Journal of Productivity Analysis (2019) 51:91–103
https://doi.org/10.1007/s11123-019-00547-8

Model averaging estimators for the stochastic frontier model

Christopher F. Parmeter[1] · Alan T. K. Wan[2] · Xinyu Zhang[3]

Published online: 4 April 2019

Abstract
Model uncertainty is a prominent feature in many applied settings. This is certainty true in the efficiency analysis realm where concerns over the proper distributional specification of the error components of a stochastic frontier model is, generally, still open along with which variables influence inefficiency. Given the concern over the impact that model uncertainty is likely to have on the stochastic frontier model in practice, the present research proposes two distinct model averaging estimators, one which averages over nested classes of inefficiency distributions and another that has the ability to average over distinct distributions of inefficiency. Both of these estimators are shown to produce optimal weights when the aim is to uncover conditional inefficiency at the firm level. We study the finite-sample performance of the model average estimator via Monte Carlo experiments and compare with traditional model averaging estimators based on weights constructed from model selection criteria and present a short empirical application.

Keywords Optimality · J-fold cross-validation · Efficiency · Model selection

1 Introduction

The stochastic frontier model (Aigner et al. 1977, Meeusen and van den Broeck 1977) has enjoyed widespread application across a diverse range of scientific milieus. Efficiency studies are useful for investigating the impact of the introduction or removal of firm regulations, constructing benchmarks by which firms are compared and in assessing improvements over time of the firm holding technology fixed. However, one of the major impediments to agreement over the results of efficiency studies is its strict adherence to distributional assumptions on the nature of efficiency. Indeed, Stone (2002) notes that any researcher who estimates the stochastic frontier model must make an arbitrary choice for the distribution of inefficiency. To combat the need for parametric assumptions authors have commonly used an array of techniques to reduce exposure to the impact that invalid assumptions can have on the analysis (for a recent example see Tsionas 2017). However, most of this analysis has focused on relaxing functional form assumptions on the shape of the frontier itself, with far less work focusing on lessening distributional assumptions pertaining to inefficiency. As Kneip et al. (2015, p. 380) note "While some central limit arguments can be advocated for the Gaussian noise, there does usually not exist any information justifying particular distributional assumptions on [inefficiency]."

Setting aside the choice of distribution for firm level inefficiency, another cause for concern for the practitioner is the exogenous factors that influence inefficiency, "determinants of inefficient". There exists considerable debate as to which variables may influence inefficiency. The recent work of Alvarez et al. (2006) and Lai and Huang (2010), develops in-depth frameworks for model selection amongst a variety of popular specifications arising from the normal-truncated normal distributional pair when determinants of inefficiency are present. However, when model uncertainty presents itself, an alternative to model selection is model averaging. Rather than selecting a single "winning" model, a model average estimator compromises across a set of competing models. Another important motivation behind model averaging is that accomplished practitioners know full well that different selection criteria favor more parsimonious models (BIC for example) while others favor models which are more heavily parameterized (the AIC

✉ Christopher F. Parmeter
cparmeter@bus.miami.edu

[1] Department of Economics, University of Miami, Miami, FL, USA

[2] Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong

[3] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 100190 Beijing, China

being the most prominent). Model averaging can represent an alternative between these two states of the world.

Traditional model averaging exercises have centered on two key features: first, prediction of an **observable** outcome, and second, averaging over a set of candidate models which are **nested**. When either of these two features do not hold then standard averaging estimators cannot be applied and alternative strategies are required. This is exactly the framework that one encounters in the efficiency arena; the main object of focus is on the construction of firm level predictions of conditional inefficiency (Jondrow et al. 1982, when no determinants are present). These measures are unobserved and commonly require distributional assumptions to recover them, many of which are non-nested (gamma versus half-normal, say). This makes direct application of standard model averaging approaches unappealing if the focus is on conditional inefficiency (as in Huang and Lai 2012).

We propose two different model averaging estimators for the stochastic frontier model where the focus is on conditional inefficiency. Both of these estimators are envisioned to be applied when researchers have access to the so-called 'determinants of inefficiency' or contextual variables of production. Little to no theoretical guidance exists on how these determinants/contextual variables enter the production process or influence inefficiency.[1] Rather, a common tactic is brute force, including all available determinants in a model of inefficiency and including the contextual variables in a linear fashion in the production technology and then engaging in inference and interpretation *ex post*. Given these forms of model uncertainty, it would seem prudent to engage in some type of model averaging exercise to combat the uncertainty with how exactly these variables enter the model.

Our first approach is to estimate firm level conditional inefficiency using as large a model as possible, and including all variables that the researcher has access to which we term inefficiency focused model averaging. This produces observable estimates of firm inefficiency which can then be used in the averaging procedure. From here submodels are penalized based on the number of parameters that they contain relative to the the largest possible model. We develop the necessary theory to show that in the setting where the focus of interest is unobserved, model averaging can still deliver optimal weights, albeit still requiring a nested distributional framework. We also demonstrate consistency of our inefficiency focused weight selection procedure.

The second model averaging estimator that we study can be thought of as the nonlinear least squares equivalent to Hansen and Racine's (2012) jackknife model averaging estimator.[2] Their estimator uses results on the form of the leave-one-observation-out hat matrix to construct a model averaging estimator. In the nonlinear regression/maximum likelihood context similar leave-one-out results do not exist and to maintain implementation ease, we introduce the J-fold cross-validation model averaging estimator (JCVMA), which omits J observations simultaneously, rather than a single observation.[3] We provide optimality of our JCVMA weights for the proposed estimator, demonstrating that our weight selection mechanism delivers weights that are as good as if we used the infeasible set of weights.

While our first model averaging estimator is potentially useful in settings where no determinants of inefficiency are present, it is limited by the scope of a nesting structure placed on the distribution of inefficiency, i.e., truncated-normal nests half-normal and gamma nests exponential, but a larger distribution which nests both truncated-normal and gamma is difficult to conceive. While this is a limitation, we provide the first attempt (to our knowledge) at developing a model averaging estimator for an unobserved criterion. Our JCVMA estimator, does not required a nesting structure, but can only be applied in settings where determinants of inefficiency are present and does not explicitly require distributional assumptions. The JCVMA estimator for the stochastic frontier model stems from the following observation: Once determinants of inefficiency are present, regardless of the distribution assumed for inefficiency, the conditional mean of output depends on these determinants and so we can once again focus on prediction of output. However, given the one-sided nature of inefficiency, the component of the conditional mean of output which depends on the determinants must be nonlinear, and hence the existing jackknife model averaging estimator will not suffice. Thus, these two estimators each offer a way around the current limitations of model averaging estimators in the context of the stochastic frontier model: the inefficiency focused model averaging estimator allows for an unobserved criterion while the JCVMA estimator allows one to potentially dispense with distributional assumptions but requires determinants of inefficiency.

While the present work can be viewed as the first serious attempt to formally construct frequentist based model

[1] For example, Lai and Huang (2010) include years of education of the primary decision maker in the household in their study of Indian farming. It is not theoretically clear if, and how, this variable should enter the production structure.

[2] Note that the model averaging estimator of Hansen and Racine's (2012) can accommodate nonlinearity of the unknown conditional mean through a sequence of bases such as orthogonal polynomials of varying order, splines of varying order and so forth, but their construction of weights is designed around a quadratic objective function with parameters which enter the model linearly. Here our focus is on the construction of weights when we have parameters which enter the model in a nonlinear fashion and/or the objective function is not quadratic.

[3] The use of $J \gg 1$ is to reduce the number of leave-one-out samples that need to be constructed to average over making the estimation more streamlined.

averaging methods specifically designed for the cross-sectional stochastic frontier model, earlier attempts have appeared providing model averaged estimates for some features of productivity or inefficiency. Specifically, Sickles (2005) takes simple averages of technical efficiency estimates for U.S. banks across a range of alternative stochastic frontier panel data models. While Sickles (2005 pg. 330) notes that this simple weighting is "clearly naïve, it does characterize the efficiency findings from the various estimators in a clear and informative summary". Huang and Lai (2012) use more formal frequentist model averaging approaches based on Buckland et al. (1997) approach, but when the object of interest is inefficiency, the weights can be viewed as *ex post* rather then the preferable *ex ante* (which is what we propose here).[4] A more comprehensive averaging analysis is performed in Sickles et al. (2014) who use a variety of weighting schemes beyond the simple averaging deployed in Sickles (2005) to forecast Asian countries' productivity growth.[5] Lastly, Olesen and Ruggiero (2018) use existing model averaging estimators to construct nonparametric production frontier estimators, ignoring the error structure completely.

The remainder of the article is setup as follows. Section 2 presents an overview of the canonical stochastic production frontier model and discusses estimation. Section 3 develops the necessary steps for constructing our alternative SFMA estimators and establishes their theoretical properties while Section 4 discusses selection of the averaging weights. Section 5 provides a Monte Carlo study illustrating the advantages of the different SFMA estimators while Section 6 applies these estimators to a commonly investigated dataset on Philippine rice farmers. Section 7 contains concluding remarks and avenues for future research. All proofs are contained in an appendix.

2 The basic stochastic frontier framework

Consider the stochastic frontier model[6]

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} - u_i + v_i, \quad \begin{array}{l} v_i \sim D\left(0, \sigma_v^2(\boldsymbol{z}_{1i};\boldsymbol{\gamma}_1)\right), \\ u_i \sim D_+\left(\mu(\boldsymbol{z}_{2i};\boldsymbol{\gamma}_2), \sigma_u^2(\boldsymbol{z}_{3i};\boldsymbol{\gamma}_3)\right) \end{array}, \tag{1}$$

where $\boldsymbol{x}_i = (1, x_{i2}, x_{i3}, \ldots, x_{ip})'$ is a $p \times 1$ vector of observed traditional inputs, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, $\boldsymbol{z}_{1i} = (1, z_{1i,2}, z_{1i,3}, \ldots, z_{1i,q_1})'$, $\boldsymbol{z}_{2i} = (1, z_{2i,2}, z_{2i,3}, \ldots, z_{2i,q_2})'$ and $\boldsymbol{z}_{3i} = (1, z_{3i,2}, z_{3i,3}, \ldots, z_{3i,q_3})'$ are vectors of observed variables, which may or may not contain elements from $\boldsymbol{x}_i$, and $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$, and $\boldsymbol{\gamma}_3$ are vectors of parameters. Here u_i captures inefficiency and v_i captures outside influences beyond the control of the producer as well as measurement error. We follow convention and assume that inputs are exogenously given. Given that u_i leads directly to a shortfall in output, it reduces output and as such it stems from a one-sided distribution. We allow overlap between $\boldsymbol{x}_i$, $\boldsymbol{z}_{1i}$, $\boldsymbol{z}_{2i}$ and $\boldsymbol{z}_{3i}$.

To recover insight about the magnitude of average inefficiency, more structure is required on the problem. The benchmark parametric stochastic production frontier was proposed independently by Aigner et al. (1977) and Meeusen and van den Broeck (1977). The standard solution is to impose distributional assumptions on both u_i and v_i (which induces a distribution for $\varepsilon_i = v_i - u_i$) and estimate all of the parameters of the model via maximum likelihood. In this framework v_i is normally distributed with $z_{1i} = 1$, and u_i is half normally or exponentially distributed with $z_2 = 0$ and $z_3 = 1$, respectively. Unequivocally, in applied SFA research v_i is assumed to be normally distributed with mean 0 and variance σ_v^2. The choice of distribution of u_i is less decisive, but the most common distributions to appear in practice are the half normal distribution, $N^+(0, \sigma_u^2)$, the Exponential distribution, $Exp(\sigma_u)$ and the truncated normal distribution, $N^+(\mu, \sigma_u^2)$.

Once distributional assumptions are in place for v_i and u_i, the density of the convoluted error term, $\varepsilon_i \equiv v_i - u_i$ is determined, and the model is estimated via maximum likelihood. Alternatively, when either $\boldsymbol{z}_{2i}$ or $\boldsymbol{z}_{3i}$ are non-constant, the model can also be estimated via nonlinear least squares. Given the one-sided nature of $D_+\left(\mu(\boldsymbol{z}_{2i};\boldsymbol{\gamma}_2), \sigma_u^2(\boldsymbol{z}_{3i};\boldsymbol{\gamma}_3)\right)$, the model can be rewritten as (Parmeter et al. 2017):

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} - g(\boldsymbol{z}_i;\boldsymbol{\gamma}) - u_i + g(\boldsymbol{z}_i;\boldsymbol{\gamma}) + v_i, \tag{2}$$

where $g(\boldsymbol{z}_i;\boldsymbol{\gamma}) = E[u|\boldsymbol{z}_i]$ and $\boldsymbol{z}_i = (\boldsymbol{z}_{2i}, \boldsymbol{z}_{3i})$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_2', \boldsymbol{\gamma}_3')'$. The exact form of $g(\boldsymbol{z}_i;\boldsymbol{\gamma})$ will depend upon the distributional assumptions placed on u (assuming that the distribution of v is symmetric). $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are both identified in this setting given the distributional assumptions placed on both the distribution of u, $D_+(\cdot,\cdot)$ as well as the exact functional forms for $\mu(\boldsymbol{z}_{2i};\boldsymbol{\gamma}_2)$ and $\sigma_u^2(\boldsymbol{z}_{3i};\boldsymbol{\gamma}_3)$. If $\boldsymbol{x}$ and $\boldsymbol{z}$ share no elements in common then it is possible to nonparametrically identify $g(\boldsymbol{z})$ (see Parmeter et al. 2017)

The model in (2) can be estimated via (generalized) nonlinear least squares (NLS) by noting that

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} - g(\boldsymbol{z}_i;\boldsymbol{\gamma}) - u_i^* + v_i = \boldsymbol{x}_i'\boldsymbol{\beta} - g(\boldsymbol{z}_i;\boldsymbol{\gamma}) + \varepsilon_i^*, \tag{3}$$

where $E[\varepsilon_i^*|\boldsymbol{x}_i, \boldsymbol{z}_i] = 0$ and $Var[\varepsilon_i^*|\boldsymbol{x}_i, \boldsymbol{z}_i]$ is non-constant; here $u_i^* = u_i - g(\boldsymbol{z}_i;\boldsymbol{\gamma})$.

Currently, we have only discussed estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, which provides information regarding the production frontier and the shape of the distribution of u_i. This information is all that is needed if interest hinges on the average level of technical inefficiency in the sample. However, if interest lies on the level of inefficiency for a given firm, knowledge of

[4] A similar strategy, in the context of productivity measurement across countries, appears in Sickles et al. (2015).
[5] See also Shang (2015).
[6] See Parmeter and Kumbhakar (2014) for a detailed account of this model.

$\mu(z_{2i};\gamma_2)$ or $\sigma_u^2(z_{3i};\gamma_3)$ is not sufficient. Thus, one typically seeks to recover an observation-specific measure of inefficiency. The primary solution when z_{2i} and z_{3i} are constant across firms, is to predict u_i with the expected value of u_i conditional on the composed error of the model, ε_i (Jondrow et al. 1982). This conditional mean of u_i given ε_i gives a point estimate of u_i. An alternative predictor of firm level inefficiency is $E[e^{-u_i}|\varepsilon_i]$ (Battese and Coelli 1988). Either of these conditional on ε measures will produce firm specific measures of inefficiency and can be used to rank firms.

2.1 The scaling case

The stochastic frontier model we have described currently involves a full parameterization of the distributions of both v and u. However, it is possible to completely avoid distributional assumptions by invoking the scaling property (Simar et al. 1994, Wang and Schmidt 2002):

$$u_i \sim g(z_{u,i};\boldsymbol{\delta}^u)u_i^*, \tag{4}$$

where $g(\cdot)\geq 0$ is a function of the exogenous variables, $z_{u,i}=(z_{2i}',z_{3i}')'$, while $u_i^*\geq 0$ is a random variable. Distributional assumptions (such as half-normal or truncated-normal) can be imposed on u_i^*; but it is assumed that u_i^* does not depend on $z_{u,i}$. When u_i follows the formulation in Eq. (4) it is then said to exhibit the scaling property.

An attractive statistical feature of the model with the scaling property imposed on the distribution of inefficiency is that it captures the idea that the *shape* of the distribution of u_i is the same for all firms (Alvarez et al. 2006). The scaling function $g(\cdot)$ essentially stretches or shrinks the horizontal axis, so that the scale of the distribution of u_i changes but its underlying shape does not. Perhaps the most important advantage of the scaling property specification is that the production frontier can be estimated without invoking distributional assumptions. Rather, nonlinear least squares (NLS) can be deployed by noting that

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} - e^{z_{u,i}'\boldsymbol{\delta}^u}\mu^* + v_i - e^{z_{u,i}'\boldsymbol{\delta}^u}(u_i^*-\mu^*) \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} - e^{z_{u,i}'\boldsymbol{\delta}^u}\mu^* + \varepsilon_i^*, \end{aligned} \tag{5}$$

where μ^* is the mean of u_i^*. This leads to the optimization criteria[7]

$$\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\delta}}^u,\hat{\mu}^*\right) = \min_{\boldsymbol{\beta},\boldsymbol{\delta}^u,\mu^*} n^{-1}\sum_{i=1}^{n}\left[y_i-\boldsymbol{x}_i'\boldsymbol{\beta}+\mu^* e^{z_{u,i}'\boldsymbol{\delta}^u}\right]^2. \tag{6}$$

[7] Given that the error term ε_i^* is heteroskedastic, $Var(\varepsilon^*|\boldsymbol{x}_i,z_{u,i}) = \sigma_v^2+\sigma_u^{2*}e^{2z_{u,i}'\boldsymbol{\delta}^u}$, where $\sigma_v^2=Var(v_i)$ and $\sigma_u^{2*}=Var(u^*)$, a generalized nonlinear least squares algorithm (though this requires distributional assumptions to disentangle σ_v^2 and σ_u^{2*}) or heteroscedasticity robust standard errors would be required to conduct valid inference.

The parameterization $e^{z_{u,i}'\boldsymbol{\delta}^u}$ is to ensure that the part of the conditional mean of y characterized by inefficiency is negative, consistent with a shortfall in output; alternative parameterizations for $g(z_{u,i};\boldsymbol{\delta}^u)$ could also be used, provided they were everywhere nonnegative. We note here that while the scaling property is attractive, it remains an assumption of the stochastic frontier model and if erroneous could lead to interpretation issues of the estimator as well as any inference conducted.

3 Stochastic frontier model averaging estimators

We propose two model averaging methods. The first works for nested model structures, such as averaging over the truncated normal and half normal families or the gamma and exponential families. The second allows averaging over a range of potentially nonnested distributions. Both methods have strengths and weaknesses. Averaging over nested model structures is similar to the framework of Huang and Lai (2012), except our proposal is to construct the model weights based on conditional inefficiency (which is unobserved), whereas their approach averages over AIC or BIC scores from the competing models, which does not necessarily provide the best estimate of inefficiency or the model weights.

Following (1), let $\boldsymbol{\theta}=(\gamma_1',\gamma_2',\gamma_3',\boldsymbol{\beta}')'$. Assume there are S candidate models containing different combinations of $\boldsymbol{x}_i$, z_{1i}, z_{2i} and z_{3i}. The s-th candidate model may be written as

$$\begin{aligned} y_i = \boldsymbol{x}_{s,i}'\boldsymbol{\beta}_s - u_{s,i} + v_{s,i}, \quad v_{s,i}\sim D\left(0,\sigma_v^2(z_{s,1i};\gamma_{s,1})\right), \\ u_{s,i}\sim D_+\left(\mu(z_{s,2i};\gamma_{s,2}),\sigma_u^2(z_{s,3i};\gamma_{s,3})\right) \end{aligned} \tag{7}$$

where $\boldsymbol{x}_{s,i}$, $z_{s,1i}$, $z_{s,2i}$ and $z_{s,3i}$ are sub-vectors of $\boldsymbol{x}_i$, z_{1i}, z_{2i} and z_{3i}, respectively. Further, let $\check{\boldsymbol{x}}_{s,i}=(\boldsymbol{x}_{s,i},z_{s,1i},z_{s,2i},z_{s,3i})'$. We note here that for our inefficiency focused model averaging estimator the families of distributions that are considered must be nested across the S different models, whereas for our J-fold cross-validation model averaging estimator we do not require the S different models to be nested. If determinants of inefficiency are not present, $z_{s,2i}$ and $z_{s,3i}$ are constant, then the JCVMA estimator is not operational and the inefficiency focused model averaging estimator is limited in scope given the nesting structure.

3.1 Inefficiency focused stochastic frontier model average estimators

Let $\boldsymbol{\theta}_s=(\gamma_{s,1}',\gamma_{s,2}',\gamma_{s,3}',\boldsymbol{\beta}_s')$ and $\widehat{\boldsymbol{\theta}}_s$ be the maximum likelihood estimator of $\boldsymbol{\theta}_s$. Let ρ_i be the focus parameter of interest, which can be $E(u_i|\varepsilon_i)$ or $E[e^{-u_i}|\varepsilon_i]$. The

corresponding estimator of ρ_i is

$$\hat{\rho}_{s,i} = \rho(y_i, \check{\boldsymbol{x}}_{s,i}, \hat{\boldsymbol{\theta}}_s). \tag{8}$$

Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$. We can write

$$\rho_i = \rho(y_i, \check{\boldsymbol{x}}_i, \boldsymbol{\theta}_0). \tag{9}$$

The model average estimator of ρ_i may be written as

$$\hat{\rho}_i(\boldsymbol{w}) = \sum_{s=1}^{S} w_s \hat{\rho}_{s,i}, \tag{10}$$

which is a weighted averages of $\hat{\rho}_{s,i}$ across the S candidate models, where $\boldsymbol{w} = (w_1, \ldots, w_S)'$ is the weight vector, belonging to the set $\mathcal{W} = \{\boldsymbol{w} \in [0,1]^S : \sum_{s=1}^{S} w_s = 1\}$.

We mention here that this specific stochastic frontier model averaging estimator is based on the focus of the averaging being the conditional (on the composed error) mean of inefficiency from the model. This is most pertinent when no determinants of inefficiency are present. However, when determinants of inefficiency are present, an alternative approach linking to jackknife cross-validation is available.

3.2 J-fold cross-validation model averaging

The second method we propose is "J-fold Cross-Validation model averaging (JCVMA), similar in spirit to Hansen and Racine's (2012) Jackknife model averaging (JMA). Beginning with the general form defined in (1) define $b_i = E(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)'$. JCVMA treats $\boldsymbol{b}$ as the target parameter.

For the sth candidate model, let

$$b_{s,i} = E_s(y_i|\check{\boldsymbol{x}}_{s,i}),$$

where E_s is the expectation operator under the assumption that the sth candidate is the correct model. We need an explicit closed form expression for $b_{s,i}$, which is readily obtainable once distributional assumptions for v_i and u_i have been made (or the scaling property is invoked). We first estimate the parameters of the sth candidate model by MLE or NLS, and then plug these estimates into $\hat{b}_{s,i}$ to obtain $\hat{b}_{s,i}$. Thus, we have the vector of estimators $\hat{\boldsymbol{b}}_s = (\hat{b}_{s,1}, \ldots, \hat{b}_{s,n})'$.

Write $\boldsymbol{w} = (w_1, \ldots, w_S)'$ as the weight vector, belonging in the set $\mathcal{W} = \{\boldsymbol{w} \in [0,1]^S : \sum_{s=1}^{S} w_s = 1\}$. The model average estimator of $\boldsymbol{b}$ is

$$\widehat{\boldsymbol{b}}(\boldsymbol{w}) = \sum_{s=1}^{S} w_s \widehat{\boldsymbol{b}}_s. \tag{11}$$

To apply JCVMA to choose weights in Eq. (11), we divide the data set into J groups such that for each group there are $M = n/J$ observations. Write $\widetilde{\boldsymbol{b}}_s^{(-j)}$ as the estimator of $(b_{s,1+(j-1)M}, \ldots, b_{s,jM})'$ with the jth group removed from the sample. Let $\widetilde{\boldsymbol{b}}_s = \text{stack}\left(\widetilde{\boldsymbol{b}}_s^{(-1)}, \ldots, \widetilde{\boldsymbol{b}}_s^{(-J)}\right)$ where the function stack$(\cdot)$ stacks the vectors on top of one another. That is, for each of the J groups, we hold out M observations, estimate the sth candidate model by MLE or NLS, and then use these estimates to predict the M observations which were excluded. This exercise is repeated a total of J times until each observation in the initial sample has been held out once. The vector $\widetilde{\boldsymbol{b}}_s$ is of the same order and length as $\boldsymbol{y}$. The JCVMA estimator of $\boldsymbol{b}$ is thus

$$\widetilde{\boldsymbol{b}}(\boldsymbol{w}) = \sum_{s=1}^{S} w_s \widetilde{\boldsymbol{b}}_s. \tag{12}$$

While there does not exist an optimal theory on the size of J, we note that Hansen and Racine's (2012) JMA exhibits a simple formulation based on leaving a single observation out due to the linear-in-parameters nature that they study. In models that are nonlinear in parameters, such a simple leave-one-observation-out solution may not exist. In these settings, JCVMA may offer a more expedient approach. We note here that the selection of J in model averaging exercises is an intensely studied topic but no concrete solutions exist at present. The appropriate selection of J is left for future research.

4 Weight choice

4.1 Inefficiency focused weight selection

We use squared error loss defined as $L(\boldsymbol{w}) = \sum_{i=1}^{n} \{\hat{\rho}_i(\boldsymbol{w}) - \rho_i\}^2$ as the basis for weight choice. Our aim is to select $\boldsymbol{w}$ such that $L(\boldsymbol{w})$ is minimized. Let $\widehat{\boldsymbol{\theta}}_{\text{full}}$ be the estimator of $\boldsymbol{\theta}$ under the full model (1), $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)'$, $\widehat{\boldsymbol{\rho}}_{\text{full}} = \left\{\hat{\rho}(y_1, \boldsymbol{x}_1, \widehat{\boldsymbol{\theta}}_{\text{full}}), \ldots, \hat{\rho}(y_n, \boldsymbol{x}_n, \widehat{\boldsymbol{\theta}}_{\text{full}})\right\}'$, $\widehat{\boldsymbol{\rho}}_s = (\hat{\rho}_{s,1}, \ldots, \hat{\rho}_{s,n})'$, and $\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) = \{\hat{\rho}_1(\boldsymbol{w}), \ldots, \hat{\rho}_n(\boldsymbol{w})\}'$. We propose to choose $\boldsymbol{w}$ by minimizing the following criterion:

$$C(\boldsymbol{w}) = \|\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \widehat{\boldsymbol{\rho}}_{\text{full}}\|^2 + n^{1/2}\log(n)\boldsymbol{k}'\boldsymbol{w}, \tag{13}$$

where $\boldsymbol{k} = (k_1, \ldots, k_S)'$, and k_s is the dimension of $\boldsymbol{\theta}_s$. The first term of $C(\boldsymbol{w})$, $\|\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \widehat{\boldsymbol{\rho}}_{\text{full}}\|^2$, can be thought of as an estimator of the squared error loss $L(\boldsymbol{w})$, whereas the second term $n^{1/2}\log(n)\boldsymbol{k}'\boldsymbol{w}$ gives rise to a penalty. Without the penalty component, a weight of unity will always be given to the largest model. The quantity $n^{1/2}\log(n)$ in the penalty term is a tuning parameter in order for the inefficiency focused model averaging estimator to satisfy consistency as stated in Theorem 1 below.

The criterion in Eq. (13) results in the empirically optimal weight vector

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{W}}{\operatorname{argmin}}\, C(\boldsymbol{w}). \tag{14}$$

We denote the true model by t as one that contains exclusively all the non-zero parameters. Any model that nests the true model is an over-fitted model, contained by the set $\mathcal{O}$. Note that the set of all candidate models is $\{1, \ldots, S\}$. We assume that S is finite and the dimension of $\boldsymbol{\theta}_s$ is fixed. Let $\mathcal{O}_1 = \{\mathcal{O} \cup \{t\}\} \cap \{1, \ldots, S\}$.

Assumption C.1. For any $s \in \{\mathcal{O} \cup \{t\}\}$, $\partial\hat{\rho}_{s,i}/\partial\widehat{\boldsymbol{\theta}}_s' \mid_{\widehat{\boldsymbol{\theta}}_s = \widetilde{\boldsymbol{\theta}}_{s,i}} = O_p(1)$ uniformly for $i = 1, \ldots, n$, *and for any* $\widetilde{\boldsymbol{\theta}}_{s,i}$ that lies between $\widehat{\boldsymbol{\theta}}_s$ and its limit. Furthermore, for any $s \in \{1, \ldots, S\}$, $\hat{\rho}_{s,i} = O_p(1)$ and $\rho_i = O_p(1)$ uniformly for $i = 1, \ldots, n$.

Theorem 1. Consistency *If* $\mathcal{O}_1$ is not an empty set and Assumption C.1 holds, then

$$n^{-1}\|\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\rho}\|^2 = O_p(n^{-1/2}\log(n)). \tag{15}$$

Let s_o and m^* be two models in $\mathcal{O}_1$. If Assumption C.1 holds and model s_o*is* nested within m^*, then

$$\hat{w}_{m^*} = O_p(\log^{-1}(n)). \tag{16}$$

By Eq. (15), the average estimation loss associated with $\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}})$ has a convergence rate of $n^{-1/2}\log(n)$. A direct implication of Eq. (16) is that if the true model is one of the candidate models, then our weight choice criterion would lead to weights assigned to the over-fitted models which converge to zero asymptotically. While Eq. (16) is no longer relevant when the true model is not nested within the candidate set, the subsequent asymptotic optimality in Theorem 2 is still valid. That is, the weight vector obtained based on our proposed method is asymptotically equivalent to that based on the infeasible optimal weight vector (even if the true model is not among the candidate models). This suggests that we can still construct asymptotically optimal weights even when the true model is not a member of the candidate set, which suggests a specific form of robustness for our inefficiency focused stochastic frontier model averaging estimator.

While the logarithmic rate of convergence for the weights in Theorem 1 may seem slow and impractical for applied work, this rate stems directly from the use of $\log(n)$ in the criterion (Eq. 13). This $\log(n)$ impacts the convergence rates of both $\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}})$ and $\hat{w}_{m^*}$. As is clear from the proof of Theorem 1, the rate of $\hat{w}_{m^*}$ can be increased by increasing $\log(n)$ in the criterion to a scalar increasing faster than $\log(n)$, however, the convergence rate of the inefficiency focused model averaging estimator, $\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}})$, will be slower. Thus, the practitioner faces a trade off between speed of convergence in the weights and speed of convergence in the model averaging estimator. Since the main aim is conceivably the construction of the inefficiency focused model averaging estimator the rate of convergence of the weights is not as important and hence, with slower convergence for the weight vector, we obtain faster convergence of the averaging estimator, hence the use $\log(n)$ in Eq. (13). Lastly, the use of $\log(n)$ is similar to model selection using BIC and the $\log(n)$ rate is the typical rate of convergence for BIC model selection (as in Lai and Huang 2010, for example).

We now develop the asymptotic optimality of $\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}})$. For any candidate model s, it can be seen from Assumptions A1–A3 of White (1982) that there exists a limit $\boldsymbol{\theta}_s^*$ such that

$$\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s^* = O_p(n^{-1/2}). \tag{17}$$

Let

$$\widehat{\boldsymbol{\rho}}_s^* = \left\{\hat{\rho}(y_1, \boldsymbol{x}_1, \boldsymbol{\theta}_s^*), \ldots, \hat{\rho}(y_n, \boldsymbol{x}_n, \boldsymbol{\theta}_s^*)\right\}',$$
$$\widehat{\boldsymbol{\rho}}^*(\boldsymbol{w}) = \left\{\widehat{\rho}_1^*(\boldsymbol{w}), \ldots, \widehat{\rho}_n^*(\boldsymbol{w})\right\}' = \sum_{s=1}^{S} w_s \widehat{\boldsymbol{\rho}}_s^*,$$

and $\xi_n = \inf_{\boldsymbol{w}\in\mathcal{W}} \|\widehat{\boldsymbol{\rho}}^*(\boldsymbol{w}) - \boldsymbol{\rho}\|^2$.

Assumption C.2. $\xi_n^{-1} n^{1/2} \log(n) = o_p(1)$.

Assumption C.2 implies all candidate models are misspecified. By this assumption, the full model cannot be one of the candidate models.

Theorem 2 Asymptotic Optimality Suppose that Assumption C.1 holds for any candidate model, Assumption C.2 is satisfied, and Eq. (17) holds. Then

$$\frac{\|\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\rho}\|^2}{\inf_{\boldsymbol{w}\in\mathcal{W}} \|\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \boldsymbol{\rho}\|^2} \xrightarrow{P} 1. \tag{18}$$

By Theorem 2, the squared error due to using the estimator $\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}})$ is asymptotically equivalent to that of the estimator based on the infeasible optimal weight vector. In other words, the model average estimator $\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}})$ is optimal in the class where the weights are restricted to $\mathcal{W}$. We label our estimator as the OPT estimator hereafter.

It might not be readily apparent that we can construct optimal weights from an object which we do not directly observe. However, the criterion in Eq. (13) should make it clear that the replacement of $\boldsymbol{\rho}$ with $\widehat{\boldsymbol{\rho}}_{\text{full}}$ presents a path forward in the construction of the weights. Moreover, both Theorems 1 and 2 demonstrate that we can obtain optimal

weights even without observing our target criterion, $\boldsymbol{\rho}$, which from the standpoint of the stochastic frontier model is technical inefficiency; this suggests in some sense robustness of our replacement of the unknown $\boldsymbol{\rho}$ with the estimator from the full model. The idea of replacing an unobservable object with an estimator is not new. For example, Mallows' C_p criterion (Mallows 1973) contains an unknown σ^2 which is routinely replaced by the estimator from the full model (although $\boldsymbol{\rho}$ and σ^2 are intrinsically different).

Aside from the intuition that $C(\boldsymbol{w})$ is a penalized estimator of the loss, $L(w)$, further intuition behind the criterion $C(\boldsymbol{w})$ exists. $C(\boldsymbol{w})$ is analogous with the Mallows' criterion of Hansen (2007). For the linear regression, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{v}$, assume that in the sth candidate model, the regressor matrix is $\boldsymbol{X}_s$, a sub-matrix of $\boldsymbol{X}$. Let $\widehat{\boldsymbol{\beta}}_s = (\boldsymbol{X}_s'\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s'\boldsymbol{y}$ and $\widehat{\boldsymbol{\beta}}_{\text{full}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Then, the Mallows' criterion in Hansen (2007) is equivalent to $\|\sum_{s=1}^{S} w_s \boldsymbol{X}_s\widehat{\boldsymbol{\beta}}_s - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{full}}\|^2 + 2\boldsymbol{k}'\boldsymbol{w}$, which uses the observable $\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{full}}$ in place of unobservable $\boldsymbol{X\beta}$. In the criterion, in Eq. (13), we use $\widehat{\boldsymbol{\rho}}_{\text{full}}$ in place of $\boldsymbol{\rho}$. Hence, the intuition behind the criterion in Eq. (13) and the Mallows' criterion are similar.

4.2 JCVMA weight selection

Given that the JCVMA estimator defined in Eq. (12) is non-operational as it depends on the unknown weights $\boldsymbol{w}$, we propose an estimator of the weight vector. To begin, define our criterion function for the selection of the weights as

$$CV_J(\boldsymbol{w}) = \left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{y}\right\|^2. \tag{19}$$

The weight vector $\widehat{\boldsymbol{w}}$ is chosen such that

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{W}}{\operatorname{argmin}}\, CV_J(\boldsymbol{w}). \tag{20}$$

The JCVMA estimator of $\boldsymbol{b}$ is thus $\widehat{\boldsymbol{b}}(\widehat{\boldsymbol{w}})$ as defined in Eq. (11).

Let $\boldsymbol{b}_s^* = \widehat{\boldsymbol{b}}_s\,|_{\widehat{\boldsymbol{\theta}}_s=\boldsymbol{\theta}_s^*}$, $\boldsymbol{b}^*(\boldsymbol{w}) = \sum_{s=1}^{S} w_s\boldsymbol{b}_s^*$, $\sigma^2 = \max_{1\le i\le n} E((y_i - b_i)^2|\boldsymbol{x}_i, \boldsymbol{z}_i)$, and $\zeta_n = \inf_{\boldsymbol{w}\in\mathcal{W}}\|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|^2$.

Assumption C.3. $\inf \zeta_n^{-1}n^{1/2} = o_p(1)$ and $\zeta_n^{-2}\sigma^2\sum_{s=1}^{S}\left\|\boldsymbol{b}_s^* - \boldsymbol{b}\right\|^2$ is uniformly integrable and equals $o_p(1)$.

Assumption C.4. $\partial\hat{b}_{s,i}/\partial\widehat{\boldsymbol{\theta}}_s'\,|_{\widehat{\boldsymbol{\theta}}_s=\widetilde{\boldsymbol{\theta}}_{s,i}} = O_p(1)$ uniformly for any s, $i = 1, \ldots, n$, and any $\widetilde{\boldsymbol{\theta}}_{s,i}$ that lies between $\widehat{\boldsymbol{\theta}}_s$ and its limit.

A direct implication of Assumption C.3 is that all candidate models are misspecified. We establish the asymptotic optimality of JCVMA in the following theorem.

Theorem 3 Asymptotic Optimality Suppose Assumptions C.3–C.4 are satisfied and Eq. (17) holds. Then

$$\frac{\left\|\widehat{\boldsymbol{b}}(\widehat{\boldsymbol{w}}) - \boldsymbol{b}\right\|^2}{\inf_{\boldsymbol{w}\in\mathcal{W}}\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}\right\|^2} \to^P 1. \tag{21}$$

The interpretation of Theorem 3 is analogous to that of Theorem 2, namely, the squared error due to using $\widehat{\boldsymbol{b}}(\widehat{\boldsymbol{w}})$ is asymptotically equivalent to the squared error resulting from the infeasible optimal weight vector.

5 Finite sample results

5.1 Inefficiency focused stochastic frontier model averaging results

This section reports results from a Monte Carlo study undertaken to compare the performance of the inefficiency focused model averaging estimator against AIC and BIC model selection (wAIC and wBIC, respectively) estimators as well as more traditional model averaging estimators based on the smoothed-AIC (s-AIC) and smoothed-BIC (s-BIC) weights, along with our inefficiency focused estimator (OPT) and the maximum likelihood estimator of the full model (FULL).

We generate the data from Eq. (1) with $D(\cdot) = N(0, \sigma_v^2)$, $\sigma_v = 2$, $D_+(\mu, \sigma_u) = Exp(\sigma_u)$, $\sigma_u = 1$, $\mu = E(u) = 2$, $\boldsymbol{\beta} = (2, -0.01, 1)'$, $\boldsymbol{x}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = (\Sigma_{j_1,j_2})$, $\Sigma_{j_1,j_2} = 0.6^{j_1-j_2}$ and $n = 50$, 100 and 200. Although the true error distributions are normal-exponential, we treat them as normal-half normal or normal-truncated normal in the estimation process. As $\boldsymbol{x}_i$ is a 3-dimensional vector, there are $S = (2^3 - 1)\times 2 = 14$ candidate models. Table 1 presents the mean and standard deviation of $\|\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\rho}\|^2/n$ for varying values of $\kappa = \sigma_u^2/\sigma_v^2$ based on 100 replications.

The results show that when $\kappa \le 0.5$, in terms of both the mean and standard deviation of the squared estimation errors, model averaging by the proposed OPT weight invariably results in the best estimates, while model selection by wBIC and wAIC always yield the worst and second worst results respectively. As well, for these values of κ, the sAIC and sBIC model average estimators generally result in estimates that are inferior to OPT but superior to other strategies; exceptions occur when $\kappa = 0.3$ and $n = 100$ and 200, for which the sBIC estimator delivers estimates with larger mean squared errors than the full model estimator. An increase in κ beyond 0.5 has the effect of worsening the relative performance of the OPT estimator. These results may be explained by noting that when κ is small, which

Table 1 Simulation results for the inefficiency focused model averaging estimator—100 simulations

			wAIC	wBIC	sAIC	sBIC	OPT	Full
$\kappa = 0.1$	$n = 50$	Mean	0.617	0.709	0.507	0.574	0.403	0.428
		s.d.	0.049	0.041	0.046	0.043	0.033	0.052
	$n = 100$	Mean	0.475	0.590	0.408	0.492	0.335	0.347
		s.d.	0.041	0.035	0.038	0.036	0.029	0.040
	$n = 200$	Mean	0.329	0.472	0.300	0.387	0.270	0.242
		s.d.	0.034	0.033	0.032	0.031	0.023	0.030
$\kappa = 0.3$	$n = 50$	Mean	0.414	0.444	0.315	0.344	0.255	0.345
		s.d.	0.032	0.030	0.032	0.031	0.024	0.035
	$n = 100$	Mean	0.318	0.371	0.241	0.278	0.205	0.268
		s.d.	0.023	0.024	0.024	0.024	0.019	0.024
	$n = 200$	Mean	0.252	0.293	0.190	0.215	0.166	0.194
		s.d.	0.018	0.018	0.019	0.020	0.015	0.018
$\kappa = 0.5$	$n = 50$	Mean	0.365	0.380	0.277	0.285	0.244	0.382
		s.d.	0.028	0.026	0.028	0.027	0.021	0.030
	$n = 100$	Mean	0.288	0.304	0.216	0.228	0.202	0.289
		s.d.	0.020	0.022	0.019	0.020	0.015	0.021
	$n = 200$	Mean	0.247	0.246	0.177	0.171	0.166	0.204
		s.d.	0.016	0.015	0.015	0.016	0.012	0.017
$\kappa = 1$	$n = 50$	Mean	0.414	0.379	0.313	0.288	0.332	0.522
		s.d.	0.031	0.029	0.026	0.025	0.022	0.035
	$n = 100$	Mean	0.379	0.270	0.253	0.215	0.286	0.386
		s.d.	0.030	0.023	0.020	0.017	0.017	0.032
	$n = 200$	Mean	0.246	0.232	0.192	0.161	0.218	0.227
		s.d.	0.025	0.023	0.021	0.015	0.016	0.027
$\kappa = 2$	$n = 50$	Mean	0.589	0.525	0.463	0.406	0.548	0.730
		s.d.	0.055	0.049	0.036	0.032	0.045	0.063
	$n = 100$	Mean	0.477	0.337	0.338	0.262	0.423	0.484
		s.d.	0.057	0.042	0.038	0.025	0.039	0.057
	$n = 200$	Mean	0.255	0.213	0.201	0.158	0.287	0.239
		s.d.	0.041	0.033	0.030	0.018	0.030	0.040

wAIC and wBIC refer to model selection through AIC and BIC, respectively, sAIC and sBIC refer to model averaging with weights calculated through the AIC and BIC criterion while OPT refer to our proposed model averaging approach. Full represents the model using all variables

implies a small σ_u relative to σ_v, it is difficult to identify a single winning model, and accordingly model averaging outperforms selection. However, As κ increases, the aforementioned difficulty dissipates and model selection becomes the preferred strategy.

5.2 JCVMA results

We generate the data from $y_i = \ln(\boldsymbol{x}_i)'\boldsymbol{\beta} + v_i - u_i$ with $v_i \sim N(0, \sigma_v^2)$, $\sigma_v = 1$, $u_i \sim e^{\boldsymbol{z}_u'\boldsymbol{\delta}^u} N_+(0, \pi/2)$, $\boldsymbol{\beta} = (0.62, 0.51)'$, $\boldsymbol{\delta}^u = (2, -0.2, 0.3)'$, $(\boldsymbol{x}_i', \boldsymbol{z}_i')' \sim N(\mu, \boldsymbol{\Sigma})$, $\mu = (4, 8, -1.0, -0.5, -0.5)'$, $\boldsymbol{\Sigma} = (\Sigma_{j_1,j_2})$, $\Sigma_{j_1,j_2} = 0.6^{|j_1 - j_2|}$ and $n = 200$, 400, 800 and 1600. Lastly, we set the variance of $\boldsymbol{z}_1 = 0.5$. We always include both $\ln x_1$ and $\ln x_2$ (i.e., we keep the production technology fixed) and average over all combinations of $\boldsymbol{z}_i$. As $\boldsymbol{z}_i$ is a 3-dimensional vector, there are $S = 2^3 - 1 = 7$ candidate models.

Table 2 presents the mean, median, and standard deviation of $\|\boldsymbol{b}(\widehat{\boldsymbol{w}}) - \boldsymbol{b}\|^2/n$, where $\boldsymbol{b} = E[\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}]$, based on AIC and BIC model selection, as well as s-AIC and s-BIC model averaging, the full (correctly specified model) and two variants of JCVMA across 1000 replications. JCVMA1 uses $\boldsymbol{b}(\widehat{\boldsymbol{w}}) = \widehat{\boldsymbol{b}}(\widehat{\boldsymbol{w}})$, as defined in Eq. (11), the fitted values from S candidate models not using hold-out samples, while JCVMA2 uses $\boldsymbol{b}(\widehat{\boldsymbol{w}}) = \widetilde{\boldsymbol{b}}(\widehat{\boldsymbol{w}})$, as defined in Eq. (12), the leave-J-observations-out fitted values from the estimated models. Note that both JCVMA1 and JCVMA2 use the same weights, $\widehat{\boldsymbol{w}}$, obtained from Eq. (20), they just conduct the averaging over different sets of fitted values. For all the simulations we leave out 10% of the sample size for our hold-out prediction (i.e., for $n = 200$, we hold out 20 observations at a time, for $n = 400$ we hold out 40 observations at a time, etc.).

Several insights are immediate from Table 2. First, as n increases, all of the methods, both selection and averaging, perform better. Second, JCVMA2 outperforms JCVMA1. It appears from Table 2 that selection or averaging over AIC or BIC offers no perceptible difference in performance. Regarding mean risk, JCVMA2 always outperforms the other methods, and also has equal standard deviation of risk. The relative gains across the methods dissipate as n increases, which is also to be expected. Overall, the results here, while limited, suggest that JCVMA offers promise.

6 Application to Philippines rice farming

We apply our JCVMA estimator to rice farming data collected in the Philippines. This data has become a benchmark example in applied efficiency analysis, serving as the dominant example in Coelli et al. (2005) and also appearing recently in Rho and Schmidt (2015). The data are composed of 43 farmers observed annually for eight years. Even though the data constitutes a panel, we will ignore this for our purposes. The output variable is tonnes of freshly threshed rice with the main input variables being area of planted rice (hectares), total labor used (man-days of family and hired-labor) and fertilizer used (kilograms). There is also a fourth input, other inputs, which is measured relative to farm 17 in the data via the Laspeyres index for 1991.[8]

We model inefficiency as depending upon several firm characteristics. For the current dataset this includes age of household head, education of household head, household size, number of adults in the household, and the percentage of area classified as bantog (upland) fields. *Ex ante* it is not clear which of these variables impacts expected inefficiency. For example, in a translog production framework,

[8] See Coelli et al. (2005, Appendix 2) for a more detailed description of the data.

Table 2 Simulation results for the JCVMA estimator—1000 simulations

		wAIC	wBIC	sAIC	sBIC	JCVMA1	JCVMA2	Full
$n = 200$	Mean	0.328	0.325	0.332	0.330	0.339	0.283	0.332
	Median	0.128	0.125	0.135	0.133	0.128	0.130	0.132
	s.d.	0.569	0.569	0.569	0.569	0.686	0.534	0.569
$n = 400$	Mean	0.260	0.260	0.261	0.262	0.261	0.224	0.260
	Median	0.101	0.103	0.104	0.106	0.103	0.107	0.103
	s.d.	0.483	0.483	0.483	0.482	0.507	0.410	0.483
$n = 800$	Mean	0.215	0.216	0.215	0.217	0.224	0.182	0.214
	Median	0.085	0.087	0.087	0.089	0.087	0.085	0.087
	s.d.	0.399	0.399	0.399	0.399	0.424	0.361	0.399
$n = 1600$	Mean	0.200	0.201	0.200	0.201	0.209	0.166	0.200
	Median	0.065	0.065	0.065	0.067	0.073	0.071	0.065
	s.d.	0.269	0.269	0.269	0.269	0.362	0.247	0.269

wAIC and wBIC refer to model selection through AIC and BIC, respectively, sAIC and sBIC refer to model averaging with weights calculated through AIC and BIC criterion while JCVMA1 and JCVMA2 refer to our proposed model averaging approach. Full represents the model using all variables

assuming a normal-half normal setup, Rho and Schmidt (2015, Table 11) find that household size and number of adults do not impact inefficiency, on average. Further, when they model the probability of being inefficient using the zero inefficiency stochastic frontier model of Kumbhakar et al. (2013), they find that none of the five variables influence inefficiency directly.

Here we confine ourselves to the normal-half normal distributional framework, but consider averaging over an array of different specifications for the variance parameter of the half normal distribution. Specifically, our distributional assumptions are $v_i \sim N(0, \sigma_v^2)$ and $u_i \sim N_+\left(0, \sigma_u^2(\boldsymbol{z}_i;\boldsymbol{\gamma})\right)$ where $\sigma_u^2(\boldsymbol{z}_i;\boldsymbol{\gamma}) = e^{\boldsymbol{z}_i'\boldsymbol{\gamma}}$. We average over every combination possible for $\boldsymbol{z}$. Including the setup where no determinants (aside from the intercept) are included this leads to 32 different models which we average over. We divide the data into eight distinct groups which suggests $J = 43$ for our hold out size. We also use five multi-starts for the estimation of every model to ensure that we are not stuck in a local optimum. Lastly, the data is shuffled prior to removing the hold out samples to ensure that the observations which are removed at each stage are done so randomly.

Table 3 lists the nine models that received non-zero weight from our model average procedure as well as the model weights. Age appears in eight of the nine models (and the one model that it does not appear in has the second smallest of the nine weights), while education only appears in three models. Both the number of adults in the household and Banrat appear four times. The model including all five of the determinants received zero weight, while eight of the nine models contained three or less of the five potential determinants of inefficiency. While this application is heuristic, given the general lack of theory on which determinants of inefficiency actually matter, our results here speak to the fact that model averaging may provide deeper insights than a kitchen sink approach.

7 Concluding remarks

Within the productivity and efficiency literature, beyond standard inputs, little in the way of theoretical guidance exists for informing how environmental and contextual variables influence the production structure. However, applied research abounds that uses these contextual variables in a wide array of manners. A natural avenue to deal with these issues of uncertainty is through model averaging. We propose two stochastic frontier model averaging estimators which can average over both environmental variables which influence the production structure directly, as well as indirectly through the determinants of inefficiency. Our selection of the weights across models is optimal provided that we do not include the full specified model amongst our set of candidate, misspecified models.

The JCVMA estimator which we propose is especially appealing as it does not necessarily require distributional assumptions to implement, if the scaling property is invoked. Further, by focusing on the conditional mean of output, a wide array of distributions can be averaged over, which our inefficiency-focused model averaging estimator is not capable of. We anticipate the JCVMA stochastic frontier estimator to have broad applicability.

Our simulations highlight that model averaging can provided estimates with lower risk than using traditional model selection procedures, which tend to favor larger models at the expense of parsimony. Natural extensions of this approach include construction of weights to average

Table 3 Models with non-zero weights selected by JCVMA for the normal-half normal Cobb-Douglas stochastic production frontier with determinants of inefficiency

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Age	1	1	1	1	1	1	1	–	1
Educ	–	–	–	1	1	–	–	–	1
HH Size	–	1	–	1	–	1	–	1	1
# Adults	–	–	–	–	1	1	1	1	–
Bantog	–	–	1	–	–	–	1	1	1
Model weights	0.0749	0.1726	0.0593	0.1080	0.2141	0.1917	0.0224	0.0500	0.1069
RTS	0.956	0.953	0.984	0.961	0.944	0.958	0.959	0.960	0.954
Median TE	1.000	0.999	1.000	1.000	1.000	1.000	0.782	0.999	0.921

All models contain an intercept. RTS refers to returns to scale while TE is the estimated level of technical efficiency of the farm. Median TE is rounded up, hence the appearance of 1.000

over uncertainty as it pertains to the distribution of inefficiency as well as alternative focus variables, for instance, measuring returns to scale or elasticities of substitution.

Acknowledgements We thank participants at the New York Camp Econometrics X, the 14th European Workshop on Efficiency and Productivity Analysis, LECCEWEPA 2015, the CEPA Workshop on Economic Measurement and the 2016 North American Productivity Workshop for valuable insight. Xinyu Zhang acknowledges the support from National Natural Science Foundation of China (Grant numbers 71522004, 11471324 and 71631008). The usual disclaimer applies.

Author contributions All three authors contributed equally to this work and the order of authorship has nothing other than alphabetical significance.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

8 Appendices

A.1. **Proof of Theorem 1**. We first decompose $C(\boldsymbol{w})$ as follows:

$$\begin{aligned} C(\boldsymbol{w}) &= \|\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \widehat{\boldsymbol{\rho}}_{\text{full}}\|^2 + n^{1/2}\log(n)\boldsymbol{k}'\boldsymbol{w} \\ &= \|\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \boldsymbol{\rho}\|^2 + \|\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\|^2 - 2\{\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \boldsymbol{\rho}\}'\{\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\} \\ &\quad + n^{1/2}\log(n)\boldsymbol{k}'\boldsymbol{w}. \end{aligned} \tag{A.1}$$

From the $\sqrt{n}$-consistency property of MLE and Assumption C.1, we have, for any $s^* \in \{\mathcal{O} \cup \{t\}\}$,

$$\|\widehat{\boldsymbol{\rho}}_{s^*} - \boldsymbol{\rho}\|^2 = \sum_{i=1}^{n} \left\| \frac{\partial \widehat{\boldsymbol{\rho}}_{s^*,i}}{\partial \widehat{\boldsymbol{\theta}}_{s^*}} \Big|_{\widehat{\boldsymbol{\theta}}_{s^*} = \widetilde{\boldsymbol{\theta}}_{s^*,i}} (\widehat{\boldsymbol{\theta}}_{s^*} - \boldsymbol{\theta}_{s^*}) \right\|^2 = O_p(1), \tag{A.2}$$

where $\widetilde{\boldsymbol{\theta}}_{s^*,i}$ lies between $\widehat{\boldsymbol{\theta}}_{s^*}$ and $\boldsymbol{\theta}_{s^*}$. By the definition of $\widehat{\boldsymbol{w}}$ in Eq. (14), we have

$$\begin{aligned} C(\widehat{\boldsymbol{w}}) &\le \|\widehat{\boldsymbol{\rho}}_{s^*} - \widehat{\boldsymbol{\rho}}_{\text{full}}\|^2 + n^{1/2}\log(n)k_{s^*} \\ &\le 2\|\widehat{\boldsymbol{\rho}}_{s^*} - \boldsymbol{\rho}\|^2 + 2\|\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\|^2 + n^{1/2}\log(n)k_{s^*}, \end{aligned} \tag{A.3}$$

which, along with (A.1), implies that

$$\begin{aligned} \|\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\rho}\|^2 - 2\{\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\rho}\}'\{\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\} &\le 2\|\widehat{\boldsymbol{\rho}}_{s^*} - \boldsymbol{\rho}\|^2 \\ &\quad + \|\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\|^2 + O(n^{1/2}\log(n)) \end{aligned},$$

and thus

$$\begin{aligned} \|\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\rho}\|^2 &\le 2\|\widehat{\boldsymbol{\rho}}_{s^*} - \boldsymbol{\rho}\|^2 + \|\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\|^2 \\ &\quad + O(n^{1/2}\log(n)) + 2\|\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\rho}\|\|\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\|. \end{aligned}$$

Therefore,

$$\begin{aligned} \{\|\widehat{\boldsymbol{\rho}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\rho}\| - \|\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\|\}^2 &\le 2\|\widehat{\boldsymbol{\rho}}_{s^*} - \boldsymbol{\rho}\|^2 \\ &\quad + O(n^{1/2}\log(n)) + 2\|\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\|^2. \end{aligned} \tag{A.4}$$

The full model belongs to $\mathcal{O}_1$ provided that any one candidate model belongs to $\mathcal{O}_1$. Hence we can obtain Eq. (15) from (A.2) and (A.4).

We next prove Eq. (16). Let

$$a_{s,m} = (\widehat{\boldsymbol{\rho}}_s - \widehat{\boldsymbol{\rho}}_{\text{full}})'(\widehat{\boldsymbol{\rho}}_m - \widehat{\boldsymbol{\rho}}_{\text{full}}) \tag{A.5}$$

and $\boldsymbol{\Phi}$ be an $S \times S$ matrix with its smth element given by

$$\Phi_{s,m} = a_{s,m} + n^{1/2}\log(n)(k_s + k_m)/2. \tag{A.6}$$

It can be easily shown that for any $\boldsymbol{w} \in \mathcal{W}$, $C(\boldsymbol{w}) = \boldsymbol{w}'\boldsymbol{\Phi}\boldsymbol{w}$. Now, define

$$\widetilde{\boldsymbol{w}} = (\widehat{w}_1, \ldots, \widehat{w}_{s_o-1}, \widehat{w}_{s_o} + \widehat{w}_{m^*}, \widehat{w}_{s_o+1}, \ldots, \widehat{w}_{m^*-1}, 0, \widehat{w}_{m^*+1}, \ldots, \widehat{w}_S)'.$$

Then we have

$$
\begin{aligned}
&0 \le C(\widetilde{\boldsymbol{w}}) - C(\widehat{\boldsymbol{w}}) \\
&= \widetilde{\boldsymbol{w}}'\boldsymbol{\Phi}\widetilde{\boldsymbol{w}} - \widehat{\boldsymbol{w}}'\boldsymbol{\Phi}\widehat{\boldsymbol{w}} \\
&= (\widetilde{\boldsymbol{w}} + \widehat{\boldsymbol{w}})'\boldsymbol{\Phi}(\widetilde{\boldsymbol{w}} - \widehat{\boldsymbol{w}}) \\
&= \{2\widehat{\boldsymbol{w}}' - (0, \ldots, 0, \widehat{w}_{m^*}, 0, \ldots, 0, -\widehat{w}_{m^*}, 0, \ldots, 0)\}\boldsymbol{\Phi}(0, \ldots, 0, \widehat{w}_{m^*}, 0, \ldots, 0, -\widehat{w}_{m^*}, 0, \ldots, 0)' \\
&= \widehat{w}_{m^*}^2(2\Phi_{s_o,m^*} - \Phi_{s_o,s_o} - \Phi_{m^*,m^*}) + 2\widehat{\boldsymbol{w}}'\boldsymbol{\Phi}(0, \ldots, 0, \widehat{w}_{m^*}, 0, \ldots, 0, -\widehat{w}_{m^*}, 0, \ldots, 0)' \\
&= \widehat{w}_{m^*}^2(2\Phi_{s_o,m^*} - \Phi_{s_o,s_o} - \Phi_{m^*,m^*}) + 2\widehat{w}_{m^*}\widehat{\boldsymbol{w}}'(\Phi_{1,s_o} - \Phi_{1,m^*}, \ldots, \Phi_{S,s_o} - \Phi_{S,m^*})' \\
&= \widehat{w}_{m^*}^2(2\Phi_{s_o,m^*} - \Phi_{s_o,s_o} - \Phi_{m^*,m^*}) + 2\widehat{w}_{m^*}\sum_{s=1}^{s}\widehat{w}_s(\Phi_{s,s_o} - \Phi_{s,m^*}) \\
&= \widehat{w}_{m^*}^2 O_p(1) + 2\widehat{w}_{m^*}\sum_{s=1}^{S}\widehat{w}_s\{O_p(n^{1/2}) + n^{1/2}\log(n)(k_{s_o} - k_{m^*})/2\} \\
&= \widehat{w}_{m^*}^2 O_p(1) + 2\widehat{w}_{m^*}O_p(n^{1/2}) + 2\widehat{w}_{m^*}n^{1/2}\log(n)(k_{s_o} - k_{m^*})/2,
\end{aligned}
$$

where the seventh equality expression is obtained using (A.2), (A.5) and (A.6). This yields

$$\widehat{w}_{m^*}n^{1/2}\log(n)(k_{m^*} - k_{s_o})/2 \le \widehat{w}_{m^*}^2 O_p(1) + \widehat{w}_{m^*}O_p(n^{1/2})$$

and hence $\widehat{w}_{m^*} = O_p(\log^{-1}(n))$, which is Eq. (16).

A.2. **Proof of Theorem 2**. Write

$$
\begin{aligned}
\|\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \boldsymbol{\rho}\|^2 = \|\widehat{\boldsymbol{\rho}}^*(\boldsymbol{w}) - \boldsymbol{\rho}\|^2 + \|\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \widehat{\boldsymbol{\rho}}^*(\boldsymbol{w})\|^2 \\
+ 2\{\widehat{\boldsymbol{\rho}}^*(\boldsymbol{w}) - \boldsymbol{\rho}\}'\{\widehat{\boldsymbol{\rho}}(\mathbf{w}) - \widehat{\boldsymbol{\rho}}^*(\mathbf{w})\}.
\end{aligned} \tag{A.7}
$$

From (A.1), (A.7), Assumption C.2, and the proof of Theorem 1′ in Wan et al. (2010), Theorem 2 holds provided that the following conditions hold:

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\|\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \widehat{\boldsymbol{\rho}}^*(\boldsymbol{w})\|^2}{\|\widehat{\boldsymbol{\rho}}^*(\boldsymbol{w}) - \boldsymbol{\rho}\|^2} = o_p(1), \tag{A.8}$$

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\left|\{\widehat{\boldsymbol{\rho}}^*(\boldsymbol{w}) - \boldsymbol{\rho}\}'\{\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \widehat{\boldsymbol{\rho}}^*(\boldsymbol{w})\}\right|}{\|\widehat{\boldsymbol{\rho}}^*(\boldsymbol{w}) - \boldsymbol{\rho}\|^2} = o_p(1), \tag{A.9}$$

and

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\left|\{\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \boldsymbol{\rho}\}'\{\widehat{\boldsymbol{\rho}}_{\text{full}} - \boldsymbol{\rho}\}\right|}{\|\widehat{\boldsymbol{\rho}}^*(\boldsymbol{w}) - \boldsymbol{\rho}\|^2} = o_p(1). \tag{A.10}$$

From Eq. (17) and Assumption C.1, we have

$$
\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\left\|\widehat{\boldsymbol{\rho}}(\boldsymbol{w}) - \widehat{\boldsymbol{\rho}}^*(\boldsymbol{w})\right\|^2}{\left\|\widehat{\boldsymbol{\rho}}^*(\boldsymbol{w}) - \boldsymbol{\rho}\right\|^2} \\
&\le \xi_n^{-1}\sum_{i=1}^{n}\sup_{\boldsymbol{w}\in\mathcal{W}}\left\{\widehat{\rho}_i(\boldsymbol{w}) - \widehat{\rho}_i^*(\boldsymbol{w})\right\}^2 \\
&= \xi_n^{-1}\sum_{i=1}^{n}\sup_{\boldsymbol{w}\in\mathcal{W}}\left\{\sum_{s=1}^{S} w_s(\widehat{\rho}_{s,i} - \widehat{\rho}_{s,i}^*)\right\}^2 \\
&\le \xi_n^{-1}\sum_{i=1}^{n} sup_{1\le s\le S}(\widehat{\rho}_{s,i} - \widehat{\rho}_{s,i}^*)^2 \\
&= \xi_n^{-1}\sum_{i=1}^{n}\sup_{1\le s\le S}\left\{\frac{\partial\widehat{\rho}_{s,i}}{\partial\widehat{\boldsymbol{\theta}}_s'}\Big|_{\widetilde{\boldsymbol{\theta}}_s = \widetilde{\boldsymbol{\theta}}_{s,i}}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s^*)\right\}^2 \\
&= O_p(\xi_n^{-1}).
\end{aligned} \tag{A.11}
$$

It follows from (A.11) and Assumption C.2 that (A.8) holds. In a similar way, we can prove that (A.9) and (A.10) hold. This proves Theorem 2.

A.3. **Proof of Theorem 3**. It can be seen that

$$
\begin{aligned}
&CV_J(\boldsymbol{w}) = \left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{y}\right\|^2 \\
&= \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b} + \widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w}) - (\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})) + \boldsymbol{b} - \boldsymbol{y}\right\|^2 \\
&\le \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}\right\|^2 + \left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|^2 + \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|^2 + \|\boldsymbol{b} - \boldsymbol{y}\|^2 \\
&+ \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}\right\|\left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\| + \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}\right\|\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\| + \left|(\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b})'(\boldsymbol{b} - \boldsymbol{y})\right| \\
&+ \left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\| + \left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|\|\boldsymbol{b} - \boldsymbol{y}\| + \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|\|\boldsymbol{b} - \boldsymbol{y}\| \\
&\le \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}\right\|^2 + \left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|^2 + \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|^2 \\
&+ \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|\left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\| + \|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|\left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\| \\
&+ \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\| + \|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\| \\
&+ \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|\|\boldsymbol{b} - \boldsymbol{y}\| + |(\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b})'(\boldsymbol{b} - \boldsymbol{y})| + \left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\| \\
&+ \left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|\|\boldsymbol{b} - \boldsymbol{y}\| + \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|\|\boldsymbol{b} - \boldsymbol{y}\| + \|\boldsymbol{b} - \boldsymbol{y}\|^2 \\
&\equiv \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}\right\|^2 + \Pi_n(\boldsymbol{w}) + \|\boldsymbol{b} - \boldsymbol{y}\|^2
\end{aligned}
$$

and

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}\right\|^2 &= \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w}) + \boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\right\|^2 \\
&= \|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|^2 + \left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|^2 + 2\left(\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right)'(\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}) \\
&= \|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|^2 + \Xi_n(\boldsymbol{w}).
\end{aligned}
$$

Hence to prove Theorem 3, it suffices to show that

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\Xi_n(\boldsymbol{w})}{\|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|^2} = o_p(1) \tag{A.12}$$

and

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\Pi_n(\boldsymbol{w})}{\|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|^2} = o_p(1). \tag{A.13}$$

Similar to the proof of (A.2), by Eq. (17) and Assumption C.4, we have

$$
\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}}\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|^2 = O_p(1) \\
&\text{and} \\
&\sup_{\boldsymbol{w}\in\mathcal{W}}\left\|\widetilde{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|^2 = O_p(1).
\end{aligned} \tag{A.14}
$$

It is readily seen that

$$\|\boldsymbol{b} - \boldsymbol{y}\|^2 = O_p(n) \tag{A.15}$$

and

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|}{\|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|^2}, = \sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\left\|\widehat{\boldsymbol{b}}(\boldsymbol{w}) - \boldsymbol{b}^*(\boldsymbol{w})\right\|}{\|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|}. \tag{A.16}$$

Table 4 Simulation results for the JCVMA estimator—1000 simulations. Holdout sample is 2.5% of n

		wAIC	wBIC	sAIC	sBIC	JCVMA1	JCVMA2	Full
$n = 200$	Mean	0.291	0.288	0.295	0.293	0.302	0.259	0.295
	Median	0.126	0.122	0.134	0.129	0.125	0.125	0.131
	s.d.	0.434	0.434	0.433	0.434	0.572	0.455	0.433
$n = 400$	Mean	0.227	0.227	0.228	0.229	0.223	0.200	0.227
	Median	0.099	0.101	0.103	0.105	0.103	0.101	0.102
	s.d.	0.349	0.348	0.348	0.348	0.363	0.323	0.349
$n = 800$	Mean	0.188	0.189	0.189	0.190	0.196	0.168	0.188
	Median	0.084	0.085	0.084	0.087	0.083	0.083	0.084
	s.d.	0.298	0.298	0.298	0.298	0.333	0.317	0.298
$n = 1600$	Mean	0.165	0.165	0.165	0.166	0.179	0.144	0.164
	Median	0.064	0.065	0.064	0.066	0.071	0.068	0.064
	s.d.	0.269	0.269	0.269	0.269	0.359	0.240	0.269

For any $\delta > 0$,

$$
\begin{aligned}
&Pr\{\sup_{\boldsymbol{w}\in\mathcal{W}} \zeta_n^{-1}|(\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b})'(\boldsymbol{b} - \boldsymbol{y})|\, gt; \delta\} \\
&\le Pr\{\sup_{\boldsymbol{w}\in\mathcal{W}} \zeta_n^{-1} \sum_{s=1}^{S} w_s|(\boldsymbol{b}_m^* - \boldsymbol{b})'(\boldsymbol{b} - \boldsymbol{y})|\, gt; \delta\} \\
&= Pr\{\max_s |(\boldsymbol{b}_s^* - \boldsymbol{b})'(\boldsymbol{b} - \boldsymbol{y})|\, gt; \zeta_n\delta\} \\
&\le \sum_{s=1}^{S} Pr\{|(\boldsymbol{b}_s^* - \boldsymbol{b})'(\boldsymbol{b} - \boldsymbol{y})|\, gt; \zeta_n\delta\} \\
&\le \sum_{s=1}^{S} E\{\zeta_n^{-2}\delta^{-2}(\boldsymbol{b}_s^* - \boldsymbol{b})'(\boldsymbol{b} - \boldsymbol{y})\}^2 \\
&= \sum_{s=1}^{S} E\left[E\{\zeta_n^{-2}\delta^{-2}(\boldsymbol{b}_s^* - \boldsymbol{b})'(\boldsymbol{b} - \boldsymbol{y})|\boldsymbol{x}, z\}^2\right] \\
&= \sum_{s=1}^{S} E\left[\zeta_n^{-2}\delta^{-2}(\boldsymbol{b}_s^* - \boldsymbol{b})'\mathrm{var}(\boldsymbol{b} - \boldsymbol{y}|\boldsymbol{x}, z)(\boldsymbol{b}_s^* - \boldsymbol{b})\right] \\
&\le \sum_{s=1}^{S} E\left[\zeta_n^{-2}\delta^{-2}\sigma^2\left\|\boldsymbol{b}_s^* - \boldsymbol{b}\right\|^2\right],
\end{aligned}
$$

where $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ and $\boldsymbol{z} = (z_1, \ldots, z_n)'$, and σ^2 is defined in the line above Assumption C.3. Together with Assumption C.3, this implies

$$
\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|(\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b})'(\boldsymbol{b} - \boldsymbol{y})|}{\|\boldsymbol{b}^*(\boldsymbol{w}) - \boldsymbol{b}\|^2} = o_p(1). \tag{A.17}
$$

By combining (A.14)–(A.17) and Assumption C.3, we can obtain (A.12) and (A.13) and hence Eq. (21). This completes the proof of Theorem 3.

A.4. **Simulation Results with Smaller Hold Out Sample**. Similar to Tables 2 and 4 presents the mean, median and standard deviation of $\|\boldsymbol{b}(\widehat{\boldsymbol{w}}) - \boldsymbol{b}\|^2/n$, where $\boldsymbol{b} = E[\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}]$, based on AIC and BIC model selection, as well as s-AIC and s-BIC model averaging, the full (correctly specified model) and two variants of JCVMA across 1000 replications. JCVMA1 uses $\boldsymbol{b}(\widehat{\boldsymbol{w}}) = \widehat{\boldsymbol{b}}(\widehat{\boldsymbol{w}})$, as defined in Eq. (11), the fitted values from S candidate models not using hold-out samples, while JCVMA2 uses $\boldsymbol{b}(\widehat{\boldsymbol{w}}) = \widetilde{\boldsymbol{b}}(\widehat{\boldsymbol{w}})$, as defined in Eq. (12), the leave-J-observations-out fitted values from the estimated models. Note that both JCVMA1 and JCVMA2 use the same weights, $\widehat{\boldsymbol{w}}$, obtained from Eq. (20), they just conduct the averaging over different sets of fitted values. For all the simulations we leave out 2.5% of the sample size for our hold out prediction (i.e., for $n = 200$, we hold out 5 observations at a time, for $n = 400$ we hold out 10 observations at a time, etc.).

Several insights are immediate from Table 4 relative to the results from Table 2. JCVMA2 still outperforms JCVMA1. JCVMA2 always outperforms the other methods in terms of mean risk, and also has equal standard deviation of risk. Comparing mean and median risk, it does not appear that the size of the hold out sample has much effect on the performance of either of the JCVMA estimators.

References

Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production functions. J Econom 6(1):21–37

Alvarez A, Amsler C, Orea L, Schmidt P (2006) Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics. J Prod Anal 25(2):201–212

Battese GE, Coelli TJ (1988) Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. J Econom 38:387–399

Buckland ST, Burnham KP, Augustin NH (1997) Model selection: an integral part of inference. Biometrics 53(4):603–618

Coelli TJ, Rao DP, O'Donnell CJ, Battese GE (2005) An Introduction to Efficiency and Productivity Analysis. Springer, New York

Hansen BE (2007) Least squares model averaging. Econometrica 75 (4):1175–1189

Hansen BE, Racine JS (2012) Jackknife model averaging. J Econom 167(1):38–46

Huang CJ, Lai H-P (2012) Estimation of stochastic frontier models based on multimodel inference. J Prod Anal 38:273–284

Jondrow J, Lovell CAK, Materov IS, Schmidt P (1982) On the estimation of technical efficiency in the stochastic frontier production function model. J Econom 19(2/3):233–238

Kneip A, Simar L, Van Keilegom I (2015) Frontier estimation in the presence of measurement error with unknown variance. J Econom 184:379–393

Kumbhakar SC, Parmeter CF, Tsionas E (2013) A zero inefficiency stochastic frontier estimator. J Econom 172(1):66–76

Lai H-P, Huang CJ (2010) Likelihood ratio tests for model selection of stochastic frontier models. J Prod Anal 34(1):3–13

Mallows CL (1973) Some comments on cp. Tehcnometrics 15:661–675

Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. Int Econ Rev 18(2):435–444

Olesen OB, Ruggiero J (2018) An improved Afriat-Diewert-Parkan nonparametric production function estimator. Eur J Operat Res 264:1172–1188

Parmeter CF, Kumbhakar SC (2014) Efficiency analysis: a primer on recent advances. Found Trends Econom 7(3-4):191–385

Parmeter CF, Wang H-J, Kumbhakar SC (2017) Nonparametric estimation of the determinants of inefficiency. J Prod Anal 47 (3):205–221

Rho S, Schmidt P (2015) Are all firms inefficient? J Prod Anal 43 (3):327–349

Shang C (2015) Essays on the use of duality, robust empirical methods, panel treatments, and model averaging with applications to housing price index construction and world productivity growth, PhD thesis, Rice University

Sickles RC (2005) Panel estimators and the identification of firm-specific efficiency levels in parametric, semiparametric and nonparametric settings. J Econom 126(2):305–334

Sickles RC, Hao J, Shang C (2014) Panel data and productivity measurement: an analysis of Asian productivity trends. J Chin Econ Bus Stud 12(3):211–231

Sickles RC, Hao J, Shang C (2015) Panel data and productivity measurement. In: Baltagi B (ed) Ch 17, Oxford Handbook fo Panel Data. Oxford University Press, New York, pp 517–547

Simar L, Lovell CAK, van den Eeckaut P (1994) Stochastic frontiers incorporating exogenous inuences on efficiency. Discussion Papers No. 9403, Institut de Statistique, Universite de Louvain

Stone M (2002) How not to measure the efficiency of public services (and how one might). J R Stat Soc Ser A 165:405–434

Tsionas EG (2017) "When, where and how" of efficiency estimation: Improved procedures for stochastic frontier modeling. J Am Stat Assoc 112:948–965

Wan ATK, Zhang X, Zou G (2010) Least squares model averaging by Mallows criterion. J Econom 156(4):277–283

Wang H-J, Schmidt P (2002) One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. J Prod Anal 18:129–144

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50(1):1–25