists a Borel measurable choice of the metric projection as a map \( pr: \mathbb{R}^n \rightarrow \mathbb{R}^n \) with the property that
\[
pr(y) \in \arg\min_{x \in K} \|y - x\|_2^2
\]
for all \( y \in \mathbb{R}^n \). Moreover, \( pr(y) \) is uniquely defined and differentiable in the extended sense for Lebesgue almost all \( y \) with \( \partial_i pr(y) \geq 0 \) for \( i = 1, \ldots, n \).

We call the set of points with a non-unique metric projection onto \( K \) the exoskeleton of \( K \), following the terminology in Hug, Last and Weil (2004), and we write
\[
exo(K) = \left\{ y \in \mathbb{R}^n \left| \arg\min_{x \in K} \|y - x\|_2 \text{ is not a singleton} \right. \right\}.
\]
This set is also called the skeleton of the open set \( K^c \) in Fremlin (1997). Theorem 1 implies that \( \exo(K) \) is a Lebesgue null set, but more is known. Theorem 1G in Fremlin (1997) gives, for instance, that \( \exo(K) \) has Hausdorff dimension at most \( n - 1 \). It should be noted that there can be points in \( K \setminus \exo(K) \) where \( pr \) is not differentiable.

As a consequence of Theorem 1, \( pr(Y) \) is uniquely defined with probability 1, and it follows from the triangle inequality that
\[
\|pr(Y)\|_2 \leq \|pr(0)\|_2 + 2\|Y\|_2.
\]
This shows, in particular, that \( pr(Y) \) has finite second moment. Moreover, Theorem 1 gives that the divergence \( \nabla \cdot pr(Y) \) is well defined and positive with probability 1. These considerations ensure that the following definition is meaningful.

Definition 2. The degrees of freedom for the metric projection is defined as
\[
df = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{cov}(Y_i, pr_i(Y)),
\]
and the Stein degrees of freedom is defined as
\[
df_S = E(\nabla \cdot pr(Y)).
\]

Our next result gives the general relation between \( df \) and \( df_S \). To this end, let
\[
\psi(y; \xi, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\|y - \xi\|_2^2}{2\sigma^2}}
\]
denote the density for the distribution of \( Y - \) the multivariate normal distribution with mean vector \( \xi \) and covariance matrix \( \sigma^2 I_n \).
Theorem 2. There exists a Radon measure $\nu$, singular w.r.t. the Lebesgue measure, such that

$$
(6) \quad df = df_S + \int_{\mathbb{R}^n} \psi(y; \xi, \sigma^2) \nu(dy).
$$

The proof of Theorem 2 is a computation of the distributional partial derivatives of $pr_i$, and subsequently an extension of the partial integration formula from test functions to a function class that includes $\psi$. As the proof reveals, the distributional partial derivative of $pr_i$ in the $i$'th direction is represented by a positive measure with Lebesgue decomposition

$$
\partial_i pr_i \cdot m_n + \nu_i,
$$

where $m_n$ denotes the Lebesgue measure on $\mathbb{R}^n$ and $\nu_i \perp m_n$. The measure $\nu$ that appears in Theorem 2 is given as $\nu = \sum_{i=1}^n \nu_i$. Note that $\nu$ depends only on the closed set $K$, and is, in particular, independent of $\xi$ and $\sigma^2$.

As a direct consequence of Theorem 2 we get the following result on the bias of the risk estimator given by (3). It implies that $\hat{\text{Risk}}$ is unbiased if and only if the measure $\nu$ is the null measure.

Corollary 1. With

$$
\hat{\text{Risk}} = ||Y - pr(Y)||^2_2 - n\sigma^2 + 2\sigma^2 \nabla \cdot pr(Y)
$$

it holds that

$$
E(\hat{\text{Risk}}) = \text{Risk} - 2\sigma^2 \int_{\mathbb{R}^n} \psi(y; \xi, \sigma^2) \nu(dy) \leq \text{Risk}.
$$

It is useful to be able to bound the support of the singular measure $\nu$. To this end we give the following proposition.

Proposition 1. If

$$
pr : \mathbb{R}^n \setminus \text{exo}(K) \to K
$$

is locally Lipschitz, and in particular if it is $C^1$, then $\text{supp}(\nu) \subseteq \text{exo}(K)$.

If $K$ is convex (in addition to being nonempty and closed) the metric projection is uniquely defined everywhere and Lipschitz continuous, see Lemma 1 in Tibshirani and Taylor (2012). Thus $\text{exo}(K) = \emptyset$ and by Proposition 1 the measure $\nu$ is the null measure. From this we get the unbiasedness of SURE for convex $K$. 

imsart-aos ver. 2014/01/08 file: nonLineardf.tex date: February 14, 2014
Fig 2. The set $K$ from Example 1 is the union of the coordinate axes (red). The metric projection is the projection onto the closest coordinate axis. The exoskeleton of $K$ (blue) is the set of points $y = (y_1, y_2) \neq (0,0)$ with either $y_1 = y_2$ or $y_1 = -y_2$. The closure of the exoskeleton equals in this example the support of the singular measure $\nu$.

**Corollary 2.** The measure $\nu$ in Theorem 2 is the null measure if $K$ is convex, in which case the risk estimator $\hat{\text{Risk}}$ is unbiased.

We provide three simple examples to illustrate the general results. The first is the projection onto the union of two orthogonal one-dimensional subspaces, which amounts to best subset selection. The second is the projection onto a convex $\ell_2$-ball, which amounts to a form of $\ell_2$-shrinkage. In the last example we consider the projection onto the $\ell_2$-sphere, which shows some interesting phenomena in the non-convex case. Example 3 shows, in particular, that $K$ need not be convex for $\nu$ to be the null measure, and thus that the support of $\nu$ can be a strict subset of $\text{exo}(K)$.

**Example 1.** We consider the case $n = 2$, $\xi = 0$, $\sigma^2 = 1$ and

$$K = \{(y_1, y_2) \in \mathbb{R}^2 \mid y_2 = 0\} \cup \{(y_1, y_2) \in \mathbb{R}^2 \mid y_1 = 0\}$$

is the union of the two orthogonal subspaces formed by the first and second coordinate axis, respectively. If we introduce the sets

$$I(z) = (-\infty, -|z|) \cup (|z|, \infty)$$

for $z \in \mathbb{R}$, we can for $y_1 \neq y_2$ write the metric projection as

$$\text{pr}(y_1, y_2) = (y_1 I(y_2)(y_1), y_2 I(y_1)(y_2)).$$
When \( y_1 \neq y_2 \) we find that

\[
\partial_1 \text{pr}_1(y) + \partial_2 \text{pr}_2(y) = 1_{I(y_2)}(y_1) + 1_{I(y_1)}(y_2) = 1,
\]

and \( \text{df}_S = 1 \). To compute the singular measure \( \nu \) we find, using Fubini’s Theorem and standard partial integration, that for \( \varphi \in C^1_c(\mathbb{R}^2) \),

\[
\int_{\mathbb{R}^2} \text{pr}_1(y) \partial_1 \varphi(y) \, dm_2(y) = \int_{\mathbb{R}} \int_{I(y_2)} y_1 \partial_1 \varphi(y_1, y_2) \, dy_1 \, dy_2
\]

\[
= - \int_{\mathbb{R}} |y_2| (\varphi(-|y_2|, y_2) + \varphi(|y_2|, y_2)) \, dy_2
\]

\[
- \int_{\mathbb{R}} \int_{I(y_2)} \varphi(y_1, y_2) \, dy_1 \, dy_2.
\]

This shows that the singular part of the distributional partial derivative of \( \text{pr}_1(y) \) w.r.t. \( y_1 \) is the measure \( \nu_1 \) determined by

\[
\int_{\mathbb{R}^2} \varphi(y) \, \nu_1(dy) = \int_{\mathbb{R}} |z| (\varphi(|z|, z) + \varphi(-|z|, z)) \, dz.
\]

The singular measure \( \nu_2 \) is determined likewise, and \( \nu = \nu_1 + \nu_2 \) is given by

\[
\int_{\mathbb{R}^2} \varphi(y) \, \nu(dy) = \int_{\mathbb{R}} |z| (\varphi(|z|, z) + \varphi(-|z|, z) + \varphi(z, |z|) + \varphi(z, -|z|)) \, dz.
\]

By choosing positive functions \( \varphi_n \in C^1_c(\mathbb{R}^2) \) such that \( \varphi_n(x) \nearrow \psi(x; 0, 1) \) for \( n \to \infty \), it follows that

\[
\int_{\mathbb{R}^2} \psi(y; 0, 1) \, d\nu(y) = \frac{2}{\pi} \int_{\mathbb{R}} |r| e^{-r^2} \, dr = \frac{2}{\pi} \int_0^\infty e^{-r} \, dr = \frac{2}{\pi}.
\]

We find that the degrees of freedom for the selection among the two one-dimensional orthogonal projections becomes

\[
\text{df} = 1 + \frac{2}{\pi} = 1.6366.
\]

In this particular case it follows directly from the covariance definition (4) that

\[
\text{df} = E(\max\{X_1, X_2\})
\]

where \( X_1 \) and \( X_2 \) are independent \( \chi^2_1 \)-distributed random variables. This concurs with findings in Ye (1998) on generalized degrees of freedom. The numerical value could in this case also be computed by computing the density of \( \max\{X_1, X_2\} \), and use this to compute the expectation \( E(\max\{X_1, X_2\}) \).
Example 2. Let $K = B(0, s)$ be the closed $\ell_2$-ball with center 0 and radius $s \geq 0$. Then
\[
\text{pr}_i(y) = \begin{cases} 
\frac{sy_i}{\|y\|_2} & \text{if } \|y\|_2 > s \\
\frac{sy_i^2}{\|y\|_2} & \text{if } \|y\|_2 \leq s
\end{cases}
\]
and
\[
\partial_i \text{pr}_i(y) = \begin{cases} 
\frac{s}{\|y\|_2} - \frac{sy_i^2}{\|y\|_2^3} & \text{if } \|y\|_2 > s \\
1 & \text{if } \|y\|_2 \leq s.
\end{cases}
\]
Since $K$ is convex
\[
df = df_S = s(n - 1)E(\|Y\|_2^{-1}1(\|Y\|_2 > s)) + nP(\|Y\|_2 \leq s).
\]
If $\xi = 0$ the expectation and probability can be expressed in terms of incomplete $\Gamma$-integrals. The unbiased estimate of $df$ is
\[
\nabla \cdot \text{pr}(Y) = \frac{s(n - 1)}{\|Y\|_2}1(\|Y\|_2 > s) + n1(\|Y\|_2 \leq s).
\]
It is interesting to compare the constrained estimator, which for fixed $s$ projects $Y$ onto the ball of radius $s$, with the linear shrinkage estimator
\[
\frac{1}{1 + \lambda}Y
\]
for a fixed $\lambda \geq 0$. The linear shrinkage estimator coincides with the metric projection onto the ball with radius
\[
s = \|Y\|_2/(1 + \lambda) \leq \|Y\|_2.
\]
It follows directly from (4) that the linear shrinkage estimator has degrees of freedom $n/(1 + \lambda)$. For the metric projection onto a ball with radius $s$ given by (7) the unbiased estimate of the degrees of freedom equals
\[
\frac{s(n - 1)}{\|Y\|_2} = \frac{n - 1}{1 + \lambda}.
\]
This is an unbiased estimate of degrees of freedom for a ball with fixed radius $s \geq 0$. The degrees of freedom for the linear shrinkage estimator is for fixed $\lambda \geq 0$. The two estimates of degrees of freedom differ because the relation $s(1 + \lambda) = \|Y\|_2$ is $Y$-dependent.
Example 3. In this example we take $K = S^{n-1}$ to be the $\ell_2$-sphere of radius 1 in $\mathbb{R}^n$, and we take $\sigma^2 = 1$ and $\xi = 0$. Then $\text{pr}(y) = y/||y||_2$ for $y \neq 0$. The metric projection is not uniquely defined for $y = 0$ and $\text{exo}(S^{n-1}) = \{0\}$. The computation of the divergence is as above with

$$\nabla \cdot \text{pr}(y) = (n - 1) \frac{1}{||y||_2}$$

for $y \neq 0$. Since

$$||\xi - \text{pr}(y)||_2^2 = \left( \frac{y}{||y||_2} \right)^2 = 1,$$

we find that $\text{Risk} = 1$. Moreover,

$$E||Y - \text{pr}(Y)||_2^2 = E\left( ||Y||_2^2 \left( 1 - \frac{1}{||Y||_2} \right)^2 \right) = E||Y||_2^2 + 1 - 2E||Y||_2$$

$$= n + 1 - 2E||Y||_2,$$

and it follows that $\text{df} = E||Y||_2$. Since $||Y||_2^2 \sim \chi_n^2$ straightforward computations give that

$$E||Y||_2 = \frac{\sqrt{2\Gamma \left( \frac{n+1}{2} \right)}}{\Gamma \left( \frac{n}{2} \right)},$$

together with

$$E\left( \frac{1}{||Y||_2} \right) = \frac{\Gamma \left( \frac{n-1}{2} \right)}{\sqrt{2\Gamma \left( \frac{n}{2} \right)}} = \frac{\sqrt{2\Gamma \left( \frac{n+1}{2} \right)}}{(n - 1)\Gamma \left( \frac{n}{2} \right)}$$

for $n \geq 2$. This shows that

$$\text{df} = E||Y||_2 = (n - 1)E\left( \frac{1}{||Y||_2} \right) = E(\nabla \cdot \text{pr}(Y))$$

for $n \geq 2$, and we conclude that $\nu$ is the null measure for $n \geq 2$. This is an example where the measure $\nu$ can be 0 in cases where the exoskeleton is nonempty.

For $n = 1$ we have $\text{df} = E|Y| = \sqrt{\frac{2}{\pi}}$, whereas $\text{pr}(y) = \text{sign}(y)$ has derivative 0 for $y \neq 0$, and thus $\text{df}_S = 0$. It follows from Proposition 1 that $\nu = c\delta_0$ (with $\delta_0$ the Dirac measure in 0) for $c \geq 0$. Since

$$\frac{c}{\sqrt{2\pi}} = c\psi(0; 0, 1) = \int \psi(y; 0, 1)\nu(dy) = \sqrt{\frac{2}{\pi}}$$

we conclude that $\nu = 2\delta_0$. Note that $\nu$ is the distributional derivative of the sign function.
3. Divergence formulas for nonlinear least squares regression.

In this section our focus changes from the abstract results concerning an arbitrary closed set $K$ in $\mathbb{R}^n$ to sets that are given in terms of a $p$-dimensional parametrization. The main purpose is to provide explicit formulas for the computation of the divergence $\nabla \cdot \text{pr}(y)$ for a given $y \in \mathbb{R}^n$ in terms of the parametrization in two different situations of practical interest. We also provide some discussion of how the results achieved are related to results from the literature on the effective degrees of freedom or the effective number of parameters. In particular, we relate our first result to TIC and our second result to the estimate of degrees of freedom for linear $\ell_1$-penalized least squares estimation.

We assume in this section that $\zeta : \mathbb{R}^p \to \mathbb{R}^n$, that $\Theta \subseteq \mathbb{R}^p$ is a closed set, and that the image $K = \zeta(\Theta)$ is closed. Note that the image is automatically closed if $\zeta$ is continuous and $\Theta$ is compact. The observation $y \in \mathbb{R}^n$ is fixed, and we make the following local regularity assumptions about the parametrization $\zeta$.

- The metric projection of $y$ onto $K$ is unique with $\text{pr}(y) = \zeta(\hat{\beta})$ for $\hat{\beta} \in \Theta$.
- The map $\zeta : \mathbb{R}^p \to \mathbb{R}^n$ is $C^2$ in a neighborhood of $\hat{\beta}$.
- The map $\zeta : \Theta \to K$ is open in $\hat{\beta}$, that is, if $V$ is a neighborhood of $\hat{\beta}$ in $\mathbb{R}^p$, there is a neighborhood $U$ of $\text{pr}(y)$ in $\mathbb{R}^n$ such that

$$U \cap K \subseteq \zeta(V \cap \Theta).$$

The inverse function theorem implies the last assumption if the derivative of $\zeta$ has rank $p$ (forcing $p \leq n$) in $\hat{\beta}$.

We introduce the two $p \times p$ matrices $G$ and $J$ by

$$G_{kl} = \sum_{i=1}^{n} \partial_k \zeta_i(\hat{\beta}) \partial_l \zeta_i(\hat{\beta})$$

and

$$J_{kl} = G_{kl} - \sum_{i=1}^{n} (y_i - \zeta_i(\hat{\beta})) \partial_k \partial_l \zeta_i(\hat{\beta}).$$

**Theorem 3.** If $\hat{\beta} \in \Theta^\circ$ and $J$ has full rank $p$, then

$$\nabla \cdot \text{pr}(y) = \text{tr} \left( J^{-1} G \right).$$
Note that under sufficient regularity assumptions, standard asymptotic arguments, see Sections 2.3 and 2.5 in Claeskens and Hjort (2008), give for $p$ fixed the expansion

$$||Y - \Pr(\xi)||_2^2 = ||Y - \Pr(Y)||_2^2 + Z + 2\sigma^2 U^T J^{-1} U + o_P(1)$$

for $n \to \infty$, with $EZ = 0$, $EU = 0$, $VU = G$,

$$G_{kl} = \sum_{i=1}^{n} \partial_k \zeta_i(\beta_0) \partial_l \zeta_i(\beta_0) \quad \text{and} \quad J_{kl} = G_{kl} - \sum_{i=1}^{n} (\xi_i - \Pr_i(\xi)) \partial_k \partial_l \zeta_i(\beta_0).$$

The parameter $\beta_0$ is defined by $\zeta(\beta_0) = \Pr(\xi)$, that is, $\zeta(\beta_0)$ is the point in the model $K = \zeta(\Theta)$ closest to $\xi$. Defining $p^* = E(U^T J^{-1} U) = \text{tr}(J^{-1} G)$ as the effective number of parameters, the generalization of AIC to misspecified models, known as Takeuchi’s information criterion, becomes

$$\text{TIC} = ||y - \Pr(y)||_2^2 + 2\sigma^2 p^*.$$ 

We recognize $J$ and $G$ as plug-in estimates of $J$ and $G$, and thus $\text{tr}(J^{-1} G)$ as an estimate of $p^*$. Theorem 3 identifies this estimate as the unbiased estimate of the Stein degrees of freedom. Corollary 1 shows, however, that $\text{TIC} - n\sigma^2$ may be negatively biased as a risk estimate, and how the bias is related to the global geometry of $K$.

We then turn our attention to the case where the parameter set is an $\ell_1$-constrained subset of $\mathbb{R}^p$. That is, we consider parameter sets of the form

$$\Theta_s = \left\{ \beta \in \mathbb{R}^p \mid \sum_{k=1}^{p} \omega_k |\beta_k| \leq s \right\}$$

for $s \geq 0$ and $\omega \in \mathbb{R}^p$ a fixed vector of nonnegative weights. With $\Pr(y) = \zeta(\hat{\beta})$ for $\hat{\beta} \in \Theta_s$, then $\hat{\beta}$ is typically on the boundary of $\Theta_s$, and the formula in Theorem 3 for the divergence does not apply. Instead we note that $\hat{\beta}$ fulfills the Karush-Kuhn-Tucker conditions

$$D\zeta(\hat{\beta})^T (y - \zeta(\hat{\beta})) = \hat{\lambda} \gamma$$

for $\gamma \in \mathbb{R}^p$ with

$$\gamma_k = \omega_k \text{sign}(\hat{\beta}_k) \quad \text{if} \quad \hat{\beta}_k \neq 0$$

$$\gamma_k \in [-\omega_k, \omega_k] \quad \text{if} \quad \hat{\beta}_k = 0$$

and $\hat{\lambda} \geq 0$ the Lagrange multiplier. We introduce the active set of parameters as

$$\mathcal{A} = \{ i \mid \hat{\beta}_i \neq 0 \},$$

and let $J_{\mathcal{A},\mathcal{A}}$ and $G_{\mathcal{A},\mathcal{A}}$ denote the submatrices of $J$ and $G$, respectively, with indices in $\mathcal{A}$. 
A solution to the Karush-Kuhn-Tucker conditions is said to fulfill the sufficient second order conditions if \( \hat{\lambda} > 0 \), \( \gamma_k \in (\omega_k, \omega_k) \) for \( k \notin A \) and \( \delta^T J_{A,A} \delta > 0 \) for all nonzero \( \delta \in \mathbb{R}^A \) satisfying \( \delta^T \gamma_A = 0 \).

Note that the sufficient second order conditions imply that a solution to the Karush-Kuhn-Tucker conditions is a local minimizer of \( ||y - \zeta(\beta)||^2 \) in \( \Theta_s \).

**Theorem 4.** If \( J_{A,A} \) has full rank \( |A| \), if \( \gamma_A^T (J_{A,A})^{-1} \gamma_A \neq 0 \) and if \( \hat{\beta} \) fulfills the sufficient second order conditions, then

\[
\nabla \cdot \text{pr}(y) = \text{tr} \left( (J_{A,A})^{-1} G_{A,A} \right) - \frac{\gamma_A^T (J_{A,A})^{-1} \gamma_A (J_{A,A})^{-1} \gamma_A}{\gamma_A (J_{A,A})^{-1} \gamma_A}.
\]

First note that \( J_{A,A} \) has full rank \( |A| \) and \( \gamma_A^T (J_{A,A})^{-1} \gamma_A \neq 0 \) if \( J_{A,A} \) is positive definite. Then observe that in the case where \( \zeta \) is locally linear around \( \hat{\beta} \) to second order, that is, \( \partial_k \partial_l \zeta(\hat{\beta}) = 0 \), we get that \( \nabla \cdot \text{pr}(y) = |A| - 1 \). The fact that we have to subtract 1 from the number of active parameters may be a little surprising – as may the second term in the general formula above. Previous results in Zou, Hastie and Tibshirani (2007) and Tibshirani and Taylor (2012) for \( \ell_1 \)-penalized linear regression give that the unbiased estimate of degrees of freedom is \( |A| \). The difference arises because we do not consider the penalized estimator for a fixed regularization parameter \( \lambda \), but the estimator constrained to \( \Theta_s \) for a fixed \( s \). See also Example 2 for a similar difference for \( \ell_2 \)-regularization. It is possible to compute the divergence of the penalized estimator under conditions similar to those above. The result is \( \text{tr} \left( (J_{A,A})^{-1} G_{A,A} \right) \) as expected. However, we cannot in an obvious way relate this quantity to the degrees of freedom of the penalized nonlinear least squares estimator. Our results hinge crucially on the fact that the estimator can be expressed in terms of a metric projection onto a closed set. If the penalized estimator can be given such a representation, e.g. via dualization as outlined in Tibshirani and Taylor (2012) in the linear case, we might be able to transfer the results to the penalized estimator, but we expect this to be difficult without convexity.

### 4. Model selection for a \( d \)-dimensional linear ODE.

In this section we investigate the use of SURE for model selection in a nontrivial example of nonlinear \( \ell_1 \)-regularized regression. The example we consider is estimation of the parameters in a system of linear ordinary differential equations. We observe \( Y_1, \ldots, Y_m \in \mathbb{R}^d \) with \( Y_i \sim \mathcal{N}(\xi_i, \sigma^2 I_d) \) and \( \xi_i = e^{t_i B} x_i \) for \( t_i > 0, x_i \in \mathbb{R}^d \) and \( e^{tB} \) denoting the matrix exponential. It is well known
that $t \mapsto e^{tB}x$ is the solution of the linear $d$-dimensional ODE

$$\frac{d}{dt} f(t) = B f(t)$$

for $t > 0$ with initial condition $f(0) = x \in \mathbb{R}^d$. The unknown parameter is $B \in \mathbb{M}(d,d)$. We collect the observations into $Y = (Y_1, \ldots, Y_m) \in \mathbb{M}(d,m)$ and we let likewise $\xi = (\xi_1, \ldots, \xi_m)$ denote the collection of expectations. We will identify the matrices $Y$ and $\xi$ with vectors in $\mathbb{R}^n$ for $n = md$, which we denote by $Y$ and $\xi$ as well (formally, the identification is made by stacking the columns). Thus $Y \sim \mathcal{N}(\xi, \sigma^2 I_n)$. We also identify $B$ with a vector in $\mathbb{R}^p$ where $p = d^2$, and the parametrization $\zeta : \mathbb{R}^p \to \mathbb{R}^n$ is given as

$$\zeta(B) = (e^{t_1 B}x_1, \ldots, e^{t_m B}x_m).$$

We take a particular interest in developing methods that work for the high-dimensional case where $d$ is large, and we note that the number of observations $n = md$ as well as the number of parameters $p = d^2$ scales with $d$. For high-dimensional applications it may be realistic to achieve a good model for a sparse $B$. Obtaining a sparse estimate of $B$ is generally useful for computational reasons, and it may also be useful for network inference and interpretations.

We will in this paper focus on the special case $t_1 = \ldots = t_m = t$, which we will refer to as the isochronal model. For the isochronal model $\zeta(B) = e^{tB}x$ with $x = (x_1, \ldots, x_m)$, in which case it is natural to parametrize the model in terms of $A = e^{tB}$. With $\hat{A}$ an estimator of $A$ we can estimate $B$ as...
\( \hat{B} = \log(\hat{A})/t \) where \( \log \) denotes the principal matrix logarithm. The least squares estimator of \( A \) amounts to ordinary linear least squares regression. We are, however, interested in obtaining sparse estimates of \( B \). Since the principal matrix logarithm doesn’t preserve sparseness in general, we will maintain the parametrization in terms of \( B \) and consider the family of \( \ell_1 \)-constrained nonlinear least squares estimators

\[
\hat{B}_s = \arg \min_{B \in \Theta_s} ||Y - e^{tB} x||^2_2
\]

where \( \Theta_s = \{ B \mid \sum_{kl} \omega_{kl} |B_{kl}| \leq s \} \) for \( s \geq 0 \) and \( \omega \in M(d,d) \) is a given weight matrix (with \( \omega_{kl} \geq 0 \)). Some technical details on the computation of derivatives and the implementation of the optimization algorithm are treated in Appendix A.

We investigated the use of SURE for model selection in a simulation study with \( t = 1 \), \( d = 5 \), \( m = 10 \), \( \sigma^2 = 0.25 \) and

\[
B = \begin{pmatrix}
-0.1 & 0.5 & 0.1 & 0.0 & -0.1 \\
-0.5 & -0.1 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & -0.1 & 0.0 & 0.5 \\
0.0 & 0.0 & 0.0 & -0.1 & 0.5 \\
0.5 & 0.0 & 0.0 & -0.5 & -0.1
\end{pmatrix}
\]

The matrix exponential of \( B \) is

\[
e^B = \begin{pmatrix}
0.78 & 0.43 & 0.09 & 0.02 & -0.06 \\
-0.43 & 0.79 & -0.02 & 0.00 & 0.02 \\
0.11 & 0.02 & 0.91 & -0.11 & 0.43 \\
0.11 & 0.02 & 0.00 & 0.79 & 0.43 \\
0.41 & 0.11 & 0.02 & -0.43 & 0.78
\end{pmatrix}
\]

The initial conditions were sampled from the 5-dimensional normal distribution \( \mathcal{N}(0, 16I) \), and we used a total of 20,000 replications. For the choice of weights (the \( \omega_{kl} \)'s) we considered two situations; either \( \omega_{kl} = 1 \), or adaptive weights, as introduced in Zou (2006), based on the MLE,

\[
\omega_{kl} = \frac{1}{|\hat{B}_{kl}|}.
\]

In the simulation study we computed the \( \ell_1 \)-constrained estimators \( \hat{B}_s \) for a range of values of \( s \) and the corresponding estimates \( \hat{\text{Risk}}(s) \) of the risk. With

\[
\hat{s} = \arg \min_s \hat{\text{Risk}}(s)
\]
Fig 4. Risks and expected values of risk estimates, computed by simulation, for several different estimators. Unit weights were used for the $\ell_1$-constrained estimators.

denoting the data driven optimal estimate of $s$, the resulting estimator of $B$ is $\hat{B}_s$. In addition, we computed the risk estimate

$$\hat{\text{Risk}}(s) = \|Y - \text{pr}(Y)\|_2^2 - n\sigma^2 + 2\sigma^2(|A| - 1)$$

based on the approximation $\nabla \cdot e^{\hat{B}_s x} \simeq |A| - 1$. From the discussion after Theorem 4 this is a good approximation if the matrix exponential is well approximated by a linear map around $\hat{B}_s$. We computed the MLE as well as a sequence of sparse($r$) solutions obtained by hard thresholding the MLE. The results of the simulation study are summarized in Figures 4, 6 and 5. If we first consider the case of constant weights, see Figure 4, the risk showed, as a function of $s$, a characteristic shape, and the constrained estimator had minimal risk around $s = 3.6$. Both risk estimates, $\hat{\text{Risk}}(s)$ and $\hat{\text{Risk}}(\hat{s})$, were, in this case, very close to being unbiased. The risks of the MLE and the sequence of thresholded MLEs were all larger than the risk for a substantial range of constrained estimators. More importantly, the risk of $\hat{B}_s$ was also smaller than the risk of any of the thresholded estimators. We should note, however, that $\hat{\text{Risk}}(\hat{s})$ did on average underestimate the actual risk of $\hat{B}_s$ somewhat.

Figure 5 zooms in on the possible biases of the risk estimates when using unit weights. The figure shows that $\hat{\text{Risk}}(s)$ was effectively unbiased. Even
with 20,000 replications a negative bias was hardly even detectable. The approximation, $\tilde{\text{Risk}}(s)$, appeared to have a slightly larger mean, but again hardly a detectable bias. We have systematically observed in our simulation studies that $E(\tilde{\text{Risk}}(s)) \geq E(\text{Risk}(s))$, but that the difference was very small, and that both risk estimates were very close to being unbiased. This suggests that the sets $\exp(\Theta_s)x$ appear to be predominantly convex with flat boundaries. Moreover, since the difference between $\tilde{\text{Risk}}(s)$ and $\tilde{\text{Risk}}(s)$ was generally found to be very small, the latter approximation can be a useful alternative to $\text{Risk}(s)$ in practice, since it is much faster to compute.

For the adaptive weights, see Figure 6, the risk behaved similarly as a function of $s$, and the optimal value of $s$, this time around $s = 11$, yielded a risk comparable to the risk obtained with unit weights. Both the risk estimates were, however, considerably downwardly biased for the adaptive weights. This was to be expected as the risk estimates do not take the data driven choice of weights into account. Despite the bias, the data driven estimate, $\hat{s}$, of the constraint resulted in an estimator $\hat{B}_{\hat{s}}$ with close to minimal risk. On the downside, $\text{Risk}(\hat{s})$ did, due to the bias of $\text{Risk}(s)$, considerably underestimate the actual risk of $\hat{B}_{\hat{s}}$.

We also observed that the estimator $\hat{B}_{\hat{s}}$ was sparser when using adaptive weights.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig5}
\caption{Estimated bias from the simulation study for the two risk estimates using unit weights.}
\end{figure}
Fig 6. Risks and expected values of risk estimates, computed by simulation, for several different estimators. Adaptive weights were used for the $\ell_1$-constrained estimators.

weights (15.6 nonzero entries on average) than when using unit weights (18.8 nonzero entries on average). Using $\hat{B}_s$ to obtain a structural estimator of the nonzero entries the accuracy (fraction of correctly estimated zero and nonzero entries) was 0.79 with adaptive weights compared to 0.69 with unit weights.

To understand better the results of the simulation study – and the nature of the nonlinear least squares problem – it would be desirable to be able visualize the image sets $\exp(\Theta_s)$, or, in particular, the images $\exp(\partial \Theta_s)$ of the boundaries of $\Theta_s$, for different choices of $s$. These are the images under the matrix exponential of the boundaries of $\ell_1$-balls. As these sets live in 25 dimensions a visualization is challenging. Figure 7 shows two selected slices of the sets by affine subspaces. The slices were constructed as follows. With

$$e^{B(a,b,c,d)} = \begin{pmatrix} a & c & 0.09 & 0.02 & -0.06 \\ b & d & -0.02 & -0.00 & 0.02 \\ 0.11 & 0.02 & 0.91 & -0.11 & 0.43 \\ 0.11 & 0.02 & 0.00 & 0.79 & 0.43 \\ 0.41 & 0.11 & 0.02 & -0.43 & 0.78 \end{pmatrix}$$

it holds that $B(0.78, -0.43, 0.43, 0.79) = B$ – the matrix that we used in the simulation. Fixing either $(b, c) = (-0.43, 0.43)$ or $(a, d) = (0.78, 0.79)$ we get
the two affine subspaces considered, which both include $B$. The slices in Figure 7 were computed as contour curves for $(a, d) \mapsto \|B(a, -0.43, 0.43, d)\|_1$ and $(c, d) \mapsto \|B(0.78, c, d, 0.79)\|_1$, respectively.

5. Proofs. In this section we give the proofs of the results stated in Sections 2 and 3. Doing so we will provide a brief account on the ideas and strategies used with some appropriate references to the literature. A further discussion of how our results and proofs are related to the literature is given in Section 6.

5.1. Proofs of results in Section 2. Central to the proofs of Theorem 1 and Theorem 2 in Section 2 is a famous theorem of Alexandrov given first in Alexandrov (1939). It loosely states that a convex function is twice differentiable except perhaps on a Lebesgue null set – Theorem 5 gives a precise version. Our first lemma is not new, but since we expect that many readers of this paper will not know about the relation between metric projections and convex functions we provide a proof.

Lemma 1. With $K \subseteq \mathbb{R}^n$ the function

$$\rho(y) = \sup_{x \in K} \{y^T x - \|x\|^2 / 2\}$$

is convex. With $\partial \rho$ denoting the subdifferential of $\rho$ then $\partial \rho(y)$ contains the set of points in $K$ closest to $y$. 
Proof. Since $\rho$ is the pointwise supremum of the affine (thus convex) functions
\[ y \mapsto y^T x - |x|^2 / 2 = |y|^2 / 2 - |y - x|^2 / 2, \]
it is convex, and
\[ \rho(y) = |y|^2 / 2 - \inf_{x \in K} |y - x|^2 / 2. \]
With
\[ \Pr(y) = \arg\min_{x \in K} |y - x|^2 \]
the nonempty set of points in $K$ closest to $y$ it follows that
\[ \rho(y) = y^T x - |x|^2 / 2 \]
for all $x \in \Pr(y)$. For $x \in \Pr(y)$
\[ \rho(y + z) = \sup_{x \in K} \{ y^T x - |x|^2 / 2 + z^T x \} \geq y^T x - |x|^2 / 2 + z^T x = \rho(y) + z^T x, \]
which shows that $\Pr(y) \subseteq \partial \rho(y)$ by definition of the subdifferential.

If $\rho$ is differentiable in $y$ we write $\nabla \rho(y)$ for the gradient. The domain,
\[ D \subseteq \mathbb{R}^n, \]
of $\nabla \rho$ is the set on which $\rho$ is differentiable. The previous lemma shows that for $y \in D$, the metric projection, $\Pr(y)$, onto $K$ is unique, and
\[ \Pr(y) = \nabla \rho(y) \quad \text{and} \quad \partial \rho(y) = \{ \Pr(y) \} = \Pr(y). \]

Observe also that if $y \in D$, if $y_n \to y$ and if $z_n \in \Pr(y_n)$ converges to $z$ then
\[ \rho(y + x) = \lim_{n \to \infty} \rho(y_n + x) \geq \lim_{n \to \infty} \rho(y_n) + x^T z_n \geq \rho(y) + x^T z \]
which implies that $z \in \partial \rho(y) = \{ \Pr(y) \}$, whence $z = \Pr(y)$. This proves a continuity property of the metric projection: If $y \in D$ and $U$ is a neighborhood of $\Pr(y)$ then $\{ z \in \mathbb{R}^n \mid \Pr(z) \subseteq U \}$ contains a neighborhood of $y$. These facts are all well known, see e.g. Theorem 3 in Asplund (1968) for a similar but abstract formulation, or Theorem 3.3 in Evans and Harris (1987) for an alternative formulation in $\mathbb{R}^n$.

We then state a version of Alexandrov’s Theorem particularly useful for our purposes.

Theorem 5. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a convex function, and let $D \subseteq \mathbb{R}^n$ denote the subset on which $g$ is differentiable. For Lebesgue almost all $y$ it holds that $y \in D$ and there exists a matrix $A$ such that
\[ (13) \quad \nabla g(x) = \nabla g(y) + A(x - y) + o(|x - y|_2) \]
for $x \in D$. The matrix $A$ is symmetric and positive semidefinite and as such uniquely determined by (13).
The theorem is a direct consequence of Theorem 2.3 and Theorem 2.8 in Rockafellar (2000). See, in addition, Chapter 13 – and Theorem 13.51 in particular – in Rockafellar and Wets (1998) for similar results. Theorem 5 also follows from Theorem 6.1 and Theorem 7.1 in Howard (1998), which is a nice self contained exposition of Rademacher’s and Alexandrov’s theorems.

In the light of Definition 1, Theorem 5 says that for a convex function $g$, $\nabla g$ is defined Lebesgue almost everywhere, and $\nabla g$ is differentiable in the extended sense Lebesgue almost everywhere. Note, however, that the differentiability points of $\nabla g$ can be a strict subset of its maximal domain of definition.

**Proof of Theorem 1.** We first prove that there is a Borel measurable selection of the set valued metric projection $Pr$, where

$$Pr(y) = \arg \min_{x \in K} ||y - x||^2$$

is defined as in the proof of Lemma 1. This follows by general arguments in Rockafellar and Wets (1998). As a set valued map, $Pr$ is outer semicontinuous by Example 5.23 in Rockafellar and Wets (1998), and combining Theorem 5.7 and Exercise 14.9 in Rockafellar and Wets (1998) it is, still as a set valued map, closed-valued and Borel measurable. Corollary 14.6 in Rockafellar and Wets (1998) implies that $Pr$ admits a Borel measurable selection, that is, there is a Borel measurable map $pr : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with

$$pr(y) \in Pr(y)$$

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

Then we prove, using Alexandrov’s Theorem, that the selection of $pr(y)$ is unique and differentiable in the extended sense for Lebesgue almost all $y$. Theorem 5 holds for the convex function $\rho$. For those $y$ where (13) holds, the differentiability of $\rho$ in $y$ assures that $pr(y) = \nabla \rho(y)$ is uniquely defined in $y$ as well as differentiable in $y$ in the sense of (13). The domain $D$ on which $pr$ is uniquely defined thus satisfies that $D^c$ is a Lebesgue null set, and $pr : D \mapsto \mathbb{R}^n$ satisfies (13) for Lebesgue almost all $y$. That is,

$$pr(x) = pr(y) + A(x - y) + o(||x - y||^2)$$

for $x \in D$, and $pr$ is differentiable in the extended sense for Lebesgue almost all $y$. By definition,

$$\partial_j pr_i(y) = A_{ij}$$

for those $y$ where $pr$ is differentiable in the extended sense, and since $A$ is positive semidefinite, $\partial_i pr_i(y) \geq 0$ for $i = 1, \ldots, n$. \qed
From hereon we assume, in accordance with Theorem 1, that a choice of pr has been made on the set where pr is not unique, such that pr : \( \mathbb{R}^n \rightarrow \mathbb{R}^n \) is Borel measurable.

We turn to the proof of Theorem 2. The relation in Theorem 2 between the degrees of freedom, df, and the Stein degrees of freedom, df_S, will be established by partial integration. However, to handle metric projections in full generality we have to turn to distributional formulations of differentiation. Partial integration holds by definition for distributional differentiation. What we need is to identify the distributional partial derivatives of the coordinates of the metric projection. For this purpose, we define a signed Radon measure to be the difference of two (positive) Radon measures. In this sense a signed Radon measure need not have bounded total variation. Though we have to be careful with such a definition to avoid the undefined “\( \infty - \infty \)”, the difference of two Radon measures does give a well defined linear functional on \( C_c(\mathbb{R}^n) \).

**Definition 4.** A function \( g \in L_{\text{loc}}^1(\mathbb{R}^n) \) is of locally bounded variation if there exist signed Radon measures \( \mu_j \) for \( j = 1, \ldots, n \) on \( \mathbb{R}^n \) such that

\[
\int_{\mathbb{R}^n} g(y) \partial_j \varphi(y) \, dy = -\int_{\mathbb{R}^n} \varphi(y) \mu_j(dy)
\]

for all \( \varphi \in C_\infty_c(\mathbb{R}^n) \).

Thus the functions of locally bounded variation are those \( L_{\text{loc}}^1 \)-functions whose distributional partial derivatives are signed Radon measures. It is easily verified that Definition 4 is equivalent to other definitions in the literature, e.g. the definition in Chapter 5 in Evans and Gariepy (1992).

**Lemma 2.** The functions \( pr_i \) for \( i = 1, \ldots, n \) are of locally bounded variation. With \( \mu_{ij} \) denoting the \( j \)'th distributional partial derivative of \( pr_i \) it holds that

- \( \mu_{ij} = \mu_{ji} \),
- \( \sum_{i,j=1}^n x_i x_j \mu_{ij} \) is a positive measure for all \( x \in \mathbb{R}^n \)
- and

\[
\int_{\mathbb{R}^n} pr_i(y) \partial_j \varphi(y) \, dy = -\int_{\mathbb{R}^n} \varphi(y) \mu_{ij}(dy)
\]

for all \( \varphi \in C^\infty(\mathbb{R}^n) \) with

\[
\sup_{y \in \mathbb{R}^n} (1 + |y|^2) \max \{|\varphi(y)|, |\partial_1 \varphi(y)|, \ldots, |\partial_n \varphi(y)|\} < \infty
\]

for all \( N \in \mathbb{N}_0 \).
Proof. First recall that
\[ |pr_i(y)| \leq \|pr(y)\|_2 \leq \|pr(0)\|_2 + \|y\|_2, \]
which proves that pr_i is in \( L^1_{\text{loc}} \). A standard mollifier argument gives that for all \( x \in \mathbb{R}^n \)
\[ \varphi \mapsto \int_{\mathbb{R}^n} \rho(y) \sum_{i,j=1}^n x_ix_j \partial_i \partial_j \varphi(y) \, dy \]
is a positive linear functional on \( C^\infty_c(\mathbb{R}^n) \) due to convexity of \( \rho \). Riesz’s representation theorem gives the existence of a Radon measure \( \mu^x \) such that
\[ \int_{\mathbb{R}^n} \rho(y) \sum_{i,j=1}^n x_ix_j \partial_i \partial_j \varphi(y) \, dy = \int_{\mathbb{R}^n} \varphi(y) \mu^x(dy). \]
Taking \( \mu_{ii} = \mu_{ei} \) and
\[ \mu_{ij} = \mu^{(e_i+e_j)/\sqrt{2}} - \mu_{ii} - \mu_{jj} \]
for \( i \neq j \) gives the existence of signed Radon measures \( \mu_{ij} \), which by construction fulfill the two first bullet points. Since \( \rho \) is convex, it is locally Lipschitz continuous, hence weakly differentiable with first weak partial derivatives coinciding with the pointwise partial derivatives, \( pr_i(y) \), for Lebesgue almost all \( y \). Hence
\[ \int_{\mathbb{R}^n} pr_i(y) \partial_j \varphi(y) \, dy = -\int_{\mathbb{R}^n} \rho(y) \partial_i \partial_j \varphi(y) \, dy = -\int_{\mathbb{R}^n} \varphi(y) \mu_{ij}(dy) \]
for all \( \varphi \in C^\infty_c(\mathbb{R}^n) \). We then prove that the partial integration formula generalizes to all \( \varphi \in C^\infty(\mathbb{R}^n) \) that fulfill (14). To this end fix a positive function \( \kappa \in C^\infty_c(\mathbb{R}^n) \) such that \( \kappa(y) = 1 \) for \( \|y\|_2 \leq 1 \). Define
\[ q_r(y) = (1 + \|y\|_2^2)^{-N} \kappa(ry), \]
then \( q_r \in C^\infty_c(\mathbb{R}^n) \) and
\[ q_r(y) \geq (1 + \|y\|_2^2)^{-N} 1(r\|y\|_2 \leq 1) \rightarrow (1 + \|y\|_2^2)^{-N} \]
for \( r \rightarrow 0 \). By monotone convergence
\[ \int_{\mathbb{R}^n} q_r(y) \mu_{ij}(dy) \rightarrow \int_{\mathbb{R}^n} (1 + \|y\|_2^2)^{-N} \mu_{ij}(dy) \]
for \( r \rightarrow 0 \). Moreover, \( \kappa(ry) = 1 \) and \( \partial_j \kappa(ry) = 0 \) for \( \|y\|_2 \leq 1/r \), hence
\[ \partial_j q_r(y) \rightarrow \partial_j (1 + \|y\|_2^2)^{-N} \]
for \( r \to 0 \). Since
\[
|pr_i(y)\partial_j q_r(y)| \leq p(y)(1 + ||y||^2)^{-2N}
\]
for some polynomial \( p(y) \) of degree \( N + 1 \) independent of \( r \) (for \( r \leq 1 \), say), and since the upper bound is integrable w.r.t. the \( n \)-dimensional Lebesgue measure for \( N \) large enough, it follows by dominated convergence that for \( N \) large enough
\[
\int_{\mathbb{R}^n} (1 + ||y||^2)^{-N} \mu_{ij}(dy) = -\int_{\mathbb{R}^n} pr_i(y)\partial_j (1 + ||y||^2)^{-N} dy.
\]
The function \( y \mapsto (1 + ||y||^2)^{-N} \) is, in particular, \( \mu_{ij} \)-integrable. If \( \varphi \in C^\infty(\mathbb{R}^n) \) fulfills (14) we let \( \varphi_r(y) = \varphi(y)\kappa(ry) \). Then \( \varphi_r \in C^\infty_c(\mathbb{R}^n) \), \( \varphi_r(y) \to \varphi(y) \) for \( r \to 0 \), and
\[
\partial_j \varphi_r(y) = \partial_j \varphi(y)\kappa(ry) + \varphi(y)r\partial_j \kappa(ry) \to \partial_j \varphi(y)
\]
for \( r \to 0 \). Moreover, for \( r \leq 1 \) there is a constant \( C_N \) such that
\[
|pr_i(y)\partial_j \varphi_r(y)| \leq C_N(1 + ||y||^2)^{-N+2}
\]
as well as
\[
|\varphi_r(y)| \leq C_N(1 + ||y||^2)^{-N}
\]
since \( \varphi \) fulfills (14). Again by Lebesgue as well as \( \mu_{ij} \)-integrability of the upper bound for \( N \) large enough, it follows from dominated convergence that
\[
\int_{\mathbb{R}^n} pr_i(y)\partial_j \varphi(y) dy = \lim_{r \to 0} \int_{\mathbb{R}^n} pr_i(y)\partial_j \varphi_r(y) dy = -\lim_{r \to 0} \int_{\mathbb{R}^n} \varphi_r(y)\mu_{ij}(dy) = -\int_{\mathbb{R}^n} \varphi(y)\mu_{ij}(dy).
\]

The first part of the proof of Lemma 2, where we establish the existence of the \( \mu_{ij} \)-measures, follows the proof of Theorem 6.3.2 in Evans and Gariepy (1992). In the remaining part we effectively prove that \( pr_i \) is a tempered distribution. This actually follows directly from the polynomial bound on
pr$_i$ by Example 7.12(c) in Rudin (1991). However, we need a little more than just the fact that the continuous linear functional

$$\varphi \mapsto \int_{\mathbb{R}^n} \text{pr}_i(y) \partial_j \varphi(y) \, dy$$

on the test functions $C_0^\infty(\mathbb{R}^n)$ extends to a continuous linear functional on the Schwartz space $S$ of rapidly decreasing functions. We also need the explicit form of the extension (the partial integration formula) as stated in Lemma 2.

To finally prove Theorem 2 we need to relate the distributional partial derivatives $\mu_{ij}$ of pr$_i$ to the pointwise partial derivatives $\partial_j \text{pr}_i$ defined Lebesgue almost everywhere. To this end we need the concept of approximate differentiability.

**Definition 5.** Let $m_n$ denote the $n$-dimensional Lebesgue measure and $B(y, r)$ the $\ell_2$-ball with center $y$ and radius $r$. A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is approximately differentiable in $y$ if there is a matrix $A$ such that for all $\varepsilon > 0$

$$\frac{1}{m_n(B(y, r))} m_n \left( \left\{ x \in B(y, r) \left| \frac{||f(x) - f(y) - A(x - y)||}{||x - y||_2} \geq \varepsilon \right. \right\} \right) \to 0$$

for $r \to 0$.

By Theorem 6.1.3 in Evans and Gariepy (1992) the matrix $A$ is unique if $f$ is approximately differentiable in $y$. It is called the approximate derivative of $f$ in $y$. Note that approximate differentiability of $f$ in $y$ is a local property, which only requires that $f$ is defined Lebesgue almost everywhere in a neighborhood of $y$.

**Lemma 3.** If $f : D \to \mathbb{R}^n$ is differentiable in $y$ in the extended sense then $f$ is approximately differentiable in $y$ with the same derivative.

**Proof.** Assume that $f$ is differentiable in $y$ in the extended sense with derivative $A$. We can then for fixed $\varepsilon > 0$ choose $r$ sufficiently small such that $D^c \cap B(y, r)$ is a Lebesgue null set and

$$\frac{||f(y) - f(x) - A(x - y)||}{||x - y||_2} < \varepsilon$$

for $x \in D \cap B(y, r)$. Choosing an arbitrary extension of $f$ to $B(y, r)$ we find that

$$\left\{ x \in B(y, r) \left| \frac{||f(y) - f(x) - A(x - y)||}{||x - y||_2} \geq \varepsilon \right. \right\} \subseteq D^c \cap B(y, r),$$
which implies that \( f \) is approximately differentiable in \( y \) with derivative \( A \).

If \( f : \mathbb{R}^n \to \mathbb{R}^n \) has coordinates of locally bounded variation with corresponding distributional partial derivatives of \( f_i \) denoted \( \mu_{ij} \) for \( j = 1, \ldots, n \) we have by Lebesgue’s decomposition theorem that

\[
\mu_{ij} = h_{ij} \cdot m_n + \nu_{ij}
\]

with \( \nu_{ij} \perp m_n \). We can now state (and subsequently use) a well known but rather deep result on approximate differentiability of functions of locally bounded variation. See Theorem 6.1.4 in Evans and Gariepy (1992).

**Theorem 6.** If \( f : \mathbb{R}^n \to \mathbb{R}^n \) has coordinates of locally bounded variation then \( f_i \) is approximately differentiable for Lebesgue almost all \( y \) with derivative \((h_{i1}(y), \ldots, h_{in}(y))\).

It is straightforward to see that if \( f \) has coordinates of locally bounded variation then it is also, as a function from \( \mathbb{R}^n \) to \( \mathbb{R}^n \), approximately differentiable for Lebesgue almost all \( y \) with derivative \((h_{ij}(y))_{i,j=1,\ldots,n}\).

**Proof of Theorem 2.** From Lemma 2, \( \text{pr}_i \) is of locally bounded variation with distributional partial derivatives \( \mu_{ij} \). Combining Theorem 1, Lemma 3 and Theorem 6 – and using that the approximate derivative is unique – we conclude that

\[
\mu_{ij} = \partial_j \text{pr}_i \cdot m_n + \nu_{ij}
\]

with \( \nu_{ij} \perp m_n \).

Letting \( \psi(y; \xi, \sigma^2) \) denote the density for the multivariate normal distribution with mean \( \xi \) and covariance matrix \( \sigma^2 I \) we have that

\[
\partial_j \psi(y; \xi, \sigma^2) = -\frac{(y_j - \xi_j)}{\sigma^2} \psi(y; \xi, \sigma^2).
\]

Since \( \psi(\cdot; \xi, \sigma^2) \in C^\infty(\mathbb{R}^n) \) fulfills (14), Lemma 2 implies that

\[
\text{cov}(Y_i, \text{pr}_i(Y)) = \int_{\mathbb{R}^n} \text{pr}_i(y)(y_i - \xi_i)\psi(y; \xi, \sigma^2) \, dy
\]

\[
= -\sigma^2 \int_{\mathbb{R}^n} \text{pr}_i(y)\partial_i \psi(y; \xi, \sigma^2) \, dy
\]

\[
= \sigma^2 \int_{\mathbb{R}^n} \psi(y; \xi, \sigma^2) \mu_{ii} \, dy
\]

\[
= \sigma^2 \int_{\mathbb{R}^n} \psi(y; \xi, \sigma^2) \partial_i \text{pr}_i(y) \, dy + \sigma^2 \int_{\mathbb{R}^n} \psi(y; \xi, \sigma^2) \nu_{i} \, dy.
\]


DEGREES OF FREEDOM FOR NONLINEAR LEAST SQUARES

Theorem (2) follows by division with \( \sigma^2 \) and summation over \( i \), which gives that

\[
\nu = \sum_{i=1}^{n} \nu_{ii}.
\]

\[\square\]

**Proof of Proposition 1.** The set \( U = \mathbb{R}^n \setminus \text{ex}(K) \) is open. If \( pr_i \) is locally Lipschitz on \( U \) Theorem 4.2.5 in Evans and Gariepy (1992) gives that \( pr_i \) is weakly differentiable, and the weak partial derivative in the \( j \)'th direction coincides with the Lebesgue almost everywhere defined \( \partial_j pr_i \). That is,

\[
\int_{\mathbb{R}^n} pr_i(y)\partial_j \varphi(y) \, dy = -\int_{\mathbb{R}^n} \partial_j pr_i(y)\varphi(y) \, dy
\]

for all \( \varphi \in C^\infty_c(\mathbb{R}^n) \). It follows that

\[
\mu_{ij} = \partial_j pr_i \cdot m_n,
\]

and all the singular measures \( \nu_{ij} \) are null measures. \[\square\]

5.2. **Proofs of the results in Section 3.** The formulas for computation of the divergence given in Section 3 will be proved using the implicit function theorem to compute the divergence of \( \zeta(\hat{\beta}) \). To connect such a local result expressed in the \( \beta \) parametrization with the divergence of the globally defined metric projection we will first establish that there is a neighborhood of \( y \) where the (global) metric projection can be found by minimizing \( ||z - \zeta(\beta)||^2 \) in a neighborhood of \( \hat{\beta} \). Note that \( Pr(z) \) denotes, as in the proof of Lemma 1, the set of metric projections of \( z \).

**Lemma 4.** If the regularity assumptions on \( \zeta \) as stated in Section 3 hold, then for all neighborhoods \( V \) of \( \hat{\beta} \) there exists a neighborhood \( N \) of \( y \) such that

\[
Pr(z) = \zeta(\arg \min_{\beta \in V \cap \Theta} ||z - \zeta(\beta)||^2_2)
\]

for \( z \in N \).

**Proof.** With \( V \) a neighborhood of \( \hat{\beta} \) there is, since \( \zeta \) was assumed to be open at \( \hat{\beta} \), a neighborhood \( U \) of \( pr(y) = \zeta(\hat{\beta}) \) such that

\[
U \cap K \subseteq \zeta(V \cap \Theta).
\]
By the continuity property of the metric projection there is a neighborhood \( N \) of \( y \) such that \( \text{Pr}(z) \subseteq U \) for \( z \in N \). By definition, \( \text{Pr}(z) \subseteq K \), hence

\[
\text{Pr}(z) \subseteq \zeta(V \cap \Theta).
\]

This proves first that \( W = \arg\min_{\beta \in V \cap \Theta} \|z - \zeta(\beta)\|_2^2 \) is not empty, and second that \( \beta \in W \) if and only if \( \zeta(\beta) \in \text{Pr}(z) \).

Below we use the implicit function theorem to show that for neighborhoods \( N \) of \( y \) and \( V \) of \( \hat{\beta} \) there exists a \( C^1 \)-map \( \hat{\beta} : N \to V \cap \Theta \) such that

\[
\{\zeta \circ \hat{\beta}(z)\} = \zeta(\arg\min_{x \in V \cap \Theta} \|z - \zeta(x)\|_2^2).
\]

It follows from Lemma 4 that

\[
pr(z) = \zeta \circ \hat{\beta}(z)
\]

for \( z \) in a neighborhood (contained in \( N \)) of \( y \). This ensures that

\[
(15) \quad \nabla \cdot pr(y) = \nabla \cdot \zeta \circ \hat{\beta}(y).
\]

The next lemma on differentiation of the quadratic loss is a straightforward computation, and its proof is left out.

**Lemma 5.** If \( \zeta \) is \( C^2 \) in a neighborhood of \( \beta \) then \( f(z, \beta) = \frac{1}{2} \|z - \zeta(\beta)\|_2^2 \) is \( C^2 \) in a neighborhood of \( (y, \beta) \) with

\[
\partial_{z_i} \partial_k f(z, \beta) = -\partial_k \zeta(\beta)
\]

and

\[
\partial_k \partial_l f(z, \beta) = J_{kl},
\]

where \( J_{kl} \) is given by (9).

Note that in the notation above, \( \partial_k \) refers to differentiation w.r.t. to \( \beta_k \) and \( \partial_{z_i} \) refers to differentiation w.r.t. \( z_i \).

**Proof of Theorem 3.** With \( f \) as in Lemma 5 the estimator \( \hat{\beta} \) fulfills

\[
\nabla f(y, \hat{\beta}) = 0,
\]

with the Jacobian of the map \( \beta \mapsto \nabla f(y, \beta) \) being \( J \) by Lemma 5. Since \( J \) has full rank by assumption the implicit function theorem implies that there
DEGREES OF FREEDOM FOR NONLINEAR LEAST SQUARES

29

is a continuously differentiable solution map \( \hat{\beta}(z) \), defined in a neighborhood of \( y \), such that

\[
\nabla \beta f(z, \hat{\beta}(z)) = 0.
\]

Moreover, \( D_z \nabla \beta f(y, \hat{\beta}) = -D_\beta \zeta(\hat{\beta})^T \) by Lemma 5, which gives by implicit differentiation that

\[
D_z \hat{\beta}(y) = J^{-1} D_\zeta(\hat{\beta})^T.
\]

Hence,

\[
D_z (\zeta \circ \hat{\beta})(y) = D_\zeta(\hat{\beta}) J^{-1} D_\zeta(\hat{\beta})^T.
\]

It follows from (15) that

\[
\nabla \cdot \text{pr}(y) = \text{tr}(D_\zeta(\hat{\beta}) J^{-1} D_\zeta(\hat{\beta})^T) = \text{tr}(J^{-1} D_\zeta(\hat{\beta})^T D_\zeta(\hat{\beta})) = \text{tr}(J^{-1} G),
\]

since \( G = D_\zeta(\hat{\beta})^T D_\zeta(\hat{\beta}) \) as defined by (8).

**Proof of Theorem 4.** With \( f \) as in Lemma 5 the estimator \( \hat{\beta} \) fulfills, by assumption,

\[
\nabla \beta f(y, \hat{\beta}) = \hat{\lambda} \gamma
\]

for \( \hat{\lambda} > 0 \), \( \gamma \in \mathbb{R}^p \), \( \gamma_k = \omega_k \text{sign}(\hat{\beta}_k) \) if \( \hat{\beta}_k \neq 0 \) and \( \gamma_k \in (-\omega_k, \omega_k) \) if \( \hat{\beta}_k = 0 \). Moreover, as \( \hat{\lambda} > 0 \) it holds that \( \sum_{k=1}^p \gamma_k \hat{\beta}_k = s \). In the following we identify any \( \mathbb{R}^A \)-vector denoted \( \beta_A \) with an \( \mathbb{R}^p \) vector with 0’s in entries with indices not in \( A \). We introduce the map

\[
R(z, \beta_A, \lambda) = \left( \frac{\nabla \beta f(z, \beta_A) - \lambda \gamma_A}{\sum_{i=1}^p \gamma_k \beta_{A,k} - s} \right),
\]

and we observe that \( R(y, \hat{\beta}_A, \hat{\lambda}) = 0 \). The derivative of \( R \) is found to be

\[
D_{\beta_A, \lambda} R(y, \hat{\beta}_A, \hat{\lambda}) = \left( \begin{array}{c} J_{A,A} \gamma_A \\ \gamma_A \end{array} \right).
\]

By the assumptions made on \( J_{A,A} \) this matrix is invertible with

\[
\left( \begin{array}{c} J_{A,A} \gamma_A \\ \gamma_A \\ 0 \end{array} \right)^{-1} = \left( \begin{array}{cc} (J_{A,A})^{-1} - (J_{A,A})^{-1} \gamma_A \gamma_A^T (J_{A,A})^{-1} \\ \gamma_A (J_{A,A})^{-1} \gamma_A \\ * \end{array} \right).
\]

It follows from the implicit function theorem that there is a neighborhood of \( y \) in which there is a continuously differentiable solution map \( (\hat{\beta}_A(z), \hat{\lambda}(z)) \) that fulfills \( R(z, \hat{\beta}_A(z), \hat{\lambda}(z)) = 0 \). By the \( C^2 \)-assumption the solution map fulfills the second order sufficient conditions in a neighborhood of \( y \), and \( \hat{\beta}_A(z) \) is a local solution to the constrained optimization problem. Since
\( D_z \nabla f(y, \hat{\beta}) = -D_{\beta} \zeta(\hat{\beta})^T \) by Lemma 5, we get by implicit differentiation that
\[
D_z \hat{\beta}_A(y) = \left( (J_{A,A})^{-1} - \frac{(J_{A,A})^{-1} \gamma_A \gamma_A^T (J_{A,A})^{-1}}{\gamma_A^2 (J_{A,A})^{-1} \gamma_A} \right) (D \zeta(\hat{\beta}),..A)^T.
\]
Since \((D \zeta(\hat{\beta}),..A)^T D \zeta(\hat{\beta}),..A = G_{A,A}\) it follows as in the proof of Theorem 3 that
\[
\nabla \cdot \text{pr}(y) = \text{tr} \left( (J_{A,A})^{-1} G_{A,A} - \frac{(J_{A,A})^{-1} \gamma_A \gamma_A^T (J_{A,A})^{-1} G_{A,A}}{\gamma_A^2 (J_{A,A})^{-1} \gamma_A} \right) = \text{tr} \left( (J_{A,A})^{-1} G_{A,A} - \frac{\gamma_A^2 (J_{A,A})^{-1} G_{A,A} (J_{A,A})^{-1} \gamma_A}{\gamma_A^2 (J_{A,A})^{-1} \gamma_A} \right).
\]

6. Discussion. Our main result obtained in this paper is Theorem 2, which implies a characterization of the possible bias of SURE. The bias is given in terms of \( d_f - d_f S \), whose magnitude is determined by how large \( \psi(y; \xi, \sigma^2) \) is on the Lebesgue null set \( N \) where the singular measure \( \nu \) is concentrated. This is, in turn, determined by the distance (scaled by \( 1/\sigma \)) from \( \xi \) to points in \( N \) in combination with the distribution of the mass of the measure \( \nu \) on \( N \). The singular measure depends only on \( K \), and it represents global geometric properties of \( K \). How the global geometry of \( K \) affects the degrees of freedom is given in a general but transparent way by (6) in Theorem 2.

We gave three simple examples where analytic computations could shed some light on the general results, and then we considered a more serious application in Section 4 on the estimation of parameters in a \( d \)-dimensional linear ODE. This example served several purposes. First we used it to test our algorithms for computing the nonlinear \( \ell_1 \)-constrained or \( \ell_1 \)-penalized least squares estimator, and we used it to test the divergence formula given in Theorem 4. For the chosen model and parameter set we concluded that \( \hat{\text{Risk}}(s) \) was, for all practical purposes, unbiased, that it was useful for selection of \( s \), and that the selected model had a lower risk than e.g. the MLE.

The use of adaptive weights did not improve on the risk in this example, but it did result in the selection of sparser models. The example also showed that in this case the approximation \(|A|-1\) to the divergence was sufficiently accurate to be a computationally cheap alternative to the formula from Theorem 4.

A central idea in previous papers, Tibshirani and Taylor (2012), Zou, Hastie and Tibshirani (2007) and Meyer and Woodroofe (2000), was that
Lipschitz continuous functions are almost differentiable, which makes Stein’s lemma applicable. This is closely related to Rademacher’s Theorem stating that Lipschitz continuous functions are differentiable almost everywhere. The metric projection onto a closed convex set is Lipschitz continuous, and what remains for computing SURE is the computation of the divergence. We relied instead on the general fact that the metric projection onto any closed set is the derivative of a convex function. We then used Alexandrov’s theorem for convex functions to establish almost everywhere differentiability of the metric projection. This is in principle well known in the mathematical literature, and Asplund provided, for instance, only a brief argument in Asplund (1973) for what is close to being Theorem 1. However, we needed to clarify in what sense the metric projection is differentiable, and the precise relationship between pointwise derivatives Lebesgue almost everywhere and distributional derivatives for which partial integration applies. The original formulation of Alexandrov’s theorem was, in particular, stated as the existence of a quadratic expansion of a convex function $g$ for Lebesgue almost all $y$. This formulation does not require a definition of differentiability of $\nabla g$ in $y$ in cases where $\nabla g$ is not defined in a neighborhood of $y$. Consequently, the conclusion cannot be formulated in terms of $\nabla g$ alone. The more recent formulation of Alexandrov’s theorem as in Theorem 5 was useful, since it allowed us to formulate Theorem 1 in terms of differentiability properties of the metric projection itself rather than as a quadratic expansion of $\rho$.

There is an extensive mathematical literature on the uniqueness, and to some extent differentiability, of the metric projection – in particular in the infinite dimensional context. Haraux (1977) showed results on the directional differentiability of the metric projection onto a closed convex set in a Hilbert space. He showed, in particular, that in finite dimensions the projection onto a polytope is directionally differentiable in $y$ for all $y$ with the directional derivative being the projection onto

$$(y - \text{pr}(y))^\perp \cap T_{\text{pr}(y)}$$

where $T_{\text{pr}(y)}$ is the tangent cone, see Haraux (1977) for the details. This is a derivative if and only if it is linear, which happens if and only if $\text{pr}(y)$ is in the relative interior of the face $(y - \text{pr}(y))^\perp \cap K$. This is also the face of smallest dimension containing $\text{pr}(y)$. If we consider an $\ell_1$-ball with radius $s$, and the solution is unique with $p(s)$ nonzero parameters, the corresponding face has dimension $p(s) - 1$. This is a curious “dimension drop” – also derived in Section 3 by different arguments – when compared to the degrees of freedom, $p(\lambda)$, for $\ell_1$-penalized linear regression with regularization parameter $\lambda$, as derived in Zou, Hastie and Tibshirani (2007) and Tibshirani and
Hansen, N.R., Sokol, A. (2012). It is explainable as a consequence of computing the degrees of freedom for fixed $s$ and not fixed $\lambda$.

Haraux (1977) also showed in his Example 2 how to compute the derivative when the boundary of the set is $C^2$. The derivative is a form of regularized projection onto the tangent plane at $pr(y)$ – the regularization being determined by the curvatures. Abatzoglou derived a similar result in Abatzoglou (1978), but without assuming convexity. These results are closely related to Theorem 3, but we chose to downplay the differential geometric content. Instead, we focused on its relation to TIC.

More recent results on differentiability of the metric projection can be found in Rockafellar and Wets (1998). Their Corollary 13.43 gives an abstract result for a specific point, $y$, where $pr(y)$ is prox-regular w.r.t. $y - pr(y)$, and the result applies, in particular, when $K$ is fully amenable (regular enough). The result by Haraux on projections onto polytopes follows from this general result – see Example 13.44 in Rockafellar and Wets (1998).

The results of this paper suggest several directions for further research. If we study the set $K$ in greater detail for specific models, we might be able to compute or bound the contribution to the risk from the singular measures. Such bounds could, perhaps in combination with concentration of measure techniques, be used to establish novel bounds on the risk. One direction to go is to study the more refined results on pointwise differentiability of the metric projection close to $K$ under regularity conditions on $K$. Under the prox-regularity assumption on $pr(y)$, as mentioned above, it is shown in Poliquin, Rockafellar and Thibault (2000) that the metric projection is Lipschitz in a neighborhood of $pr(y)$. Thus in this neighborhood the singular measures are 0. This can be a path for bounding the risk if $\xi$ is close to $K$.

In addition to the theoretical directions one important direction of our future research will be to study more systematically the use of $\ell_1$-constrained least squares estimation of parameters in linear as well as nonlinear ODEs.

APPENDIX A: ALGORITHMS AND IMPLEMENTATION

The general implementation that computes $\ell_1$-penalized nonlinear least squares estimates, as well as the implementation of computations specifically related to linear ODEs are available in the R package smde. See http://www.math.ku.dk/~richard/smde/ for information on obtaining the R package and the R code used for the results reported in Section 4.

In the following sections we describe some of the technical results behind our implementation. In particular, the computation of derivatives related to the matrix exponential.
A.1. Differentiation of the matrix exponential. The map $A \rightarrow e^A$ is well known to be $C^\infty$ as a map from $\mathbb{M}(d,d)$ to $\mathbb{M}(d,d)$. Moreover, its first and second partial derivatives can be efficiently computed. We summarize a few useful results from the literature.

We denote by $L(A,F)$ the directional derivative of the matrix exponential in $A \in \mathbb{M}(d,d)$ in the general direction $F \in \mathbb{M}(d,d)$. It has the analytic integral representation

\begin{equation}
L(A,F) = \int_0^1 e^{(1-u)A} F e^{uA} \, du. \tag{16}
\end{equation}

See e.g. (10.15) in Higham (2008). If we use $\partial_{kl}$ to denote the partial derivative w.r.t. the $(k,l)'th$ entry, and if $E_{kl}$ denotes the $(k,l)'th$ unit matrix, we have $\partial_{kl} e^A = L(A,E_{kl})$. This gives the identity

\begin{equation}
\text{tr}(\partial_{kl} e^A M) = \text{tr} \left( E_{kl} \int_0^1 e^{uA} M e^{(1-u)A} \, du \right) = L(A,M)_{l,k}. \tag{17}
\end{equation}

for any $M \in \mathbb{M}(d,d)$. We will use this formula in the following section. Efficient algorithms exist for computing $L(A,F)$ for general matrices. It holds, for instance, that

\begin{equation*}
\exp \left( \begin{bmatrix} A & F \\ 0 & A \end{bmatrix} \right) = \begin{bmatrix} e^A & L(A,F) \\ 0 & e^A \end{bmatrix},
\end{equation*}

see (10.43) in Higham (2008), so if we can efficiently compute matrix exponentials, we can compute the derivative. The `expmFrechet` function in the `expm` R package, Goulet et al. (2012), implements a faster algorithm that avoids the dimension doubling.

For the second partial derivatives it follows from (16) that

\[
\partial_{hr} \partial_{kl} e^A = H(A,E_{hr},E_{kl}) + H(A,E_{kl},E_{hr}),
\]

where

\[
H(A,F,G) = \int_0^1 \int_0^u e^{(1-u)A} F e^{(u-s)A} G e^{sA} \, ds \, du.
\]

The computation of these iterated integrals is based on Theorem 1 in Van Loan (1978), which implies that

\[
\exp \left( \begin{bmatrix} A & F & 0 \\ 0 & A & G \\ 0 & 0 & A \end{bmatrix} \right) = \begin{bmatrix} e^A & L(A,F) & H(A,F,G) \\ 0 & e^A & L(A,G) \\ 0 & 0 & e^A \end{bmatrix}.
\]
From the integral representation of $H(A, F, G)$ we find that for $M \in \mathbb{M}(d, d)$

$$
\text{tr}(\partial_{hr}\partial_{kl}e^AM) = \text{tr}(E_{hr}H(A, E_{kl}, M)) + \text{tr}(E_{kl}H(A, E_{hr}, M)) = H(A, E_{kl}, M)_{r,h} + H(A, E_{hr}, M)_{l,k},
$$

which was used for the computation of the $J$ matrix that enters in the formula in Theorem 4.

**A.2. Coordinate descent algorithm and sufficient transformations.** To solve the optimization problem

$$
\min_\beta \|y - \zeta(\beta)\|^2_2 + \lambda \sum_{k=1}^{p} \omega_k |\beta_k|
$$

for a decreasing sequence of $\lambda$’s we have implemented a plain coordinate wise descent algorithm based on a standard Gauss-Newton-type quadratic approximation of the loss function. That is, for given $\beta \in \Theta$ we approximate the loss in the $k$'th direction as

$$
\|y - \zeta(\beta)\|^2_2 \simeq \|r(\beta) - \partial_k \zeta(\beta) \delta\|^2_2 = \|r(\beta)\|^2_2 - 2 \langle r(\beta), \partial_k \zeta(\beta) \delta \rangle + \|\partial_k \zeta(\beta)\|^2_2 \delta^2
$$

where $r(\beta) = y - \zeta(\beta)$. The coordinate wise penalized quadratic optimization problem can be solved explicitly, and we then iterate over the coordinates until convergence. We implemented two versions of the algorithm. Algorithm A is a generic algorithm that relies on two auxiliary functions for computing $\zeta(\beta)$ and $D\zeta(\beta)$. Algorithm B is specific to the example in Section 4. For this example with $m$ observations solving a $d$-dimensional linear ODE, the computation time for Algorithm A scales linearly with $m$, but the computation of $e^{tB}x$ and $De^{tB}x$ can be implemented to take advantage of sparseness of $B$, in which case the computation time does scale reasonably well with $d$, if $B$ is sparse. Algorithm B relies, on the other hand, on the precomputation of three sufficient statistics, being $d \times d$ matrices, as outlined below. Algorithm B cannot take the same advantage of a sparse $B$ and does not scale as well with $d$, but after the precomputation of the sufficient statistics, all other computation times are independent of $m$.

Since the loss is generally not convex, the steps may not be descent steps if the quadratic approximation is poor. We implemented Armijo backtracking as described in Tseng and Yun (2009) to ensure sufficient decrease and hence convergence.

As mentioned above, Algorithm B for the linear ODE example relies on sufficient statistics for the computation of the loss as well as the quadratic
approximation. We give here a brief derivation of the necessary formulas. On $M(d, d)$ the inner product can be expressed in terms of the trace,
\[ (A, B) = \text{tr}(A^T B). \]
The corresponding norm, often referred to as the Frobenius norm, is the ordinary 2-norm when matrices are identified with vectors in $\mathbb{R}^{d^2}$. For the linear ODE example, $\zeta(B) = e^{tB}x$, and
\[ \|y - \zeta(B)\|_2^2 = \text{tr}(yy^T) - 2\text{tr}(e^{tB}xy^T) - \text{tr}(e^{tB^T}e^{tB}xx^T), \]
which depends on the data through the three cross products $yy^T$, $xy^T$ and $xx^T$ only. These are $d \times d$ sufficient transformations. We also find that
\[ \langle r(B), \partial_{kl} \zeta(B) \rangle = \text{tr}(\partial_{kl} e^{tB}(y^T - xx^T e^{tB^T})), \]
by (17). Consequently, the entire gradient of the quadratic loss can be computed as $-2tL(tB, xy^T - xx^T e^{tB^T})^T$, which amounts to computing a single directional derivative of the exponential map.

We also need to compute inner products of the derivatives, $\partial_{kl} \zeta(B)$, of $\xi$, and to this end we observe that
\[ \langle \partial_{kl} \zeta(B), \partial_{hr} \zeta(B) \rangle = \text{tr}(x^T(\partial_{kl} e^{tB})^T \partial_{hr} e^{tB}x) \]
\[ = \text{tr}((\partial_{kl} e^{tB})^T \partial_{hr} e^{tB}xx^T) \]
\[ = t^2 L(tB^T, L(tB, E_{hr})xx^T)_{k,l}. \]
That is, an entire column (or row) of the matrix of inner products can be computed by computing two directional derivatives of the exponential map.

**A.3. Penalized vs. constrained optimization.** As mentioned above, our algorithms solve the penalized optimization problem for a given sequence of $\lambda$’s. A solution, $\hat{\beta}_\lambda$, for a given $\lambda$ is also a solution to the constrained optimization problem
\[ \min_{\beta \in \Theta_s(\lambda)} \|y - \zeta(\beta)\|_2^2 \]
where $s(\lambda) = \sum_{k=1}^p \omega_k |\hat{\beta}_{\lambda,k}|$ and
\[ \Theta_s = \left\{ \beta \left| \sum_{k=1}^p \omega_k |\hat{\beta}_k| \leq s \right. \right\}. \]
The value of $s(\lambda)$ is decreasing in $\lambda$. Thus the algorithm provides a sequence of solutions to the constrained problems for increasing values of $s$. If the sequence of $\lambda$’s is fixed, the sequence of $s$’s will, however, be random. This is a small nuisance in the simulation study where we want to compute the degrees of freedom repeatedly for a fixed $s$. In practice we have solved this by linear interpolation to compute $\hat{\text{Risk}}(s)$ for a fixed set of constraints $s$.

REFERENCES

Abatzoglou, T. J. (1978). The minimum norm projection on $C^2$-manifolds in $\mathbb{R}^n$. Trans. Amer. Math. Soc. 243 115–122. MR502897 (80a:58006)

Alexandrov, A. D. (1939). Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it. Leningrad State University Annals [Uchenye Zapiski] Mathematical Series 6 3-35.

Asplund, E. (1968). Fréchet differentiability of convex functions. Acta Math. 121 31–47. MR0231199 (37 #6754)

Asplund, E. (1973). Differentiability of the metric projection in finite-dimensional Euclidean space. Proc. Amer. Math. Soc. 38 218–219. MR0310150 (46 #9252)

Burnham, K. P. and Anderson, D. R. (2002). Model selection and multimodel inference, Second ed. Springer-Verlag, New York. A practical information-theoretic approach. MR1919620

Claeskens, G. and Hjort, N. L. (2008). Model selection and model averaging. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. MR2431297 (2009k:62005)

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. Journal of the American Statistical Association 99–467.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. Ann. Statist. 32 407–499. With discussion, and a rejoinder by the authors. MR2060166 (2005d:62116)

Evans, L. C. and Gariepy, R. F. (1992). Measure theory and fine properties of functions. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL. MR1158660 (93f:28001)

Evans, W. D. and Harris, D. J. (1987). Sobolev embeddings for generalized ridged domains. Proc. London Math. Soc. (3) 54 141–175. MR872254 (88b:46056)

Fremlin, D. H. (1997). Skeletons and central sets. Proc. London Math. Soc. (3) 74 701–720. MR1434446 (97m:54059)

Goulet, V., Dutang, C., Maechler, M., Firth, D., Shapira, M., Stadelmann, M. and expm-developers@lists.R-forge.R-project.org (2012). expm: Matrix exponential R package version 0.99-0.

Haraux, A. (1977). How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities. J. Math. Soc. Japan 29 615–631. MR0481060 (58 #1207)

Higham, N. J. (2008). Functions of Matrices: Theory and Computation. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Howard, R. (1998). ALEXANDROV’S THEOREM ON THE SECOND DERIVATIVES OF CONVEX FUNCTIONS VIA RADEMACHER’S THEOREM ON THE FIRST DERIVATIVES OF LIPSCHITZ FUNCTIONS. Unpublished lecture notes.

Hug, D., Last, G. and Weil, W. (2004). A local Steiner-type formula for general closed sets and applications. Math. Z. 246 237–272. MR2031455 (2005a:53130)
Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** 1083–1104. MR1810920 (2002c:62069)

Poliquin, R. A., Rockafellar, R. T. and Thibault, L. (2000). Local differentiability of distance functions. *Trans. Amer. Math. Soc.* **352** 5231–5249. MR1694378 (2001b:49024)

Rockafellar, R. T. (2000). Second-order convex analysis. *J. Nonlinear Convex Anal.* **1** 1–16. MR1751725 (2001c:49030)

Rockafellar, R. T. and Wets, R. J. B. (1998). *Variational analysis. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **317**. Springer-Verlag, Berlin. MR1491362 (98m:49001)

Rudin, W. (1991). *Functional analysis*, second ed. *International Series in Pure and Applied Mathematics*. McGraw-Hill Inc., New York. MR1157815 (92k:46001)

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR630098 (83a:62080)

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

Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.* **117** 387–423. MR2421312 (2009g:90075)

Van Loan, C. F. (1978). Computing integrals involving the matrix exponential. *IEEE Trans. Automat. Control* **23** 395–404. MR0494865 (58 #13648)

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131. MR1614596 (99a:62097)

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 (2008d:62024)

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967 (2009d:62096)

E-mail: Niels.R.Hansen@math.ku.dk

E-mail: alexander@math.ku.dk

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF COPENHAGEN
UNIVERSITETSPARKEN 5
2100 COPENHAGEN
DENMARK