Statistica Sinica **27** (2017), 859-877
doi:http://dx.doi.org/10.5705/ss.202015.0212

ESTIMATION UNDER MODEL UNCERTAINTY

Nicholas T. Longford

SNTL and Imperial College

Abstract: Model selection has had a virtual monopoly on dealing with model uncertainty ever since models were identified as important conduits for statistical inference. Model averaging alleviates some of its deficiencies, but does not offer a practical solution in all settings. We propose an alternative based on linear combinations of the candidate models' estimators. The general proposal is elaborated for ordinary regression and is illustrated with examples. Some estimators based on invalid models contribute to efficient estimation of certain quantities.

Key words and phrases: Basis estimator, composite estimation, model selection, ordinary regression, propensity matching.

1. Introduction

The search for a parsimonious valid model to fit to a dataset is often justified by the belief that the corresponding estimator of any relevant parameter or another target is efficient (Burnham and Anderson (2002)). It implies estimation in two stages. First we associate each candidate model with an estimator. Then we identify the best fitting model and apply the associated estimator. Statements made about such *post-selection* (ps) estimators are not valid when they are conditional on the selected model and its validity, or assume that the selection is ignorable. This is obvious from the representation of a ps estimator as the mixture

$$\hat\theta_S \;=\; i_0\hat\theta_0 + i_1\hat\theta_1 + \cdots + i_K\hat\theta_K \;\;= \mathbf{i}^\top\hat{\boldsymbol\theta}\,, \tag{1.1}$$

where $\hat{\boldsymbol\theta} = (\hat\theta_0\,, \hat\theta_1\,, \ldots, \hat\theta_K)^\top$ are the estimators of a target θ based on respective models $M_0\,, M_1\,, \ldots,\; M_K\,,$ and the elements of $\mathbf{i} = (i_0\,, i_1\,, \ldots, i_K)^\top$ indicate the selected model S; S is a multinomial random variable defined in $\mathcal{M} = \{M_0\,, M_1\,, \ldots, M_K\}$, or in $\{0, 1, \ldots, K\}$, as $S = k$ when $i_k = 1$. A selection process S is said to be ignorable for model k when $\hat\theta_k$ has the same distribution as $(\hat\theta_k \,|\, S = k)$. Only some trivial and esoteric selection processes are ignorable.

We want to estimate θ with as small mean squared error (MSE) as possible. We assume that such an estimator $\tilde\theta$ is the final product of an analysis, and of interest is solely its realized value and an (approximately) unbiased estimator of its MSE. We set aside all issues of interpretation, such as which models k (and

corresponding estimators $\hat{\theta}_k$) contribute to $\tilde{\theta}$, and in what way. In contrast, lasso (Tibshirani (1996)) yields a single model that is intended for all inferences based on the dataset. Some of the weaknesses of the lasso are addressed by Zou (2006) and Li and Shao (2015), but it remains a ps estimator. The SCAD selection process (Fan and Li (2001)) has good (asymptotic) properties of selecting the important covariates and estimating regression coefficients. Efron (2014) studies the properties of ps estimators by resampling methods. Rolling and Yang (2014) seek a single model on which to base estimation of a narrow class of quantities related to non-constant treatment effects.

Using ps estimators is in discord with two principles. First, they ignore the consequences of the errors in selection; their distributions are distorted by selection. The selection should be informed by the target of estimation, θ, but in the current practice it usually is not. Second, estimation under model uncertainty is an application of the EM algorithm (Dempster, Laird and Rubin (1977)) in which the appropriate model is the missing information. The E-step of the algorithm estimates the conditional probabilities of the alternative models, and the M-step combines the estimators in $\hat{\boldsymbol{\theta}}$ with these probabilities as the weights. Thus, $\hat{\theta}_k$ and M_k that do not win the selection contest should have a role in estimation of θ. In a ps estimator they do not. In model averaging, such runners-up are incorporated, but with weights not sensitive to the target θ. This criticism is in line with Claeskens and Hjort (2003) and Longford (2012) who concluded that no single process S yields efficient estimators $\hat{\theta}_S$ for a wide range of targets θ, and estimators have to be combined differently for one target than for another.

We award no credit for unbiasedness (nor for small bias), or for analytical or distributional simplicity of the estimator. We eschew any asymptotic arguments or derivations, because we regard model uncertainty as a sub-asymptotic problem. Oracle properties of an estimator are also irrelevant because they refer to asymptotics, and we find some invalid models useful for estimation in finite samples. In brief, we separate estimation from the discovery of the structure underlying the data, and justify this approach by both examples and theoretical arguments. Our derivations resemble those of Liang et al. (2011) but, unlike us, they pursue model averaging as a compromise. We show that greater flexibility is essential and the averaging has to be target-specific.

Our solution is based on composite estimators, defined as linear combinations

$$\tilde{\theta} \,=\, \mathbf{c}^\top \hat{\boldsymbol{\theta}} \,=\, c_0\hat{\theta}_0 + c_1\hat{\theta}_1 + \cdots + c_K\hat{\theta}_K\,, \tag{1.2}$$

where $\mathbf{c}$ is a vector of constants that add up to unity: $\mathbf{c}^\top\mathbf{1} = 1$, where $\mathbf{1}$ is the vector of unities of length implied by the context. We use similarly $\mathbf{0}$ for the vector of zeros and $\mathbf{I}$ for the identity matrix. In some cases, model 0 is a submodel of each model $1, \ldots, K$, models $k = 1, \ldots, K-1$ are all submodels of model K,

and model K is assumed to be valid a priori. Models are solely the sources of candidate estimators $\hat{\theta}_k$, and all that matters are their MSE-related properties, $\mathrm{E}(\hat{\boldsymbol{\theta}})$ and $\mathbf{V} = \mathrm{Var}\,(\hat{\boldsymbol{\theta}})$, evaluated under the assumption that a specified model M^* is valid. These moments may depend on some unknown parameters. We assume that estimator $\hat{\theta}^*$ based on M^*, and $\hat{\theta}_K$ when $M_K = M^*$, is unbiased. The validity of any other model is neither assumed nor inferred. We never make the assumption that $\hat{\theta}_k$, $k < K$, would be unbiased if M_k were valid. We adhere to the frequentist paradigm, operating with sampling distributions, but we could switch to the Bayesian paradigm by working with posterior distributions instead.

We exploit the fact that some submodels of M^*, even if patently invalid, are useful for estimating *some* targets. In the problem of estimating a linear combination of regression parameters in ordinary regression with K covariates, we first reduce the list of candidates from 2^K models (and their estimators) to $K + 1$, and then reduce the problem further to a composition of only $\hat{\theta}_0$, based on the simplest model M_0, and $\hat{\theta}^*$. Then the only ps issue is the selection of M_0.

In the next section, we highlight the breakdown of established methods for model selection and averaging, with the exception of the focused information criterion (Claeskens and Hjort (2008)). Section 3 presents our general proposal and Section 4 gives details of estimating a linear predictor $\mathbf{x}\boldsymbol{\beta}$ in ordinary regression. Illustrations and examples of the method are presented in Section 5. Section 6 explores an extension to generalized linear models (GLM). The concluding section discusses the implications of the method on the everyday practice of statistics.

2. Motivating Examples

Suppose outcomes $\{y_{jh}\}$, $j = 1, \ldots, n_h$ and $h = 1, \ldots, H$, are generated by the model of analysis of variance with the standard assumptions of normality, independence and constant within-group variance $\sigma^2 > 0$. The obvious estimator of the expected outcome in group 1, μ_1, is its sample mean $\hat{\mu}_1 = \sum_j y_{j1}/n_1$. When n_1 is small the overall sample mean $\hat{\mu} = \sum_h \sum_j y_{jh}/n$, where $n = \sum_h n_h$, is a credible alternative. It is biased for μ_1, but $\mathrm{Var}\,(\hat{\mu}) = \sigma^2/n$ is much smaller than $\mathrm{Var}\,(\hat{\mu}_1) = \sigma^2/n_1$. Choosing $\hat{\mu}_1$ or $\hat{\mu}$, as in (1.1) with $K = 1$, may be inferior to $\tilde{\mu}_1 = (1 - c)\hat{\mu}_1 + c\hat{\mu}$ with a suitable constant c. Let $\Delta n_1 = 1/n_1 - 1/n$ and $\mu = \mathrm{E}(\hat{\mu})$. The optimal constant is $c^* = \Delta n_1/\{\Delta n_1 + (\mu_1 - \mu)^2/\sigma^2\}$. It has to be estimated, but $\tilde{\mu}_1(\hat{c}^*)$ is more efficient than the ps estimator for a wide range of values of $|\mu_1 - \mu|/\sigma$; see Longford (2008), Chapter 1. The constant c^* depends on n_1, so we combine $\hat{\mu}_k$ and $\hat{\mu}$ differently for one group than for another that has a different sample size. Model averaging (Kass and Raftery (1995), Hoeting et al. (1999), and Hansen (2007)) uses the same set of weights for the two targets.

For estimating σ^2, we proceed similarly. The common unbiased estimator of σ^2, denoted $\hat{\sigma}_1^2$, is based on the model with unrelated means μ_k. It is associated with the χ^2_{n-H} distribution. The estimator $\hat{\sigma}_0^2$ based on the submodel with $\mu_1 = \ldots = \mu_H$ has bias $(n-1)^{-1}\sum_h n_h(\mu_h - \mu)^2$, but its distribution has $H-1$ additional degrees of freedom. When $H - 1 \ll n - H$, we should combine the two estimators of σ^2 with a large weight given to $\hat{\sigma}_1^2$ because a few degrees of freedom gained are a poor trade for the likely bias of $\hat{\sigma}_0^2$.

Ordinary regression is commonly applied to compare two treatments for a medical or some other condition; $y_{jh} = \mu_h + x_{jh}\,\beta + \varepsilon_{jh}$, where $j = 1, \ldots, n_h$ and $h = 0, 1$, with $\varepsilon_{jh} \sim \mathcal{N}(0, \sigma^2)$ independently. The covariate X is a background variable, with its values x_{jh} not affected by the treatment assignment. We combine the contrast of the within-group means, $\Delta\bar{y} = \bar{y}_1 - \bar{y}_0$, based on the possibly false assumption that $\beta = 0$, with the ordinary least squares (OLS) estimator $\Delta\hat{\mu}$ of $\Delta\mu = \mu_1 - \mu_0$ as $\Delta\tilde{\mu} = (1-c)\Delta\hat{\mu} + c\Delta\bar{y}$ with a suitable coefficient c. When the within-group means $\bar{x}_0$ and $\bar{x}_1$ differ the optimal value of c is

$$c^* \;=\; \frac{1}{1 + (T_0 + T_1)\,\rho^2}\,,$$

where $T_h = \sum_j (x_{jh} - \bar{x}_h)^2$, $h = 0, 1$, and $\rho = \beta/\sigma$. When $\bar{x}_1 = \bar{x}_0$, $\Delta\hat{\mu} = \Delta\bar{y}$ and then the choice between $\Delta\hat{\mu}$ and $\bar{y}$, or how they are combined, is immaterial; X can be ignored. A connection with randomisation and post-observation design (Rosenbaum (2010)) is discussed in Supplementary Materials, Section A.

In summary, our disposition to the candidate models should be informed by the design (e.g., the sample sizes n_h) and the target of estimation. This exposes a profound weakness shared by model selection and model averaging.

3. Composition

For the composite estimator $\tilde{\theta}$ given by (1.2), $\mathrm{MSE}(\tilde{\theta}; \theta) = (\mathbf{b}^\top \mathbf{c})^2 + \mathbf{c}^\top \mathbf{V} \mathbf{c}$, where the vector of biases $\mathbf{b} = \mathrm{E}(\hat{\boldsymbol{\theta}}) - \theta\mathbf{1}$ and $\mathbf{V} = \mathrm{Var}\,(\hat{\boldsymbol{\theta}})$ are evaluated under an a priori specified valid model M^*, not necessarily one in the collection $\mathcal{M}$. Estimator $\hat{\theta}^*$ associated with M^* is unbiased, so $\hat{\mathbf{b}} = \hat{\boldsymbol{\theta}} - \hat{\theta}^*\mathbf{1}$ is unbiased for $\mathbf{b}$. If $\mathbf{V}$ is nonsingular, then $\mathrm{MSE}(\tilde{\theta}; \theta)$ attains its minimum for

$$\mathbf{c}^* \;=\; \frac{1}{\mathbf{1}^\top \left(\mathbf{V} + \mathbf{b}\mathbf{b}^\top\right)^{-1} \mathbf{1}} \left(\mathbf{V} + \mathbf{b}\mathbf{b}^\top\right)^{-1} \mathbf{1}\,,$$

derived by the method of Lagrange multipliers. If $\mathbf{V}$ is singular, we drop redundant estimators from $\hat{\boldsymbol{\theta}}$. Let $B_0 = \mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1}$, $B_1 = \mathbf{1}^\top \mathbf{V}^{-1} \mathbf{b}$, and $B_2 = \mathbf{b}^\top \mathbf{V}^{-1} \mathbf{b}$. The identity $(\mathbf{V} + \mathbf{b}\mathbf{b}^\top)^{-1} = \mathbf{V}^{-1} - (1 + B_2)^{-1} \mathbf{V}^{-1} \mathbf{b}\mathbf{b}^\top \mathbf{V}^{-1}$ implies the expressions

$$\mathbf{c}^* = \frac{1}{B_0\,(1+B_2) - B_1^2}\,\left\{(1+B_2)\,\mathbf{V}^{-1}\mathbf{1} - B_1\mathbf{V}^{-1}\mathbf{b}\right\},$$

$$\mathbf{c}^{*\top}\hat{\boldsymbol{\theta}} = \hat{\theta}^* + \frac{1}{B_0\,(1+B_2) - B_1^2}\,\left\{(1+B_2)\,\mathbf{1}^\top\mathbf{V}^{-1}\hat{\mathbf{b}} - B_1\mathbf{b}^\top\mathbf{V}^{-1}\hat{\mathbf{b}}\right\},$$

$$\mathrm{MSE}\left(\mathbf{c}^{*\top}\hat{\boldsymbol{\theta}};\theta\right) = \frac{1+B_2}{B_0\,(1+B_2) - B_1^2}\,. \tag{3.1}$$

The essence of our proposal is to substitute naive estimators for all terms in the expression for $\mathbf{c}^{*\top}\hat{\boldsymbol{\theta}}$. This results in the estimator

$$\tilde{\theta} \;=\; \hat{\theta}^* + \frac{\hat{B}_1}{\hat{B}_0\left(1+\hat{B}_2\right) - \hat{B}_1^2}\,. \tag{3.2}$$

The right-hand side of (3.1), denoted by $\mathrm{MSE}^\dagger$, understates the MSE of $\tilde{\theta}$. In the next section, we derive simple expressions for $\tilde{\theta}$ in a standard problem in ordinary regression that reduce the problem to a composition of $\hat{\theta}_0$ and $\hat{\theta}^*$. Estimation of MSE is addressed in Section 4.3.

4. Linear Predictor in Ordinary Regression

Borrowing the terminology from linear algebra, we refer to $\hat{\boldsymbol{\theta}}$ in (1.2) as the basis that generates a space of composite estimators. This space is not linear, nor is it a simplex, because negative coefficients c_k are permitted. A basis is called non-redundant if the space it generates becomes smaller after excluding any one of its elements. A basis is called complete for a collection of estimators, or the related models, if every estimator in the collection belongs to the space.

Suppose an ordinary regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, is valid and $\mathbf{X}$, of dimensions $n \times (K+1)$, has orthogonal columns and full rank $K+1$. Define $\mathbf{T} = \mathbf{X}^\top\mathbf{X}$ and $\mathbf{t} = \mathrm{diag}(\mathbf{T}^{-1}) = (t_0\,, t_1\,, \ldots, t_K)^\top$. We consider only submodels defined by constraining some elements of $\boldsymbol{\beta}$ to zero. For given $\mathbf{x}_0 = (x_{0,0}\,, \ldots, x_{0,K})$, we estimate the target $\theta = \mathbf{x}_0\,\boldsymbol{\beta}$ by OLS as $\hat{\theta} = \mathbf{x}_0\,\hat{\boldsymbol{\beta}}$, using a selection of models. Then we form a composition of these estimators.

The OLS estimator $\hat{\theta}^*$ based on model M^* with no constraints on $\boldsymbol{\beta}$ is unbiased and $\mathrm{Var}\,(\hat{\theta}^*) = \sigma^2\mathbf{x}_0\mathbf{T}^{-1}\mathbf{x}_0^\top$. For each excluded covariate X_k, the estimator is changed by subtracting $x_{0,k}\,\hat{\beta}_k$, incurring (additional) bias $-x_{0,k}\,\beta_k$, but reducing the variance by $\sigma^2 t_k\,x_{0,k}^2$. The biases may (partly) cancel, but the variance reductions accumulate. A complete non-redundant basis for the OLS estimators based on the 2^K models comprises $K+1$ estimators because every estimator of $\mathbf{x}_0\,\boldsymbol{\beta}$ can be combined from any basis estimator and the estimators of the elementary biases $-x_{0,k}\,\hat{\beta}_k$. We consider two such bases: A, the intercept-only model M_0 and models formed by adding one variable at a time, in arbitrary order, ending up with the unconstrained model $M_K = M^*$; B, M_0, and all the simple

regressions, comprising one covariate each. All these models have an intercept. The empty model M_0 can be replaced by the model $y = \varepsilon$, or by a model with one or several covariates that are included unconditionally.

The spaces generated by bases A and B coincide and can be described as all compositions of the estimators $\hat{\theta}_0$ and $\hat{\theta}^*$, or equivalently, as adjustments of one of them by a fraction (shrinkage) of the estimated bias of $\hat{\theta}_0$. This result is derived next, and such simple compositions are studied further in Section 4.3. Section B of Supplementary Materials discusses estimation of the variance σ^2.

4.1. Nested sequence of models (basis A)

Let $\hat{\boldsymbol{\theta}}$ be the set of OLS estimators of $\theta = \mathbf{x}_0 \boldsymbol{\beta}$ based on a nested sequence of ordinary regression models in basis A. Let $b_k = \mathrm{E}(\hat{\theta}_k) - \theta$ and $v_k = \mathrm{Var}\,(\hat{\theta}_k)$, and write the elementary biases as $\Delta b_k = b_{k-1} - b_k = -x_{0,k}\,\beta_k$, the variance inflations as $\Delta v_k = v_k - v_{k-1} = \sigma^2 t_k\, x_{0,k}^2$, and their ratios as $\rho_k = \Delta b_k / \Delta v_k$, $k = 1, \ldots, K$. The covariance of two estimators in $\hat{\boldsymbol{\theta}}$ is equal to the variance of the estimator based on the covariates that the two associated models have in common. Therefore the elements of $\mathbf{V} = \mathrm{Var}\,(\hat{\boldsymbol{\theta}})$ are $v_{hk} = v_{\min(h,k)}$, $0 \le h, k \le K$. The inverse of $\mathbf{V}$ is tridiagonal ($\hat{\boldsymbol{\theta}}$ is a Markov chain), with diagonal elements $u_{00} = 1/v_0 + 1/\Delta v_1$, $u_{kk} = 1/\Delta v_k + 1/\Delta v_{k+1}$ for $k = 1, \ldots, K-1$, and $u_{KK} = 1/\Delta v_K$, and $u_{k,k+1} = u_{k+1,k} = -1/\Delta v_k$ next to the diagonal. For example, for $K = 3$,

$$\mathbf{V} = \begin{pmatrix} v_0 & v_0 & v_0 & v_0 \\ v_0 & v_1 & v_1 & v_1 \\ v_0 & v_1 & v_2 & v_2 \\ v_0 & v_1 & v_2 & v_3 \end{pmatrix}$$

$$\mathbf{V}^{-1} = \begin{pmatrix} \frac{1}{v_0} + \frac{1}{\Delta v_1} & -\frac{1}{\Delta v_1} & 0 & 0 \\ -\frac{1}{\Delta v_1} & \frac{1}{\Delta v_1} + \frac{1}{\Delta v_2} & -\frac{1}{\Delta v_2} & 0 \\ 0 & -\frac{1}{\Delta v_2} & \frac{1}{\Delta v_2} + \frac{1}{\Delta v_3} & -\frac{1}{\Delta v_3} \\ 0 & 0 & -\frac{1}{\Delta v_3} & \frac{1}{\Delta v_3} \end{pmatrix}.$$

The identities $\mathbf{V}^{-1}\mathbf{1} = (1/v_0\,, 0, \ldots, 0)^\top$ and $\mathbf{V}^{-1}\mathbf{b} = (\rho_1 + b_0/v_0\,, \rho_2 - \rho_1\,, \ldots, \rho_K - \rho_{K-1}\,, -\rho_K\,)^\top$ imply that $B_0 = 1/v_0$, $B_1 = b_0/v_0$ and $B_2 = R + b_0^2/v_0$, where $R = r_1 + \cdots + r_K$ and $r_k = \Delta b_k\, \rho_k = \Delta b_k^2 / \Delta v_k$. Hence

$$\mathbf{c}^* \;=\; \frac{1}{1+R}\Big\{1 + R - b_0\,\rho_1\,,\; b_0\,(\rho_1 - \rho_2)\;,\;\; \ldots\;,\; b_0\,(\rho_{K-1} - \rho_K)\;,\; b_0\,\rho_K\Big\}$$

and, since $\hat{b}_0 = \hat{\theta}_0 - \hat{\theta}_K$, the composite estimator $c^{*\top}\hat{\boldsymbol{\theta}}$ is

$$\tilde{\theta} \;=\; \hat{\theta}_K + \frac{\hat{b}_0}{1+R} \;\;=\;\; \hat{\theta}_0 - \frac{R\hat{b}_0}{1+R} \;\;=\;\; \frac{\hat{\theta}_0 + R\hat{\theta}_K}{1+R}\,. \tag{4.1}$$

In practice, R has to be estimated. The consequent inflation of MSE can be explored by simulations. In basis A, $r_k = \beta_k^2/(t_k\sigma^2)$ and r_k does not depend on $\mathbf{x}_0$, although the case $x_{0,k} = 0$ has to be treated separately. Therefore R does not depend on $\mathbf{x}_0$ and neither do the weights assigned to $\hat{\theta}_0$ and $\hat{\theta}_K$. By dropping estimator $\hat{\theta}_k$, $0 < k < K$, from basis A, the expression for $\tilde{\theta}$ is altered only by replacing $r_k + r_{k+1}$ in R with $(\Delta b_k + \Delta b_{k+1})^2/(\Delta v_k + \Delta v_{k+1})$, and then $x_{0,k}$ and $x_{0,k+1}$ cancel out only when $x_{0,k}\, x_{0,k+1} = 0$. By dropping $\hat{\theta}_k$, R is reduced by

$$\frac{\Delta v_k\, \Delta v_{k+1}}{\Delta v_k + \Delta v_{k+1}}\ (\rho_k - \rho_{k+1})^2\ \geq\ 0\,.$$

In an incomplete basis, R and $\tilde{\theta}$ may depend on $\mathbf{x}_0$.

We explore $\mathrm{MSE}(\tilde{\theta};\theta)$ as a function of R:

$$\mathrm{MSE}^{\circ}\Big(\tilde{\theta};\theta\Big)\ =\ v_0 + \frac{R^2(v_K - v_0)}{(1+R)^2} + \frac{b_0^2}{(1+R)^2}\,; \tag{4.2}$$

the circle $^{\circ}$ indicates that the evaluation ignores the uncertainty about R. For very large R, MSE° is close to v_K; for very small R it is close to $\mathrm{MSE}(\hat{\theta}_0\,;\theta) = v_0 + b_0^2$. The minimum of $\mathrm{MSE}^{\circ}(\tilde{\theta};\theta)$ is attained for $R^* = b_0^2/(v_K - v_0)$ and the minimum value is

$$\mathrm{MSE}^{\circ}\Big(\tilde{\theta}^*;\theta\Big)\ =\ v_0 + \frac{b_0^2}{1 + b_0^2/(v_K - v_0)}\,. \tag{4.3}$$

This coincides with the optimal composition of $\hat{\theta}_0$ and $\hat{\theta}_K$, making all the intermediate models $1,\ldots,K-1$ redundant. Note that $R^* = R^*(\theta)$ is target-specific.

By substituting B_0, B_1, and B_2 in the MSE in (3.1), we obtain the identity

$$\mathrm{MSE}^{\dagger}\Big(\tilde{\theta};\theta\Big)\ =\ v_0 + \frac{b_0^2}{1+R}\,,$$

so $\mathrm{MSE}^{\dagger}$ and MSE° have the same form, $v_0 + b_0^2/(1+m)$, with $m = R$ and $m = b_0^2/(v_K - v_0)$, respectively. Thus, $\mathrm{MSE}^{\dagger}$ becomes smaller as more variables are considered, whereas MSE° is inflated if v_K is increased. Neither MSE is uniformly smaller than the other but $\mathrm{MSE}^{\dagger} = \mathrm{MSE}^{\circ} = v_0$ when $b_0 = 0$.

If $\hat{\theta}_0$ is dropped from the basis, $\hat{\theta}_1$ takes over its role as the simplest estimator. Then R is reduced by r_1, and $\hat{\theta}_1$ and $\hat{b}_1$ are substituted in (4.1) for $\hat{\theta}_0$ and $\hat{b}_0$, respectively. The expression for MSE° in (4.2) implies that this is worthwhile when $b_1^2 \ll b_0^2$ and $r_1 \gg 0$. Comparison of the expressions (4.3) for the original and reduced bases yields the condition

$$b_1^2\ <\ \frac{(v_k - v_1)^2}{(v_k - v_0)^2}\ \big(v_k - v_0 + b_0^2\big) - (v_k - v_1) \tag{4.4}$$

for usefully dropping $\hat{\theta}_0$. This rule can be applied by using estimates for b_0^2 and b_1^2. Note that dropping or not dropping $\hat{\theta}_0$ amounts to post-selection.

We do not regard selection of M^* as a ps issue, because a data-based selection normally entails uncertainty. Nevertheless, if submodel $M_{K'}$ of M^* is selected, eliminating models $M_{K'+1}$, ..., M_K from the basis, MSE$^\circ$ in (4.3) is reduced because $v_{K'} < v_K$. In contrast, MSE† in (3.1) is increased, as it is a minimum over a smaller space of estimators.

4.2. Simple regressions (basis B)

Basis B, with the empty and simple regressions, generates the same space of compositions as basis A, so the same estimator with minimum MSE† is obtained for the two bases. The proof that parallels the derivations in Section 4.1 is given in Supplementary Materials, Section C.

A covariate h can be included in M_0. Then every other model in the basis has to be supplemented with covariate h. Let $\nabla v_h = v_h - v_0$ and $b'_0 = b_0 + x_{0,h}\beta_h$; b'_0 is the bias of the 'new' estimator $\hat{\theta}_0$. By comparing the expressions (4.3) for the original and the new bases, with v_K replaced by v^*, we obtain the condition

$$b_0'^{\,2} < \frac{b_0^2 \nabla v_h^2}{(v^* - v_0)^2} + \frac{\nabla v_h^2 - 2b_0^2 \nabla v_h}{v^* - v_0} + b_0^2 - \nabla v_h$$

for when moving covariate h to M_0 is worthwhile. As a (quadratic) function of ∇v_h, the right-hand side attains its minimum at $\nabla v_h^* = \frac{1}{2}(v^* - v_0)(v^* - v_0 + 2b_0^2)/(v^* - v_0 + b_0^2)$, and the minimum, $-\frac{1}{4}(v^* - v_0)^2/(v^* - v_0 + b_0^2)$, is negative. Without this derivation, an attractive choice for the revised model 0 might be the regression on one or several covariates known to be strongly associated with the outcome. In Section 5.1 we give an example in which this is a poor choice.

For either complete basis, A or B with $K \geq 1$, the average $\hat{R}/K$ has a noncentral F distribution with K and $n-K$ degrees of freedom and noncentrality parameter $\lambda = R/K$. For fixed degrees of freedom, both the mean and variance of the noncentral F are increasing functions of λ. The ratio of the variance and squared mean,

$$2\,\frac{(K+\lambda)^2 + (K+2\lambda)(n-K-2)}{(K+\lambda)^2(n-K-4)},$$

is decreasing in λ. It converges to $2/(n-K-4)$ for $\lambda \to +\infty$, so it is small for large λ when $n \gg K$. In fact, the ratio is smaller than a positive constant c even when $\lambda = 0$ for $K > 2(n-2)/\{c(n-K-4)\} > 2/c$. Therefore, a large value of $\hat{R}$ is very likely a consequence of large λ and implies that composition cannot improve much on $\hat{\theta}^*$. With a basis reduced to $H \leq K$ estimators, $\hat{R}/H$ has a noncentral $F_{H,n-H}$ distribution, but the two reduced-basis estimators differ.

4.3. Simple composition

Estimation with bases A and B reduces to composing two estimators, based on the valid model M^* and the smallest submodel M_0. Since $\hat{\theta}_0$ and the estimator

of its bias, $\hat{b}_0 = \hat{\theta}_0 - \hat{\theta}^*$, are independent, we need to estimate efficiently only the transformed bias $b_0 R/(1+R) = b_0^3(b_0^2+v^*-v_0)$. We consider shrinkage estimators $\tilde{b}_0(C) = C\hat{b}_0\,\hat{R}/(1+\hat{R})$. No generality is lost by assuming that $s = v^* - v_0 = 1$; otherwise we work with $b_0/\sqrt{s}$.

For sufficiently large sample size n, when the difference between the t_{n-K-1} and $\mathcal{N}(0,1)$ distributions and the uncertainty about σ^2 are negligible, we evaluate the empirical MSE of $\tilde{b}_0/\sigma$ on a fine two-way grid of values of C and b_0 and identify the coefficient $C(b_0)$ for which the empirical MSE is minimized. For smaller n, we simulate the values of $\hat{b}_0/\hat{\sigma}$ from the appropriate t distribution. The results are summarized in panels A and B of Figure 1 by the functions $C(b_0)$ and the corresponding empirical root-MSEs (rMSE) for $\mathcal{N}(0,1)$ and a few t distributions. The optimal shrinkage C increases with b_0/σ, converging at a slow rate to 1.0 as $|b_0| \to +\infty$. The rMSEs increase, have a flat maximum, and then slowly decrease toward 1.0. For large n, MSE < 1.0 up to about $b_0 = 1.6\sigma$. Composition is not useful for $b_0 > 1.6$ because $\mathrm{Var}\,(\hat{b}_0/\sigma) > 1$; choosing $\hat{\theta}^*$ unconditionally is then prefered. When $\hat{b}_0/\hat{\sigma}$ has a noncentral t_h distribution, $\mathrm{Var}\,(\hat{b}_0/\hat{\sigma}) \geq h/(h-2) > 1$, but the MSE is also greater, so the borderline b_0^* up to which composition is useful is about the same. However, b_0/σ is not known.

Panels C and D compare the shrinkage coefficients and the rMSE functions for the t distributions by relating them to $\mathcal{N}(0,1)$. Panel C is curtailed because of the effects of coarse optimization. For a given scaled bias b_0/σ, the optimal shrinkage increases with the degrees of freedom and converges to the shrinkage that is optimal with $\mathcal{N}(0,1)$. The minimum rMSE decreases with the degrees of freedom toward its 'normal' counterpart. The ratio $\sqrt{\mathrm{MSE}/\mathrm{MSE}^\circ}$ can be regarded as a measure of deception caused by ignoring the uncertainty about R. Its reciprocal is plotted by dots in panel D. The deception is greatest at $b_0/\sigma \doteq \pm 2.3$, equal to 1/0.84=1.19. Composition is not useful for $|b_0|/\sigma > 1.6$, so in practice the upper bound on deception is lower.

5. Examples

We explored composite estimation of a cubic regression for observations made at points $x = 1, 2, \ldots, 40$. The targets were the fitted values in the range $-10 < x < 50$, including some extrapolation. The basis for composite estimation was formed by the sample mean (model 0, or A) and estimators based on the linear (L), quadratic (Q), and cubic regressions on x (model $K = 3$ or C). We simulated datasets from the ordinary regression model $y = 25 + x_1' + 0.01x_2' - 0.0012x_3' + \varepsilon$ with $\mathrm{Var}\,(\varepsilon) = 100$, $x_0' = 1$, $x_1' = x - \bar{x}$, $x_2' = (x-\bar{x})^2 - (n^2-1)/12$, and $x_3' = (x-\bar{x})^3 - (3n^2-7)(x-\bar{x})/20$, where $\bar{x} = 20.5$. These polynomials are orthogonal for $x = 1, \ldots, n = 40$. For each of 1,000 replications, we fitted the four basis models and evaluated the composite estimators for the complete basis

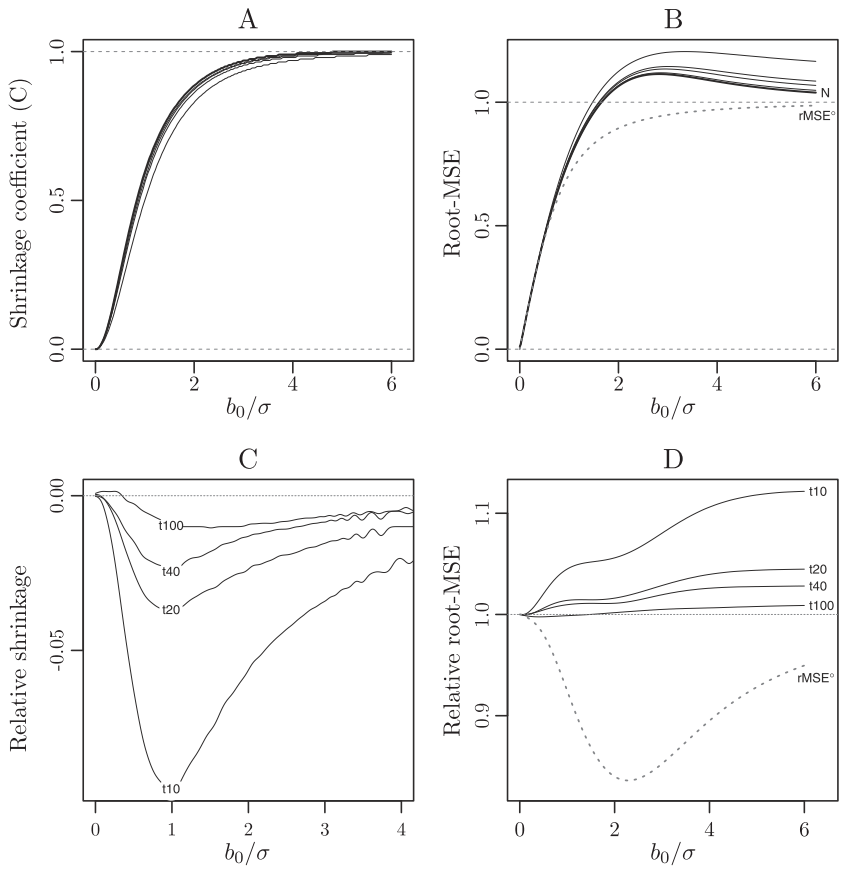

Figure 1. Optimal shrinkage coefficients and the corresponding rMSEs of the simple composition with $v^* - v_0 = 1$, as functions of the bias b_0 for the normal and a selection of t distributions. The relative shrinkage in panel C is defined as the difference of the shrinkage coefficients for a t and the $\mathcal{N}(0,1)$ distribution. The relative rMSE in panel D is defined as the ratio of the rMSEs for a t and the $\mathcal{N}(0,1)$ distribution.

(ALQC), with $\hat{\theta}_0$ dropped (LQC), and the simple compositions AC, LC, and QC. The ps estimator (psE) was defined by testing the null-hypothesis for the cubic coefficient, using the conventional test size of 0.05.

Estimator ALQC was slightly more efficient than estimator C around $x = 20.5$ and less efficient around $x = 6$ and 35. Estimator AC was more efficient than C and ALQC only around $x = 20.5$ and for extrapolation ($x < 1$ and $x > 40$). Except for a few narrow intervals of x, psE was less efficient than both C and Q, as were estimators defined by other established model selection (information) criteria. Submodel Q was selected in only 13.4% of replications, and yet was more efficient than ALQC, AC, C, and psE in a wide range around $x = 31$. We

regard both ALQC and AC as failures because they are less efficient than Q or C in wide ranges of values of x. But psE has not much to commend either.

A large value of r_1 suggests that estimator A should be dropped from the basis, although (4.4) is a more appropriate criterion. Composite estimators LQC and LC were uniformly more efficient than C, and LC was also superior to Q, except for $x \in (31.5, 35)$. Estimator LQC was more efficient than LC only by a narrow margin for $x < 3$ and $x \in (11, 19)$. Apart from intervals $(2, 9)$ and $(34, 36)$, the estimator with shrinkage applied to LC (LCshr) was more efficient than LC. The gain in efficiency was small throughout, except for $x > 40$, where the reduction of rMSE was by nearly 4%. The shrinkage coefficient was based on the 90th percentile of t_{36}. For higher percentiles, the gains would be smaller but positive throughout.

Selection of the basis is addressed in Figure 2, where the percentages of the decisions to drop $\hat{\theta}_0$ from the basis (and use estimator LC instead of AC), and to drop also $\hat{\theta}_1$ (and use QC) are plotted. The shaded region represents the choice of LC, and the regions above and below are for AC and QC, respectively. The choices were based on the inequality in (4.4). The estimator with minimum (empirical) rMSE is marked by gray horizontal dashes drawn at the heights -1 (to use QC), 51 (LC), and 102 (AC). Thick segments indicate 'clear winners' (e.g., AC for $-10 < x < -2.5$), for which the second best choice has rMSE greater at least 1.025 times. Thinner segments are used for 'narrow winners' (e.g., QC for $-2 < x < 0.2$) and are accompanied by the 'narrow losers' marked by even thinner segments (e.g., QC for $0.3 < x < 2.2$). Their rMSEs are greater than for the corresponding winners less than 1.025 times. Thus, LC is a winner or a narrow loser throughout the range $(-2.5, 50)$, except for a narrow interval around 20.5, where AC is the winner. Estimator QC is not a clear winner anywhere in $(-10, 50)$ and AC is a clear loser in much of the range. Inspection of $\hat{\mathbf{r}}$ would make us choose LC, a good choice overall, but not uniformly the best.

The pointwise selected compositions do not form a cubic function, but the fit is pointwise more efficient than any cubic function. This indicates both a problem with interpreting the fit and a hindrance to efficient estimation caused by insisting on a single (basis) estimator or a universal composition (model averaging). Owing to symmetry of x, the sample mean is the most efficient of the basis estimators for $x = 20.5$ ($x' = 0$); the interval of x around 20.5 where A is efficient is narrow.

In all replications, $\hat{r}_1/\hat{R} > 0.67$ — $\hat{r}_1$ dominated in $\hat{\mathbf{r}}$. The values of $\hat{R}$ had mean 61.7, median 58.7 and standard deviation (sd) 21.7; the mean of $\hat{r}_2 + \hat{r}_3$ was only 2.66, the median was 3.54 and sd was 3.33. One might conclude from a single realization that $\hat{\theta}_0$ is not useful, except at $x = 20.5$, and choose estimator LC. Equation (4.4) credits $\hat{\theta}_0$ for some values of x, although sometimes incorrectly.

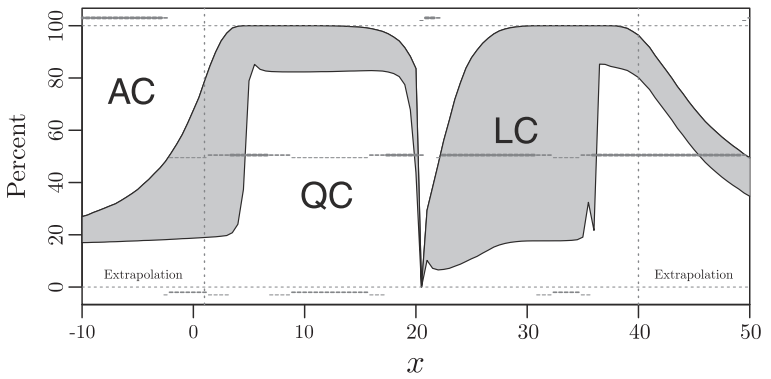

Figure 2. Percentages of decisions to drop from the basis estimator $\hat{\theta}_0$ (and use LC) and also $\hat{\theta}_1$ (and use QC), as functions of the target x. Based on simulations with 1,000 replications.

Any model selection or averaging would disqualify $\hat{\theta}_0$. Clearly, good model fit and efficiency are distinct criteria that should not be confused.

5.1. Prostate cancer data

We reanalyse the data on prostate cancer studied originally by Stamey et al. (1989) and used by Tibshirani (1996) to illustrate lasso. The dataset is available in `R` (R Core Team (2013)). The covariates are the cancer volume (lcavol), prostate weight (lweight), age in years, the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), capsular penetration (lcp), Gleason score (gleason) and the percentage of Gleason scores 4 and 5 (pgg45). The outcome variable, prostate specific antigen (lpsa), has values in the range $(-0.43, 5.58)$; their sample mean and sd are 2.48 (1.15). The variables lcavol, lweight, lbph, lcp, and lpsa are log-transformed. The variables lbph, lcp, and pgg45 are continuous, but their respective minima of -1.386, -1.386, and zero occur frequently, for 43, 45, and 35 of the 97 cases. The binary variable svi has the frequencies 76 (value 0) and 21 (1). All but six cases have values of gleason equal to 6 (35 cases) or 7 (56 cases).

We studied prediction for ages $40-80$ years and a fixed set of values of the covariates listed in the first row of Table 1. Tibshirani (1996) selected the covariates lcavol, lweight, and svi by both subset selection and lasso. We applied ordinary regression with intercept only (M0), all the covariates (M1), and the covariates selected by lasso (MT). The compositions of M0 and M1 and of MT and M1 are denoted by C0 and CT, respectively. Further, Cp and Cpt refer to the respective versions of C0 and CT averaged over 1,000 replicate sets of plausible values. The estimators are assessed by their predictions and estimated rMSEs in

Table 1. Values of the covariates used in prediction ('**?**' indicates the target). Prostate cancer data.

	Variable								
	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
Value	1.35	3.65	40−80	-1.40	0	-1.40	7.00	0.0	**?**
Mean	1.35	3.65	63.9	0.10	0.22	-0.18	6.75	24.4	2.48
St. dev.	1.18	0.50	7.45	1.45	0.41	1.40	0.72	28.2	1.15

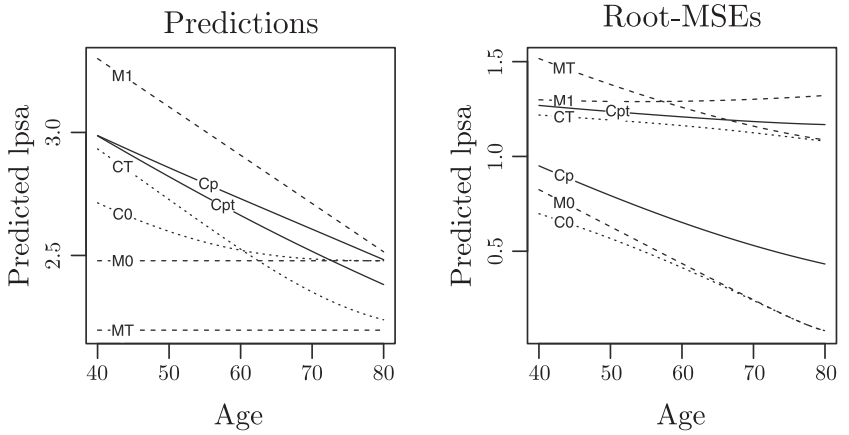

Figure 3. Prediction of lpsa for ages 40 – 80 years, with covariate values listed in the first row of Table 1. Prostate cancer data.

Notes: M denotes OLS and C composite estimators; 0 stands for the intercept-only, 1 for the unconstrained, and T for the model selected by lasso; p indicates that the estimate is based on 1,000 replicate sets plausible values.

Figure 3, drawn as functions of age. With model M1 we obtain a (decreasing) linear function of age. Compositions C0 and CT yield nonlinear functions of age.

The biases (and rMSEs) of the predictions based on OLS were estimated with reference to model M1. The estimates ignore the uncertainty about R^*, which was in fact very modest throughout the range of age, because $\hat{R}^*$ was quite small. Even with a considerable leeway for error, estimators M0 and C0 were uniformly far superior to CT. The estimates of rMSE obtained by bootstrap (not drawn in the diagram) decreased from 1.06 for 40 years to 0.98 for 80 years of age, still smaller than the estimated rMSE of CT; they decreased from 1.22 to 1.08 for rMSE$^\circ$ and from 1.37 to 1.26 for the bootstrap overestimate.

Setting M_0 to the model with the variables identified by lasso is not useful because prediction is made at the sample means of lcavol and lweight, and at a value of svi not far away from its mean, 0.22. With these variables in M_0, we import a lot of sampling variation for next to no bias reduction. The small

contributions from the 'unimportant' variables happen to add up and generate a prediction strongly related to age. The estimators of rMSE based on 1,000 sets of plausible values of $\boldsymbol{\beta}$ and σ^2 are represented in Figure 3 by solid lines marked Cp (for composition of M0 and M1) and Cpt (MT and M1). Their means are plotted in the left-hand panel. The rMSE for Cpt is only slightly, though uniformly, greater than rMSE°. In contrast, much greater inflation takes place for C0, especially for older ages, although it does not affect our assessment that C0 is much more efficient than CT. In conclusion, the trivial model with no covariates should be retained in the basis of the composition (as M_0), even if it were known that lcavol, lweight, and svi are important predictors. In Supplementary Materials, Section D, we relate composite estimation to propensity matching analysis.

6. Composition of GLM Estimators

We explore composition outside the confines of ordinary regression by estimating the linear predictor $\theta = \mathbf{x}_0\boldsymbol{\beta}$ in logistic regression for binary outcomes. We note first that efficient estimation of $\mathbf{x}_0\boldsymbol{\beta}$ and of its inverse-link transform $g^{-1}(\mathbf{x}_0\boldsymbol{\beta})$ are different tasks when the link function g is nonlinear, as is the case when some of the probabilities fitted by logistic regression are extreme.

The reduction of multi-estimator compositions to simple compositions does not carry over to GLM. The identity $\mathrm{Cov}\,(\hat{\theta}_1\,,\hat{\theta}_2) = \mathrm{Var}\,(\hat{\theta}_1)$ for estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ based on respective models M_1 and M_2 holds only when M_1 is a submodel of M_2, and M_2 is valid. The latter clause is not required with the identity link, as in ordinary regression. An extreme departure arises when the iterative weights in two (invalid) models differ substantially. For example, the weights in a logistic regression may be clustered around 0.25 for one model and be much smaller for another. Balance of the covariates is a property not only of $\mathbf{X}$ but also of the model, through the weights it implies. The model-based estimators do not form a basis as defined in Section 4. Equation (3.2) applies generally, but the intermediate estimators do not cancel out as they do in (4.1).

Nevertheless, we may consider the composition of estimators $\hat{\theta}_0$ and $\hat{\theta}^*$. We give an example of a simple composition that is more efficient than both basis estimators. We generated a 500×3 regression matrix with the intercept column, a random sample from the convolution of $\mathcal{N}(0,1)$ rounded to integers with the uniform distribution on $(0,1)$, and a column comprising 100 repeats of the regular sequence (1, 2, 3, 4, 5). This (fixed) matrix $\mathbf{X}$ was used in 10,000 replications of generating binary outcomes according to the logistic regression with $\boldsymbol{\beta} = (-1, 0.4, -0.25)^\top$. We estimated the linear predictors for $\mathbf{x}_0 = (1, 0, k)$, $k = 1, \ldots, 5$, using the models with intercept only (M_0), with the first two columns of $\mathbf{X}$ (model M_1), and with the entire matrix $\mathbf{X}$ (model M_2). The covariate X_2 had the range $(-2.68, 3.42)$, mean 0.40, median 0.425, and 0 at the percentile 36.6.

Table 2. Simulation of linear prediction in logistic regression; $\beta_0 = 1$, $\beta_1 = 0$; 10,000 replications.

	β_2				
	1	2	3	4	5
$\hat{\theta}_2 = \hat{\theta}^*$ — estimator with M_2					
Bias	-0.012	-0.014	-0.016	-0.018	-0.021
rMSE	0.204	0.155	0.147	0.185	0.249
rMSE°	0.200	0.152	0.145	0.184	0.249
$\tilde{\theta}_{12}$ — composition of $\hat{\theta}_1$ and $\hat{\theta}_2$					
Bias	−0.055	−0.031	0.001	0.023	0.047
rMSE	0.214	0.158	0.145	0.185	0.254
rMSE°	0.193	0.150	0.144	0.178	0.236
E($\hat{c}$)	0.127	0.114	0.397	0.171	0.157
$\tilde{\theta}_{02}$ — composition of $\hat{\theta}_0$ and $\hat{\theta}_2$					
Bias	−0.067	−0.006	0.010	0.019	0.039
rMSE	0.196	0.136	0.145	0.185	0.254
rMSE°	0.169	0.130	0.143	0.180	0.240
E($\hat{c}$)	0.406	0.662	0.107	0.079	0.088

The values of the target, $-1.25, -1.70, \ldots, -2.25$, correspond to probabilities 0.223, 0.182, 0.148, 0.119 and 0.095.

We formed simple compositions of estimators $\hat{\theta}_0$ and $\hat{\theta}_2$ and of $\hat{\theta}_1$ and $\hat{\theta}_2$, based on the estimated variance matrices of the estimators of $\mathbf{x}_0\boldsymbol{\beta}$ and their biases, assuming that $\hat{\theta}_2 = \mathbf{x}_0\hat{\boldsymbol{\beta}}_2$ is unbiased. The empirical biases and rMSEs are listed in Table 2. The analytical rMSEs in the table are the averages of the values of rMSE° based on each replicate fit. Note that the variation of the iterative weights across the replications is a source of uncertainty additional to that in ordinary regression. The values of the average shrinkage coefficient c are listed in the fourth row of each block for a composite estimator.

The table shows that compositions $\tilde{\theta}_{02}$ and $\tilde{\theta}_{12}$ are not useful for $\beta_2 > 3$, but $\tilde{\theta}_{02}$ is more efficient than $\hat{\theta}_2$ for $\beta_2 = 1$ and 2. We would be deceived by rMSE°, especially for $\beta_2 = 1$. Estimators $\hat{\theta}_0$ and $\hat{\theta}_1$ are not competitive, except for narrow ranges where their bias is small; details are omitted. The shrinkage coefficient is on average very large (0.662) for $\tilde{\theta}_{02}$ at $\beta_2 = 2$. Its sd is 0.260, so the coefficient is rarely small. For prediction at $\mathbf{x}_0 = (1, -3, k)$, $k = 1, \ldots, 5$, $\mathrm{rMSE}(\tilde{\theta}_{12}; \theta) < \mathrm{rMSE}(\tilde{\theta}_{02}; \theta)$ in all five cases, but $\mathrm{rMSE}(\tilde{\theta}_{12}; \theta)$ differs from $\mathrm{rMSE}(\hat{\theta}_2; \theta)$ by less than 1%. The composite estimators entail much more deception as the targets are much smaller probabilities, decreasing from 0.08 to 0.03. An application is presented in Supplementary Materials, Section E.

7. Conclusion

We summarize our alternative to model selection by two key points related to model-based estimation: efficiency in finite samples is not equivalent to the combination of inferred validity and parsimony, and composition has a greater potential than selecting one of the models. This potential is relatively easy to realize in standard ordinary regression problems. Estimation should not be an eliminatory contest of models, but a cooperative effort of estimators, with small MSE (or a similar criterion) as the sole arbiter of quality. We re-interpret the well known quip of Box (1976) that 'All models are wrong, but some are useful.' by pointing to the value of some grossly invalid models in a composition.

We defined a basis of composition as a finite set of model-based estimators of a quantity (a function of parameters). A basis can be formed by estimators that are not linked to any models. For example, the estimators may be defined by alternative methods, such as (full, restricted or penalized) maximum likelihood, moment matching and a non-parametric method, or different adjustments of the sampling weights (and equal weights) in survey analysis. For an application of this idea to mixed models, see Longford (2015).

In ordinary regression, composite estimators offer a dramatic reduction of the dimensionality of the problem, from 2^K to K, where K is the number of covariates. Further reduction is afforded by studying the quantities $\hat{r}_k = \Delta\hat{b}_k^2/\Delta\hat{v}_k$ and reducing model selection to specifying the simplest basis model. Our estimator of choice is a composition of this and the valid (the most complex) model M^*. It is advantageous to define a more parsimonious model M^*, but its validity is imperative. The outcome of the analysis is an estimate, not a model to be adopted, so any scientific problem has to be converted to one or several sets of quantities to be estimated. Relations among these estimators usually have no inferential meaning because they may be based on different (collections of) models. We suggest to study relations among the underlying quantities.

Our method is not suitable for interpreting the results, that is, to infer the plausible structure underlying the data from the analytical form of the estimator. However, when some elements of the structure are characterized by numeric quantities, their efficient estimation is paramount; then our approach is relevant. For other formats of inferential statements, such as confidence intervals, our approach is not immediately adaptable. For example, $\mu+c\sigma$ may be estimated more efficiently than by standard methods, but the distribution of such a confidence limit can only be approximated. In contrast, it is known with precision when based on the established estimators. Minimising MSE also implies a disregard for any constraints on the targets, even when all the basis estimators satisfy them.

We do not have a closed-form expression for the MSE of the composition or for its unbiased estimator. A closed-form expression, MSE° in (4.2), is available for simple composition in ordinary regression, assuming that the sum of

scaled squared biases, R, is known. The naive estimator of MSE$^\circ$ is too optimistic because it ignores the uncertainty due to estimating R. The extent of this underestimation can be explored by simulations. The multiplicative bias (deception) is not greater than 1.19. The bias is close to this bound when R is large, and composition is then ineffective — it is less efficient than the valid-model estimator $\hat{\theta}^*$. Therefore, the level of deception is in practice lower than 1.19. However, this statement ignores the ps nature of estimation, arising from the choice between composition and the valid model.

Bootstrap, Efron and Tibshirani (1993) and Davison and Hinkley (1997), and crossvalidation, Picard and Cook (1984) and Shao (1993), are not suitable for MSE estimation because composition is conditioned on the (joint) distribution of covariates, which cannot be held fixed in replications. Also, a different composition is optimal for a subset of the data because it has a different balance of the bias (unchanged) and variance (increased) than the original dataset.

Selection from a finite set of options is an operation frequently encountered in practice (Lindley (1985)) but the calculus used in everyday statistical evaluations is more amenable to linear operations, such as composition, and its properties are easier to explore. In particular, composition is a smooth (differentiable) operation, whereas selection is discontinuous.

Models used in the analysis of experiments and with post-observation design tend to fit poorly, because they do not involve any background variables. Models that fit better may yield less efficient estimators and, ironically, raise concerns about model validity. Composition reduces the divide between the analysis of experiments, in which the (joint) distribution of the covariates is controlled, and observational studies in which it is not. It leads to a simple analysis of treatment effects when the within-group distributions are (nearly) identical, both when this is arranged by design and when it happens to arise. Balance with respect to a subset of the covariates is effectively exploited by composition. Matched-group analysis has some important advantages over modelling for estimating treatment effects in observational studies, but composition reduces this advantage by involving invalid models.

Supplementary Materials

The online supplementary materials contain five sections, dealing with post-observation design (Section A), composite estimation of the residual variance in ordinary regression (B), composite estimation with basis B (C), propensity matching with the prostate cancer data analysed in Section 5.1 (D), and an application of composition to GLM contrasted with matched-pairs analysis (E).

Acknowledgements

Comments on an earlier version of the manuscript made by Aleix Ruiz de Villa and his encouragement are acknowledged. Anonymous referees were instrumental in improvements on the earlier versions of the manuscript.

References

Box, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71**, 791-799.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference. A Practical Information-Theoretical Approach*. 2nd edition. Springer-Verlag, New York.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98**, 900-945.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, New York.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications.* Cambridge University Press, Cambridge, UK.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Efron, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109**, 991-1007.

Efron, B. and Tibshirani, R. (1993). *An Introduction to Bootstrap.* Chapman and Hall, New York.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Hansen, B. E. (2007). Notes and comments. Least squares model averaging. *Econometrica* **75**, 1175-1180.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky C. T. (1999). Bayesian model averaging: a tutorial. *Statist. Sci.* **14**, 382-401.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors and model uncertainty. *J. Amer. Statist. Assoc.* **90**, 773-795.

Li, Q., and Shao, J. (2015). Regularizing lasso: a consistent variable selection method. *Statist. Sinica* **25**, 975-992.

Liang, H., Zou, G., Wan, T. K. and Zhang, X. (2011). Optimal weight choice for frequentist model averaging estimators. *J. Amer. Statist. Assoc.* **106**, 1053-1066.

Lindley, D. V. (1985). *Making Decisions.* Wiley, Chichester, UK.

Longford, N. T. (2008). *Studying Human Populations. An Advanced Course in Statistics.* Springer-Verlag, New York.

Longford, N. T. (2012). 'Which model?' is the wrong question. *Statist. Neerlandica* **66**, 237-252.

Longford, N. T. (2015). On the inefficiency of the restricted maximum likelihood estimator. *Statist. Neerlandica* **69**, 171-196.

Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *J. Amer. Statist. Assoc.* **79**, 575-583.

R Core Team (2013). R: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rolling, A. C. and Yang, Y. (2014). Model selection for estimating treatment effects. *J. Roy. Statist. Soc. Ser. B* **76**, 749-769.

Rosenbaum, P. R. (2010). *Design of Observational Studies.* Springer-Verlag, New York.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, ii: Radical prostatectomy treated patients. *J. Urology* **16**, 1076-1083.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

SNTL and Imperial College, London, UK

E-mail: sntlnick@sntl.co.uk

(Received June 2015; accepted May 2016)