Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Fence method for nonparametric small area estimation

by Jiming Jiang, Thuan Nguyen and J. Sunil Rao

June 2010

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 3-11
Statistics Canada, Catalogue No. 12-001-X

Fence method for nonparametric small area estimation

Jiming Jiang, Thuan Nguyen and J. Sunil Rao [1]

Abstract

This paper considers the problem of selecting nonparametric models for small area estimation, which recently have received much attention. We develop a procedure based on the idea of fence method (Jiang, Rao, Gu and Nguyen 2008) for selecting the mean function for the small areas from a class of approximating splines. Simulation results show impressive performance of the new procedure even when the number of small areas is fairly small. The method is applied to a hospital graft failure dataset for selecting a nonparametric Fay-Herriot type model.

Key Words: Fay-Herriot Model; Fence method; Nonparametric model selection; Penalized spline; Small area estimation.

1. Introduction

Small area estimation (SAE) has received increasing attention in recent literature. Here the term small area typically refers to a population for which reliable statistics of interest cannot be produced due to certain limitations of the available data. Examples of small areas include a geographical region (*e.g.*, a state, county, municipality, *etc.*), a demographic group (*e.g.*, a specific age × sex × race group), a demographic group within a geographic region, *etc*. In absence of adequate direct samples from the small areas, methods have been developed in order to "borrow strength". Statistical models, especially mixed effects models, have played important roles in SAE. See Rao (2003) for a comprehensive account of various methods used in SAE.

While there is extensive literature on inference about small areas using mixed effects models, including estimation of small area means which is a problem of mixed model prediction, estimation of the mean squared error (MSE) of the empirical best linear unbiased predictor (EBLUP; see Rao 2003), and prediction intervals (*e.g.*, Chatterjee, Lahiri and Li 2007), model selection in SAE has received much less attention. However, the importance of model selection in SAE has been noted by prominent researchers in this field (*e.g.*, Battese, Harter and Fuller 1988, Ghosh and Rao 1994). Datta and Lahiri (2001) discussed a model selection method based on computation of the frequentist's Bayes factor in choosing between a fixed effects model and a random effects model. They focused on the following one-way balanced random effects model for the sake of simplicity: $y_{ij} = \mu + u_i + e_{ij}$, $i = 1, ..., m$, $j = 1, ..., k$, where the u_i's and e_{ij}'s are normally distributed with mean zero and variances σ_u^2 and σ_e^2, respectively. As noted by the authors, the choice between a fixed effects model and a random effects one in this case is equivalent to testing the following one-sided hypothesis $H_0: \sigma_u^2 = 0$ *vs* $H_1: \sigma_u^2 > 0$. Note that, however, not all model selection problems can be formulated as hypothesis testing. Fabrizi and Lahiri (2004) developed a robust model selection method in the context of complex surveys. Meza and Lahiri (2005) demonstrated the limitations of Mallows' C_p statistic in selecting the fixed covariates in a nested error regression model (Battese, Harter and Fuller 1988), defined as $y_{ij} = x_{ij}'\beta + u_i + e_{ij}$, $i = 1, ..., m$, $j = 1, ..., n_i$, where y_{ij} is the observation, x_{ij} is a vector of fixed covariates, β is a vector of unknown regression coefficients, and u_i's and e_{ij}'s are the same as in the model above considered by Datta and Lahiri (2001). Simulation studies carried out by Meza and Lahiri (2005) showed that the C_p method without modification does not work well in the current mixed model setting when the variance σ_u^2 is large; on the other hand, a modified C_p criterion developed by these latter authors by adjusting the intra-cluster correlations performs similarly as the C_p in regression settings. It should be pointed out that all these studies are limited to linear mixed models, while model selection in SAE in a generalized linear mixed model (GLMM) setting has never been seriously addressed.

Recently, Jiang *et al.* (2008) developed a new strategy for model selection, called *fence methods*. The authors noted a number of limitations of the traditional model selection strategies when applied to mixed model situations. For example, the BIC procedure (Schwarz 1978) relies on the effective sample size which is unclear in typical situations of SAE. To illustrate this, consider the nested error regression model introduced above. Clearly, the effective sample size is not the total number of observations $n = \sum_{i=1}^m n_i$, neither is proportional to m, the number of small areas unless all the n_i are equal and fixed. The fence methods avoid such

1. Jiming Jiang, University of California, Davis. E-mail: jiang@wald.ucdavis.edu; Thuan Nguyen, Oregon Health and Science University; J. Sunil Rao, Case Western Reserve University.

limitations, and therefore are suitable to mixed model selection problems, including linear mixed models and GLMMs. The basic idea of fence is to build a statistical fence to isolate a subgroup of what are known as the correct models. Once the fence is constructed, the optimal model is selected from those within the fence according to a criterion which can incorporate quantities of practical interest. More details about the fence methods are given below.

The focus of this paper is nonparametric models for SAE. These models have received much recent attention. In particular, Opsomer, Breidt, Claeskens, Kauermann and Ranalli (2007) proposed a spline-based nonparametric model for SAE. The idea is to approximate an unknown nonparametric small-area mean function by a penalized spline (P-spline). The authors then used a connection between P-splines and linear mixed models (Wand 2003) to formulate the approximating model as a linear mixed model, where the coefficients of the splines are treated as random effects. Consider, for simplicity, the case of univariate covariate. Then, a P-spline can be expressed as

$$\tilde{f}(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + ... + \gamma_q (x - \kappa_q)_+^p, \quad (1)$$

where p is the degree of the spline, q is the number of knots, $\kappa_j, 1 \le j \le q$ are the knots, and $x_+ = x1_{(x>0)}$. Clearly, a P-spline is characterized by p, q, and also the location of the knots. Note that, however, given p, q, the location of the knots can be selected by the space-filling algorithm implemented in R [*cover.design*()]. But the question how to choose p and q remains. The general "rule of thumb" is that p is typically between 1 and 3, and q proportional to the sample size, n, with 4 or 5 observations per knot (Ruppert, Wand and Carroll 2003). But there may still be a lot of choices given the rule of thumb. For example, if $n = 200$, the possible choices for q range from 40 to 50, which, combined with the range of 1 to 3 for p, gives a total of 33 choices for the P-spline. Our new adaptive fence method offers a data-driven approach for choosing p and q for the spline-based SAE model.

The rest of the paper is organized as follows. The fence methods are described in section 2. In section 3 we develop an adaptive fence procedure for the nonparametric model selection problem. In section 4 we demonstrate the finite sample performance of the new procedure with a series of simulation studies. In section 5 we consider a real-life data example involving a dataset from a medical survey which has been used for fitting a Fay-Herriot model (Fay and Herriot 1979). Some technical results are deferred to the appendix.

2. Fence methods

As mentioned, the basic idea of fence is to construct a statistical fence and then select an optimal model from those within the fence according to certain criterion of optimality, such as model simplicity. Let $Q_M = Q_M(y, \theta_M)$ be a measure of lack-of-fit, where y represents the vector of observations, M indicates a candidate model, and θ_M denotes the vector of parameters under M. Here by lack-of-fit we mean that Q_M satisfies the basic requirement that $\mathrm{E}(Q_M)$ is minimized when M is a true model, and θ_M the true parameter vector under M. Then, a candidate model M is in the fence if

$$\hat{Q}_M \le \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}, \quad (2)$$

where $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} Q_M$, Θ_M being the parameter space under M, $\tilde{M}$ is a model that minimizes $\hat{Q}_M$ among $M \in \mathcal{M}$, the set of candidate models, and $\hat{\sigma}_{M,\tilde{M}}$ is an estimate of the standard deviation of $\hat{Q}_M - \hat{Q}_{\tilde{M}}$. The constant c_n on the right side of (2) can be chosen as a fixed number (*e.g.*, $c_n = 1$) or adaptively (see below).

The calculation of $\hat{Q}_M$ is usually straightforward. For example, in many cases Q_M can be chosen as the negative log-likelihood, or residual sum of squares. On the other hand, the computation of $\hat{\sigma}_{M,\tilde{M}}$ can be quite challenging. Sometimes, even if an expression can be obtained for $\hat{\sigma}_{M,\tilde{M}}$, its accuracy as an estimate of the standard deviation cannot be guaranteed in a finite sample situation. Jiang, Nguyen and Rao (2009) simplified an adaptive fence procedure proposed by Jiang *et al.* (2008). For simplicity, we assume that $\mathcal{M}$ contains a full model, M_{f}, of which each candidate model is a submodel. It follows that $\tilde{M} = M_{\mathrm{f}}$. In the simplified adaptive procedure, the fence inequality (2) is replaced by

$$\hat{Q}_M - \hat{Q}_{M_{\mathrm{f}}} \le c_n, \quad (3)$$

where c_n is chosen adaptively as follows. For each $M \in \mathcal{M}$, let $p^*(M) = \mathrm{P}^*\{M_0(c) = M\}$ be the empirical probability of selection for M, where $M_0(c)$ denotes the model selected by the fence procedure based on (3) with $c_n = c$, and P^* is obtained by bootstrapping under M_{f}. For example, under a parametric model one can estimate the model parameters under M_{f} and then use a parametric bootstrap to draw samples under M_{f}. Suppose that B samples are drawn, then $p^*(M)$ is simply the sample proportion (out of a total of B samples) that M is selected by the fence procedure based on (3) with the given c_n. Let $p^* = \max_{M \in \mathcal{M}} p^*(M)$. Note that p^* depends on c_n. Let c_n^* be the c_n that maximizes p^* and this is our choice. Jiang *et al.* (2008) offers the following explanation of the motivation behind adaptive fence. Suppose that there is a true model among the candidate models, then, the optimal model is the one from which the data is generated, and

therefore should be the most likely given the data. Thus, given c_n, one is looking for the model (using the fence procedure) that is most supported by the data or, in other words, one that has the highest (posterior) probability. The latter is estimated by bootstrapping. Note that although the bootstrap samples are generated under M_{f}, they are almost the same as those generated under the optimal model. This is because the estimates corresponding to the zero parameters are expected to be close to zero, provided that the parameter estimators under M_{f} are consistent. One then pulls off the c_n that maximizes the (posterior) probability and this is the optimal choice.

There are two extreme cases corresponding to $c_n = 0$ and $c_n = \infty$ (*i.e.*, very large). Note that if $c_n = 0$, then $p^* = 1$. This is because when $c_n = 0$ the procedure always chooses M_{f}. Similarly, if there is a unique simplest model (*e.g.*, model with minimum dimension), say, M_*, then $p^* = 1$ for very large c_n. This is because, when c_n is large enough, all models are in the fence, hence the procedure always chooses M_*, if simplicity is used as the criterion of optimality for selecting the model within the fence. These two extreme cases are handled carefully in Jiang *et al.* (2008) and Jiang *et al.* (2009). However, as noted by Jiang *et al.* (2008), the procedures to handle the extreme cases, namely, the screen tests and baseline adjustment/threshhold checking, are rarely needed in practice. For example, in most applications there are a (large) number of candidate variables, and it is believed that only a (small) subset of them are important. This means that the optimal model is neither M_* nor M_{f}. Therefore, there is no need to worry about the extreme cases, and the procedures to handle these cases can be skipped. In most applications a plot of p^* against c_n is W-shaped with the peak in the middle corresponding to c_n^*.

The left plot of Figure 2 provides an illustration. This is a plot of p^* against c_n for the example discussed in section 5. The plot shows the typical "W" shape, as described, and the peak in the middle corresponds to where the optimal c_n, *i.e.*, c_n^* is.

Jiang *et al.* (2009) established consistency of the simplified adaptive fence and studied its finite sample performance.

3. Nonparametric SAE model selection

For the simplicity of illustration we consider the following SAE model:

$$y_i = f(X_i) + B_i u_i + e_i, \quad i = 1, \ldots, m, \tag{4}$$

where y_i is an $n_i \times 1$ vector representing the observations from the i^{th} small area; $f(X_i) = [f(x_{ij})]_{1 \le j \le n_i}$ with $f(x)$ being an unknown (smooth) function; B_i is an $n_i \times b$ known matrix; u_i is a $b \times 1$ vector of small-area specific random effects; and e_i is an $n_i \times 1$ vector of sampling errors. It is assumed that u_i, e_i, $i = 1, \ldots, m$ are independent with $u_i \sim N(0, G_i)$, $G_i = G_i(\theta)$, and $e_i \sim N(0, R_i)$, $R_i = R_i(\theta)$, θ being an unknown vector of variance components. Note that, besides $f(X_i)$, the model is the same as the standard "longitudinal" linear mixed model (*e.g.*, Laird and Ware 1982, Datta and Lahiri 2000).

The approximating spline model is given by replacing $f(x)$ by $\tilde{f}(x)$ in (1), where the coefficients β's and γ's are estimated by penalized least squares, *i.e.*, by

$$\text{minimizing} \quad |y - X\beta - Z\gamma|^2 + \lambda |\gamma|^2, \tag{5}$$

where $y = (y_i)_{1 \le i \le m}$, the $(i, j)^{\text{th}}$ row of X is $(1, x_{ij}, \ldots, x_{ij}^p)$, the $(i, j)^{\text{th}}$ row of Z is $[(x_{ij} - \kappa_1)_+^p, \ldots, (x_{ij} - \kappa_q)_+^p]$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, and λ is a penalty, or smoothing, parameter. To determine λ, Wand (2003) used the following interesting connection to a linear mixed model. To illustrate the idea, let us consider a simple case in which $B_i = 0$ (*i.e.*, there is no small-area random effects), and the components of e_i are independent and distributed as $N(0, \tau^2)$. If the γ's are treated as random effects which are independent and distributed as $N(0, \sigma^2)$, then the solution to (5) are the same as the best linear unbiased estimator (BLUE) for β, and the best linear unbiased predictor (BLUP) for γ, if λ is identical to the ratio τ^2/σ^2. Thus, the value of λ may be estimated by the maximum likelihood (ML), or restricted maximum likelihood (REML) estimators of σ^2 and τ^2 (*e.g.*, Jiang 2007). However, there has been study suggesting that this approach is biased towards undersmoothing (Kauermann 2005). Consider, for example, a special case in which $f(x)$ is, in fact, the quadratic spline with two knots given by (10). (Note that this function is smooth in that it has a continuous derivative.) It is clear that, in this case, the best approximating spline should be $f(x)$ itself with only two knots, *i.e.*, $q = 2$ (of course, one could use a spline with many knots to "approximate" the two-knot quadratic spline, but that would seem very inefficient in this case). However, if one uses the above linear mixed model connection, the ML (or REML) estimator of σ^2 is consistent only if $q \to \infty$ (*i.e.*, the number of appearances of the spline random effects goes to infinity). The seeming inconsistency has two worrisome consequences: (i) the meaning of λ may be conceptually difficult to interpret; (ii) the behavior of the estimator of λ may be unpredictable.

The fence method offers a natural approach to choosing the degree of the spline, p, the number of knots, q, and the smoothing parameter, λ at the same time. Note, however, a major difference from the situations considered in Jiang *et al.* (2008) and Jiang *et al.* (2009) in that the true underlying model is not among the class of candidate models, *i.e.*, the approximating splines (1). Furthermore, the

role of λ in the model should be made clear: λ controls the degree of smoothness of the underlying model. A natural measure of lack-of-fit is $Q_M = |y - X\beta - Z\gamma|^2$. However, $\hat{Q}_M$ is not obtained by minimizing Q_M over β and γ without constraint. Instead, we have $\hat{Q}_M = |y - X\hat{\beta} - Z\hat{\gamma}|^2$, where $\hat{\beta}$ and $\hat{\gamma}$ are the solution to (5), and hence depends on λ. The optimal λ is to be selected by the fence method, together with p and q, as described below.

Another difference is that there may not be a full model among the candidate models. Therefore, the fence inequality (3) is replaced by the following:

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \le c_n, \tag{6}$$

where $\tilde{M}$ is the candidate model that has the minimum $\hat{Q}_M$. We use the following criterion of optimality within the fence which combines model simplicity and smoothness. For the models within the fence, choose the one with the smallest q; if there are more than one such models, choose the model with the smallest p. This gives the best choice of p and q. Once p, q are chosen, we choose the model *within the fence* with the largest λ. Once again, note that λ is part of the model M that is selected (or "estimated") by the fence method. The tuning constant c_n is chosen adaptively using the simplified adaptive procedure of Jiang *et al.* (2009), where parametric bootstrap is used for computing p^* (see section 2).

The following theorem is proved in Appendix. For simplicity, assume that the matrix $W = (X\ Z)$ is of full rank. Let $P_{W^\perp} = I_n - P_W$, where $n = \sum_{i=1}^m n_i$ and $P_W = W(W'W)^{-1}W'$.

Theorem. Computationally, the above fence procedure is equivalent to the following: (i) first use the (adaptive) fence to select p and q using (6) with $\lambda = 0$ and $\hat{Q}_M = y' P_{W^\perp} y$ (see Lemma below), and same criterion as above for choosing p, q within the fence; (ii) let M_0^* denotes the model corresponding to the selected p and q, find the maximum λ such that

$$\hat{Q}_{M_0^*,\lambda} - \hat{Q}_{\tilde{M}} \le c_n^*, \tag{7}$$

where for any model M with the corresponding X and Z, we have

$$\begin{aligned}
\hat{Q}_{M,\lambda} &= |y - X\hat{\beta}_\lambda - Z\hat{\gamma}_\lambda|^2,\\
\hat{\beta}_\lambda &= (X'V_\lambda^{-1}X)^{-1}X'V_\lambda^{-1}y,\\
\hat{\gamma}_\lambda &= \lambda^{-1}(I_q + \lambda^{-1}Z'Z)^{-1}Z'(y - X\hat{\beta}_\lambda),\\
X'V_\lambda^{-1}X &= X'X - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'X,\\
X'V_\lambda^{-1}y &= X'y - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'y,
\end{aligned}$$

and c_n^* is chosen by the adaptive fence procedure described in section 2 (V_λ is defined below but not directly needed here for the computation because of the last two equations).

Note that in step (i) of the Theorem one does not need to deal with λ. The motivation for (7) is that this inequality is satisfied when $\lambda = 0$, so one would like to see how far λ can go. In fact, the maximum λ is a solution to the equation $\hat{Q}_{M_0^*,\lambda} - \hat{Q}_{\tilde{M}} = c_n^*$. The purpose of the last two equations is to avoid direct inversion of $V_\lambda = I_n + \lambda^{-1}ZZ'$, whose dimension is equal to n, the total sample size. Note that V_λ does not have a block diagonal structure because of ZZ', so if n is large direct inversion of V_λ may be computationally burdensome.

The proof of the Theorem requires the following lemma, whose proof is given in Appendix.

Lemma. For any M and y, $\hat{Q}_{M,\lambda}$ is an increasing function of λ with $\inf_{\lambda>0} \hat{Q}_{M,\lambda} = \hat{Q}_M$.

4. Simulations

We consider an extension of the Fay-Herriot model (Fay and Herriot 1979) in a nonparametric setting. The model can be expressed as

$$y_i = f(x_i) + v_i + e_i,\ i = 1, \ldots, m, \tag{8}$$

where $v_i, e_i, i = 1, \ldots, m$ are independent such that $v_i \sim N(0, A)$, $e_i \sim N(0, D_i)$, where A is unknown but the sampling variance D_i is assumed known. The main difference from the traditional Fay-Herriot model is $f(x_i)$, where $f(x)$ is an unknown smooth function.

For simplicity we assume $D_i = D$, $1 \le i \le m$. Then, the model can be expressed as

$$y_i = f(x_i) + \varepsilon_i,\ i = 1, \ldots, m, \tag{9}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma^2 = A + D$, which is unknown. Thus, the model is the same as the nonparametric regression model.

We consider three different cases that cover various situations and aspects. In the first case, Case 1, the true underlying function is a linear function, $f(x) = 1 - x$, $0 \le x \le 1$, hence the model reduces to the traditional Fay-Herriot model. The goal is to find out if fence can validate the traditional Fay-Herriot model in the case that it is valid. In the second case, Case 2, the true underlying function is a quadratic spline with two knots, given by

$$f(x) = 1 - x + x^2 - 2(x-1)_+^2 + 2(x-2)_+^2,\ 0 \le x \le 3 \tag{10}$$

(the shape is half circle between 0 and 1 facing up, half circle between 1 and 2 facing down, and half circle between 2 and 3 facing up). Note that this function is smooth in that it has a continuous derivative. Here we intend to investigate whether the fence can identify the true underlying function in the "perfect" situation, *i.e.*, when $f(x)$ itself is a spline. The last case, Case 3, is perhaps the most practical situation,

in which no spline can provide a perfect approximation to $f(x)$. In other words, the true underlying function is not among the candidates. In this case $f(x)$ is chosen as $0.5\sin(2\pi x),\ 0 \le x \le 1,$ which is one of the functions considered by Kauermann (2005).

We consider situations of small or medium sample size, namely, $m = 10, 15$ or 20 for Case I, $m = 30$, 40 or 50 for Case 2, and $m = 10, 30$ or 50 for Case 3. The covariate x_i are generated from the Uniform$[0, 1]$ distribution in Case 1, and from Uniform$[0, 3]$ in Case 2; then fixed throughout the simulations. Following Kauermann (2005), we let x_i be the equidistant points in Case 3. The error standard deviation σ in (9) is chosen as 0.2 in Case 1 and Case 2. This value is chosen such that the signal standard deviation in each case is about the same as the error standard deviation. As for Case 3, we consider three different values for $\sigma, 0.2, 0.5$ and 1.0. These values are also of the same order as the signal standard deviation in this case.

The candidate approximating splines for Case 1 and Case 2 are the following: $p = 0, 1, 2, 3,\ q = 0$ and $p = 1, 2, 3,\ q = 2, 5$ (so there are a total of 10 candidates). As for Case 3, following Kauermann (2005), we consider only linear splines (*i.e.*, $p = 1$); furthermore, we consider the number of knots in the range of the "rule of thumb" (*i.e.*, roughly 4 or 5 observations per knot; see section 1), plus the intercept model $(p = q = 0)$ and the linear model $(p = 1,\ q = 0)$. Thus, for $m = 10,\ q = 0,\ 2,\ 3;$ for $m = 30,\ q = 0, 6, 7, 8;$ and for $m = 50,\ q = 0, 10, 11, 12, 13.$

Table 1 shows the results based on 100 simulations under Case 1 and Case 2. As in Jiang *et al.* (2009), we consider both the highest peak, that is, choosing c_n with the highest p^*, and 95% lower bound (L.B.), that is, choosing a smaller c_n corresponding to a peak of p^* in order to be conservative, if the corresponding p^* is greater than the 95% lower bound of the p^* for any larger c_n that corresponds to a peak of p^*. It is seen that performance of the adaptive fence is satisfactory even with the small sample size. Also, it appears that the confidence lower bound method works better in smaller sample, but makes almost no difference in larger sample. These are consistent with the findings of Jiang *et al.* (2009).

Table 1
Nonparametric model selection - Case 1 and Case 2. Reported are empirical probabilities, in terms of percentage, based on 100 simulations that the optimal model is selected

	Case 1			Case 2		
Sample size	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$
Highest Peak	62	91	97	71	83	97
Confidence L.B.	73	90	97	73	80	96

Table 2 shows the results for Case 3. Note that, unlike Case 1 and Case 2, here there is no optimal model (an optimal model must be a true model according to our definition). So, instead of giving the empirical probabilities of selecting the optimal model, we give the empirical distribution of the selected models in each case. It is apparent that, as σ increases, the distribution of the models selected becomes more spread out. A reverse pattern is observed as m increases. The confidence lower bound method appears to perform better in picking up a model with splines. Within the models with splines, fence seems to overwhelmingly prefer fewer knots than more knots.

Table 2
Nonparametric model selection - Case 3. Reported are empirical distributions, in terms of percentage, of the selected models

	Sample Size	$m = 10$		$m = 30$		$m = 50$	
	# of Knots	**0, 2, 3**		**0, 6, 7, 8**		**0, 10, 11, 12, 13**	
		(p, q)	%	(p, q)	%	(p, q)	%
$\sigma = 0.2$	Highest Peak	(0, 0)	1	(1, 0)	9	(1, 10)	100
		(1, 0)	31	(1, 6)	91		
		(1, 2)	68				
	Confidence L.B.	(1, 0)	24	(1, 0)	9	(1, 10)	100
		(1, 2)	76	(1, 6)	91		
$\sigma = 0.5$	Highest Peak	(0, 0)	14	(1, 0)	21	(1, 0)	13
		(1, 0)	27	(1, 6)	77	(1, 10)	84
		(1, 2)	56	(1, 7)	2	(1, 11)	2
		(1, 3)	3			(1, 12)	1
	Confidence L.B.	(0, 0)	8	(1, 0)	8	(1, 0)	2
		(1, 0)	23	(1, 6)	89	(1, 10)	94
		(1, 2)	65	(1, 7)	3	(1, 11)	2
		(1, 3)	4			(1, 12)	2
$\sigma = 1$	Highest Peak	(0, 0)	27	(0, 0)	15	(0, 0)	10
		(1, 0)	20	(1, 0)	18	(1, 0)	26
		(1, 2)	49	(1, 6)	63	(1, 10)	60
		(1, 3)	4	(1, 7)	4	(1, 11)	2
						(1, 12)	2
	Confidence L.B.	(0, 0)	20	(0, 0)	1	(0, 0)	2
		(1, 0)	13	(1, 0)	13	(1, 0)	13
		(1, 2)	59	(1, 6)	82	(1, 10)	80
		(1, 3)	8	(1, 7)	4	(1, 11)	2
						(1, 12)	3

Note that the fence procedure allows us to choose not only p and q but also λ (see section 3). In each simulation we compute $\hat{\beta} = \hat{\beta}_\lambda$ and $\hat{\gamma} = \hat{\gamma}_\lambda$, given below (7), based on the λ chosen by the adaptive fence. The fitted values are calculated by (1) with β and γ replaced by $\hat{\beta}$ and $\hat{\gamma}$, respectively. We then average the fitted values over the 100 simulations. Figure 1 shows the average fitted values for the three cases $(m = 10, 30, 50)$ with $\sigma = 0.2$ under Case 3. The true underlying function values, $f(x_i) = 0.5\sin(2\pi x_i)$, $i = 1, \ldots, m$ are also plotted for comparison.

5. A real-life data example

We consider a dataset from Morris and Christiansen (1995) involving 23 hospitals (out of a total of 219 hospitals) that had at least 50 kidney transplants during a 27 month period (Table 3). The y_i's are graft failure rates for kidney transplant operations, that is, y_i = number of graft failures $/n_i$, where n_i is the number of kidney transplants at hospital i during the period of interest. The variance for graft failure rate, D_i, is approximated by $(0.2)(0.8)/n_i$, where 0.2 is the observed failure rate for all hospitals. Thus, D_i is assumed known. In addition, a severity index x_i is available for each hospital, which is the average fraction of females, blacks, children and extremely ill kidney recipients at hospital i. The severity index is considered as a covariate.

Table 3
Hospital data from Morris and Christiansen (1995)

Area	y_i	x_i	$\sqrt{D_i}$
1	0.302	0.112	0.055
2	0.140	0.206	0.053
3	0.203	0.104	0.052
4	0.333	0.168	0.052
5	0.347	0.337	0.047
6	0.216	0.169	0.046
7	0.156	0.211	0.046
8	0.143	0.195	0.046
9	0.220	0.221	0.044
10	0.205	0.077	0.044
11	0.209	0.195	0.042
12	0.266	0.185	0.041
13	0.240	0.202	0.041
14	0.262	0.108	0.036
15	0.144	0.204	0.036
16	0.116	0.072	0.035
17	0.201	0.142	0.033
18	0.212	0.136	0.032
19	0.189	0.172	0.031
20	0.212	0.202	0.029
21	0.166	0.087	0.029
22	0.173	0.177	0.027
23	0.165	0.072	0.025

Ganesh (2009) proposed a Fay-Herriot model for the graft failure rates. as follows: $y_i = \beta_0 + \beta_1 x_i + v_i + e_i$, where the v_i's are hospital-specific random effects and e_i's are sampling errors. It is assumed that v_i, e_i are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$. Here the variance A is unknown. Based on the model Ganesh obtained credible intervals for selected contrasts. However, inspections of the raw data suggest some nonlinear trends, which raises the question on whether the fixed effects part of the model can be made more flexible in its functional form.

To answer this question, we consider the Fay-Herriot model as a special member of a class of approximating spline models discussed in section 3. More specifically, we assume

$$y_i = f(x_i) + v_i + e_i, \quad i = 1, \ldots, m, \tag{11}$$

where $f(x)$ is an unknown smooth function and everything else are the same as in the Fay-Herriot model. We then consider the following class of approximating spline models:

$$\hat{f}(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + \ldots + \gamma_q (x - \kappa_q)_+^p \tag{12}$$

with $p = 0, 1, 2, 3$ and $q = 0, 1, \ldots, 6$ ($p = 0$ is only for $q = 0$). Here the upper bound 6 is chosen according to the "rule-of-thumb" (because $m = 23$, so $m/4 = 5.75$). Note that the Fay-Herriot model corresponds to the case $p = 1$ and $q = 0$. The question is then to find the optimal model, in terms of p and q, from this class.

We apply the adaptive fence method described in section 3 to this case. Here to obtain the bootstrap samples needed for obtaining c_n^*, we first compute the ML estimator under the model $\tilde{M}$, which minimizes $\hat{Q}_M = y' P_{W^\perp} y$ among the candidate models [*i.e.*, (12); see Theorem in section 3], then draw parametric bootstrap samples under model $\tilde{M}$ with the ML estimators treated as the true parameters. This is reasonable because $\tilde{M}$ is the best approximating model in terms of the fit, even though under model (11) there may not be a true model among the candidate models. The bootstrap sample size is chosen as 100.

The fence method selects the model $p = 3$ and $q = 0$, that is, a cubic function with no knots, as the optimal model. To make sure that the bootstrap sample size $B = 100$ is adequate, we repeated the analysis 100 times, each time using different bootstrap samples (recall in the adaptive fence one needs to draw bootstrap samples in order to determine c_n^*, so the question is whether different bootstrap samples lead to different results of model selection). All results led to the same model: a cubic function with no knots (even though the bootstrap-derived intermediate quantities, such as p^* and c_n^*, varied across bootstraps). We also ran the data analysis using $B = 1{,}000$, and selected model remained the same. Thus, it appears that the bootstrap sample size $B = 100$ is adequate. The left figure of Figure 2 shows the plot of p^* against c_n in the adaptive fence model selection.

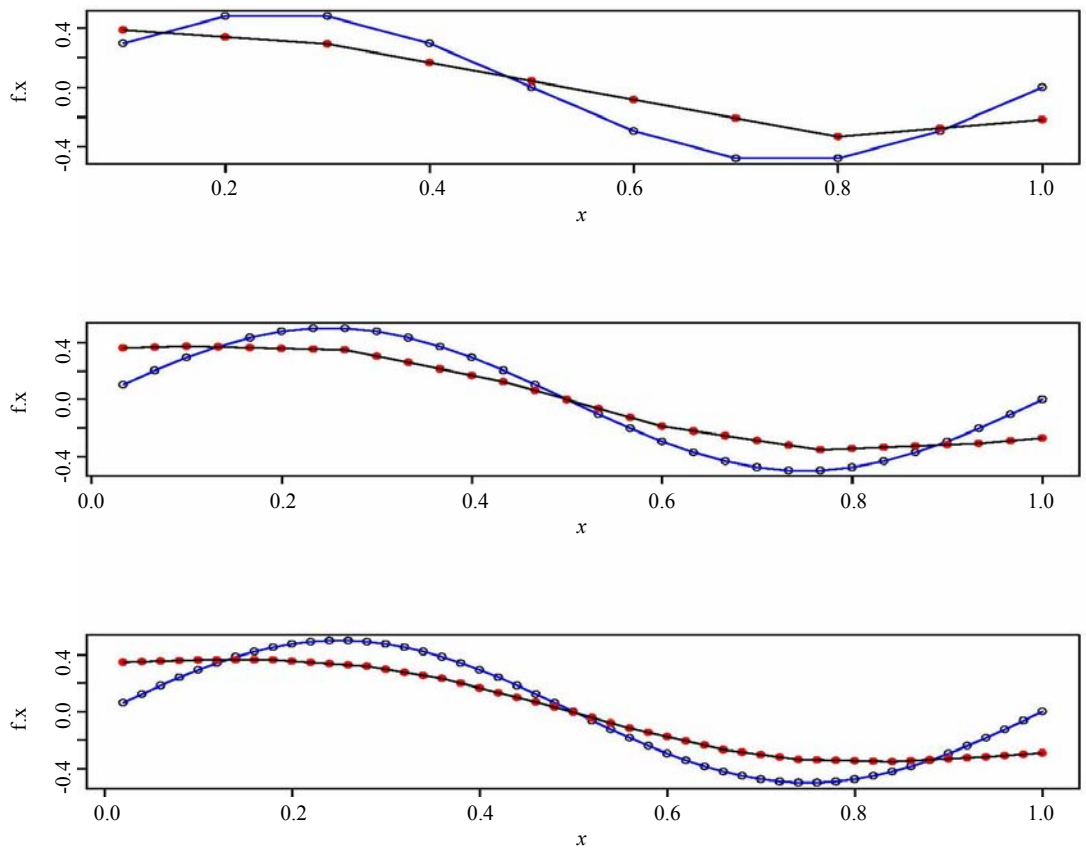

Figure 1 Case 3 Simulation. Top figure: Average fitted values for $m = 10$. Middle figure: Average fitted values for $m = 30$. Bottom figure: Average fitted values for $m = 50$. In all cases, the dots represent the fitted values, while the circles correspond to the true underlying function

A few comparisons are always helpful. Our first comparison is to fence itself but with a more restricted space of candidate models. More specifically, we consider (12) with the restriction to linear splines only, *i.e.*, $p = 1$, and knots in the range of the "rule of thumb", *i.e.*, $q = 4, 5, 6$, plus the intercept model $(p = q = 0)$ and the linear model $(p = 1,\ q = 0)$. In this case, the fence method selected a linear spline with four knots (*i.e.*, $p = 1,\ q = 4$) as the optimal model. The value of λ corresponding to this model is approximately equal to 0.001. The plot of p^* against c_n for this model selection is very similar to the left figure of Figure 2, and therefore omitted. In addition, the right figure of Figure 2 shows the fitted values and curves under the two models selected by the fence from within the different model spaces as well as the original data points.

A further comparison can be made by treating (11) as a generalized additive model (GAM) with heteroscedastic errors. A weighted fit can be obtained with the amount of smoothing optimized by using a generalized cross-validation (GCV) criterion. Here the weights used are $w_i = 1/(A + D_i)$ where the maximum likelihood estimate for A is used as a plug-in estimate. Recall that the D_i's are known. This fitted function is also overlayed in the right figure of Figure 2. Notice how closely this fitted function resembles the restricted space fence fit.

To expand the class of models under consideration by GCV-based smoothing, we used the BRUTO procedure (Hastie and Tibshirani 1990) which augments the class of models to look at a null fit and a linear fit for the spline function; and embeds the resulting model selection (*i.e.*, null, linear or smooth fits) into a weighted backfitting algorithm using GCV for computational efficiency. Interestingly here, BRUTO finds simply an overall linear fit for the fixed effects functional form. While certainly an interesting comparison, BRUTO's theoretical properties for models like (11) have not really been studied in depth.

Finally, as mentioned in section 3, by using the connection between P-spline and linear mixed model one can formulate (12) as a linear mixed model, where the spline coefficients are treated as random effects. The problem then becomes a (parametric) mixed model selection problem, hence the method of Jiang *et al.* (2009) can be applied. In fact, this was our initial approach to this dataset, and the model we found was the same as the one by BRUTO. However, we have some reservation about this approach, as explained in section 3.

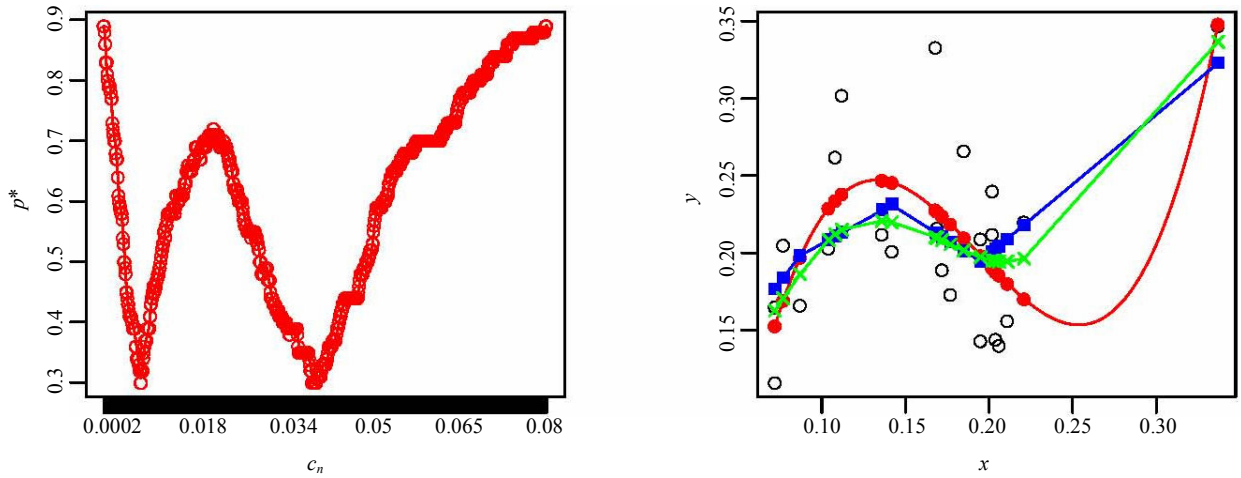

Figure 2 Left: A plot of p^* against c_n from the search over the full model space. Right: The raw data and the fitted values and curves; dots and their curve correspond to the cubic function resulted from the full model search; squares and their lines correspond to the linear spline with 4 knots resulted from the restricted model search; green X's and their lines represent the GAM fits

6. Concluding remarks

Although the focus of the current paper is nonparametric SAE model selection, our method may be applicable to spline-based mixed effects model selection problems in other areas, for example, in the analysis of longitudinal data (*e.g.*, Wang 2005).

In the case where a true model exists among the candidate models, such as Cases 1 and 2 in section 4, consistency of the proposed fence model selection method can be established in the same way as in Section 3 of Jiang *et al.* (2009) (although the result of the latter paper does not directly apply). However, practically, the situation that non-parametric modeling is most useful is when a true model does not exist, or is not among the candidates, such as Case 3 in section 4. In this case, no result of consistency can be proved, of course. It remains unclear what is a desirable asymptotic behavior to study in the latter case.

Acknowledgements

Jiming Jiang is partially supported by NSF grants DMS - 0203676 and DMS - 0402824. J. Sunil Rao is partially supported by NSF grants DMS - 0203724, DMS - 0405072 and NIH grant K25-CA89868.

Appendix

1. *Proof of Lemma.* Write $g(\lambda) = \hat{Q}_{M,\lambda}$. It can be shown (detail omitted) that $g'(\lambda) = 2\lambda y' B_\lambda A_\lambda B_\lambda' y$, where $A_\lambda = B'(W'W + \lambda BB')^{-1}B$, $B_\lambda = W(W'W + \lambda BB')^{-1}B$ with $B' = (0\ I_q)$ and $W = (X\ Z)$. Hence $g'(\lambda) \geq 0$ for $\lambda > 0$. Also $Q_{M,\lambda} \to \hat{Q}_M$ as $\lambda \to 0$.

2. *Proof of Theorem.* Consider the fence inequality

$$\hat{Q}_{M,\lambda} - \hat{Q}_{\tilde{M},\tilde{\lambda}} \leq c_n, \qquad \text{(A.1)}$$

where $(\bar{M}, \bar{\lambda})$ minimizes $\hat{Q}_{M,\lambda}$. Also consider the fence inequality using $\hat{Q}_M = y' P_{W^\perp} y$, which is

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \leq c_n. \qquad \text{(A.2)}$$

By Lemma, we must have $\bar{\lambda} = 0$, and $\bar{M} = \tilde{M}$, hence $\hat{Q}_{\bar{M},\bar{\lambda}} = \hat{Q}_{\tilde{M}}$. It follows, again by Lemma, that for the same c_n, (A.2) holds if and only if (A.1) holds for some λ. Therefore, the models within the fence, in terms of p and q, are the same under both procedures. It is then easy to see, according to the selection criterion, that the same model $M_0 = M_0(c_n)$, in terms of p and q, will be selected under both procedures for the given c_n. It then follows that the c_n^* selected using the adaptive procedure will be the same under both procedures. Then, once again using the above argument, the optimal model M_0^*, in terms of p and q, will be the same under both procedures.

The formulae below (7) can be derived using the expressions of BLUE and BLUP (*e.g.*, Jiang 2007, §2.3.1) and the following identity (*e.g.*, Sen and Srivastava 1990, page 275): If U is $n \times q$ and V is $q \times n$, then $(P + UV)^{-1} = P^{-1} - P^{-1}U(I_q + VP^{-1}U)^{-1}VP^{-1}$ so long as the inverses exist.

References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 80, 28-36.

Chatterjee, S., Lahiri, P. and Li, H. (2007). Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Annals of Statistics*, to appear.

Datta, G.S., and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

Datta, G.S., and Lahiri, P. (2001). Discussions on a paper by Efron & Gous. (Ed., P. Lahiri) *Model Selection*, IMS Lecture Notes/Monograph 38.

Fabrizi, E., and Lahiri, P. (2004). A new approximation to the Bayes information criterion in finite population sampling. Technical Report, Dept. of Math., Univ. of Maryland.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Ganesh, N. (2009). Simultaneous credible intervals for small area estimation problems. *Journal of Multivariate Analysis*, in press.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.

Hastie, T., and Tibshirani, R.J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

Jiang, J., Rao, J.S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, 36, 1669-1692.

Jiang, J., Nguyen, T. and Rao, J.S. (2009). A simplified adaptive fence procedure. *Statistics and Probability Letters*, 79, 625-629.

Kauermann, G. (2005). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning and Inference*, 127, 53-69.

Laird, N.M., and Ware, J.M. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963-974.

Meza, J., and Lahiri, P. (2005). A note on the C_p statistic under the nested error regression model. *Survey Methodology*, 31, 105-109.

Morris, C.N., and Christiansen, C.L. (1995). Hierarchical models for ranking and for identifying extremes with applications. *Bayes Statistics* 5, Oxford Univ. Press.

Opsomer, J.D., Breidt, F.J., Claeskens, G., Kauermann, G. and Ranalli, M.G. (2007). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society B*, to appear.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Ruppert, R., Wand, M. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Univ. Press.

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

Sen, A., and Srivastava, M. (1990). *Regression Analysis*. New York: Springer.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18, 223-249.

Wang, J.-L. (2005). Nonparametric regression analysis of longitudinal data. *Encyclopedia of Biostatistics*, 2nd Ed.