Biometrics 71, 131–138
March 2015
DOI: 10.1111/biom.12243

On the Selection of Ordinary Differential Equation Models with Application to Predator-Prey Dynamical Models

Xinyu Zhang,[1] Jiguo Cao,[2,*] and Raymond J. Carroll[3,4]

[1]International School of Economics and Management, Capital University of Economics and Business, Beijing 100070, China
[2]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A1S6
[3]Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, U.S.A.
[4]Department of Mathematics and Statistics, University of Technology, P.O. Box 123, Broadway, Sydney, New South Wales 2007, Australia
email: jiguo_cao@sfu.ca

Summary. We consider model selection and estimation in a context where there are competing ordinary differential equation (ODE) models, and all the models are special cases of a "full" model. We propose a computationally inexpensive approach that employs statistical estimation of the full model, followed by a combination of a least squares approximation (LSA) and the adaptive Lasso. We show the resulting method, here called the LSA method, to be an (asymptotically) oracle model selection method. The finite sample performance of the proposed LSA method is investigated with Monte Carlo simulations, in which we examine the percentage of selecting true ODE models, the efficiency of the parameter estimation compared to simply using the full and true models, and coverage probabilities of the estimated confidence intervals for ODE parameters, all of which have satisfactory performances. Our method is also demonstrated by selecting the best predator-prey ODE to model a lynx and hare population dynamical system among some well-known and biologically interpretable ODE models.

Key words: Adaptive LASSO; Least squares approximation; Spline modeling; Variable selection.

1. Introduction

Ordinary differential equation (ODE) models are widely used to describe complicated dynamical systems in ecology and many other scientific areas, because ODEs model the rates of change of the dynamical systems and quantify the underlying mechanisms of the dynamical systems. Typically, ODEs have no analytic solution. The ODE solution is not unique unless we specify the initial condition, which is defined as the value of the dynamical system at the starting point. After the initial condition of the dynamical system is specified, ODEs can be solved using numerical methods such as the Euler method and the Runge–Kutta method (Stoer and Bulirsch, 2002).

Parameters in these ODE models often have scientific interpretations, but their values are usually unknown. Therefore, it is necessary to estimate ODE parameters from measurements of the dynamical systems in the presence of measurement errors. This is a difficult problem to solve, because the numerical solution of ODEs is computationally challenging when analytical solution is not available. Several methods have been proposed to address this problem. For example, the least squares method is often used to find the optimal estimates of ODE parameters by fitting the numerical solution of ODEs to data (Bard, 1974; Biegler, Damiano and Blau, 1986; Williams and Kalogiratou, 1993). Alternatively, a two-step method (Ramsay and Silverman, 2005; Brunel, 2008; Chen and Wu, 2008) does not need to solve ODEs numerically, allowing for very efficient computation. Ramsay et al. (2007) proposed a parameter cascading method, in which the dynamical process is estimated with penalized smoothing splines, with the roughness penalty term defined by ODEs. This method retains satisfactory numerical performance for ODE parameter estimates from finite data samples (Cao, Fussman, and Ramsay, 2008). Qi and Zhao (2010) showed the consistency and asymptotic normality of the ODE parameter estimate using the parameter cascading method. More recently, Hall and Ma (2014) proposed a kernel based one-step estimation method.

In practice, based on different understandings of the dynamical systems, it is often the case that there are multiple competing ODE models that can be used to describe the same dynamical system. For example, as described in more detail in Section 2, there are several ODEs that have been proposed to model predator-prey dynamical systems, an application upon which we focus, although our method is far more broadly applicable. For example, the molecular mechanism of a particular organism can be modeled as a complex metabolic network by using various ODE models based on different understandings of the structure of the metabolic network (Voit and Almeida, 2004). In general, there exist full models that include all the competing ODEs as special cases. Estimators based on full models by using the parameter cascading or the two-stage methods are consistent under regularity conditions. However, the resulting estimators may not be efficient, because some parameters equal, or are very close to, zero.

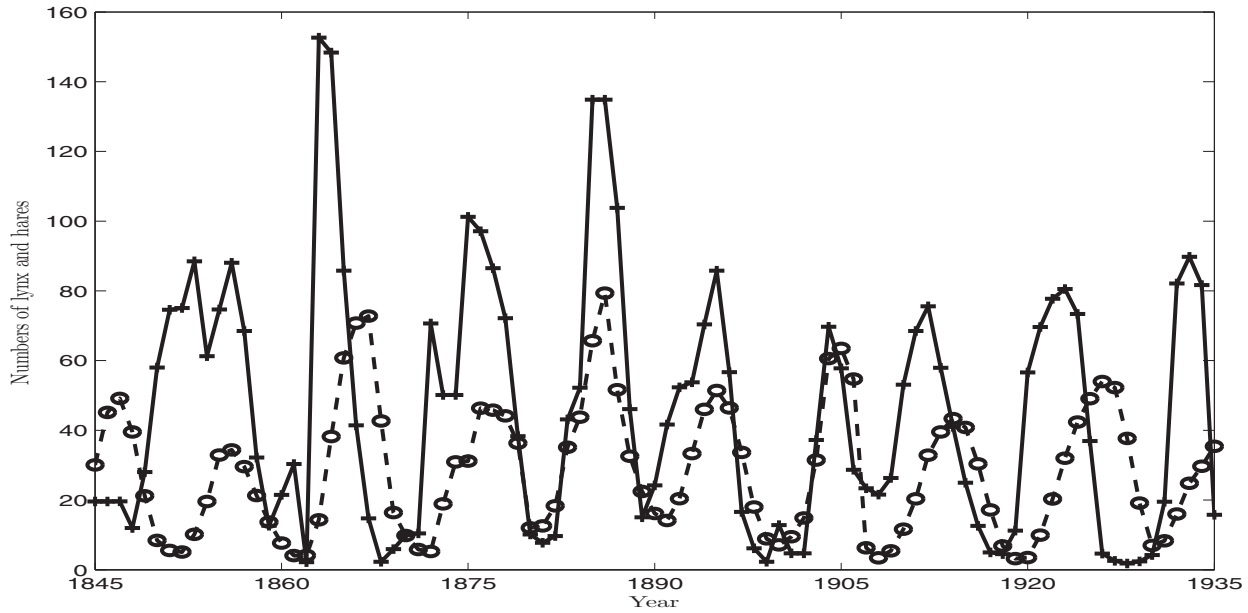

Figure 1. The numbers of Canadian lynx (dashed line with o) and snowshoe hares (solid line with +) from 1845 to 1935 recorded by the Hudson Bay company.

Therefore, our main goal is to develop a computationally inexpensive approach for selecting an appropriate ODE model without estimating all competing ODE models.

Traditional subset selection methods such as the Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978) and corrected AIC (Hurvich and Tsai, 1989) can be used to select an ODE model; see, for example, Miao et al. (2009, 2012). But when we need to examine all possible ODE models, the number of candidate models increases exponentially with the dimensionality of the parameter space. This problem becomes more serious when selecting ODE models, because the computational cost of estimating parameters in each ODE model is much higher in comparison with regular statistical models. Our proposed method is far less computationally costly than these alternatives.

In this article, we extend the least squares approximation (LSA) method (Wang and Leng, 2007) for the selection of ODE models. In LSA, a least squares type function is used to approximate an original loss function, and then an adaptive LASSO (Zou, 2006) type penalty is added to the approximated function. We call the combined functions the LSA criterion. Estimators obtained by minimizing the LSA criterion can identify zero and nonzero parameters consistently, under some regularity conditions. However, the estimators themselves may not have the oracle property (Fan and Li, 2001), because the technical covariance assumption of Wang and Leng (2007) may not be satisfied, as explained in Section 4. Thus, we propose to re-estimate parameters in the ODE model selected by LSA, and then the oracle property can be achieved.

The rest of the article is organized as follows. As a motivation of our ODE model selection problem, in Section 2 we introduce a problem of selecting a predator-prey ODE to model a lynx and hare population dynamical system among some well-known and biologically interpretable ODE model candidates. In Section 3, we present the parameter cascading method for estimating parameters for a given ODE model. Our ODE model selection method by LSA is developed in Section 4. Section 5 applies our method to the lynx and hare example. The finite sample performance of the LSA method is investigated with Monte Carlo simulations in Section 6. Finally, we conclude with some remarks in Section 7.

2. Motivation: Selecting Predator-Prey ODE Models

The Canadian lynx is a type of wild felid, or cat, which is found in northern forests across almost all of Canada and Alaska. Canadian lynx feed predominantly on snowshoe hares. This pair of interacting populations is a classic example of the predator-prey dynamical system. Figure 1 displays the numbers of Canadian lynx and snowshoe hares between 1845 and 1935, recorded by the Hudson Bay company (Odum and Barrett, 2004). It shows the oscillating behavior of both populations.

The population dynamical system of the interacting predator and prey species is popularly modeled by ordinary differential equations. Murdoch, Briggs, and Nisbet (2003) reviewed some competing predator-prey ODE models. Four such models are listed below, in which H and P denote the population sizes of the prey and predator, respectively, and r, a, e, v, g, b, and z are unknown parameters.

- The *Lotka–Volterra model* is

$$\frac{\mathrm{d}H}{\mathrm{d}t} = rH - aHP;$$
$$\frac{dP}{\mathrm{d}t} = eHP - vP. \tag{1}$$

- The *logistic prey model* is

$$\frac{\mathrm{d}H}{\mathrm{d}t} = rH(1 - H/g) - aHP;$$
$$\frac{\mathrm{d}P}{\mathrm{d}t} = eHP - vP. \tag{2}$$

- *The density-dependent predator death model is*

$$\frac{\mathrm{d}H}{\mathrm{d}t} = rH - aHP;$$
$$\frac{\mathrm{d}P}{\mathrm{d}t} = eHP - vP - bP^2. \tag{3}$$

- The *predator-dependent functional response model is*

$$\frac{\mathrm{d}H}{\mathrm{d}t} = rH - aHP/(1 + zP);$$
$$\frac{\mathrm{d}P}{\mathrm{d}t} = eHP/(1 + zP) - vP. \tag{4}$$

All four predator-prey ODE models are well studied and have their own biological interpretations. Our goal is to select the most appropriate ODE model that describes the population dynamical system of Canadian lynx and snowshoe hares based on the data displayed in Figure 1. Although the four predator-prey ODE models have various forms, and they are not nested models, these four ODE models are all special cases of the ODE model

$$\frac{\mathrm{d}H}{\mathrm{d}t} = rH - aHP - kH^2 + az\{P/(1 + zP)\}HP;$$
$$\frac{\mathrm{d}P}{\mathrm{d}t} = eHP - vP - bP^2 - ez\{P/(1 + zP)\}HP. \tag{5}$$

For instance, when $k = z = b = 0$, the ODE model (5) reduces to the Lotka–Volterra model (1); when $z = b = 0$, the ODE model (5) reduces to the logistic prey model (2); when $k = z = 0$, the ODE model (5) reduces to the density-dependent predator death model (3); and when $k = b = 0$, the ODE model (5) reduces to the predator-dependent functional response model (4). We call the ODE model (5) the full model in this article because each of four ODE models (1)–(4) are special cases of it. There are seven unknown parameters r, a, k, z, e, v, and b in the full model.

3. Estimating ODE Parameters

Consider a dynamical system, $\mathbf{X}(t)$, of dimension p, which can be modeled with a general ODE

$$\frac{\mathrm{d}\mathbf{X}(t)}{\mathrm{d}t} = \mathbf{f}\{\mathbf{X}(t)|\boldsymbol{\theta}\}, \tag{6}$$

where $\mathbf{f}(\cdot)$ is a p-dimensional function. The d-dimensional parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^{\mathrm{T}}$ may possibly have some zero elements. Let $\mathbf{Y}_i$, $i = 1, \ldots, n$, be the measurements of the dynamical system in the presence of measurement errors, so that

$$\mathbf{Y}_i = \mathbf{X}(t_i) + \epsilon_i, \tag{7}$$

where ϵ_i are independent and identically distributed measurement errors with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Delta} = \sigma^2\mathbf{I}_p$.

Our proposed ODE model selection method does not depend on the ODE parameter estimation method, as long as the ODE parameter estimate is root-n consistent and asymptotically normal. In this article, for specificity, the parameter cascading method (Ramsay et al., 2007) is chosen to estimate ODE parameters.

The parameter cascading method estimates the dynamical process $\mathbf{X}(t)$ as a linear combination of basis functions, so that

$$\mathbf{X}(t) = \left\{\sum_{j=1}^{J_1}\phi_{j1}(t)c_{j1}, \ldots, \sum_{j=1}^{J_p}\phi_{jp}(t)c_{jp}\right\}^{\mathrm{T}}$$
$$= \left\{\boldsymbol{\phi}_1(t)^{\mathrm{T}}\mathbf{c}_1, \ldots, \boldsymbol{\phi}_p(t)^{\mathrm{T}}\mathbf{c}_p\right\}^{\mathrm{T}} = \boldsymbol{\Phi}(t)\mathbf{c}, \tag{8}$$

where $\boldsymbol{\phi}_u(t) = \{\phi_{1u}(t), \ldots, \phi_{J_u u}(t)\}^{\mathrm{T}}$ is a vector of basis functions for the uth component of $\mathbf{X}(t)$, $\mathbf{c}_u = (c_{1u}, \ldots, c_{J_u u})^{\mathrm{T}}$ is the corresponding vector of basis coefficients, $\mathbf{c} = (\mathbf{c}_1^{\mathrm{T}}, \ldots, \mathbf{c}_p^{\mathrm{T}})^{\mathrm{T}}$, and $\boldsymbol{\Phi}(t) = \mathrm{diag}\{\boldsymbol{\phi}_1(t)^{\mathrm{T}}, \ldots, \boldsymbol{\phi}_p(t)^{\mathrm{T}}\}$. We use B-spline basis functions because of their compact support property (de Boor, 2001); precisely, they are non-zero only in short subintervals. This feature ensures their ability to provide only local adjustments and greatly increases the computational efficiency by using sparse matrix computations in Matlab (Matlab, 2013). The number of basis functions must be large enough to adequately represent $\mathbf{X}(t)$. This number can be systematically increased from a chosen starting value in any desired fashion until the estimated $\widehat{\mathbf{X}}(t)$ adequately approximates some numerical solutions of the ODE using the smoothing splines method; see Chapter 5 in Ramsay and Silverman (2005). As a rule of thumb, we find it adequate to use cubic B-spline basis functions with one knot put at each data: we use this approach in our application and simulation studies.

The parameter cascading method estimates the basis coefficient $\mathbf{c}$ and the ODE parameter $\boldsymbol{\theta}$ in two nested levels of optimization. In the inner level of optimization, the basis coefficient $\mathbf{c}$ can be estimated, for any given ODE parameter $\boldsymbol{\theta}$, by minimizing

$$J(\mathbf{c}|\boldsymbol{\theta}) = \sum_{i=1}^{n}\{\mathbf{Y}_i - \mathbf{X}(t_i)\}^{\mathrm{T}}\{\mathbf{Y}_i - \mathbf{X}(t_i)\}$$
$$+ \lambda\int[\mathbf{X}'(t) - \mathbf{f}\{\mathbf{X}(t)|\boldsymbol{\theta}\}]^{\mathrm{T}}[\mathbf{X}'(t) - \mathbf{f}\{\mathbf{X}(t)|\boldsymbol{\theta}\}]\mathrm{d}t, \tag{9}$$

where the smoothing parameter λ controls the trade-off between fitting the data and fidelity to the ODE model. The integral term in (9) also serves as the penalty on the roughness of the fitted curve, $\mathbf{X}(t)$, because $\mathbf{X}(t)$ has to be smooth enough to ensure the derivative, $\mathbf{X}'(t)$, is close to $\mathbf{f}\{\mathbf{X}(t)|\boldsymbol{\theta}\}$. In fact, in the special case when $\mathbf{f}\{\mathbf{X}(t)|\boldsymbol{\theta}\} \equiv \mathbf{0}$, the criterion

(9) is used in the smoothing spline method (Ramsay and Silverman, 2005). The integral in (9) usually has no closed-form expression, but it can be conveniently evaluated with numerical quadrature methods. The composite Simpson's rule is used in our research, which provides an adequate approximation to the exact integral (Burden and Douglas, 2000). We choose λ using the algorithm in Section 2 of Qi and Zhao (2010).

The estimate for the basis coefficient, $\widehat{\mathbf{c}}$, can be treated as an implicit function of $\boldsymbol{\theta}$ and expressed as $\widehat{\mathbf{c}}(\boldsymbol{\theta})$. Hence, the estimated dynamical process, $\widehat{\mathbf{X}}(t)$, can also be treated as an implicit function of $\boldsymbol{\theta}$, so that

$$\widehat{\mathbf{X}}(t|\boldsymbol{\theta}) = \boldsymbol{\Phi}(t)\widehat{\mathbf{c}}(\boldsymbol{\theta}). \tag{10}$$

In the outer level of optimization, the ODE parameter $\boldsymbol{\theta}$ is estimated by minimizing

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{\mathbf{Y}_i - \widehat{\mathbf{X}}(t_i|\boldsymbol{\theta})\}^{\mathrm{T}} \{\mathbf{Y}_i - \widehat{\mathbf{X}}(t_i|\boldsymbol{\theta})\}. \tag{11}$$

The optimization is implemented with the Gauss–Newton algorithm, which is a modification of the Newton algorithm and has the advantage of not requiring the computation of the Hessian matrix, improving computational efficiency. It nevertheless has the disadvantage that the algorithm can only be used to minimize functions in the form of sum of squared functions. Starting with the initial estimate, $\boldsymbol{\theta}^{(0)}$, the Gauss–Newton algorithm proceeds with the iterations

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \left\{ \left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}\boldsymbol{\theta}} \right)^{\mathrm{T}} \left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}\boldsymbol{\theta}} \right) \right\}^{-1} \left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}\boldsymbol{\theta}} \right)^{\mathrm{T}} \mathbf{r},$$

where $\mathbf{r}$ is a vector of length np by stacking $\mathbf{Y}_i - \widehat{\mathbf{X}}(t_i|\boldsymbol{\theta})$ together, and uses the chain rule

$$\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}\boldsymbol{\theta}} = \frac{\partial \mathbf{r}}{\partial \boldsymbol{\theta}} + \left(\frac{\mathrm{d}\widehat{\mathbf{c}}}{\mathrm{d}\boldsymbol{\theta}} \right)^{\mathrm{T}} \frac{\partial \mathbf{r}}{\partial \widehat{\mathbf{c}}}.$$

When the variances of the components of the measurement error ϵ_i are different, that is, $\boldsymbol{\Delta} = \mathrm{diag}(\sigma_1^2, ..., \sigma_p^2)$, it is necessary to change the first term of $\{\mathbf{Y}_i - \mathbf{X}(t_i)\}^{\mathrm{T}}\{\mathbf{Y}_i - \mathbf{X}(t_i)\}$ of (9) to $\{\mathbf{Y}_i - \mathbf{X}(t_i)\}^{\mathrm{T}}\mathbf{W}\{\mathbf{Y}_i - \mathbf{X}(t_i)\}$ and change $\{\mathbf{Y}_i - \widehat{\mathbf{X}}(t_i|\boldsymbol{\theta})\}^{\mathrm{T}}\{\mathbf{Y}_i - \widehat{\mathbf{X}}(t_i|\boldsymbol{\theta})\}$ of (11) to $\{\mathbf{Y}_i - \widehat{\mathbf{X}}(t_i|\boldsymbol{\theta})\}^{\mathrm{T}}\mathbf{W}\{\mathbf{Y}_i - \widehat{\mathbf{X}}(t_i|\boldsymbol{\theta})\}$, where $\mathbf{W}$ is a diagonal matrix, depending on $\boldsymbol{\Delta}$, and is typically $\boldsymbol{\Delta}^{-1}$.

4. Model Selection by Least Squares Approximation

Let $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,d})^{\mathrm{T}}$ be the true value of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{full}} = (\widehat{\theta}_{\mathrm{full},1}, \ldots, \widehat{\theta}_{\mathrm{full},d})^{\mathrm{T}}$ be the estimate of $\boldsymbol{\theta}$ under the full model using the parameter cascading method of Section 3. Qi and Zhao (2010) show that under some regularity conditions,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{full}} - \boldsymbol{\theta}_0) \to \mathrm{Normal}(0, \boldsymbol{\Sigma}), \tag{12}$$

as $n \to \infty$, where $\boldsymbol{\Sigma}$ is the variance–covariance matrix of the limiting distribution. Let $\widehat{\boldsymbol{\Sigma}}$ be the estimate of $\boldsymbol{\Sigma}$, which may be obtained according to Appendix A.3 of Ramsay et al. (2007).

The least squares approximation (LSA) method is used to select ODE models. Specifically, we minimize the LSA criterion with an adaptive LASSO-type penalty, which is defined as follows

$$Q(\boldsymbol{\theta}|\rho) = (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{full}})^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{full}}) + \rho \sum_{j=1}^{d} |\widehat{\theta}_{\mathrm{full},j}|^{-\gamma} |\theta_j|, \tag{13}$$

where ρ is a tuning parameter, and γ is a prespecified positive number. The least angle regression (LARS) algorithm proposed by Efron et al. (2004) can be used to find the entire solution path in minimizing $Q(\boldsymbol{\theta}|\rho)$.

Let $\widehat{\boldsymbol{\theta}}_{\mathrm{LSA}}(\rho) = \{\widehat{\theta}_{\mathrm{LSA},1}(\rho), \ldots, \widehat{\theta}_{\mathrm{LSA},d}(\rho)\}^{\mathrm{T}}$ denote the parameter value which minimizes $Q(\boldsymbol{\theta}|\rho)$, $\mathcal{A} = \{j : \theta_j \neq 0\}$, and $\widehat{\mathcal{A}}(\rho) = \{j : \widehat{\theta}_{\mathrm{LSA},j}(\rho) \neq 0\}$. The following theorem shows the selection consistency of $\widehat{\boldsymbol{\theta}}_{\mathrm{LSA}}(\rho)$.

THEOREM 1. *As $n \to \infty$, if $n^{1/2}\rho \to 0$, $n^{(1+\gamma)/2}\rho \to \infty$, and $\widehat{\boldsymbol{\Sigma}}$ converges to a positive definite matrix $\boldsymbol{\Sigma}^*$ in probability, then*

$$\Pr\{\widehat{\mathcal{A}}(\rho) = \mathcal{A}\} \to 1. \tag{14}$$

The proof of Theorem 1 is provided in the supplementary file.

From Fan and Li (2001), an estimation procedure δ is called to have the oracle property if the nonzero coefficient set $\widehat{\mathcal{A}}_\delta$, determined by δ, and the estimated coefficient $\widehat{\boldsymbol{\theta}}_\delta$ have the following properties:

- **a.** It identifies the right subset model: $\Pr(\widehat{\mathcal{A}}_\delta = \mathcal{A}) \to 1$ as $n \to \infty$,
- **b.** It has the optimal estimation rate: $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\delta,\widehat{\mathcal{A}}_\delta} - \boldsymbol{\theta}_{\mathcal{A}}) \to \mathrm{Normal}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{A}})$ in distribution as $n \to \infty$, where $\boldsymbol{\theta}_{\mathcal{A}}$ consists of all nonzero components of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_{\delta,\widehat{\mathcal{A}}_\delta}$ is its estimator by the procedure δ, and $\boldsymbol{\Sigma}_{\mathcal{A}}$ is the corresponding variance–covariance matrix when all the nonzero components of $\boldsymbol{\theta}$ are known.

Although $\widehat{\boldsymbol{\theta}}_{\mathrm{LSA}}(\rho)$ has selection consistency, it may not have the oracle property, because the covariance assumption of Wang and Leng (2007) is not satisfied, which is explained as follows. Let $\boldsymbol{\Sigma}$ denote the variance–covariance matrix of the limiting distribution of the ODE parameter estimates for the full model. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, and $\boldsymbol{\Omega}^{(s)}$ be the sub-matrix of $\boldsymbol{\Omega}$ associated with the sub-model s. Let $\boldsymbol{\Sigma}_s$ denote the variance–covariance matrix of the limiting distribution of the ODE parameter estimates for the sub-model s, and $\boldsymbol{\Omega}_s = \boldsymbol{\Sigma}_s^{-1}$. The covariance assumption of Wang and Leng (2007) is $\boldsymbol{\Omega}_s = \boldsymbol{\Omega}^{(s)}$. The expression of $\boldsymbol{\Sigma}$, given by Formula (5.24) of Qi and Zhao (2010), has a sandwich form, and thus $\boldsymbol{\Omega}_s \neq \boldsymbol{\Omega}^{(s)}$ generally. Therefore, the covariance assumption of Wang and Leng (2007) is not satisfied here.

However, we can further estimate parameters under the selected ODE model by using the parameter cascading method introduced in Section 3. Formula (14) means that the model selected by minimizing LSA criterion is the true model with probability approaching to one when $n \to \infty$. When estimating ODE parameters under the true model using the parameter cascading method, the ODE parameter estimates has the

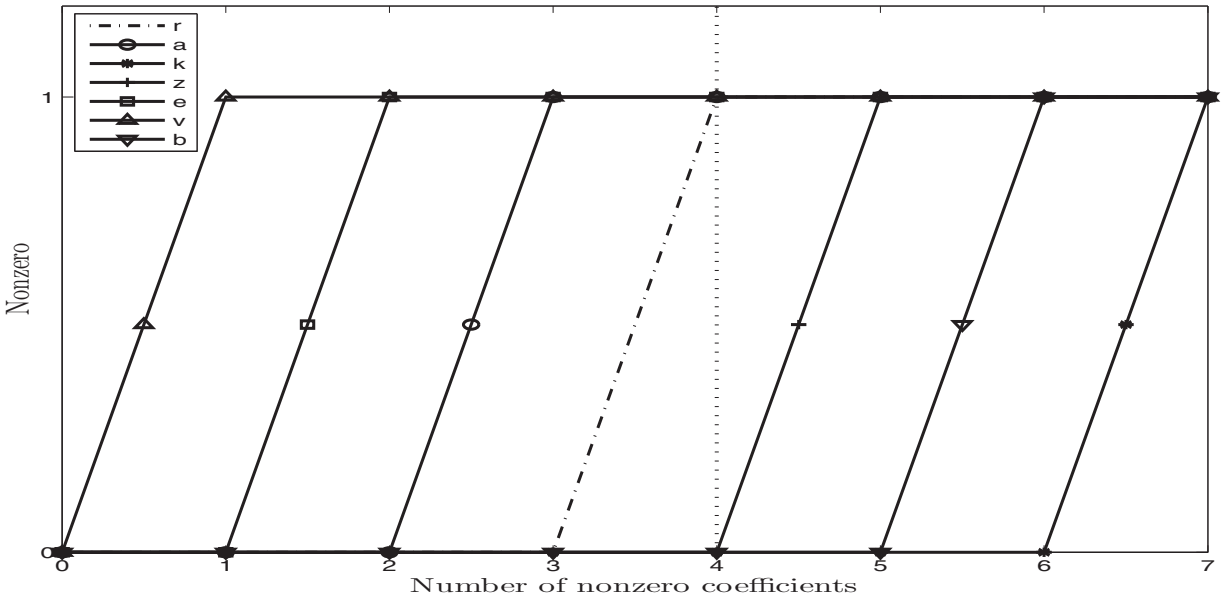

Figure 2. ODE model selection path obtained with the least squares approximation method. The y-axis indicates whether the ODE parameter estimates are zero or not. The x-axis is the number of nonzero parameter estimates. The dashed line indicates the ODE model selected by BIC.

optimal estimation rate (Qi and Zhao 2010). Therefore, the resulting ODE parameter estimates have the oracle property.

As proposed by Wang and Leng (2007), we set $\gamma = 1$ and choose the tuning parameter ρ by minimizing the BIC criterion

$$\mathrm{BIC}(\rho) = \left\{\widehat{\boldsymbol{\theta}}_{\mathrm{LSA}}(\rho) - \widehat{\boldsymbol{\theta}}_{\mathrm{full}}\right\}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}^{-1} \left\{\widehat{\boldsymbol{\theta}}_{\mathrm{LSA}}(\rho) - \widehat{\boldsymbol{\theta}}_{\mathrm{full}}\right\} + \log(n) g_\rho / n, \tag{15}$$

g_ρ is the number of nonzero parameters in $\widehat{\boldsymbol{\theta}}_{\mathrm{LSA}}(\rho)$. According to whether the resulting model $\widehat{\mathcal{A}}(\rho)$ is overfitted, underfitted or the true model, we partition $(0, \infty)$ into the following three mutually exclusive regions:

$$\mathbb{R}^{\mathrm{over}} = \{\rho \in (0,\infty) : \mathcal{A} \subset \widehat{\mathcal{A}}(\rho)\},$$

$$\mathbb{R}^{\mathrm{under}} = \{\rho \in (0,\infty) : \mathcal{A} \not\subseteq \widehat{\mathcal{A}}(\rho)\},$$

and

$$\mathbb{R}^{\mathrm{true}} = \{\rho \in (0,\infty) : \mathcal{A} = \widehat{\mathcal{A}}(\rho)\}.$$

Let $\rho^* = n^{-2/3}$. Then ρ^* satisfies the conditions of Theorem 1. So

$$\Pr\{\widehat{\mathcal{A}}(\rho^*) = \mathcal{A}\} \to 1, \tag{16}$$

as $n \to \infty$. Then we obtain the following theorem.

THEOREM 2. *As $n \to \infty$, if $\widehat{\boldsymbol{\Sigma}}$ converges to a positive definite matrix $\boldsymbol{\Sigma}^*$ in probability, then*

$$\Pr\{\inf_{\rho \in \mathbb{R}^{\mathrm{under}} \cup \mathbb{R}^{\mathrm{over}}} BIC(\rho) > BIC(\rho^*)\} \to 1. \tag{17}$$

The proof of Theorem 2 is provided in the supplementary file.

The results (16) and (17) imply that any ρ failing to identify the true model cannot be selected with probability approaching to one, that is, the BIC is consistent in selecting the tuning parameter and thus the ODE model. The computation for selecting the tuning parameter is also very efficient, because the LARS algorithm is used to find the solution path in minimizing $Q(\boldsymbol{\theta}|\rho)$, and we only need to compare BIC values of solutions in the solution path of LARS.

5. Application

The populations of the Canadian lynx and snowshoe hares may be modeled with any of the five predator-prey ODE models introduced in Section 2. We apply the least squares approximation method to select one ODE model using the observed populations of two species. Figure 1 displays the numbers of Canadian lynx and snowshoe hares between 1845 and 1935.

We estimate the ODE parameters in the full predator-prey ODE model (5) using the parameter cascading method, which is denoted as $\widehat{\boldsymbol{\theta}}_{\mathrm{full}}$. In the parameter cascading method, we choose the cubic B-spline basis functions with one knot put in each data point. Then we minimize the least squares approximation criterion (13) to estimate the ODE parameters $\boldsymbol{\theta}_0$. Because the least squares approximation (LSA) criterion (13) contains the adaptive LASSO-type penalty, some of the ODE parameter estimates may be zero, so that the full predator-prey ODE model (5) may be reduced to one of the four simplified ODE models, as discussed in Section 2.

Figure 2 displays the ODE model selection path obtained by minimizing the LSA criterion (13). Because the values of ODE parameter estimates have different orders of magnitude, the y-axis in Figure 2 does not indicate the estimated parameter values that appear in commonly-used solution path plots. Instead, it only indicates whether the ODE parameter estimates are zero or not. Figure 2 shows that estimates for the

Table 1
The estimates, standard errors, and p-values for the four parameters in the selected ODE Model (1) with the parameter cascading method from the real data

	$\widehat{r}$	$\widehat{a}$	$\widehat{e}$	$\widehat{v}$
Estimate	94.748	4.135	1.331	50.430
Standard error	35.767	1.561	0.686	25.986
p-value	0.008	0.008	0.052	0.052

ODE parameters change from zero to nonzero in the order v, e, a, r, z, b, k.

BIC defined in (15) chooses the model with four non-zero parameters, which are r, a, e, and v. The full ODE model (5) is then reduced to the ODE model (1). Table 1 displays the estimates for the four non-zero ODE parameters, their standard errors, and p-values. It shows that the estimates for r and a are statistically significant at the 0.01 significance level, and the p-values for the estimates of e and v are also very close to 0.05.

6. Simulations

6.1. *Simulation of Coefficient Estimation*

In this subsection, the finite sample performance of the proposed LSA method is investigated via Monte Carlo simulations, which are also compared with the estimation based on the full and true models.

The simulated data are generated by adding random measurement errors to the numeric solution of the ODE model (5). ODE (5) is solved by setting the true initial values $H(0) = 20$ and $P(0) = 30$, and the true parameter values $(r, a, k, z, e, v, b) = (3.0, 0.4, 0.0, 0.0, 0.3, 2.0, 0.0)$, which means that the true ODE model is the simplified model (1). The random measurement errors are generated based on the bivariate normal distribution with mean $(0, 0)^{\mathrm{T}}$, and variance–covariance matrix $\mathbf{\Delta} = \sigma^2\mathbf{I}_2$, where $\mathbf{I}_2$ is an identity matrix of size 2. The finite sample performance of the proposed LSA is investigated in four scenarios by varying $\sigma = 0.02, 0.05$ and the sample size $n = 50, 100$. We run 1000 simulation replications in each scenario. The ODE parameters are estimated with the parameter cascading method by using the cubic B-spline basis functions with one knot put in each data point.

We summarize the simulation results in Table 2. The finite sample performance of the parameter estimates are evaluated by calculating the squared error $\sum_{j=1}^{d}(\widehat{\theta}_j - \theta_{0j})^2$ and the absolute error $\sum_{j=1}^{d}|\widehat{\theta}_j - \theta_{0j}|$, where θ_{0j} is the true value of the jth parameter. It is seen that our method performs much better than the parameter estimation based on the full model. Specifically, when $\sigma = 0.02$, the LSA method selects the true model in all 1000 simulation replications, and the medians and means of the squared error and the absolute error of parameter estimates using the LSA method and true model are exactly the same. When $\sigma = 0.05$, the percentages of LSA selecting the true model are also very high, being 94.7% and 99.9% for $n = 50$ and $n = 100$, respectively. Figures 3 and 4 shows boxplots of the squared errors and the absolute errors of parameter estimates in the 1000 simulation replications.

In this simulation, we have also tried using the AIC and BIC methods proposed by Miao et al. (2009). However, we found that these two methods were very time consuming for

Table 2
The means, medians, and standard deviations (SD) of the squared error $\sum_{j=1}^{d}(\widehat{\theta}_j - \theta_{0j})^2$ and the absolute error $\sum_{j=1}^{d}|\widehat{\theta}_j - \theta_{0j}|$ of the parameter estimates by using the full model, LSA method, and true model. The third row shows the percentage of the simulation replications that the LSA method chooses the correct model.

SD of measurement errors			$\sigma = 0.02$		$\sigma = 0.05$	
Sample size			50	100	50	100
Percentages of LSA selecting the correct model			100.0%	100.0%	94.7%	99.9%
Squared error	Mean	Full model	0.131	0.074	0.848	0.458
		LSA	0.019	0.010	0.583	0.074
		True model	0.019	0.010	0.118	0.065
	Median	Full model	0.061	0.036	0.370	0.219
		LSA	0.009	0.005	0.063	0.029
		True model	0.009	0.005	0.057	0.029
	SD	Full model	0.181	0.099	1.346	0.624
		LSA	0.026	0.015	1.999	0.298
		True model	0.026	0.015	0.163	0.094
Absolute error	Mean	Full model	0.330	0.249	0.827	0.619
		LSA	0.116	0.085	0.431	0.214
		True model	0.116	0.085	0.289	0.212
	Median	Full model	0.279	0.215	0.692	0.537
		LSA	0.101	0.071	0.262	0.176
		True model	0.101	0.071	0.249	0.176
	SD	Full model	0.237	0.175	0.625	0.438
		LSA	0.084	0.064	0.662	0.183
		True model	0.084	0.064	0.209	0.160

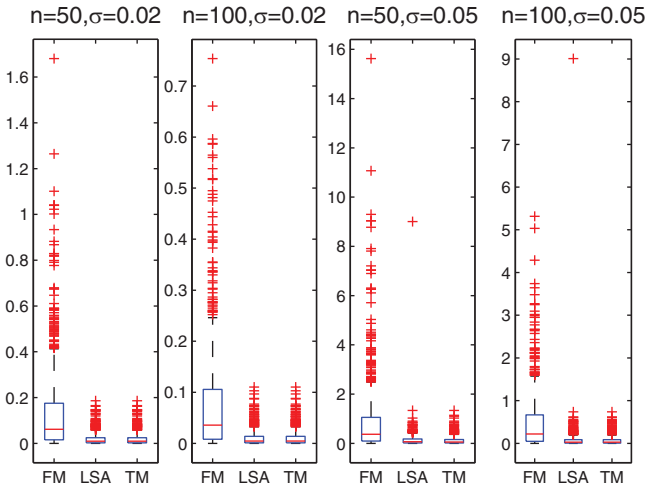

Figure 3. Boxplots for the sum squared error $\sum_{j=1}^{d}(\widehat{\theta}_j - \theta_{0j})^2$ of the parameter estimates by using the full model ("FM"), the LSA method, and the true model ("TM") in 1000 simulation replications when varying the sample size $n = 50, 100$, and the standard deviation of data noise $\sigma = 0.02, 0.05$.

two reasons. The first reason is that the full ODE model has 7 unknown parameters, so there are $2^7 = 128$ candidate models that need to be estimated to calculate AIC and BIC values. The second reason is that Miao et al. (2009) suggested using numerical solutions of the ODE model with estimated parameter values for calculating AIC and BIC values. However, the parameters are estimated very poorly (they are far away from the true values) in some candidate ODE models, and as a result, it may take a long time to solve the corresponding ODE model numerically.

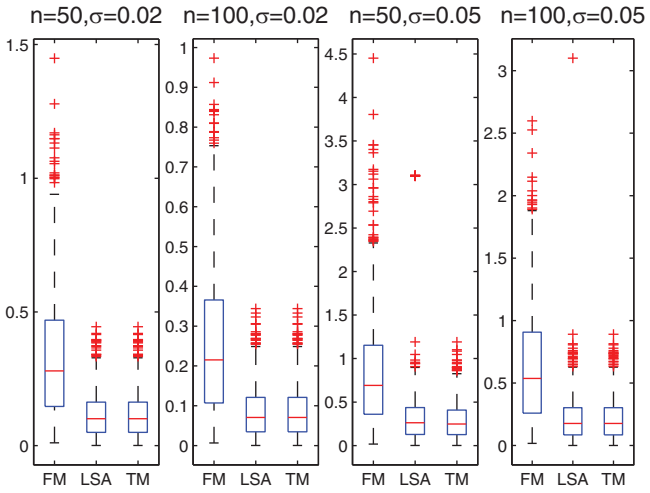

Figure 4. Boxplots for the sum absolute errors $\sum_{j=1}^{d}|\widehat{\theta}_j - \theta_{0j}|$ of the parameter estimates by using the full model ("FM"), the LSA method, and the true model ("TM") in 1000 simulation replications when varying the sample size $n = 50, 100$, and the standard deviation of data noise $\sigma = 0.02, 0.05$.

6.2. *Simulation of Coverage Probability*

Although the estimate $\widehat{\mathbf{\Sigma}}$, proposed by Ramsay et al. (2007), is used in the literature (Cao et al., 2008; Cao, Wang, and Xu, 2011), few studies have explicitly examined its performance by simulation. Since $\widehat{\mathbf{\Sigma}}$ plays an important role in the LSA criterion (13), we conduct simulation experiments to evaluate the performance of the estimate $\widehat{\mathbf{\Sigma}}$.

The simulated data are generated by adding random measurement errors to the numerical solution of the ODE model (1). The ODE model (1) is numerically solved by setting the true parameters $(r, a, e, v) = (0.3, 0.1, 0.4, 0.02)$ and the initial values $H(0) = 20$ and $P(0) = 30$. The random measurement errors are generated based on the bivariate normal distribution with the mean $(0, 0)^{\mathrm{T}}$ and the variance–covariance matrix $\mathbf{\Delta} = \sigma^2\mathbf{I}_2$, where $\mathbf{I}_2$ is an identity matrix of size 2. In our simulation studies, we investigate four scenarios by varying $\sigma = 0.02, 0.05$ and the sample size $n = 50, 100$. We run 500 simulation replications in each scenario. The parameters in the ODE model (1) are estimated with the parameter cascading method by using the cubic B-spline basis functions with one knot put in each data point.

We use the estimated variance–covariance matrix $\widehat{\mathbf{\Sigma}}$ to construct the $100(1-\alpha)\%$ confidence interval for the jth parameter θ_j as $[\widehat{\theta}_j - z_{\alpha/2}\widehat{\Sigma}_{jj}^{1/2}, \widehat{\theta}_j + z_{\alpha/2}\widehat{\Sigma}_{jj}^{1/2}]$, where $\widehat{\Sigma}_{jj}$ is the jth diagonal element of $\widehat{\mathbf{\Sigma}}$, and $z_{\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile of the standard normal distribution. We then calculate the coverage probabilities of the $100(1-\alpha)\%$ confidence interval as

$$\mathrm{CP}(\theta_j) = \sum_{t=1}^{500} I(\theta_{0j} \in [\widehat{\theta}_j - z_{\alpha/2}\widehat{\Sigma}_{jj}^{1/2}, \widehat{\theta}_j + z_{\alpha/2}\widehat{\Sigma}_{jj}^{1/2}]),$$

where $I(\cdot)$ is an indicator function, and θ_{0j} is the true value of θ_j. Table 3 displays the coverage probabilities of the 90% and 95% confidence intervals for the four parameters in the ODE model (1) when the standard deviation of the measurement errors $\sigma = 0.02, 0.05$ and the sample size $n = 50, 100$. It shows that all coverage probabilities are close to their nominal levels. When the sample size n increases or the variance of measurement errors decreases, the coverage probabilities generally become closer to their nominal levels.

7. Concluding Remarks

We have focused on predator-prey modeling, largely because there are well-known competing models in that field and also because we have an interesting data set to work with. However, our LSA method is completely general: it needs merely that there be competing ODE models and a general model that includes all of them.

As remarked in Section 1, the computational cost of the use of LSA is much lower than that of subset selection methods. This is because (a) we need not estimate many candidate ODE models; and (b) the LARS algorithm can be used to find the entire solution path in minimizing the LSA criterion.

An alternative is to directly add the adaptive LASSO type penalty to the original loss function (11). Such a method, while appealing, is likely to be extremely computationally challenging, and much more computationally expensive than LSA.

Table 3
The coverage probabilities of the 90% *and* 95% *confidence intervals for the four parameters in the ODE model* (1) *when the standard deviation of measurement errors* $\sigma = 0.02, 0.05$ *and the sample size* $n = 50, 100$

		90%				95%			
		r	a	e	v	r	a	e	v
$\sigma = 0.02$	$n = 50$	0.876	0.870	0.906	0.888	0.936	0.932	0.950	0.942
	$n = 100$	0.902	0.906	0.884	0.896	0.942	0.944	0.948	0.948
$\sigma = 0.05$	$n = 50$	0.864	0.862	0.870	0.858	0.926	0.922	0.926	0.920
	$n = 100$	0.902	0.884	0.878	0.876	0.958	0.950	0.946	0.924

8. Supplementary Materials

Proofs of theorems referenced in Section 4 and the R and Matlab code are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors would like to thank Prof Yanyuan Ma for providing the Canadian lynx and snowshoe hares data. They also thank the editor, the associate editor and three referees for their very constructive comments. Zhang's work was supported by National Natural Science Foundation of China (Grant numbers 71101141, 11271355 and 11471324). Cao's research was supported by a discovery grant (PIN: 328256) from the Natural Science and Engineering Research Council of Canada (NSERC). Carroll's research was supported by a grant from the National Cancer Institute (R37-CA057030). This article was finished during the first author's visit to Texas A&M University.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Bard, Y. (1974). *Nonlinear Parameter Estimation*. New York: Academic Press.

Biegler, L., Damiano, J. J., and Blau, G. E. (1986). Nonlinear parameter estimation: A case study comparison. *AIChE Journal* **32**, 29–45.

Brunel, N. J. (2008). Parameter estimation of ODE's via nonparametric estimators. *Electronic Journal of Statistics* **2**, 1242–1267.

Burden, R. L. and Douglas, F. J. (2000). *Numerical Analysis*, 7th edition. Pacific Grove, CA: Brooks/Cole.

Cao, J., Fussman, G., and Ramsay, J. O. (2008). Estimating a predator-prey dynamical model with the parameter cascades method. *Biometrics* **64**, 959–967.

Cao, J., Wang, L., and Xu, J. (2011). Robust estimation for ordinary differential equation models. *Biometrics* **67**, 1305–1313.

Chen, J. and Wu, H. (2008). Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. *Journal of the American Statistical Association* **103**, 369–383.

de Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–840.

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

Hall, P. and Ma, Y. (2014). Quick and easy kernel based one-step estimation of parameters in differential equations. *Journal of the Royal Statistical Society, Series B* **76**, 735–748.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Matlab (2013). *Version R2013b*. Natick, MA: The MathWorks, Inc.

Miao, H., Dykes, C., Demeter, L., and Wu, H. (2009). Differential equation modeling of hiv viral fitness experiments: Model identification, model selection, and multi-model inference. *Biometrics* **65**, 292–300.

Miao, H., Jin, X., Perelson, A. S., and Wu, H. (2012). Evaluation of multitype mathematical models for CFSE-labeling experiment data. *Bulletin of Mathematical Biology* **74**, 300–326.

Murdoch, W., Briggs, C., and Nisbet, R. (2003). *Consumer-Resource Dynamics*. New York: Princeton University Press.

Odum, E. P. and Barrett, G. W. (2004). *Fundamentals of Ecology*. Boston, MA: Brooks Cole.

Qi, X. and Zhao, H. (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Annals of Statistics* **38**, 435–481.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society, Series B* **69**, 741–796.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd edition. New York: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Stoer, J. and Bulirsch, R. (2002). *Introduction to Numerical Analysis*, 3rd edition. Berlin, New York: Springer-Verlag.

Voit, E. O. and Almeida, J. (2004). Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics* **20**, 1670–1681.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association* **102**, 1039–1048.

Williams, J. and Kalogiratou, Z. (1993). Least squares and chebyshev fitting for parameter estimation in ODEs. *Advances in Computational Mathematics* **1**, 357–366.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Received November 2013. *Revised May* 2014.
Accepted June 2014.