Model Selection and Risk Estimation with Applications to Nonlinear Ordinary Differential Equation Systems

PhD thesis

Frederik Vissing Mikkelsen

Department of Mathematical Sciences
University of Copenhagen

This thesis has been submitted to the PhD School of the Faculty of Science,
University of Copenhagen.

Frederik Vissing Mikkelsen
Email: frm@math.ku.dk

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

Academic advisor:	Prof. Niels Richard Hansen, University of Copenhagen.
Assessment committee:	Prof. Michael Stumpf, Imperial College London; Assoc. Prof. Ryan Tibshirani, Carnegie Mellon University; Prof. Carsten Wiuf, University of Copenhagen.
Submission date:	October 31, 2017.
ISBN:	978-87-7078-901-1

Preface

The following thesis was submitted as part of the formal requirements to obtain a PhD degree from the University of Copenhagen. The work was carried out from November 2014 to October 2017. Most of my research took place at the Department of Mathematical Sciences, University of Copenhagen. Additionally, an inspiring stay as visiting researcher at Stanford University had a significant impact on the content.

I would like to express my genuine gratitude to my PhD advisor, Prof. Niels Richard Hansen, for guiding me through three tough, but exciting years and for helping me see the various facets of research, communication and teaching. Your endless support and feedback has been invaluable. Furthermore, I am grafeful to Prof. Jonathan Taylor for inviting me to stay at the Deparment of Statistics at Stanford University, for giving me the opportunity to experience their world leading scientific environment and for the many insightful discussions.

To all my fellow PhD candidates and postdocs in Copenhagen, thank you for creating such a great atmosphere. Special thanks go to my office mates: Niels Olsen, Emil Jørgensen, Mads Raad, Adam Lund, Rune Christiansen and Kang Li. Thank you for great times on and off campus and for all the small everyday chats. Also warm thanks to Roman Croessmann, Charles Zheng and Chaojun Wang for great times at Stanford.

Last but not least, my deepest thanks go to my family for all your support and in particular to Mia for her endless love and encouragement. Without you this work would have little to no meaning.

Frederik Vissing Mikkelsen

Abstract

Broadly speaking, this thesis is devoted to model selection applied to ordinary differential equations and risk estimation under model selection. A model selection framework was developed for modelling time course data by ordinary differential equations. The framework is accompanied by the R software package, *episode*. This package incorporates a collection of sparsity inducing penalties into two types of loss functions: a squared loss function relying on numerically solving the equations and an approximate loss function based on inverse collocation methods. The goal of this framework is to provide effective computational tools for estimating unknown structures in dynamical systems, such as gene regulatory networks, which may be used to predict downstream effects of interventions in the system. A recommended algorithm based on the computational tools is presented and thoroughly tested in various simulation studies and applications.

The second part of the thesis also concerns model selection, but focuses on risk estimation, i.e., estimating the error of mean estimators involving model selection. An extension of Stein's unbiased risk estimate (SURE), which applies to a class of estimators with model selection, is developed. The extension relies on studying the degrees of freedom of the estimator, which for a broad class of estimators decomposes into two terms: one ignoring the selection step and one correcting for it. The classic SURE assumes that the estimator in question is almost differentiable and it therefore only accounts for the first term of the decomposition. In order to account for the second term the continuum of models arising when the selection procedure has a tuning parameter is studied. By exploiting the duality between varying the tuning parameter for fixed observations and perturbing the observations for fixed tuning parameter, an identity is derived for a class of estimators which support the extension of SURE. The resulting corrected version of SURE is generally fast to compute and for the lasso-OLS estimator it shows promising results when compared to risk estimation via cross validation.

Resumé

Denne afhandling omhandler modelselektion i ordinære differentialligninger og *risk*-estimation under modelselektion. Vi etablerer et estimationssetup for ordinære differentialligninger hvori modellen både selekteres og estimeres. Dette setup er understøttet af R pakken *episode*, som inkorporerer en familie af penaliseringsmetoder der inducerer modelselektion. Disse penaliseringsmetoder kan anvendes på to typer af inferens: en baseret på numeriske løsninger af differentialligningerne og en baseret på en approksimation ved hjælp af *inverse collocation*. Pakken indeholder værktøjerne til at estimere ukendte strukturer i dynamiske systemer, som for eksempel genregulatoriske netværk, og gør det derved muligt at prædiktere interventionseffekter i systemet. En konkret algoritme baseret på disse værktøjer præsenteres og anvendes, samt gennemtestes i en række simulationsstudier.

Den anden halvdel af afhandlingen tager også udgangspunkt i modelselektion, men fokus er flyttet til risk-estimation, altså estimation af fejlen for middelværdiestimatorer som involverer modelselektion. En udvidelse af Stein's unbiased risk estimat (SURE), som kan anvendes på en række estimatorer med indbygget modelselektion, etableres. Denne udvidelse beror på frihedsgraderne for estimatoren, som kan dekomponeres i to led: ét som ignorerer modelselektionen og ét som korrigerer for den. Den klassiske SURE har den stående antagelse at estimatoren er næsten differentiabel og tager derfor kun højde for det første led. For at estimere det andet led er det nøvendigt at nærstudere det kontinuum af modeller der opstår når modelselektionen afhænger af en tuningparameter. Ved at udnytte den dualitet der er imellem at justere tuningparameteren for fast observation og perturbere observationen for fast tuningparameter, opnår vi en ligning som holder for en række modelselektionsmetoder. Den korrigerede udgave af SURE er relativ hurtig at evaluere og for lasso-OLS estimatoren viser den lovende resultater sammenlignet med risk-estimation ved hjælp af krydsvalidering.

Contents

1 Introduction

Planetary motions, changes in atmospheric pressure, raindrops falling to the ground, cellular dynamics and chemical reactions. These are all concrete examples of dynamical systems; they exist all around us, ranging from macroscopic to microscopic scales. Dynamical systems involve concrete or abstract objects, which evolve over time according to some fixed "rules". The focus of this thesis is twofold: to develop a *model selection* framework for ordinary differential equations (a particular type of dynamical systems) and to quantify and estimate the *risk* in model selection.

To motivate the first, consider *gene regulatory networks* in systems biology, which we briefly outline: Within living cells segments of DNA are copied into mRNA (messenger ribonucleic acid) sequences by the enzyme RNA polymerase. This is called the *transcription* process. The mRNA sequences reaching the ribosomes determine the production of *gene products* (RNA and proteins). This is called the *translation* process. Some of these gene products, as well as other molecular regulators within the cell, may in turn regulate the transcription of the mRNA sequences from their respective site on the DNA. The system evolves over time according to these internal and external gene regulations ([5]).

While the uncertainty of planetary motions and the changes in atmospheric pressure can primarily be ascribed to unknowns in the state of the system, the uncertainty in gene regulatory networks is of a different nature — it is structural. The primary challenge is identifying which genes influence the transcription of given mRNA sequences (through their associated gene products). The first point of interest in this thesis is therefore learning ODE systems with unknown structures.

The second statistical topic addressed in this thesis is risk estimation, i.e., estimation of the error of an estimator. The framework considered is risk estimation of mean estimators under a squared loss. This is the framework in which Steins unbiased risk estimate (SURE) applies, provided that the mean estimator is *almost differentiable* (see [7] and [19]). While all almost differentiable estimators are continuous, the results in this thesis also apply to certain mean estimators with discontinuities. Discontinuous estimators are rather common in data adaptive model selection and the results are therefore particularly suited for risk estimation under model selection.

1.1 Systems of ordinary differential equations

Ordinary differential equation (ODE) systems are continuous dynamical systems with the simplifying assumption that the infinitesimal change of a state variable x over time t only depends on the current position of x. Mathematically, let $x(t)$ denote the position of the state, $x \in \mathbb{R}^d$, at time t. Then $x : \mathbb{R} \to \mathbb{R}^d$, viewed as a function of t, is a solution to an ODE if

$$\frac{dx}{dt} = f(x), \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a smooth d-dimensional field. Much mathematical research has been put into showing existence and uniqueness of solutions to (1.1) under various conditions, see e.g., [15] and [18]. However, existence and uniqueness of the systems in question will be given little attention in this thesis. Most attention will be directed at the use of ODEs in statistical modelling and model selection. The following subsections motivate the use of ODEs for modelling aggregated systems.

1.1.1 Aggregated dynamics

The use of ordinary differential equations for modelling dynamical systems is often motivated by some underlying physical laws governing the observed system. Even though thermodynamic laws dictates that certain biological or chemical systems follow stochastic (partial) differential equations, the dynamic is often reduced to follow that of ordinary differential equations with no intrinsic noise. This reduction is typically justified through an aggregation argument. Two of such are outlined below: one focusing on aggregating the dynamics of chemical combonents in a population of similarly behaving cells and another one focusing on chemical kinetics in a well stirred solution in which the number of molecules is large. A reduction to deterministic ODE systems simplifies the inference and considerably reduces the computational burden.

Aggregation of cellular dynamics

Reduction of stochastic cell dynamics to deterministic systems is considered in [14]. Here the dynamic of a single cell is assumed to follow the dynamic of a stochastic delay differential equation (SDDE)

$$dX = f(\mathcal{F}_X)dt + \sigma(\mathcal{F}_X)dB_t, \tag{1.2}$$

where X is a d-dimensional process, $\mathcal{F}_X(t) = \{X(s)|s \leq t\}$ denotes its history and f and σ are drift and diffusion functions, respectively. Realistically, the experimental setup only allows for observing the aggregate population of cells. Luckily, the diffusion term of the average of independent realisations of (1.2) vanishes as the population size increases and thus the aggregate population dynamics follow a deterministic delayed differential equation.

However, as pointed out in [14] unless the drift and expectation commute ($E(f(\mathcal{F}_X)) = f(\mathcal{F}_{E(X)})$) the aggregate population dynamics will not share the same drift as that of the

individual cells. Most nonlinear drifts do not commute with the expectation and thus (1.2) might need to be simplified by a linearisation.

Linear noise approximation in chemical kinetics

Another approach of reducing stochastic systems to deterministic systems is the linear noise approximation (LNA) used in chemical kinetics, as described in [22]. On molecular level the abundances of d chemical species (e.g., H_2O, NaCl, etc) in a well stirred solution are assumed to follow a d-dimensional jump markov process, N_t.

Among these chemical species R reactions drive the transition of one set of species (the reactants) to another (the products), e.g., $HCl + NaOH \to H_2O + NaCl$. We let $v_r \in \mathbb{R}^d$ denote the net change in number of molecules, given that reaction r takes place. The jump intensity, λ_r, of reaction r depends on the abundance of its reactants prior to the time of reaction. Let $P(n, t|N_0)$ denote the probability distribution of $N_t = n$ given the initialisation $N(0) = N_0$. The Kolmogorov forward equations then states that:

$$\frac{\partial P(n,t|N_0)}{\partial t} = \sum_{r=1}^{R} \lambda_r(n - v_r)P(n - v_r, t|N_0) - \lambda_r(n)P(n, t|N_0) \tag{1.3}$$

The linear noise approximation relies on an increasing total number of molecules, which we assume is proportional to the volume of the solution, Ω. The concentration process $X = N/\Omega$ then assymptotically follows the distribution of $x + \xi/\sqrt{\Omega}$, as $\Omega \to \infty$ (for details see [22]). Here x follows a deterministic ordinary differential equation (ODE) given by the field and initial conditions:

$$\frac{dx}{dt} = \sum_{r=1}^{R} v_r \gamma_r(x), \qquad x(0) = x_0 \tag{1.4}$$

where $\gamma_r(x) \coloneqq \lambda_r(\Omega x)/\Omega$ are the rate functions and the limit $x_0 = \lim_{\Omega\to\infty} N_0/\Omega$ is assumed to exist. The noise process then follows a time-inhomogenous Ornstein Uhlenbeck process

$$d\xi_t = \left(\sum_{r=1}^{R} v_r \partial\gamma_r(x_t)\right)\xi_t dt + \sum_{r=1}^{R} v_r \sqrt{\gamma_r(x_t)} dB_t^r, \qquad \xi(0) = 0 \tag{1.5}$$

where $(B^r)_{r=1}^R$ are i.i.d. brownian motions. Thus, ξ is a zero mean Gaussian process with a computationally tractable time-dependent spatial covariance structure. However, the covariance structure is only analytically available if the matrices $(v_r \partial\gamma_r(x))_{r,x}$ commute. Conclusively, as $\Omega \to \infty$, the concentration process X is the sum of a solution to an ODE and a Gaussian noise process.

Mass action kinetics refers to a class of polynomial ODE systems on the form (1.4), where γ_r is a monomial for each r. Each index r represents a chemical reaction on the form:

$$a_1 X_1 + ... + a_d X_d \to b_1 X_1 + ... + b_d X_d, \tag{1.6}$$

where $(a_i)_{i=1}^d$ and $(b_i)_{i=1}^d$ are non-negative integers, called the *stoichiometric coefficients*. The net change of molecules due to reaction (1.6) is $v = (a_i - b_i)_{i=1}^d$. For each reaction $r = 1, ..., R$ let $A_r \in \mathbb{N}^d$ denote the reactant stoichiometric coefficients and let $B_r \in \mathbb{N}^d$ denote the product stoichiometric coefficients. In mass action kinetics the rate function γ_r for reaction r is given by the monomial $\gamma_r(x) = k_r x^{A_r} = k_r \prod_{i=1}^d x_i^{A_r(i)}$ for some non-negative rate constant $k_r \geq 0$. The full system thus reads:

$$\frac{dx}{dt} = \sum_{r=1}^{R} k_r (A_r - B_r) x^{A_r}, \qquad x(0) = x_0. \tag{1.7}$$

Extending the above chemical kinetics framework to include rational rate functions is often required when modelling gene regulatory networks. These rate functions are also called *hill functions* or *hill-type functions*. The rational form typically appears when reducing a larger mass action kinetics model with latent coordinates to the smaller system consisting of the observed coordinates only. This reduction is carried out through a quasi-stationary approximation, in which certain reactions are assumed to equilibrise instantly, see e.g., [16].

1.1.2 Learning ODE systems

We consider a parametrised d-dimensional ODE given by:

$$\frac{dx}{dt} = f(x, \theta), \qquad x(0) = x_0 \tag{1.8}$$

with initial condition $x_0 \in \mathbb{R}^d$ and a smooth function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$. This thesis primarily focuses on parametric frameworks, however non-parametric methods for learning ODE systems is also an active area of research, most notably are the methods by [4] and [23].

Let $(y_i)_{i=1}^n$ be noisy observations at discrete time points $(t_i)_{i=1}^n$. Inference based on the squared loss

$$\ell_y(\theta) = \frac{1}{2} \sum_{i=1}^{n} \|y_i - x(t_i)\|_2^2, \text{ subjected to } x(t_i) - x_0 = \int_0^{t_i} f(x(s), \theta)\, ds, \text{ for } i = 1, ..., n \tag{1.9}$$

is a classic approach ([1], [13]) and is refered to as *the least squares method* or *the gold standard method*. Note that the constraint in (1.9) implies that the solution curve $x(t)$ depends on $\theta \in \mathbb{R}^p$ and $x_0 \in \mathbb{R}^d$. The least squares method has a straight forward interpretation: it favours mean structures within the ODE class, (1.8), which provide small empirical variances.

Evaluating the loss function requires solving the ODE system, which in most nonlinear systems are carried out using a numerical solver (see e.g. [17] for a comprehensive overview). If ℓ_y is optimised using gradient based methods then the computational burden primarily lies in evaluating the derivatives of x with respect to θ and x_0. This is because the derivatives $z_\theta = \partial_\theta x$ and $z_{x_0} = \partial_{x_0} x$ are solutions to even larger ODE systems on $\mathbb{R}^{d \times p}$ and $\mathbb{R}^{d \times d}$, respectively. These ODEs are called the *sensitivity equations* and are given by:

$$\begin{aligned} \frac{dz_\theta}{dt} &= \partial_x f(x, \theta) z_\theta + \partial_\theta f(x, \theta), \quad z_\theta(0) = \mathbf{0}_{d \times p}, \\ \frac{dz_{x_0}}{dt} &= \partial_x f(x, \theta) z_{x_0}, \quad z_{x_0}(0) = I_d, \end{aligned} \tag{1.10}$$

where $\mathbf{0}_{d\times p}$ denotes the $d\times p$-dimensional 0-matrix and I_d is the d-dimensional identity matrix.

Avoiding this computational burden is a motivation for considering the *inverse collocation methods*, which completely avoid numerically solving the ODE systems. The name is derived from *collocation methods* in numerical analysis, which refer to a class of methods for solving ODE systems numerically. It goes as follows; in the ODE system

$$\frac{dx}{dt} = f(x(t), \theta), \qquad x(0) = x_0, \tag{1.11}$$

the parameter vector θ is assumed known. Moreover, a finite set of collocation points $\mathcal{C} \subseteq \mathbb{R}$ are chosen, as well as a vector space $\mathcal{V}$ of functions on $\mathbb{R}$ with values in $\mathbb{R}^d$. A numerical solution, $\tilde{x} \in \mathcal{V}$, is sought that minimises the distance $||\frac{d\tilde{x}}{dt}(t) - f(\tilde{x}(t), \theta)||$ between $\left(\frac{d\tilde{x}}{dt}(t)\right)_{t\in\mathcal{C}}$ and $(f(\tilde{x}(t), \theta))_{t\in\mathcal{C}}$. Typically, $\mathcal{V}$ consists of functions with coordinates in $\text{span}(\varphi_j)$ for a choice of finitely many basis functions $\varphi_j : \mathbb{R} \to \mathbb{R}$ and the norm is the 2-norm on $\mathbb{R}^d$. Collocation methods thus solve the forward problem of computing the solution of (1.11) for a known θ.

By *inverse collocation methods* we refer to a class of estimators of the parameter θ, given the observed trajectory x, that solve the inverse problem using the collocation idea. These methods exist in many versions ([2], [3], [6], [9], [12], [21]) and are known under many other names, e.g., *gradient matching*, *trajectory matching* or *smooth-and-match estimators*. However, they all rely on the same two-step procedure: 1) approximate the data, y, by an element in $\mathcal{V}$ to get an estimate of the full trajectory $\hat{x}$; 2) base the estimation of θ on the trajectory $\hat{x}$ as if it was the true trajectory, by minimising the distance between the position, gradient or integral at a given set of collocation points. Typically, $\hat{x}$ is obtained as a smoother or an approximation of y via a basis expansion.

One example of an inverse collocation method is the *gradient matching method* ([2], [21]), which minimises the approximate loss function:

$$\frac{1}{2}\sum_{t\in\mathcal{C}} \left\| \frac{d\hat{x}}{dt}(t) - f(\hat{x}(t), \theta) \right\|_2^2, \tag{1.12}$$

where the coordinates of $\hat{x}$ are approximated via a basis expansion. Though the *inverse collocation methods* are faster to compute, the resulting estimates of θ may still strongly depend on y through the choice of smoothing scheme $y \mapsto \hat{x}$.

Each parameter coordinate in (1.8) corresponds to a smaller component in the ODE structure. For instance, in the mass action kinetics framework each rate parameter k_r encode the rate of a single reaction. And more importantly, that reaction is only present in the system if k_r is strictly positive. If little to no prior knowledge of the system exists, the parameter space must be large enough to provide a sufficiently flexible ODE system. Choosing the level of flexibility is up to the data analyst. For instance in a mass action kinetics framework, one could choose the set of all *catalytic* reactions, i.e.,

$$X_s + X_c \to X_p + X_c$$

with $s, c, p \in \{1, ..., d\}$ and $s \neq c \neq p$. The experimental setting or underlying theory may reduce this set of reactions to those that are physically possible, or perhaps suggest a more appropriate set of reactions.

In this parametric framework, learning the structure of the dynamics of x from noisy observations thus amounts to sparse estimation of θ. There exist a vast selection of tools for sparse estimation, particularly within linear regression. One approach — which is used in this thesis — is to include a sparsity enforcing penalty function in the loss function. Examples of such penalties include: lasso ([20]), elastic net ([25]), smoothly clipped absolute deviation (SCAD, by [8]) or minimax concave penalty (MCP, by [24]).

1.2 Risk estimation under model selection

In risk estimation we are concerned with the following problem: let Y be an n-dimensional random variable with finite expectation $\mu \in \mathbb{R}^n$. For an estimator of the mean vector $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}^n$, which given the observation Y returns the estimate of μ, we define the *risk* of $\hat{\mu}$ as

$$\text{Risk}(\hat{\mu}) \coloneqq E\|\mu - \hat{\mu}(Y)\|_2^2, \tag{1.13}$$

provided that $\hat{\mu}(Y)$ has finite second moment. The risk of an estimator is a measure of how close its estimates are to the true mean vector on average. Risks relying on other loss functions than the squared loss also exist, but are not considered here.

If Y also has finite second moment we can expand the squared norm in (1.13) and obtain:

$$\text{Risk}(\hat{\mu}) = E\|Y - \hat{\mu}(Y)\|_2^2 - \sum_{i=1}^{n} VY_i + 2\sum_{i=1}^{n} \text{cov}(Y_i, \hat{\mu}_i(Y)). \tag{1.14}$$

The first term is the expected *residual sum of squares*, the second term only depends on the marginal variances of Y and the last term measures the linear dependence between Y and its fitted values under $\hat{\mu}$.

Cross validation and bootstrapping are two general methods for estimating risks. In cross validation, the data is split into folds and the estimator is trained on subsets of the folds (assuming that it is possible) and validated on the remaining folds. In bootstrapping the risk is estimated by applying the estimator to replicated data drawn from some distribution, which appropriately resemples the distribution of Y. Both methods have their advantages and drawbacks, but both rely on applying the estimator multiple times, which may be computationally intensive. In this thesis we will focus on an alternative risk estimator, the computationally lighter Stein's unbiased risk estimator (SURE, see [7] and [19]).

1.2.1 Degrees of freedom and Steins unbiased risk estimate

Assuming that $VY_i = \sigma^2$ for each $i = 1, ..., n$, (1.14) reads

$$\text{Risk}(\hat{\mu}) = E\|Y - \hat{\mu}(Y)\|_2^2 - n\sigma^2 + 2\sigma^2 \text{df}(\hat{\mu}), \tag{1.15}$$

where

$$\text{df}(\hat{\mu}) \coloneqq \sum_{i=1}^{n} \frac{\text{cov}(Y_i, \hat{\mu}_i(Y))}{\sigma^2} \tag{1.16}$$

are the *degrees of freedom* of the estimator.

Consider the example of an estimator $\hat{\mu} = \Pi_V$, which is the orthogonal projection onto some fixed subspace V of $\mathbb{R}^n$. This example includes all ordinary least squares estimators in regression with pre-specified design matrices. If the coordinates of Y are uncorrelated, then the degrees of freedom of $\hat{\mu}$ is

$$\mathrm{df}(\Pi_V) = \frac{\mathrm{tr}(\mathrm{Cov}(Y, \Pi_V Y))}{\sigma^2} = \mathrm{tr}(\Pi_V) = \dim(V), \tag{1.17}$$

where tr denotes the trace operator. This example shows that (1.16) coincides with the usual terminology from linear models that the degrees of freedom equals the dimension of predictor space. It is important to stress that the intuitive notion of degrees of freedom as the dimension of the "fitting" space does not extend to nonlinear estimators. In fact even if Y is Gaussian, degrees of freedom are not guaranteed to be finite, see e.g. [10] for examples and a discussion of this.

Assuming that Y is Gaussian with uncorrelated coordinates, i.e., $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, one can recover the degrees of freedom from the local behaviour of $\hat{\mu}$ if it is *almost differentiable*. By Definition 1 in [19] almost differentiability means that for each coordinate function $\hat{\mu}_i : \mathbb{R}^n \to \mathbb{R}$ there exists a function $g_i : \mathbb{R}^n \to \mathbb{R}^n$ such that for all $z \in \mathbb{R}^n$

$$\hat{\mu}_i(y + z) - \hat{\mu}_i(y) = \int_0^1 \langle z; g_i(y + tz) \, dt, \tag{1.18}$$

for almost all $y \in \mathbb{R}^n$. Almost differentiable functions are tightly linked to *Sobolev functions* — functions for which weak derivatives exist having this partial integration property. The function g_i is essentially the weak derivative of $\hat{\mu}_i$ and we let $\partial_i \hat{\mu}_i$ denote the i^{th} coordinate of g_i. With these we define the *divergence* of $\hat{\mu}$:

$$\mathrm{div}(\hat{\mu}) \coloneqq \sum_{i=1}^{n} \partial_i \hat{\mu}_i. \tag{1.19}$$

If the estimator is almost differentiable and $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, then Stein's Lemma (Lemma 2 in [19]) yields

$$\mathrm{df}(\hat{\mu}) = E(\mathrm{div}(\hat{\mu})). \tag{1.20}$$

We refer to $\mathrm{df_S} \coloneqq E(\mathrm{div}(\hat{\mu}))$ as *Stein's degrees of freedom*. Combined with (1.15) it yields Stein's unbiased risk estimate (SURE):

$$\mathrm{SURE} \coloneqq \|Y - \hat{\mu}(Y)\|_2^2 - n\sigma^2 + 2\sigma^2 \mathrm{div}(\hat{\mu})(Y). \tag{1.21}$$

The divergence operator originates from physics and has an interesting interpretation: for a smooth field $f : \mathbb{R}^n \to \mathbb{R}^n$, define the flow $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ by

$$F(t, x) = x + \int_0^t f(F(s, x)) \, ds, \tag{1.22}$$

which has the property $\partial_t F(0,x) = f(x)$. For a compact borel measurable set $\Omega \subseteq \mathbb{R}^n$ let $\mathrm{vol}_F(t,\Omega) \coloneqq \mathcal{L}^n(F(t,\Omega))$ denote its volume. Here $\mathcal{L}^n$ denotes the Lebesgue measure on $\mathbb{R}^n$. By the Change of Variable Theorem

$$\mathrm{vol}_F(t,\Omega) = \int_\Omega JF(t,F(-t,z))\ dz, \tag{1.23}$$

where $JF \coloneqq |\det(\partial_x F)|$ denotes the Jacobian of $x \mapsto F(t,x)$. Standard results from analysis and linear algebra yields

$$\begin{aligned} \det(\partial_x F(t,x)) &= \det\left((\partial_x F)(0,x) + t\partial_t\partial_x F(0,x) + t\varepsilon(t,x)\right) \\ &= \det\left(I + t\partial_x f(x) + t\varepsilon(t,x)\right) \\ &= 1 + t\cdot\mathrm{tr}(\partial_x f(x) + \varepsilon(t,x)) + \mathcal{O}(t^2), \end{aligned} \tag{1.24}$$

for some ε vanishing at $t=0$. Hence $\mathrm{div}(f)(x) = \mathrm{tr}(\partial_x f(x)) = \partial_t JF(0,x)$, which in terms of (1.23) shows

$$\partial_t \mathrm{vol}_F(0,\Omega) = \int_\Omega \mathrm{div}(f)(z)\ dz. \tag{1.25}$$

In other words, $\mathrm{div}(f)$ measures the infinitesimal change in volumes due to moving them via the flow F. In light of this interpreation, the intuition behind the expectation $E(\mathrm{div}(\hat{\mu})(Y))$ is that it measures of the average strength of the dependence between Y and $\hat{\mu}(Y)$, if $\hat{\mu}$ is sufficiently smooth.

In order for (1.20) to hold the almost differentiability requirement is absolutely essential, as the main argument in proving (1.20) is a partial integration argument. If for instance the estimator is discontinuous, which is often the case for data adaptive model selection, the partial integration argument no longer applies and the identity likely breaks. The second goal of this thesis is therefore to characterise the difference $\mathrm{df}(\hat{\mu}) - \mathrm{df}_{\mathrm{S}}(\hat{\mu})$ for a class of discontinuous estimators $\hat{\mu}$. Furthermore, the goal is also to investigate if such a characterisation can be used to extend SURE to estimators involving data adaptive model selection.

2 Summaries and contributions

The thesis consists of five research papers written during my PhD studies at University of Copenhagen in the time period from November 2014 until October 2017. To clear out any potential confusion, during this period I changed my name from Frederik Riis Mikkelsen to Frederik Vissing Mikkelsen. Each of the five papers can be read independently and their respective titles and summaries are as follows:

I **Learning Large Scale Ordinary Differential Equation Systems**

This is the primary paper concerning model selection in ODE systems. Here we present the following modelling framework: assume that time course data, y, is drawn from a finite set of environments $\{1, ..., E\}$ representing the interventions, i.e., $y = (y^e)_{e=1}^E$ with y^e consisting of n_e time points in environment e. With similarly organised weights $w = (w^e)_{e=1}^E$, we consider the loss function

$$\ell_y(\theta) \coloneqq \frac{1}{2} \sum_{e=1}^{E} \sum_{i=1}^{n_e} \sum_{l=1}^{d} w_{i,l}^e \left(y_l^e(t_i) - x_l^e(t_i, \theta \circ c_e) \right)^2 + \lambda \sum_{j=1}^{p} v_j \text{pen}(\theta_j), \tag{2.1}$$

where $\text{pen}(\cdot)$ is a sparsity enforcing penality function with coordinate-wise weights $(v_j)_j$ and x^e solves (1.8). We assume that the effective parameter $\theta \circ c_e$ in environment e is a hadamard product of the baseline parameter θ and the environment-specific scales c_e. This framework is used in the accompanying R package *episode* and it allows for different types of time course data, including both perturbed and intervened data, to enter the estimation procedure simultaneously.

We propose the adaptive integral matching (AIM) algorithm, which first employes an inverse collocation method to produce initial parameter estimates for a family of smoothers. The initial estimates are then used to adapt weights and scales in (2.1) and finally the initial estimates are refitted by passing them as initialisations in minimising (2.1). Through extensive simulation studies AIM shows strong network and reaction recovery in both mass action kinetics systems and a full scale model of glycolysis in *Saccharomyces cerevisiae*. Furthermore, AIM proves state-of-the-art network recovery for the *in silico* phosphoprotein abundance data from the eighth DREAM challenge.

The supplementary material referenced in the paper can be found at https://arxiv.org/abs/1710.09308.

II **Computational Aspects of Parameter Estimation in Ordinary Differential Equation Systems**

This short paper primarily concerns the computational aspects of fitting ODE systems to time course data. The most popular methods are outlined and discussed and a combination of two of these methods is proposed. In this combined approach the computationally light integral matching method is used to propose descent directions for the more computationally intensive least squares method. Consequently, the combined method enjoys the statistical properties of the least squares approach while being faster. These results partly laid the foundation for reducing the computational burden in the R package *episode*.

III **A Model Based Rule for Selecting Spiking Thresholds in Neuron Models**

This paper addresses the problem of selecting spiking thresholds in single neuron ODE based models. When modelling the electrical potential of neurons via ODE systems, the electrical potential exhibits all its local maxima on the null cline of the system. The magnitude of these local maxima are exactly what determines the spiking behaviour of the neuron. In this paper approximate separatrices mimicking the 'all-or-none' principle of neuron spiking are found by identifying the points on the null cline with maximal divergence. We propose spiking thresholds obtained by projecting these maximisers onto the coordinate representing the electrical potential. The proposed method is applied to six different single neuron ODE models and yields spiking thresholds in line with those typically used in practice. Besides dealing with ODE systems in a neuron modelling framework, this paper also relates to the rest of the thesis in its use and interpretation of the divergence operator. This operator is essential in the papers concerning risk estimation.

IV **Degrees of Freedom for Piecewise Lipschitz Estimators**

In this paper we develop a representation of the difference $\mathrm{df}(\hat{\mu}) - \mathrm{df}_{\mathrm{S}}(\hat{\mu})$ for a wide class of mean estimators $\hat{\mu}$. This class concerns estimators which can be written on the form

$$\hat{\mu} = \sum_{i=1}^{N} 1_{U_i} \hat{\mu}_i, \tag{2.2}$$

where $(U_i)_{i=1}^N$ is a finite collection of disjoint open sets on $\mathbb{R}^n$ with $\bigcup_i \overline{U}_i = \mathbb{R}^n$ and each $\hat{\mu}_i : \overline{U}_i \to \mathbb{R}^n$ is locally Lipschitz. An estimator on the form (2.2) is allowed to be discontinuous, in the sense that $\hat{\mu}_i$ and $\hat{\mu}_j$ may disagree on their common boundary $\overline{U}_i \cap \overline{U}_j$ for $i \neq j$. This class fits perfectly in the framework of data adaptive model selection; the regions $(U_i)_i$ represent selection events and the estimators $(\hat{\mu}_i)_i$ represent the estimation procedures conditional on the selection, i.e., the *post-selection* estimators.

Under a weak set of regularity conditions one can establish

$$\mathrm{df}(\hat{\mu}) = \mathrm{df}_{\mathrm{S}}(\hat{\mu}) + \frac{1}{2} \sum_{i \neq j} \int_{\partial U_i \cap \partial U_j} \psi \langle \eta_i ; \hat{\mu}_j - \hat{\mu}_i \rangle \, d\mathcal{H}^{n-1}, \tag{2.3}$$

where ψ denotes the density function of $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, η_i denotes the unit outer normal to ∂U_i and $\mathcal{H}^{n-1}$ is the $n-1$-dimensional Hausdorff measure. On its own (2.3)

yields no immediate way of estimating the quantity $\mathrm{df}(\hat{\mu}) - \mathrm{df_S}(\hat{\mu})$ for a single realisation y of Y. However, it is possible to construct an estimator of $\mathrm{df}(\hat{\mu}) - \mathrm{df_S}(\hat{\mu})$ using (2.3) for the lasso-OLS estimator in regression. The lasso-OLS, $\hat{\mu}^{\lambda}_{\mathrm{l-OLS}}$, applies the OLS estimator restricted to the predictors chosen by the lasso estimator ([20]) with tuning parameter $\lambda > 0$. For this specific estimator we derive the identity

$$\mathrm{df}(\hat{\mu}^{\lambda}_{\mathrm{l-OLS}}) = \mathrm{df_S}(\hat{\mu}^{\lambda}_{\mathrm{l-OLS}}) - \lambda \partial_{\lambda} \mathrm{df_S}(\hat{\mu}^{\lambda}_{\mathrm{l-OLS}}). \tag{2.4}$$

The Steins degrees of freedom for the lasso-OLS estimator, $\mathrm{df_S}(\hat{\mu}^{\lambda}_{\mathrm{l-OLS}})$, equals the expected dimension of the space selected by the lasso estimator. Using this and (2.4) we develop a corrected version of SURE. In an extensive simulation study we show that tuning the lasso-OLS estimator using the corrected SURE yields estimates which, compared to 5- and 10-fold cross validation, are close to the true mean.

V Extending 'SURE' to Estimators with Data Adaptive Model Selection via Flows

In this paper we seek to extend SURE to other data adaptive model selection procedures than lasso-OLS. This is achieved by establishing a framework in which identities similar to (2.4) can be obtained. We assume that the estimator $\hat{\mu}^t$ depends on a tuning parameter $t \in \mathbb{R}$ and that it is on the form:

$$\hat{\mu}^t = \sum_{i=1}^{N} 1_{F(t,U_i)} \hat{\mu}^t_i. \tag{2.5}$$

Here $(U_i)_{i=1}^N$ still represents the selection events and $\hat{\mu}^t_i : \overline{U}_i \to \mathbb{R}^n$ a locally Lipschitz post-selection estimator. The essential component of (2.5) is the function $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, which we assume is a *flow* (see e.g., [11] for a definition).

Under a set of regularity conditions and assumptions on the boundaries $(\partial U_i)_{i=1}^N$ and the *field* $x \mapsto \partial_t F(0, x)$, one can establish

$$\mathrm{df}(\hat{\mu}) = \mathrm{df_S}(\hat{\mu}) + \partial_t E(H(t, Y)), \tag{2.6}$$

where $H : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ is a function only depending on t and the observation y. An explicit formula for H is provided in the paper.

Four different estimators for which (2.5) holds are considered: marginal screening, relaxed lasso, best subset selection and singular value decomposition with a hard threshold on the singular values. For all but best subset selection, (2.6) is established. Moreover, for best subset selection $\partial_t E(H(t, Y))$ can still be viewed as a partial correction of $\mathrm{df_S}(\hat{\mu})$.

Bibliography

[1] L. T. Biegler, D. J. J., and B. G. E. Nonlinear parameter estimation: A case study comparison. *AIChE Journal*, 32:29–45, 1986.

[2] N. J.-B. Brunel. Parameter estimation of odes via nonparametric estimators. *Electron. J. Statist.*, 2:1242–1267, 2008.

[3] B. Calderhead, M. Girolami, and N. D. Lawrence. Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 217–224. Curran Associates, Inc., 2009.

[4] S. Chen, A. Shojaie, and D. M. Witten. Network reconstruction from high dimensional ordinary differential equations. *Journal of the American Statistical Association*, 2016.

[5] E. H. Davidson and I. S. Peter. Chapter 2 - gene regulatory networks. In E. H. Davidson and I. S. Peter, editors, *Genomic Control Process*, pages 41 – 77. Academic Press, Oxford, 2015.

[6] F. Dondelinger, D. Husmeier, S. Rogers, and M. Filippone. Ode parameter inference using adaptive gradient matching with gaussian processes. In C. M. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 216–228, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

[7] B. Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

[8] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[9] S. Gugushvili and C. A. Klaassen. n-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing. *Bernoulli*, 18(3):1061–1098, 08 2012.

[10] L. Janson, W. Fithian, and T. J. Hastie. Effective degrees of freedom: a flawed metaphor. *Biometrika*, 102(2):479–485, 2015.

[11] J. M. Lee. *Integral Curves and Flows*, pages 205–248. Springer New York, New York, NY, 2012.

[12] H. Liang and H. Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008.

[13] B. M. Parameter fitting in dynamic models. *Ecological Modelling*, 6:97–115, 1979.

[14] C. J. Oates and S. Mukherjee. Network inference and biological dynamics. *The Annals of Applied Statistics*, 6(3):1209–1235, 2012.

[15] L. Perko. *Differential Equations and Dynamical Systems*, volume 7. Springer-Verlag New York, 3 edition, 2001.

[16] M. Santillán. On the use of the hill functions in mathematical models of gene regulatory networks. *Math. Model. Nat. Phenom.*, 3:85–97, 2008.

[17] T. Sauer. *Numerical Analysis*. Pearson, Boston, 2006.

[18] T. Sideris. *Ordinary Differential Equations and Dynamical Systems*. Atlantis Studies in Differential Equations. Atlantis Press, 2013.

[19] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 11 1981.

[20] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[21] J. M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Stat. Comput.*, 3:28–46, 1982.

[22] E. W. J. Wallace, D. T. Gillespie, K. R. Sanft, and L. R. Petzold. Linear noise approximation is valid over limited times for any chemical system that is sufficently large. *IET Systems Biology*, 6(4):102–115, 2012.

[23] H. Wu, T. Lu, H. Xue, and H. Liang. Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109(506):700–716, 2014.

[24] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 04 2010.

[25] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67:301–320, 2005.

Papers

I

Learning Large Scale Ordinary Differential Equation Systems

Frederik Vissing Mikkelsen
Department of Mathematical Sciences
University of Copenhagen

Niels Richard Hansen
Department of Mathematical Sciences
University of Copenhagen

Publication details

In review.

LEARNING LARGE SCALE ORDINARY DIFFERENTIAL EQUATION SYSTEMS

FREDERIK VISSING MIKKELSEN AND NIELS RICHARD HANSEN

Abstract. Learning large scale nonlinear ordinary differential equation (ODE) systems from data is known to be computationally and statistically challenging. We present a framework together with the adaptive integral matching (AIM) algorithm for learning polynomial or rational ODE systems with a sparse network structure. The framework allows for time course data sampled from multiple environments representing e.g. different interventions or perturbations of the system. The algorithm AIM combines an initial penalised integral matching step with an adapted least squares step based on solving the ODE numerically. The R package *episode* implements AIM together with several other algorithms and is available from CRAN. It is shown that AIM achieves state-of-the-art network recovery for the *in silico* phosphoprotein abundance data from the eighth DREAM challenge with an AUROC of 0.74, and it is demonstrated via a range of numerical examples that AIM has good statistical properties while being computationally feasible even for large systems.

1. Introduction

We consider the problem of modelling time course data sampled from a dynamical system in different environments. We model data via ordinary differential equations (ODEs), with a particular emphasis on learning the network of the system's constituents. This setting arises for instance in systems biology with biochemical reactions and molecular networks (Wilkinson 2006, Oates & Mukherjee 2012, Babtie et al. 2014, Hill et al. 2016), where a reaction network or a gene regulatory network may either be partially known or completely unknown. Learning such ODE networks from data is highly relevant as they provide a means for predicting downstream effects of interventions in the system.

Our main contribution is a framework and the corresponding R package *episode* for learning polynomial and rational ODE systems, which is directly applicable to experimental data. Extensive numerical experiments have lead us to propose the adaptive integral matching (AIM) algorithm, though the R package includes several other learning algorithms. The framework and the learning algorithm AIM are useful when there exists little to no prior knowledge of the system in question and a fully data-driven network recognition is needed. However, the framework does also allow for incorporating prior knowledge into the estimation procedure as will be illustrated.

The paper is organised as follows. Section 2 motivates the ODE network estimation problem with the small EnvZ/OmpR system. Section 3 defines the statistical

Key words and phrases. ODE; Network inference; Inverse collocation; Nonlinear least squares; Systems biology; Chemical kinetics.

framework and Section 4 reviews two standard approaches to parameter estimation in ODE systems: *the least squares method* and *the inverse collocation methods*. Then the AIM algorithm that combines both approaches is presented together with the functionality of the R package *episode* developed for this paper. In Section 5 we apply our proposed method to two large scale dynamical systems: the *in silico* protein phosphorylation network used in the eighth DREAM challenge (Hill et al. 2016), and a full scale model of glycolysis in *Saccharomyces cerevisiae* (Hynne et al. 2001). Finally, in Section 6 we present two extensive simulation studies that compare the performance of AIM to other methods proposed in the literature.

2. The ODE Network Estimation Problem

We illustrate the problem addressed in this paper by a simple and concrete dynamical system: the EnvZ/OmpR system. It is present among a wide range of bacteria and is particularly well studied in *Escherichia coli* (Bernardini et al. (1990), Batchelor & Goulian (2003), Shinar & Feinberg (2010)). In this system the histidine kinase EnvZ responds to changes in the osmolarity resulting from extracellular impermeable compounds. It responds by controlling the phosphorylation of the regulator, OmpR, which itself proceeds to regulate the transcription of certain genes, including *ompF* and *ompC*. These two genes act as *porins* responsible for regulating the cellular diffusion across the membrane.

The EnvZ/OmpR system is heavily studied and the whole regulation process described above is the product of numerous studies, each of which were carefully designed to isolate specific mechanisms and investigate them individually. However, various networks in systems biology are only partially understood or not even discovered yet. In this paper we do not assume that the system in question was carefully dissected and studied as a sum of local mechanisms. We simply assume that the system was observed under different perturbations or interventions and solve the network estimation problem globally.

The EnvZ/OmpR system is driven by the six coupled ordinary differential equations (see e.g., Batchelor & Goulian (2003)):

$$
\begin{aligned}
\frac{d[\text{(EnvZ-P)OmpR}]}{dt} &= k_1[\text{EnvZ-P}][\text{OmpR}] - (k_{-1} + k_t)[\text{(EnvZ-P)OmpR}] \\
\frac{d[\text{EnvZ(OmpR-P)}]}{dt} &= k_2[\text{EnvZ}][\text{OmpR-P}] - (k_{-2} + k_p)[\text{EnvZ(OmpR-P)}] \\
\frac{d[\text{EnvZ-P}]}{dt} &= k_{-1}[\text{(EnvZ-P)OmpR}] - k_{-k}[\text{EnvZ-P}] + k_k[\text{EnvZ}] \\
&\quad - k_1[\text{EnvZ-P}][\text{OmpR}] \\
\frac{d[\text{EnvZ}]}{dt} &= k_{-k}[\text{EnvZ-P}] - k_k[\text{EnvZ}] + (k_p + k_{-2})[\text{EnvZ(OmpR-P)}] \\
&\quad + k_t[\text{(EnvZ-P)OmpR}] - k_2[\text{EnvZ}][\text{OmpR-P}] \\
\frac{d[\text{OmpR}]}{dt} &= k_{-1}[\text{(EnvZ-P)OmpR}] - k_1[\text{EnvZ-P}][\text{OmpR}] \\
&\quad + k_p[\text{(EnvZ)OmpR-P}] \\
\frac{d[\text{OmpR-P}]}{dt} &= k_t[\text{(EnvZ-P)OmpR}] - k_2[\text{EnvZ}][\text{OmpR-P}] \\
&\quad + k_{-2}[\text{EnvZ(OmpR-P)}]
\end{aligned}
\tag{1}
$$

which is a *mass action kinetics* (MAK) system. See Section 6 for details. In these equations, EnvZ-P and OmpR-P denote the phosphorylation of EnvZ and OmpR,

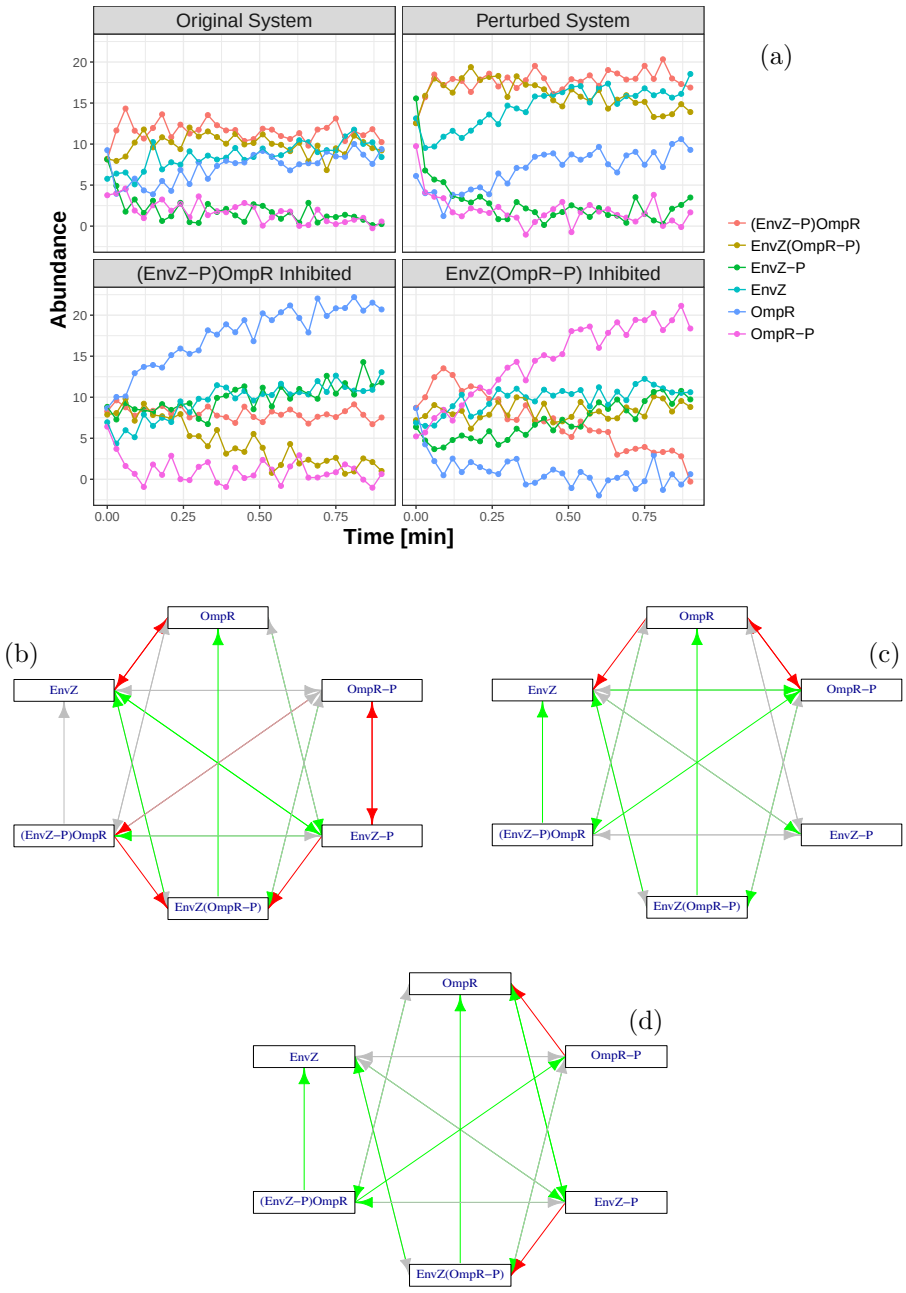


FIGURE 1. Simulated time course data (a) from the EnvZ/OmpR system with two perturbations and two interventions. Networks estimated from the perturbed data (b), the intervened data (c) and all data (d). Edges colouring scheme: true positive (green), false positive (red), false negative (gray).

respectively. These systems are characterised by a set of reactions, e.g.,

$$\text{EnvZ(OmpR-P)} \xrightarrow{k_p} \text{OmpR} + \text{EnvZ}.$$

The AIM algorithm works by searching through a large set of candidate reactions.

Figure 1 shows simulated data at 26 time points and the network recovered from these data via the AIM algorithm (specifically, Algorithm 4.2 in Section 4.3). The networks were recovered from a search space consisting of all MAK systems constructed from reactions on the form

$$X \to Y, \quad X + Y \to Z \quad \text{or} \quad Z \to X + Y, \tag{2}$$
$$\text{with } X, Y, Z \in \left\{ \begin{matrix} \text{EnvZ(OmpR-P)}, \text{(EnvZ-P)OmpR}, \\ \text{EnvZ}, \text{EnvZ-P}, \text{OmpR}, \text{OmpR-P} \end{matrix} \right\}.$$

The true parameter values in (1) were drawn at random from a normal distribution with mean 3 and the initial conditions were drawn uniformly at random from the interval $[5, 10]$. The AIM algorithm was here tuned to report reaction networks consisting of eight reactions.

This example illustrates that correct recovery of the network of reactions can benefit from combining several types of data sets. It was thus paramount to develop statistical and computational tools for recovering the network from time course data sampled under different perturbations and/or interventions, thus unifying the estimation process and circumventing the need for highly specific and specialised experiments with individual estimation procedures.

3. Statistical Framework

We consider a d-dimensional ODE given by:

$$\frac{dx}{dt} = f(x(t), \theta), \qquad x(0) = x_0 \tag{3}$$

with initial condition $x_0 \in \mathbb{R}^d$ and the smooth field $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ parameterised by $\theta \in \mathbb{R}^p$. In terms of f we define a corresponding network with nodes $1, \ldots, d$ and an edge from node l to node i if and only if $\partial f_i / \partial x_l \neq 0$. For many parameterised ODE systems a nonzero coordinate in θ corresponds to the presence of one or a few edges in the network, thus if we enforce sparsity in θ we also enforce sparsity in the network. This is, for instance, the case for the polynomial and rational fields that are currently implemented in the R package *episode*, see Table 1. In the setting of this paper, the focus is therefore on p being large but the true parameter being sparse. In some of the examples we consider, p is of the order $12,000$ with θ having as little as 0.65% of the parameters being nonzero.

We assume that the process x is observed at discrete time points $(t_i)_{i=1}^n$ with i.i.d. noise $(\varepsilon_i)_i$,

$$y(t_i) = x(t_i) + \varepsilon_i.$$

Using a sparsity enforcing penalty function pen, e.g., ℓ^1, elastic net, SCAD or MCP, we will consider the penalised least squares loss function

$$\ell_y(\theta) \coloneqq \frac{1}{2} \sum_{i=1}^{n} \sum_{l=1}^{d} w_{i,l} (y_l(t_i) - x_l(t_i, \theta))^2 + \lambda \sum_{j=1}^{p} v_j \text{pen}(\theta_j), \tag{4}$$

where $v = (v_j)_j$ are penalty weights and $w = (w_{i,l})_{i,l}$ are observation weights. Strong distributional assumptions on the errors, $\varepsilon_i \in \mathbb{R}^d$, are not necessary, but we

note that the least squares loss doesn't account for potential correlation among the d coordinates. However, differences in the error variances among the coordinates are accounted for by the observation weights, which are chosen adaptively by the proposed AIM algorithm.

Finally, we allow for observations of the same system under different interventions. We assume that the interventions are encoded in the ODE system through a Hadamard product of the parameter θ. More precisely, let $\{1, ..., E\}$ be a finite set of environments representing the interventions and let the data y consist of E sub-datasets $(y^e)_{e=1}^E$ with n_e time points in environment e. Define the environment specific observation weights similarly. The effective parameter of the ODE system in environment e is $\theta \circ c_e$, where $\theta \in \mathbb{R}^p$ is the baseline parameter corresponding to the unconstrained/unintervened system and $c_e \in \mathbb{R}^p$ is a vector of coordinatewise scale factors.

Typically, the scale factors c_e are binary. For instance, if in environment e the l^{th} coordinate of x is inhibited from affecting the i^{th} coordinate, then coordinate j of c_e is set to zero if and only if $\partial^2 f_i / \partial\theta_j \partial x_l \neq 0$. This inhibiting mode-of-action of an intervention is commonly used in gene regulatory networks in which certain proteins can inhibit the translation of some genes (see e.g., Fire (1999), Elbashir et al. (2001)). The loss function taking this type of intervention into account thus reads

$$\ell_y(\theta) \coloneqq \frac{1}{2} \sum_{e=1}^{E} \sum_{i=1}^{n_e} \sum_{l=1}^{d} w_{i,l}^e \left(y_l^e(t_i) - x_l^e(t_i, \theta \circ c_e) \right)^2 + \lambda \sum_{j=1}^{p} v_j \text{pen}(\theta_j). \tag{5}$$

Direct optimisation of (5) is challenging as this is generally a non-convex optimisation problem with many local minima, and most nonlinear ODEs will have to be solved numerically just to evaluate (5). In the following section we will introduce methods that mitigate some of the difficulties.

4. Methods

4.1. **The least squares method.** Direct minimisation of (5) above is called the (penalised) *least squares* method. This is sometimes referred to as the *gold standard* approach, see, e.g., Chen et al. (2016). As noted above, evaluating x in (5) typically requires a numerical ODE solver, which makes the least squares method computationally heavy. We refer to Sauer (2006) for a comprehensive overview of numerical ODE solvers, and to Appendix A for details on how to optimise (5) while keeping computation time to a minimum.

4.1.1. *Issues.* The penalised least squares method suffers from three main problems: it is computational demanding, it is a non-convex optimisation problem, and the solution depends on the choice of parameter scale (the choice of penalty weights).

The numerical solution of (3) is fundamentally a sequential problem, thus each evaluation of x is computationally heavy with only limited parallelisation options. Moreover, the derivative of x with respect to θ or x_0 solves another ODE, called the *sensitivity equations*, of dimensions d^2 and dp, respectively (see Appendix A for details).

The loss function (4) is non-convex even in the simplest case of a linear ODE, since linearity of the vector field f does not imply that the solution to the ODE is

linear. For nonlinear ODE systems we cannot even expect that (5) has a unique local minimiser for small λ.

The dependence on parameter scale is a general problem for penalised nonlinear least squares. The scales on which the parameters are penalised are essential for what parameters the sparsity inducing penalty selects. All other equal, parameters for which x is more sensitive is typically chosen over those for which x is less sensitive. This is a clear issue for correct network inference. In linear regression, it is common to standardise the predictors to bring the parameters on a common scale, but no immediate method exists for standardising the parameters in the nonlinear least squares function (5). It appears that any such method would depend on the unknown θ.

The inverse collocation methods introduced below address the three main problems of the least squares method.

4.2. **Inverse collocation methods.** In numerical analysis, *collocation methods* are a class of methods for solving ODE systems numerically. It goes as follows; in the ODE system

$$\frac{dx}{dt} = f(x(t), \theta), \qquad x(0) = x_0 \tag{6}$$

the parameter vector θ is assumed known. Moreover, a finite set of collocation points $\mathcal{C} \subseteq \mathbb{R}$ are chosen, as well as a vector space $\mathcal{V}$ of functions. A numerical solution, $\tilde{x} \in \mathcal{V}$, is sought that makes $\|\frac{d\tilde{x}}{dt}(t) - f(\tilde{x}(t), \theta)\|$ small in the collocation points for some norm. That is, the numerical solution is found by minimising a distance between $\left(\frac{d\tilde{x}}{dt}(t)\right)_{t\in\mathcal{C}}$ and $(f(\tilde{x}(t), \theta))_{t\in\mathcal{C}}$. Typically, $\mathcal{V} = \text{span}(\varphi_j)$ for a choice of finitely many basis functions $\varphi_j : \mathbb{R} \to \mathbb{R}$ and the norm is the 2-norm on $\mathbb{R}^d$. Collocation methods thus solve the forward problem of computing the solution of (6) for a known θ.

By *inverse collocation methods* we refer to a class of estimators of the parameter θ, given the observed trajectory x, that solve the inverse problem using the collocation idea. These methods exist in many versions (Varah (1982), Brunel (2008), Liang & Wu (2008), Calderhead et al. (2009), Gugushvili & Klaassen (2012), Dondelinger et al. (2013)) and are known under many other names, e.g., *gradient matching*, *trajectory matching* or *smooth-and-match estimators*. However, they all rely on the same two-step procedure: 1) approximate the data, y, by an element in $\mathcal{V}$ to get an estimate of the full trajectory $\hat{x}$; 2) base the estimation of θ on the trajectory $\hat{x}$ as if it were the true trajectory, by minimising the distance between the position, gradient or integral at a given set of collocation points. Typically, $\hat{x}$ is obtained as a smoother or an approximation of y via a basis expansion.

One example of an inverse collocation method is the *gradient matching method* (Varah (1982), Brunel (2008)), which minimises the approximate loss function:

$$\frac{1}{2}\sum_{t\in\mathcal{C}} \left\| \frac{d\hat{x}}{dt}(t) - f(\hat{x}(t), \theta) \right\|_2^2. \tag{7}$$

This method considerably reduces the computational cost compared to the least squares method, since it does not require solving the ODE system. Moreover, if f is linear in θ the optimisation problem becomes a linear least squares problem, which thus avoids all the three problems with the least squares method.

Later Dattner & Klaassen (2015) proposed minimising

$$\frac{1}{2}\sum_{t\in\mathcal{C}}\left\|\hat{x}(t)-x_0-\int_0^t f(\hat{x}(s),\theta)\,ds\right\|_2^2, \tag{8}$$

since the ODE system can be characterised as solving

$$x(t_2)-x(t_1)=\int_{t_1}^{t_2} f(x(s),\theta)\,ds, \qquad \text{for all } t_1,t_2\in\mathbb{R}. \tag{9}$$

instead. This requires numerical integration, which is often more stable than numerical differentiation, and under certain assumptions $\sqrt{n}$-consistency is guaranteed, as by Gugushvili & Klaassen (2012). Also, in this method the smoothed trajectory $\hat{x}$ does not have to be differentiable.

Note that in all of the above methods the collocation time points in $\mathcal{C}$ do not have to coincide with the observation time points of y. However, adding more time points in (7) and (8) will not necessarily decrease the variance of the estimator, as that mostly comes down to the y-$\hat{x}$ relation, i.e., the smoothing operation.

Nonparametric inverse collocation methods also exist, most notably are those by Wu et al. (2014) and Chen et al. (2016). Here the authors do not assume a parametric form of the field f, but approximate it by a nonparametric basis. In the former the authors consider the loss function

$$\frac{1}{2}\sum_{t\in\mathcal{C}}\sum_{l=1}^{d}\left(\frac{d\hat{x}_l}{dt}(t)-\sum_{j,k}\psi_k(\hat{x}_j(t))\theta_{ljk}\right)^2, \tag{10}$$

with $(\psi_k)_{k=1}^K$ a finite set of basis functions and $(\theta_{ljk})_{ljk}$ estimable parameters. In Chen et al. (2016) the integrals are considered instead:

$$\frac{1}{2}\sum_{t\in\mathcal{C}}\sum_{l=1}^{d}\left(\hat{x}_l(t)-x_l(0)-\sum_{j,k}\Psi_k(\hat{x}_j)(t)\theta_{ljk}\right)^2, \tag{11}$$

with $\Psi_k(x)(t) \coloneqq \int_0^t \psi_k(x(s))\,ds$. Note that both nonparametric methods assume f to be additive in the coordinates of x.

Finally, we note that the generalised profiling method described by Ramsay et al. (2007) is another variation on the inverse collocation method. It is inspired by functional data analysis and the main difference lies in that the smoothing step is θ-dependent and thus becomes part of the optimisation step.

Penalised versions of the inverse collocation methods – as alternatives to minimising (4) – have also been proposed by e.g., Lu et al. (2011) and Wu et al. (2014) to promote sparse solutions.

4.2.1. *Issues.* Though the inverse collocation methods remedy most issues of the least squares approach (in fact all of those discussed above, if the ODE is θ-linear), the inverse collocation methods also have their share of issues. Most notably, the results become dependent on the initial approximation (the smoother), which will introduce a bias without a clear trade-off in terms of a reduced variance. To illustrate this we present a small simulation study. Consider the classic Michaelis-Menten kinetics modelling the chemical reaction system (see Michaelis & Menten

(1913))

$$\mathrm{E} + \mathrm{S} \underset{k_r}{\overset{k_f}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_{cat}} \mathrm{E} + \mathrm{P} \tag{12}$$

in which the enzyme (E) forms a complex (ES) through a binding interaction with the substrate (S), which further releases the product (P) along with the freed enzyme. The abundances of the four compounds $x = (x_{\mathrm{E}}, x_{\mathrm{ES}}, x_{\mathrm{P}}, x_{\mathrm{S}})$ satisfy an ODE with $p = 3$ positive parameters (k_f, k_r, k_{cat}):

$$\begin{aligned} \frac{dx_{\mathrm{E}}}{dt} &= -k_f x_{\mathrm{E}} x_{\mathrm{S}} + k_r x_{\mathrm{ES}} + k_{cat} x_{\mathrm{ES}} & \frac{dx_{\mathrm{P}}}{dt} &= k_{cat} x_{\mathrm{ES}} \\ \frac{dx_{\mathrm{ES}}}{dt} &= k_f x_{\mathrm{E}} x_{\mathrm{S}} - k_r x_{\mathrm{ES}} - k_{cat} x_{\mathrm{ES}} & \frac{dx_{\mathrm{S}}}{dt} &= -k_f x_{\mathrm{E}} x_{\mathrm{S}} + k_r x_{\mathrm{ES}}. \end{aligned} \tag{13}$$

This classical ODE model is linear in the parameters and thus well suited for the inverse collocation methods. We generated data at $n = 10, 25, 100$ equidistant time points from the true trajectory with i.i.d. additive Gaussian noise. The data set was replicated 250 times and for each of them we applied a Gaussian kernel smoother with a range of bandwidths followed by the method proposed by Dattner & Klaassen (2015) to obtain parameter estimates. A summary of the resulting estimators is presented in Figure 2.

From Figure 2 we notice a bias which severely increases with the bandwidth, while the variance is only moderately reduced. Moreover, the bias hardly seems to change with the number of observations, unless the bandwidth is zero (equivalent to a linear interpolation smoother). Intuitively, this is no surprise: the purpose of smoothers, as indicated by their name, is to smooth the data. This is often manifested in a reduced pointwise variance, $\mathrm{V}(\hat{x}_y(t)) \leq \mathrm{V}(y(t))$ for $t \in \mathbb{R}$, and an increased autocovariance, $\mathrm{Cov}(\hat{x}_y(t), \hat{x}_y(s)) \geq \mathrm{Cov}(y(t), y(s))$, for $t, s \in \mathbb{R}$ close. Together this results in underestimated slopes. Since the slopes are essentially what is being modelled in ODE systems we would expect the resulting parameter estimates to have a large bias.

The least squares method and inverse collocation with zero bandwidth have the smallest biases. However for moderate and large noise levels the variance of the least squares method decreases faster with the number of observations. Though the inverse collocation methods with large bandwidths have slightly smaller variance, the least squares method still outperforms them, except for some settings with $n = 10$ and $\sigma = 0.1$.

Finally, inverse collocation methods suffer from one additional issue; they require fully observed processes to work. There is no obvious way of producing smoothed curves for latent coordinates and all coordinates are required in (7) and (8). This problem is revisited in Section 5.3.

4.3. **Adaptive Integral Matching.** We propose combining an inverse collocation method with the least squares method in such a way that we benefit from both methods. Inverse collocation methods are computationally lighter and produce good approximate parameter estimates, while not fully enjoying the statistical qualities of the least squares estimator. The least squares method is computationally expensive and suffers heavily from multiple local minimas, while generally performing better if the latter problems are alleviated.

Before presenting our suggestion of a combined estimator, we introduce a modification of the inverse collocation method by Dattner & Klaassen (2015). We propose

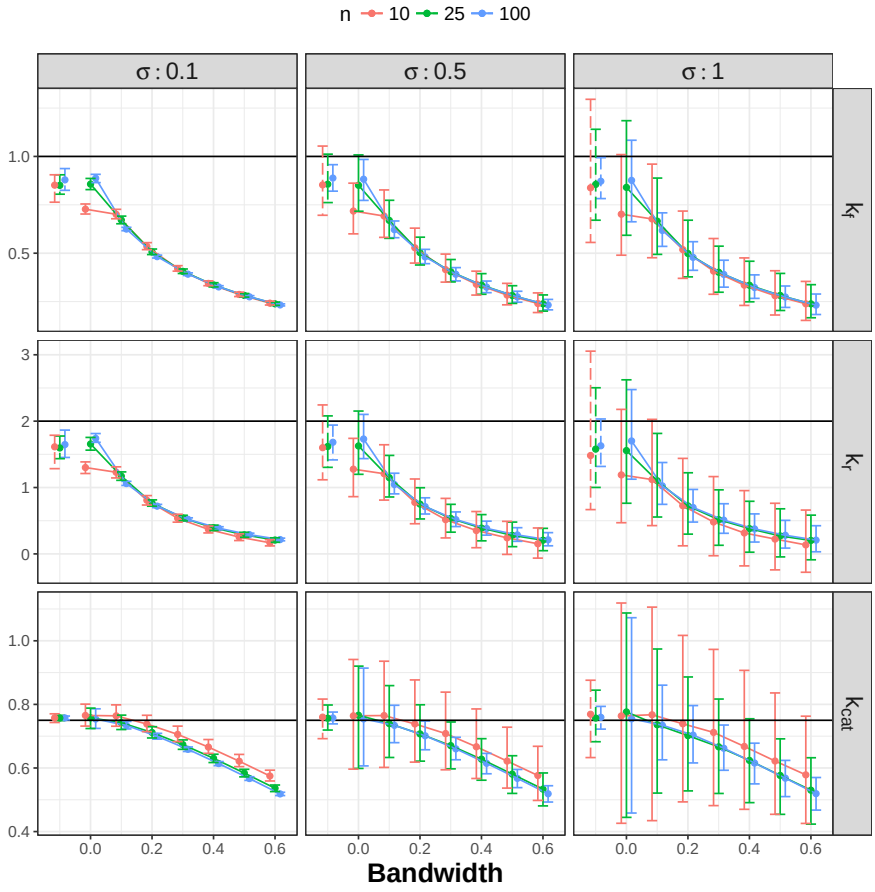


FIGURE 2. Medians and 5% and 95% percentiles of inverse collocation estimator considered by Dattner & Klaassen (2015). The kernels were scaled such that the quartiles are at $\pm 0.25 \times$ bandwidth. Data is simulated from (13) with $x_0 = (10, 2, 2, 10)$ and a time range of 1. The true parameters are marked with horizontal lines. The dashed lines on the left are the corresponding medians and percentiles of the least squares method.

the collocation method that consists of minimising the following approximate loss function

$$\tilde{\ell}^{\hat{x}}(\theta) := \frac{1}{2} \sum_{e=1}^{E} \sum_{i=1}^{n_e - 1} \sum_{l=1}^{d} w_{i,l}^{e} \left(\hat{x}_l^e(t_{i+1}) - \hat{x}_l^e(t_i) - \int_{t_i}^{t_{i+1}} f_l(\hat{x}^e(s), \theta \circ c_e) \, ds \right)^2 + \lambda \sum_{j=1}^{p} v_j \mathrm{pen}(\theta_j), \tag{14}$$

where $\hat{x}^e$ is the smoothed curve based on the data from environment e, and $\hat{x} = (\hat{x}^e)_{e=1}^E$ denotes the collection of smoothed curves for each environment. The above differs from (8) by integrating between consecutive time points instead of integrating

from 0 to t. This has two positive side effects: 1) it prevents errors between the true trajectory and its estimate $\hat{x}$ from accumulating; 2) the initial condition x_0 is no longer estimated. This is highly preferable as the initial condition is often a nuisance parameter and in a penalised setup the optimisation procedure often sets x_0 to compensate for the restricted freedom in θ. We refer to the estimator

$$\hat{\theta}^{\hat{x}}_{\lambda} := \arg\min_{\theta} \tilde{\ell}^{\hat{x}}(\theta) \tag{15}$$

as the *integral matching estimator* and stress that it depends on the smoother, $\hat{x}$.

From an integral matching estimate, $\hat{\theta}^{\hat{x}}_{\lambda}$, we adapt the scales $(c_e)_e$ and, optionally, the weights $(w^e)_e$. The new adapted scales are proportional to

$$c_e \circ \left(\left\| \left(\int_{t_i}^{t_{i+1}} \partial_{\theta_j} f(\hat{x}^e(s), \hat{\theta} \circ c_e) \, ds \right)_{i,e} \right\|_2^{-1} \right)_j, \quad \text{for } e = 1, ..., E. \tag{16}$$

If the field is linear in θ, then the updated scales only depend on the smoother. If the field is not θ-linear one uses $\hat{\theta} = \hat{\theta}^{\hat{x}}_{\lambda}$ for a small λ. The scales are thus standardised by the column norms of the first order Taylor approximation of the integrals in (14). If f is linear in θ, this coincides with standardising the columns in a penalised linear least squares problem. This adaptation of the scales ensures that parameters are locally on the same scale and thus penalised in a fair manner in the subsequent least squares estimation. Similarly, the new adapted weights are proportional to

$$\frac{(w^e_{i,l})_{i,e}}{\sum_{e=1}^{E} \sum_{i=1}^{n_e - 1} w^e_{i,l} \left(\hat{x}^e_l(t_{i+1}) - \hat{x}^e_l(t_i) - \int_{t_i}^{t_{i+1}} f_l(\hat{x}^e(s), \theta \circ c_e) \, ds \right)^2} \tag{17}$$

for $l = 1, ..., d$, i.e., inversely proportional to the empirical variances for each species. This adapts the variance structure across species for the subsequent estimation. This leads to the *adaptive integral matching* (AIM) algorithm:

Algorithm 4.1 (AIM). *Input: Time course data from E environments, $y = (y^e)_{e=1}^E$, each sampled at $(t_i)_{i=1}^{n_e}$ timepoints. Similarly structured observation weights $w = (w^e)_{e=1}^E$, along with penalty weights, $v \in \mathbb{R}^p_+$, and environment-specific scales $(c_e)_{e=1}^E$. Smoothed trajectories $(\hat{x}^e)_e$ evaluated on a fine grid of time points.*

1. *Apply the integral matching estimator, (15), to obtain initial estimates $\hat{\theta}^{\hat{x}}_{\lambda}$ for a sequence of λ values.*
2. *Adapt the scales and weights according to (16) and (17).*
3. *Refit by minimising (5) using the adapted weights and $\hat{\theta}^{\hat{x}}_{\lambda}$ as initial value.*

In step (3) the penalty term may be scaled down or removed entirely to reduce the bias induced by the penalty, and the parameter space may be restricted to reduce the computational costs. Algorithm 4.2 below presents a particular incarnation of the refitting step. In Appendix A additional techniques to reduce the computation time are presented.

4.3.1. *Implementation.* As part of this paper, software for optimising (5) and (14) (used in Algorithm 4.1) is available in the R package *episode*. In the latter optimisation problem the user supplies the smoothed trajectories $(\hat{x}^e)_e$ evaluated on a fine grid of time points and the software then optimises (14) using numerical integration over the supplied grid. By keeping this modular form, the user has complete

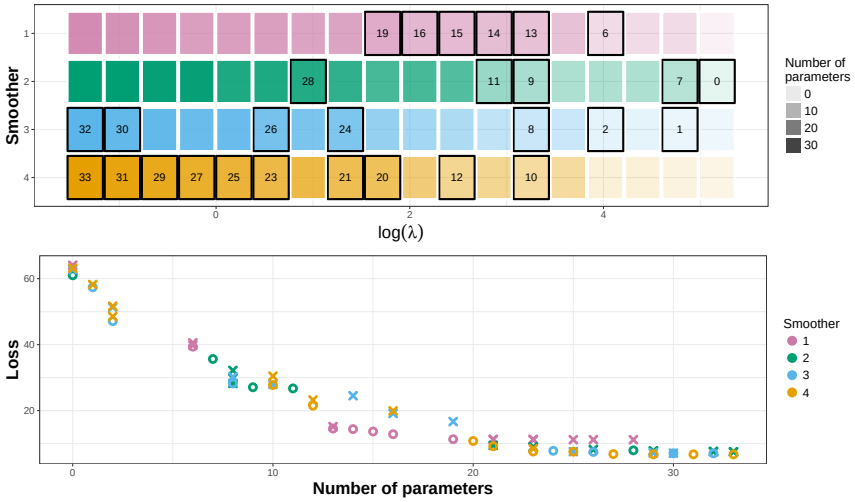


FIGURE 3. Visualisation of the stratified ranking in Algorithm 4.2 applied to the EnvZ/OmpR data from Section 2. Four smoothers were employed and each proposed a sequence of candidate models for varying tuning parameter (top). The loss values of the candidate models are stratified according to model size (bottom). For each model size the candidate with minimal loss is found and marked with a black border (top).

freedom in choosing a suitable smoother. It is possible not to smooth the data at all, which corresponds to $\hat{x}^e$ linearly interpolating the observations y^e.

We recommend subjecting a whole family of smoothers to Algorithm 4.1 in order to alleviate potentially high variance and multiple local minima issues. The resulting version of the AIM algorithm that we suggest and have tested extensively consists of the following steps:

Algorithm 4.2. *Input: Time course data from E environments, $y = (y^e)_{e=1}^E$, each sampled at $(t_i)_{i=1}^{n_e}$ timepoints. Similarly structured observation weights $w = (w^e)_{e=1}^E$, along with penalty weights, $v \in \mathbb{R}_+^p$, and environment-specific scales $(c_e)_{e=1}^E$.*

(1) Produce a family of smoothed curves $\{\hat{x}\}$, from data y, where the smoother is applied to each environment separately: $\hat{x} = (\hat{x}_e)_{e=1}^E$.

(2) For each $\hat{x}$ apply Algorithm 4.1 with the refitting step implemented as follows: define the support estimator $\hat{S}_\lambda^{\hat{x}} = \operatorname{supp}(\hat{\theta}_\lambda^{\hat{x}})$ and compute the unpenalised least squares estimate

$$\tilde{\theta}_\lambda^{\hat{x}} := \underset{\theta:\operatorname{supp}(\theta)=\hat{S}_\lambda^{\hat{x}}}{\arg\min} \frac{1}{2} \sum_{e=1}^{E} \sum_{i=1}^{n_e} \sum_{l=1}^{d} w_{i,l}^e \left(y_l^e(t_i) - x_l^e(t_i, \theta \circ c_e)\right)^2. \tag{18}$$

over the restricted parameter space determined by λ and $\hat{x}$.

(3) Stratify the refitted estimates $(\tilde{\theta}_\lambda^{\hat{x}})_{\lambda,\hat{x}}$ by the number of non-zero parameters. For each strata rank the resulting estimates by their loss value at optimum. See Figure 3 for an illustration of this step.

The purpose of the stratified ranking is to produce a sequence of models indexed by the number of nonzero parameters. This is primarily important for comparison purposes in the subsequent sections.

Currently, the R package *episode* implements AIM and other learning algorithms for mass action kinetics (described below), which encode all polynomial fields, power law kinetics, which encode all polynomial fields in a different way and two larger classes of ODE systems assuming a rational form of the field. As for penalties, ℓ^1, ℓ^2, elastic net, SCAD and MCP are implemented. Moreover, the package handles missing values and allows for box-constrained optimisation as well. Table 1 provides a schematic overview of the features in *episode*. The tools in *episode* are flexible and modular and the Algorithms 4.1 and 4.2 are primarily recommendations on how to combine them. When using the *episode* package for the least squares method, i.e., optimising (5), suitable initialisations are required and the resulting estimates may depend on these. The tools are thus designed to easily pass the integral matching estimates as initialisations for the least squares method.

	ODE Models
MAK	Mass Action Kinetics $$\frac{dx}{dt} = (B - A)^T \mathrm{diag}(x^A)k,$$ with $A, B \in \mathbb{N}_0^{r \times d}$ fixed and $k \in \mathbb{R}_+^r$ estimable.
PLK	Power Law Kinetics $$\frac{dx}{dt} = \theta x^A,$$ with $A \in \mathbb{N}_0^{r \times d}$ fixed and $\theta \in \mathbb{R}^{d \times r}$ estimable.
RLK	Rational Law Kinetics $$\frac{dx}{dt} = \theta \frac{x^A}{1 + x^B},$$ with $A, B \in \mathbb{N}_0^{r \times d}$ fixed, the fraction evaluated elementwise and $\theta \in \mathbb{R}^{d \times r}$ estimable.
RMAK	Rational Mass Action Kinetics $$\frac{dx}{dt} = C^T \frac{\theta_1 x^A}{1 + \theta^2 x^A},$$ with $A \in \mathbb{N}_0^{b \times d}$ and $C \in \mathbb{N}_0^{r \times d}$ fixed, the fraction evaluated elementwise and $\theta_1, \theta_2 \in \mathbb{R}^{r \times b}$ estimable.
	Data Structures
Inhibition	Species i is inhibited from reacting with species j in environment e: Set l^{th} coordinate of $c_e \in \mathbb{R}^p$ to 0 if $\partial_{\theta_l} f_{ij} \neq 0$, and 1 otherwise.
Activation	Species i only reacts with species j in environment e: If $\partial_{\theta_l} f_{ij} \neq 0$ set $c_e(l) = 1$ and $c_{e'}(l) = 0$ for all $e \neq e'$.
Stimulation	Reaction rate of reaction l is increased by factor k in environment e: Set $c_e(l) = k$.
Misc	Missing data. Partially observed processes only supported by exact estimation.
	Estimation Components
Penalties	ℓ^1, ℓ^2, elastic net, SCAD, MCP and no penalty.
Weights	Both observation and penalty weights.
Loss	Can minimise both least squares loss (5) and integral matching loss (14). The minimiser of the latter loss function can easily be passed as initialisation for minimising the former.
Parameter Constraints	Box constraints for all estimable parameters are available.
Misc.	Automatic adaptation of parameter scales and observation weights.

TABLE 1. Overview of features in the R package *episode*.

5. Applications

In this section we study two concrete large scale dynamical systems. One is the *in silico* protein phosphorylation network used in the eighth DREAM challenge, and the other is glycolysis in *Saccharomyces cerevisiae*. Both of these systems are like the EnvZ/OmpR system based on *mass action kinetics*, which is first reviewed briefly. However, in these applications it is not all components of the mass action system that is observed, and rational fields are used to model the dynamics of the observed species.

5.1. Mass Action Kinetics.

We consider a chemical kinetics framework of ODE systems. Assuming that we have d chemical species, e.g., NaCl, H_2O or proteins, labelled $X = (X_1, ..., X_d)$. A set of $r = 1, ..., R$ reactions on the form:

$$a_1X_1 + ... + a_dX_d \to b_1X_1 + ... + b_dX_d, \tag{19}$$

govern the dynamics of the species. Here $(a_i)_{i=1}^d$ and $(b_i)_{i=1}^d$ are non-negative integers, called the stoichiometric coefficients. For reaction $r = 1, ..., R$ let $A_r, B_r \in \mathbb{N}_0^d$ denote the vector of left hand and right hand side stoichiometric coefficients, respectively. The net change of molecules due to reaction r is $v_r = B_r - A_r$.

Let $x = (x_1, ..., x_d) \in \mathbb{R}_+^d$ denote the vector of abundances of each chemical species. If the total number of molecules is sufficiently large, we can model the dynamics of x as

$$\frac{dx}{dt} = \sum_{r=1}^{R} v_r\gamma_r(x(t)), \qquad x(0) = x_0. \tag{20}$$

See Wallace et al. (2012) for details on its derivation. The laws of mass action kinetics (see, e.g., Horn & Jackson (1972)) states that

$$\gamma_r(x) = k_r x^{A_r}, \tag{21}$$

where $k_r \geq 0$ is a rate constant and x^a is shorthand for $\Pi_{i=1}^d x_i^{a_i}$ for any two non-negative vectors in $\mathbb{R}^d$. The *stoichiometric matrices* A and B are the $R \times d$-dimensional matrices with the r^{th} row being A_r and B_r respectively. The matrix notation of (20) is

$$\frac{dx}{dt} = (B - A)^T \text{diag}(x^A)k, \qquad x(0) = x_0, \tag{22}$$

where $k = (k_r)_{r=1}^R$ and $x^A = (x^{A_r})_{r=1}^R$.

Ideally, all chemical reaction systems should approximately be a mass action kinetics system. However, in complex reaction networks this may not be the case for the observable species. For gene regulatory networks, say, some proteins may exist in different forms depending on whether an inhibitor or activator is bound to its associated sites, which is not directly observable. In such cases a *quasi-stationary approximation* is often employed to reduce a full mass action system to a system for the observable variables only. The quasi-stationary approximation assumes that the chemical species, X, can be divided into two subsets, X_L and X_O, the *latent* and *observed* species:

$$\begin{aligned} \frac{dx_L}{dt} &= f_L(x_L, x_O) \\ \frac{dx_O}{dt} &= f_O(x_L, x_O). \end{aligned} \tag{23}$$

Under the quasi-stationary assumption, i.e., the dynamics of x_L is faster than x_O, the dynamics of x can be approximated by the ODE system

$$\frac{dx_O}{dt} = f_O(\tilde{x}_L(x_O), x_O), \tag{24}$$

where $\tilde{x}_L(x_O)$ is the restriction of x_L to the manifold $\mathcal{M}_{x_O} \coloneqq \{x_L \mid f_L(x_L, x_O) = 0\}$ for all values of x_O. In certain settings, including fast binding on gene-sites in gene regulatory networks, this approximation is reasonable and the right hand side of (24) is rational. See Santillan (2008) for a detailed treatment. This is the main motivation for including rational systems in our framework and in the R package *episode*, and its usage will be illustrated by the two applications below.

5.2. ***in silico* phosphoprotein abundance data.** In this section we compare AIM to state-of-the-art network inference methods in systems biology. The eighth DREAM challenge (Hill et al. (2016)) aimed at advancing causal inference of signalling networks in protein phosphorylation. One of the challenges presented the participants with time course data from a complex *in silico* dynamical model of a protein signalling network. The species were given anonymous labels and thus no prior knowledge of the network was given.

The data consisted of 20 environments produced using combinations of three inhibitors (or no inhibitor) and two types of stimuli each with two strengths. The targets of the inhibitors were provided and encoded in AIM through the scales $(c_e)_e$ in Algorithm 4.2. In light of the rational ODE systems discussed in Section 6, AIM fitted the ODE system given by the field

$$\frac{dx}{dt} = \theta \mathrm{diag}(x^A)\mathrm{diag}(1 + x^B)^{-1}, \tag{25}$$

where A and B are $R \times d$-dimensional matrices and $\theta \in \mathbb{R}^{d \times R}$ estimable coefficients. The rows of A and B ($(a_r)_r$, $(b_r)_r$), ran over all non-negative integer d-tuples summing to at most one. Thus the search space consisted of first order rational functions.

Besides the final DREAM challenge submissions, AIM was compared to two additional methods. The first was the integral matching (IM) estimator, given in (15). This method represents the use of a penalised inverse collocation method to select the network. The second method was the least squares estimator using a SCAD penalty (SCAD), which was obtained by optimising (5) for a decreasing sequence of λ, initialised in $\theta = 0$. The continuation principle was used, i.e., the optimum found at the previous value of λ was re-used as initialisation for next value of λ.

The performance of AIM was assessed using the *DREAMTools* Python package provided by Cokelaer et al. (2015) and containing the tools used to assess the original challenge submissions. AIM got a AUROC score of 0.737, which makes AIM the second best solution overall among the 65 submissions and notably better than the two ODE-based submissions. An overview of the performances of AIM, IM and SCAD, along with the final submissions for the eighth DREAM challenge is presented in Figure 4.

5.3. **Glycolysis in *Saccharomyces cerevisiae*.** Hynne et al. (2001) presented a full scale chemical kinetics model for glycolysis in *Saccharomyces cerevisiae*, constructed from experimental substrate measurements. While Hynne et al. (2001)

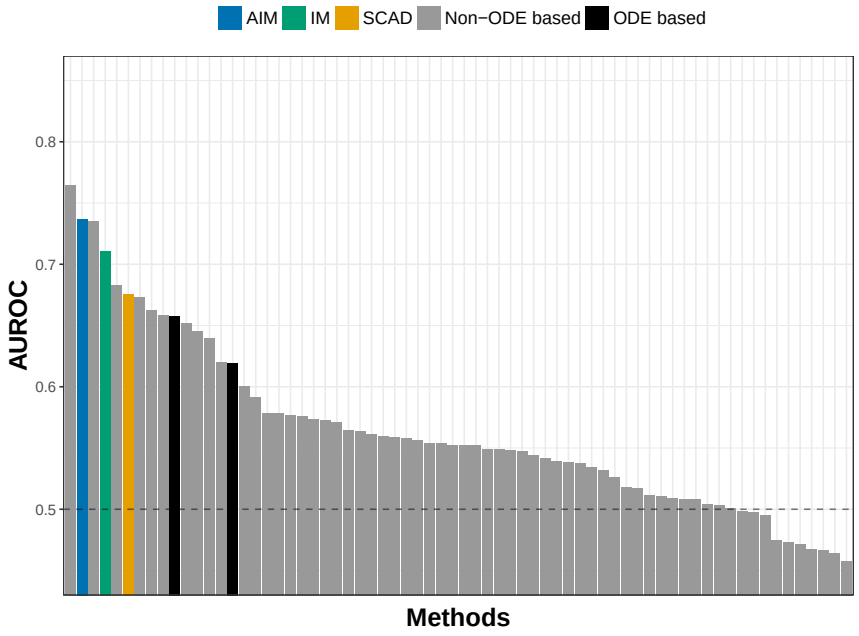


FIGURE 4. The AUROC scores of all final submissions in the *in silico* network recognition challenge in the eighth DREAM challenge (Hill et al. (2016)) (gray and black bars), along with three of the methods considered in this paper.

knew the metabolic pathway a priori and focused on estimating unknown rate parameters, we will apply AIM to identify the network from simulated data. In total, $d = 22$ chemical species enter the glycolysis cycle in an elaborate metabolic network, see Figure 6.

The dynamical model considered by Hynne et al. (2001) does not fall into the class of mass action kinetics models. All mass action kinetics models have polynomial fields, but the ODE field consider by Hynne et al. (2001) is rational. More precisely, the field is (20) with rate functions on the form

$$\gamma_r(x) = \frac{\langle a_r; x^{A_r}\rangle}{\langle b_r; x^{B_r}\rangle}, \qquad A_r \in \mathbb{N}_0^{\alpha_r \times d}, B_r \in \mathbb{N}_0^{\beta_r \times d}, \alpha_r \text{ and } \beta_r \in \mathbb{N}, \tag{26}$$

with $\langle \cdot; \cdot \rangle$ denoting the standard inner product and $a_r \in \mathbb{R}^{\alpha_r}, b_r \in \mathbb{R}^{\beta_r}$ estimable coefficients.

For a parametric model on the form (26) to be generic enough to include the model considered by Hynne et al. (2001), the polynomials $\langle a_r; x^{A_r}\rangle, \langle b_r; x^{B_r}\rangle$ need to have an order of at least 3. Hence, if no prior knowledge on the glycolysis is given, at least $p = 2d(1 + d + d^2 + d^3) = 490,820$ parameters are needed. It is possible to use AIM with half a million parameters for polynomial systems as given by (21). However, the rational ODE systems are far more sensitive than the polynomial, which in practice results in far longer computations for the numerical solvers and a higher variance of the resulting estimator. Thus a model search space of dimension 490,820 is currently not feasible for rational systems, and we will

therefore consider three scenarios for fitting this system using either prior knowledge or an approximate and smaller model search space.

We consider two different prior knowledge scenarios: 1) knowing what *complexes* can be formed in the system, i.e., what terms $A_r \in \mathbb{N}_0^{\alpha_r \times d}, B_r \in \mathbb{N}_0^{\beta_r \times d}$ may appear in the rational field. Even with this prior knowledge, we know very little about the network, since we do not know what complexes drive what reactions. In the system of Hynne et al. (2001) there are in total 46 complexes. 2) we know a superset of the complexes. In this setting we include an additional 46 false complexes drawn at random.

In the third scenario we restrict AIM to a smaller parametric model, which will not include the true model. Hence the purpose of this scenario is partly to study the performance of AIM on large and realistic ODE systems and partly to study the robustness to model misspecification. The restricted model space assumes rate functions on the form

$$\gamma_r(x) = \frac{k_r x^{a_r}}{1 + x^{b_r}}, \tag{27}$$

with $a_r \in \mathbb{N}_0^d$ and $b_r \in \mathbb{N}_0^d$ covering all first order terms (i.e., all combinations of non-negative integers summing to at most 1) and k_r estimable coefficient. This produces a total of $p = d(d+1)^2 = 11,638$ parameters. By assuming fixed coefficients in the denominator of the rate functions, we obtain an ODE field which is linear in the parameters.

5.3.1. *Simulation study design.* Using the reactions and rate functions listed in Table 1 and 2 in Hynne et al. (2001), we numerically solved the ODE system with parameters in Tables 4-7 in Hynne et al. (2001).

We considered $E = 5, 10, 15, 20$ environments each given its own inhibition. These were produced as follows: 20 distinct chemical species were selected at random, one for each of the maximal number of environments. In each environment the selected species were inhibited, i.e., the species did not form any complexes with the other species and were thus prevented from reacting with the other species.

The trajectories ran for 5 minutes, at which the system had settled at a stationary point. The trajectories were observed at 30 log-equidistant time points with additive Gaussian noise, with standard deviations $\sigma = 0.1, 0.25, 0.5$. The signal of this system is approximately 3, hence the lower noise level.

Each prior knowledge setting had an associated model search space for which AIM was applied. The data was separated into environments, in each of which all but the inhibited species evolved over time. AIM was applied to each environment individually, and the resulting subnetworks were averaged to produce the full network estimates.

5.3.2. *Results.* The ROC curves for the network estimator were calculated for each of the 100 replications. The average curves are in Figure 5. Not surprisingly the performance decreased with increasing noise, but more importantly we see a clear improvement with the number of environments. The estimated network and the true network are summarised in Figure 6. We note that the approximate model has the overall worst performance in terms of network recovery, while the two models that incorporate prior knowledge by restricting the search space perform better. Though we do identify aspects of the network reasonably well, it is also evident that there is room for improvement, especially when no prior knowledge is used.

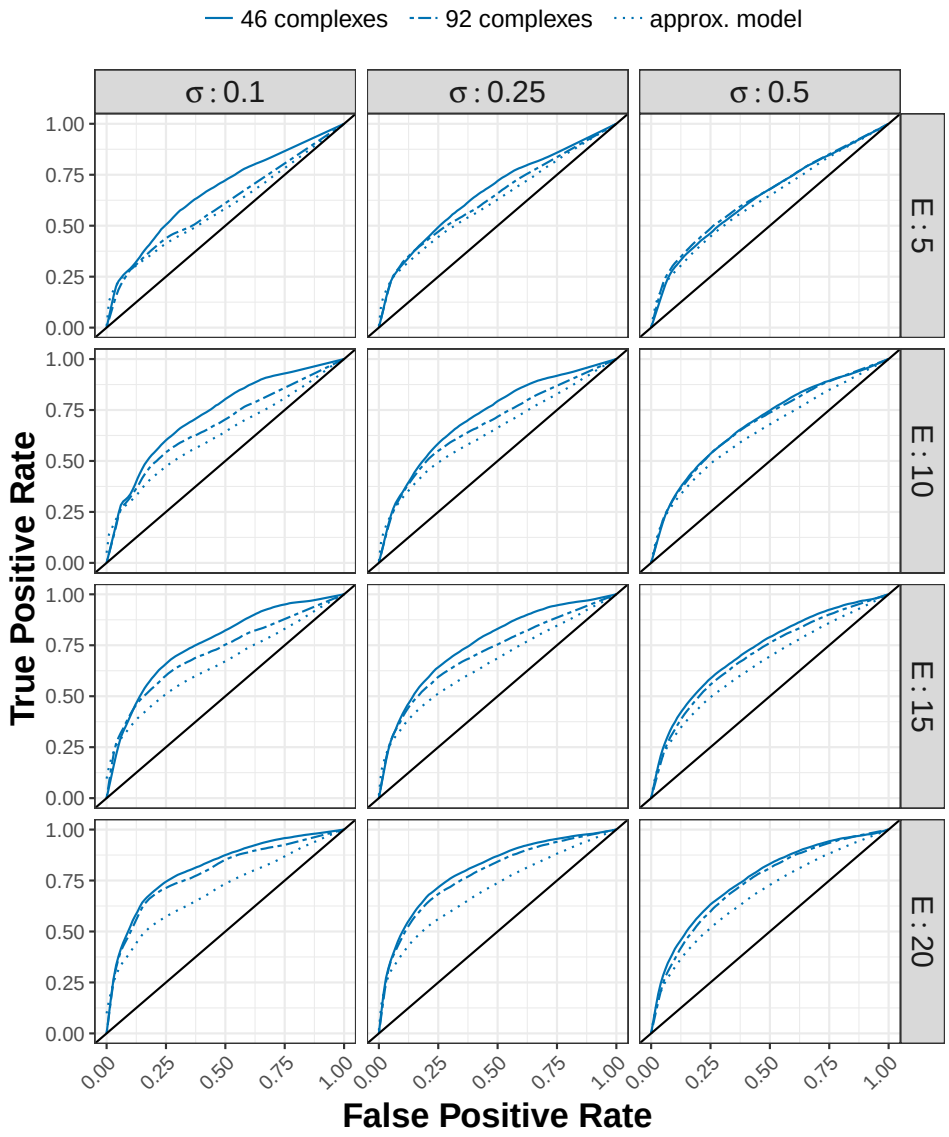


FIGURE 5. Pointwise average of the ROC curves, stratified according to noise level and number of environments.

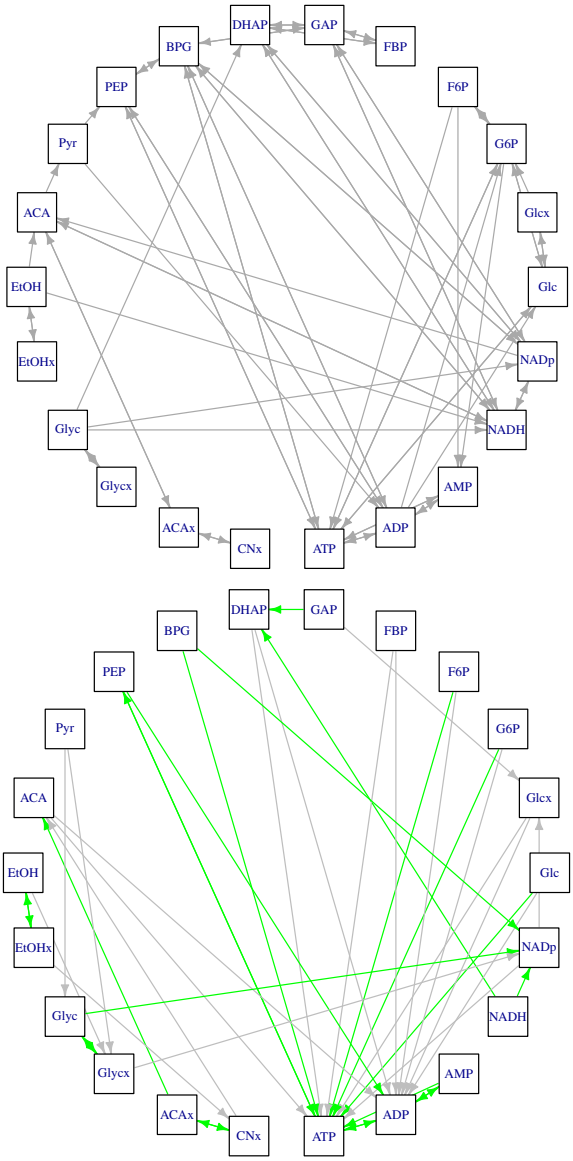


FIGURE 6. The true glycolysis network (upper). The graph with the two most reported children for each node (lower). True edges are green. $E = 20$, $\sigma = 0.25$ and second setting with superset of complexes known.

6. Simulation Studies

In this section we return to the mass action kinetics systems:

$$\frac{dx}{dt} = (B - A)^T \text{diag}(x^A)k, \qquad x(0) = x_0, \tag{28}$$

where $k = (k_r)_{r=1}^R$ and $x^A = (x^{A_r})_{r=1}^R$, and $R \in \mathbb{N}$ denotes the number of reactions. When the stoichiometric matrices A and B are either not known at all or only partially known, we seek to identify them from a larger set of candidate reactions. We test the performance of AIM in such a challenge through two simulation studies.

6.1. **Simulation Study I.**

In this section we compare AIM to another ODE network recovery algorithm GRADE, provided by Chen et al. (2016). GRADE is a nonparametric inverse collocation method. It replaces a parametric form of f with a basis function expansion assuming an additive form, i.e., any given coordinate of f depends on the other coordinates in an additive manner. GRADE was shown quite effective in simulation studies and applications by Chen et al. (2016).

6.1.1. *Simulation study design.*

The setting of this simulation study is a replicate of the simulation study in Section 5.3 in Chen et al. (2016). We consider five independent Lotka-Volterra systems, i.e., for $k = 1, ..., 5$ we let

$$\begin{aligned} \frac{dx_{2k-1}}{dt} &= 2x_{2k-1}(t) - vx_{2k-1}(t)x_{2k}(t) \\ \frac{dx_{2k}}{dt} &= vx_{2k-1}(t)x_{2k}(t) - 2x_{2k-1}(t). \end{aligned} \tag{29}$$

Note that the above ODE can be cast as a mass action kinetics system with 10 species and 15 reactions.

For each of the E environments we drew the initialisation uniformly at random from $[0, 4]$ and solved (29) for $t \in [0, 5]$. Observations were extracted at $t = 0, 0.1, 0.2, ..., 5$ with additive Gaussian noise. AIM was applied with a single linear interpolation smoother and GRADE used a spline smoother for smoothing the data and a monomial basis expansion of size 3 in (11).

AIM searched ODE solutions using mass action kinetics reactions on the form

$$\begin{aligned} X_i + X_j &\to 2X_i, && i, j = 1, ..., 10, \ i \neq j. \\ X_i &\to 2X_i, && i = 1, ..., 10. \\ X_i &\to 0, && i = 1, ..., 10. \end{aligned} \tag{30}$$

The search space thus consisted of $p = 110$ reactions.

The following simulation parameters were used:

Parameter	Values				Description
E	2	4	8		Number of environments
v	1	3	5	7	Interaction parameter
σ	0.5	1	2		Standard deviation of additive noise

The noise level was intentionally relatively large, as this ODE system is far easier to recover than those of the other systems considered in this paper. Each simulation was replicated 100 times.

6.1.2. *Results.* The ROC curves were derived for each simulation setting and method. A summary of the ROC curves is presented in Figure 7 for $v = 5$. Similar summaries for the remaining values of v are found in the supplementary material.

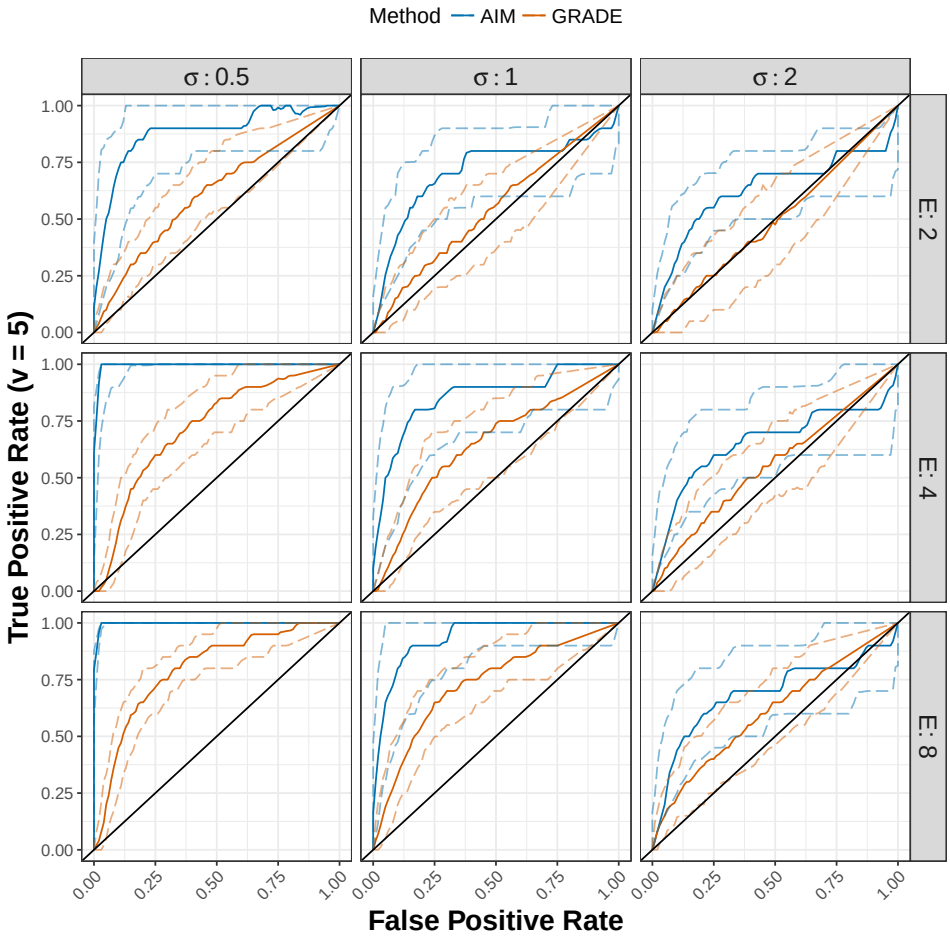


FIGURE 7. Pointwise median and 5^{th} and 95^{th} percentiles of ROC curves for the Lotka-Volterra system with $v = 5$, stratified according to noise level and number of environments.

Across all noise levels, number of environments and interaction parameters we see that AIM generally performs better than GRADE. We ascribe this to the additivity assumption in GRADE, as we see improvements for decreasing v. Surprisingly, AIM works to an acceptable degree even for $\sigma = 2$, which corresponds to a signal-to-noise ratio of 1.

6.2. **Simulation Study II.** In this section we report the results from an extensive simulation study, whose purpose was to quantify how well AIM (in its concrete form of Algorithm 4.2) identifies the correct reaction network.

6.2.1. *Estimators.* AIM was compared to an exhaustive gradient matching method (EGM), inspired by Babtie et al. (2014). See Appendix A for details on its implementation. Even though it relies on an inverse collocation method, this approach is computationally expensive as it selects the reactions based on best subset selection applied to each species separately. This computer intensive inverse collocation method attempts to get the most information out of the approximate loss function, by finding global minima on lower dimensional subspaces.

Solving the best subset selection problem for each species separately only induces the global best subset selection solution if each coordinate of θ induces a single edge in the network. This property holds for linear ODE systems and does not hold for most mass action kinetics systems. This simulation study restricts the attention to reactions on the form $X_i + X_j \to 2X_i$, $i \neq j$, hence each reaction corresponds to a bidirectional edge between node i and j – as well as a self-edge in both nodes. Thus, in this particular simulation study, EGM will provide the same networks as a best subset selection performed over all possible reactions. EGM was not applied to the examples considered in Section 5, as the number of species was too large for an exhaustive search to be computationally feasible.

Each method reported estimated reaction networks consisting of up to $5d$ reactions. EGM used the Gaussian process smoother described in Babtie et al. (2014), IM used a linear interpolation smoother and AIM used both smoothers. In order to produce additional initialisations for AIM, the integral matching estimates from each smoother were produced with and without standardising the coordinates of the process. Both AIM and IM used the elastic net penalty (Zou & Hastie (2005)) with $\alpha = 0.25$.

6.2.2. *Simulation study design.* Data was drawn from reaction networks composed of reactions on the form:

$$X_i + X_j \to 2X_i \tag{31}$$

where $i, j = 1, ..., d$ and $i \neq j$, with a total number of reactions at $p = d(d-1)$. Time course data from E environments were drawn. Each environment was given its own initial condition produced as follows: between one and four distinct chemical species were selected at random. In each environment all but the selected species were knocked down by 50% from their equilibrium value and the initial abundance of the selected species were increased by the total mass knocked down. The initial conditions were rescaled to have an average of 5. Since the total number of molecules is preserved by reactions on form (31), the average signal strength is approximately 5.

Data were sampled at $t = 0$ and $t = 2^{i/2}$, for $i = -5, -4, ..., 2, 3$, all with additive Gaussian noise. Each species $i = 1, ..., d$ was given $\alpha = 1, 2$ true reactions, i.e., a total of $d\alpha$ reactions on the form (31) had rate parameter 1 and the remaining 0.

The following simulation parameters were used:

Parameter	Values			Description
d	7	9	11	Number of species
α	1	2		Number of true reactions per species
E	2	4	8	Number of environments
n	10			Number of data points per environment
σ	0.1	0.5	1	Standard deviation of additive noise

For each combination of the simulation parameters, 100 replications of the above simulation experiments were conducted.

6.2.3. *Results.* We first report the recovery of the true network. For each replicate and simulation parameter combination the receiver operating characteristic (ROC) curves of the network were derived. Pointwise averages over the replicates are illustrated in Figure 8 for $d = 9$ and $\alpha = 1$. The remaining curves can be found in the supplementary material.

From Figure 8 we see that AIM consistently recovered the network better than the other methods. IM and SCAD were the worst performing methods with SCAD improving the most with increasing number of environments, though not reaching the level of EGM and AIM.

These tendencies are repeated in the other figures in the supplementary material, with an overall decrease in performance for $\alpha = 2$. Figure 9 provides an overview of the area under the ROC curves (AUROC) across simulation settings. AIM generally had the largest median AUROC values across all settings. For all methods we also see improvements when increasing the number of environments and that increasing the number of species for most scenarios decreases the performance. Generally, for all methods the network recovery performances dropped considerably for $\alpha = 2$.

Next we report the recovery of the true reactions. We visualise their performance by their precision-recall curves. In Figure 10 the pointwise averaged precision-recall curves for $d = 9$ and $\alpha = 1$ are presented. The remaining curves can be found in the supplementary material.

From Figure 10 we see that EGM recovered most correct reactions early in the recovery for E large. But after recovering 20–35% of the true reactions AIM surpassed EGM in reaction recovery performance. All methods improved considerably with increasing number of environments. These results match what we observed for the network recovery to some degree.

The network ROC curves and the reaction precision-recall curves together suggest that EGM recovers the first few reactions and network edges accurately, but AIM is more accurate when more reactions are reported.

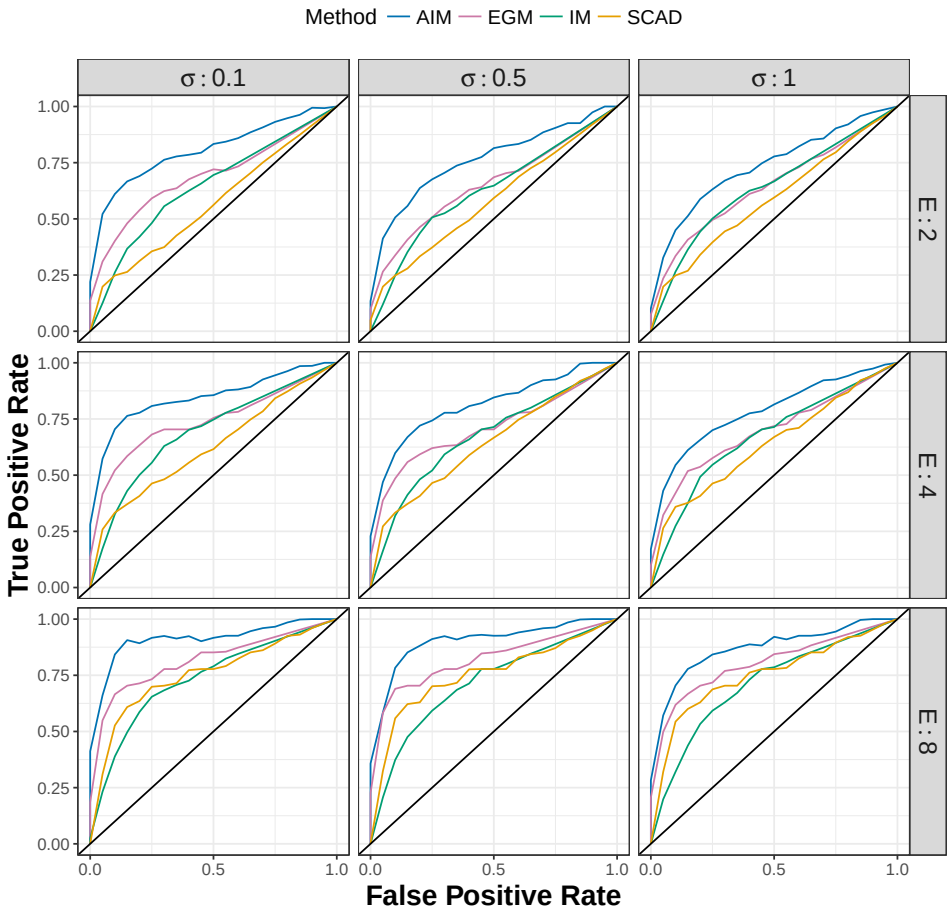


FIGURE 8. Pointwise averaged ROC curves of the network estimates for $d = 9$ and $\alpha = 1$, stratified according to noise level and number of environments.

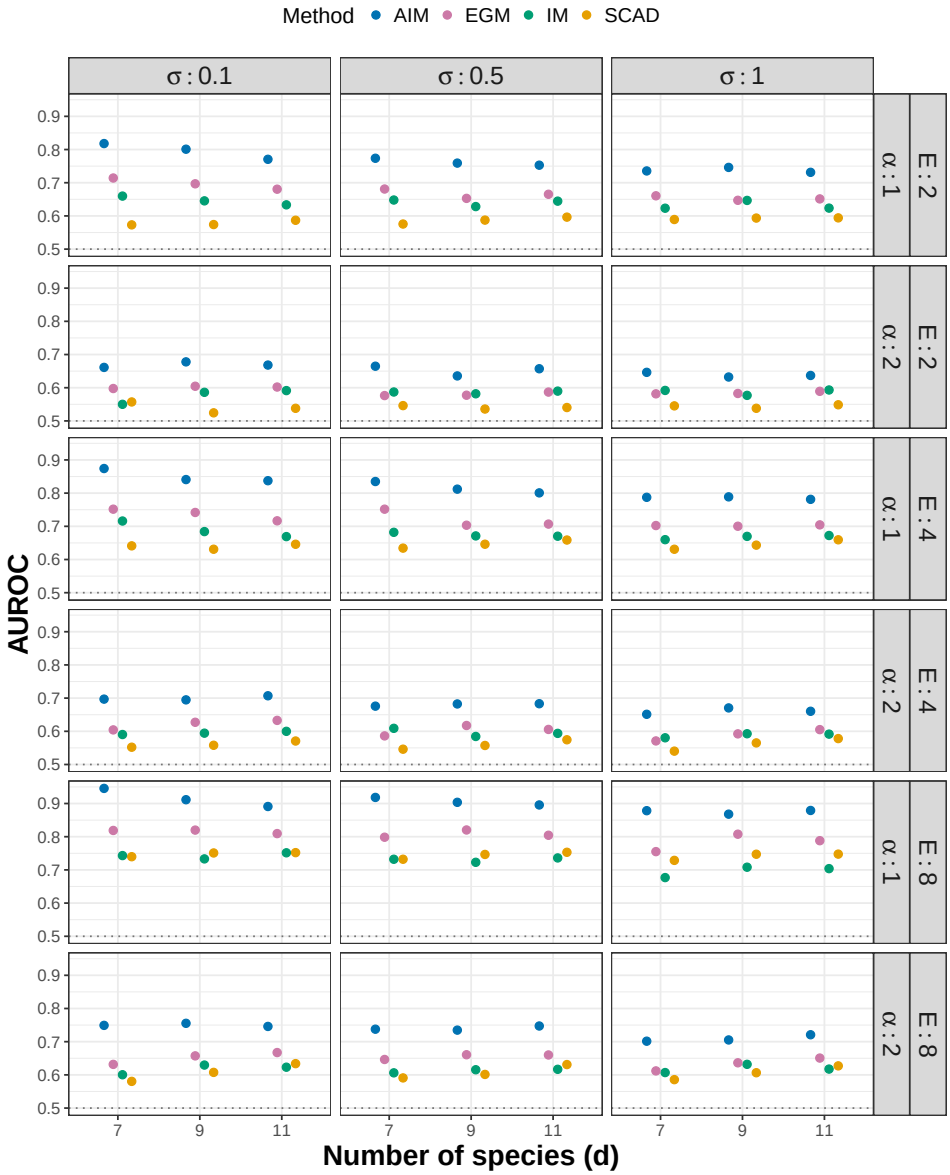


FIGURE 9. Median AUROC across the 100 replications and stratified according to the simulation settings.

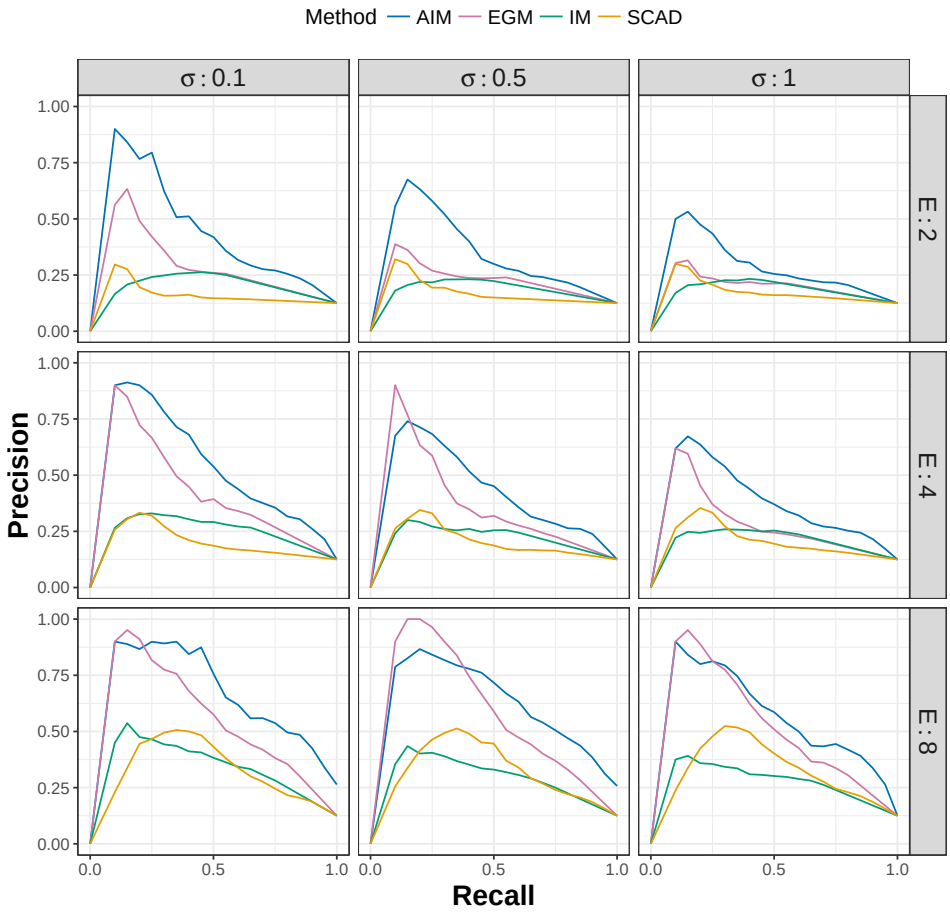


FIGURE 10. Pointwise averaged precision-recall curves of the reactions recovered for $d = 9$ and $\alpha = 1$, stratified according to noise level and number of environments.

The mean squared error of the estimated reaction networks were also assessed. A single model was selected for each method by minimising the mean squared error on an independent test set. The squared error between the trajectories produced by the selected model and the true trajectory at each time point was derived. Medians of the mean squared error are presented in Figure 11.

We see that the methods using the ODE-based loss (AIM and SCAD) have much smaller mean squared error than the methods based on the approximate loss. That the mean squared error is so large for IM and EGM can be explained as a bias phenomenon similar to the one observed for the Michaelis-Menten example as illustrated in Figure 2. IM without a penalty and a linear interpolation smoother – as used in this simulation study – is expected to be relatively unbiased but with a large variance. However, the sparsity enforcing penalty introduced an additional bias, and the resulting trajectories of the fitted ODE did not match the truth very well in general (data not shown). For EGM the conclusion is the same, but the argument is the other way around. This method used a Gaussian process smoother, which should decrease the variance of the parameter estimates, but the under-estimated slopes introduced a stronger bias. Again, the resulting trajectories of the ODE fitted using EGM did not match the truth very well. Though the mean squared error suggests that the approximate loss functions provide quantitatively incorrect estimates, we did find qualitatively correct network and reaction recovery for those methods.

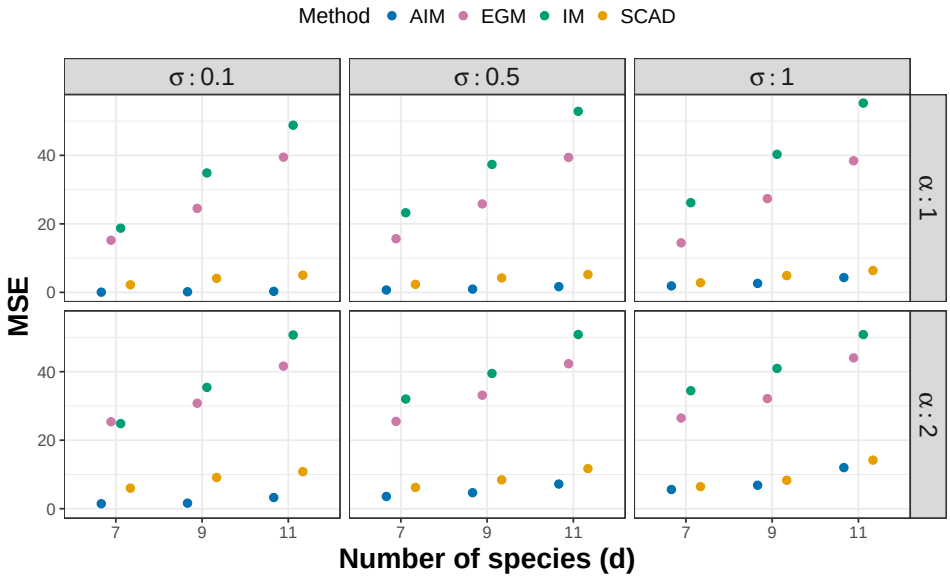


FIGURE 11. Medians of the mean squared error between the tuned trajectories and the true trajectory for $E = 4$.

Finally we report computation times. The median computation time over 10 replications can be found in Figure 12 for two collections of reactions: $X_i + X_j \to 2X_i$, $i \neq j$ and $X_i + X_j \to X_i + X_k$, $j \neq i \neq k$. The model search space size of the latter grows faster with the number of species and it quickly becomes a challenge

for EGM. In fact EGM was excluded for $d > 5$ due to infeasible computation times. For d small, AIM is somewhat slow, however in terms of scalability with d AIM resembles IM more than EGM.

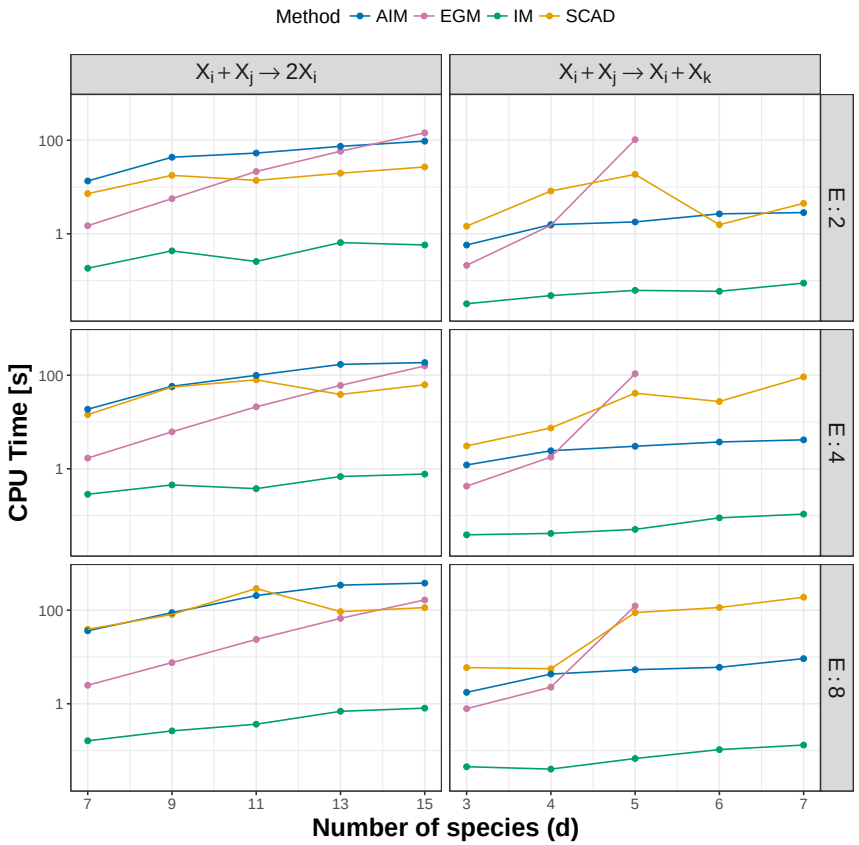


FIGURE 12. Median computation time for the two reaction collections. $\sigma = 0.5$ and $\alpha = 1$.

7. Discussion

Collocation based estimation of parameters in ODE systems is computationally less demanding than the least squares method that relies on numerical solutions of the ODE systems. Calderhead et al. (2009) demonstrated this in a Bayesian setting and proposed a collocation method based on Gaussian processes, Babtie et al. (2014) relied on a Gaussian process collocation based search of the model space similar to the EGM method that we have implemented, and Chen et al. (2016) relied on penalised collocation based estimation for their method GRADE (Graph Reconstruction via Additive Differential Equations). Based on these and similar results we sought to develop a scalable inference framework for mass action systems, but we found several challenges, and the present paper represents a synthesis of how we dealt with these challenges. We discuss below how the most important challenges were addressed.

7.1. **Bias.** We found that penalised collocation methods were computationally fast, but even if they did recover qualitatively the correct network and reactions for realistic data sizes, the resulting parameter estimates were biased. The bias was induced partly by the initial smoothing and partly by the penalisation, and the fitted model would not reproduce very accurately the solution trajectories of the true data generating ODE system. Moreover, the results would be sensitive to the precise choice of initial smoother. We found that among the collocation methods, our proposed integral matching (IM) estimator obtained by minimising (14) has reasonable statistical properties.

7.2. **Penalised least squares.** To test if penalised least squares methods are feasible for large systems we implemented a number of algorithms for numerical minimisation of the penalised least squares loss including the proximal gradient algorithm with screening as presented in Appendix A.2. Sparsity and screening combined with fast solvers of the sensitivity equations makes it possible to apply these algorithms even for fairly large systems. However, the sparsity inducing penalty still induces a bias of the resulting estimates, which can also be quite dependent on the initialisation of the optimisation algorithm due to local minima of the objective function. We illustrated that least squares with the SCAD penalty achieved rather accurate estimates in terms of mean squared error from the true trajectories, but in terms of network recovery it was inferior to the other methods considered – in particular IM, which is much faster.

7.3. **Parameter scale.** A parameter in an ODE system typically controls the rate of a reaction, and the bias induced by the penalty results in reaction rates being underestimated. It is our experience that the bias induced by the penalty can have quite substantial effects for nonlinear ODE systems, and the choice of parameter scale determines this bias together with the combined effect of the penalty term. Standardisation as used in regression models, e.g. in the R package glmnet (Friedman et al. 2010), for bringing the parameters on a common scale is not directly applicable. We suggest adaptive rescaling as given by (16), which does require a pilot estimate of the unknown parameter unless f is linear in θ. However, we did not find this to be a data-driven panacea for the choice of parameter scale, and we ended up concluding that the unpenalised estimator given by (18) had better statistical properties in our experiments.

7.4. **Combining methods.** Our combined AIM algorithm uses the fast collocation method IM to obtain good initialisation parameters for the least squares method. Moreover, AIM in the form of Algorithm 4.2 – which we have extensively tested – uses multiple smoothers to achieve an even greater variety of initialisations, and it introduces sparsity in the least squares method by restricting the parameter space. The stratified ranking was proposed as a way to aggregate the resulting models into a sequence of models indexed by the number of nonzero parameters. Clearly, alternative aggregations are possible, e.g. using a weighted average. Moreover, the simulation study in Section 6.2 found that EGM performed slightly better than AIM for the first couple of reactions. As EGM performs the first couple of search iterations fairly quickly, a hybrid approach for initialisation suggests itself using EGM for the first couple of reactions and IM for the remaining reactions. We have not investigated if alternative aggregation schemes or hybrid approaches for initialisation could further improve the statistical properties of the algorithm.

7.5. **Network recovery.** We demonstrated that AIM has good network recovery properties in a number of different examples and compared to several alternatives. In the Lotka-Volterra example it was, for instance, demonstrated that AIM was far superior to GRADE (Chen et al. 2016). This is perhaps unsurprising given that GRADE assumes an additive form of f, but we emphasise this to argue that additivity is a quite strong assumption, which is unlikely to hold for many ODE systems of practical relevance.

AIM also performed well in the recovering of the *in silico* network of protein phosphorylation, and it was superior in terms of AUROC to IM and SCAD considered in this paper as well as the two ODE-based solutions that participated in the eighth DREAM challenge. We did not participate in the challenge, but AIM would have been ranked second among all participants. We note that the top-ranked submissions including the winning team did not rely on a mechanistic model – the submission only required network edge weights. The winning team constructed edge weights via tests for nonlinear functional relations without the constraints of an ODE system, and were in this way better able to capture the correct network structure (see Supplementary material on Team 7 in Hill et al. (2016)). However, such methods are not capable of predicting e.g. intervention or perturbation effects. It is clearly of interest to utilise such network estimates as prior information for learning ODE systems, and we demonstrated how this can be done in our framework for the discovery of the glycolysis network. For the DREAM challenge it would make an unfair comparison if we were to piggyback on the published top-ranked network for this particular data set, hence we ran AIM in this example without any prior network restrictions.

7.6. **Conclusion.** The AIM algorithm was presented and demonstrated to have good statistical properties for realistic data structures and sizes. The implementation of AIM also demonstrated that it is possible to learn large ODE systems via least squares methods – even if this is computationally heavy. Further improvements may be possible, e.g. to account for a more complicated noise structure than additive, uncorrelated noise. In the light of the linear noise approximation, described in detail by Wallace et al. (2012), the noise in mass action kinetics systems scales with the signal. We have partially addressed this by rescaling the observation weights as given by (17), which will adjust the weights according to the

variance of each species. However, we have not investigated ways to adjust for a more complicated variance structure.

Our intention with AIM and the associated R package *episode* is to provide a thoroughly tested, applicable and useful framework for learning ODE systems using state-of-the-art methods. This should be of use to experimentalists, and it should be able to serve as a benchmark for further developments. The R package currently supports polynomial and rational systems in certain parameterisations, and it is implemented in a modular way that allows for easy addition of new parameterised families of ODE systems. Doing so, the entire framework consisting of data structures, ODE solvers and optimisers including AIM are then directly available.

Appendix A. Computational Aspects

A.1. **Sensitivity equations and approximative gradients.** Let $x : \mathbb{R} \to \mathbb{R}^d$ solve the ODE

$$\frac{dx}{dt} = f(x, \theta), \qquad x(0) = x_0. \tag{32}$$

The derivative of x with respect to $\theta \in \mathbb{R}^p$, i.e., the matrix valued function $x_\theta : \mathbb{R} \to \mathbb{R}^{d\times p}$, solves another ODE system:

$$\frac{dx_\theta}{dt} = \frac{\partial f}{\partial x}(x, \theta)x_\theta + \frac{\partial f}{\partial \theta}(x, \theta), \qquad x_\theta(0) = \mathbb{O}_{d\times p}, \tag{33}$$

where $\mathbb{O}_{d\times p}$ is the $d \times p$-dimensional zero-matrix. Analogously, the derivative of x with respect to x_0, $x_{x_0} : \mathbb{R} \to \mathbb{R}^{d\times d}$, solves the ODE:

$$\frac{dx_{x_0}}{dt} = \frac{\partial f}{\partial x}(x, \theta)x_{x_0}, \qquad x_{x_0}(0) = \mathbb{I}_d, \tag{34}$$

with $\mathbb{I}_d$ the n-dimensional identity matrix. The equations (33) and (34) are called the *sensitivity equations* of (32). Notice that once the original system is solved, the columns of the sensitivity equations can be solved independently.

The sensitivity equations are often solved simultaneously with the original system (32). Even if (32) requires a computationally intensive solver (e.g., a solver with adaptive step length or implicit solvers), the sensitivity equations often only require simple solvers like the Euler scheme to be accurate. There are two reasons for this. Firstly, the sensitivity equations are (time-inhomogeneous) affine ODE systems which are often less sensitive. Secondly, the exact gradient is not necessary to optimise a smooth function – an approximate gradient pointing in roughly the same direction will suffice.

A method for deriving even faster approximate gradients to

$$\ell_y := \sum_{t\in\mathcal{C}} \|y_t - \int_0^t f(x(s, \theta), \theta)ds\|_2^2 \tag{35}$$

was proposed by Mikkelsen (2015) and inspired by inverse collocation methods. It goes as follows: assuming that θ_0 is the current value of θ in the optimisation, then minimise

$$\theta \mapsto \sum_{t\in\mathcal{C}} \|y_t - \int_0^t f(x(s, \theta_0), \theta)ds\|_2^2 \tag{36}$$

to produce the next step. Though these approximate solutions are not guaranteed to improve the original loss function, they still produce fast and approximate descent directions. If the approximate solution does not improve the loss function, it is suggested to take one classic gradient-based step before retrying the approximate solution.

The above approach is equivalent to using the Gauss-Newton method on the original loss function, but ignoring the first term of the right hand side of (33), when calculating the differentials.

A.2. **Proximal gradient and screening methods.** The penalised ODE loss function

$$\ell_y(\theta) := \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{d} w_{i,l}(y_l(t_i) - x_l(t_i, \theta))^2 + \lambda\sum_{j=1}^{p} v_j \mathrm{pen}(\theta_j), \tag{37}$$

is optimised using a proximal gradient method, as described in Hale et al. (2008) for ℓ^1 penalties. For non-convex penalties, like SCAD and MCP, this method is combined with the majorisation method by Fan & Li (2001).

The proximal gradient method for (37) thus starts with initialisation θ^0 and the proceeds with

$$\theta^{k+1} = \text{prox}(\theta^k, p^k, \tau), \qquad \text{for } k = 0, 1, ... \tag{38}$$

where the proximal operator is defined as

$$\text{prox}(\theta, p, \tau) \coloneqq \text{sign}(\theta - \tau p) \circ \max(0, |\theta - \tau p| - \lambda\mu(\theta)). \tag{39}$$

The vector p^k is the derivative of $\frac{1}{2}\sum_{i=1}^n \sum_{l=1}^d w_{i,l}(y_l(t_i) - x_l(t_i, \theta))^2$ at θ^k, i.e.,

$$p^k \coloneqq -\sum_{i=1}^n \sum_{l=1}^d w_{i,l}(y_l(t_i) - x_l(t_i, \theta^k)) \frac{dx_l}{d\theta}|_{\theta=\theta^k}(t_i) \tag{40}$$

and the sensitivity equation is solved using the approximative methods described above. For non-convex penalties, the majorisation amounts to replacing the penalty weights in each step by $v_j \circ \frac{d^2 \text{pen}}{d\theta_j^2}(\theta_j)$.

In (38) the step length τ is chosen through backtracking. Moreover, not all coordinates of θ changes in each step of (38). This is due to the sparsity inducing property of the proximal operator. In practice this means that many coordinates of the derivatives p^k are calculated (using computationally intensive numerical solvers) and then never used. Computations are thus saved if occasionally the ODE system is screened for strong variables as follows: at every n^{th} step all coordinates of p^k are evaluated. If $\theta^k = \text{prox}(\theta^k, p^k, 1)$ (up to some numerical accuracy) then stop, else identify the active set $\mathcal{A} = \{i \mid \theta_i^k \neq 0 \text{ or } \theta_i^k \neq \text{prox}(\theta_i^k, p_i^k, 1)\}$ and run proximal gradient algorithm on $\mathcal{A}$ only until next screening.

A.3. **Exhaustive Gradient Matching.** Inspired by Babtie et al. (2014) exhaustive gradient matching applies a best subset selection of parameters for explaining the dynamics of each chemical species individually. The individual results are then combined into a parameter estimate of the full ODE system.

For an ODE system given by the field $f(x, \theta)$, then each coordinate of the solution satisfies

$$\frac{dx_l}{dt} = f_l(x, \theta), \quad l = 1, ..., d \tag{41}$$

where f_l is the l^{th} coordinate of f and $\theta \in \mathbb{R}^p$. Given smoothed curves $\hat{x} = (\hat{x}_l)_{l=1}^d$ for each coordinate, then the approximate inverse collocation loss function is

$$\ell(\theta) \coloneqq \frac{1}{2}\sum_{l=1}^d \sum_{t\in\mathcal{C}} \left(\frac{d\hat{x}_l}{dt}(t) - f_l(\hat{x}(t), \theta)\right)^2. \tag{42}$$

If the field is linear in the parameters the above becomes the sum of squares,

$$\ell(\theta) = \sum_{l=1}^d \|Y_l - X_l\theta\|_2^2. \tag{43}$$

where $Y_l = \left(\frac{\hat{x}_l}{dt}(t)\right)_{t\in\mathcal{C}}$ and $X_l = \left(\frac{\partial f_l}{\partial\theta}(\hat{x}(t))\right)_{t\in\mathcal{C}}$ is a concatenation of the θ-gradients of the field over the time points.

The exhaustive gradient matching method (EGM) goes as follows: for each $l = 1, ..., d$ construct $Y_l = \left(\frac{\hat{x}_l}{dt}(t)\right)_{t\in\mathcal{C}}$ and $X_\cdot = \left(\frac{\partial f_l}{\partial\theta}(\hat{x}(t))\right)_{t\in\mathcal{C}}$. For any $\mathcal{K} \subseteq \{1, ..., p\}$ let $X_l^{\mathcal{K}}$ denote the $\mathcal{K}$ columns of X_l and let $\theta^{\mathcal{K}} \in \mathbb{R}^{|\mathcal{K}|}$. For $k = 1, ..., K$ find the subset $\mathcal{K}_k^l \subseteq \{1, ..., p\}$ with $|\mathcal{K}_k^l| = k$ such that

$$\min_{\theta^{\mathcal{K}_k^i}} \frac{1}{2}\|Y_l - X_l^{\mathcal{K}_k}\theta^{\mathcal{K}_k}\|_2^2 \tag{44}$$

is minimal.

Each species now has a sequence of subsets of increasing size, $(\mathcal{K}_k^l)_{k=0}^K$. They are combined into a sequence of subsets representing the full system, $(\mathcal{K}_k)_{k=0}^{dK}$, by the union

$$\mathcal{K}_k := \bigcup_{i=1}^{d} \mathcal{K}_{\alpha_i(k)}^l. \tag{45}$$

The k-dependent tuple $\alpha(k) = (\alpha_l(k))_{l=1}^d \in \{1, ..., K\}^d$ is given by the recursion

$$\alpha(0) = (0)_{l=1}^d, \qquad \alpha(k+1) = \alpha(k) + e_{l^*}, \quad k = 0, ..., Kd-1 \tag{46}$$

where the increments e_{l^*} is 1 at coordinate l^* and zero elsewhere. The coordinate l^* is chosen such that

$$\min_{\theta: j\notin\bigcup_{l=1}^d \mathcal{K}_{\gamma_i}^l \Rightarrow \theta_j = 0} \ell(\theta), \qquad \gamma = \alpha(k) + e_{l^*} \tag{47}$$

is minimal, i.e., the species whose next subset improves the loss the most determines the next full subset $\mathcal{K}_{k+1}$.

The EGM estimator becomes the best subset selection estimator of (42) if and only if each coordinate of the parameter vector θ affects only one edge in the network. This is the case for linear ODE systems and, if ignoring self-edges, also the case for the systems studied in Section 6.2. However, for most ODE systems a single coordinate often affects multiple network edges simultaneously.

References

Babtie, A. C., Kirk, P. & Stumpf, M. P. H. (2014), 'Topological sensitivity analysis for systems biology', *Proc Natl Acad Sci USA* **111**, 18507–18512.

Batchelor, E. & Goulian, M. (2003), 'Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system', *PNAS* **100**, 691–696.

Bernardini, M. L., Fontaine, A. & Sansonetti, P. J. (1990), 'The two-component regulatory system ompr-envz controls the virulence of shigella flexneri.', *J. Bacteriol* **172**, 6274–6281.

Brunel, N. J.-B. (2008), 'Parameter estimation of odes via nonparametric estimators', *Electron. J. Statist.* **2**, 1242–1267.
URL: *http://dx.doi.org/10.1214/07-EJS132*

Calderhead, B., Girolami, M. & Lawrence, N. D. (2009), Accelerating bayesian inference over nonlinear differential equations with gaussian processes, *in* D. Koller, D. Schuurmans, Y. Bengio & L. Bottou, eds, 'Advances in Neural Information Processing Systems 21', Curran Associates, Inc., pp. 217–224.
URL: *http://papers.nips.cc/paper/3497-accelerating-bayesian-inference-over-nonlinear-differential-equations-with-gaussian-processes.pdf*

Chen, S., Shojaie, A. & Witten, D. M. (2016), 'Network reconstruction from high dimensional ordinary differential equations', *Journal of the American Statistical Association* .

Cokelaer, T., Bansal, M., Bare, C., Bilal, E., Bot, B. M., Neto, E. C., Eduati1, F., Gönen, M., Hill, S. M., Hoff, B., Karr, J. R., Küffner, R., Menden, M. P., Meyer, P., Norel, R., Pratap, A., Prill, R. J., Weirauch, M. T., Costello, J. C., Stolovitzky, G. & Saez-Rodriguez, J. (2015), 'Dreamtools: a python package for scoring collaborative challenges [version 1; referees: 3 approved with reservations]', *F1000Research* **4**.

Dattner, I. & Klaassen, C. A. J. (2015), 'Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters', *Electron. J. Statist.* **9**(2), 1939–1973.

Dondelinger, F., Husmeier, D., Rogers, S. & Filippone, M. (2013), Ode parameter inference using adaptive gradient matching with gaussian processes, *in* C. M. Carvalho & P. Ravikumar, eds, 'Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics', Vol. 31 of *Proceedings of Machine Learning Research*, PMLR, Scottsdale, Arizona, USA, pp. 216–228.
URL: *http://proceedings.mlr.press/v31/dondelinger13a.html*

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K. & Tuschl, T. (2001), 'Duplexes of 21-nucleotide rnas mediate rna interference in cultured mammalian cells', *Nature* **411**, 494–498.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fire, A. (1999), 'Rna-triggered gene silencing', *Trends in Genetics* **15**(9), 358–363.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.
URL: *http://www.jstatsoft.org/v33/i01/*

Gugushvili, S. & Klaassen, C. A. (2012), 'n-consistent parameter estimation for

systems of ordinary differential equations: bypassing numerical integration via smoothing', *Bernoulli* **18**(3), 1061–1098.
URL: *http://dx.doi.org/10.3150/11-BEJ362*

Hale, E. T., Yin, W. & Zhang, Y. (2008), 'Fixed-point continuation for ℓ_1-minimization: Methodology and convergence', *SIAM J. Optim.* **19**, 1107–1130.

Hill, S. M., Heiser, L. M., Cokelaer, T., Unger, M., Nesser, N. K., Carlin, D. E., Zhang, Y., Sokolov, A., Paull, E. O., Wong, C. K., Graim, K., Bivol, A., Wang, H., Zhu, F., Afsari, B., Danilova, L. V., Favorov, A. V., Lee, W. S., Taylor, D., Hu, C. W., Long, B. L., Noren, D. P., Bisberg, A. J., HPN-DREAM-Consortium, Mills, G. B., Gray, J. W., Kellen, M., Norman, T., Friend, S., Qutub, A. A., Fertiq, E. J., Guan, Y., Song, M., Stuart, J. M., Spellman, P. T., Koeppl, H., Stolovitzky, G., Saez-Rodriguez, J. & Mukherjee, S. (2016), 'Inferring causal molecular networks: empirical assessment through a community-based effort', *Nat. Meth.* **13**, 310–318.

Horn, F. & Jackson, R. (1972), 'General mass action kinetics', *Archive for Rational Mechanics and Analysis* **47**(2), 81–116.
URL: *https://doi.org/10.1007/BF00251225*

Hynne, F., Danø, S. & Sørensen, P. G. (2001), 'Full-scale model of glycolysis in *Saccharomyces cerevisiae*', *Biophysical Chemistry* **94**, 121–163.

Liang, H. & Wu, H. (2008), 'Parameter estimation for differential equation models using a framework of measurement error in regression models', *Journal of the American Statistical Association* **103**(484), 1570–1583.
URL: *http://www.jstor.org/stable/27640205*

Lu, T., Liang, H., Li, H. & Wu, H. (2011), 'High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification', *Journal of the American Statistical Association* **106**(496), 1242–1258.

Michaelis, L. & Menten, M. L. (1913), 'Die kinetik der invertinwirkung', *Biochem Z.* **49**, 333–369.

Mikkelsen, F. R. (2015), 'Computational aspects of parameter estimation in ordinary differential equation systems', *Proceedings of the 19th European Young Statisticians Meeting* pp. 94–99.

Oates, C. J. & Mukherjee, S. (2012), 'Network inference and biological dynamics', *The Annals of Applied Statistics* **6**(3), 1209–1235.

Ramsay, J. O., Hooker, G., Campbell, D. & Cao, J. (2007), 'Parameter estimation for differential equations: a generalized smoothing approach', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(5), 741–796.
URL: *http://dx.doi.org/10.1111/j.1467-9868.2007.00610.x*

Santillan, M. (2008), 'On the use of the hill functions in mathematical models of gene regulatory networks', *Math. Model. Nat. Phenom.* **3**, 85–97.

Sauer, T. (2006), *Numerical Analysis*, Pearson, Boston.

Shinar, G. & Feinberg, M. (2010), 'Structural sources of robustness in biochemical reaction networks', *Science* **327**(5971), 1389–1391.
URL: *http://science.sciencemag.org/content/327/5971/1389*

Varah, J. M. (1982), 'A spline least squares method for numerical parameter estimation in differential equations', *SIAM J. Sci. Stat. Comput.* **3**, 28–46.

Wallace, E. W. J., Gillespie, D. T., Sanft, K. R. & Petzold, L. R. (2012), 'Linear noise approximation is valid over limited times for any chemical system that is sufficently large.', *IET Systems Biology* **6**(4), 102–115.

Wilkinson, D. J. (2006), *Stochastic modelling for systems biology*, Chapman & Hall/CRC Mathematical and Computational Biology Series, Chapman & Hall/CRC, Boca Raton, FL.

Wu, H., Lu, T., Xue, H. & Liang, H. (2014), 'Sparse additive ordinary differential equations for dynamic gene regulatory network modeling', *Journal of the American Statistical Association* **109**(506), 700–716.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society. Series B* **67**, 301–320.

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

E-mail address, Corresponding author: frm@math.ku.dk

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

II

Computational Aspects of Parameter Estimation in Ordinary Differential Equation Systems

Frederik Vissing Mikkelsen
Department of Mathematical Sciences
University of Copenhagen

Publication details

Published in *Proceedings of the 19th European Young Statisticians Meeting* (2015).

Computational Aspects of Parameter Estimation in Ordinary Differential Equation Systems

Frederik Riis Mikkelsen[*1]

[1]*Department of Mathematical Sciences, University of Copenhagen, Denmark*

Abstract: Ordinary differential equation (ODE) systems are widely applicable in many branches of the natural sciences. They are especially valuable for analysing entire networks of processes with no internal noise. Though simple from a statistical point of view, the applicability of these models are usually hindered by their computational complexity. In this work I present a selection of current methods to cope with the computational aspects of estimating parameters in ODE systems. Based on some of these methods, I present an algorithm for finding maximum likelihood estimates (MLE) with certain computational qualities.

Keywords: ODE systems, parameter estimation, non-linear least squares, computational statistics
AMS subject classifications: 62J02.

1 Introduction

We have in mind a d-dimensional ordinary differential equation system:

$$\dot{x} = f(x, \theta), \qquad x \in \mathbb{R}^d \tag{1}$$

parametrised by a p-dimensional vector $\theta \in \mathbb{R}^p$. For given θ, a solution to (1) is a function $\psi_\theta : \mathbb{R} \to \mathbb{R}^d$, such that

$$\psi_\theta(t) = \psi_\theta(0) + \int_0^t f(\psi_\theta(s), \theta)\, ds, \qquad \text{for all } t \in \mathbb{R}. \tag{2}$$

We observe the state of the system at discrete time points $0 = t_1 < t_2 < ... < t_n$ with independent Gaussian noise:

$$y_j = \psi_\theta(t_j) + \varepsilon_j, \qquad \varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d) \tag{3}$$

for $j = 1, ..., n$. The negative log-likelihood is directly available (σ^2 is omitted):

$$\ell_y(\theta) = \frac{1}{2} \sum_{j=1}^{n} \|y_j - \psi_\theta(t_j)\|_2^2. \tag{4}$$

This modelling framework for ODE systems is therefore quite simple from a statistical point of view. However, as explained below, optimising (4) is rather difficult from a computational angle.

*Corresponding author: frm@math.ku.dk

2 The Optimisation Problem

Finding the maximum likelihood estimator reduces to solving a non-linear least squares problem, due to (4). However, evaluating the likelihood requires solutions of the underlying ODE system. Various numerical methods for finding approximative solutions exist, but are relatively time consuming. Specifically, employing an explicit Runge-Kutta scheme of size s (see e.g. section 17 in [5] for details) the number of evaluations of f is $\frac{sT}{\delta}$. Here δ is the step size and T is the time span. For such a scheme the global truncation error is $\mathcal{O}(T\delta^p)$, for some scheme-dependent $p \leq s$. Using an implicit Runge-Kutta scheme, instead, leads to a smaller global truncation error, but raises the number of evaluations of f and is $\mathcal{O}\left(\frac{dsT}{\delta}\right)$ in best case scenarios.

The number of f-evaluations is substantial for assessing the computational complexity of evaluating ℓ_y. Though linear in each variable (considering $1/\delta$ as measuring the mesh), the number of f-evaluations is typically large. Consequently, in order to optimise (4) efficiently, a minimal number of evaluations of ℓ_y is preferable, especially when the observed time points cover a large time span or the ODE system is stiff.

3 Methods

3.1 Gauss-Newton Approach (shooting)

From a numerical optimisation perspective, ℓ_y has the valuable property of being a sum of squares. Thus the classical Gauss-Newton algorithm is typically a first choice for the optimisation scheme. The Gauss-Newton algorithm has the same rate of convergence as most second order approximation algorithms, but requires no computations of the hessian matrix (see e.g. section 10 in [6] for details). However, calculating the gradient of ℓ_y:

$$\nabla_\theta \ell_y(\theta) = -\sum_{j=1}^{n} (y_j - \psi_\theta(t_j))' D_\theta \psi(t_j) \tag{5}$$

amounts to deriving $D_\theta \psi$. This differential is typically only available as a solution to the matrix differential equation system

$$\dot{D_\theta \psi} = \frac{\partial f}{\partial x}(\psi(t), \theta) D_\theta \psi + \frac{\partial f}{\partial \theta}(\psi(t), \theta). \tag{6}$$

Consequently, employing the Gauss-Newton algorithm requires solving (1) and (6) simultaneously at each step. Using an explicit Runge-Kutta scheme of size s, this amounts to evaluating f, $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial \theta}$, $\frac{sT}{\delta}$ times at each step of the optimisation. Subsequently, for large and complex systems, evaluating the gradient of ℓ_y is either extremely time consuming or close to impossible.

There are numerous variations of the above approach (see, e.g., [3] for a more sophisticated version). They are often referred to as *shooting* methods, inspired by the shooting method from boundary value problems. These methods typically

rely on general optimisation algorithms and will therefore not exploit all essential features of ODE systems. The remaining algorithms incorporate these features by considering, e.g., the functional nature of the data.

3.2 Generalised Smoothing Approach (collocation)

Certain implicit Runge-Kutta schemes are so-called *collocation* methods. They rely on the principle that an approximative solution to (1) can be found in some finite dimensional function space, typically spanned by a set of spline functions. The collocation method therefore amounts to finding an element of the function space that satisfy (1) in some pre-specified time points, called *collocation points*. This approach is the inspiration to various parameter estimation methods in ODE systems. The following method is due to Ramsay et. al, see [4]:

This approach relies on the approximation of ψ_θ given by

$$\psi_\theta(t) \approx \varphi(t)'\hat{c}_\theta \qquad \text{for } t \in [0, t_n]. \tag{7}$$

Here φ is a vector of univariate basis functions, which combined with the vector $\hat{c}_\theta$ of coefficients yields an approximative solution to (1). The parameter dependence $\theta \mapsto \psi_\theta$ is therefore passed on to $\theta \mapsto \hat{c}_\theta$. The least squares criterion:

$$J(c, \theta) = \sum_j \|y_j - \varphi(t_j)'c\|_2^2 + \lambda \int_{t_1}^{t_n} \|\dot{\varphi}(t)'c - f(\varphi(t)'c, \theta)\|_2^2 \, dt \tag{8}$$

is proposed, where $\lambda > 0$ is a tuning parameter shifting the weight between the data fitting criterion and the so-called *fidelity measure* of $\varphi'c$. Applying profiling methods to (8), $\hat{c}_\theta$ appears as the minimum of $c \mapsto J(c, \theta)$. If f is linear in x, the minimisation problem reduces to a linear least squares problem, thus providing an analytical expression for $\theta \mapsto \hat{c}_\theta$.

By introducing this approach, some of the tools of functional data analysis is suddenly available, which provide new insightful views on the estimation problem. However, it is worth considering the influence of the choice of φ on the inference. Moreover, the relation between optimising a family of semi-norms parametrised by λ (the criterion J in (8)) and the actual MLE defined through (4) is not completely clear. Finally, for non-linear systems, optimising $c \mapsto J(c, \theta)$ and $\theta \mapsto J(\hat{c}_\theta, \theta)$ using gradient based methods still require evaluating $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial \theta}$ many times (depending on how the integral in (8) is approximated).

3.3 Gradient/integral Matching

The core principle of this method is: if the whole noiseless curve ψ is observed, then θ can be inferred by minimising

$$\int_{t_1}^{t_n} \left\| \dot{\psi}(t) - f(\psi(t), \theta) \right\|_2^2 dt, \quad \text{or} \quad \int_{t_1}^{t_n} \left\| \psi(t) - \psi(t_1) - \int_{t_1}^{t} f(\psi(s), \theta) ds \right\|_2^2 dt. \tag{9}$$

We therefore consider the estimator, that takes a non-parametric estimate of ψ, $\hat{\psi}$, and returns the value of θ minimising (9):

$$\hat{\psi} \mapsto \arg\min_{\theta} \int_{t_1}^{t_n} \left\| \hat{\psi}(t) - \psi(t_1) - \int_{t_1}^{t} f(\hat{\psi}(s), \theta) ds \right\|_2^2 dt. \tag{10}$$

If $\psi(t_1)$ is unknown, it can be included in the parameter vector θ. In [1] the author proves that if the above map is applied to a consistent non-parametric estimator, the resulting estimator of θ is also consistent, under mild regularity assumptions. Additionally, he also finds conditions for asymptotic normality.

This approach truly flourishes when applied to systems in which f is linear in θ (and not necessarily linear in x). In such cases (10) reduces to a linear least squares problem, which can be solved even for very large and complex systems, i.e., for large n and p. Furthermore, one can introduce, e.g., ℓ^1-penalties to (10) and apply the method to systems with $p >> nd$ and still have computationally stable methods for finding solutions.

Brewer et al. ([2]) proposed an iterative procedure exploiting the qualities of this type of gradient matching. More precisely, they consider a fitting criterion resembling that of [4]:

$$\sum_j \|y_j - \varphi(t_j)'c\|_2^2 + \lambda \sum_r \|\dot{\varphi}(t_r)'c - f(\varphi(t_r)'\tilde{c}, \theta)\|_2^2 \tag{11}$$

where r is allowed to run over a finer (or coarser) grid than j. The iterations consist of letting $\tilde{c}$ be fixed and then estimate (c, θ) as the linear least squares estimates of (11). The estimate of c then enters as $\tilde{c}$ in the next iteration. By applying gradient matching iteratively one avoids choosing a specific $\hat{\psi}$, as opposed to a non-iterative gradient matching.

Similarly to the generalised smoothing approach, this method has the following important strength: the optimisation problem and the ODE-solution problem are separated. Thus evaluating the fitting criterion is inexpensive and evaluating the gradient (typically) requires less calculations of $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial \theta}$.

The method described above is applicable to many non-trivial systems and can handle large and sparse models. However, there are things to consider: the iterative procedure is still dependent on the choice of φ. It is also unclear how optimising the criterion (11) is related to the MLE given through (4). Finally, it is nontrivial whether this sequence of iterative estimates of (c, θ) converge to the optimum of (11) (if it converges at all).

4 Combining Algorithms

Returning to the original problem of minimising (4), we required relatively few evaluations of ℓ_y and $\nabla \ell_y$. In this section we consider a new algorithm based on the above, which yields the actual MLE (a quality of the shooting methods) and still exploits the computationally attractive aspects of gradient matching.

Firstly, given a current estimate of θ, denoted θ_k in the iterative procedure, we calculate an approximative solution curve ψ_{θ_k}. Then we perform integral matching

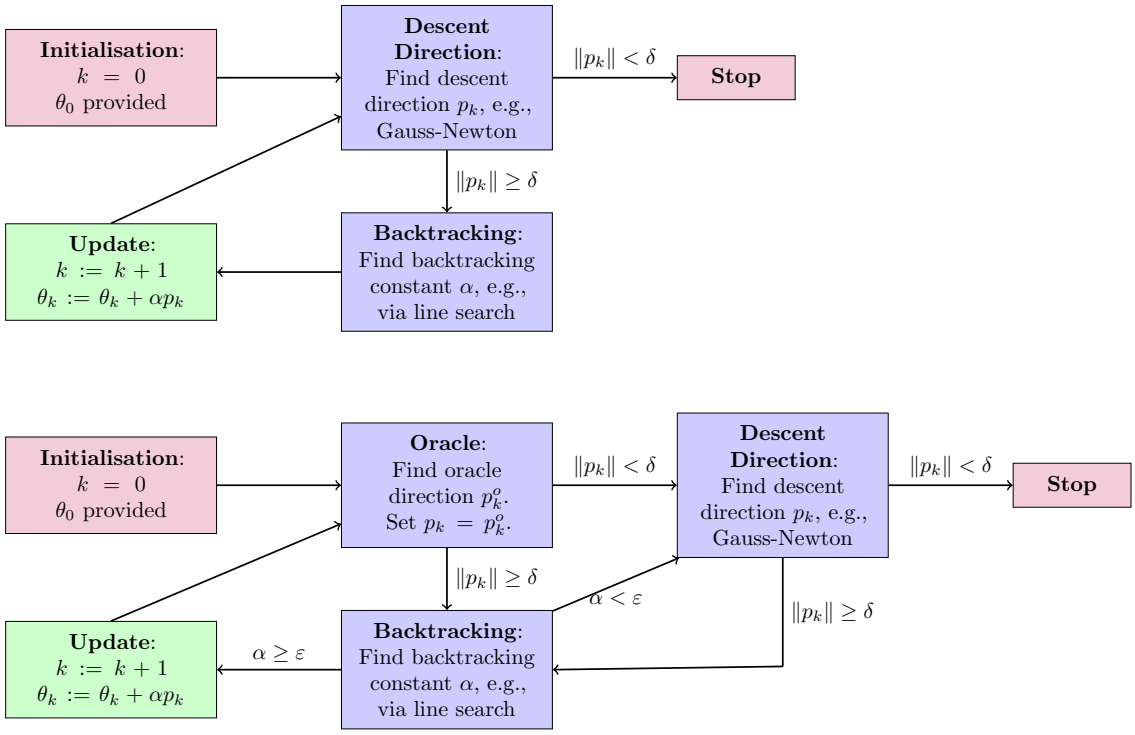


Figure 1: Flowcharts of a generic line search algorithm with and without an oracle.

between the curve and the observations. The resulting estimate of θ:

$$\theta_k^{\text{oracle}} = \arg\min_{\theta} \sum_{j=1}^{n} \left\| y_j - \psi(t_1) - \int_{t_1}^{t_j} f(\psi_{\theta_k}(s), \theta) ds \right\|_2^2. \tag{12}$$

is called the *oracle* estimate. We denote $p_k^{\text{oracle}} = \theta_k^{\text{oracle}} - \theta_k$ the *oracle* direction. In general it is not certain that $\ell_y(\theta_k^{\text{oracle}}) < \ell_y(\theta_k)$, hence a backtracking of p_k^{oracle} must be employed in order to gain a descent:

$$\theta_{k+1} = \theta_k + \alpha p_k^{\text{oracle}}$$

for some $\alpha \in [0, 1]$. However, it is not even certain that p_k^{oracle} is a descent direction! In which case, the backtracking will fail to find a positive α within numerical tolerance. In this case, no benefit from the oracle is gained. The algorithm then passes on to some classic optimisation scheme, e.g., Gauss-Newton. Once a single Gauss-Newton update is done, the algorithm returns to the oracle for the next iterate. A generic line search optimisation algorithm with and without an oracle are visualised by two flowcharts in figure 1.

This algorithm maintains the convergence properties of the Gauss-Newton algorithm, while benefiting from computational advantages possessed by the oracle. In practice the oracle mostly provides excellent descent directions and the Gauss-Newton part will only be invoked to verify that the final iterate is an approximative local minima.

5 Discussion and Further Work

The new combined algorithm presented above has been implemented and tested on simulated data from mass action kinetics models. The results look promising both for large and small σ^2, along with high and low frequency data. In these studies the oracle always provided a descent direction. Intuitively this is not true in general, as the algorithm will perform poorly for *stiff* or *chaotic* systems. The computationally heavy part of the algorithm is the Gauss-Newton part. Consequently, it is of high interest to find conditions that ensure the oracle alone provides the descent directions necessary to find MLE.

Additionally, the algorithm can be extended to parameter estimation with forced sparsity, e.g., using ℓ^p penalties. This is relevant for estimating unknown model structures. However, when introducing such penalties the algorithm has to be revised in order to accommodate potential lack of smoothness.

Acknowledgements: A great thanks to Professor Niels Richard Hansen and Martin Vincent of Department of Mathematical Sciences at University of Copenhagen for guidance and many insightful discussions.

References

[1] N. J-B. Brunel. Parameter estimation of ODE's via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267, 2008.

[2] D. Brewer, M. Barenco, R. Collard, M. Hubank and J. Stark. Fitting ordinary differential equations to short time course data. *Philos. Transact. A Math. Phys. Eng. Sci.*, 366:519–544, 2008.

[3] Z. Li, M. R. Osborne and T. Prvan. Parameter estimation of ordinary differential equations. *IMA Journal of Numerical Analysis*, 25(2):264–285, 2005.

[4] J. O. Ramsay, G. Hooker, D. Campbell and J. Cao. Parameter estimation for differential equations: a generalised smoothing approach. *J. R. Statist. Soc. B*, 69(5):741–796, 2007.

[5] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling. *Numerical Recipes: The Art of Scientific Computing* (3^{rd} ed). Cambridge University Press, 2007.

[6] J. Nocedal and S. J. Wright. *Numerical Optimization* (2^{nd} ed). Springer, 2006.

III

A Model Based Rule for Selecting Spiking Thresholds in Neuron Models

Frederik Vissing Mikkelsen
Department of Mathematical Sciences
University of Copenhagen

Publication details

Published in *Mathematical Biosciences and Engineering*, 13(3), (2016).

Manuscript submitted to
AIMS' Journals
Volume **13**, Number **3**, June **2016**

doi:10.3934/mbe.2016008

pp. **569–578**

A MODEL BASED RULE FOR SELECTING SPIKING THRESHOLDS IN NEURON MODELS

Frederik Riis Mikkelsen

Department of Mathematical Sciences, University of Copenhagen
Universitetsparken 5
Copenhagen, 2100, Denmark

Abstract. Determining excitability thresholds in neuronal models is of high interest due to its applicability in separating spiking from non-spiking phases of neuronal membrane potential processes. However, excitability thresholds are known to depend on various auxiliary variables, including any conductance or gating variables. Such dependences pose as a double-edged sword; they are natural consequences of the complexity of the model, but proves difficult to apply in practice, since gating variables are rarely measured.

In this paper a technique for finding excitability thresholds, based on the local behaviour of the flow in dynamical systems, is presented. The technique incorporates the dynamics of the auxiliary variables, yet only produces thresholds for the membrane potential. The method is applied to several classical neuron models and the threshold's dependence upon external parameters is studied, along with a general evaluation of the technique.

1. **Introduction.** One of the most essential properties of a neuronal model is its ability to capture both the *active* spiking phases (fast large-amplitude oscillations) and the *inactive* resting phases (weakly nonlinear oscillations), [4]. The concept of *excitability thresholds*, which in general is not well defined, is an *ad hoc* characteristic separating these two "domains". Excitability threshold are, nevertheless, essential in many applications, e.g., when studying membrane potentials data is often separated into active and inactive phases and analysed accordingly.

A vast selection of literature exists on the subject of choosing spiking thresholds in neuronal models. These include both experimentally based approaches and purely theoretical constructions based on some class of models. The focus of this paper is the latter. For experimentally based solutions see, e.g., [19] for a thorough presentation and benchmarking of some of the most common methods.

Some of the theoretical model based approaches rely on differential geometry and prove highly relevant in the context of detecting distinctive behaviour in dynamical systems, see e.g., [7]. A classical approach is studying the *inflection* sets of the system, i.e., the region of the state space at which trajectories have vanishing curvature. This approach for detecting excitability thresholds was first proposed in 1976 ([15]), where it was applied to the BonhoefferVan der Pol model. It was later reintroduced in [16] in connection to *canards* in chemical systems. For thorough

2010 *Mathematics Subject Classification.* Primary: 37C10, 65L05; Secondary: 92C20.

Key words and phrases. Dynamical Systems, Hodgkin-Huxley, Excitability, Neuron Modelling, Spiking, Threshold Selection.

treatment in relation to excitability and canards, see e.g., [4] and [21]. As discussed in [4], inflection methods are limited to planar systems.

Another type of model based threshold characterisation relies on studying the steepest slope of the membrane potential process. This, however, crucially depends on the state of the gating variables. Such dependences is studied in [18], in which the authors provide a threshold equation that yields an instantaneous threshold value as a function of the underlying ionic channel conductance. In [20] the excitability threshold in neuronal models is successfully captured as manifolds in the state space and thus stressing the dependence of the threshold on activation and inactivation variables.

The technique presented in this paper focuses on multidimensional neuronal models, in which an excitability threshold solely for the membrane potential is desired. This is favourable, as the gating variables are rarely measured. Though the technique produces a threshold independent of the gating variables, it still captures the overall dynamics of the system.

The paper is outlined as follows: in section 2 the general framework is established along with the fundamental assumptions of the model. Moreover, the excitability threshold technique is motivated and derived. In section 3 the threshold rule is applied to six different neuron models for varying parameter settings. Finally, the advantages, drawbacks and general evaluation of the method are considered in section 4.

2. Framework and Construction.

Let E and Θ denote open subsets of $\mathbb{R}^d$ and $\mathbb{R}^p$, respectively. Let $f : E \times \Theta \mapsto \mathbb{R}^d$ be a C^1-function and consider the *initial value problem*:

$$\dot{x} = f(x, \theta), \quad x(0) = x_0, \tag{1}$$

with $x_0 \in E$ and $\theta \in \Theta$. A *solution* or *trajectory* of (1) is a function $\varphi : \mathbb{R} \times E \times \Theta \to E$ satisfying

$$\varphi(t, x_0, \theta) = \varphi(0, x_0, \theta) + \int_0^t f(\varphi(s, x_0, \theta), \theta)\, ds, \qquad \text{for all } t \in \mathbb{R}. \tag{2}$$

We say that f constitutes a *dynamical system* on the state space E. Standard existence and uniqueness results for solutions to (1) can be found in, e.g., [17]. Unless relevant, θ will be dropped from the notation.

We will consider any neuronal model given as a dynamical system. We assume the first coordinate of x represents the electrical potential, denoted by v. Let u denote the additional variables. We therefore have the tensor structure: $x = (v, u)$, $f = (f_v, f_u)$ and $\varphi = (\varphi_v, \varphi_u)$. Additionally, we assume f is C^2.

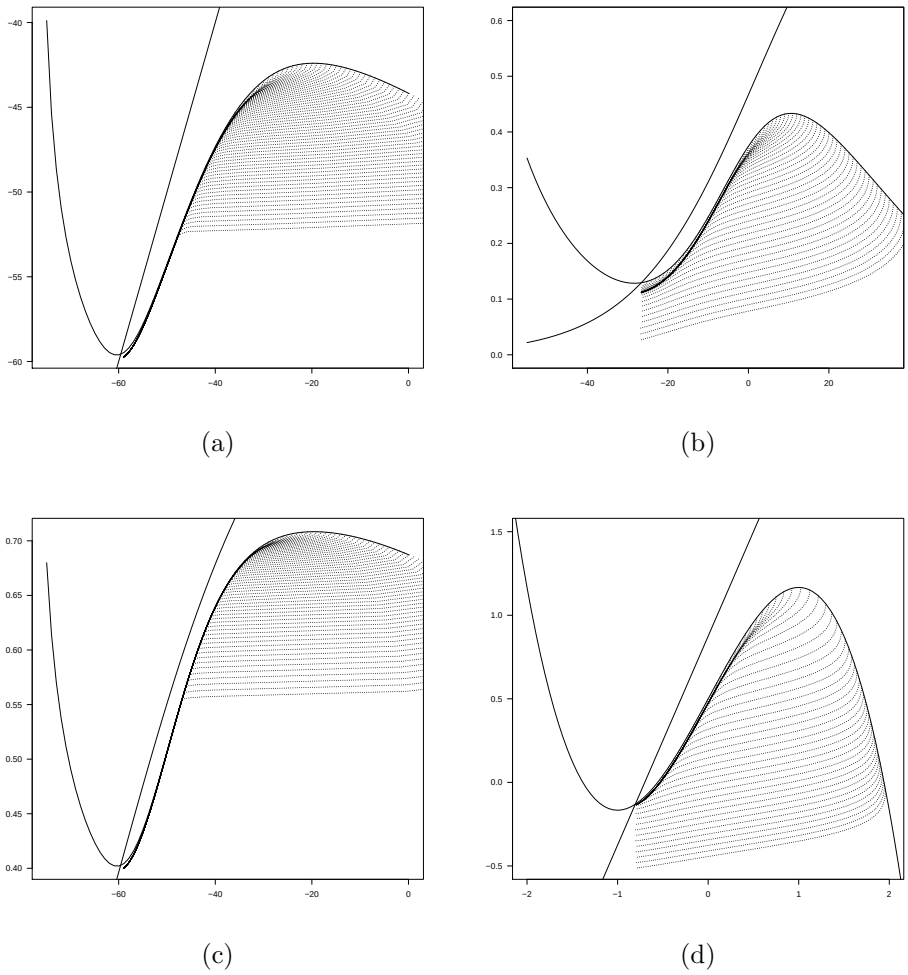
In the following we need the manifold:

$$\mathcal{N} := \{x \in E \mid f_v(x) = 0\}, \tag{3}$$

known as the *v-nullcline*. Any trajectory of the system has its marginal v-stationary points on $\mathcal{N}$. Consequently, all spikes occur on the manifold $\mathcal{N}$, thus stressing the importance of $\mathcal{N}$ in relation to excitability thresholds.

In figure 1 trajectories in reverse time of different system are plotted (see section 3 and appendix A for details). Though they are initialised on equally spaced points on $\mathcal{N}$ we observe clustering of the trajectories.

Because the flow is in reverse, trajectories will be highly sensitive towards initial conditions if initialised close to the clustered trajectories. Trajectories to the left curve towards the inactive regime and trajectories to the right curve towards the



(a) (b)

(c) (d)

FIGURE 1. State space of different bi-dimensional models with drawn nullclines. The dotted lines are trajectories initialised on $\mathcal{N}$ and runs in negative time flow. (a) the Abbott-Kepler reduction, (b) the Morris-Lecar model, (c) the Kokoz-Krinskii reduction, (d) the Fitzhugh-Nagumo model.

active regime. Hence, in this figure, the clustered trajectories are closely related to the inflection sets of the state spaces. Additionally, they are also closely linked to the manifolds studied in [20]. This clustering of trajectories is therefore closely linked to the transition between active and inactive phases and thus builds the foundation of the threshold technique outlined below.

Generally, the clustering is characterised by how the flow scales volumes. In figure 1 equally spaced initialisations on $\mathcal{N}$ (corresponding to equal volumes) are squeezed together or separated from each other by the flow. Consequently, clustering of trajectories occurs when volumes are considerably scaled down by the reverse flow, or, scaled up by the non-reversed flow.

Formally, the scaling of volumes is quantified by the Jacobian of the flow $x \mapsto \varphi(t,x)$:

$$J(t,x) := \left|\det(\partial_x \varphi(t,x))\right|. \tag{4}$$

In order to make this characteristic tractable, we consider the infinitesimal scaling of volumes at $t = 0$. Standard results yield:

$$\begin{aligned} \det(\partial_x \varphi(t,x)) &= \det\left(\partial_x \varphi(0,x) + t\partial_t \partial_x \varphi(0,x) + t\varepsilon(t,x)\right) \\ &= \det\left(I + t\left[\partial_x f(x) + \varepsilon(t,x)\right]\right) \\ &= 1 + t \cdot \operatorname{trace}(\partial_x f(x) + \varepsilon(t,x)) + \mathcal{O}(t^2) \end{aligned} \tag{5}$$

for some function ε vanishing at $t = 0$. Consequently,

$$\partial_t J(0,x) = \nabla f(x), \tag{6}$$

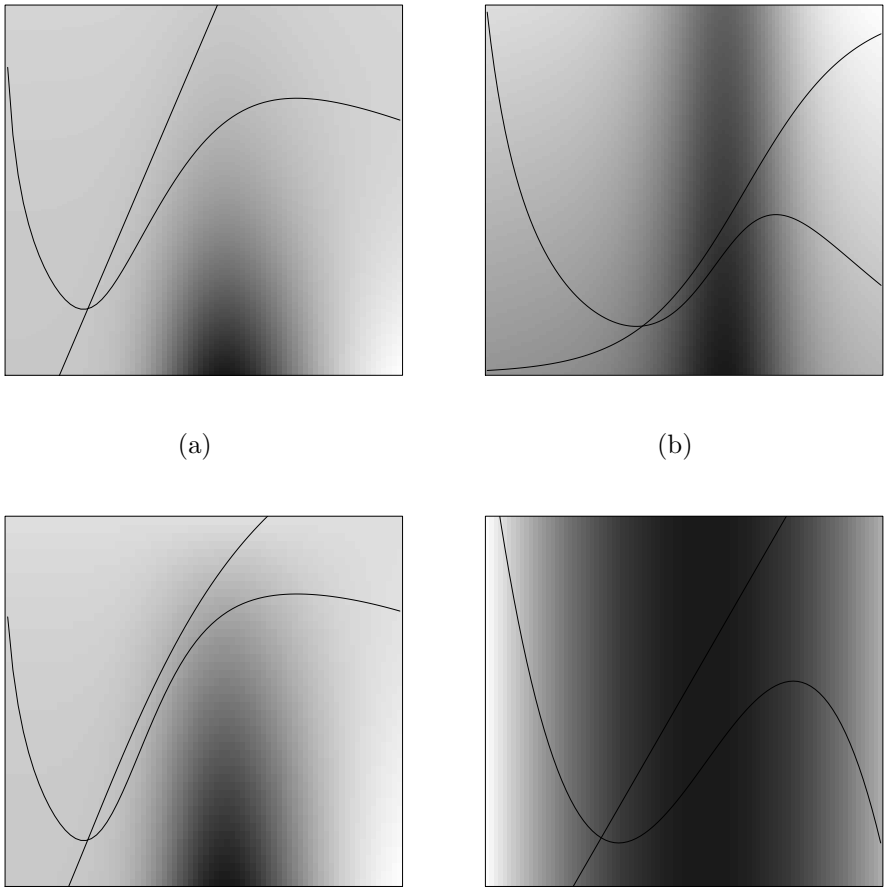
where ∇ is the divergence operator. Motivated by the above, the proposed threshold, v_{thr}, is given by

$$v_{\text{thr}} := \pi_v \left(\operatorname*{argmax}_{x \in \mathcal{N}} \nabla f(x) \right), \tag{7}$$

where π_v is the projection onto the v-coordinate.

Clearly, the applicability of v_{thr} relies on existence and uniqueness of a maximal argument for $\nabla f(x)$ on $\mathcal{N}$. For instance, v_{thr} does not exist in linear systems. However, as seen in section 3, v_{thr} is well defined in the classical neuron models. Moreover, heat maps of ∇f for four bi-dimensional models is presented in figure 2.

As mentioned, one of the most important properties of a neuronal model is its ability to capture both active and inactive phases and their separate distinctive behaviour. Furthermore, the transition between the two regimes must be fast, otherwise the model does not sufficiently reflect the "all-or-none"-principle of excitability in neurons. From the model's perspective this implies that only finely tuned initialisations exhibits local v-maxima that can neither be considered active nor inactive. Such crossings of $\mathcal{N}$ are exactly those having large values of ∇f and thus captured by (7). Therefore v_{thr} is well defined if the model considered sufficiently mimics the "all-or-none" principle.



(a) (b)

(c) (d)

FIGURE 2. Heat maps of ∇f on the state space of different bi-dimensional models with drawn null-clines: (a) the Abbott-Kepler reduction, (b) the Morris-Lecar model, (c) the Kokoz-Krinskii reduction, (d) the Fitzhugh-Nagumo model. Darker colours mean larger ∇f value.

3. **Examples.** In the following we investigate v_{thr} for different models and parameter configurations. In practice v_{thr} is determined by studying the marginalisation

$$v \mapsto \sup_{u:(v,u)\in\mathcal{N}} \nabla f(v,u). \tag{8}$$

The models considered here are: the Hodgkin-Huxley (HH) model ([8]), the Connor-Stevens (CS) model ([2]), the Abbot-Kepler (AK) reduction ([1], [10]), the Kokoz-Krinskii (KK) reduction ([11]), the Morris-Lecar (ML) model ([13]) and the Fitzhugh-Nagumo (FHN) model ([6], [14]). All models considered are specified in appendix A.

In the case of the Hodgkin-Huxley and the Connor-Stevens model computing v_{thr} is especially simple, as evaluating (8) amounts to solving a linear programming (lp) problem. All models, except for the Fitzhugh-Nagumo model, do not admit closed form expressions for v_{thr} and are evaluated numerically. The results for varying input current, I, are visualised in figure 3. The excitability threshold suggested by the above technique lies at the typically proposed level for the different models. Moreover, the threshold increases with I.

For the Hodgkin-Huxley and the Connor-Stevens model a sudden change occur around $I = 3.3$ and $I = -4.9$, respectively. As indicated, these are not discontinuities, but are the results of changing active constraints in the lp problem of evaluating (8). The sudden change is not associated with bifurcations (the closest bifurcation takes place at $I = 9.78$ for the HH model, see [12]). In fact, the threshold rule seems to be unaffected by any of the bifurcations occurring when tuning I. This emphasises that v_{thr} is not related to whether the system promotes spiking behaviour or not, but how local v-maxima are separated by the dynamics.

Finally, we consider the Fitzhugh-Nagumo model. Straightforward calculations yield $v_{\text{thr}} = 0$, hence the threshold in this particular model is independent of the parameters. Again, this stresses the interpretation of v_{thr}; it measures where the mimicking of the "all-or-none"-principle is most prominent in the model. Hence, for varying parameters, $v = 0$ still acts as the separation of local v-maxima.

4. **Discussion.** In the above an excitability threshold for the membrane potential in neuronal models has been presented. It applies to multidimensional neuronal models given as ODEs and is relatively easy to evaluate. The threshold rule relies on the same classic considerations behind other threshold rules, e.g., [4] and [20]. While still incorporating the full dynamics of the system, it provides thresholds in v only. Additionally, requirements of bi-dimensionality, imposed in e.g., [4], is not necessary.

A drawback of the threshold rule is that it may not fully capture the complexity of models, as more sophisticated manifold based threshold rules do, e.g., as in [20]. However, the above presented technique applies to situations in which only the membrane potential is observable.

Finally, the main drawback of the technique is that it only applies whenever $\operatorname{argmax}_{x\in\mathcal{N}} \nabla f(x)$ is well defined. However, as pointed out in section 2, if the neuron model captures the "all-or-none"-principle sufficiently well, then ∇f will be large whenever a transition from inactive to active phases occurs.

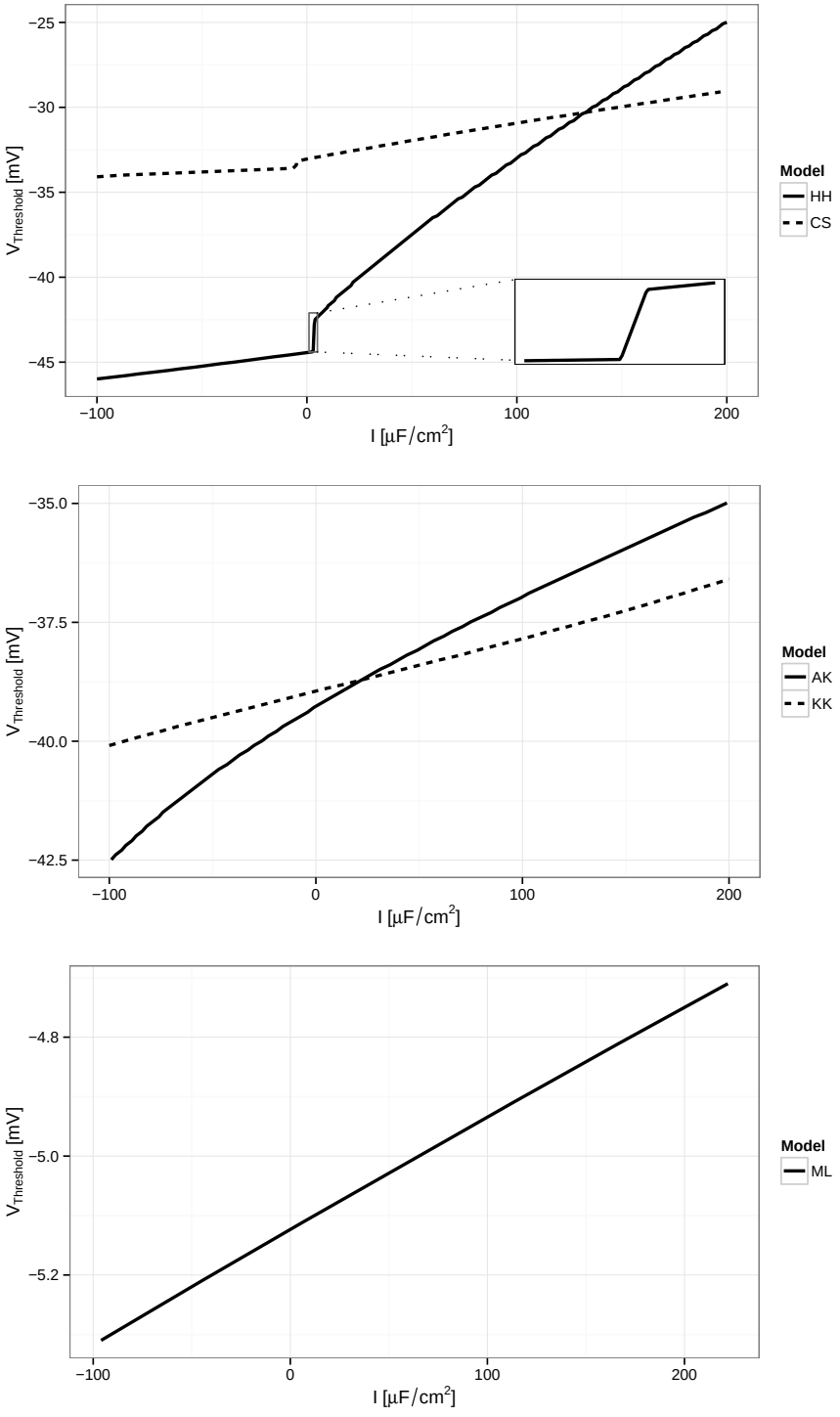


FIGURE 3. Values of v_{thr} for varying input current I and different models. The models are Hodgkin-Huxley (HH), Connor-Stevens (CS), Abbott-Kepler (AK), Kokoz-Krinskii (KK) and Morris-Lecar (ML).

Acknowledgments. A great thanks to Susanne Ditlevsen for supervising the project that led to these results.

REFERENCES

[1] L. Abbott and T. Kepler, *Lect. Notes Phys.*, **368** (1990), 5.
[2] J.A. Connor and C.F. Stevens, Prediction of repetitive firing behaviour from voltage clamp data on an isolated neurone soma., *J. Physiol*, **213** (1971), 31–53.
[3] P. Dayan and L. Abbott, *Theoretical Neuroscience - Computational and Mathematical Modeling of Neural Systems*, The MIT Press, 2005.
[4] M. Desroches, M. Krupa and S. Rodrigues, Inflection, canards and excitability threshold in neuronal models, *J. Math. Biol.*, **67** (2012), 989–1017.
[5] S. Ditlevsen and P. Greenwood, The morris-lecar neuron model embeds a leaky integrate-and-fire model, *J. Math. Biol.*, **67** (2012), 239–259.
[6] R. Fitzhugh, Impulses and physiological states in theoretical models of nerve membrane, *Biophys. J*, **1** (1961), 445–466.
[7] J. Ginoux and B. Rossetto, Differential geometry and mechanics: applications to chaotic dynamical systems, *Int. J. Bifurcat. Chaos*, **16** (2006), 887–910.
[8] A. Hodgkin and A. Huxley, A quantitative description of the membrane current and application to conduction and excitation in nerve, *J. Physiol.*, **117** (1952), 500–544.
[9] E. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, The MIT Press, 2007.
[10] T. Kepler, L. Abbott and E. Marder, Membranes with the Same Ion Channel Populations but Different Excitabilities, *Biol. Cybern.*, **66** (1992), 381.
[11] V.I. Krinsky and Yu.M. Kokoz, *Biofizika*, **18** (1973), 506.
[12] C. Meunier, Two and three dimensional reductions of the Hodgkin-Huxley system: separation of time scales and bifurcations, *Biol. Cybern.*, **67** (1992), 461–468.
[13] C. Morris and H. Lecar, Voltage oscillations in the barnacle giant muscle fiber, *Biophys. J*, **35** (1981), 193–213.
[14] J. Nagumo, S. Arimoto and S. Yoshizawa, An active pulse transmission line simulating nerve axon, *Proc IRE*, **50** (1962), 2061–2070.
[15] M. Okuda, New method of nonlinear analysis for shaping and threshold actions, *J Phys. Soc. Jpn*, **41** (1976), 1815–1816.
[16] B. Peng, V. Gaspar and K. Showalter, False bifurcations in chemical systems: canards, *Phil. Trans. R Soc. Lond A*, **337** (1991), 275–289.
[17] L. Perko, *Differential Equations and Dynamical Systems*, Texts in Applied Mathematics 7, Springer, 3rd edition, 2000.
[18] J. Platkiewicz and R. Brette, A threshold equation for action potential initiation, *PLoS Comput. Biol.*, **6** (2010).
[19] M. Sekerli, C. Del Negro, R. Lee and R. Butera, Estimating action potential thresholds from neuronal time-series: New metrics and evaluation of methodologies, *IEEE T Bio. Med. Eng.*, **51** (2004), 1665–1672.
[20] A. Tonnelier, Threshold curve for the excitability of bidimensional spiking neurons, *Phys. Rev. E*, **90**, 2, 022701 (2014).
[21] M. Wechselberge, J. Mitry and J. Rinzel, Canard Theory and Excitability, *Nonautonomous Dynamical Systems in the Life Sciences*, Ch. 3, Springer, 2013.

Appendix A. Models.

A.1. **The Hodgkin-Huxley Model.** The Hodgkin-Huxley model, first presented in 1952 (see [8]), is the most influential model in neuroscience. It is a four-dimensional dynamical model governed by the following dynamics:

$$\begin{aligned} \dot{v} &= \frac{I - g_{\mathrm{L}}(v - V_{\mathrm{L}}) - g_{\mathrm{Na}} m^3 h (v - V_{\mathrm{Na}}) - g_{\mathrm{K}} n^4 (v - V_{\mathrm{K}})}{C}, \\ \dot{n} &= \frac{n_\infty(v) - n}{\tau_n(v)}, \quad \dot{m} = \frac{m_\infty(v) - m}{\tau_m(v)}, \quad \dot{h} = \frac{h_\infty(v) - h}{\tau_h(v)}. \end{aligned} \tag{9}$$

Here $\tau_x = 1/(\alpha_x + \beta_x)$ and $x_\infty = \alpha_x/(\alpha_x + \beta_x)$ for $x = n, m, h$ and

$$\begin{aligned} \alpha_n(v) &= \frac{\frac{v+55}{100}}{1 - \exp\left(-\frac{v+55}{10}\right)}, & \beta_n(v) &= \frac{\exp\left(-\frac{v+55}{10}\right)}{8}, \\ \alpha_m(v) &= \frac{\frac{v+40}{10}}{1 - \exp\left(-\frac{v+40}{10}\right)}, & \beta_m(v) &= 4\exp\left(-\frac{v+65}{18}\right), \\ \alpha_h(v) &= 0.07\exp\left(-\frac{v+65}{20}\right), & \beta_h(v) &= \frac{1}{1 + \exp\left(-\frac{v+35}{10}\right)}. \end{aligned} \tag{10}$$

The parameters are listed in table 1 and taken from [12].

TABLE 1. Parameter values for the Hodgkin-Huxley model.

Parameter	Value	Parameter	Value
g_{Na}	120 mS/cm^2	V_{Na}	50 mV
g_{K}	36 mS/cm^2	V_{K}	-77 mV
g_{L}	0.3 mS/cm^2	V_{L}	-54.4 mV
C	1 μF/cm^2		

A.2. **The Fitzhugh-Nagumo Model.** The first phenomenological model resembling the Hodgkin-Huxley dynamics was proposed by Fitzhugh ([6]) and Nagumo et. al ([14]) independently. It reads:

$$\begin{aligned} \dot{v} &= v - \frac{v^3}{3} - u + I, \\ \dot{u} &= \frac{v + a - bu}{\tau}. \end{aligned} \tag{11}$$

The parameters used in figure 1 and 2 are $a = 0.7$, $b = 0.8$, $I = 0.5$ and $\tau = 12.5$.

A.3. **The Connor-Stevens Model.** The Connor-Stevens model ([2]) extends the model of Hodgkin and Huxley with a transient potassium current.

$$\begin{aligned} \dot{v} &= \frac{I - g_{\mathrm{L}}(v - V_{\mathrm{L}}) - g_{\mathrm{Na}} m^3 h (v - V_{\mathrm{Na}}) - g_{\mathrm{K}} n^4 (v - V_{\mathrm{K}}) - g_A a^3 b (v - V_A)}{C}, \\ \dot{n} &= \frac{n_\infty(v) - n}{\tau_n(v)}, \quad \dot{m} = \frac{m_\infty(v) - m}{\tau_m(v)}, \quad \dot{h} = \frac{h_\infty(v) - h}{\tau_h(v)}, \\ \dot{a} &= \frac{a_\infty(v) - a}{\tau_a(v)}, \quad \dot{b} = \frac{b_\infty(v) - b}{\tau_b(v)}. \end{aligned} \tag{12}$$

Here $\tau_x = 1/(\alpha_x + \beta_x)$ and $x_\infty = \alpha_x/(\alpha_x + \beta_x)$ for $x = n, m, h$ and

$$\begin{aligned}
\alpha_n(v) &= \frac{0.02(v+45.7)}{1-\exp\left(-0.1(v+45.7)\right)}, & \beta_n(v) &= 0.25\exp\left(-0.0125(v+55.7)\right),\\
\alpha_m(v) &= \frac{0.38(v+29.7)}{1-\exp\left(-0.1(v+29.7)\right)}, & \beta_m(v) &= 15.2\exp\left(-0.0556(v+54.7)\right),\\
\alpha_h(v) &= 0.266\exp\left(-0.05(v+48)\right), & \beta_h(v) &= \frac{3.8}{1+\exp\left(-0.1(v+18)\right)},\\
a_\infty(v) &= \left[\frac{0.0761\exp(0.0314(v+94.22))}{1+\exp(0.0346(v+1.17))}\right]^{\frac{1}{3}}, & \tau_a(v) &= 0.3632 + \frac{1.158}{1+\exp(0.0497(v+55.96))},\\
b_\infty(v) &= \left[\frac{1}{1+\exp(0.0688(v+53.3))}\right]^4, & \tau_b(v) &= 1.24 + \frac{2.678}{1+\exp(0.0624(v+50))}.
\end{aligned} \tag{13}$$

The parameters are listed in table 2 and taken from [3].

TABLE 2. Parameter values for the Connor-Stevens model.

Parameter	Value	Parameter	Value
g_{Na}	120 mS/cm^2	V_{Na}	55 mV
g_{K}	20 mS/cm^2	V_{K}	-72 mV
g_{L}	0.3 mS/cm^2	V_{L}	-17 mV
g_{A}	47.7 mS/cm^2	V_{A}	-75 mV
C	1 μF/cm^2		

A.4. **The Kokoz-Krinskii Reduction.** Kokoz and Krinskii provided a more realistic model mimicking the dynamics of the Hodgkin-Huxley model in [11]. It relies on two reductions: m is assumed instantaneous and the sum of the slower variables h and n remain constant at level K. The dynamics therefore reduces to

$$\begin{aligned}
\dot{v} &= \frac{I - g_{\mathrm{L}}(v-V_{\mathrm{L}}) - g_{\mathrm{Na}}m_\infty(v)^3(K-n)(v-V_{\mathrm{Na}}) - g_{\mathrm{K}}n^4(v-V_{\mathrm{K}})}{C},\\
\dot{n} &= \frac{n_\infty(v)-n}{\tau_n(v)}.
\end{aligned} \tag{14}$$

We set $K = 0.8$ and the rest of the specifications are as in section A.1.

A.5. **The Morris-Lecar Model.** Another classical example of a conductance based neuron model is the Morris-Lecar model, see [13] for details. The dynamics are as follows:

$$\begin{aligned}
\dot{v} &= \frac{I - g_{\mathrm{L}}(v-V_{\mathrm{L}}) - g_{\mathrm{Ca}}m_\infty(v)(v-V_{\mathrm{Ca}}) - g_{\mathrm{K}}u(v-V_{\mathrm{K}})}{C},\\
\dot{u} &= \alpha(v)(1-u) - \beta(v)u,
\end{aligned} \tag{15}$$

where

$$\begin{aligned} m_\infty(v) &= \frac{1}{2}\left(1+\tanh\left(\frac{v-V_1}{V_2}\right)\right), \\ \alpha(v) &= \frac{1}{2}\phi\cosh\left(\frac{v-V_3}{2V_4}\right)\left(1+\tanh\left(\frac{v-V_3}{V_4}\right)\right), \\ \beta(v) &= \frac{1}{2}\phi\cosh\left(\frac{v-V_3}{2V_4}\right)\left(1-\tanh\left(\frac{v-V_3}{V_4}\right)\right). \end{aligned} \tag{16}$$

The parameter values used in this example are taken from [5] and are given in table 3.

TABLE 3. Parameter values for the Morris-Lecar model.

Parameter	Value	Parameter	Value	Parameter	Value
V_1	-1.2 mV	g_{Ca}	4.4 μS/cm^2	V_{Ca}	120mV
V_2	18 mV	g_{K}	8 μS/cm^2	V_{K}	-84 mV
V_3	2 mV	g_{L}	2 μS/cm^2	V_{L}	-60 mV
V_4	30 mV	C	20 μF/cm^2	ϕ	0.04 ms^{-1}

A.6. **The Abbot-Kepler Reduction.** The reduction of the Hodgkin-Huxley model given below is just one of many possible reductions. They all follow the same principle proposed by Abbot and Kepler ([1], [10]). In this paper we consider

$$\begin{aligned} \dot{v} &= \frac{I - g_{\mathrm{L}}(v-V_{\mathrm{L}}) - g_{\mathrm{Na}}m_\infty(v)^3h_\infty(u)(v-V_{\mathrm{Na}}) - g_{\mathrm{K}}n_\infty(u)^4(v-V_{\mathrm{K}})}{C}, \\ \dot{u} &= \alpha(v,u)\frac{h_\infty(v)-h_\infty(u)}{\tau_h(v)h'_\infty(u)} + (1-\alpha(v,u))\frac{n_\infty(v)-n_\infty(u)}{\tau_n(v)n'_\infty(u)}, \end{aligned} \tag{17}$$

where

$$\alpha(v,u) = \frac{\left(g_{\mathrm{Na}}m_\infty(v)^3h'_\infty(u)(v-V_{\mathrm{Na}})\right)^2}{\left(g_{\mathrm{Na}}m_\infty(v)^3h'_\infty(u)(v-V_{\mathrm{Na}})\right)^2 + (4g_{\mathrm{K}}n_\infty(u)^3n'_\infty(u)(v-V_{\mathrm{K}}))^2}. \tag{18}$$

The rest is specified in section A.1.

Received xxxx 20xx; revised xxxx 20xx.

E-mail address: frm@math.ku.dk

IV
Degrees of Freedom for Piecewise Lipschitz Estimators

Frederik Vissing Mikkelsen
Department of Mathematical Sciences
University of Copenhagen

Niels Richard Hansen
Department of Mathematical Sciences
University of Copenhagen

Publication details

Accepted in *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques* (2017).

DEGREES OF FREEDOM FOR PIECEWISE LIPSCHITZ ESTIMATORS

FREDERIK RIIS MIKKELSEN AND NIELS RICHARD HANSEN

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.

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.

2010 *Mathematics Subject Classification.* 62J05, 62J07.

Key words and phrases. best subset selection, lasso-OLS, degrees of freedom, Stein's Lemma.

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 μ the unknown parameter, and we let $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}^n$ denote an estimator of μ. 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 μ without reference to the regression setup.

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

$$\mathrm{Risk}(\hat{\mu}) \coloneqq E\|\mu - \hat{\mu}(Y)\|_2^2,$$

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 ∂_i denoting the partial derivative w.r.t. the ith 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*

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

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

$$\operatorname{df}_{\mathrm{S}}(\hat{\mu}) := E(\operatorname{div}(\hat{\mu})(Y)). \tag{2}$$

A simple expansion of the risk yields

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

Hence $\|Y - \hat{\mu}(Y)\|_2^2 - n\sigma^2 + 2\sigma^2 \widehat{\operatorname{df}}$ is an unbiased risk estimate if $\widehat{\operatorname{df}}$ is an unbiased estimate of $\operatorname{df}(\hat{\mu})$. In practice, σ^2 must be estimated as well and a bias of $\widehat{\operatorname{df}}$ 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 $\operatorname{df}(\hat{\mu}) = \operatorname{df}_{\mathrm{S}}(\hat{\mu})$ due to Stein's lemma (Lemma 2 in [21]), in which case $\operatorname{div}(\hat{\mu})(Y)$ is an unbiased estimate of $\operatorname{df}(\hat{\mu})$. However, most estimators with discontinuities are not almost differentiable, and for such estimators it is not clear if $\operatorname{div}(\hat{\mu})(Y)$ is a useful estimate of the degrees of freedom. Indeed, our main result, Theorem 2.4, provides a representation of $\operatorname{df}(\hat{\mu}) - \operatorname{df}_{\mathrm{S}}(\hat{\mu})$, 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 $\operatorname{df}(\hat{\mu})$ and $\operatorname{df}_{\mathrm{S}}(\hat{\mu})$ for the lasso-OLS estimator, which can be used to derive an estimate of $\operatorname{df}(\hat{\mu})$. 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 $i = 1, ..., N$:*

(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 ∂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 μ_i and μ_j on the boundary.
b) **Degrees of freedom.** Assumption 2.2(a) implies by Rademacher's theorem (Theorem 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 $\mathrm{df}(\hat{\mu})$ and $\mathrm{df}_{\mathrm{S}}(\hat{\mu})$ 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 η_i is defined on a subset of ∂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 ∂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 minimisation 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. Appendix 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*

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

where η_i denotes the measure theoretic outer unit normal to ∂U_i.

The proof is in Appendix 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 $\mathrm{df}(\hat{\mu}) = \mathrm{df}_{\mathrm{S}}(\hat{\mu})$, 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, 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 $A \subseteq \{1,...,p\}$, 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

$$\mathcal{S} := \{S = \mathrm{col}(X_A) \mid A \subseteq \{1,...,p\}\}$$

denote the set of subspaces spanned by columns of X. The orthogonal projection onto a subspace $S \in \mathcal{S}$ is denoted by Π_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, $\text{supp}(\hat{\beta}^{\lambda}) \subseteq \{1, ..., p\}$, is called an *active set*. The lasso estimator $\hat{\mu}^{\lambda}_{\text{lasso}}(y) = X\hat{\beta}^{\lambda}$ belongs to the space $\text{col}(X_A)$ for $A = \text{supp}(\hat{\beta}^{\lambda})$, and it follows by Lemma 7 in [25] that there exists a Lebesgue null set N, such that $\text{col}(X_A)$ 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 $\text{col}(X_A)$ when there is a solution $\hat{\beta}^{\lambda}$ with active set $A = \text{supp}(\beta^{\lambda})$ is therefore well defined. The *lasso-OLS* estimator $\hat{\mu}^{\lambda}_{\text{l-OLS}} \coloneqq \Pi_{\widehat{S}^{\lambda}}$ 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

$$U^{\lambda}_S \coloneqq (\widehat{S}^{\lambda} = S)$$

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^{λ}_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^{λ}_S, its divergence equals $\dim(S)$, hence Stein's degrees of freedom is

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

From Lemma 3 in [23] it follows that $\dim(\widehat{S}^{\lambda}) = |\text{supp}(\hat{\beta}^{\lambda})|$ 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 $\text{df}(\hat{\mu})$ can be used for computing or estimating the degrees of freedom. The first term of (4), $\text{df}_S(\hat{\mu})$, may be estimated by $\text{div}(\hat{\mu})(Y)$, but the second term is more difficult. In this section we show how this second term can be related to the derivative of $\lambda \mapsto \text{df}_S(\hat{\mu}^{\lambda}_{\text{l-OLS}})$ 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.

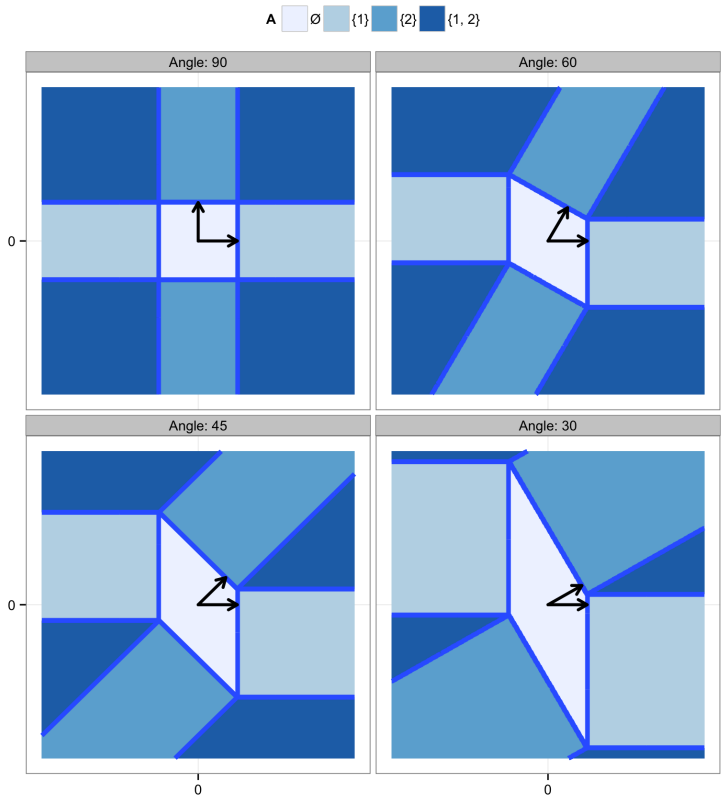


FIGURE 1. Illustrations of the decomposition of $\mathbb{R}^2$ into the four sets $U^1_\emptyset$, $U^1_{\{1\}}$, $U^1_{\{2\}}$ and $U^1_{\{1,2\}}$ according to the lasso estimator with $\lambda = 1$. The set $U^1_\emptyset$ consists of the points shrunk to zero, the sets $U^1_{\{1\}}$ and $U^1_{\{2\}}$ 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.

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,

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

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

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

and as in [24]

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

Letting ∂_λ denote the differential operator with respect to λ we observe that

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

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 μ and is not readily estimable. But we have the divergence estimator, $\sum_i 1_{(|y_i|>\lambda)}$, of $\mathrm{df}_{\mathrm{S}}(\hat{\mu}^\lambda_{\text{l-OLS}})$, 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*

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

where ∂_λ denotes differentiation w.r.t. λ.

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

$$h(\delta) \coloneqq \mathrm{df}_{\mathrm{S}}(\hat{\mu}^{\exp(\delta)}_{\text{l-OLS}}),$$

then

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

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, λ_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 Δ_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 Δ_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 $-\lambda\partial_\lambda \mathrm{df}_{\mathrm{S}}(\hat{\mu}^{\lambda}_{\text{l-OLS}})$ 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

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

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 parametrised 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*

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

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 $i = 1, ..., N$. If $\hat{\mu}^1$ satisfies Assumption 2.2 then

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

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 λ. Next, the change of variable formula yields

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

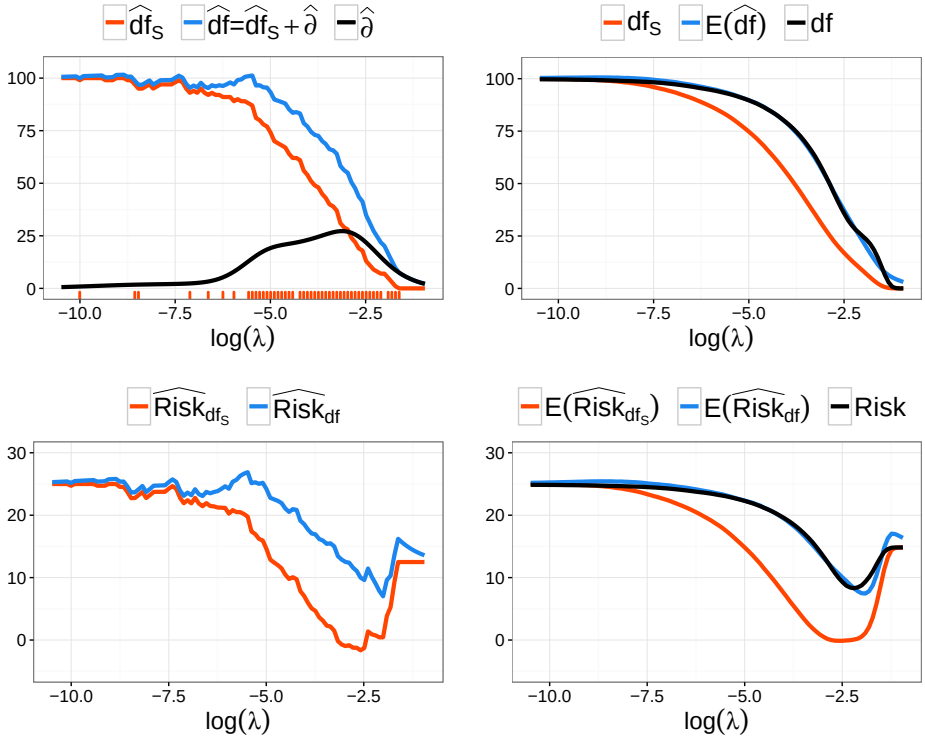


FIGURE 2. Left: Realization of the estimates of degrees of freedom $\hat{\mathrm{df}}_S = \dim(\hat{S}^\lambda)$ and $\hat{\mathrm{df}} = \dim(\hat{S}^\lambda) + \widehat{\partial}$ as well as the correction term $\widehat{\partial}$ as a function of $\log(\lambda)$ (top) and corresponding estimates 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 $\mathrm{df} = \mathrm{df}(\hat{\mu}^\lambda_{\text{l-OLS}})$ 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).

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

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

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

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

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

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

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

for each $A \subseteq \{1, ..., p\}$. 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

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

Hence $U_A^\lambda = \lambda^{\frac{1}{2-k}} U_A^1$ holds for all $A \subseteq \{1, ..., p\}$ 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^λ 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

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

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^λ_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 $S_1, S_2 \in \mathcal{S}$

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

For $A \subseteq \{1, ..., p\}$ 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

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

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

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

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

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

for $S_1, S_2 \in \mathcal{S}$. Consequently, if $S_1 \neq S_2$ then $\mathcal{H}^{n-1}\left(\overline{U}^\lambda_{S_1} \cap \overline{U}^\lambda_{S_2}\right) = 0$, 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

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

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 $\widehat{\text{Risk}}_{\text{df}}$ given by (7) performs as an estimate of the risk and in terms of selecting the penalty parameter λ. 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

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

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

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

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

The lasso estimator was tuned by minimising the risk estimate

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

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

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

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

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

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

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

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

All simulations were carried out assuming either that σ^2 was known or using the following estimator of σ^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

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

and σ^2 was finally estimated as

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

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 γ and α.

Two simulation designs were implemented with parameters as follows:

Parameter	Values for simulation study I					Values for simulation study II				
σ	0.5					0.1	0.2	0.5	1	2
α	0.1					0	0.05	0.1	0.3	0.5
n	50	100	200	400	800	100	200			
p	200	2000	20000			n				
γ	1					1	0.9			
X	S					O	S	E		
ρ	0.1					0	0.1	0.4	0.7	

The parameter ρ 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 240×377 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 σ^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 λ 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 λ-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.

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 λ (the λ 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 $p = 2000, 20000$. 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 σ does not affect the MSE of the risk estimates much.

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.

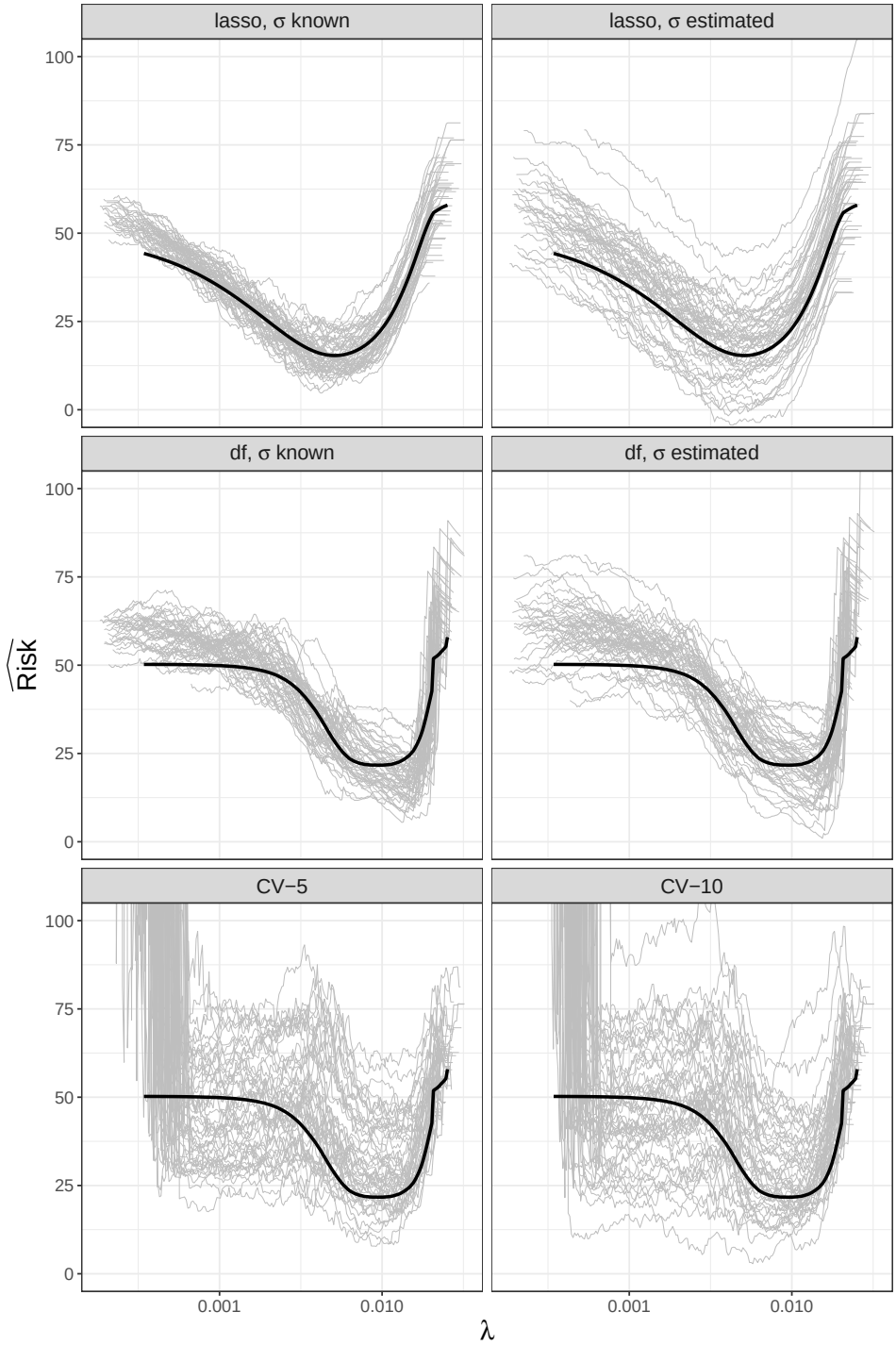


FIGURE 3. Risk estimates $\widehat{\text{Risk}}_{\text{df}}$, $\widehat{\text{Risk}}_{\text{CV-5}}$, $\widehat{\text{Risk}}_{\text{CV-10}}$ and $\widehat{\text{Risk}}_{\text{lasso}}$ (gray lines) for 50 samples as a function of λ. The black lines are Monte Carlo estimates of the true risks. The design 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).

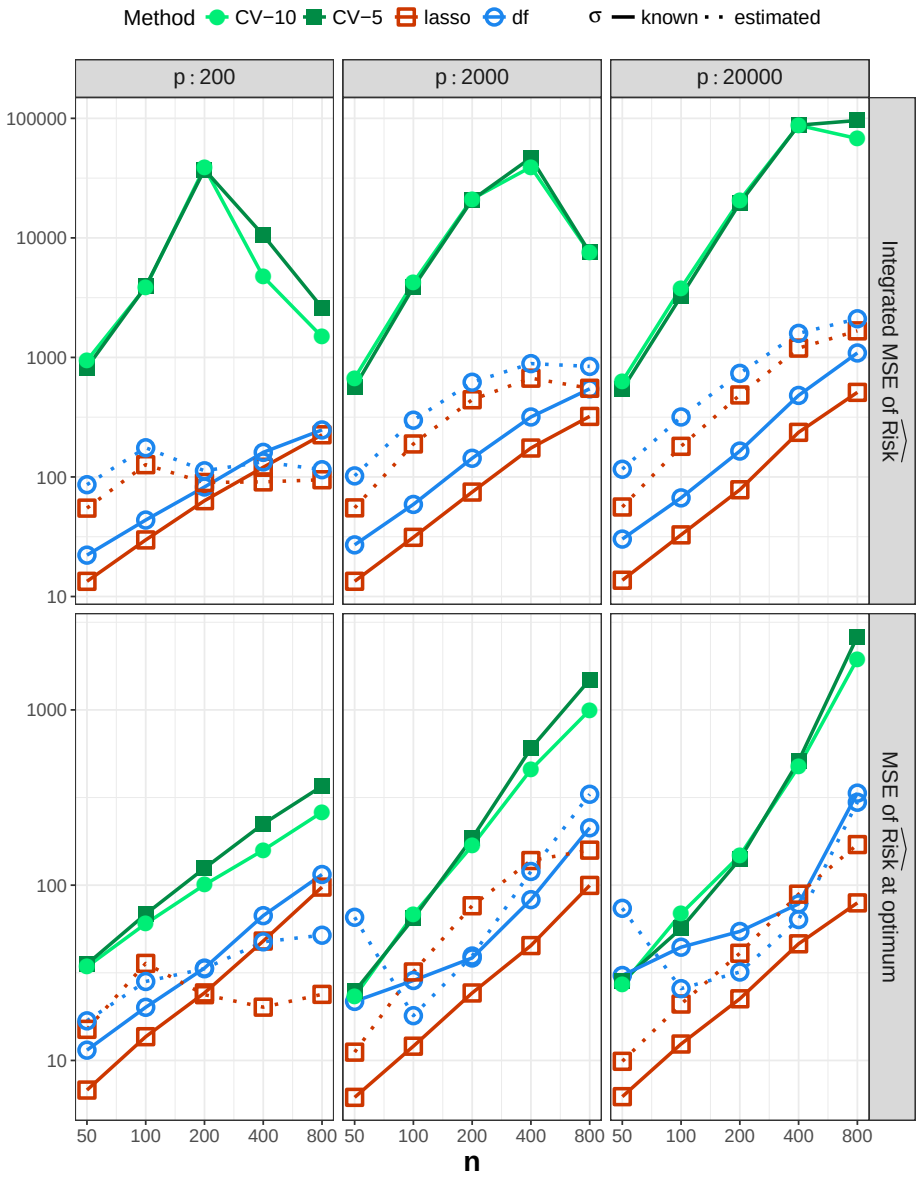


FIGURE 4. Integrated mean squared error (top) and mean squared error at the optimal value of λ, $\hat{\lambda}$ (bottom) of the risk estimates $\widehat{\text{Risk}}_{\text{df}}$, $\widehat{\text{Risk}}_{\text{CV-5}}$, $\widehat{\text{Risk}}_{\text{CV-10}}$ and $\widehat{\text{Risk}}_{\text{lasso}}$. 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)

4.4. **Results from study II.** Firstly, we discuss the comparison of the two tuning methods df and $\mathrm{df_S}$ 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 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 λ 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, $\gamma = 0.9$, α large and σ large), $\hat{\mu}^{\hat{\lambda}_{\mathrm{df}}}_{\text{l-OLS}}$ clearly outperforms the oracle-OLS estimator, while $\hat{\mu}^{\hat{\lambda}_{\mathrm{df_S}}}_{\text{l-OLS}}$ 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 $\widehat{\mathrm{Risk}}_{\text{CV-K}}$ is constant in λ. 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 σ 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 γ. 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 (α large, σ small and $\gamma = 1$). Cross-validation appears to perform best for highly correlated designs (ρ 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.

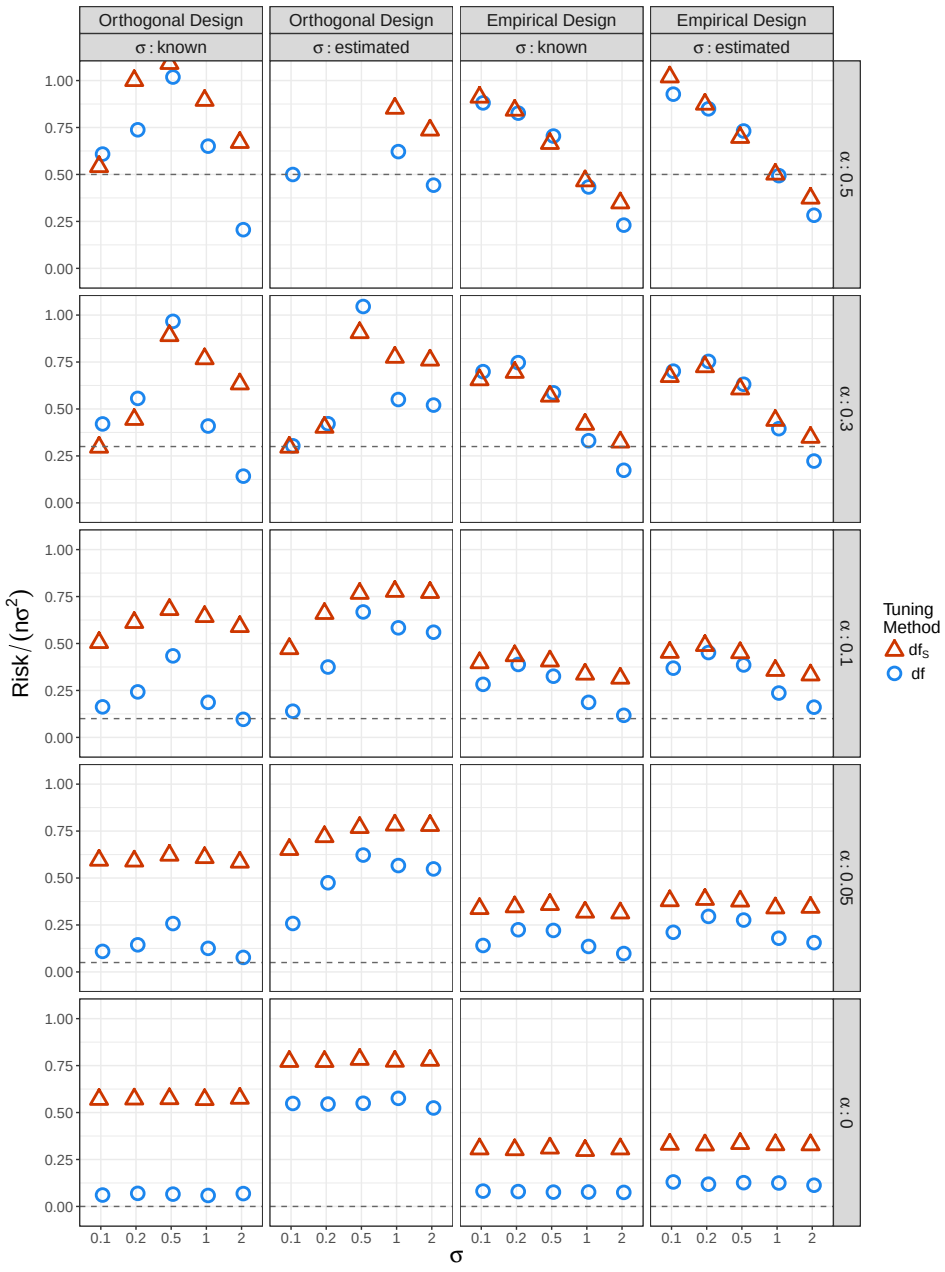


FIGURE 5. Risk relative to $\sigma^2 n$ for the estimators $\hat{\mu}^{\hat{\lambda}_{\mathrm{df_S}}}_{\text{l-OLS}}$ and $\hat{\mu}^{\hat{\lambda}_{\mathrm{df}}}_{\text{l-OLS}}$ for orthogonal and empirical designs with $n = 100$ and $\gamma = 1$. The dashed line is $\lceil n\alpha \rceil / n \simeq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

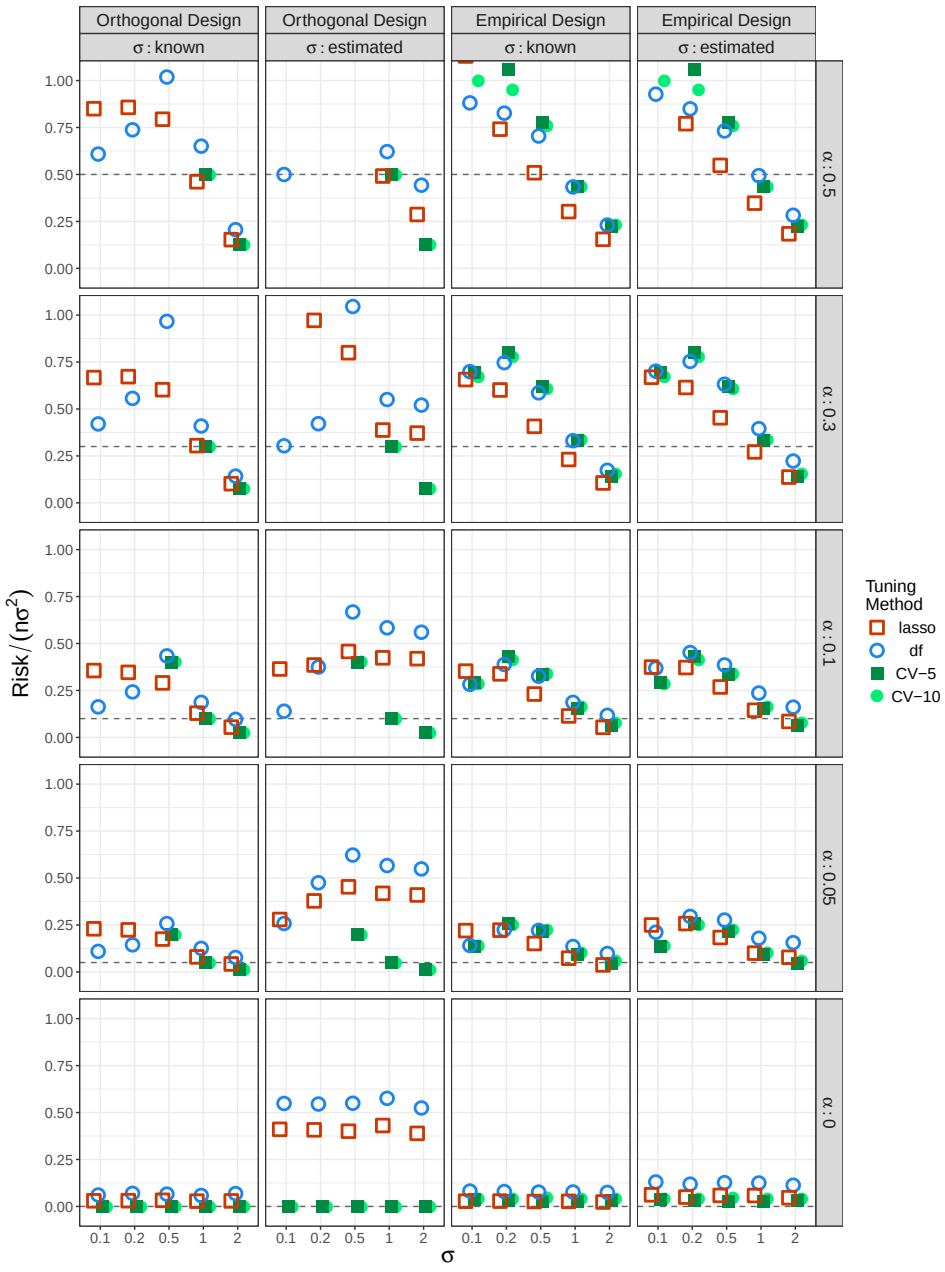


FIGURE 6. Risk relative to $\sigma^2 n$ for the estimators $\hat{\mu}^{\hat{\lambda}_{\mathrm{df}}}_{\text{l-OLS}}$, $\hat{\mu}^{\hat{\lambda}_{\text{CV-5}}}_{\text{l-OLS}}$, $\hat{\mu}^{\hat{\lambda}_{\text{CV-10}}}_{\text{l-OLS}}$ and $\hat{\mu}^{\hat{\lambda}_{\text{lasso}}}_{\text{lasso}}$ for orthogonal 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.

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 $\mathrm{Pen}(\cdot) = \|\cdot\|_0$ 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 μ with tuning parameter $\lambda > 0$, denoted by $\hat{\mu}^\lambda_{\mathrm{bs}}$, is

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

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

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

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

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

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

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

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

$$\mathcal{A}_1 := \left\{ A_1, A_2 \subseteq \{1,\dots,p\} \;\middle|\; \begin{array}{c} \mathrm{col}(X_{A_1}) \subseteq \mathrm{col}(X_{A_2}) \text{ and} \\ \mathrm{rank}(X_{A_2}) = \mathrm{rank}(X_{A_1}) + 1 \end{array} \right\} \quad \text{and}$$

$$\mathcal{A}_2 := \left\{ A_1, A_2 \subseteq \{1,\dots,p\} \;\middle|\; \mathrm{col}(X_{A_1}) \neq \mathrm{col}(X_{A_2}) \text{ and } \begin{array}{c} (A_1, A_2) \notin \mathcal{A}_1 \\ (A_2, A_1) \notin \mathcal{A}_1 \end{array} \right\},$$

then

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

where

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

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

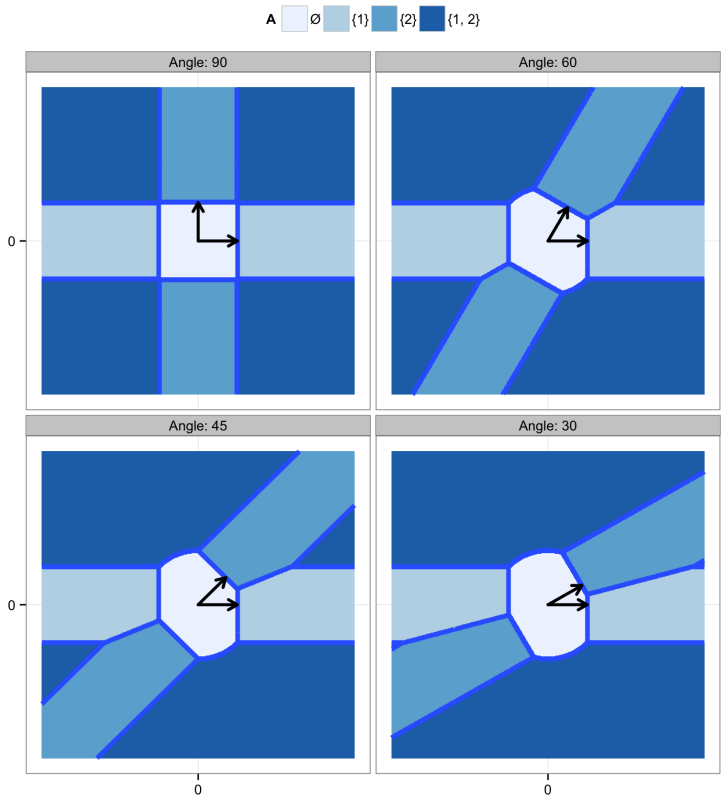


FIGURE 7. Illustrations of the decomposition of $\mathbb{R}^2$ into the four sets $U^1_\emptyset$, $U^1_{\{1\}}$, $U^1_{\{2\}}$ and $U^1_{\{1,2\}}$ according to the best subset selection estimator in the Lagrangian formulation with $\lambda = 1$. The set $U^1_\emptyset$ consists of the points projected 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.

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

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

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

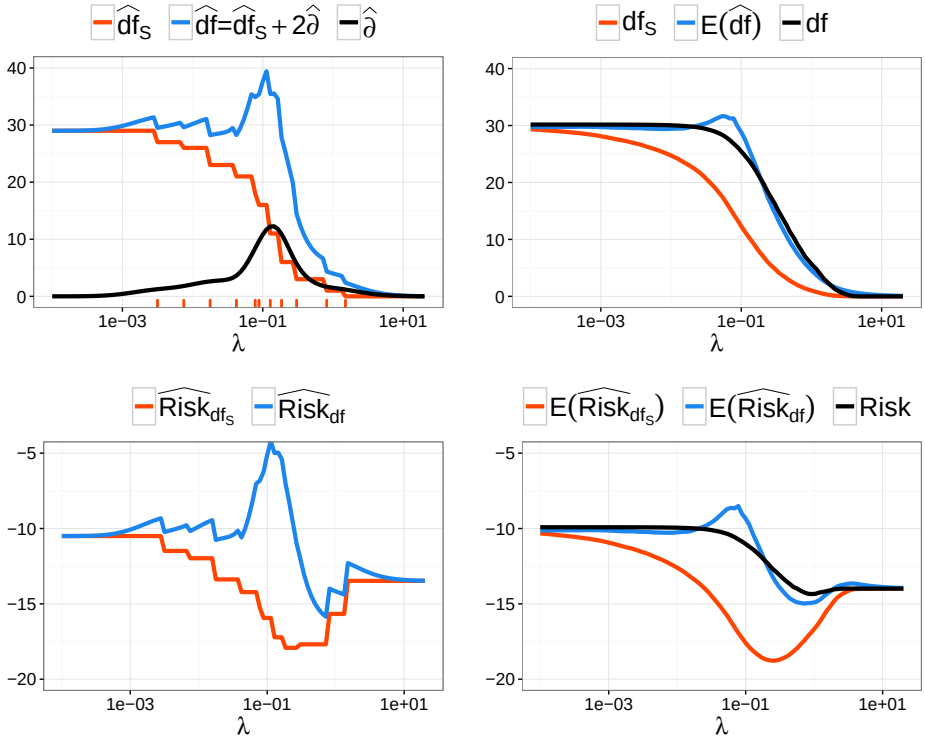


FIGURE 8. Left: Realization of the estimates of degrees of freedom $\hat{\mathrm{df}}_S = \dim(\hat{S}^\lambda)$ and $\hat{\mathrm{df}} = \dim(\hat{S}^\lambda) + 2\widehat{\partial}$ as well as the correction term $\widehat{\partial}$ as a function of $\log(\lambda)$ for best subset selection (top) and 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).

(20) was actually a good approximation of $\mathrm{df}(\hat{\mu}^\lambda_{\mathrm{bs}})$. Figure 8 shows the 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 $\lambda \mapsto \mathrm{df}_{\mathrm{S}}(\hat{\mu}^{\lambda})$ to the discontinuities of the estimator $\hat{\mu}^{\lambda}$ as expressed via the integral representation of $\mathrm{df}(\hat{\mu}^{\lambda}) - \mathrm{df}_{\mathrm{S}}(\hat{\mu}^{\lambda})$. This does, indeed, make some intuitive sense as the first expresses the mean jump of degrees of freedom per unit change of λ and the other (in some sense) the mean discontinuity of degrees of freedom per unit change of y. Changing λ for fixed y or changing y for fixed λ 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 $\mathrm{df}_{\mathrm{S}}(\hat{\mu}^{\lambda})$ that corresponds to jumps that can be related to $\mathrm{df}(\hat{\mu}^{\lambda}) - \mathrm{df}_{\mathrm{S}}(\hat{\mu}^{\lambda})$. 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 $\mathrm{df}(\hat{\mu}^{\lambda}) - \mathrm{df}_{\mathrm{S}}(\hat{\mu}^{\lambda})$. 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 $\mathrm{df}(\hat{\mu}^{\lambda}) - \mathrm{df}_{\mathrm{S}}(\hat{\mu}^{\lambda})$. 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 A. Additional results and Proofs

A.1. **Semialgebraic sets.** Observe that for A and B subsets of $\mathbb{R}^n$ it holds that

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

Especially, the family of sets

$$\left\{ E \in \mathcal{B}(\mathbb{R}^n) \;\middle|\; \begin{array}{l} r \mapsto \mathcal{H}^{n-1}\left(\partial E \cap B(0,r)\right) \\ \text{is polynomially bounded} \end{array} \right\} \tag{22}$$

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 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 $r \mapsto \mathcal{H}^{n-1}\left(\partial E \cap B(0,r)\right)$ is polynomially bounded.*

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

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

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 η_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* $i = 1, ..., N$.
(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 η_i on ∂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:

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

holds for all i, j, k distinct.

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

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

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

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

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

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

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 $i = 1, ..., N$ 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 η_i denotes the outer unit normal of ∂U_i, which is well defined and nonzero on a subset of ∂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

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

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

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

Since η_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. *Journal of the American Statistical Association*, 87(419):738–754, 1992.

[2] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[3] A. Dalalyan and A. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

[4] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[5] B. Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

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

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

[8] J. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.*, 42(3):819–849, 06 2014.

[9] H. Federer. *Geometric measure theory*. Grundlehren der mathematischen Wissenschaften. Springer, 1969.

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

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

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

[13] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1990.

[14] K. Kato. On the degrees of freedom in shrinkage estimation. *Journal of Multivariate Analysis*, 100(7):1338 – 1352, 2009.

[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):907–927, 06 2016.

[16] T. Loi and P. Phien. Bounds of Hausdorff measures of tame sets. *Acta Mathematica Vietnamica*, 39(4):637–647, 2014.
[17] N. Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374 – 393, 2007.
[18] M. Meyer and M. Woodroofe. On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, 28(4):1083–1104, 2000.
[19] F. Mikkelsen and N. Hansen. Supplementary material for "Degrees of freedom for piecewise Lipschitz estimators". http://doi.org/10.5281/zenodo.321847, 2017.
[20] S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, 26(1):35–67, 2016.
[21] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 11 1981.
[22] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.
[23] R. J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Statist.*, 7:1456–1490, 2013.
[24] R. J. Tibshirani. Degrees of freedom and model search. *Statistica Sinica*, 25(3):1265–1296, 2015.
[25] R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *Ann. Statist.*, 40(2):1198–1232, 04 2012.
[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):1417–1423, 2014.
[27] J. Ye. On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.*, 93(441):120–131, 1998.
[28] H. Zou, T. Hastie, and R. Tibshirani. On the degrees of freedom of the lasso. *Ann. Statist.*, 35(5):2173–2192, 10 2007.

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark
E-mail address, Corresponding author: frm@math.ku.dk

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

V

Extending SURE to Estimators with Data Adaptive Model Selection via Flows

Frederik Vissing Mikkelsen
Department of Mathematical Sciences
University of Copenhagen

Niels Richard Hansen
Department of Mathematical Sciences
University of Copenhagen

Publication details

Draft

EXTENDING 'SURE' TO ESTIMATORS WITH DATA ADAPTIVE MODEL SELECTION VIA FLOWS

FREDERIK VISSING MIKKELSEN AND NIELS RICHARD HANSEN

Abstract. For a class of estimators of mean value parameters in $\mathbb{R}^n$, which involve data adaptive model selection, we present a representation of the degrees of freedom. This representation readily yields an estimator of the degrees of freedom, which subsequently provides a natural extension of Stein's unbiased risk estimate (SURE) to a class of estimators with data adaptive model selection. Four examples of estimators, for which the classic SURE does not apply, are considered in detail in this paper: marginal screening, relaxed lasso, best subset selection and singular value decomposition with a hard threshold on the singular values. The representation of the degrees of freedom relies on linking the data to a tuning parameter via a flow. Using such flows, the dependence between the data and the estimates for fixed tuning parameter can be understood by the dual operation, i.e., having the data fixed and varying the tuning parameter.

1. Introduction

Risk estimation is a key concept in statistical modelling. It provides means for assessing the error of a given estimator and in statistical settings with multiple competing estimators it provides model selection criteria. The empirical risk does not account for the flexibility of the model and is often too optimistic. For classic model selection criteria, such as AIC and Mallows's C_p, the dimension of the parameter space is used to adjust for the optimism of the empirical risk and provide a fair model score across different dimensions. These selection criteria are suitable for linear models, in which the dimension of the parameter space is directly linked to the flexibility of the linear predictor. Extensions to models or methods outside of this setting exist. An example of such is the use of the divergence of sufficiently differentiable estimators based on Stein's lemma as described by Efron (2004). Stein's lemma was also used by Zou et al. (2007) and Tibshirani & Taylor (2012) 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 similar information criteria correctly adjust for the optimism resulting from selecting one model from a collection 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 (1992), who proposed a bootstrap based method for risk estimation as an alternative. Ye (1998) defined the notion of generalised 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 involving 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 by Stein (1981)), then applying the divergence operator to the estimator yields an unbiased estimate of the generalised degrees of freedom. Correcting the empirical risk by this estimate yields Stein's unbiased risk estimate (SURE), as used by Donoho & Johnstone (1995) and Xie et al. (2012). Furthermore, Meyer & Woodroofe (2000), Zou et al. (2007), Kato (2009), Tibshirani & Taylor (2012) and Candès et al. (2013) among others used the divergence operator 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 no longer be an unbiased estimate of the degrees of freedom. Discontinuities of the estimator in particular appear 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 in general lead to non-ignorable discontinuities. A variable selection procedure effectively divides the sample space into a number of disjoint regions, with the estimator being, say, a projection on each region. The resulting estimator will generally be discontinuous on the boundaries between regions.

Recent developments in computations of degrees of freedom for discontinuous estimators include that of Tibshirani (2015). Here the author considered linear regression models with orthogonal design and showed how to compute the degrees of freedom for the hard threshold operator. This operation is equivalent to the Lagrangian formulation of best subset selection for orthogonal designs. Additionally, an extension of Stein's lemma to some discontinuous estimators was presented, though it was not shown if this extension applies to subset selection estimators. Mikkelsen & Hansen (2017) recently derived a general representation of degrees of freedom for piecewise Lipschitz estimators and used it to compute the degrees of freedom for the lasso-OLS estimator: an estimator which applies the OLS estimator restricted to the predictors selected by the lasso estimator by Tibshirani (1996).

The main contribution of this paper is Theorem 4.6 and its associated Corollary 4.7. These results provide a formula for the contribution to the degrees of freedom that are due to discontinuities of the selection procedure. They rely on perturbing the observation space by a flow. For a broad class of selection procedures we define a function H, depending on the observation y and a tuning parameter t. Through studying the behaviour of H in the tuning parameter direction, one is able to recover the associated behaviour in the y direction. By the specific construction of H the recovered behaviour in the y direction equals the part of the degrees of freedom that are due to the discontinuities arising from the selection procedure.

Before defining the function H in Section 3 we motivate the use of degrees of freedom in risk estimation in Section 2. The assumptions required for recovering of the degrees of freedom via H are presented in Section 4, followed by four examples in Section 5: marginal screening, relaxed lasso, best subset selection and singular value decomposition with a hard threshold on the singular values. All examples, except for best subset selection, satisfy all the conditions for Corollary 4.7. For best subset selection, only a partial recovery of the degrees of freedom is possible. Proofs and some auxiliary results are in the appendix.

2. Model Selection and Stein's Unbiased Risk Estimate

We consider a multivariate normally distributed random variable $Y \sim \mathcal{N}(\mu, \sigma^2 I)$ and let $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}^n$ denote an estimator of the unknown mean vector μ. The risk of the estimator is defined as

$$\text{Risk}(\hat{\mu}) \coloneqq E\|\mu - \hat{\mu}(Y)\|_2^2,$$

provided that $\hat{\mu}(Y)$ has finite second moment. The risk measures the average squared error on the estimator and is therefore desirable to estimate. Risk estimators with good statistical properties can for instance be used to select tuning parameters.

Our main interest is to estimate the risk under the Gaussian model, which we obtain through two types of *degrees of freedom*. When defining these, $\psi(y; \mu, \sigma^2)$ denotes the density of Y and $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^n$. Furthermore, the divergence operator is defined as

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

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

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

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

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

$$\text{df}_{\text{S}}(\hat{\mu}) \coloneqq E(\text{div}(\hat{\mu})(Y)). \tag{2}$$

The degrees of freedom naturally arise through a simple expansion of the risk

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

Hence $\|Y - \hat{\mu}(Y)\|_2^2 - n\sigma^2 + 2\sigma^2 \widehat{\text{df}}$ is an unbiased risk estimate provided that $\widehat{\text{df}}$ is an unbiased estimate of $\text{df}(\hat{\mu})$. Having a bias in $\widehat{\text{df}}$ may also be preferable if it sufficiently reduces the variance.

If $\hat{\mu}$ is *almost differentiable* then $\text{df}(\hat{\mu}) = \text{df}_{\text{S}}(\hat{\mu})$ due to Stein's lemma (Lemma 2 in Stein (1981)), in which case $\text{div}(\hat{\mu})(Y)$ becomes an unbiased estimate of $\text{df}(\hat{\mu})$ and yields SURE (Stein's Unbiased Risk Estimator):

$$\|Y - \hat{\mu}(Y)\|_2^2 - n\sigma^2 + 2\sigma^2 \text{div}(\hat{\mu})(Y) \tag{4}$$

The problem, however, is that most estimators involving data adaptive model selection are discontinuous. In particular they are not almost differentiable, and for such estimators it is not clear if $\text{div}(\hat{\mu})(Y)$ is a useful estimate of the degrees of freedom. In this paper, we will study the bias

$$\text{df}(\hat{\mu}) - \text{df}_{\text{S}}(\hat{\mu}) \tag{5}$$

for a range of estimators involving model selection. Recently, Mikkelsen & Hansen (2017) provided a representation of (5), which involves a sum of boundary integrals. This representation on its own yields no immediate estimator of (5), but the

authors were able to do so for the specific *lasso-OLS estimator*, $\hat{\mu}^{\lambda}_{\text{l-OLS}}$. For this particular estimator, which applies the OLS estimator restricted to the set of predictors selected by the lasso estimator (Tibshirani (1996)) with tuning parameter $\lambda > 0$, we have the striking identity:

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

where ∂_{λ} denotes the differential with respect to λ. This equation is quite interesting, because $\mathrm{df}_{\mathrm{S}}(\hat{\mu}_{\text{l-OLS}})$ is the expected dimension of the model selected by the lasso (Lemma 3 in Tibshirani (2013)). These dimensions are readily available, when evaluating the lasso estimator. The identity (6) was used by Mikkelsen & Hansen (2017) to derive a correction for SURE. Furthermore, through an extensive simulation study the authors measured the effect of tuning lasso-OLS via the corrected SURE, the uncorrected SURE and 5- and 10-fold cross validation. The results showed that tuning via the corrected SURE led to estimates that were generally closer to the true mean than estimates tuned via uncorrected SURE. Moreover, the corrected SURE and both types of cross validation led to estimates that were comparably close to the true mean. However, the SURE based methods were faster to compute than the cross validated methods.

In this paper we seek to establish identities similar to (6) for other estimators involving model selection. We show that for some model selection estimators $\hat{\mu}^t$, in which the selection events are determined by a tuning parameter $t \in \mathbb{R}$ in a certain way, one has the identity

$$\mathrm{df}(\hat{\mu}^t) - \mathrm{df}_{\mathrm{S}}(\hat{\mu}^t) = \partial_t E(H(t, Y)), \tag{7}$$

where H is a real-valued function defined on $\mathbb{R} \times \mathbb{R}^n$. We provide an explicit formula for H in the following section. In the lasso-OLS case, $-H(t, y)$ reduces to the dimension of the model selected by the lasso estimator for tuning parameter $\lambda = e^t$ (up to an additive constant, see Example 5.2 for details). This agrees with (6). A visualisation of (7) applied to marginal screening in Example 5.1 is presented in Figure 1.

3. Constructing and applying H

Consider a finite set of models $\mathcal{M} = \{M_1, ..., M_N\}$. A *selection procedure* with tuning parameter $t \in \mathbb{R}$ is a map $\widehat{M}_t : \mathbb{R}^n \to \mathcal{M}$; given an observation y model $\widehat{M}_t(y)$ is chosen. The sets $(\widehat{M}_t = M)_{M \in \mathcal{M}}$ are called the *selection events*.

We assume that each model $M \in \mathcal{M}$ has an associated *post-selection estimator* $\hat{\mu}_M$, which is applied to the observation y given that $\widehat{M}_t(y) = M$, and we assume that each of these are locally Lipschitz. In a linear regression setting, a typical example of a post-selection estimator is an OLS estimator on some subset of the predictors provided by the selection procedure $\widehat{M}_t$. As we will see, the specific choice of the post-selection estimators are not that important. They do matter in terms of how H is defined, but the validity of (7) is mostly a question of the selection procedure.

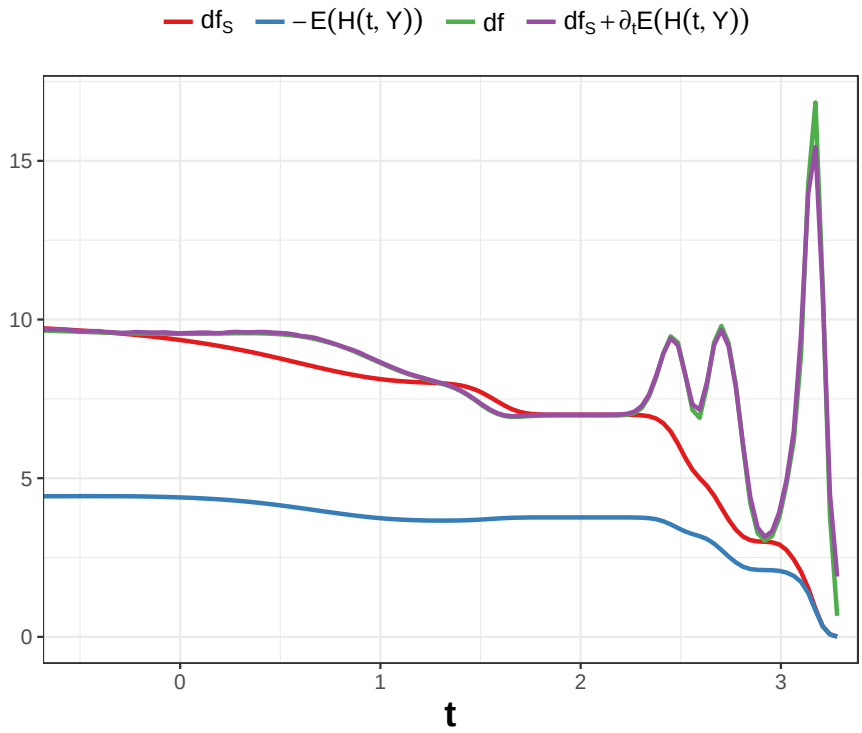


FIGURE 1. Monte Carlo estimates of df, df_S and $-E(H(t,Y))$ for marginal screening in Example 5.1 using 10.000 replications. H was shifted vertically to be in the frame. The design matrix, $X \in \mathbb{R}^{12\times 10}$, had i.i.d. standard Gaussian entries, $\sigma = 0.25$ and $\mu = X\beta$, where the first four coordinates of β were 1, the remaining 0. The derivative of $E(H(t,Y))$ was approximated by finite differencing.

The most important property we require for the selection procedure $\widehat{M}_t$ in order for (7) to hold is the following: there exists a C^2 *flow*, i.e., a C^2-function $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ with the properties

$$\begin{aligned} F(0, y) &= y, \quad \text{for all } y \in \mathbb{R}^n, \\ F(t + s, y) &= F(s, F(t, y)), \quad \text{for all } s, t \in \mathbb{R} \text{ and } y \in \mathbb{R}^n, \end{aligned} \tag{8}$$

connecting the tuning parameter and the selection procedure:

$$\widehat{M}_t(y) = \widehat{M}_0(F(-t, y)), \quad \text{for all } t \in \mathbb{R} \text{ and almost all } y \in \mathbb{R}^n. \tag{9}$$

The above connection is visualised in Figure 2. Note that all flows are invertible in the second coordinate: $F(-t, F(t, y)) = y$ for all $t \in \mathbb{R}$ and $y \in \mathbb{R}^n$, i.e., $F(-t, \cdot)$ is the inverse of $F(t, \cdot)$. The function $f : \mathbb{R}^n \to \mathbb{R}^n$ given by $f(y) = \partial_t F(0, y)$ is called the *field* of the flow.

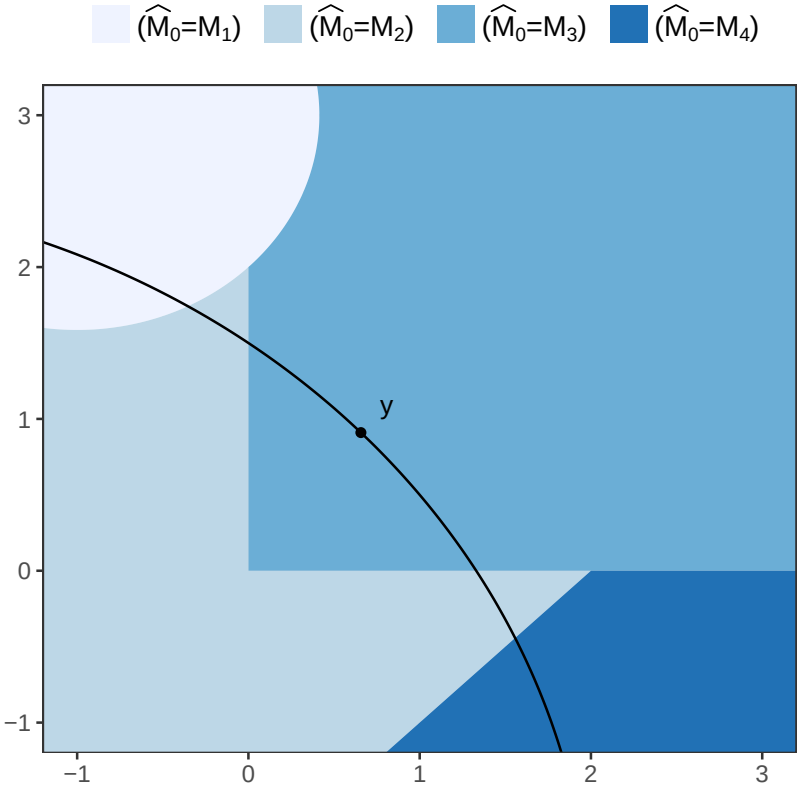


FIGURE 2. Partition of $\mathbb{R}^2$ by some generic selection procedure. An observation y lies on its trajectory $t \mapsto F(-t, y)$ (black line). Under (9), the selection events for $t = t_0$ are transformations of the selection events for $t = 0$. However, it is often more useful to think of y being transformed instead.

Example 3.1. Condition (9) does not hold for arbitrary selection procedures, but Example 3.4 in Mikkelsen & Hansen (2017) showcases some examples. More precisely, they consider a regression with X an $n \times p$ design matrix and the penalised loss function

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

where $\mathrm{Pen} : \mathbb{R}^p \to \mathbb{R}$ is *positive homogeneous* of degree $k \in [0, 2)$ (this includes the quasi-norms $\mathrm{Pen}(\beta) = \|\beta\|_p^k$ with $0 < k < 2$, $p > 0$ and $\mathrm{Pen}(\beta) = \|\beta\|_0 = |\mathrm{supp}(\beta)|$). If the selection procedure selects the model based on the support of β at optimum, i.e.,

(10)
$$(\widehat{M}^t = M) = \left\{ y \in \mathbb{R}^n \;\middle|\; \inf_{\beta : \mathrm{supp}(\beta) = M} \ell(y, \beta, \lambda) = \inf_{\beta} \ell(y, \beta, \lambda) \right\}, \quad M \subseteq \{1, ..., p\},$$

then (9) holds with the flow $F(t, y) = e^{\frac{t}{2-k}} y$.

For $M \in \mathcal{M}$ let $U_M = \mathrm{int}(\widehat{M}_0 = M)$ denote the interior of the selection event for $t = 0$. If the selection event $(\widehat{M}_0 = M)$ has locally finite perimeter (which will be made precise in Section 4) we can assume that the boundary $\partial(\widehat{M}_0 = M)$ is Lebesgue null. Most meaningful selection events have locally finite perimeter; one would typically need either a fractal-like structure or boundaries that oscillate with increasing frequency in order for this assumption to break. So for any meaningful selection procedure, U_M represents the selection event $U_M = \mathrm{int}(\widehat{M}_0 = M)$ almost surely. If (9) holds then $\mathrm{int}(\widehat{M}_t = M) = F(t, U_M)$ for all $t \in \mathbb{R}$ and $M \in \mathcal{M}$. The combined estimator $\hat{\mu}^t$, which applies the selection procedure and the subsequent post-selection estimators is on the form

$$\hat{\mu}^t(y) = \sum_{M \in \mathcal{M}} 1_{F(t, U_M)}(y) \hat{\mu}_M(t, y), \tag{11}$$

almost surely. Note that the post-selection estimators are allowed to depend on the tuning parameter $t \in \mathbb{R}$.

3.1. **Definition of H.** In order to define H, we assume that the selection procedure satisfies (9) and that for almost all y the left and right limits of $t \mapsto \widehat{M}_t(y)$ exist in all $t \in \mathbb{R}$. Let $M_t^-(y) \coloneqq \lim_{s \nearrow t} \widehat{M}_s(y)$ and $M_t^+(y) \coloneqq \lim_{s \searrow t} \widehat{M}_s(y)$ denote the left and right limit of the selection procedure, respectively.

For a given observation y, let $(t_k)_k$ denote the jump points of the selection procedure, i.e, the values of t for which $\widehat{M}_t^- \neq \widehat{M}_t^+$. Let $U^{t_k-} \coloneqq F(t_k, U_{M_{t_k}^-})$ and $U^{t_k+} \coloneqq F(t_k, U_{M_{t_k}^+})$ denote the associated left and right selection events. For each jump point t_k, y belongs to the intersection of the boundaries

$$\partial U^{t_k-} \cap \partial U^{t_k+},$$

which is an $(n-1)$-dimensional surface in $\mathbb{R}^n$. These surfaces are typically defined by some equation arising from a set of KKT conditions, some threshold operator or proximal operator. Either way, a (y-dependent) normal vector η_k of this surface in y is almost surely well defined (up to a scalar). Now, define the value of H as

$$H(t,y) = \begin{cases} \sum_{k:0<t_k<t} \frac{\langle \eta_k; \hat{\mu}^{t_k-0}(y) - \hat{\mu}^{t_k+0}(y) \rangle}{\langle \eta_k; f(y) \rangle} & \text{if } t > 0 \\ \sum_{k:t<t_k<0} \frac{\langle \eta_k; \hat{\mu}^{t_k+0}(y) - \hat{\mu}^{t_k-0}(y) \rangle}{\langle \eta_k; f(y) \rangle} & \text{if } t < 0. \end{cases} \tag{12}$$

If a term has zero denominator we set that term to zero by convention. Note that whether η_k points inwards or outwards is irrelevant for the value of H, so is its length (as long as it is non-zero). See Figure 3 for a visualisation of H.

In practice, if $\partial U^{t_k-} \cap \partial U^{t_k+}$ is characterised by a manifold $(G_k = 0)$ for some smooth $G_k : \mathbb{R}^n \to \mathbb{R}$, then the gradient ∇G_k qualifies as a normal vector (provided it is non-zero). In order to evaluate H we only need to evaluate the trajectory $t \mapsto \hat{\mu}^t$ and identify the flow, the jump points and a normal vector at each jump point.

We interpret $t \mapsto H(t,y)$ as the accumulated difference between neighbouring post-selection estimators on the trajectory $t \mapsto F(t,y)$ relative to how the field penetrates the boundary between the neighbouring regions. The denominators only depend on the selection events $(U_M)_{M \in \mathcal{M}}$ and the flow F, while the numerators depend on the post-selection estimators as well. The contribution from a given term is minimal if the field is perpendicular to the boundary.

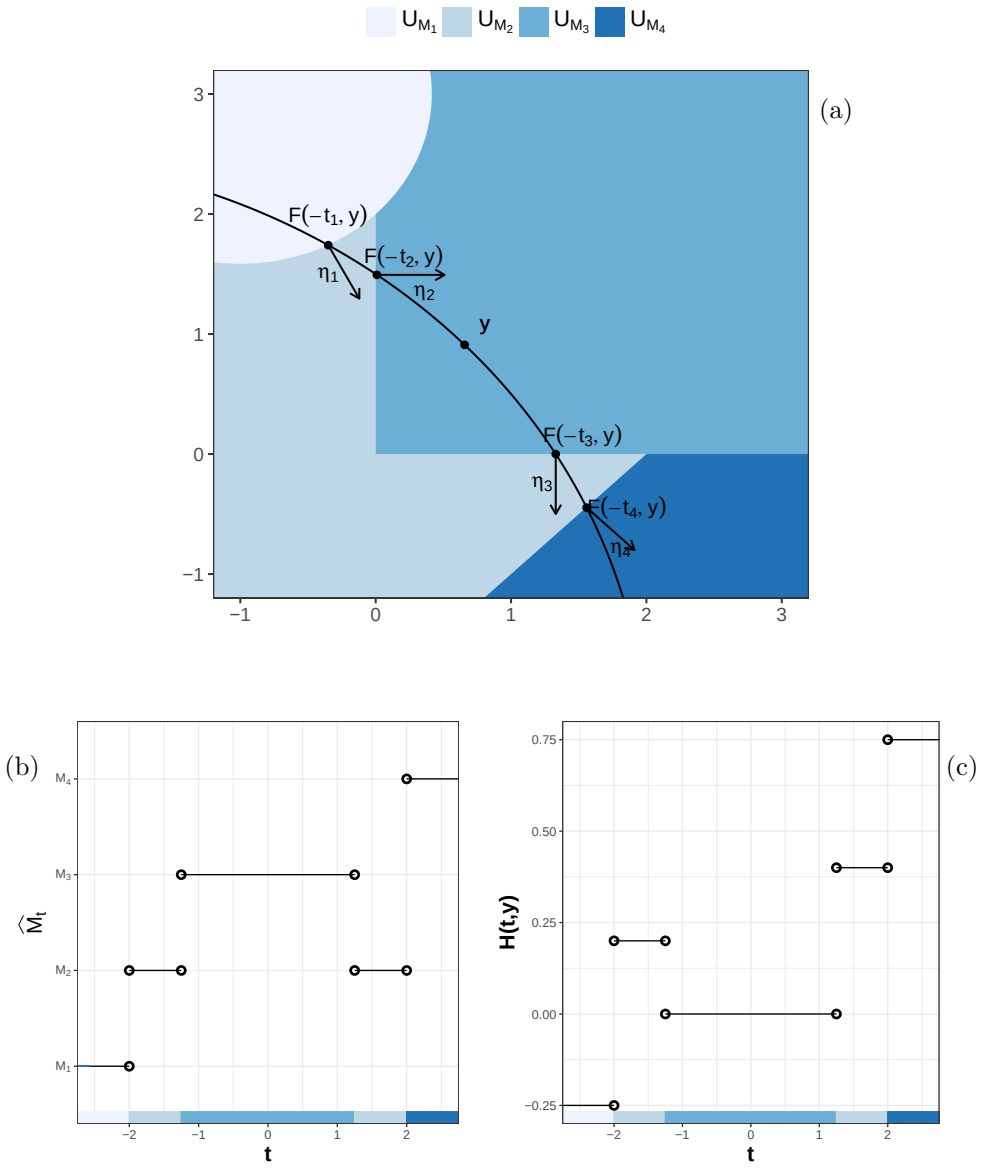


FIGURE 3. Visualisation of H in a generic setting on $\mathbb{R}^2$. (a) The jump points are identified when the trajectory $t \mapsto F(-t, y)$ intersects the boundaries. The normal vectors are marked at the crossings. (b) The selection procedure evaluated on the trajectory $t \mapsto F(-t, y)$. (c) The function $t \mapsto H(t, y)$, the jump sizes are determined by the normal vectors, the field and the difference in post-selection estimators across neighbouring regions.

3.2. **Estimating $\partial_t E(H(t,Y))$.** For fixed y the function $t \mapsto H(t,y)$ is a step function. Thus estimating $\partial_t E(H(t,Y))$ from a single realisation of Y is not straightforward. One method is to extend the approach of Mikkelsen & Hansen (2017) to this setting. Their approach relies on smoothing the step function before differentiating, which is analogous to estimating a weighted intensity for a jump process. The resulting estimate is not guaranteed to be unbiased, but smoothing appropriately may considerably reduce the variance. This approach was proven quite fruitful in the simulation study of Mikkelsen & Hansen (2017).

Let $\hat{\mu}$ be an estimator on the form (11) and let H be defined as in (12). If $t \mapsto E(H(t,Y))$ is differentiable (which it is under Assumption 4.4 given below), then the following algorithm provides and estimate of $\partial_t E(H(t,Y))$ given a single realisation y of Y:

Algorithm 3.2. *Input: observation y, estimator $\hat{\mu}^t$ given by (11) with selection procedure $\widehat{M}_t$ and flow F satisfying (9).*

(1) Evaluate the field f at y.

(2) Identify the jump points $(t_k)_k$, i.e., the $t \in \mathbb{R}$ where $\widehat{M}_t^- \neq \widehat{M}_t^+$.

(3) Evaluate the outer normal vectors $(\eta_k)_k$ of the boundaries

$$\partial U^{t_k-} \cap \partial U^{t_k+},$$

at y for each jump point t_k.

(4) Evaluate the weights

$$w_k = \frac{\langle \eta_k; \hat{\mu}^{t_k-0}(y) - \hat{\mu}^{t_k+0}(y) \rangle}{\langle \eta_k; f(y) \rangle}$$

for each t_k.

(5) Apply a weighted kernel density smoother to (t_k) with weights (w_k) for all k where $w_k > 0$. Denote the kernel density estimate by $p^+(t)$. Analogously, let $p^-(t)$ denote a kernel density estimate applied to (t_k) with weights $(-w_k)$ for all k where $w_k < 0$.

(6) Scale p^+ with the total sum of positive weights, i.e., $\sum_{k:w_k>0} w_k$. And similarly, scale p^- with the total sum of negative weights, i.e., $\sum_{k:w_k<0} w_k$

Output: $\widehat{\partial}_t := p^+(t) + p^-(t)$, an estimate of $(\partial_t E(H(t,Y)))_t$.

In the next section we describe how to show (7) for an estimator $\hat{\mu}$. If (7) is established, we propose the risk estimator

$$\widehat{\mathrm{Risk}} := \|Y - \hat{\mu}^t(Y)\|_2^2 - n\sigma^2 + 2\sigma^2(\mathrm{div}(\hat{\mu}^t)(Y) + \widehat{\partial}_t), \tag{13}$$

where $\widehat{\partial}_t$ is obtained from Algorithm 3.2. (13) is a natural extension of SURE, (4), to estimators with data adaptive model selection.

4. Properties of H

For notational simplicity, from this section and onwards we index selection events and post-selection estimators by $i = 1, ..., N$ instead of $M \in \mathcal{M} = \{M_1, ..., M_N\}$. Before presenting the conditions and the associated proof of (7), we first recall the decomposition of the degrees of freedom given by Mikkelsen & Hansen (2017). In this paper we extend the original framework to include a wider range of estimators.

4.1. **Decomposition of degrees of freedom.** Consider $i = 1, 2, ...$ models, each represented by an open selection event $U_i \subseteq \mathbb{R}^n$ and a post-selection estimator $\hat{\mu}_i : \overline{U}_i \to \mathbb{R}^n$. We assume that the selection events are disjoint and $\bigcup_i \overline{U}_i = \mathbb{R}^n$. We then define the estimator $\hat{\mu} = \sum_i 1_{U_i} \hat{\mu}_i$, which given a selection event $y \in U_i$ applies the post-selection estimator $\hat{\mu}_i$ to the observation y. Let $\mathcal{H}^{n-1}$ denote the $n-1$ dimensional Haussdorf measure on $\mathbb{R}^n$ and $\mathcal{L}^n$ denote the Lebesgue measure on $\mathbb{R}^n$. The estimator $\hat{\mu}$ is required to satisfy the following rather weak assumptions:

Assumption 4.1. *The estimator $\hat{\mu}$ can be written as $\hat{\mu} = \sum_{i=1}^{\infty} 1_{U_i} \hat{\mu}_i$ for a collection of open and disjoint sets $\{U_i\}_{i=1}^{\infty}$ with $\bigcup_{i=1}^{\infty} \overline{U}_i = \mathbb{R}^n$ and a collection of post-selection estimators $\{\hat{\mu}_i\}_{i=1}^{\infty}$, where each $\hat{\mu}_i : \overline{U}_i \to \mathbb{R}^n$ is locally Lipschitz. Additionally:*

(a) The relative boundaries are covered, i.e.,

$$\overline{U}_i \setminus \operatorname{int}(\overline{U}_i) \subseteq \bigcup_{j \neq i} \overline{U}_j, \quad i = 1, 2, ...$$

(b) The selection events have locally finite perimeter, i.e.,

$$\mathcal{H}^{n-1}(\partial U_i \cap K) < \infty, \quad i = 1, 2, ...$$

for all $K \subseteq \mathbb{R}^n$ compact.

(c) The random variable $\langle Y - \mu; \hat{\mu}(Y) \rangle$ has finite first moment and the random variable $\operatorname{div}(\hat{\mu})(Y) = \sum_{i=1}^{\infty} 1_{U_i}(Y) \operatorname{div}(\hat{\mu}_i)(Y)$ is either almost surely non-negative or has finite first moment.

(d) The boundary integrals are finite:

$$\sum_{i=1}^{\infty} \int_{\partial U_i} |\langle \eta_i; \hat{\mu}_i \rangle| \psi \, d\mathcal{H}^{n-1} < \infty$$

where η_i denotes the unit outer normal to ∂U_i.

Remark 4.2. Note that the above assumptions differ from the original assumptions by Mikkelsen & Hansen (2017) in that we allow for a countably infinite collection of disjoint selection events and associated post-selection estimators and that the polynomial bounds have been replaced by integrability assumptions. Moreover, note

(a) The relative boundary condition, (a), automatically holds when the collection $\{U_i\}_i$ is finite, since $\bigcup_{j \neq i} \overline{U}_j$ is closed and $(\overline{U}_i)_i$ cover the whole space. For the infinite case, the condition may break in some pathological examples. Moreover, this condition is only used in the proof of Lemma A.1 and for that proof to hold we only need

$$\overline{U}_i \setminus \left(\operatorname{int}(\overline{U}_i) \cup \bigcup_{j \neq i} \overline{U}_j \right)$$

to be $\mathcal{H}^{n-1}$-null for each $i = 1, 2...$

(b) The locally finite perimeter condition, (b), automatically holds for semi-algebraic sets, i.e., if the selection events are determined by a finite number of polynomial equalities and inequalities, see Lemma A.3. One can weaken the condition to

$$\mathcal{H}^{n-1}(\partial U_i \cap (\langle \eta_i; \hat{\mu}_i \rangle \neq 0) \cap K) < \infty, \quad i = 1, 2, ...$$

for all $K \subseteq \mathbb{R}^n$ compact. The price of this weakened condition is that it depends on the post-selection estimator and not only the selection events.

(c) The moment conditions, (c), are standard and completely analogous to the original moment conditions for Stein's Lemma.

(d) The convergence of the absolute boundary integrals is often the most difficult condition to verify, but it automatically follows from the other conditions via Lemma A.4 if a few additional conditions hold.

The decomposition theorem of Mikkelsen & Hansen (2017) states

Theorem 4.3. *If $\hat{\mu}$ satisfies Assumption 4.1 then*

$$\mathrm{df}(\hat{\mu}) = \mathrm{df}_{\mathrm{S}}(\hat{\mu}) + \frac{1}{2} \sum_{i \neq j} \int_{\overline{U}_i \cap \overline{U}_j} \langle \hat{\mu}_j - \hat{\mu}_i, \eta_i \rangle \psi(\,\cdot\,; \mu, \sigma^2)\, d\mathcal{H}^{n-1}, \tag{14}$$

where η_i denotes the measure theoretic outer unit normal to ∂U_i. Moreover, all quantities in (14) are finite.

We provide the proof of Theorem 4.3 under the weaker assumptions in the appendix. The above theorem has been extended to the case of countably infinite selection events, however we will only prove properties of $\partial_t E(H(t, Y))$ when it is constructed from an estimator $\hat{\mu}$ with a finite number of selection events. The reason for the extended version of the decomposition theorem, is that we apply it to a possibly infinite case in the *proof* of Theorem 4.6 given below.

4.2. **Establishing $\mathrm{df} - \mathrm{df}_{\mathrm{S}} = \partial_t E(H(t, Y))$.** We return to the setting in Section 3 with a finite number of selection events $(U_i)_{i=1}^N$. Before proceeding we define the set

$$\Phi := \left\{ z \in \bigcup_i \partial U_i \;\middle|\; \langle \eta(z); f(z) \rangle \neq 0 \right\}. \tag{15}$$

Geometrically, $z \in \Phi$ if it belongs to some boundary ∂U_i and the field at z is not tangent to the boundary in z. We stress that the set Φ only depends on the selection events $(U_i)_i$ and the flow F, in particular it does not depend on the post-selection estimator. The importance of Φ will soon become apparent.

The following assumptions are required to prove Theorem 4.6 below.

Assumption 4.4. *The estimator $\hat{\mu} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ with tuning parameter $t \in \mathbb{R}$ can be written as $\hat{\mu}(t, y) = \sum_{i=1}^N 1_{F(t,U_i)}(y) \hat{\mu}_i(t, y)$ for a collection of open and disjoint sets $\{U_i\}_{i=1}^N$ with $\bigcup_{i=1}^N \overline{U}_i = \mathbb{R}^n$ and a C^2 flow $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. Furthermore,*

(a) For each $t \in \mathbb{R}$ the estimator $\hat{\mu}(t, \cdot)$ satisfies Assumption 4.1.

(b) For each $t \in \mathbb{R}$ the mapping $G_t : y \mapsto H(t, y) f(y)$ satisfies Assumption 4.1.

(c) Almost surely the number of jump points, $(t_k)_k$, is finite.

(d) The surface ∂U_i is $\mathcal{H}^{n-1}$-almost-everywhere C^2 for all $i = 1, ..., N$.

(e) For each $t_0 \in \mathbb{R}$ there exists $\delta > 0$ and $g_1, g_2 : \mathbb{R}^n \to \mathbb{R}$ such that $g_1(Y), g_2(Y)$ both have finite first moment and

$$|\langle G_t(y); y\rangle| \leq g_1(y), \quad |\mathrm{div}(G_t)(y)| \leq g_2(y)$$

for all (t, y) with $t \in [t_0 - \delta; t_0 + \delta]$ and $y \in F(t, \Phi)$.

Remark 4.5. Condition (d) is an extremely weak condition and almost follows from (a). More precisely, Theorem 5.7.2 in Evans & Gariepy (1992) gives that ∂U_i is $\mathcal{H}^{n-1}$-almost-everywhere C^1, whenever U_i has locally finite perimeter. Assumption 4.4 is most easily verified via Lemma 4.8 given below.

Theorem 4.6. *If $\hat{\mu}$ satisfies Assumption 4.4 then*

$$\partial_t E(H(t, Y)) = \frac{1}{2} \sum_{i \neq j} \int_{\overline{F(t,U_i)} \cap \overline{F(t,U_j)} \cap F(t,\Phi)} \langle \hat{\mu}_j - \hat{\mu}_i, \eta_i \rangle \psi(\,\cdot\,; \mu, \sigma^2)\, d\mathcal{H}^{n-1}, \tag{16}$$

where η_i denotes the measure theoretic outer unit normal to $\partial F(t, U_i)$.

Once we have established (16) via Theorem 4.6, we see the importance of Φ — it characterises which parts of the boundary integrals we can recover via H. Comparing (16) with

$$\mathrm{df}(\hat{\mu}^t) - \mathrm{df}_S(\hat{\mu}^t) = \frac{1}{2} \sum_{i \neq j} \int_{\overline{F(t,U_i)} \cap \overline{F(t,U_j)}} \langle \hat{\mu}_j - \hat{\mu}_i, \eta_i \rangle \psi(\,\cdot\,; \mu, \sigma^2)\, d\mathcal{H}^{n-1}$$

from Theorem 4.3, we immediately see that if the integrand in (16) vanishes outside of $F(t, \Phi)$, then we can conclude (7). We state this as a Corollary:

Corollary 4.7. *If $\hat{\mu}$ satisfies Assumption 4.4 and*

$$\langle \eta_i; \hat{\mu}_j - \hat{\mu}_i \rangle = 0 \quad \mathcal{H}^{n-1}\text{-a.e. on } \overline{U}_i \cap \overline{U}_j \cap \Phi^c \tag{17}$$

for all $i \neq j$, then

$$\mathrm{df}(\hat{\mu}^t) = \mathrm{df}_S(\hat{\mu}^t) + \partial_t E(H(t, Y)). \tag{18}$$

Verifying conditions (a) and (c) in Assumption 4.4 is quite easy and often shown using Lemma A.3, Lemma A.4 and Remark 4.2. However, condition (b) and (e) in Assumption 4.4 may prove difficult and the following lemma may be consulted:

Lemma 4.8. *If the estimator $\hat{\mu} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ with tuning parameter $t \in \mathbb{R}$ can be written as $\hat{\mu}(t, y) = \sum_{i=1}^N 1_{F(t,U_i)}(y) \hat{\mu}_i(t, y)$, where:*

(a) The selection events $(U_i)_i$ are semi-algebraic and open.

(b) The function F is a C^2-flow and for each $t \in \mathbb{R}$ both $y \mapsto \|F(t, y)\|$ and $y \mapsto \|\partial_y F(t, y)\|$ are polynomially bounded. Moreover, almost surely the number of jumps $(t_k)_k$ is bounded.

(c) For each $t \in \mathbb{R}$ and $i = 1, ..., N$ the post-selection estimator $\hat{\mu}_i(t, \cdot) : F(t, \overline{U}_i) \to \mathbb{R}^n$ is locally Lipschitz and $\|\hat{\mu}_i\|$ is polynomially bounded. Moreover, either the random variables $(1_{F(t,U_i)} \mathrm{div}(\hat{\mu}_i(t, Y)))_i$ are all almost surely non-negative or they all have finite first moment.

(d) For each $i, j = 1, ..., N$ with $j \neq i$ the function

$$h_{ij}(t, y) = \frac{\langle \eta_i(y); \hat{\mu}_i(t, y) - \hat{\mu}_j(t, y) \rangle}{\langle \eta_i(y); f(y) \rangle} \tag{19}$$

defined on the manifold $(t,y) \in \{(t,y) \mid y \in \overline{F(t,U_i)} \cap \overline{F(t,U_j)} \cap F(t,\Phi)\}$ *is locally Lipschitz and for each* $t_0 \in \mathbb{R}$ *there exist* $\delta > 0$*, such that both* $\sup_{t\in[t_0-\delta;t_0+\delta]} |h_{ij}|$ *and* $\sup_{t\in[t_0-\delta;t_0+\delta]} \|\partial_y h_{ij}\|$ *are polynomially bounded.*

Then $\hat{\mu}$ *satisfies Assumption 4.4.*

The above lemma is applied throughout the paper and it is the main method for applying Theorem 4.6 and Corollary 4.7. The proof of the lemma is postponed to the appendix. Note that the assumptions in Lemma 4.8 are quite transparent in terms of which conditions depend on the selection events and which depend on the post-selection estimators. If a given estimator $\hat{\mu}$ satisfies the assumptions of Lemma 4.8, then one would only need to verify (c) and (d) if the post-selection estimators are replaced.

5. Applications in selection procedures with tuning parameters

In this section we consider estimators $\hat{\mu}^t$ with data adaptive selection procedures and a tuning parameter $t \in \mathbb{R}$, such that

$$\mathrm{df}(\hat{\mu}^t) = \mathrm{df}_{\mathrm{S}}(\hat{\mu}^t) + \partial_t E(H(t,Y)) \tag{20}$$

can be proven via Corollary 4.7. There are plenty of other estimators for which Theorem 4.6 applies. Assumption 4.4 is generally quite weak, thus the main requirement of the estimator is that the selection procedure and the tuning parameter interplays through a *flow*, i.e., model i is selected if and only if $Y \in F(t,U_i)$ for some smooth flow F. If such a flow exists and Assumption 4.4 holds, then Corollary 4.7 only amounts to checking (17). Even if (17) is not true we can still conclude that

$$\begin{aligned}\mathrm{df}(\hat{\mu}^t) =& \mathrm{df}_{\mathrm{S}}(\hat{\mu}^t) + \partial_t E(H(t,Y)) \\ &+ \frac{1}{2}\sum_{i\neq j} \int_{\overline{F(t,U_i)}\cap\overline{F(t,U_j)}\setminus F(t,\Phi)} \langle \hat{\mu}_j - \hat{\mu}_i, \eta_i\rangle \psi(\,\cdot\,;\mu,\sigma^2)\, d\mathcal{H}^{n-1}\end{aligned}$$

and thus think of $\mathrm{df}_{\mathrm{S}}(\hat{\mu}^t) + \partial_t E(H(t,Y))$ as a higher order approximation of $\mathrm{df}(\hat{\mu}^t)$.

In the following we consider four examples: marginal screening, relaxed lasso, best subset selection and singular value decompositions with a hard threshold on the singular values. Three first are all examples from linear regression, in which we consider a fixed $n \times p$-dimensional design matrix X and an estimatable coefficient vector $\beta \in \mathbb{R}^p$. The columns of X are refered to as *predictors*. In regression, model selection comes down to selecting the predictors. Hence the models are indexed by subsets $A \subseteq \{1,...,p\}$ of the predictor indeces. For every such subset $A \subseteq \{1,...,p\}$, with size $|A|$, let X_A denote the design matrix restricted to the columns listed in A and let $\beta_A \in \mathbb{R}^{|A|}$ denote the coefficients restricted to the coordinates listed in A. Similarly, X_{-A} denotes $X_{\{1,...,p\}\setminus A}$ and β_{-A} denotes $\beta_{\{1,...,p\}\setminus A}$. Finally, let Π_{X_A} denote the orthogonal projection onto X_A.

Additionally, we introduce the following linear algebra notation: for $n \times m$-matrices B let B^+ denote the Moore-Penrose inverse of B and let $\mathrm{span}(B)$ denote the span, i.e., the image of linear function $x \mapsto Bx$ defined on $\mathbb{R}^m$. Similarly, let $\mathrm{span}_>(B)$ and $\mathrm{span}_\geq(B)$ denote the image of $\mathbb{R}^n_{>0} := \{x \in \mathbb{R}^m \mid x > 0\}$ and $\mathbb{R}^n_{\geq 0} := \{x \in \mathbb{R}^m \mid x \geq 0\}$, respectively.

In the examples we will also provide explicit formulas for H. Changing (12) by an additive constants does not change properties of $\partial_t E(H(t,Y))$. We therefore

specify H on the form

$$H(t,y) = \sum_{k:t_k<t} \frac{\langle \eta_k; \hat{\mu}^{t_k-0}(y) - \hat{\mu}^{t_k+0}(y)\rangle}{\langle \eta_k; f(y)\rangle}, \quad t \in \mathbb{R}$$

instead.

Example 5.1 (Marginal screening)**.** We make the assumption on X that no pair of the columns $(x_i)_{i=1}^p$ are identical up to a change of sign. *Marginal screening* is a two-step estimator with tuning parameter $t \in \mathbb{R}$. At first the inner products $\langle y; x_i\rangle$ between the observation y and the columns of X are evaluated. The inner products can be viewed as measures of marginal association between the observations and the predictors. The first step of the *marginal screening* estimator identifies the predictors whose inner product exceeds e^t in magnitude. This is called the *active set*:

$$\mathcal{A}_t \coloneqq \left\{1 \le i \le p \;\middle|\; |\langle y; x_i\rangle| > e^t\right\}. \tag{21}$$

In the second step, a linear model is fitted to the *active set* only, i.e., $\hat{\mu}^t_{\text{ms}} \coloneqq \Pi_{X_{\mathcal{A}_t}}$.

Firstly, we identity the flow. For each $A \subseteq \{1, ..., p\}$ let U_A denote the affine sets

$$U_A \coloneqq \left\{ y \in \mathbb{R}^n \;\middle|\; \begin{array}{l} |\langle y; x_i\rangle| > 1 \text{ if } i \in A \\ |\langle y; x_i\rangle| < 1 \text{ if } i \notin A \end{array} \right\}. \tag{22}$$

Clearly, U_A is open and $\bigcup_{A\subseteq\{1,...,p\}} \overline{U}_A = \mathbb{R}^n$. Moreover, $y \in e^t U_A$ implies $\mathcal{A}_t = A$ and we see that

$$\hat{\mu}^t_{\text{ms}} = \sum_{A\subseteq\{1,...,p\}} 1_{F(t,U_A)} \Pi_{X_A}$$

holds Lebesgue almost everywhere, where $F(t,y) = e^t y$ is the flow. For $y \in \mathbb{R}^n$ the number of crossings between regions of $t \mapsto F(-t,y)$ is at most p.

For the outer normal vectors, note that

$$\partial F(t_0, U_A) \subseteq \{y \mid |\langle y; x_i\rangle| = e^{t_0} \text{ for some } 1 \le i \le p\}.$$

Clearly the outer normal at y, $\eta(y)$, is proportional to x_i for an x_i with $|\langle y; x_i\rangle| = e^{t_0}$. The set of y for which multiple x_i satisfy the equality is an affine set of dimension $n-2$, thus η is well defined $\mathcal{H}^{n-1}$ almost everywhere on $\partial F(t_0, U_A)$. The field is simply $f(y) = \partial_t F(0,y) = y$. These two observations together show that for $y \in \partial F(t, U_A)$ we have $\langle \eta(y); f(y)\rangle \propto e^t$ and we conclude that $\Phi \setminus \bigcup_A \partial U_A$ is $\mathcal{H}^{n-1}$-null.

From the above we conclude (a)-(c) in Lemma 4.8. To verify condition (d) observe that if $y \in \partial F(t_0, U_{A_1}) \cap \partial F(t_0, U_{A_2})$ for some $t_0 \in \mathbb{R}$ and $A_1, A_2 \subseteq \{1, ..., p\}$, then we must have $|\langle y; x_j\rangle| = e^{t_0}$ for all $j \in A_1 \triangle A_2$. By a dimensionality argument we conclude that $A_1 \triangle A_2$ is a singleton and we assume without loss of generality that $A_2 = A_1 \cup \{j\}$ for some $j \notin A_1$. Therefore,

$$h_{A_1,A_2}(t_0, y) = \frac{\langle x_j; (\Pi_{A_1} - \Pi_{A_2})y\rangle}{\langle x_j; y\rangle} = \frac{\langle \Pi_{A_1} x_j; y\rangle}{\langle x_j; y\rangle} - 1, \tag{23}$$

which is always between -1 and 0, so it is polynomially bounded. Moreover, its derivative with respect to y is

$$\partial_y h_{A_1,A_2}(t,y) = \frac{\Pi_{A_1} x_j - x_j h_{A_1,A_2}(t,y)}{\langle x_j; y\rangle}. \tag{24}$$

So $\sup_{t\in[t_0-\delta;t_0+\delta]} \|\partial_y h_{A_1,A_2}(t,y)\| \leq 2\|x_j\|_2 e^{-t_0+\delta}$ and the right hand side is constant in y and especially polynomially bounded. Lemma 4.8 holds and we conclude that

$$\mathrm{df}(\hat{\mu}^t_{\mathrm{ms}}) = \mathrm{df}_{\mathrm{S}}(\hat{\mu}^t_{\mathrm{ms}}) + \partial_t E(H(t,Y)), \tag{25}$$

where

$$H(t,y) = \sum_{i:|\langle y;x_i\rangle|<e^t} 1 - \frac{\langle y; \Pi_{\mathcal{A}_{|\langle y;x_i\rangle|}} x_i\rangle}{\langle y; x_i\rangle}. \tag{26}$$

Example 5.2 (Relaxed lasso)**.** This example is an extension of Example 2.5 by Mikkelsen & Hansen (2017), in which the *lasso-OLS estimator* is considered. This estimator performs the OLS estimator on the model selected by the *lasso estimator* (by Tibshirani (1994)). An extension of this estimator is the *relaxed lasso estimator* proposed by Meinshausen (2007). This example showcases how changing the post-selection estimators will change H, but not the validity of (20).

The *lasso estimator* $\hat{\mu}^t_{\mathrm{lasso}}(y)$ with tuning parameter $t \in \mathbb{R}$ is defined as $\hat{\mu}^t_{\mathrm{lasso}}(y) = X\hat{\beta}^t$ where

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

Likewise, $\hat{\mu}^t_{\mathrm{lasso},A}(y)$ denotes the lasso estimator with support restricted to $A \subseteq \{1,...,p\}$, i.e., $\hat{\mu}^t_{\mathrm{lasso},A}(y) = X_A\hat{\beta}^t_A$ where

$$\hat{\beta}^t_A \in \arg\min_{\beta_A} \frac{1}{2}\|y - X_A\beta_A\|_2^2 + e^t\|\beta_A\|_1.$$

We make the assumption that the columns of X are in general position, thus making $\hat{\beta}^t$ unique for all y and all $t \in \mathbb{R}$ (see, e.g., Lemma 3 by Tibshirani (2013)). With this assumption we can define the *active set* $\mathcal{A}_t := \mathrm{supp}(\hat{\beta}^t)$ for all y and $t \in \mathbb{R}$.

The *relaxed lasso estimator* is given by $\hat{\mu}^{t,\phi}_{\mathrm{rl}} := X\hat{\beta}^{t+\log(\phi)}_{\mathcal{A}_t}$, where $\phi \in [0,1)$ is a second tuning parameter. In other words, the relaxed lasso estimator first applies the lasso estimator and then refits data using another lasso estimator with the smaller tuning parameter $t + \log(\phi)$ and the support determined by the first lasso step. Having $\phi = 0$ corresponds to the lasso-OLS estimator. It was shown by Meinshausen (2007) that the relaxed lasso has attractive statistical properties and that it is computationally lighter than *bridge estimators*, which minimises the loss given in Example 3.1 with the concave penalty $\mathrm{Pen}(\beta) = \|\beta\|_\gamma^\gamma$, $0 < \gamma < 1$.

Fix $\phi \in [0,1)$ and consider the selection events for $\hat{\mu}^{t,\phi}_{\mathrm{rl}}$. For $A \subseteq \{1,...,p\}$ let

$$\begin{aligned} U_A :=& \mathrm{int}(\mathcal{A}_t = A) \\ =& \mathrm{int}\left\{ y \in \mathbb{R}^n \,\middle|\, \inf_{\beta_A} \frac{1}{2}\|y - X_A\beta_A\|_2^2 + \|\beta_A\|_1 = \inf_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \|\beta\|_1 \right\} \end{aligned} \tag{27}$$

denote the interior of the selection event $(\mathcal{A}_t = A)$ for the first lasso step. We refer to Example 2.5 in Mikkelsen & Hansen (2017) for a proof that U_A is an affine set and thus semi-algebraic. Since $\|\cdot\|_1$ is a positive homogeneous penalty we see from Example 3.1 that $F(t,U_A)$ is the selection event $(\mathcal{A}_t = A)$, where $F(t,y) = e^t y$.

Once the first lasso step has been executed, the post-selection estimator is another evaluation of the lasso estimator or, in the case $\phi = 0$, the OLS estimator

with fixed support. Thus given the selection event $(\widehat{A}_t = A)$, the post-selection estimator is

$$\hat{\mu}_A = \begin{cases} \hat{\mu}_{\text{lasso},A}^{t+\log\phi} & \text{if } 0 < \phi < 1 \\ \Pi_A & \text{if } \phi = 0 \end{cases}$$

In either case the post-selection estimator is Lipschitz continuous. That $\hat{\mu}_{\text{lasso},A}^{t+\log\phi}$ is Lipschitz continuous follows from Lemma 3 in Tibshirani & Taylor (2012), which states that $\text{id} - \hat{\mu}_{\text{lasso},A}^{t+\log\phi}$ is a projection onto a convex polytope. Lipschitz continuous functions have bounded derivatives, thus conditions (a)-(c) holds in Lemma 4.8.

Before characterising the normal vectors, note that in light of Sections 2.2 and 2.3 in Tibshirani (2013) we can re-write the selection events as

$$\begin{aligned} U_A &= \bigcup_{s \in \{-1,1\}^{|A|}} U_{A,s}, \quad \text{with} \\ U_{A,s} &= \text{span}_{>0}(X_A \text{diag}(s)) + \{w \in \text{col}(X_A)^\perp \mid \|X_{-A}^T w\|_\infty < 1\} + (X_A^T)^+ s. \end{aligned} \tag{28}$$

Hence U_A can be decomposed into regions representing different signs, s, of the active parameters. Each of these regions can be decomposed into an element in the positive span of $X_A \text{diag}(s)$, an element w from a polytope embedded in the the orthogonal complement of the column space of X_A and a constant vector depending on the sign, $(X_A^T)^+ s$.

For pairs of subsets $A_1, A_2 \subseteq \{1, ..., p\}$ with U_{A_1}, U_{A_2} non empty and distinct, consider $\overline{U}_{A_1} \cap \overline{U}_{A_2}$. If y belongs to this intersection, then it is the limit of two sequences from each set, i.e.,

$$\begin{aligned} y &= \lim_{n\to\infty} X_{A_1} \text{diag}(s_{A_1}) \alpha_n^1 + \lim_{n\to\infty} w_n^1 + (X_{A_1}^T)^+ s_{A_1} \\ y &= \lim_{n\to\infty} X_{A_2} \text{diag}(s_{A_2}) \alpha_n^2 + \lim_{n\to\infty} w_n^2 + (X_{A_2}^T)^+ s_{A_2} \end{aligned} \tag{29}$$

for some $s_{A_1} \in \{-1,1\}^{|A_1|}$, $s_{A_2} \in \{-1,1\}^{|A_2|}$ fixed and sequences $(\alpha_n^1)_n \subset \mathbb{R}_{>0}^{|A_1|}$, $(\alpha_n^2)_n \subset \mathbb{R}_{>0}^{|A_2|}$, $(w_n^1)_n \subset \text{col}(X_{A_1})^\perp$ and $(w_n^2)_n \subset \text{col}(X_{A_2})^\perp$. By continuity of the unique lasso solutions the fitted value of y is in the span of $X_{A_1 \cap A_2}$ and consequently we have

$$\overline{U}_{A_1} \cap \overline{U}_{A_2} \subseteq \bigcup_{s \in \{-1,1\}^{|A_1 \cup A_2|}} \begin{pmatrix} \text{span}_{\geq 0}(X_{A_1 \cap A_2} \text{diag}(s_{A_1 \cap A_2})) + (X_{A_1 \cup A_2}^T)^+ s \\ + \{w \in \text{col}(X_{A_1 \cup A_2})^\perp \mid \|X_{-(A_1 \cup A_2)}^T w\|_\infty \leq 1\} \end{pmatrix} \tag{30}$$

where $s_{A_1 \cap A_2}$ are the signs in $s \in \{-1,1\}^{|A_1 \cup A_2|}$ corresponding to the indeces $A_1 \cap A_2$. Analogously, any element in the right hand side of (30) also belongs to $\overline{U}_{A_1} \cap \overline{U}_{A_2}$ by suitably choosing sequences in (29). In order for $\overline{U}_{A_1} \cap \overline{U}_{A_2}$ to have dimension $n-1$ it must hold that either $A_1 \subset A_2$ or $A_2 \subset A_1$ and the ranks of X_{A_1} and X_{A_2} differ by one. In this case, we see from the right hand side of (30) that the normal vector belongs to $\text{col}(X_{A_2}) \ominus \text{col}(X_{A_1})$ if $A_1 \subset A_2$. Provided that it is non-zero $\eta = (\Pi_{A_2} - \Pi_{A_1})y$ is a candidate normal vector at $y \in \overline{U}_{A_1} \cap \overline{U}_{A_2}$. To show that it is non-zero, observe that by (30) there is some sign vector $s \in \{-1,1\}^{|A_2|}$ such that

$$(\Pi_{A_2} - \Pi_{A_1})y = (\Pi_{A_2} - \Pi_{A_1})(X_{A_2}^T)^+ s = (\Pi_{A_2} - \Pi_{A_1})(\text{diag(s)} X_{A_2}^T)^+ \mathbf{1}. \tag{31}$$

If the right hand side is zero, then it means that $(\text{diag(s)} X_{A_2}^T)^+ \mathbf{1}$ belongs to $\text{col}(X_{A_1})$, which contradicts the general position assumption. Hence $\eta = (\Pi_{A_2} -$

$\Pi_{A_1})y$ is non-zero and only depends on y through the active set and sign vector and

$$\langle \eta; y\rangle = e^t \|(\Pi_{A_2} - \Pi_{A_1})(\mathrm{diag(s)} X_{A_2}^T)^+ \mathbf{1}\|_2^2, \text{ for all } y \in \overline{F(t,U_{A_1})} \cap \overline{F(t,U_{A_2})} \tag{32}$$

where s is the sign vector $\mathrm{sign}(X_A^T(y - \hat{\mu}^{e^t}_{\mathrm{lasso}}(y)))$. This in particular shows that $\Phi \setminus \bigcup_A \partial U_A$ is $\mathcal{H}^{n-1}$-null.

If $y \in \overline{F(t,U_{A_1})} \cap \overline{F(t,U_{A_2})}$ the jump quantity (19) reads

$$\begin{aligned} h_{A_1,A_2}(t,y) &= \frac{\langle \eta; \hat{\mu}_{A_1}(y) - \hat{\mu}_{A_2}(y)\rangle}{\langle \eta; y\rangle} = \frac{\langle (\Pi_{A_2} - \Pi_{A_1})y; \hat{\mu}_{A_1}(y) - \hat{\mu}_{A_2}(y)\rangle}{\langle (\Pi_{A_2} - \Pi_{A_1})y; y\rangle} \\ &= \begin{cases} \frac{\langle (\Pi_{A_2} - \Pi_{A_1})y; \hat{\mu}_{A_1}(y) - \hat{\mu}_{A_2}(y)\rangle}{\langle (\Pi_{A_2} - \Pi_{A_1})y; y\rangle} & \text{if } 0 < \phi < 1 \\ -1 & \text{if } \phi = 0 \end{cases} \end{aligned} \tag{33}$$

or, equivalently:

$$h_{A_1,A_2}(t,y) = -\frac{\langle \hat{\mu}^{-\infty}_{\mathrm{lasso},A_2}(y) - \hat{\mu}^{-\infty}_{\mathrm{lasso},A_1}(y); \hat{\mu}^{t+\log(\phi)}_{\mathrm{lasso},A_2}(y) - \hat{\mu}^{t+\log(\phi)}_{\mathrm{lasso},A_1}(y)\rangle}{\langle \hat{\mu}^{-\infty}_{\mathrm{lasso},A_2}(y) - \hat{\mu}^{-\infty}_{\mathrm{lasso},A_1}(y); \hat{\mu}^{-\infty}_{\mathrm{lasso},A_2}(y) - \hat{\mu}^{-\infty}_{\mathrm{lasso},A_1}(y)\rangle}, \tag{34}$$

where $\hat{\mu}^{-\infty}_{\mathrm{lasso},A} = \Pi_A$ corresponds to the limit of the relaxed lasso as $\phi \to 0$. We saw in (32) that the denominator equals $e^t c$, where c is one of finitely many positive constants. In particular, for all $T \in \mathbb{R}$ the functions

$$y \mapsto \sup_{t\in[-T;T]} |h_{A_1,A_2}(t,y)|, \quad y \mapsto \sup_{t\in[-T;T]} \|\partial_y h_{A_1,A_2}(t,y)\|$$

are polynomially bounded by the Lipschitz continuity of the lasso estimator. By Lemma (4.8) and Corollary 4.7 we conclude that (20) holds.

Let $\mathcal{A}_t^-$ and $\mathcal{A}_t^+$ denote the left and right limit of $t \mapsto \mathcal{A}_t$. For given y with jump points $(t_k)_k$ from the first lasso step, the function H reads:

$$H(t,y) = \sum_{k:t_k<t} (1_{\mathcal{A}^+_{t_k} \subset \mathcal{A}^-_{t_k}} - 1_{\mathcal{A}^+_{t_k} \supset \mathcal{A}^-_{t_k}}) \frac{\langle (\Pi_{\mathcal{A}^-_{t_k}} - \Pi_{\mathcal{A}^+_{t_k}})y; \hat{\mu}^{t_k+\log(\phi)}_{\mathrm{lasso},\mathcal{A}^-_{t_k}}(y) - \hat{\mu}^{t_k+\log(\phi)}_{\mathrm{lasso},\mathcal{A}^+_{t_k}(y)}\rangle}{\|\Pi_{\mathcal{A}^-_{t_k}} y - \Pi_{\mathcal{A}^+_{t_k}} y\|_2^2}. \tag{35}$$

Example 5.3 (Best subset selection)**.** The Lagrange formulation of the best subset selection estimator of μ with tuning parameter $t \in \mathbb{R}$, $\hat{\mu}^t_{\mathrm{bs}}$, is given by

$$\hat{\mu}^t_{\mathrm{bs}} = X\hat{\beta}^t \quad \text{where} \quad \hat{\beta}^t = \arg\min_\beta \frac{1}{2}\|Y - X\beta\|_2^2 + e^t|\mathrm{supp}(\beta)|. \tag{36}$$

This estimator also fits the framework of Theorem 4.6. However, we will soon realise that (18) does not hold.

By Example 3.1 we have that $\hat{\mu}^t_{\mathrm{bs}} = \sum_{A\subseteq\{1,\dots,p\}} 1_{F(t,U_A)}\Pi_A$ (Lebesgue a.e.), where $F(t,y) = e^{t/2}y$ and

$$\begin{aligned} U_A &:= \left\{ y \in \mathbb{R}^n \;\middle|\; |A| - \frac{1}{2}\|\Pi_A y\|_2^2 < \min_{B\in\mathbb{P}_A} |B| - \frac{1}{2}\|\Pi_B y\|_2^2 \right\}, \\ &\text{where } \mathbb{P}_A := \{B \subseteq \{1,\dots,p\} \mid \mathrm{col}(X_B) \neq \mathrm{col}(X_A)\}, \end{aligned} \tag{37}$$

for all $A \subseteq \{1,\dots,p\}$. The sets $(U_A)_A$ are directly determined by quadratic equations and are thus semi-algebraic. The post-selection estimators are linear, so we have already verified condition (a)-(c) in Lemma 4.8.

To verify the last condition of Lemma 4.8, we identify the normal vectors. Assume $y \in \overline{F(t,U_{A_1})} \cap \overline{F(t,U_{A_2})}$ for some $A_1, A_2 \subseteq \{1,...,p\}$. If the column space of X_{A_1} and X_{A_2} are equal, $\mathrm{col}(X_{A_1}) = \mathrm{col}(X_{A_2})$, then

$$\int_{\overline{F(t,U_{A_1})} \cap \overline{F(t,U_{A_2})}} \langle \hat{\mu}_{A_2} - \hat{\mu}_{A_1}; \eta_{A_1} \rangle \, d\mathcal{H}^{n-1} = 0,$$

and we can safely ignore that case. By construction we have

$$-\frac{1}{2}\|\Pi_{A_1} y\|_2^2 + e^t |A_1| = -\frac{1}{2}\|\Pi_{A_2} y\|_2^2 + e^t |A_2|, \tag{38}$$

or equivalently, $\|\Pi_{A_2} y\|_2^2 - \|\Pi_{A_1} y\|_2^2 = 2e^t(|A_2| - |A_1|)$. Thus $\eta = \Pi_{A_2} y - \Pi_{A_1} y$ qualifies as a normal vector and

$$\langle \eta; f(y) \rangle = \frac{1}{2}\langle \Pi_{A_2} y - \Pi_{A_1} y; y \rangle = \frac{1}{2}(\|\Pi_{A_2} y\|_2^2 - \|\Pi_{A_1} y\|_2^2) = e^t(|A_2| - |A_1|)$$

which is 0 if and only if $|A_1| = |A_2|$. Hence

$$\bigcup_i \partial U_i \setminus \Phi = \bigcup_{(A_1, A_2) \in \mathcal{P}} \overline{U}_{A_1} \cap \overline{U}_{A_2},$$

where

$$\mathcal{P} := \left\{ A_1, A_2 \subseteq \{1,...,p\} \;\middle|\; |A_1| = |A_2|, \ \mathrm{col}(X_{A_1}) \neq \mathrm{col}(X_{A_2}) \right\}$$

denotes all pairs of subsets of $\{1,...,p\}$ having same size and but indexing different column spaces. Let $(A_1, A_2) \notin \mathcal{P}$. The jump term h_{A_1,A_2} from (19) defined whenever $y \in \overline{F(t,U_{A_1})} \cap \overline{F(t,U_{A_2})}$ simplifies to

$$h_{A_1,A_2}(t,y) = -\frac{\|\Pi_{A_2} y - \Pi_{A_1} y\|_2^2}{e^t(|A_2| - |A_1|)}.$$

Differentiating h_{A_1,A_2} with respect to y gives

$$\partial_y h_{A_1,A_2}(t,y) = -\frac{2(\Pi_{A_2} - \Pi_{A_1})y}{e^t(|A_2| - |A_1|)}.$$

For any compact subset of the tuning parameter space, $K \subseteq \mathbb{R}$, there exists a polynomial p bounding both $\sup_{t \in K} \|h_{A_1,A_2}\|$ and $\sup_{t \in K} \|\partial_y h_{A_1,A_2}\|$. Assumption 4.4 holds by Lemma 4.8.

For given y we derive the value of $H(t,y)$. For $k = 0, ..., \min(n,p)$ let $\widehat{A}_k$ denote the best subset of size k, i.e., $|\widehat{A}_k| = k$ and

$$\|\Pi_{\widehat{A}_k} y\|_2^2 \geq \|\Pi_A y\|_2^2, \quad \text{for all } A \subseteq \{1,...,p\} \text{ with } |A| = k.$$

Then in the Lagrange formulation of the best subset selection problem, (36), the jump times $(t_k)_{k=0}^{\min(n,p)-1}$ used in H are given by

$$2e^{t_k} = \|\Pi_{\widehat{A}_{k+1}} y\|_2^2 - \|\Pi_{\widehat{A}_k} y\|_2^2, \tag{39}$$

using (38). Hence H reads:

$$H(t,y) = 2 \sum_{k : t_k < t} \frac{\|\Pi_{\widehat{A}_{k+1}} y - \Pi_{\widehat{A}_k} y\|_2^2}{\|\Pi_{\widehat{A}_{k+1}} y\|_2^2 - \|\Pi_{\widehat{A}_k} y\|_2^2}, \tag{40}$$

up to an additive constant.

From Theorem 4.6 and the above we conclude that

$$\begin{aligned}\mathrm{df}(\hat{\mu}^t_{\mathrm{bs}}) =& \mathrm{df}_{\mathrm{S}}(\hat{\mu}^t_{\mathrm{bs}}) + \partial_t E(H(t,Y)) \\ &+ \frac{1}{2}\sum_{(A_1,A_2)\in\mathcal{P}} \int_{\overline{F(t,U_{A_1})}\cap\overline{F(t,U_{A_2})}} \|\Pi_{A_2}y - \Pi_{A_1}y\|_2 \; d\mathcal{H}^{n-1}(y).\end{aligned}$$

Here H is given in (40) and $\mathrm{df}_{\mathrm{S}}(\hat{\mu}^t_{\mathrm{bs}}) = E(\mathrm{div}(\hat{\mu}^t_{\mathrm{bs}})(Y)) = E(\mathrm{rank}(X_{\mathrm{supp}(\hat{\beta}^t)}))$, i.e., the expected size of the selected model.

Example 5.4 (Singular value decomposition with hard threshold on the singular values)**.** Let $\mathbf{Y}$ be a $p \times q$-dimensional matrix with $p \geq q$ and $n \coloneqq pq$. Assuming that

$$\mathrm{vec}(\mathbf{Y}) \sim \mathcal{N}(\mathrm{vec}(\mu), \sigma^2 I_n),$$

for some $p \times q$ matrix μ, this framework still fits in the above setting by the vectorisation-operator, vec. The singular value decomposition (SVD) of $\mathbf{Y}$ is given by

$$\mathbf{Y} = \sum_{k=1}^{q} d_k u_k v_k^T$$

with ordered singular values $d_1 \geq d_2 \geq ... \geq d_q \geq 0$. With probability one the singular values are distinct. The *hard threshold SVD estimator* of μ with tuning parameter $t \in \mathbb{R}$ is given by

$$\hat{\mu}^t_{\mathrm{h.svd}} \coloneqq \sum_{k=1}^{q} 1_{d_k \geq e^t} d_k u_k v_k^T,$$

which only selects components whose singular value exceeds the threshold e^t. We will argue that the *hard threshold SVD estimator* satisfy (20) through Lemma 4.8 and Corollary 4.7.

Firstly, we realise that the selection events are:

$$\begin{aligned} U^t_0 &\coloneqq \{\mathbf{Y} \in \mathbb{R}^n \mid d_1 < e^t\} \\ U^t_r &\coloneqq \{\mathbf{Y} \in \mathbb{R}^n \mid d_{r+1} < e^t < d_r\}, \quad r = 1, ..., q-1, \end{aligned} \tag{41}$$

and the post-selection estimators are the *reduced rank estimators*:

$$\hat{\mu}_r = \sum_{k=1}^{r} d_k u_k v_k^T, \qquad r = 0, ..., q-1.$$

The divergence of the reduced rank estimator is:

$$\mathrm{div}(\hat{\mu}_r) = pr + \sum_{i=1}^{r}\sum_{j=r+1}^{q} \frac{d_i^2 + d_j^2}{d_i^2 - d_j^2},$$

which is easily derived from e.g. formula (9) in Candès et al. (2013). We see that the divergence of $\hat{\mu}_r$ is almost surely non-negative.

Let $p(\mathbf{Y}, \lambda) = \det(\mathbf{Y}^T\mathbf{Y} - \lambda I)$ denote the characteristic polynomial of $\mathbf{Y}^T\mathbf{Y}$, as a function of both the matrix $\mathbf{Y}$ and $\lambda \in \mathbb{R}$. Then λ is an eigenvalue of $\mathbf{Y}^T\mathbf{Y}$ if and only if $p(\mathbf{Y}, \lambda) = 0$. Since the squared singular values are the eigenvalues of $\mathbf{Y}^T\mathbf{Y}$ we can conclude two things: 1) $U^t_i = e^t U^0_i$ for $i = 0, ..., q-1$ and $t \in \mathbb{R}$. 2)

$$U_i \coloneqq U^0_i = \{\mathbf{Y} \in \mathbb{R}^n \mid p(\mathbf{Y}, d^2_{i+1}) = 0 = p(\mathbf{Y}, d^2_i),\ d^2_{i+1} < 1 < d^2_i\}$$

is semi-algebraic for each $i = 1, ..., q-1$ and a similar argument applies to $U_0 \coloneqq U_0^0$. Hence $\hat{\mu}^t_{\text{h.svd}} = \sum_{i=0}^{q-1} 1_{F(t,U_i)} \hat{\mu}_i$, where $F(t, y) = y$.

We have therefore verified all conditions in Lemma 4.8, except for (d). For that we first derive the normal vectors. Let $\mathbf{Y} \in F(t, \overline{U}_i) \cap F(t, \overline{U}_j)$ with $i < j$ and $t \in \mathbb{R}$ fixed. Then

$$d_{j+1} \leq e^t \leq d_j \leq d_{i+1} \leq e^t \leq d_i,$$

thus $d_k = e^t$ for all k with $j - 1 \leq k \leq j$. Hence $F(t, \overline{U}_i) \cap F(t, \overline{U}_j)$ is at most $n-(j-i)$-dimensional. We are therefore only concerned with $F(t, \overline{U}_{j-1}) \cap F(t, \overline{U}_j)$ for $j = 1, ..., q-1$. Moreover, by the same dimensionality argument we can assure that for $\mathcal{H}^{n-1}$-almost all $\mathbf{Y}$ in $F(t, \overline{U}_{j-1}) \cap F(t, \overline{U}_j)$

$$\begin{cases} d_2 < d_1 = e^t & \text{if } j = 1 \\ d_{j+1} < d_j = e^t < d_{j-1} & \text{if } j = 2, ..., q-2 \\ d_{q-1} = e^t < d_{q-2} & \text{if } j = q-1 \end{cases}$$

Let $\mathbf{Y}$ be such a point and let $u_j v_j^T$ be the j^{th} component of $\mathbf{Y}$. By continuity of the components, we can find an $\varepsilon > 0$ such that $\mathbf{Y} + \varepsilon \mathbf{Z} \in F(t, \overline{U}_{j-1}) \cap F(t, \overline{U}_j)$ for all $\text{vec}(\mathbf{Z}) \perp \text{vec}(u_j v_j^T)$. This implies that $\eta_j = \text{vec}(u_j v_j^T)$ serves as a normal vector of $F(t, \overline{U}_{j-1}) \cap F(t, \overline{U}_j)$ at $\mathbf{Y}$. In particular, the inner product $\langle \eta_j; f(y) \rangle = \langle \eta_j; \mathbf{Y} \rangle = d_j = e^t$ is positive. We therefore have $\Phi \setminus \bigcup_i \partial U_i$ is $\mathcal{H}^{n-1}$-null and the jump sizes are:

$$h_{j,j-1}(\mathbf{Y}) = \frac{\langle u_j v_j^T; \hat{\mu}_j(\mathbf{Y}) - \hat{\mu}_{j-1}(\mathbf{Y}) \rangle}{\langle u_j v_j^T; \mathbf{Y} \rangle} = \frac{\langle u_j v_j^T; d_j u_j v_j^T \rangle}{\langle u_j v_j^T; \mathbf{Y} \rangle} = 1 \tag{42}$$

which trivially satisfy (d). We conclude that (20) holds by Lemma 4.8 and Corollary 4.7. From (42) we see that H simply equals (up to an additive constant) the number of singular values smaller than e^t.

6. Discussion

We have provided a method for estimating $\text{df} - \text{df}_S$ for a class of estimators involving data adaptive model selection. For these estimators we have extended the classic divergence operator approach for estimating the degrees of freedom and thereby obtained the adjustments needed for applying SURE to certain estimators with model selection. The method relies on the function H with the property that, under certain regularity conditions, $\partial_t E(H(t, Y))$ equals some or all of the boundary integrals representing $\text{df} - \text{df}_S$ in Theorem 4.3. This formulation of $\text{df} - \text{df}_S$ through $\partial_t E(H(t, Y))$ expresses the duality between changing the tuning parameter t for fixed observation y and changing y for fixed t.

Four examples of estimators with data adaptive model selection were considered. Three of which had the identity $\text{df} - \text{df}_S = \partial_t E(H(t, Y))$. For the remaining example, best subset selection, it was still possible to derive an approximation which is of higher order than that in Section 5 of Mikkelsen & Hansen (2017), which only covers nested model. The new approximation also covers non-nested models with different dimensions.

For all of the examples considered in Section 5 the tuning parameter t in H coincides with the tuning parameter already present in the selection procedure (through the reparametrisation $t = \log \lambda$). This is not a requirement for applying Theorem 4.6 and Corollary 4.7; one can always fix the originial tuning parameter

of the estimator (if any) and perturb the estimator by some flow. As long as the conditions of Theorem 4.6 and Corollary 4.7 apply, this approach is valid. However, if the tuning parameter of the estimator coincides with the time parameter for some flow, then one can apply the results for all tuning parameter values collectively.

The degrees of freedom estimator, and associated risk estimator, derived in Mikkelsen & Hansen (2017) for the lasso-OLS estimator was proven quite fruitful in practice. It is still ongoing research to investigate if the extended degrees of freedom and associated risk estimator presented in this paper perform just as well. Simulation studies similar to those by Mikkelsen & Hansen (2017) need to be employed to compare the risk estimator to other risk estimators, such as those deriving from cross validation or bootstrap based methods.

Acknowledgement

Frederik Vissing Mikkelsen would like to thank Prof. Jonathan Taylor for numerous insightful discussions during the initial development process that led to the above results.

Appendix A. Proofs and Additional Results

A.1. **Proof of Theorem 4.3.** The following lemma characterises the outer unit normal vectors η_i for $i = 1, 2, \ldots$ and the proof is almost identical to its finite equivalent (Lemma A.2 by Mikkelsen & Hansen (2017)):

Lemma A.1. *Under Assumption 4.1 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* $i = 1, 2, \ldots$.
(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 η_i on ∂U_i vanishes outside *the measure theoretic boundary* $\partial_* U_i$, see Definition 5.8 in Evans & Gariepy (1992). Moreover, these two types of boundaries relates to *the reduced boundary* $\partial^* U_i$ (see Definition 5.7 in Evans & Gariepy (1992)) 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 Evans & Gariepy (1992)). All in all, we see that the Lemma holds if we can show the following claims:

$$\begin{aligned} \partial^* U_i &\subseteq \bigcup_{l \neq i} \overline{U}_l \\ \eta_i &= -\eta_j \text{ on } \partial^* U_i \cap \partial^* U_j \\ \partial^* U_i \cap \partial^* U_j \cap \partial^* U_k &= \emptyset \end{aligned} \tag{43}$$

holds for all i, j, k distinct.

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

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

Note that $\{U_i^r(x)\}_i$ are still disjoint. By Theorem 5.7.1 in Evans & Gariepy (1992)

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

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

$$1_{U_i^r(x) \cup U_j^r(x) \cup U_k^r(x)} \xrightarrow{r \to 0} 1_{H_i(x)} + 1_{H_j(x)} + 1_{H_k(x)} \text{ in } L^1_{\text{loc}}(\mathbb{R}^n), \tag{44}$$

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 $\eta_i = 0$ on $\partial U_i \cup \text{int}(\overline{U}_i)$. The first claim thus follows directly from the relative boundary assumption. □

We now present the proof of Theorem 4.3 in the extended framework of a countable set of selection events:

Proof of Theorem 4.3. For $i = 1, 2, \ldots$ Gauss-Green's formula (see Theorem 5.8.1 in Evans & Gariepy (1992) and Theorem 4.5.6 in Federer (1969)) gives that

$$\int_{U_i} \text{div}(f) \, d\mathcal{L}^n = \int_{\partial U_i} \langle f, \eta_i \rangle \, d\mathcal{H}^{n-1} \tag{45}$$

for all Lipschitz continuous vector fields f with compact support. Here η_i denotes the outer unit normal of ∂U_i, which is well defined and nonzero on a subset of ∂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 (45) applied to f_r yields

$$\int_{\partial U_i} g_r \psi \langle \hat{\mu}_i, \eta_i \rangle \, d\mathcal{H}^{n-1} = \int_{U_i} g_r \psi \mathrm{div}(\hat{\mu}_i) \, d\mathcal{L}^n + \int_{U_i} \langle g_r D\psi + \psi D g_r, \hat{\mu}_i \rangle \, d\mathcal{L}^n. \tag{46}$$

Due to Assumption 4.1 all integrands above are either dominated by integrable functions or monotonely increasing in r. By letting $r \to \infty$ Lebesgue's Dominated Convergence Theorem and the Monotone Convergence Theorem yields

$$\int_{\partial U_i} \psi \langle \hat{\mu}_i, \eta_i \rangle \, d\mathcal{H}^{n-1} = \int_{U_i} \psi \mathrm{div}(\hat{\mu}_i) \, d\mathcal{L}^n + \int_{U_i} \langle D\psi, \hat{\mu}_i \rangle \, d\mathcal{L}^n. \tag{47}$$

When summing over i, both the left-most and right-most sums are absolutely convergent by assumption. The middle term is either absolutely convergent or a sum of positive terms. Either way, all terms are absolutely convergent. Thus Lebesgues Dominated Convergence Theorem shows that

$$\mathrm{df}(\hat{\mu}) = \mathrm{df_S}(\hat{\mu}) - \sum_{i=1}^{\infty} \int_{\partial U_i} \psi \langle \hat{\mu}_i, \eta_i \rangle d\mathcal{H}^{n-1}, \tag{48}$$

with all terms finite. Using the absolute convergence we see from Lemma A.1 that

$$\begin{aligned} \mathrm{df}(\hat{\mu}) &= \mathrm{df_S}(\hat{\mu}) - \sum_{i=1}^{\infty} \sum_{j:j\neq i} \int_{\partial U_i \cap \partial U_j} \psi \langle \hat{\mu}_i, \eta_i \rangle d\mathcal{H}^{n-1} \\ &= \mathrm{df_S}(\hat{\mu}) + \frac{1}{2} \sum_{j \neq i} \int_{\partial U_i \cap \partial U_j} \langle \hat{\mu}_j - \hat{\mu}_i, \eta_i \rangle \psi d\mathcal{H}^{n-1}. \end{aligned} \tag{49}$$

Since η_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 (14). □

A.2. **Proof of Theorem 4.6 and Lemma 4.8.** Before proving Theorem 4.6 we first present and prove the following result regarding derivatives of integrals of functions over parametrised integration domains:

Theorem A.2. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and $F : I \times \mathbb{R}^n \to \mathbb{R}^n$ be C^2, where $I \subseteq \mathbb{R}$ open and $F_\lambda := F(\lambda; \cdot)$ is a diffeomorphism for each $\lambda \in I$. Let $U \subseteq \mathbb{R}^n$ be open with locally finite perimeter. For each $\lambda \in I$ define the function $f_\lambda : \mathbb{R}^n \to \mathbb{R}^n$ by*

$$f_\lambda(y) = \partial_\lambda F|_{(\lambda, F_\lambda^{-1}(y))}, \quad y \in \mathbb{R}^n. \tag{50}$$

If for each $\lambda_0 \in I$

1. *$h f_{\lambda_0}$ is Lebesgue integrable on $F_{\lambda_0}(U)$.*
2. *$h\langle f_{\lambda_0}; \eta \rangle$ is $\mathcal{H}^{n-1}$-integrable on $\partial F_{\lambda_0}(U)$, where η is the unit outer normal on $\partial F_{\lambda_0}(U)$.*

(3) There exists a neighbourhood $N \subseteq I$ of λ_0 and $g : \mathbb{R}^n \to \mathbb{R}$ Lebesgue integrable such that

$$1_{F_\lambda(U)}|\mathrm{div}(hf_\lambda)| \leq g, \quad \textit{for all } \lambda \in N$$

then

$$\partial_\lambda \int_{F_\lambda(U)} h(y)\ dy = \int_{\partial F_\lambda(U)} h \cdot \langle \eta; f_\lambda \rangle\ d\mathcal{H}^{n-1}, \quad \textit{for each } \lambda \in I. \tag{51}$$

Proof. Fix $\lambda_0 \in I$ and let N be the associated neighbourhood. By the change of variable formula we have

$$\int_{F_{\lambda_0}(U)} h(y)\ dy = \int_U h(F_{\lambda_0}(z)) \cdot JF_{\lambda_0}(z)\ dz \tag{52}$$

where $JF_\lambda = |\det(DF_\lambda)|$ denotes the Jacobian of the change-of-variable mapping. We now study the derivative of the integrand with respect to λ in the right hand side of (52):

(53)
$$\begin{aligned}
\partial_\lambda \left((h \circ F_\lambda) \cdot JF_\lambda\right) &= (h \circ F_\lambda) \cdot \partial_\lambda JF_\lambda + \langle Dh \circ F_\lambda; \partial_\lambda F_\lambda \rangle \cdot JF_\lambda \\
&= (h \circ F_\lambda) \cdot JF_\lambda \cdot \mathrm{tr}\left((DF_\lambda)^{-1} \partial_\lambda DF_\lambda\right) + \langle Dh \circ F_\lambda; \partial_\lambda F_\lambda \rangle \cdot JF_\lambda \\
&= [(h \cdot \mathrm{div}(f_\lambda) + \langle Dh; f_\lambda \rangle) \circ F_\lambda] \cdot JF_\lambda \\
&= [\mathrm{div}(hf_\lambda) \circ F_\lambda] \cdot JF_\lambda
\end{aligned}$$

Hence

$$\int_U \left|\partial_\lambda \left((h \circ F_\lambda) \cdot JF_\lambda\right)\right|(z)\ dz\Big|_{\lambda=\lambda_0} = \int_{F_{\lambda_0}(U)} |\mathrm{div}(hf_{\lambda_0})(y)|\ dy.$$

By assumption the integrand of the right hand side is bounded over N by an integrable function and we can interchange differentiation and integration of the right hand side in (52):

$$\partial_\lambda \int_{F_\lambda(U)} h(y)\ dy\Big|_{\lambda=\lambda_0} = \int_{F_{\lambda_0}(U)} \mathrm{div}(hf_{\lambda_0})(y)\ dy. \tag{54}$$

Carrying out the above argument for all $\lambda \in I$ we get

$$\partial_\lambda \int_{F_\lambda(U)} h(y)\ dy = \int_{F_\lambda(U)} \mathrm{div}(hf_\lambda)(y)\ dy. \tag{55}$$

For the final step, we fix $\lambda \in I$ and use Gauss-Green's formula (see Theorem 5.8.1 in Evans & Gariepy (1992) and Theorem 4.5.6 in Federer (1969)), which states

$$\int_U \mathrm{div}(\varphi)\ d\mathcal{L}^n = \int_{\partial U} \langle \varphi; \eta \rangle\ d\mathcal{H}^{n-1} \tag{56}$$

for all Lipschitz continuous vector fields φ with compact support and η denote the unit outer normal on ∂U and m the Lebesgue measure. Let $(e_r)_r$ be a sequence of smooth functions with

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

and $(e_r)_r$ and $(De_r)_r$ uniformly bounded. Then $\varphi_r = e_r hf_\lambda$ is Lipschitz continuous with compact support. Then (56) applied to φ_r and $F_\lambda(U)$ (which also has locally

finite perimeter) yields

$$\int_{F_\lambda(U)} e_r \mathrm{div}(hf_\lambda) + \langle De_r; hf_\lambda\rangle \, d\mathcal{L}^n = \int_{\partial F_\lambda(U)} e_r h\langle f_\lambda, \eta\rangle \, d\mathcal{H}^{n-1}.$$

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

$$\int_{F_\lambda(U)} \mathrm{div}(hf_\lambda) \, d\mathcal{L}^n = \int_{\partial F_\lambda(U)} h\langle f_\lambda, \eta\rangle \, d\mathcal{H}^{n-1}$$

which combined with (55) ends the proof. □

Using this intermediate result we prove Theorem 4.6:

Proof of Theorem 4.6. Firstly, we consider the field $y \mapsto \partial_t F(0, y)$, the normal vectors and the set Φ. If $y \in \partial F(t, U_i)$ for some $i = 1, ..., N$ and $t \in \mathbb{R}$, then $y = F(t, z)$ for some unique $z \in \partial U_i$, $(z = F(-t, y))$. Let $\eta_i(y)$ denote a normal vector of $\partial F(t, U_i)$ in y and let $\eta_i(z)$ denote a normal vector of ∂U_i in z. An application of the chain rule shows that $\eta_i(y)$ equals $\partial_y F(-t, y)^T \eta_i(z)$ (up to a scalar). Additionally, by differentiating the identity $F(-t, F(t, y)) = y$ with respect to t we obtain $\partial_y F(-t, y)\partial_t F(t, F(-t, y)) = \partial_t F(-t, y)$ for all $(t, y) \in \mathbb{R} \times \mathbb{R}^n$. Hence,

(57)
$$\langle \eta_i(y); f(y)\rangle = \langle \eta_i(y); \partial_t F(t, F(-t, y))\rangle \propto \langle \eta_i(z); \partial_t F(-t, y)\rangle = \langle \eta_i(z); f(z)\rangle.$$

For the two equalities we used the fundamental property of smooth flows that the trajectories, $t \mapsto F(t, y)$, are integral curves of the field, $y \mapsto \partial_t F(0, y)$, see e.g., Proposition 9.7 by Lee (2012), which means $\partial_t F(t, y) = \partial_t F(0, F(t, y))$ holds for all $(t, y) \in \mathbb{R} \times \mathbb{R}^n$.

Besides being important for interpreting (50) when F is a flow, (57) also shows that

$$F(t, \Phi) = \left\{ y \in \bigcup_i F(t, \partial U_i) \;\middle|\; \langle \eta(y); f(y)\rangle \neq 0 \right\}, \tag{58}$$

which illustrates the purpose of Φ: for a given $t \in \mathbb{R}$ the set $F(t, \Phi)$ consists exactly of all points y for which the jump terms of H equal

$$H(t + 0, y) - H(t - 0, y) = \frac{\langle \eta(y); \hat{\mu}^{t+0}(y) - \hat{\mu}^{t-0}(y)\rangle}{\langle \eta(y); f(y)\rangle}. \tag{59}$$

Outside of $F(t, \Phi)$ the difference in H is zero by convention. Let

$$\Psi := \left\{ y \in \mathbb{R}^n \;\middle|\; F(t, y) \in \Phi, \text{ whenever } F(t, y) \in \bigcup_i \partial U_i \right\},$$

i.e., points y for which no jump term in H have zero denominator.

For $z \in \mathbb{R}^n$ define the function

$$\iota(z) := \sum_{i=1}^{N} i 1_{U_i}(z), \tag{60}$$

which indicates the selection event, i.e., $y \in F(t, U_i)$ if and only if $\iota(F(-t, y)) = i$. None of $i = 1, ..., N$ are selected if $y \in \bigcup_i \partial F(t, U_i)$, which holds if and only if

$\iota(F(-t,y))=0$. For $y\in\Psi$ set $\tau_0(y)\coloneqq 0$ and $a_0^+(y)=\iota(y)$ and define recursively:

$$\begin{aligned}
\tau_k^+(y) &\coloneqq \inf\{t>\tau_{k-1}^+(y)\mid \iota(F(-t,y))\neq a_{k-1}^+(y)\}\\
a_k^+(y) &\coloneqq \begin{cases}\lim_{s\searrow\tau_k^+(y)}\iota(F(-s,y)) & \text{if } \tau_k^+(y)<+\infty\\ a_{k-1}^+(y) & \text{else}\end{cases}\\
\tau_k^-(y) &\coloneqq \sup\{t<\tau_{k-1}^-(y)\mid \iota(F(-t,y))\neq a_{k-1}^-(y)\}\\
a_k^-(y) &\coloneqq \begin{cases}\lim_{s\nearrow\tau_k^-(y)}\iota(F(-s,y)) & \text{if } \tau_k^-(y)>-\infty\\ a_{k-1}^-(y) & \text{else}\end{cases}
\end{aligned}\qquad k=1,2,... \tag{61}$$

The above limits are well defined since the field of the flow is not tangent with the boundaries. Starting at $t=0$ the sequences $(\tau_k^\pm(y))_{k=1}^{k=\infty}$ denotes the jump times of $t\mapsto\iota(F(-t,y))$, which can also be interpreted as the time of crossing of y across the boundaries $\{\partial F(t,U_i)\}_i$. The two neighbouring regions that y cross at $\tau_k^+(y)$, are the left and right limits, $\lim_{t\to\tau_k^+(y)}\iota(F(-t,y))$, or $a_{k-1}^+(y)$ and $a_k^+(y)$ (reverse order for τ_k^-).

Consider the set of double-ended sequences in $\{1,...,N\}$, $\{1,...,N\}^{\mathbb{Z}\setminus\{0\}}$ and let $\alpha:\Psi\to\{1,...,N\}^{\mathbb{Z}\setminus\{0\}}$ denote the function $y\mapsto((a_k^-)_\infty^{k=1},(a_k^+)_{k=1}^\infty)$. On $\{1,...,N\}^{\mathbb{Z}\setminus\{0\}}$ define the equivalence relation given by: $a\sim b$ if and only if $s_l(a)=b$ for some $l\in\mathbb{Z}$, where s_l is the shift operator. Let Δ denote the set of equivalence classes. For each $a\in\Delta$ define the sets

$$V_a\coloneqq\{y\in\Psi\mid\exists t\in\mathbb{R}:\alpha(F(t,y))\sim a\}\,. \tag{62}$$

These are again equivalence classes on Ψ given by the relation: $x\sim y$ if and only if the trajectories of x and y traverse the same sequence of selection events up to a time shift. To see this, notice that $\alpha(y)\sim\alpha(F(t,y))$ for all $t\in\mathbb{R}$ and $y\in\Psi$. Consequently, $(V_a)_{a\in\Delta}$ are *invariant* with respect to the flow:

$$F(t,V_a)=V_a,\qquad\text{for all } t\in\mathbb{R}, \tag{63}$$

for each $a\in\Delta$. By Assumption 4.4(c) there is a countable collection of equivalence classes, $(V_a)_{a\in A}$, covering almost all of Ψ, i.e., $\mathcal{L}^n(\Psi\setminus\bigcup_{a\in A}V_a)=0$ where $A\subseteq\Delta$ countable.

Next, fix $a\in A$. The flow passes through finitely many of the boundaries ∂U_{a_k} indexed by a. By Theorem 5.7.2 in Evans & Gariepy (1992) each boundary can be decomposed in to

$$\partial U_i=\bigcup_{j=1}^{\infty}K_j^i\cup N_i \tag{64}$$

where K_j^i is a compact subset of a C^1 hypersurface and a N_i is a $\mathcal{H}^{n-1}$ null set. Since the flow only passes through a finite number of boundaries, there are at most countably many possible sequences of C^1 hypersurface to pass through, $(K_{j_k}^{a_k})_k$. Hence by further refining the cover $(V_a)_{a\in A}$, with $A\subseteq\Delta\times\mathbb{N}$ countable, we can assure that $\mathcal{L}^n(\Psi\setminus\bigcup_{a\in A}V_a)=0$ still holds and for each $a\in A$ the trajectories $F(\mathbb{R},x)$ pass through the same C^1 boundary-segments for all $x\in V_a$. Since the refinement of the cover $(V_a)_{a\in A}$ did not mix the different trajectories, the invariance (63) still holds.

Next, we wish to show that if $y \in \partial F(t, U_i)$ for some $t \in \mathbb{R}$ and $y \in \Psi$, then there exists a map $y \mapsto t(y)$ defined in some neighbourhood in Ψ of y, N_y, such that

$$w \in \partial F(t(w), U_i), \quad \text{for all } w \in N_y \tag{65}$$

Firstly, for some j and $z \in K_j^i$ we have $y = F(t, z)$ and $K_j^i \subseteq (G = 0)$ for some C^2 function $G : \mathbb{R}^n \to \mathbb{R}$. Hence in some sufficently small neighbourhood of y, (65) is equivalent to $G(F(-t(w), w)) = 0$, for all w in that neighbourhood. The *Implicit Function Theorem*, (Theorem 9.27 by Rudin (1976)), shows that the map $t(\cdot)$ is C^2 if $DG|'_{-t,y}\partial_t F(-t, y) \neq 0$. To show this, note that $F(-t, F(t, z)) = z$ for all $t \in \mathbb{R}$ and $z \in \mathbb{R}^n$, hence

$$\partial_t F(-t, y) = \partial_x F(-t, y)\partial_t F(t, z), \text{ for all } t \in \mathbb{R}, \ y, z \in \mathbb{R}^n \text{ s.t. } y = F(t, z). \tag{66}$$

Since $y \in \Psi$ we therefore get

$$DG|'_{-t,y}\partial_t F(-t, y) = DG|'_{-t,y}\partial_x F(-t, y)\partial_t F(t, z) = \langle \eta(y); \partial_t F(t, F(-t, y))\rangle \neq 0,$$

and the Implicit Function Theorem applies. From (65) and Assumption 4.4(d) we conclude that each term entering H is locally Lipschitz on V_a. Moreover, for $i = 1, ..., N$ and $t \in \mathbb{R}$ fixed, the terms entering H are fixed on $V_a \cap F(t, U_i)$. Consequently, on $V_a \cap F(t, U_i)$ the function H is differentiable and only depends implicitly on t, through the domain $V_a \cap F(t, U_i)$.

Fix $a \in A$ and $i = 1, ..., N$, then Theorem A.2 applied to $h = H\psi$ and $F(t, U) = F(t, V_a \cap U_i) = V_a \cap F(t, U_i)$ gives

$$\partial_t \int_{V_a \cap F(t,U_i)} \psi(y)H(y)\ dy = \int_{\partial(V_a \cap F(t,U_i))} \psi H \cdot \langle \eta; f\rangle\ d\mathcal{H}^{n-1}, \tag{67}$$

since $f_t(y) = \partial_t F|_{t,F^{-1}(y)} = \partial_t F|_{t,F(-t,y)} = \partial_t F(0, y) = f(y)$. All the integrability conditions in Theorem A.2 follow directly from Assumption 4.4(b)+(e) and Theorem 4.3.

Consider the integration domain of the right hand side of (67); $\partial(V_a \cap F(t, U_i)) \subseteq \partial V_a \cup \partial F(t, U_i)$. For a point $y \in \partial V_a \setminus \partial F(t, U_i)$ we have $\langle \eta(y); f(y)\rangle = 0$, since $\eta(y)$ must be orthogonal to the flow (a consequence of the invariance, $F(t, \partial V_a) = \partial V_a$, $t \in \mathbb{R}$). Hence (67) becomes

$$\partial_t \int_{V_a \cap F(t,U_i)} \psi(y)H(y)\ dy = \int_{V_a \cap \partial F(t,U_i)} \psi H \cdot \langle \eta_i; f\rangle\ d\mathcal{H}^{n-1}. \tag{68}$$

Next, we sum over $a \in A$. For that, fix $t_0 \in \mathbb{R}$ and let δ and g_1, g_2 be given as in Assumption 4.4(e). Applying Theorem 4.3 to G_t we have

$$\int_{\partial(V_a \cap F(t,U_i))} \psi H \cdot \langle \eta; f\rangle\ d\mathcal{H}^{n-1} = \int_{V_a \cap F(t,U_i)} \psi(y) \left\langle G_t(y); \frac{y - \mu}{\sigma^2} \right\rangle\ dy - \int_{V_a \cap F(t,U_i)} \psi(y)\mathrm{div}(G_t)(y)\ dy.$$

Hence

$$\left|\partial_t \int_{V_a \cap F(t,U_i)} \psi(y)H(y)\ dy\right| \leq \int_{V_a \cap F(t,U_i)} \psi(y)g_1(y)\ dy + \int_{V_a \cap F(t,U_i)} \psi(y)g_2(y)\ dy.$$

for all $t \in [t_0 - \delta; t_0 + \delta]$. By the Weierstrass M-test and Lebesgues Dominated Convergence Theorem we conclude that

$$\partial_t \int_{F(t,U_i)} \psi(y) H(t,y)\, dy = \int_{\partial F(t,U_i)} \psi H \cdot \langle \eta_i; f \rangle \, d\mathcal{H}^{n-1}. \tag{69}$$

Finally, we sum over $i = 1, ..., N$ and (69) becomes

(70)

$$\begin{aligned}\partial_t E(H(t,Y)) &= \sum_i \int_{\partial F(t,U_i)} \psi(y) H(t,y) \langle \eta_i; f(y) \rangle \, d\mathcal{H}^{n-1}(y) \\ &= \frac{1}{2} \sum_{i \neq j} \int_{\overline{F(t,U_i)} \cap \overline{F(t,U_j)} \cap F(t,\Phi)} \psi(y) \langle \eta_i(y); \hat{\mu}_j(y) - \hat{\mu}_i(y) \rangle \, d\mathcal{H}^{n-1}(y).\end{aligned}$$

In the above we used that for $\mathcal{H}^{n-1}$-almost all $y \in \overline{F(t,U_i)} \cap \overline{F(t,U_j)}$, $\eta_i = -\eta_j$ (Lemma A.1) and that the difference in H across the neighbouring regions $F(t,U_i)$ and $F(t,U_j)$, given in (59) is exactly the term accounting for this crossing, i.e., $\frac{\langle \eta_i(y); \hat{\mu}_j(y) - \hat{\mu}_i(y) \rangle}{\langle \eta_i(y); f(y) \rangle}$ if $y \in F(t,\Phi)$, and 0 if $y \notin F(t,\Phi)$. □

Finally, we present the proof of Lemma 4.8:

Proof of Lemma 4.8. Firstly, if the number of crossings is bounded it is also finite for almost all y. Secondly, the boundaries are $\mathcal{H}^{n-1}$ almost everywhere C^2, since they are semi-algebraic. This shows Assumption 4.4(c)+(d). Next we show that $\hat{\mu}(t,\cdot)$ satisfies Assumption 4.1 for all t. Condition 4.1(a) holds by Remark 4.2(a) since we only consider finitely many selection events. Condition 4.1(b) follows from the polynomial bound provided by Lemma A.3. Since both $\hat{\mu}(Y)$ and $\operatorname{div}(\hat{\mu})(Y)$ are finite sums of stochastic variables with finite first moment (or, alternatively, in the latter case non-negative) Assumption 4.1(c) is clear. Finally, Assumption 4.1(d) is a direct application of Lemma A.4, which applies because of the polynomial bound provided by Lemma A.3.

Next, we show Assumption 4.4(e) and that G_t fulfills Assumption 4.1 for all t. Reconstruct the refined cover of $\mathbb{R}^n$, $(V_a)_a$, from the proof of Theorem 4.6. Let $W_{a,i} = U_i \cap V_a$ and $\tilde{\mu}_{a,i}(t,y) = H(t,y) f(y)$, then

$$G_t = \sum_{a,i} 1_{F(t,W_{a,i})} \tilde{\mu}_{a,i}.$$

We will prove that this representation of G_t satisfies Assumption 4.1. On each $F(t,W_{a,i})$ the finite number of terms added in H are fixed and among $((h_{ij})_{j \neq i})_i^N$. Each of which are locally Lipschitz, thus $\tilde{\mu}_{a,i}$ is also locally Lipschitz. Furthermore, by assumption both $\tilde{\mu}_{a,i}$ and $\operatorname{div}(\tilde{\mu}_{a,i})$ are polynomially bounded.

The selection events $(W_{a,i})_{a,i}$ are again disjoint (since $(U_i)_i$ are disjoint) and their closure cover $\mathbb{R}^n$. Each boundary ∂U_i can be divided into a finite number of smooth manifolds, since U_i is semi-algebraic. That combined with the bounded number of crossings of $(\partial U_i)_i$ shows that the cover $(V_a)_a$ is finite. Thus $(W_{a,i})_{a,i}$ is also finite and both G_t and $\operatorname{div}(G_t)$ are polynomially bounded, which imply Assumption 4.4(e). Consequently, Assumption 4.1(c) also holds for G_t. Furthermore, Assumption 4.1(a) also holds for G_t by Remark 4.2(a).

For Assumption 4.1(b), note that $\partial F(t,W_{a,i}) = \partial(V_a \cap F(t,U_i)) \subseteq \partial V_a \cap F(t,U_i) \cup V_a \cap \partial F(t,U_i)$. As pointed out in the proof of Theorem 4.6 $\langle \eta_i; \tilde{\mu}_{a,i} \rangle =$

$H(t,y)\langle \eta_{a,i}; f(y)\rangle = 0$ on $\partial V_a \cap F(t,U_i)$. This and Lemma A.3 yields

$$\mathcal{H}^{n-1}\left(\partial F(t,W_{a,i}) \cap (\langle \eta_i; \tilde{\mu}_{a,i}\rangle \neq 0) \cap B(0,r)\right) \leq \mathcal{H}^{n-1}\left(\partial F(t,U_i)\cap B(0,r)\right) \leq p(r) \tag{71}$$

for some polynomial p. Therefore, in light of Remark 4.2(b), $(W_{a,i})_{a,i}$ satisfies Assumption 4.1(b). Finally, Lemma A.4 applies and yields Assumption 4.1(d) for G_t. This concludes the proof. □

A.3. **Additional Lemmas.** Concerning condition (b) in Assumption 4.1 the following observation is useful: for A and B subsets of $\mathbb{R}^n$ and F a flow it holds that

$$\begin{aligned} \partial F(t,A) &= \partial(F(t,A)^c), \\ \partial F(t,A\cup B) &\subseteq \partial F(t,A)\cup \partial F(t,B), \\ \partial F(t,A\cap B) &\subseteq \partial F(t,A)\cup \partial F(t,B). \end{aligned} \tag{72}$$

Especially, the family of sets

$$\begin{aligned} &\left\{ E\in\mathcal{B}(\mathbb{R}^n) \,\middle|\, \mathcal{H}^{n-1}\left(\partial F(t,E)\cap B(0,r)\right) < +\infty, \quad \text{for all } r>0 \right\} \\ &\left\{ E\in\mathcal{B}(\mathbb{R}^n) \,\middle|\, \begin{array}{c} \text{there exists a polynomial } p \text{ s.t.} \\ \mathcal{H}^{n-1}\left(\partial F(t,E)\cap B(0,r)\right) \leq p(r), \quad \text{for all } r>0 \end{array} \right\} \end{aligned} \tag{73}$$

are all stable under complement, finite union and finite intersection. Here $\mathcal{B}(\mathbb{R}^n)$ denote the Borel σ-algebra on $\mathbb{R}^n$.

Below we present a lemma for verifying Assumption 4.1(b) for semi-algebraic selection regions. Recall that 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.3. *Let $F:\mathbb{R}^n\to\mathbb{R}^n$ be a C^1-diffeomorphism with $\|F^{-1}\|$ and $\|DF\|$ polynomially bounded. For every semi-algebraic set $U\subseteq\mathbb{R}^n$ there exist a polynomial $p:\mathbb{R}\to\mathbb{R}$ such that*

$$\mathcal{H}^{n-1}\left(\partial F(U)\cap B(0,r)\right) \leq p(r)$$

for all $r>0$.

Proof. By the stability under finite set operations of the family given by (73) it suffices to show that $r\mapsto \mathcal{H}^{n-1}\left(F((P=0))\cap B(0,r)\right)$ is polynomially bounded for any nonzero polynomial P. Let $\pi_j:\mathbb{R}^n\to\mathbb{R}$ denote the projection onto the j^{th} coordinate and let $\pi_{-j}:\mathbb{R}^n\to\mathbb{R}^{n-1}$ denote the projection onto all but the j^{th} coordinate. Fix a non-zero polynomial $P:\mathbb{R}^n\to\mathbb{R}$. There exists a finite number of polynomials $(p_i)_{i=1}^N$ on $\mathbb{R}^{n-1}$ and coordinates $(j_i)_{i=1}^N\in\{1,...,n\}^N$, such that if $P(x)=0$ then for some $i=1,...,N$ we have $\pi_{j_i}(x) = f_i(\pi_{-j_i}x)$. For notational simplicity we assume $j_i=1$ for all i. Defining $g_i:\mathbb{R}^{n-1}\to\mathbb{R}^n$ as $g_i(z)=(f_i(z),z)$, then

$$F((P=0)) \subseteq \bigcup_i \operatorname{im}(F\circ g_i).$$

With $A_r = \pi_{-j_i}(F^{-1}(B(0,r))) \subseteq \mathbb{R}^{n-1}$ the area formula (Theorem 3.3.1 in Evans & Gariepy (1992)) now yields

$$\begin{aligned}\mathcal{H}^{n-1}\left(F((P=0)) \cap B(0,r)\right) &\leq \sum_{i=1}^{N} \int \mathcal{H}^0\left(A_r \cap (F \circ g_i)^{-1}(\{y\})\right)\, d\mathcal{H}^{n-1}(y) \\ &= \sum_{i=1}^{N} \int_{A_r} J(F \circ g_i)(z)\, dz \\ &= \sum_{i=1}^{N} \int_{A_r} (JF)(g_i(z))\sqrt{1+\|Dg_i(z)\|^2}\, dz. \end{aligned} \tag{74}$$

Since $r \mapsto \sup_{z \in A_r} \|z\|$, $\|g_i\|$, $\|Dg_i\|$ and JF are polynomially bounded, we conclude that the left hand side of (74) is polynomially bounded in r. □

Assumption 4.1(d) is likely the most difficult to verify. The following lemma provides a short cut:

Lemma A.4. *Condition (d) in Assumption 4.1 follows from the remaining conditions in Assumption 4.1 if $\|\hat{\mu}_i\|$ and*

$$r \mapsto \mathcal{H}^{n-1}\left(\partial U_i \cap (\langle \eta_i; \hat{\mu}_i \rangle \neq 0) \cap B(0,r)\right), \quad r > 0$$

are polynomially bounded for each $i = 1, 2, \ldots$ and either the collection $(U_i)_i$ is finite or $\sum_i 1_{U_i} \mathrm{div}(\hat{\mu}_i)(Y)$ has finite first moment.

Proof. For some $x \in \partial U_i$ with $\langle \eta_i(x); \hat{\mu}_i(x) \rangle < 0$ we can find $r_x \in (0,1)$ such that $\langle \eta_i(y); \hat{\mu}_i(y) \rangle < 0$ holds for all $y \in \overline{B(x, r_x)} \cap \partial U_i$. Let $A \subseteq \partial U_i \cap (\langle \eta_i; \hat{\mu}_i \rangle > 0)$ denote the set where it is possible. Due to continuity of $\hat{\mu}_i$ and that η_i is $\mathcal{H}^{n-1}$-a.e. continuous (Theorem 5.7.2 in Evans & Gariepy (1992)) we conclude that $\mathcal{H}^{n-1}(\partial U_i \setminus A) = 0$. Next, construct the set

$$N_i := \bigcup_{x \in A} \overline{B(x, r_x)} \cap U_i.$$

Let $R > 0$ be arbitrary. Using the Vitali covering lemma there exists $C_R \subseteq A \cap B(0,R)$ countable such that $N_i \cap B(0,R) \subseteq \bigcup_{x \in C_R} \overline{B(x, 5r_x)}$ and $(B(x, r_x))_{x \in C_R}$ disjoint. If $\mathcal{H}^{n-1}(\partial U_i \cap (\langle \eta_i; \hat{\mu}_i \rangle < 0) \cap B(0,R)) > 0$ then

$$\begin{aligned}\mathcal{H}^{n-1}\left(\partial N_i \cap B(0,R)\right) &\leq \mathcal{H}^{n-1}\left(\bigcup_{x \in C_R} \partial B(x, 5r_x) \cap B(0,R)\right) \\ &\leq \frac{\mathcal{H}^{n-1}\left(\bigcup_{x \in C_R} \partial B(x, 5r_x) \cap B(0,R)\right)}{\mathcal{H}^{n-1}\left(\partial U_i \cap (\langle \eta_i; \hat{\mu}_i \rangle < 0) \cap B(0,R)\right)} \mathcal{H}^{n-1}\left(\partial U_i \cap B(0,R)\right) \\ &\leq \frac{\mathcal{H}^{n-1}\left(\bigcup_{x \in C_R} \partial B(x, r_x) \cap B(0,R)\right)}{\mathcal{H}^{n-1}\left(\bigcup_{x \in C} \partial U_i \cap B(x, r_x) \cap B(0,R)\right)} 5^{n-1} \mathcal{H}^{n-1}\left(\partial U_i \cap B(0,R)\right) \\ &\leq 2\pi\Gamma\left(\frac{n}{2}\right) 5^{n-1} \mathcal{H}^{n-1}\left(\partial U_i \cap B(0,R)\right). \end{aligned} \tag{75}$$

In the second last step we used the following two observations: for $x \in B(0,R)$ the fraction of the sphere $\partial B(x, cr_x)$ which is inside $B(0,R)$ increases when c decreases. Secondly, the surface measure of the full sphere is proportional to c^{n-1}. In the final step of (75) we used the following: if a surface, ∂U_i, passes through a ball such that it intersects the origin, then the surface area contained inside the ball is at least

equal to the surface area of a hyperplane having the same property. Thus the ratio in the second last step is at most equal to the ratio between the surface area of a ball and the surface area of a cross section through the origin, which is $2\pi\Gamma\left(\frac{n}{2}\right)$.

From (75) we conclude that if Assumption 4.1(b) holds for U_i, then so it does for N_i. Consequently, we can replicate the first half of the proof of Theorem 4.3 for N_i right up until (46):

$$\int_{\partial N_i} g_r \psi \langle \hat{\mu}_i, \eta_i \rangle \, d\mathcal{H}^{n-1} = \int_{N_i} g_r \psi \mathrm{div}(\hat{\mu}_i) \, d\mathcal{L}^n + \int_{N_i} \langle g_r D\psi + \psi D g_r, \hat{\mu}_i \rangle \, d\mathcal{L}^n,$$

where $(g_r)_r$ and $(Dg_r)_r$ uniformly bounded and $g_r = 1$ on $B(0,r)$ and vanishes outside $B(0,r+1)$. The right hand side integrands are either bounded by integrable functions or monotonely increasing in R due to Assumption 4.1(c). As for the left hand side, note

$$\int_{\partial N_i} \psi |\langle \hat{\mu}_i, \eta_i \rangle| \, d\mathcal{H}^{n-1} \leq \sum_{r=1}^{\infty} \int_{\partial N_i \cap B(0,r) \setminus B(0,r-1)} |\langle \hat{\mu}_i, \eta_i \rangle| \, d\mathcal{H}^{n-1} e^{-ar}$$

for some $a > 0$. By assumption the integrals in the last sum are bounded by a polynomial only depending on r, and the whole sum is finite. Lebesgues Dominated Convergence Theorem and The Monotone Convergence Theorem yield

$$\int_{\partial N_i} \psi \langle \hat{\mu}_i, \eta_i \rangle \, d\mathcal{H}^{n-1} = \int_{N_i} \psi \mathrm{div}(\hat{\mu}_i) \, d\mathcal{L}^n + \int_{N_i} \langle D\psi, \hat{\mu}_i \rangle \, d\mathcal{L}^n. \tag{76}$$

A similar argument applies to $P_i := U_i \setminus \overline{N}_i$. Using this and Lemma A.1 we get

$$\begin{aligned}
\int_{\partial U_i} \psi |\langle \eta_i; \hat{\mu}_i \rangle| \, d\mathcal{H}^{n-1} &= \int_{\partial P_i \cap \partial U_i} \psi \langle \eta_i; \hat{\mu}_i \rangle \, d\mathcal{H}^{n-1} - \int_{\partial N_i \cap \partial U_i} \psi \langle \eta_i; \hat{\mu}_i \rangle \, d\mathcal{H}^{n-1} \\
&= \int_{\partial P_i} \psi \langle \eta_i; \hat{\mu}_i \rangle \, d\mathcal{H}^{n-1} - \int_{\partial N_i} \psi \langle \eta_i; \hat{\mu}_i \rangle \, d\mathcal{H}^{n-1} \\
&= \int_{P_i} \psi \mathrm{div}(\hat{\mu}_i) \, d\mathcal{L}^n + \int_{P_i} \langle D\psi, \hat{\mu}_i \rangle \, d\mathcal{L}^n \\
&\quad - \int_{N_i} \psi \mathrm{div}(\hat{\mu}_i) \, d\mathcal{L}^n - \int_{N_i} \langle D\psi, \hat{\mu}_i \rangle \, d\mathcal{L}^n \\
&\leq \int_{U_i} \psi |\mathrm{div}(\hat{\mu}_i)| \, d\mathcal{L}^n + \int_{U_i} |\langle D\psi, \hat{\mu}_i \rangle| \, d\mathcal{L}^n
\end{aligned}$$

By summing over i we get

$$\sum_{i=1}^{\infty} \int_{\partial U_i} \psi \, |\langle \eta_i; \hat{\mu}_i \rangle| \, d\mathcal{H}^{n-1} \leq \int \psi |\mathrm{div}(\hat{\mu})| \, d\mathcal{L}^n + \int |\langle D\psi, \hat{\mu} \rangle| \, d\mathcal{L}^n < \infty$$

by Assumption 4.1(c). If $\psi |\mathrm{div}(\hat{\mu})|$ is not Lebesgue integrable by assumption, then $(U_i)_i$ is finite and thus $\psi |\mathrm{div}(\hat{\mu})|$ is a finite sum of terms, each of which are Lebesgue integrable due to (76) and its P_i-equivalent. □

References

Breiman, L. (1992), 'The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error', *Journal of the American Statistical Association* **87**(419), 738–754.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475276*

Candès, E. J., Sing-Long, C. & Trzasko, J. D. (2013), 'Unbiased risk estimates for singular value thresholding and spectral estimators', *IEEE Trans. Signal Processing* **61**(19), 4643–4657.
URL: *https://doi.org/10.1109/TSP.2013.2270464*

Donoho, D. L. & Johnstone, I. M. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *Journal of the American Statistical Association* **90**(432), 1200–1224.
URL: *http://www.jstor.org/stable/2291512*

Efron, B. (2004), 'The estimation of prediction error: Covariance penalties and cross-validation', *Journal of the American Statistical Association* **99**(467), 619–632.
URL: *http://dx.doi.org/10.1198/016214504000000692*

Evans, L. & Gariepy, R. (1992), *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, Taylor & Francis.

Federer, H. (1969), *Geometric measure theory*, Grundlehren der mathematischen Wissenschaften, Springer.

Kato, K. (2009), 'On the degrees of freedom in shrinkage estimation', *Journal of Multivariate Analysis* **100**(7), 1338 – 1352.
URL: *http://www.sciencedirect.com/science/article/pii/S0047259X08002753*

Lee, J. M. (2012), *Integral Curves and Flows*, Springer New York, New York, NY, pp. 205–248.

Meinshausen, N. (2007), 'Relaxed lasso', *Computational Statistics & Data Analysis* **52**(1), 374 – 393.
URL: *http://www.sciencedirect.com/science/article/pii/S0167947306004956*

Meyer, M. & Woodroofe, M. (2000), 'On the degrees of freedom in shape-restricted regression', *Ann. Statist.* **28**(4), 1083–1104.
URL: *http://dx.doi.org/10.1214/aos/1015956708*

Mikkelsen, F. R. & Hansen, N. R. (2017), 'Degrees of freedom for piecewise lipschitz estimators', *Annales de l'Institut Henri Poincaré* .

Rudin, W. (1976), *Principles of Mathematical Analysis*, International Series in Pure and Applied Mathematics, 3 edn, McGraw-Hill Inc.

Stein, C. M. (1981), 'Estimation of the mean of a multivariate normal distribution', *Ann. Statist.* **9**(6), 1135–1151.
URL: *http://dx.doi.org/10.1214/aos/1176345632*

Tibshirani, R. (1994), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B* **58**(1), 267–288.
URL: *http://www.jstor.org/stable/2346178*

Tibshirani, R. J. (2013), 'The lasso problem and uniqueness', *Electron. J. Statist.* **7**, 1456–1490.
URL: *http://dx.doi.org/10.1214/13-EJS815*

Tibshirani, R. J. (2015), 'Degrees of freedom and model search', *Statistica Sinica*

25(3), 1265–1296.

Tibshirani, R. J. & Taylor, J. (2012), 'Degrees of freedom in lasso problems', *Ann. Statist.* **40**(2), 1198–1232.
URL: *http://dx.doi.org/10.1214/12-AOS1003*

Xie, X., Kou, S. & Brown, L. (2012), 'Sure estimates for a heteroscedastic hierarchical model', *Journal of the American Statistical Association* **107**, 1465–1479.

Ye, J. (1998), 'On measuring and correcting the effects of data mining and model selection', *J. Amer. Statist. Assoc.* **93**(441), 120–131.
URL: *http://dx.doi.org/10.2307/2669609*

Zou, H., Hastie, T. & Tibshirani, R. (2007), 'On the degrees of freedom of the lasso', *Ann. Statist.* **35**(5), 2173–2192.
URL: *http://dx.doi.org/10.1214/009053607000000127*

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

E-mail address, Corresponding author: frm@math.ku.dk

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