
**Centre for Efficiency and Productivity Analysis**

**Working Paper Series**
**No. WP01/2023**

# BAYESIAN ARTIFICIAL NEURAL NETWORKS FOR FRONTIER EFFICIENCY ANALYSIS

Mike Tsionas, Christopher F. Parmeter and Valentin Zelenyuk

**Date: January 2023**

**School of Economics**
**University of Queensland**
**St. Lucia, Qld. 4072**
**Australia**

**ISSN No. 1932 - 4398**

# BAYESIAN ARTIFICIAL NEURAL NETWORKS FOR FRONTIER EFFICIENCY ANALYSIS

MIKE TSIONAS, CHRISTOPHER F. PARMETER, AND VALENTIN ZELENYUK

Abstract. Artificial neural networks have offered their share of econometric insights, given their power to model complex relationships. One area where they have not been readily deployed is the estimation of frontiers. The literature on frontier estimation has seen its share of research comparing and contrasting data envelopment analysis (DEA) and stochastic frontier analysis (SFA), the two workhorse estimators. These studies rely on both Monte Carlo experiments and actual data sets to examine a range of performance issues which can be used to elucidate insights on the benefits or weaknesses of one method over the other. As can be imagined, neither method is universally better than the other. The present paper proposes an alternative approach that is quite flexible in terms of functional form and distributional assumptions and it amalgamates the benefits of both DEA and SFA. Specifically, we bridge these two popular approaches via Bayesian artificial neural networks while accounting for possible endogeneity of inputs. We examine the performance of this new machine learning approach using Monte Carlo experiments which is found to be very good, comparable to, or often better than, the current standards in the literature. To illustrate the new techniques, we provide an application of this approach to a data set of large US banks.

**Key Words**: Machine Learning; Simulation; Flexible Functional Forms; Bayesian Artificial Neural Networks; Banking; Efficiency Analysis.

This working paper is a substantially revised version of CEPA WP08/2021. Feedback is very welcome.

---

Mike Tsionas, Montpellier Business School Université de Montpellier, Montpellier Research in Management and Lancaster University Management School, LA1 4YX, U.K.; m.tsionas@lancaster.ac.uk. Christopher F. Parmeter, Miami Herbert Business School, University of Miami, Miami FL; cparmeter@bus.miami.edu. Valentin Zelenyuk, School of Economics, The University of Queensland, Brisbane, Australia; v.zelenyuk@uq.edu.au. This paper was previously ciculated under the title "Bridging the Divide? Bayesian Artificial Neural Networks for Frontier Efficiency Analysis." We acknowledge support from the Australian Research Council (FT170100401). We also thank Evelyn Smart, Bao Hoang Nguyen and Zhichao Wang for their feedback. Comments from the editor, Elie Tamer, as well as an associate editor and two referees are warmly acknowledged. The usual caveat applies.

## 1. Introduction

Neural networks constitute a set of powerful machine learning tools to tackle myriad data-supported real world problems. However the uncertainty associated with predictions stemming from these methods is often challenging to quantify. Bayesian statistics offers a formalism to quantify the uncertainty associated with deep neural network predictions. There are many interesting uses of deep learning methods in economics, each potentially having their own specific implementation. Here we will focus on adapting these methods to the area of production frontier estimation and efficiency analysis, which to the best of our knowledge is novel to this extensive field of research. In particular, here we study how Bayesian neural networks can be used, coupled with the construction of robust priors, to make predictions about firm level characteristics.

Frontier methods continue to have a pronounced impact through myriad of applications in various fields of economics/econometrics and management science/operations research. They have been deployed for analyzing performance of airlines, banks, energy generation and distribution, farms, fisheries, hotels, healthcare providers, universities, etc.[1] Aside from their frequent use in academia, they have also become part of the standard toolbox used by regulators and economics/management consultants when crafting policies and strategies aimed at improving performance. Currently, the two main approaches are stochastic frontier analysis (SFA) and data envelopment analysis (DEA). Since the inception of SFA and DEA, there have been robust and lively debates on the merits of each method relative to the other. On the one hand, DEA can estimate a multi-output multi-input production technology in a fully nonparametric fashion, enforcing axioms of production and identify firm level efficiency. On the other, parametric SFA does not run into dimensionality issues that can potentially plague reliable inference of DEA and is robust to the presence of stochastic noise that is likely to exist in economic production environments. Both methods have their supplicants and detractors.[2]

While the relative merits of one method over the other can be debated, what is less controversial is the importance of understanding and quantifying firm efficiency. These methods have

[1]See various examples provided in Kumbhakar, Parmeter and Zelenyuk (2021a).

[2]This language may seem harsh, however, many empirical applications studying efficiency commonly use one (but not both of the methods) behaving as though the other method does not exist. In our view this does a crippling disservice to the field as both methods have been shown to have tangible benefits (Badunenko, Henderson and Kumbhakar, 2012) and can offer useful empirical insights.

enjoyed widespread adoption across a range of economic milieus, seeing deployment for a host of analyses related to regulation and benchmarking and determining frontier individuals, firms, states, governments and countries. Still, the debate rages within this area, should one use DEA, which assumes that no stochastic noise is present in the production process or SFA, which usually requires rigid parametric assumptions. As it turns out the answer is complicated. Both methods have seen a wide variety of improvements since their initial incarnations. Parmeter and Zelenyuk (2019) have recently reviewed and compared a range of methods that embrace the benefits of both DEA and SFA, demonstrating that such a binary choice (SFA or DEA) is no longer an empirically relevant question. What is relevant is how best to combine or embrace the virtues of both methods. There still exists a widespread lack of consensus on a go to, practical approach that researchers can deploy to tackle the important issue of frontier efficiency analysis.

Here, taking insights from the extant DEA literature, using its axiomatic underpinnings and linear programming apparatus, we try to improve traditional Bayesian estimation of the stochastic frontier model (e.g. see van den Broeck et al. 1994; Griffiths and Hajargasht 2016). Standard parametric SFA, for example, assumes that the frontier is linear in the parameters, the two-sided error term is Normally distributed, while the one-sided error component (which stands for inefficiency) is distributed according to the Half Normal, Truncated Normal or Gamma distributions (for example). Clearly, this involves many assumptions which policy and decision makers may be uncomfortable with. To make things worse, these assumptions are rarely tested in practice.

This limitation of strict reliance on potentially fragile parametric assumptions suggests that the use of SFA can benefit from the inclusion of some of the best virtues of DEA. In particular, here we use DEA to benchmark a good "prior" for the parameters of inefficiency and/or the frontier parameters themselves, although we do not make distributional assumptions about the one-sided error term. We show how to impose monotonicity and curvature conditions (if desired). Central to our framework is the notion of a "smooth mixtures of Normals" (SMN), developed by Geweke and Keane (1997, 2007) and Villani et al. (2009). The use of SMN has been shown to have excellent asymptotic properties (Norets, 2010) and the notion of heterogeneity in the data that can be captured using endogenously determined groups based on SMN. To our knowledge this approach has yet to be applied to the frontier literature. Further, our ability to combine DEA

constructed priors inside of SMN is a novel approach to further enhance the appeal of the approach for practitioners of efficiency and productivity analysis.

The use of DEA priors within a stochastic frontier setting may seem nonstandard. However, our aim to use a "robust" prior when informing the stochastic frontier model. As an alternative to direct use of DEA priors we also rely on a least favorable prior (LFP) setup that, under mild conditions, is known to be minimax. Specifically, a conditional expectation (optimal Bayes estimator) evaluated with a LFP distribution is also a minimax estimator. Finding least favorable distributions can be challenging due to the inherent optimization over the space of probability distributions, which is infinite-dimensional. To reduce the dimensionality we only consider priors constructed from both classic DEA and SFA settings. Other types of alternatives for robust Bayes approaches exist as well (see Giacomini and Kitagawa, 2021 and Kline and Tamer, 2016), though we leave discussion and analysis of this for future research.

Our results from the Monte Carlo experiments show that the new techniques perform well in the presence of endogeneity. An accompanying appendix also demonstrates that the new estimator is competitive to alternative nonparametric SFA approaches and in some cases behaves better in terms of root mean-squared error. The new method seems to perform well with nearly non-informative or, practically, flat priors.

The question of how to combine DEA and SFA (certainly, an old one) is addressed via the use of both strict DEA-based priors and LFPs and then the assessment of model weights in optimal predictive pools (Geweke and Amisano, 2011 a,b). To illustrate, we provide an empirical application of the proposed approach to a data set of large US banks. The main findings show that the new SMN models receive, practically, the lion's share in such predictive analyses. This suggests that these two approaches, once thought to be in conflict, can actually work together in empirical settings to produce useful insights. Our empirical application also illustrates the usefulness of the newly proposed approach and, in particular, the importance of accounting for endogeneity in frontier estimation. *Inter alia*, it also hints that some of the findings from earlier frontier studies about efficiency and economies scale in the US banking (or, in fact, any other industry) that did not account for endogeneity might need revisiting. Such revisiting may confirm or change our understanding of the banking sector.

## 2. Model

Our main interest is the benchmark stochastic frontier model (SFM):

$$y_i = f(\boldsymbol{x}_i) + v_i - u_i,\ i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{x}_i \in \mathbb{R}^K$ is the vector of inputs, $y_i = \log Y_i$ is log output, $f : \mathbb{R}^K \to \mathbb{R}$ is the production frontier, $v_i$ represents noise (a random variable supported in $\mathbb{R}$), and $u_i$ is a random variable supported on $\mathbb{R}_+$ representing technical inefficiency.[3] Moreover, cost frontiers can be considered using $+u_i$ instead of $-u_i$ in (1). For ease of exposition straightforward. Moreover, in model (1) it is assumed that the inputs and output are in log terms, and $\boldsymbol{x}_i$ may contain functions of the inputs as well (for example, squared terms and interactions terms as in the case of a translog production function).

In addition, suppose there is a vector $\boldsymbol{z}_i \in \mathbb{R}^{d_z}$ of "environmental" variables. These variables are commonly thought of as only influencing inefficiency but for our purposes they could also be a subvector of $\boldsymbol{x}_i$ or influence $v_i$ through heteroskedasticity. These add minimal complications to our discussion below.[4] We denote $\boldsymbol{w}_i = [\boldsymbol{x}_i', \boldsymbol{z}_i']' \in \mathbb{R}^{d_w}$.

There are five fundamental issues in the estimation of the SFM: (i) Specification of the structure of $u_i$; (ii) Specification of the structure of $v_i$; (iii) Specification of the functional form $f(\cdot)$; (iv) Possible endogeneity of the covariates $\boldsymbol{x}_i$; and (v) Imposition of smoothness constraints to enforce axioms of production.

The literature studying the SFM has offered proposals attempting to lessen the impact of some of these points, either simultaneously or individually. For example, numerous authors have studied how to estimate the production technology in a nonparametric fashion (see Parmeter and Zelenyuk, 2019 for a review) but only Simar and Zelenyuk (2011), Parmeter and Racine (2012) and Kuosmanen and Kortelainen (2013) have sought to enforce axioms of production. Other authors have turned their attention to dispensing with distributional assumptions on $u$. Papers by Hall and Simar (2002), Tran and Tsionas (2009), Horrace and Parmeter (2011) and Parmeter, Wang and Kumbhakar (2017) fully dispense with distributional assumptions on inefficiency. For endogeneity,

[3]This formulation nests input and output-oriented distance functions with appropriate redefinition of the covariates.
[4]The existence of such predetermined or environmental variables is assumed in place. With panel data a time trend and lags of inputs and outputs can be used along with any other available environmental variables. With cross-sectional data, one can proceed using artificial instruments as in Lewbel (1996).

only recently has the literature started to devote attention to solving this pernicious problem. Amsler, Prokhorov and Schmidt (2016, 2017) are two of the most prominent, but they do not focus on any of the other issues which impact the estimation of SFMs. To date, there does not yet exist an approach that can handle all of these important assumptions in a direct manner. In what follows, we will consider each of these issues within the confines of the approach we develop.

## 3. Estimation Issues of the Stochastic Frontier Model

Estimation of the benchmark SFM is by now well known and well understood (Kumbhakar, Parmeter and Zelenyuk, 2021a,b). Here we will detail the salient estimation issues as they pertain to five points above and how our framework can accommodate these important issues.

### 3.1. **Modelling Inefficiency.**

3.1.1. *Deterministic Inefficiency.* As $r_i \equiv e^{-u_i} \in (0,1]$ is technical efficiency, it is reasonable to parameterize $r_i = \Phi(\boldsymbol{w}_i'\boldsymbol{\delta})$ for $\boldsymbol{w}_i'\boldsymbol{\delta} \in \mathbb{R}$, where $\Phi(\cdot)$ is a cumulative distribution function (the standard Normal in this study), and $\boldsymbol{\delta} \in \mathbb{R}^{d_\delta}$. Note that in this setup, inefficiency is treated as deterministic instead of random. The benefit is that this switch eliminates the need for distributional assumptions; the obvious downside is that the selected $\boldsymbol{w}$ variables fully explain inefficiency so an omitted variable bias may arise.

The roots of this formulation go back to at least Simar et al. (1994) and were further elaborated on from various vantages, more recently by Paul and Shankar (2018) and Tsionas and Mamatzakis (2019). This formulation can be adapted to a semiparametric context by constructing a weighted average of these CDFs across $G$ distinct groups. Naturally, as $G$ increases the weighted average takes on more variation, but also reduces bias due to the assumed parametric form of the CDF. More specifically we have

$$r_i = \sum_{g=1}^{G} \Phi(\boldsymbol{w}_i'\boldsymbol{\delta}_{g,1})\delta_{g,2} \Rightarrow -u_i = \log \sum_{g=1}^{G} \Phi(\boldsymbol{w}_i'\boldsymbol{\delta}_{g,1})\delta_{g,2},\ i = 1,\ldots,n, \tag{2}$$

where $\delta_{g,2} \in \mathbb{R}$, $g \in \mathbb{G} = \{1,\ldots,G\}$; $\boldsymbol{\delta}_{g,1} \in \mathbb{R}^{d_w}$ are unknown parameters and the number of groups, $G$, and the weights $\delta_{g,2}$ are also unknown. We redefine $\boldsymbol{\delta} = [\boldsymbol{\delta}_{g,1}', \delta_{g,2}]'$ for $g \in \mathbb{G}$.

From the approximation theory surrounding artificial neural networks (ANN) (e.g., see Hornik, 1993 and Hornik et al., 1994), the formulation in Equation (2) can approximate arbitrarily

well the unknown functional form between efficiency and the covariates $\boldsymbol{w}_i$. This specification can help alleviate any biases arising due to estimation of the SFM with one of the usual distributional assumptions commonly deployed, viz. that the one-sided error term follows a specific distribution like the Half Normal, Truncated Normal or Gamma, for example. In our formulation $G$ is selected in a data-driven fashion, making the proposed approach adaptive.

3.1.2. *Stochastic Inefficiency.* If one were uncomfortable making the distinctation that technical inefficiency were purely deterministic with respect to a set of covariates, and alternative would be to use the transformation $s_i = \ln \frac{1-r_i}{r_i}$ where $r_i = e^{-u_i} \in (0,1)$ represents technical efficiency, and it is defined in $\mathbb{R}$ unlike $r_i$ which is between zero and one. We assume

$$\ln \tfrac{1-e^{-u_i}}{e^{-u_i}} | \mu_i, \sigma \sim \mathcal{N}(\mu_i, \sigma^2),\ i = 1, \ldots, n, \tag{3}$$

where $\mu_i$ and $\sigma$ are, respectively, location and scale parameters.

An alternative to (3) is to use the environmental variables in $\boldsymbol{w}_i$:

$$\ln \tfrac{1-e^{-u_i}}{e^{-u_i}} | \boldsymbol{w}_i, \boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{w}_i'\boldsymbol{\gamma}, \sigma_u^2), i = 1, \ldots, n, \tag{4}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{d_w}$ is a parameter vector and $\sigma_u$ is a scale parameter. This setup is close in spirit to the deterministic setup in Equation (2), but allows for omitted variation.

3.2. **Specification of the Noise Distribution.** The unknown distribution of noise can be parameterized, generically, using smooth mixtures of Normal distributions (Geweke and Keane, 1997, 2007; Villani et al., 2009; Norets, 2010).[5] To adapt this interesting and novel approach to our context, we assume there are $M$ groups in the data and

$$y_i | \boldsymbol{w}_i, u_i \sim \mathcal{N}\left(\boldsymbol{x}_i'\boldsymbol{\beta}_m - u_i,\ \sigma_{vim}^2(\boldsymbol{w}_i; \boldsymbol{\nu}_m)\right), \text{with probability } p_{im}(\boldsymbol{w}_i; \boldsymbol{\eta}_m), \tag{5}$$

where $m \in \mathbb{M} = \{1, \ldots, M\}$, with $M$ denoting the number of groups, while $\boldsymbol{\beta}_m \in \mathbb{R}^K$ and $\boldsymbol{\nu}_m \in \mathbb{R}^{d_{\nu_m}}$ contain unknown parameters, which we denote as a vector $\boldsymbol{\nu} = [\nu_1', \ldots, \nu_M']' \in \mathbb{R}^{d_{\nu_m}}$. Notice that the variance in each group and the probability that observation $i$ is in this group are both functions of the covariates with respective parameters $\boldsymbol{\nu}_m$ and $\boldsymbol{\eta}_m$ ($m \in \mathbb{M}$).

[5]Alternatively, one could try other approximation devices, provided they integrate to one to preserve the construction of a density.

We specify

$$p_{im}(\boldsymbol{w}_i;\boldsymbol{\eta}_m) = \frac{e^{\boldsymbol{w}_i'\boldsymbol{\eta}_m}}{\sum\limits_{m_0\in\mathbb{M}} e^{\boldsymbol{w}_i'\boldsymbol{\eta}_{m_0}}}, \quad m\in\mathbb{M},$$

using, without loss of generality, $\boldsymbol{\eta}_1 = \mathbf{0}$.[6] Furthermore, we denote $\boldsymbol{\beta} = [\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_M']' \in \mathbb{R}^{d_\beta}$, $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1',\ldots,\boldsymbol{\gamma}_M']' \in \mathbb{R}^{d_\gamma}$, and $\boldsymbol{\eta} = [\boldsymbol{\eta}_1',\ldots,\boldsymbol{\eta}_M']' \in \mathbb{R}^{d_\eta}$, while for the variances, we assume

$$\sigma^2_{vim}(\boldsymbol{w}_i;\boldsymbol{\nu}_m) = e^{\boldsymbol{w}_i'\boldsymbol{\nu}_m}, \quad m\in\mathbb{M}, \tag{6}$$

where $\boldsymbol{\delta}_m \in \mathbb{R}^{d_{\nu_m}}$ is a vector of parameters.[7] As pointed out by Norets (2010), these are special cases of so-called "mixtures of experts" models in statistics and computer science.

3.3. **Unknown functional form.** The framework in Equation (5) allows for adaptive estimation of the production frontier through the number of groups $M$. As $M$ is allowed to increase, this will increase the ability of the smoothed mixture of Normals to detect local curvature of the frontier.

The key to allow for full flexibility for the estimation of the frontier is to allow $M$ to grow as $n$ increases. This will ensure that the frontier is treated akin to a fully nonparametric framework. The approach here is to select $M$ using the marginal likelihood. To explain the marginal likelihood, suppose we have data $Y$, parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{d_\theta}$, a likelihood function $\mathcal{L}(\boldsymbol{\theta};Y)$ and a prior $p(\boldsymbol{\theta})$. By Bayes' theorem the posterior is $p(\boldsymbol{\theta}|Y) \propto \mathcal{L}(\boldsymbol{\theta};Y)p(\boldsymbol{\theta})$. The marginal likelihood is the normalizing constant of the posterior, viz. $\mathcal{M}(Y) = \int_\Theta \mathcal{L}(\boldsymbol{\theta};Y)p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$. Given two models say "1" and "2" with different parameters, likelihoods and priors, the Bayes factor in favor of model "1" and against model "2", given the same data, when the prior odds are 1:1, is $BF_{1:2} = \frac{\mathcal{M}_1(Y)}{\mathcal{M}_2(Y)}$, also known as the posterior odds ratios.[8] The marginal likelihood can be estimated using the so-called "candidate's formula" which is an identity for all $\boldsymbol{\theta} \in \Theta$, viz. $\mathcal{M}(Y) = \frac{\mathcal{L}(\boldsymbol{\theta};Y)p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|Y)}$. The numerator is easy to compute at, say, the posterior mean. As the denominator is unknown it can be evaluated using the Laplace approximation (DiCiccio et al., 1997).

The rate of growth of $M$ as $n$ increases is a theoretical construction as we have a given $n$ so we need to select the best value of $M$ using some criterion (the marginal likelihood in our case).

[6]Our approach here is not reliant on a logit specification and could be deployed, alternatively, with a probit or any other flexible function that is bounded between 0 and 1.

[7]Here, it is worth recalling that Parmeter and Zelenyuk (2019, section 2.3) put special emphasis on heteroskedasticity in the one-sided error term although they do discuss heteroskedasticity of $v_i$ as well, and here we will also look at these from different angles.

[8]See Zellner (1971, chapter 10, in particular p. 293); and Kass and Raftery (1995).

As noted in Fong and Holmes (2020) marginal likelihood can be demonstrated to be equivalent to a near exhaustive leave-$p$-out cross-validation framework, provided this leave-$p$-out procedure is averaged over all values of $p$ and all hold out prediction sets. Doing so allows the use of the log posterior model predictive probability to operate as a scoring rule. Further, Fong and Holmes (2020) show that this log posterior predictive score is the only coherent scoring rule under data exchangeability. Our use here of their implementation offers new insights into nonparametric Bayesian applications by formally linking the marginal likelihood (a Bayesian construct) to cross-validation (a Frequentist construct).

Lastly, we end by noting that the number of smoothed Normals that we have assumed can differ between the estimation of the frontier and the estimation of the distribution of $v$. While we have implicitly left the number of smooth Normals as $M$, in practice we would have $M_1$ mixtures for the frontier and $M_2$ mixtures for the unknown noise distribution. This introduces some subtle notational complexities but does not impact the overall construction or use of the estimator.

3.4. **Endogeneity.** Although endogeneity is not the main focus in the vast majority of SFA and DEA studies (including Parmeter and Zelenyuk, 2019), in practice, it is an important issue if one is concerned with obtaining consistent and reliable estimates. To understand the importance of endogeneity, we write (5) as follows

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_m - u_i + \sigma_{vim}(\boldsymbol{w}_i;\boldsymbol{\nu}_m)\xi_{im}, \text{ with probability } p_{im}(\boldsymbol{w}_i;\boldsymbol{\eta}_m),\ m \in \mathbb{M}, \tag{7}$$

where $\xi_{im}$ are mutually independent for all $i$ and $m$ with zero (conditional) mean and unit (conditional) variance (assuming $\boldsymbol{w}_i$ contains an intercept) but not necessarily independent of $\boldsymbol{x}_i$.

The endogeneity of inputs is well known in the production economics literature and has received a lot of attention over the years, see Gandhi, Navarro, and Rivers (2020) and the references cited therein for a recent discussion. Only recently has the efficiency community embraced the importance of endogeneity.[9] Our approach is to couple Equation (7) with a reduced form

$$\boldsymbol{x}_i = \Pi(\boldsymbol{q}_i;\boldsymbol{\varpi}) + \boldsymbol{V}_i, \tag{8}$$

[9]See Kutlu (2015), Amsler, Prokhorov and Schmidt (2016, 2017), Griffiths and Hajargasht (2016) and the special issue of *Journal of Econometrics* (Kumbhakar and Schmidt, 2016) on this topic.

where $\Pi(\cdot;\boldsymbol{\varpi}) : \mathbb{R}^{d_q} \to \mathbb{R}^K$, $\boldsymbol{\varpi} \in \mathbb{R}^{d_\varpi}$ is a vector of parameters, and $\boldsymbol{V}_i$ is an error term supported in $\mathbb{R}^K$. The reduced form relates the endogenous variables $\boldsymbol{x}_i$ to the predetermined variables $\boldsymbol{q}_i$ via a possibly nonlinear form.[10] To account for endogeneity, it is not enough to relate the $\boldsymbol{x}_i$s to the $\boldsymbol{q}_i$s. In fact, we need to account for dependence between $\boldsymbol{V}_i$ and $\boldsymbol{\xi}_i = [\xi_{i1}, \ldots, \xi_{iM}]'$.[11] If, in fact, $\boldsymbol{z}_i$ in (6) does *not* contain an intercept then the problem is simplified as we do not have to assume that the elements of $\boldsymbol{\xi}_i$ have unit variance. This is convenient as we can now assume:

$$[\boldsymbol{V}_i', \boldsymbol{\xi}_i']' \sim \mathcal{N}_{K+M}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is a covariance matrix whose $M \times M$ southeastern submatrix is diagonal to take into account the mutual independence of elements of $\boldsymbol{\xi}_i$. In general we have

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{VV}} & \boldsymbol{\Sigma}_{\boldsymbol{V\xi}} \\ \boldsymbol{\Sigma}_{\boldsymbol{V\xi}}' & \boldsymbol{\Sigma}_{\boldsymbol{\xi\xi}} \end{bmatrix},$$

where the submatrix $\boldsymbol{\Sigma}_{\boldsymbol{V\xi}}$ accounts for correlations between elements of $\boldsymbol{V}_i$ and $\boldsymbol{\xi}_i$, and $\boldsymbol{\Sigma}_{\boldsymbol{\xi\xi}}$ is diagonal with elements $\sigma^2_{\xi 1}, \ldots, \sigma^2_{\xi M}$. These are, of course, unknown. In modeling endogeneity, the elements of $\boldsymbol{\Sigma}_{\boldsymbol{V\xi}}$ are essential.

As the specification of (8) is important, we consider two strategies: (i) a linear reduced form (LRF) and a nonlinear reduced form (NLRF). The LRF can be modeled as:

$$\boldsymbol{x}_i = \Pi \boldsymbol{q}_i + \boldsymbol{V}_i. \tag{9}$$

This setup follows the common convention in the majority of applied economic work. An alternative is to use an ANN which we can write as follows

$$x_{ki} = \boldsymbol{q}_i' \boldsymbol{\varpi}_{k0} + \sum_{g=1}^{G} \psi(\boldsymbol{q}_{it}' \boldsymbol{\varpi}_{kg1}) \varpi_{kg2} + V_{ik},\ k = 1, \ldots, K, \tag{10}$$

where $\boldsymbol{\varpi}_{k0}$ and $\boldsymbol{\varpi}_{kg1} \in \mathbb{R}^{d_q}$, $\varpi_{kg2} \in \mathbb{R}$ $(g = 0, 1, \ldots, G)$ are unknown parameters, and the "activation function" is $\psi(s) = \frac{1}{1+e^{-s}}$ for $s \in \mathbb{R}$. An assumption that $G = 0$ means that only the linear component is present in (10). We redefine $\boldsymbol{\varpi} \in \mathbb{R}^{d_\varpi}$ to include all parameters in the set of

[10] Notice that if the relationship were linear we would have $\Pi(\boldsymbol{q}_i;\boldsymbol{\varpi}) = \Pi \boldsymbol{q}_i$, where the reduced form parameters $\Pi \in \mathbb{R}^{d_q \times d_q}$.

[11] On this point, see Kleibergen and van Dijk (1993).

equations in (10). Finally, $V_{ik}$ is the $k^{th}$ element of $\boldsymbol{V}_i$. For simplicity in implementation, we keep the same number of ANN nodes ($G$) throughout, although assuming different values is possible.

Of course, there are certain issues[12] with the reduced form that arise in implementation, whether linear or not. The first issue is whether we have "weak instruments" which can result in erratic behavior of parameter estimates and posterior densities in finite samples (Kleibergen and van Dijk, 1993). The second issue is whether the instruments ($\boldsymbol{q}_i$) are correlated with the endogenous variables ($\boldsymbol{x}_i$), viz. whether they are "relevant". These issues are taken up in Appendix A.3. Finally, we note that we could also allow elements of $\boldsymbol{z}$ to be endogenous as well. We leave this as an extension for future research and refer the reader to Amsler, Prokhorov and Schmidt (2017) for further details in a parametric, frequentist setup.

3.5. **Imposing shape constraints.** With the setup described above we can fully estimate the stochastic production frontier without distributional assumptions on $v$ or $u$[13] and we can estimate the unknown frontier in a nonparametric fashion. However, even with this flexibility, we still face the potential that our estimated frontier may not respect traditional axioms of production (monotonicity/convexity of a given input). Thus it is imperative to focus on how to impose these types of constraints on the estimated frontier.

The average frontier is given as

$$\mathcal{E}\left(y_i|\boldsymbol{x}_i,\boldsymbol{z}_i\right)=\sum_{m\in\mathbb{M}}p_{im}(\boldsymbol{w}_i;\boldsymbol{\eta}_m)\left(\boldsymbol{x}_i'\boldsymbol{\beta}_m-u_i\right), \tag{11}$$

where $u_i$ is given by (2), and $\mathcal{E}(\cdot)$ denotes expectation (and $\mathcal{E}(\cdot|\cdot)$ denotes conditional expectation). One important class of models is when, in fact, $p_{im}(\boldsymbol{w}_i;\boldsymbol{\eta}_m)$ depends only on $\boldsymbol{z}_i$ (a form of separability). In this case, if $\boldsymbol{x}_i'\boldsymbol{\beta}_m$ satisfies monotonicity and concavity, then the average frontier in (11) will automatically satisfy these properties as well. Using the convention that $y_i$ and $\boldsymbol{x}_i$ are in logs, the simplest possible choice is a Cobb-Douglas production function with coefficients $\boldsymbol{\beta}_m$ ($m\in\mathbb{M}$). Although this choice is convenient it can, potentially, result in a large number of groups ($M$) especially with big (wide) data. Therefore, we explore a more general formulation – a

[12]See Staiger and Stock (1997) and Stock et al. (2002).
[13]Technically we require a parametric mixture, but allowing $G$ to grow effectively mutes this.

translog functional form:

$$f(\boldsymbol{x}_i;\boldsymbol{\beta}) = \beta_{0,m} + \boldsymbol{\beta}'_{1,m}\boldsymbol{x}_i + \tfrac{1}{2}\boldsymbol{x}'_i\boldsymbol{B}_m\boldsymbol{x}_i,\ \forall m \in \mathbb{M},$$

where $\beta_{0,m} \in \mathbb{R}$, $\boldsymbol{\beta}_{1,m} \in \mathbb{R}^K$ and $\boldsymbol{B}_m \in \mathbb{R}^{K\times K}$ is a symmetric matrix. All these parameters[14] are subsumed in the definition of $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$.

For the translog setup the parameter space $\mathcal{B}$ depends on the monotonicity and curvature restrictions, and our approach will seek to impose the desired constraints on a selected grid of points, using rejection sampling within our MCMC.[15] That is, we select a set $\mathcal{X} = [\boldsymbol{x}_\iota,\ \iota = 1,\dots,\bar{I}]$ from which to check the requisite curvature conditions of our translog functional form. One of the points is the component-wise mean of the data, and the other $\bar{I}-1$ points are randomly selected. Furthermore, for simplicity we proceed on the assumption that $p_{im}(\boldsymbol{w}_i;\boldsymbol{\eta}_m)$ and the variances in (6) depend only on $\boldsymbol{z}_i$. At each draw of the chain, we check if the curvature conditions are satisfied everywhere on the grid. If the constraints are satisfied, this set of parameters is kept, otherwise it is discarded.

The choice of a translog within a group is convenient because: (i) the translog is a second-order approximation to arbitrary functions;[16] (ii) groups are relatively homogeneous with monotone – concave frontiers which (locally, at least) the translog can approximate very well; (iii) with the translog it is relatively straightforward to impose monotonicity and concavity at a specified number of points.[17]

## 4. The Use of DEA in Estimation

As we have seen above, an estimation methodology exists that allows us to circumvent many of the unpopular or contentious assumptions that are levied against the stochastic frontier model. Even with these impediments removed (or mitigated), one may still prefer DEA to our proposed

[14]It is, of course, possible to include firm- and time-specific effects or a time trend when we have panel data. The time trend should be included along with its square and interaction with all other $\boldsymbol{x}_i$s.
[15]Flexibility of the algorithm will be compromised if we make the grid too fine.
[16]At a substantial increase in (computational/notational) complexity due to the number of interactions needed to truly approximate a multivariate problem. alternative functional forms could be deployed. Moreover, Sickles and Zelenyuk (2019, pg. 181) discuss the approximation powers of the translog functional form.
[17]When the log data are in deviations from their means, then monotonicity holds if the first order coefficients are non-negative. For imposition of concavity, see Diewert and Wales (1987).

model. Rather, we advocate that DEA can be used to craft robust priors for the stochastic frontier estimator just described. This avenue offers a best-of-both worlds appeal.

To construct these DEA based priors, we need to estimate inefficiency from a baseline deterministic frontier model. The obvious route is either FDH or DEA. Specifically, the deterministic frontier estimator of the Farrell output oriented technical efficiency for an observation $(\boldsymbol{x}^o, Y^o)$ is given by

$$\widehat{OTE}(\boldsymbol{x}^o, Y^o) = \max_{\lambda,\varrho_1,\ldots,\varrho_n} \{\lambda : \sum_{i=1}^{n} \varrho_i Y_i \geq \lambda Y^o; \quad \sum_{i=1}^{n} \varrho_i \boldsymbol{x}_i \leq \boldsymbol{x}^o; \quad \lambda \geq 0; \quad (\varrho_1, \ldots, \varrho_n) \in \mathcal{Z}\}, \tag{12}$$

where $\mathcal{Z}$ is the set of restrictions on the intensity variables $(\varrho_1, \ldots, \varrho_n)$, which can be used to impose various shape constraints on the production relationship. In particular, if $\mathcal{Z} = \{(\varrho_1, \ldots, \varrho_n) : \kappa_1 \leq \sum_{i=1}^{n} \varrho_i \leq \kappa_2, \varrho_i \geq 0, \forall \varrho_i\}$ and if $\kappa_1 = 0$ and $\kappa_2 = \infty$ (i.e., unbounded) then the assumption of constant returns to scale (CRS) is imposed on the estimated production relationship. Meanwhile, if $\kappa_1 = 0$ and $\kappa_2 = 1$ then non-increasing returns to scale is assumed, while if $\kappa_1 = 1$ and $\kappa_2 = 1$ then variable returns to scale (VRS) is assumed. Finally, if in addition to requiring $\kappa_1 = \kappa_2 = 1$, we require $\varrho_i \in \{0, 1\}$ $\forall i$, convexity is relaxed leading to the FDH estimator (Deprins et al., 1984).[18]

Our use for these estimates is to calibrate a prior for the coefficients in (2). As we use the same data for calibrating the prior and performing posterior analysis, this is, at best, an empirical Bayes procedure. Our intention is to compare results across priors computed using various deterministic frontier estimators. Our benchmark prior is

$$p(\boldsymbol{\theta}) \propto \mathbb{I}_{\mathcal{B}}(\boldsymbol{\beta}) \cdot \prod_{m=1}^{M} \sigma_{\varrho,m}^{-1} \cdot |\boldsymbol{\Sigma}|^{-(M+K+1)/2} \cdot p(\boldsymbol{\theta}_*), \tag{13}$$

where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d_\theta}$ is the entire parameter vector, $\Theta$ is the parameter space (mainly affected by $\mathcal{B}$), $\boldsymbol{\theta}_*$ is the parameter vector excluding $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, its prior is $p(\boldsymbol{\theta}_*)$, and the scale parameters, $\mathbb{I}_{\mathcal{B}}(\boldsymbol{\beta}) = 1$ if $\boldsymbol{\beta} \in \mathcal{B}$, and zero otherwise. The prior for all other parameters $(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\varpi})$ is flat, and the prior for the scale parameters is improper and (relatively) uninformative (see Zellner, 1971, p. 225, equation (8.9) for $\boldsymbol{\Sigma}$).

To craft the DEA-based prior, we estimate the functional form in (2) using $\hat{r}_{i(j)}$ from DEA to obtain estimates $\hat{\boldsymbol{\gamma}}_{(j)}$ and their covariance matrix $\hat{V}_{(j)}$, where $(j)$ corresponds to the $j^{\text{th}}$ sub-sample

[18]For the statistical properties of the DEA and FDH estimators see the review of Simar and Wilson (2015).

from $\mathbb{J} = \{1, \ldots, J\}$ where $J$ is the total number of subsamples that we consider to calibrate a DEA prior.[19] Let $\hat{\boldsymbol{\gamma}}_* = [\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(J)}]'$. Our prior is then

$$\boldsymbol{\gamma} \sim \mathcal{N}_{d_\gamma}(a\hat{\gamma},\ h\hat{V}), \tag{14}$$

where $\hat{\boldsymbol{\gamma}} = \boldsymbol{\omega}'\hat{\boldsymbol{\gamma}}_{(j)}$, $\boldsymbol{\omega} \in \mathcal{S} = \{\boldsymbol{\omega} \in \mathbb{R}_+^J : \boldsymbol{\omega}'\mathbf{1}_J = 1\}$, $\mathbf{1}_J$ is a $J \times 1$ vector of ones, $\mathcal{S}$ is the boundary of the unit simplex in $\mathbb{R}^J$, $\boldsymbol{\omega}$ is a vector of weights to reflect the importance of each different method, $a \in \mathbb{R}$ and $h > 0$ are parameters whose selection will be defined below, and $\hat{V} = \omega\hat{V}_*\omega'$, $\hat{V}_* = \text{diag}[\hat{V}_{(j)}, j \in \mathbb{J}]$.

Parameters $a$ and $h$ are introduced to improve the fit of the model along with the weights $\boldsymbol{\omega}$ as follows. We randomly take 10,000 draws for $\boldsymbol{\omega}$ from $\mathcal{S}$ and 10,000 draws for $a$ and $h$, and we select the set of weights that maximizes the marginal likelihood or "evidence" of the model (see our earlier discussion of the marginal likelihood). We draw $a$ from a Normal distribution with mean one and diagonal covariance matrix whose diagonal elements are all equal to $10^4$; for $h$ we draw from $\log h \sim \mathcal{N}(0, 10^4)$. Regarding the selection of $G$, we specify $G = G_{\max} = 10$.[20] This procedure produces optimal values for $a$, $h$ and weights $\boldsymbol{\omega}$ which are used to calibrate our prior in (14). Regarding $\boldsymbol{\theta}_*$, which includes only $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$, we assume the following prior:

$$\boldsymbol{\theta}_* \sim \mathcal{N}_{dim(\boldsymbol{\theta}_*)}\left(\bar{\boldsymbol{\theta}}_*,\ h_*^2\boldsymbol{I}_{dim(\boldsymbol{\theta}_*)}\right), \tag{15}$$

where $\bar{\boldsymbol{\theta}}_*$ is the prior mean vector, the prior covariance matrix is a diagonal matrix with all diagonal elements equal to $h_*^2$, and $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix. In our benchmark prior, we have $\bar{\boldsymbol{\theta}}_* = \mathbf{0}_{dim(\boldsymbol{\theta}_*)}$ and $h_* = 10^4$, but we intend to perform posterior sensitivity analysis with respect to these prior parameters including $a$, $h$ and $\boldsymbol{\omega}$ (see Appendix A.4 where we also take up the issue of weak instruments and the validity of instruments in relation to (9) and (10)).

Alternatively, instead of conducting estimation separately for DEA- and SFA-based priors we may combine the two priors:

$$p(\cdot) = \phi p_{DEA}(\cdot) + (1 - \phi)p_{SFA}(\cdot), \tag{16}$$

[19] In the simulations and application we set $J = 20\%$ of the total number of observations.

[20] This is necessitated by the fact that in our MCMC procedure we try different values of $G$ in the interval $\{1, \ldots, G_{\max}\}$ to finally select the optimal value of $G$. For this reason we need to have a prior for $\boldsymbol{\gamma}$ that is fully defined for all $G$.

where $\phi \in [0,1]$, and the arguments are the same as in (14) and (15). We can treat $\phi$ as a parameter with a uniform distribution in the unit interval. We denote this prior as DEA/SFA. Values of $\phi$ will be useful in assessing whether DEA or SFA based priors are more useful.

4.1. **Constructing a Least Favorable Prior.** Ideally we would seek to use a more general prior, preferably a robust or least favorable prior. To make our discussion more precise, and in a more general notation, suppose $\theta \in \Theta \subset \mathbb{R}^p$ is a parameter vector and we are interested in estimating a function $g(\theta)$. Moreover, $d(y)$ is a decision rule that depends on the observation $y = \{y_i\}_{i=1}^n$, $\mathcal{D}$ is the set of feasible decision rules, $\{P_\theta : \theta \in \Theta\}$ denotes a set of prior distributions on the parameter, and the distribution of the observation given the parameters is denoted by $p(y;\theta)$. Let $\mathrm{d}Q_\pi(y,\theta) = \mathrm{d}P_\theta(y)\,\mathrm{d}\pi(\theta)$ and $S = \{Q_\pi : \pi \in \Gamma\}$ where $\Gamma$ is a set of "reasonable" or "relevant" distributions on $\Theta$. The estimation problem is

$$\min_{d\in\mathcal{D}} \max_{\theta\in\Theta} \int (g(\theta) - d(y))^2 \,\mathrm{d}P_\theta(y). \tag{17}$$

Alternatively, it is

$$\min_{d\in\mathcal{D}} \max_{\pi\in\Gamma} \int (g(\theta) - d(y))^2 \,\mathrm{d}P_\theta(y)\,\mathrm{d}\pi(\theta). \tag{18}$$

The interpretation is that "nature" chooses a least-favorable prior $\pi$ of $\Gamma$ and the researcher determines a decision rule (an estimator) which makes the maximum risk as small as possible. We consider a finite set of prior distributions $\{\pi_1, \ldots, \pi_J\}$ and take $\Gamma$ to be the convex hull

$$\Gamma = \left\{ \sum_{j=1}^J \delta_j \pi_j : 0 \le \delta_j \le 1,\ j = 1, \ldots, J,\ \sum_{j=1}^J \delta_j = 1 \right\}.$$

Similarly, we work with a grid $\{\theta_1, \ldots, \theta_J\} \subset \Theta$, $Q_j = P_{\theta_j}$ and $\mathcal{S}$ being the convex hull of $\{Q_1, \ldots, Q_J\}$. If the set of decision rules $\mathcal{D}$ is unrestricted, then, via Bayes rule, the posterior mean of $g$ is:

$$d^\delta(y) = \frac{\sum_{j=1}^J g(\theta_j) p(y;\theta_j)\delta_j}{\sum_{j=1}^J p(y;\theta_j)\delta_j}, \tag{19}$$

with $\pi_j$ assigning unit mass to $\theta_j$. The risk of this estimator is

$$r(Q^{\delta}, d) = \sum_{j=1}^{J} \delta_j r(P_{\theta_j}, d), \tag{20}$$

where $\delta_j$s are least-favorable prior probabilities, and $r(P_\theta, d) = \int [g(\theta) - d(y)]^2 \, p(y;\theta) \, \mathrm{d}y$. We can use Monte Carlo simulation to estimate $r(P_\theta, d)$. If $\{\upsilon(\theta, s)\}_{s=1}^{S}$ is a set of independent and identically distributed draws from $p(y;\theta)$, we obtain

$$r(P_\theta, d) \simeq S^{-1} \sum_{s=1}^{S} [g(\theta) - d(\upsilon(\theta, s))]^2 . \tag{21}$$

In turn, we compute $\rho(\delta) = r(Q^{\delta}, d^{\delta})$ from (19) and (20). Therefore, the maximizing value $\delta^o$ will be the least-favorable prior, and $\rho(\delta^o)$ gives the minimax value for risk, relative to $\{\theta_1, \ldots, \theta_J\}$.

Suppose the DEA inefficiency estimates are $\hat{u}_i$ $(i = 1, \ldots, n)$. We use these estimates along with, say, the specification in Equation (3) to construct a least-favorable prior inefficiency distribution. Let

$$\mu_i = a + b\hat{u}_i, \; \sigma^2 = h^2 S^2, \tag{22}$$

where $S^2$ is the sample variance of $\ln \frac{1-e^{-\hat{u}_i}}{e^{-\hat{u}_i}}$, where $a$, $b$ and $h$ are parameters. The meaning of this prior is that in (22) we use DEA estimates $\hat{u}_i$ but we adjust them through the parameters $a$, $b$, and $h$. Parameter $a$ stands for an overall bias correction, parameter $b$ for a slope correction, while parameter $h$ adjusts the variance of the DEA estimates. If the DEA estimates provide a least-favorable approximation we expect that $a = 0$, $b = 1$ and $h = 1$. Otherwise, the least-favorable prior has to be defined through an appropriate adjustment of the DEA inefficiency estimates.

## 5. Monte Carlo Simulations

A benefit of our Bayesian ANN approach is that it can easily accommodate endogeneity if instruments are available. For our simulation framework we use the DGP of Amsler, Prokohorov and Schmidt (2017, APS hereafter).[21] APS proposed two ways for estimating their model with endogeneity: IV and MLE. They acknowledged that both approaches are not simple "... because a specific copula must be assumed to model the correlation of $u_i^o$ with the endogenous variables,

[21]For performance when endogeneity is not present see Appendix A.1.

and because simulation methods are necessary to form the IV criterion function or the likelihood." Given that the current menu of semi- or nonparametric stochastic frontier estimators does not explicitly deal with endogeneity, we only focus on our proposed estimator.

The DGP of APS is as follows. They consider the baseline SFM with the scaling property invoked

$$y_i = \beta_0 + x_i'\boldsymbol{\beta} + v_i - u_i, \quad u_i = u_i^o e^{q_i'\boldsymbol{\delta}},$$

where the basic inefficiency term (Wang and Schmidt, 2002) $u_i^o$ is distributed as Half Normal, $\mathcal{N}_+(0, \sigma_u^2)$, and $v_i$ is distributed as $\mathcal{N}(0, \sigma_v^2)$. To introduce endogeneity, APS partition both $x$ and $q$ into exogenous and endogenous components

$$\boldsymbol{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} \quad \boldsymbol{q}_i = \begin{bmatrix} q_{1i} \\ q_{2i} \end{bmatrix},$$

where $x_{1i}$ and $q_{1i}$ are exogenous and $x_{2i}$ and $q_{2i}$ are endogenous. The instruments are termed $\boldsymbol{z}_i = [1, x_{1i}, q_{1i}, w_i]'$, where $w_i$ are the outside instruments. Finally, the endogenous variables are generated as

$$\begin{aligned} x_{2i} =& \Pi_x'\boldsymbol{z}_i + \eta_i \\ q_{2i} =& \Pi_q'\boldsymbol{z}_i + \tau_i. \end{aligned}$$

Endogeneity is introduced by allowing $v_i$, $\tau_i$ and $\eta_i$ to be correlated with each other, defined as $\rho$. Relative to APS we have a different model as, among other things, our reduced form estimator is more flexible but we avoid the use of copulas which places certain limitations on the type of dependence we can model. Moreover, from our point of view, in terms of inefficiency, our approach is broader as we do not anchor to a particular model but we couple it with efficiency priors that can, potentially, deliver better results in finite samples.

For our simulations we use the small variations of the "base case" of APS which consists of setting $\beta_0 = 0$, $\boldsymbol{\beta} = (0.61, 0.61)'$, and $\Pi_x = \Pi_q = (0, \alpha, \alpha, \alpha, \alpha)$ where $\alpha \in \{0, 0.116, 0.316\}$. We generate our exogenous covariates, $x_{1i}$ and $q_{1i}$ along with the instruments $w_{1i}$ and $w_{2i}$ as $\mathcal{N}(0, 1)$ marginals and correlation equal to 0.5. The errors, $\eta$, $\tau$ and $v$ are all distributed $\mathcal{N}(0, 1)$ with correlation $\rho \in \{0, 0.25, 0.50\}$. We select $\delta_1 \in \{0.0, 0.5, 1.0\}$, define $\boldsymbol{\delta} = (\delta_1, \delta_2)$ where we always

TABLE 1. Mean RMSE of conditional frontier across 200 simulations.

| $\alpha$ | $n \downarrow$ | $\boldsymbol{\delta} = 0$ | | | $\boldsymbol{\delta} = 0.5$ | | | $\boldsymbol{\delta} = 1.00$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho \rightarrow$ | 0 | 0.25 | 0.5 | 0 | 0.25 | 0.5 | 0 | 0.25 | 0.5 |
| | 100 | 0.251 | 0.289 | 0.31 | 0.333 | 0.394 | 0.42 | 0.479 | 0.51 | 0.699 |
| | | **0.288** | **0.303** | **0.325** | **0.33** | **0.329** | **0.351** | **0.557** | **0.368** | **0.402** |
| | 200 | 0.182 | 0.209 | 0.224 | 0.248 | 0.294 | 0.313 | 0.348 | 0.37 | 0.507 |
| **0** | | **0.214** | **0.226** | **0.242** | **0.242** | **0.242** | **0.258** | **0.395** | **0.26** | **0.285** |
| | 400 | 0.13 | 0.149 | 0.16 | 0.175 | 0.208 | 0.221 | 0.25 | 0.266 | 0.364 |
| | | **0.159** | **0.168** | **0.18** | **0.173** | **0.172** | **0.184** | **0.28** | **0.184** | **0.202** |
| | 800 | 0.09 | 0.103 | 0.111 | 0.123 | 0.146 | 0.156 | 0.177 | 0.189 | 0.258 |
| | | **0.114** | **0.12** | **0.128** | **0.119** | **0.119** | **0.127** | **0.195** | **0.129** | **0.141** |
| | 100 | 0.387 | 0.394 | 0.421 | 0.498 | 0.504 | 0.623 | 0.666 | 0.706 | 0.76 |
| | | **0.338** | **0.392** | **0.466** | **0.625** | **0.465** | **0.531** | **0.73** | **0.581** | **0.714** |
| | 200 | 0.29 | 0.295 | 0.316 | 0.374 | 0.379 | 0.467 | 0.5 | 0.53 | 0.57 |
| **0.116** | | **0.253** | **0.293** | **0.349** | **0.46** | **0.342** | **0.39** | **0.554** | **0.441** | **0.543** |
| | 400 | 0.21 | 0.214 | 0.229 | 0.271 | 0.275 | 0.343 | 0.363 | 0.389 | 0.419 |
| | | **0.185** | **0.215** | **0.256** | **0.33** | **0.246** | **0.281** | **0.417** | **0.332** | **0.408** |
| | 800 | 0.15 | 0.153 | 0.162 | 0.194 | 0.194 | 0.244 | 0.256 | 0.277 | 0.298 |
| | | **0.136** | **0.157** | **0.187** | **0.235** | **0.175** | **0.2** | **0.295** | **0.235** | **0.289** |
| | 100 | 0.423 | 0.458 | 0.474 | 0.494 | 0.508 | 0.525 | 0.762 | 0.771 | 0.91 |
| | | **0.349** | **0.51** | **0.467** | **0.433** | **0.624** | **0.536** | **0.478** | **0.659** | **0.659** |
| | 200 | 0.32 | 0.348 | 0.359 | 0.375 | 0.384 | 0.409 | 0.578 | 0.6 | 0.708 |
| **0.316** | | **0.269** | **0.392** | **0.36** | **0.336** | **0.485** | **0.417** | **0.37** | **0.51** | **0.51** |
| | 400 | 0.241 | 0.263 | 0.27 | 0.283 | 0.29 | 0.307 | 0.437 | 0.451 | 0.532 |
| | | **0.204** | **0.298** | **0.273** | **0.243** | **0.35** | **0.301** | **0.276** | **0.38** | **0.38** |
| | 800 | 0.171 | 0.189 | 0.192 | 0.203 | 0.206 | 0.222 | 0.314 | 0.325 | 0.384 |
| | | **0.15** | **0.219** | **0.201** | **0.176** | **0.253** | **0.218** | **0.202** | **0.278** | **0.278** |

Notes: Numbers in regular font correspond to the LRF specification while numbers in bold represent the NLRF specification.

have $\delta_1 = \delta_2$ and fix $\sigma_u^2 = \pi/(\pi - 2)$. The results are presented for sample sizes 100, 200, 400 and 800.

Tables 1 and 2 contain basic simulation results when we follow APS and allow for endogeneity into both the covariates entering the frontier and the scaling function of the inefficiency term, respectively. We compare the root mean square error (RMSE) of the conditional frontier and the conditional mean of inefficiency in this setting across 200 simulations.

There are several interesting features stemming from these two tables. First, the performance of the NLRF specification works as well or better than the LRF for the majority of parameter combinations, even though the true model for the frontier is linearly specified. Second, we see reasonable decay in RMSE for both estimators as the sample size increases. We also observe that

TABLE 2. Mean RMSE of conditional mean of inefficiency across 200 simulations.

| $\alpha$ | $n \downarrow$ | $\boldsymbol{\delta} = 0$ | | | $\boldsymbol{\delta} = 0.5$ | | | $\boldsymbol{\delta} = 1.00$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho \rightarrow$ | 0 | 0.25 | 0.5 | 0 | 0.25 | 0.5 | 0 | 0.25 | 0.5 |
| | 100 | 0.344 | 0.383 | 0.405 | 0.407 | 0.424 | 0.496 | 0.563 | 0.661 | 0.804 |
| | | **0.345** | **0.383** | **0.405** | **0.432** | **0.449** | **0.53** | **0.568** | **0.586** | **0.715** |
| | 200 | 0.241 | 0.269 | 0.284 | 0.301 | 0.305 | 0.367 | 0.416 | 0.475 | 0.578 |
| **0** | | **0.258** | **0.287** | **0.303** | **0.317** | **0.33** | **0.389** | **0.421** | **0.434** | **0.53** |
| | 400 | 0.165 | 0.184 | 0.194 | 0.21 | 0.219 | 0.256 | 0.29 | 0.341 | 0.416 |
| | | **0.191** | **0.212** | **0.224** | **0.229** | **0.238** | **0.281** | **0.301** | **0.31** | **0.379** |
| | 800 | 0.113 | 0.126 | 0.133 | 0.144 | 0.152 | 0.176 | 0.199 | 0.238 | 0.289 |
| | | **0.14** | **0.156** | **0.164** | **0.165** | **0.17** | **0.201** | **0.212** | **0.219** | **0.267** |
| | 100 | 0.356 | 0.408 | 0.484 | 0.568 | 0.633 | 0.665 | 0.684 | 0.841 | 1.063 |
| | | **0.42** | **0.437** | **0.448** | **0.457** | **0.462** | **0.49** | **0.636** | **0.667** | **0.935** |
| | 200 | 0.268 | 0.303 | 0.364 | 0.422 | 0.464 | 0.487 | 0.514 | 0.625 | 0.778 |
| **0.116** | | **0.317** | **0.33** | **0.341** | **0.348** | **0.351** | **0.371** | **0.484** | **0.508** | **0.712** |
| | 400 | 0.194 | 0.221 | 0.264 | 0.308 | 0.334 | 0.351 | 0.373 | 0.456 | 0.56 |
| | | **0.222** | **0.231** | **0.254** | **0.258** | **0.259** | **0.26** | **0.361** | **0.373** | **0.523** |
| | 800 | 0.137 | 0.157 | 0.186 | 0.219 | 0.234 | 0.246 | 0.263 | 0.325 | 0.394 |
| | | **0.155** | **0.161** | **0.181** | **0.181** | **0.184** | **0.187** | **0.261** | **0.262** | **0.368** |
| | 100 | 0.415 | 0.641 | 0.641 | 0.725 | 0.779 | 0.889 | 0.963 | 0.99 | 1.007 |
| | | **0.478** | **0.507** | **0.507** | **0.53** | **0.552** | **0.612** | **0.665** | **0.784** | **0.831** |
| | 200 | 0.316 | 0.477 | 0.488 | 0.552 | 0.58 | 0.662 | 0.738 | 0.759 | 0.772 |
| **0.316** | | **0.362** | **0.384** | **0.384** | **0.39** | **0.406** | **0.469** | **0.489** | **0.601** | **0.637** |
| | 400 | 0.24 | 0.35 | 0.37 | 0.418 | 0.426 | 0.486 | 0.548 | 0.563 | 0.573 |
| | | **0.256** | **0.272** | **0.272** | **0.288** | **0.299** | **0.357** | **0.361** | **0.458** | **0.485** |
| | 800 | 0.169 | 0.254 | 0.261 | 0.295 | 0.309 | 0.353 | 0.392 | 0.403 | 0.41 |
| | | **0.18** | **0.191** | **0.191** | **0.206** | **0.214** | **0.258** | **0.26** | **0.334** | **0.354** |

Notes: Numbers in regular font correspond to the LRF specification while numbers in bold represent the NLRF specification.

the performance of the estimator degrades as the level of endogeneity increases ($\rho$ increases), the impact of determinants of inefficiency increases ($\boldsymbol{\delta}$ increases) and the strength of the instruments manifests (the elements of $\Pi_x$ increase). The rate of decay of RMSE also appears to be connected to the configuration of the parameters. There also does not appear to be a large difference in terms of RMSE regarding estimation of the frontier versus the conditional mean of inefficiency.

## 6. Empirical application

6.1. **Data and results.** We use the same data as in Malikov et al. (2015) where we have five inputs, five outputs. The data is an unbalanced panel with 2,397 observations for 285 large U.S. commercial banks (2001-2010), whose total assets were in excess of one billion dollars (in 2005 U.S. dollars) in their first three years in the data. The data come from Call Reports of the Federal

Reserve Bank of Chicago and are as follows. Outputs are $y_1$ = Consumer Loans, $y_2$ = Real Estate Loans, $y_3$ = Commercial & Industrial Loans, $y_4$ = Securities and $y_5$ = Off-Balance Sheet Activities Income. The inputs are: $x_1$ = Labor, number of full time Employees, $x_2$ = Physical Capital (Fixed Assets), $x_3$ = Purchased Funds, $x_4$ = Interest-Bearing Transaction Accounts and $x_5$ = Non-Transaction Accounts. We also have Equity Capital and Total Assets (TA) as additional covariates. All monetary valued inputs and outputs are in thousands of real USD (deflated to 2005 using CPI (all urban consumers)).

We estimate an output distance function of the form

$$y_{it1} = f(\tilde{\boldsymbol{x}}_{it}; \boldsymbol{\beta}) + v_{it} - u_{it},$$

where $\tilde{\boldsymbol{x}}$ includes all logs of inputs as well as $\tilde{y}_{it\ell} \equiv y_{it\ell} - y_{it1}$ $(\ell = 1, \ldots, F)$, a time trend, the log of equity, the log of non-performing loans and the log of total assets. All output variables are in logs as well. We estimate our model using different values for $G \in \{1, \ldots, G_{\max}\}$ and obtain the marginal likelihood values, $\mathcal{M}_G(\boldsymbol{S})$ where $\boldsymbol{S}$ denotes the entire data set. In turn, we obtain posterior model probabilities

$$\mathcal{P}_G(\boldsymbol{S}) = \frac{\mathcal{M}_G(\boldsymbol{S})}{\sum_{g'=1}^{G_{\max}} \mathcal{M}_{G'}(\boldsymbol{S})}, \text{ for } G \in \{1, \ldots, G_{\max}\}. \tag{23}$$

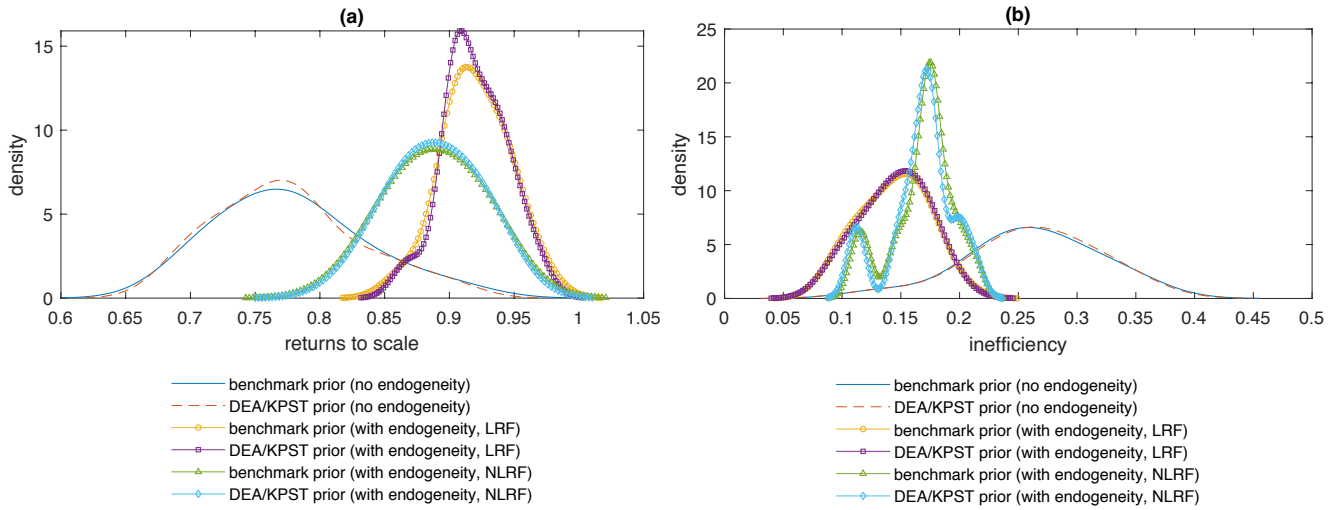
Next, all measures and quantities of interest can be obtained by combining results for different values of $G$ using (23). To implement MCMC, we use 150,000 iterations omitting the first 50,000 in a burn-in phase to mitigate possible start up effects. We report the results for both our benchmark prior as well as the prior that anchors on DEA or KPST in (14). We also examine the results under endogeneity of $\tilde{y}_{it\ell}$. $\boldsymbol{w}_{it}$, as in (2), includes the log of total assets, the log of equity, the log of non-performing loans, a time trend and all of their squares and interactions. Sample distributions of posterior mean estimates of key functions of interest are reported in Figure 1.[22]

A general important finding is that ignoring endogeneity results in very different estimates of returns to scale (RTS) and inefficiency.[23] For example, in panel (a), we see that the estimates of RTS are surprisingly low and close to 0.75 for the models that ignore endogeneity and technical

[22]See Table 4 for the full set of posterior means and standard deviations in Appendix A.2.

[23]RTS is estimated as the negative of the ratio of the sum of the input elasticities to the sum of the output elasticities, see Malikov et al. (2016, equation 39).

FIGURE 1. Sample distributions of posterior mean estimates. DEA and KPST stand for Data Envelopment Analysis, and Kumbhakar, Park, Simar and Tsionas (2007, KPST), respectively. LRF is the linear reduced form in (9) and NLRF corresponds to the flexible nonlinear specification of the reduced form in (10).



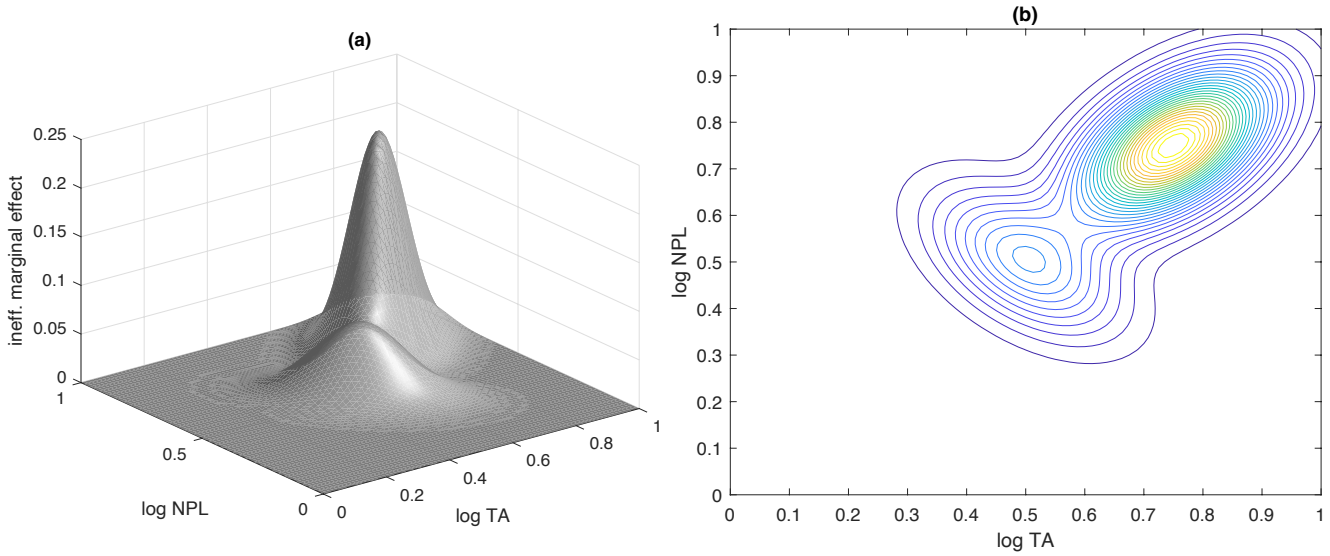

inefficiency (panel (b)) averages close to 30%. Taking endogeneity into account, the RTS estimates average slightly above 0.9 for LRF and slightly below 0.9 for NLRF. Meanwhile, estimated technical efficiency averages around 15% for LRF and 17% for NLRF. We note that for LRF the optimal $G$ is 3 and for NLRF $G$ is 2. Although one might have expected larger values, one has to keep in mind that this is a relatively homogeneous data set of US banks. This similarity of the estimates from LRF and NLRF suggests robustness to the specification of the form of endogeneity.

One important point that comes out of the empirical application is that the behavior of the new methods does not depend on whether we benchmark the prior in (14) using DEA. In this sense it appears that it does not matter too much what is the prior of $\boldsymbol{\gamma}$ provided it can be dominated by the data/likelihood and this is what happens in our application. This conclusion is supported by posterior sensitivity analysis which we undertake in Appendix A.4 and is another indication of robustness.

In Figure 2, we report sample distributions of posterior mean estimates of inefficiency effects (marginal effects) derived from (2) with respect to total assets (TA) and non-performing loans

(NPL) using the NLRF setup described above.[24] Both variables are in logs and they are normalized to lie in the interval $(0, 1)$ for visual clarity.

FIGURE 2. Sample distributions of posterior mean estimates of inefficiency effects via NLRF.




The interesting feature of the marginal effects, besides having a positive effect on inefficiency, is that there are two groups in the data. The first group contains banks with near average values of total assets and NPL, and the second group contains bank observations with a larger size and higher values of NPL. This can testify to the conjecture or claim that the larger banks among the large US banks were more exposed to risk. This is also consistent with the findings on raw inefficiency scores, where we saw a bimodal distribution (panel (b) in Figure 1 when using NLRF).

We also used two least favorable prior setups to assess the impact that weighting between DEA and SFA priors would have on the analysis. These results are contained in Appendix A.6. The main feature of the use of LFPs is that the location parameter $a$ is negative and the rescaling of DEA scores is based on a scale parameter $h$ that exceeds one with significant probability. Overall, the LFP results for RTS and inefficiency lead to similar conclusions: average estimated RTS is near 0.9 while average technical inefficiency is near 16%.

[24]Similar results were found using the LRF.

6.2. **A bridge too far?** Combining DEA and SFA in our context, may or may not be a "bridge too far". To decide whether this is the case, we use optimal model pools (Geweke and Amisano, 2011a,b). Given a set of models, say $m \in \mathbb{M} = \{1, 2, \ldots, M\}$ whose posterior predictive densities are $p_m(\boldsymbol{S}_o|\boldsymbol{S})$, where $\boldsymbol{S}$ is the data that we use for posterior inference and $\boldsymbol{S}_o$ is an out-of-sample set of observations that we want to use for prediction and/or model validation.

We consider predictive densities of the form:

$$\sum_{m\in\mathbb{M}} w_m p_m(\boldsymbol{S}_o|\boldsymbol{S}_{1:t});\ \sum_{m\in\mathbb{M}} w_m = 1;\ w_m \geq 0\ \forall m \in \mathbb{M}, \tag{24}$$

known as *linear opinion pools*. The data up to and including period $t$ are denoted $\boldsymbol{S}_{1:t}$. We consider using the log predictive score function:

$$\sum_{(i)} \log\left[\sum_{m\in\mathbb{M}} w_m p_m(\boldsymbol{S}_o|\boldsymbol{S}_{(i)})\right], \tag{25}$$

where $\boldsymbol{S}_{(i)}$ denotes the sample of available observations for posterior inference, indexed by $(i)$. The index $(i)$ depends on which observations are actually used in-sample.

The log predictive score function is a measure of the out-of-sample prediction record of the model. Maximizing (25) subject to (24) is quite feasible given nonlinear programming software. Therefore, the problem is:

$$\max_{\{w_m, m\in\mathbb{M}\}} \sum_{(i)} \log\left[\sum_{m=1}^{M} w_m p_m(\boldsymbol{S}_o|\boldsymbol{S}_{(i)})\right],$$
$$\text{subject to}$$
$$\sum_{m\in\mathbb{M}} w_m = 1;\ w_m \geq 0\ \forall m \in \mathbb{M}.$$

The objective is to maximize weighted log predictive scoring (or, equivalently, maximize out-of-sample performance) subject to the usual portfolio-like constraints on the weights attached to different models. The models considered do not have to contain the same parameters (or have any parameters for that matter) so this approach is quite general.[25]

[25]In our computations we used the `fortran 77` version of library `lbfgs` in `netlib`. The problem can be converted to unconstrained optimization by reparametrizing $w_m = \frac{\exp\left(-\omega_m^2\right)}{1+\sum_{m_0>1}\exp\left(-\omega_{m_0}^2\right)}$, and the $\omega_i$s are defined on the real line. Convergence was quite fast given any initial conditions and no other local optima were found.

In our application, we omit 10% of observations randomly and we obtain the posterior predictive density as follows:

$$p_m(\boldsymbol{S}_o|\boldsymbol{S}) = \int p_m(\boldsymbol{S}_o, \boldsymbol{\theta}_m|\boldsymbol{S})\mathrm{d}\boldsymbol{\theta} = \int p_m(\boldsymbol{S}_o|\boldsymbol{\theta}_m, \boldsymbol{S})p(\boldsymbol{\theta}_m|\boldsymbol{S})\mathrm{d}\boldsymbol{\theta}_m \simeq S^{-1}\sum_{s=1}^{S} p_m(\boldsymbol{S}_o|\boldsymbol{\theta}_m^{(s)}, \boldsymbol{S}),$$

where $\{\boldsymbol{\theta}_m^{(s)}, s = 1, \ldots, S\}$ is a set of MCMC draws that converges in distribution to the distribution whose density is the posterior $p_m(\boldsymbol{\theta}|\boldsymbol{S})$ ($m \in \mathbb{M}$). As we use SMN, $p_m(\boldsymbol{S}_o|\boldsymbol{\theta}_m^{(s)}, \boldsymbol{S})$ can be evaluated point-wise as it is available in closed form.

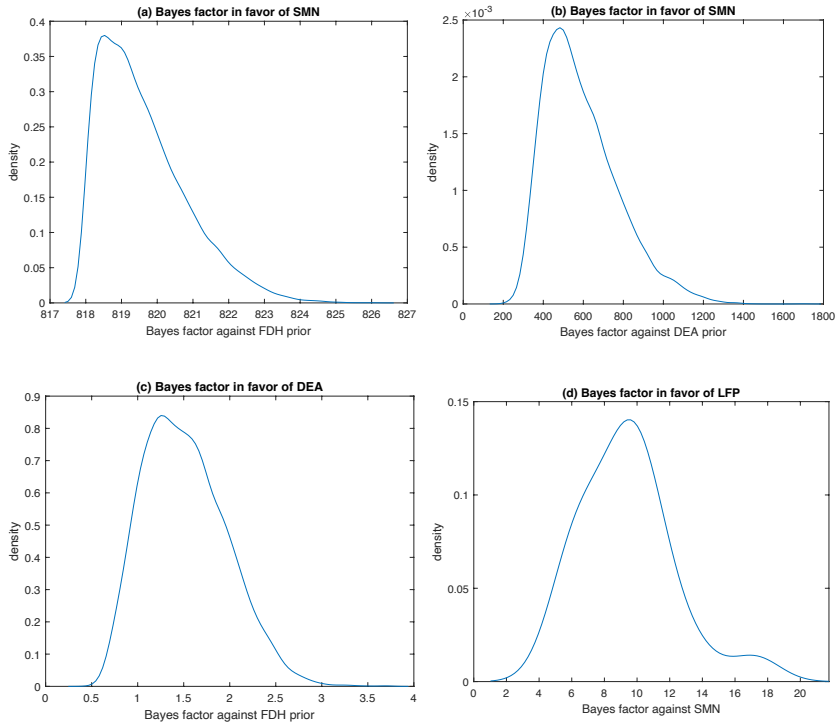
We repeat omitting 10% of the observations, randomly, 1,000 times and use as competing models (i) our model, (ii) DEA prior, (iii) FDH prior, (iv) KPST prior, and (v) the LFP. We account for endogeneity using (10).

It remains to consider a prior where the $\boldsymbol{\beta}_m$s and $\boldsymbol{\delta}_m$ as well as the inefficiency parameters are crafted using the DEA and FDH priors. The procedure has been described at the end of Section 2. We apply this procedure in 1,000 different sub-samples of the data omitting 10% of observations randomly, re-applying MCMC and computing the marginal likelihood of the SMN versus the marginal likelihood of the DEA and FDH priors. We take as a benchmark the FDH case so its marginal likelihood is normalized to 1. The resulting distributions of Bayes factors (ratios of marginal likelihood of DEA and SMN to the marginal likelihood of FDH) are reported in Figure 3.

The marginal likelihood in favor of the SMN model (and against the FDH prior, reported in panel (a)) ranges from 818 to 826 and has a median value of 819. The marginal likelihood in favor of the SMN model with the default prior (and against the FDH prior, reported in panel (b)) ranges from roughly from 220 to 1700 and the median is 560. As the Bayes factor in favor of DEA and against the FDH prior is $BF_{DEA:FDH} = \frac{BF_{SMN:DEA}}{BF_{SMN:FDH}}$, the respective density is presented in panel (c) of Figure 3; it ranges from 0.98 to 3.70 with a median 1.47 which does not provide strong evidence for or against either model. On the contrary, the posterior odds in favor of the SMN model and against DEA or FDH priors are overwhelming (Kass and Raftery, 1995).

We also see some evidence that the least favorable prior (LFP-I) has a small preference against the SMN prior (panel (d)): the Bayes factor ranges from 2 to about 20, with a median of just under 10. The similarity of the main technological characteristics of the results using DEA and

FIGURE 3. Log Bayes factors against alternative priors in 1,000 subsamples using the NLRF.




KPST priors, indicates that our core findings are robust across various priors. The results are also similar whether we account for endogeneity via LRF or NLRF, which are however substantially different from the results when ignoring the presence of endogeneity. It appears the key is that we must account for endogeneity, an aspect that many applied efficiency papers (both deterministic and stochastic) have ignored when studying bank performance with frontier methods.

## 7. Concluding Remarks

The contrast between DEA and SFA is likely to persist given the distinct differences in the formal models they estimate; and yet nonparametric methods provide a bridge that can be usefully deployed in practice based on Bayesian ANN. In this work, we propose a novel flexible model that (i) does not require distributional assumption on inefficiency; (ii) is based on a flexible functional form for the frontier using smooth mixtures of Normals which can approximate arbitrarily well any smooth function; (iii) accounts for endogeneity in a flexible manner; (iv) can incorporate

information from DEA, SFA, or both, in benchmarking an empirical Bayes prior and (v) has the ability to enforce axioms of production on the technology. We further supplement the approach by offering a least favorable prior within a class of frontier models which can be estimated prior to the Bayesian ANN. This allows the estimator to be minmax over this class of priors and simplifies the optimization (given the discreteness of the set of priors).

In the stochastic DEA approach (e.g., Simar and Zelenyuk, 2011), Parmeter and Zelenyuk (2019) demonstrate that one can first filter out the noise from the data using semi- or nonparametric SFA and then apply variants of DEA. We agree that this line of attack can be successfully implemented in practice and has its advantages although DEA-like methods like stochastic DEA or concave nonparametric least squares are designed to deliver the same thing. In this study, we have proposed a different use of DEA (possibly along with other similar methods) to craft an empirical Bayes prior for the parameters of inefficiency. This method has been found to work acceptably well in Monte Carlo simulations and the empirical application. However, it seems that our novel model does not depend critically on this aspect as the prior allows "domination" by the data/likelihood (despite the fact that it is based on the data themselves). It is in very small samples (close to 100 observations) where this use of DEA is found to be rather useful which is, in itself, an important point for practitioners.

While there exists a natural conflict between the underlying assumptions of the deterministic and stochastic frontier models, our use of combinations across model estimates to construct priors to inform an adaptive and flexible estimator should not be overlooked. We have presented some evidence and tools that can be used in empirical applications to answer such questions about how best to combine priors (and if it matters) which clearly extends beyond the field of efficiency analysis. Our proposal of using optimal model pools is designed to answer precisely such questions regarding choice of prior.

## References

[1] Amsler, C., Prokhorov, A., Schmidt, P., (2016). Endogeneity in stochastic frontier models. Journal of Econometrics 190 (2), 280–288.

[2] Amsler, C., Prokhorov, A., Schmidt, P., (2017). Endogenous environmental variables in stochastic frontier models. Journal of Econometrics 199 (1), 131–140.

[3] Badunenko, O., Henderson, D. J., Kumbhakar, S. C., (2012). When, where and how to perform efficiency analysis. Journal of the Royal Statistical Society Series A 175 (4), 863–892.

[4] DiCiccio, T. J., Kass, R. E., Raftery, A., Wasserman, L., (1997). Computing Bayes factors by combining simulation and asymptotic approximations. Journal of the American Statistical Association, 92, 903–15.

[5] Diewert, W. E., Wales, T. J., (1987). Flexible functional forms and global curvature conditions. Econometrica 55, 43–68.

[6] Fan, Y, Li, Q., Weersink, A., (1996). Semiparametric estimation of stochastic production frontier models. Journal of Business and Economic Statistics 14 (4), 460–468.

[7] Fong, E., Holmes, C. C., (2020). On the marginal likelihood and cross-validation. Biometrika 107 (2), 489–496.

[8] Gandhi, A., Navarro, S., Rivers, D. A. (2020). On the Identification of Gross Output Production Functions. Journal of Political Economy, 128 (8), 2973–3016.

[9] Geweke, J., Amisano, G., (2011a). Optimal prediction pools. Journal of Econometrics 164, 130–141.

[10] Geweke, J., Amisano, G., (2011b). Hierarchical Markov normal mixture models with applications to financial asset returns. Journal of Applied Econometrics 26, 1–29.

[11] Geweke, J., Keane, M., (1997). Mixture of Normals Probit Models. Federal Reserve Bank of Minneapolis, Research Department, Staff Report 237.

[12] Geweke, J., Keane, M. (2007). Smoothly mixing regressions. Journal of Econometrics 138, 252–290.

[13] Giacomini, R., Kitagawa, T. (2021). Robust Bayesian inference for set identified models. Econometrica 89 (4) 1519–1556.

[14] Griffiths, W. E., Hajargasht, G., (2016). Some models for stochastic frontiers with endogeneity. Journal of Econometrics, 190, 341–348.

[15] Hornik, K., (1993). Some new results on neural network approximation. Neural Networks 6 (8), 1069–1072.

[16] Hornik, K., Stinchcombe, M., White, H., Auer, P., (1994). Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Neural Computation 6 (6), 1262–1275.

[17] Kass, R. E., Raftery, A. E., (1995). Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.

[18] Kline, B., Tamer, E. (2016). Bayesian inference in a class of partially identified models. Quantitative Economics 7 (2), 329–366.

[19] Kumbhakar, S. C., Schmidt, P. (2016). Endogeneity Problems in Econometrics. Journal of Econometrics 190 (2), 209–374.

[20] Kumbhakar, S. C., Park, B. U., Simar, L., Tsionas, M. G. (2007). Nonparametric stochastic frontiers: A local maximum likelihood approach. Journal of Econometrics 137 (1), 1–27.

[21] Kumbhakar, S. C., Parmeter, C.F., Zelenyuk, Z., (2021a). Stochastic Frontier Analysis: Foundations and Advances I. In: S.C. Ray, R. Chambers and S. C. Kumbhakar (eds.) *Handbook of Production Economics*. Springer, Singapore.

[22] Kumbhakar, S. C., Parmeter, C.F., Zelenyuk, Z., (2021b). Stochastic Frontier Analysis: Foundations and Advances II. In: S.C. Ray, R. Chambers and S. C. Kumbhakar (eds.) *Handbook of Production Economics*. Springer, Singapore.

[23] Kleibergen, F., van Dijk, H. K., (1993). On the shape of the likelihood/posterior in cointegration models. Econometric Theory 10 (3–4), 514–551.

[24] Lewbel, A. (1996). Constructing Instruments for Regressions With Measurement Error When no Additional Data are Available, with An Application to Patents and R&D. Econometrica 65 (5), 1201–1213.

[25] Malikov, E., Kumbhakar, S. C., Tsionas, M. G. (2015). A Cost System Approach to the Stochastic Directional Technology Distance Function with Undesirable Outputs: The Case of US Banks in 2001-2010. Journal of Applied Econometrics 31, 1407–1429.

[26] Martins-Filho, C. B., Yao, F., (2015). Semiparametric stochastic frontier estimation via profile likelihood. Econometric Reviews 34 (4), 413–451.

[27] Norets, A., (2010). Approximation of conditional densities by smooth mixtures of regressions. The Annals of Statistics 38 (3), 1733–1766.

[28] Olley, S., Pakes, A., (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. Econometrica, 64 (6), 1263–1298.

[29] Park, B. U., Simar, L., Zelenyuk, V., (2015). Categorical data in local maximum likelihood: Theory and applications to productivity analysis. Journal of Productivity Analysis 43 (1), 199–214.

[30] Parmeter, C. F., Kumbhakar, S. C., (2014). Efficiency analysis: A primer on recent advances. Foundations and Trends in Econometrics 7 (3–4), 191–385.

[31] Parmeter, C. F., Simar, L., Van Keilegom, I., Zelenyuk, V., (2021). Inference for nonparametric stochastic frontier models. Manuscript.

[32] Parmeter, C. F., Zelenyuk, V., (2019). Combining the Virtues of Stochastic Frontier and Data Envelopment Analysis. Operations Research, 67 (6), 1628–1658.

[33] Paul, S., Shankar, S., (2018). On estimating efficiency effects in a stochastic frontier model. European Journal of Operational Research 271 (2), 769–774.

[34] Rubin, D. B., (1987). Comment on "The calculation of posterior distributions by data augmentation", by M. A. Tanner and W. H. Wong". Journal of the American Statistical Association 82, 543–546.

[35] Rubin, D. B., (1988). Using the SIR Algorithm to Simulate Posterior Distributions, in Bayesian Statistics 3, ed. J. M. Bernardo, M. H. DeGroot , D. V. Lindley, and A. F. M. Smith, 395–402. Oxford: Oxford University Press.

[36] Sickles, R., Zelenyuk, V., (2019). *Measurement of Productivity and Efficiency: Theory and Practice*. New York, NY: Cambridge University Press.

[37] Simar, L., Zelenyuk, V., (2011). Stochastic FDH/DEA estimators for frontier analysis. Journal of Productivity Analysis 36 (1), 1–20.

[38] Simar, L., Van Keilegom, I., Zelenyuk, V., (2017). Nonparametric least squares methods for stochastic frontier models. Journal of Productivity Analysis 47(3), 189–204.

[39] Smith, A. F. M., Gelfand, A. E., (1992). Bayesian statistics without tears: A sampling–resampling perspective. Journal of the American Statistical Association 46 (2), 84–88.

[40] Staiger, D., Stock, J. B., (1997). Instrumental variables regressions with weak instruments. Econometrica, 65, 557–586.

[41] Stock, J. H., Wright, J. H., Yogo, M., (2002). A survey of weak instruments and weak identification in generalized method of moments. Journal of Business and Economic Statistics, 20, 518 – 29.

[42] Tsionas, M. G., (2021). Estimating monotone concave stochastic production frontiers, Journal of Business & Economic Statistics, forthcoming.

[43] Tsionas, M. G., Mamatzakis, E., (2019). Further results on estimating inefficiency effects in stochastic frontier models. European Journal of Operational Research 275 (3), 1157–1164.

[44] van den Broeck, J., Koop, G., Osiewalski, J., Steel, M. F. J., (1994). Stochastic frontier models: A Bayesian perspective. Journal of Econometrics, 61, 273–303.

[45] Villani, M., Kohn, R., Giordani, P., (2009). Regression density estimation using smooth adaptive Gaussian mixtures. Journal of Econometrics 153, 155–173.

[46] Voglis, C. E., Hadjidoukas, P. E. , Lagaris, I. E., Papageorgiou, D. G., (2009). A numerical differentiation library exploiting parallel architectures. Computer Physics Communications 180 (8), 1404–1415.

[47] Zellner, A., (1971). *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.

## Technical Appendices – Not Intended for Publication

### A.1. Additional Monte Carlo Results

We use the same simulation designs as in Parmeter and Zelenyuk (2019) and we compare with the best results in each of the cases reported in their Tables 1-5. We consider five different data generating processes (DGP). DGP I is a linear frontier with Normal-Half Normal (NHN) errors generated as $y_i = x_i^{1/2} + v_i - u_i$, $v_i \sim \mathcal{N}(0, \sigma_v^2)$, $u_i \sim \mathcal{N}_+(0, \sigma_u^2)$ with $\sigma_v = 0.2$ and $\lambda = \frac{\sigma_u}{\sigma_v} \in \{0.5, 1, 2\}$, and $i = 1, \ldots, n$, where the sample size ($n$) takes the values 50, 100, 200, and 400. The covariate $x_i$ is generated uniformly in the interval (0.5, 10). In DGP II (to assess the impact of distribution misspecification), the inefficiency distribution is specified as Gamma with parameters (1, 5), (9, 5), and (9, 34). In DGP III, inefficiency has a skedastic function $\sigma_u(x) = 0.95e^{-0.5x}$. DGP IV is a nonlinear frontier with

$$y_i = a_1 x_i + a_2 \sin x_i + v_i - u_i, \quad i = 1, \ldots, n, \tag{A.1}$$

with $a_1 = a_2$ to ensure monotonicity. In DGP V, Parmeter and Zelenyuk (2019) have a nonlinear frontier (adopted from Martins-Filho and Yao (2015)):

$$y_i = b_1 + b_2 \arctan[b_3(x_i - b_4)] + v_i - u_i, \quad i = 1, \ldots, n, \tag{A.2}$$

where $b_1 = 1$, $b_2 = 0.5$, $b_3 = 20$, and $b_4 = 0.5$. The two-sided and one-sided error components are specified as before. We report results for the benchmark $FLW_{DD}$ estimator ("DD" stands for "data driven") which relies on plug-in likelihood estimation with least-squares cross-validation and performs well in the simulations reported in Tables 1-5 of Parmeter and Zelenyuk (2019, pp. 1649–1652). In all of our Monte Carlo experiments we use $\bar{I} = 10$.

We also use 200 Monte replications, as in their study, and we focus on the more difficult cases, viz. $\lambda = 0.5$, $Gamma(9, 5)$ to save space, although results for other values of $\lambda$ and the parameters of the Gamma distribution gave similar conclusions. For all the DGPs, we estimate the frontier at each data point in each replication and then compute the square root of the mean squared error (RMSE) over all the data points and take the median over all the replications of Monte Carlo to be consistent with Parmeter and Zelenyuk (2019). We do not correct for endogeneity using (9) or (10) as there is none in these DGPs.

TABLE 3. RMSE results, Comparison with Tables 1 - 5 of Parmeter and Zelenyuk (2019), $FLW_{DD}$ method.

| | I. Linear NHN frontier | II. Linear frontier, *gamma* distribution of inefficiency | III. Linear frontier, heteroskedastic inefficiency | IV. Nonlinear frontier, NHN distribution$^{(a)}$ | V. Nonlinear frontier, NHN distribution$^{(b)}$ |
|---|---|---|---|---|---|
| $n = 50$ | 0.113 | 0.417 | 0.180 | 0.138 | 0.120 |
| | *0.062* | *0.070* | *0.035* | *0.049* | *0.032* |
| | 0.113 | 0.129 | 0.135 | 0.135 | 0.120 |
| | **0.058** | **0.065** | **0.030** | **0.048** | **0.030** |
| $n = 100$ | 0.096 | 0.377 | 0.182 | 0.114 | 0.095 |
| | *0.035* | *0.055* | *0.022* | *0.032* | *0.011* |
| | 0.095 | 0.111 | 0.132 | 0.110 | 0.097 |
| | **0.039** | **0.051** | **0.020** | **0.030** | **0.010** |
| $n = 200$ | 0.070 | 0.382 | 0.171 | 0.092 | 0.073 |
| | *0.025* | *0.032* | *0.017* | *0.025* | *0.005* |
| | 0.070 | 0.097 | 0.120 | 0.087 | 0.072 |
| | **0.020** | **0.030** | **0.015** | **0.022** | **0.005** |
| $n = 400$ | 0.047 | 0.376 | 0.164 | 0.059 | 0.059 |
| | *0.015* | *0.021* | *0.011* | *0.011* | *0.001* |
| | 0.047 | 0.096 | 0.097 | 0.049 | 0.058 |
| | **0.014** | **0.020** | **0.010** | **0.010** | **0.001** |

Notes: (a) This refers to (A.1). (b) This refers to (A.2). Numbers in regular font correspond to root mean squared errors (RMSE) from Tables 1–5 in Parmeter and Zelenyuk (2019, pp. 1649–1652) and they refer to $FLW_{DD}$ (the Fan et al., 1997 estimator) which relies on plug-in likelihood estimation with least-squares cross validation. KPST estimates of inefficiency to calibrate the prior in (14). Numbers in italics refer to results of this study based on using DEA estimates of inefficiency to calibrate the prior in (14). Underlined figures do not use information from DEA or KPST and are based on our benchmark prior (13). As FDH priors did not yield better results compared to DEA or KPST they are omitted. Numbers in bold represent results from a prior anchored on KPST.

The $FLW_{DD}$ estimator performs well for DGPs I, IV and V (as expected) of Parmeter and Zelenyuk (2019) although the parametric estimator performs slightly better as expected in DGP I. The techniques proposed here perform slightly better than FLW in scenarios I, II, and III but RMSEs are considerably lower for DGPs III and IV. Although DGP V has a nonlinear frontier, it

can, nevertheless, be approximated globally by a linear function so the parametric SFA estimator (Parmeter and Zelenyuk, 2020, p. 1652, Table 1) performs well. In this instance, our results are also the same or slightly better compared to FLW. Our results are considerably better than FLW in the relatively "difficult" DGP IV. Under misspecification of the inefficiency distribution (DGP II) our results are, again, slightly better than FLW. The methods proposed here perform rather well under scenario DGP where inefficiency is heteroskedastic albeit not in the same way that we model heteroskedasticity. This shows that approximations like (6) are likely to perform very well in practice. Our method also performs very well under a misspecified $Gamma(9,\, 5)$ distribution for inefficiency where the parametric methods perform better than FLW but still they have (like FLW) large RMSEs.

When we do not use DEA-based information (see underlined figures in Table 1) but instead we use the prior in (13), RMSEs are larger than RMSEs delivered by the DEA prior, mostly in scenarios II and III. In III, as we have heteroskedasticity, the RMSEs corresponding to the DEA prior are significantly lower so, in practice, using DEA to benchmark a prior like (14) may prove to be beneficial.

Our implementation in scenario III assumes that $x_i$ is in the skedastic function (6). We have experimented with a version where we try $z_i$ in the skedastic function, where $z_i = x_i + \sigma_z \zeta_i$, $\zeta_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ $(i = 1, \ldots, n)$. For moderate values of $\sigma_z$ we find[26] that the DEA prior loses its advantage over (13). Specifically, this happens for values of $\sigma_z$ greater than about 0.5 which is, approximately, 10% of the average value of $x_i$ (which is 5.25 in this instance). The RMSEs are still smaller compared to FLW even when $\sigma_z$ exceeds 1 or 2, suggesting that using DEA to benchmark such prior may prove to be beneficial.

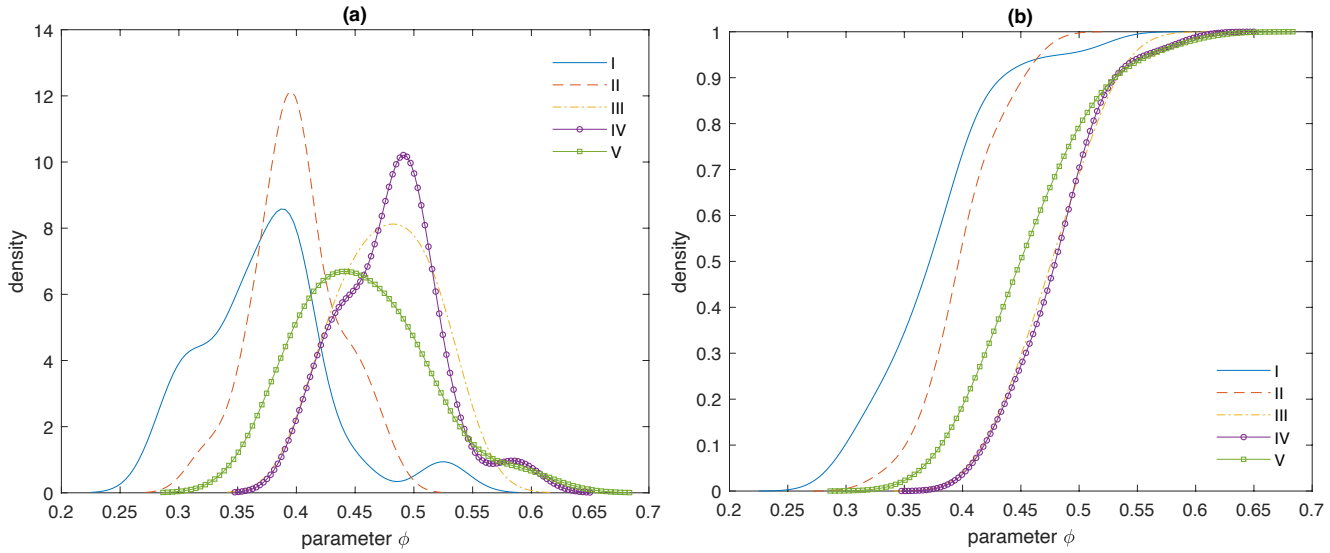
Finally, we also find that in the case of small samples ($n$ is 50 or 100) that the DEA prior is important in reducing RMSEs. For larger samples, despite the fact that the prior comes from the data, it seems to be "dominated" by the likelihood.[27] Moreover, results for FDH are not reported as they were always inferior relative to DEA.

The estimated densities (across simulations) of $\phi$ are reported in panel (a) of Figure A.1. In panel (b) of Figure A.1 we report the corresponding estimated distribution functions. From

[26]These results are not reported here but are available on request.

[27]Clearly, this "domination" is due to the fact that we allow a very large variance in (15), as $h_* = 10^4$.

FIGURE A.1. Distributions of $\phi$ for the five models.

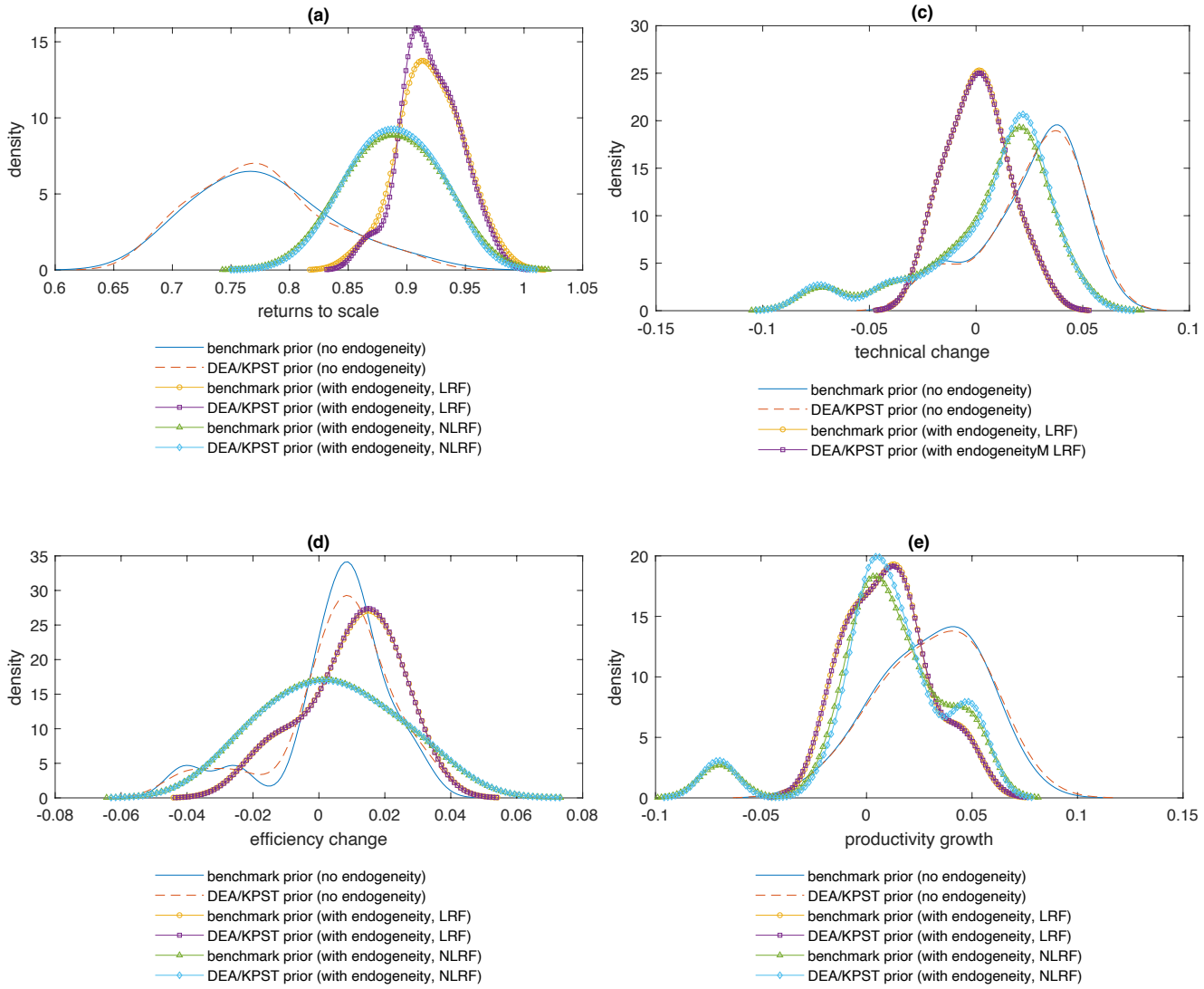



this evidence it turns out that DEA and KPST receive approximately equal weight (as we do not have first-order stochastic dominance in models I and II). In Models III, IV, and V, again, we do not have first-order stochastic dominance, although compared to models I and II, $\phi$ is likely to be more concentrated around 0.5. This type of exercise is important as, rarely if ever, do we have the luxury of dividing the data, using one part to craft the prior and the second to perform posterior inference.

To conclude this discussion, we note that while the KPST prior is slightly better than the DEA prior (and both give much better results than the others), obtaining the DEA prior is much easier/simpler as it does not involve setting a bandwidth, which may be fraught with difficulties, especially in high dimensions, and for which the results might be sensitive to a particular choice. On the other hand, DEA might be sensitive if too much noise or outliers are present in the data (as KPST might be too). Hence, it might be wise to use both and potentially with other priors (or mix them as described above), to confirm the robustness of results or to identify potential issues to explore in more details.

For our proposed estimator we see vast improvements relative to FLW-type estimators for all sample sizes when we anchor to the KPST results (rather than to DEA to calibrate (14) or (15) without any information from either DEA and KPST). The reason can be easily attributed to the fact that the estimation of inefficiency in finite samples is a difficult exercise on its own, and informative priors, that is priors that carry useful information about inefficiency, can be of great help in reducing root mean squared errors. A notable feature of such priors is that their effect does not converge to zero as the sample size becomes arbitrarily large. For example, DEA- and KPST-based priors remain informative even in large samples, that is DEA and KPST can be usefully employed to learn about efficiency even in large samples. In this sense, and since the number of parameters in the inefficiency distribution increases with the sample size, anchoring priors remain informative.

FIGURE A.2. Sample distributions of posterior mean estimates. DEA and KPST stand for Data Envelopment Analysis, and Kumbhakar, Park, Simar and Tsionas (2007, KPST), respectively. LRF is the linear reduced form in (9) and NLRF corresponds to the flexible nonlinear specification of the reduced form in (10).




## A.2. Posterior Means and Standard Deviations For Application

Figure A.2 presents the remainder of the estimated posterior means and standard deviations for various features of the banking technology. Table 4 provides the specific numeric values for all of these technological features for the both the posterior mean and the posterior standard deviation.

TABLE 4. Posterior means and standard deviations of functions of interest.

| | posterior mean | posterior s.d. |
|---|---|---|
| **Returns to Scale** | | |
| Benchmark Prior (No Endogeneity) | 0.777 | 0.056 |
| DEA/KPST Prior (No Endogeneity) | 0.774 | 0.055 |
| Benchmark Prior (With Endogeneity, LRF) | 0.920 | 0.025 |
| DEA/KPST Prior (With Endogeneity, LRF) | 0.919 | 0.024 |
| Benchmark Prior (With Endogeneity, NLRF) | 0.889 | 0.032 |
| DEA/KPST Prior (With Endogeneity, NLRF) | 0.890 | 0.030 |
| **Inefficiency** | | |
| Benchmark Prior (No Endogeneity) | 0.262 | 0.055 |
| DEA/KPST Prior (No Endogeneity) | 0.260 | 0.050 |
| Benchmark Prior (With Endogeneity, LRF) | 0.144 | 0.028 |
| DEA/KPST Prior (With Endogeneity, LRF) | 0.145 | 0.030 |
| Benchmark Prior (With Endogeneity, NLRF) | 0.170 | 0.028 |
| DEA/KPST Prior (With Endogeneity, NLRF) | 0.167 | 0.029 |
| **Technical Change** | | |
| Benchmark Prior (No Endogeneity) | 0.026 | 0.023 |
| DEA/KPST Prior (No Endogeneity) | 0.025 | 0.025 |
| Benchmark Prior (With Endogeneity, LRF) | 0.0005 | 0.014 |
| DEA/KPST Prior (With Endogeneity, LRF) | 0.0006 | 0.015 |
| Benchmark Prior (With Endogeneity, NLRF) | 0.005 | 0.030 |
| DEA/KPST Prior (With Endogeneity, NLRF) | 0.006 | 0.031 |
| **Efficiency Change** | | |
| Benchmark Prior (No Endogeneity) | 0.0046 | 0.018 |
| DEA/KPST Prior (No Endogeneity) | 0.0051 | 0.017 |
| Benchmark Prior (With Endogeneity, LRF) | 0.0097 | 0.013 |
| DEA/KPST Prior (With Endogeneity, LRF) | 0.0096 | 0.014 |
| Benchmark Prior (With Endogeneity, NLRF) | 0.033 | 0.017 |
| DEA/KPST Prior (With Endogeneity, NLRF) | 0.034 | 0.019 |
| **Productivity Growth** | | |
| Benchmark Prior (No Endogeneity) | 0.0046 | 0.018 |
| DEA/KPST Prior (No Endogeneity) | 0.0051 | 0.017 |
| Benchmark Prior (With Endogeneity, LRF) | 0.0097 | 0.013 |
| DEA/KPST Prior (With Endogeneity, LRF) | 0.0097 | 0.014 |
| Benchmark Prior (With Endogeneity, NLRF) | 0.0033 | 0.018 |
| DEA/KPST Prior (With Endogeneity, NLRF) | 0.0044 | 0.019 |
| $\phi$ | 0.112 | 0.015 |

## A.3. MCMC methods

We use a recent advance on the Metropolis Adjusted Langevin Algorithm (MALA) called fast MALA (fMALA), see Durmus et al. (2017). Suppose we have a parameter vector $\boldsymbol{\theta} \in \Re^d$, and we target $\pi(\boldsymbol{\theta})$ which represents the posterior, omitting the dependence on data to ease notation. We consider a Langevin diffusion defined by:

$$\mathrm{d}\boldsymbol{\theta}_t = \tfrac{1}{2}\boldsymbol{\Sigma} \cdot \nabla \ln \pi(\boldsymbol{\theta}_t) + \boldsymbol{\Sigma}^{1/2}\mathrm{d}\boldsymbol{W}_t, \tag{A.3}$$

where $\{\boldsymbol{W}_t, t \geq 0\}$ is a standard $d$-dimensional Brownian motion, and $\boldsymbol{\Sigma}$ is a given positive definite self-adjoint matrix. Under appropriate assumptions on $\pi$ one can show that the dynamics generated by (A.3) are ergodic and result in $\pi(\boldsymbol{\theta})$ as a unique invariant distribution. A standard approach is to discretize (A.3) using a one step integrator, and sample using the averages over the numerical trajectories. This approach introduces a bias because the posterior does not coincide in general with the exact $\pi$.

An alternative way of sampling from $\pi$ exactly, i.e. that is not biased by discretizing (A.3), is by using the Metropolis-Hastings algorithm (Hastings, 1970). The idea is to construct a Markov chain $\{\boldsymbol{\theta}_j\}$, where at each step $j$, given $\boldsymbol{\theta}_j$, a new sample proposal $\boldsymbol{\theta}^c$ is generated from the Markov chain with transition kernel $q(\boldsymbol{\theta}, \cdot)$. This proposal is then accepted ($\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}^c$) with probability $\alpha(\boldsymbol{\theta}_j, \boldsymbol{\theta}^c)$ and rejected ($\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j$) otherwise. If we have

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^c) = \min\left\{1,\ \frac{\pi(\boldsymbol{\theta}^c)q(\boldsymbol{\theta}^c, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\theta}^c)}\right\}, \tag{A.4}$$

then the resulting Markov chain $\{\theta_j\}$ is $\pi$-invariant and will, for large $j$ generate samples from $\pi$ under mild ergodicity assumptions. In general, a candidate is generated as:

$$\boldsymbol{\theta}^c = \boldsymbol{\mu}(\boldsymbol{\theta}, h) + \boldsymbol{S}(\boldsymbol{\theta}, h)\boldsymbol{\zeta}, \tag{A.5}$$

where $\boldsymbol{\zeta} \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{I}_d)$. The specific fMALA proposal has

$$\boldsymbol{\mu}(\boldsymbol{\theta}, h) = x + \tfrac{h}{2}f(\boldsymbol{\theta}) - \tfrac{h^2}{24}\nabla f(\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}) + \{\boldsymbol{\Sigma} : \nabla^2 f(\boldsymbol{\theta})\}, \tag{A.6}$$

$$\boldsymbol{S}(\boldsymbol{\theta}, h) = \left(h^{1/2}\boldsymbol{I}_d + \tfrac{h^{3/2}}{12}Df(\boldsymbol{\theta})\right)\boldsymbol{\Sigma}^{1/2}, \tag{A.7}$$

where $f(\boldsymbol{\theta}) \triangleq \boldsymbol{\Sigma} \cdot \nabla \ln \pi(\boldsymbol{\theta})$, $\nabla f(\boldsymbol{\theta})$ and $\nabla^2 f(\boldsymbol{\theta})$ are the $d \times d$ Jacobian and $d \times d^2$ Hessian of $f(\boldsymbol{\theta})$, respectively, and $\boldsymbol{\Sigma} = \boldsymbol{S}(\boldsymbol{\theta}, h)$. Let $\nabla^2 f(\boldsymbol{\theta}) = [\boldsymbol{H}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{H}_d(\boldsymbol{\theta})]$ where $[\boldsymbol{H}_i(\boldsymbol{\theta})]_{jk} = \frac{\partial f_i(\theta)}{\partial \theta_k \partial \theta_j}$. Then, $\{\boldsymbol{\Sigma} : \nabla^2 f(\boldsymbol{\theta})\}_i \triangleq tr\left[\boldsymbol{\Sigma}' \boldsymbol{H}_i(\boldsymbol{\theta})\right]$. The scaling constant has been investigated in Durmus et al. (2017) and it is related directly to the discretization of (A.3). Durmus et al. (2017) recommend $h = \varepsilon d^{-1/5}$ for some positive constant, $\varepsilon$. The optimal acceptance rate maximizing the first-order efficiency is very close to the 0.704 predicted in Theorem 3.2 of Durmus et al. (2017). Therefore, one can calibrate the constant $\varepsilon$ (during the burn-in phase) so that the acceptance rate is close to 0.70.

This approach has been found to perform excellently once $\varepsilon$ and $h$ are calibrated correctly during the burn-in phase. All derivatives are computed numerically[28] during the burn-in phase, and they are interpolated[29] in the main phase of the MCMC algorithm. This results in dramatic computational savings and, as a matter of fact, different chains can be run in parallel in computers with multiple nodes. We run ten different chains starting from randomly selected initial conditions and we compare the chains after 150,000 iterations with a burn-in phase consisting of 50,000 iterations. Our transition density $q(\boldsymbol{\theta}, \boldsymbol{\theta}^c)$ is a $d$-dimensional Student-$t$ distribution with five degrees of freedom. We monitor convergence using the standard diagnostics of Geweke (1992).

[28]We use the `Fortran77` subroutines in package `NDL` of Voglis et al. (2009). Specifically we use version 2.0 of Hadjidoukas et al. (2014), `https://data.mendeley.com/datasets/j2fhmszg85/1`, see also `http://cpc.cs.qub.ac.uk/summaries/AEDG_v1_0.html`

[29]We use the `Fortran` subroutines in `finterp` by Jacob Williams in `https://github.com/jacobwilliams/finterp/blob/master/README.md` Alternatively, we use for comparison `RBF_INTERP_ND` in `https://people.sc.fsu.edu/~jburkardt/f_src/rbf_interp_nd/rbf_interp_nd.html`. `RBF_INTERP_ND` is a `Fortan90` library by John Burkardt which defines and evaluates radial basis function (RBF) interpolants to multidimensional data.

## A.4. Posterior sensitivity analysis

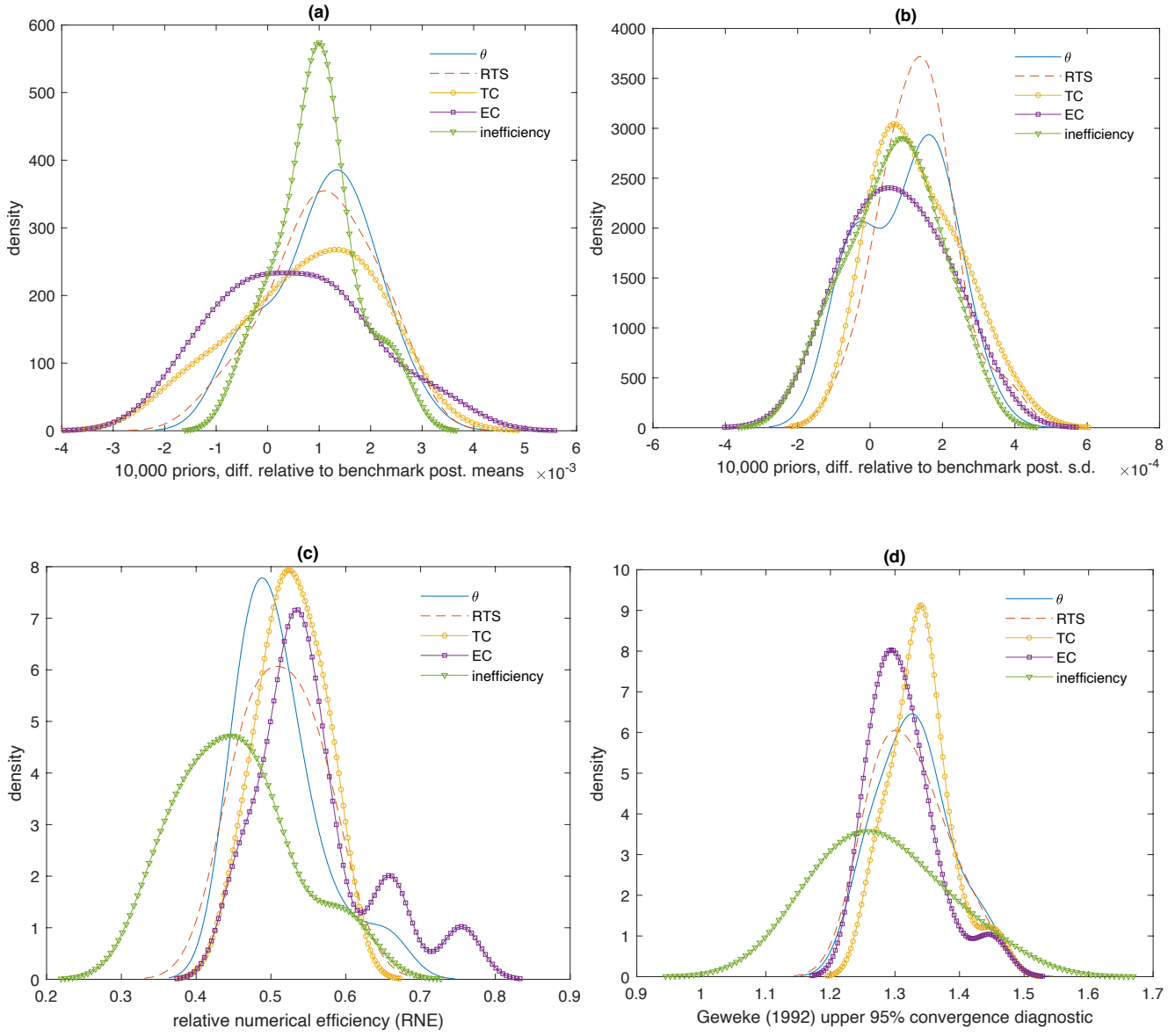
To perform posterior sensitivity analysis, we replace (13) and (14) with the following. Suppose $\vartheta \in \mathbb{R}^{d_\vartheta}$ denotes all elements of $\boldsymbol{\theta}$ with the exception of scale parameters. Then we assume

$$\vartheta \sim \mathcal{N}_{d_\vartheta}\left(\boldsymbol{a},\ \boldsymbol{V}\right). \tag{A.8}$$

We draw the elements of the mean vector ($\boldsymbol{a}$) from Normal distributions with zero mean and standard deviation 100. For $\boldsymbol{V}$ we assume it is a diagonal matrix whose elements are draws as $\log V_{ii} \sim \mathcal{N}(0, 100)$. We consider 10,000 prior specifications, viz. 10,000 different draws for $\boldsymbol{a}$ and $\boldsymbol{V}$, and we use the Sampling-Importance-Resampling (SIR) approach (Rubin, 1987, 1988; Smith and Gelfand, 1992) with 20% sub-samples of the original MCMC sample for re-weighting and we compute posterior moments corresponding to the new priors.

The differences relative to benchmarks are reported in panels (a) and (b) of Figure A.3, respectively. We report results for parameters ($\theta$), returns to scale (RTS), technical change (TC), efficiency change (EC) and inefficiency itself. In panel (c) we report Geweke's (1992) relative numerical efficiency (RNE) which should be equal to one if one could draw IID samples from the posterior. In panel (d), we report upper 95% values of Geweke's (1992) convergence diagnostic which is asymptotically Normal in the number of draws. As the upper 95% values of this $z$-test are less than 1.96 we can be relatively confident that the MCMC chains have converged and, despite autocorrelation, the RNEs are not extremely low to prevent us from a thorough exploration of the posterior.

FIGURE A.3. Posterior sensitivity relative to benchmark prior



## A.5. Validity and weakness of instruments

We focus on Equation (10) which we repeat here in the interest of clarity:

$$x_{ki} = \boldsymbol{q}_i'\varpi_{k0} + \sum_{g=1}^{G} \psi(\boldsymbol{q}_i'\varpi_{kg1})\varpi_{kg2} + V_{ik},\ k = 1, \ldots, K;\ i = 1, \ldots, n. \tag{A.9}$$

We can obtain the linear reduced form (LRF) in (9) using appropriate restrictions. To make sure (to the extent feasible) we rely on squared correlation coefficients between actual and fitted values from (A.9). As a matter of fact we can use a single measure, the generalized R-squared,

$$R_*^2 = 1 - \tfrac{|\boldsymbol{\Sigma_{VV}}|}{|\boldsymbol{S}|}, \tag{A.10}$$

where $\boldsymbol{\Sigma_{VV}}$ is the covariance matrix of $V_{ik}$s and $\boldsymbol{S}$ is the empirical covariance matrix of $x_{ki}$s. For each MCMC draw, we save the generalized R-squared, say $R^2_{*(s)}$, and we present its posterior distribution based on these draws.

The issue of valid instruments is based on ideas from Generalized Method of Moments (GMM) estimation. Our (A.9) requires the moment conditions

$$\tfrac{1}{n}\sum_{i=1}^{n}\left(x_{ki} - \boldsymbol{q}_i'\varpi_{k0} - \sum_{g=1}^{G}\psi(\boldsymbol{q}_i'\varpi_{kg1})\varpi_{kg2}\right)\boldsymbol{q}_i = 0,\ k = 1, \ldots, K. \tag{A.11}$$

We can write these moment conditions compactly as follows:

$$\tfrac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{S}_i; \varpi_{1,0}, \hat{\varpi}_*) = \boldsymbol{0}, \tag{A.12}$$

where the notation is intended to make clear that we fix part of the parameter vector.

We have $Kd_z$ moment conditions but the number of parameters is $KG[(1 + d_z) + 1]$ so we need to expand the number of moment conditions or reduce the number of parameters. We decide to set $\varpi_{kg1}$ and $\varpi_{kg2}$ to their posterior means, set $\varpi_{k0} = \varpi_{1,0}$ $(k = 2, \ldots, K)$ and redo the MCMC using only a common $\varpi_{1,0}$ in the moment conditions.[30] Then we have only $d_z$ parameters.

The test statistic that we use is the Sargan-Hansen test statistic given by

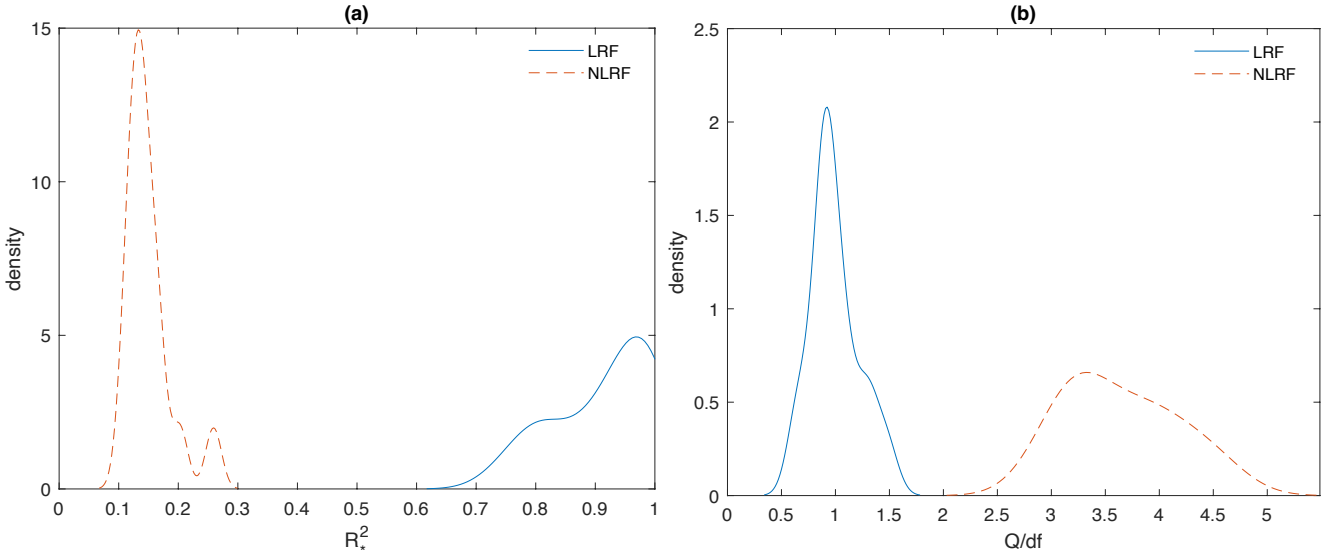
$$Q = \left[\tfrac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{S}_i; \varpi_{1,0}, \hat{\varpi}_*)\right]\boldsymbol{W}\left[\tfrac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{S}_i; \varpi, \hat{\varpi}_*),\right] \xrightarrow{d} \chi^2_{(K-1)d_z}, \tag{A.13}$$

[30] $\varpi_{1,0}$ is estimated with least-squares for each MCMC draw.

where $\boldsymbol{W} = \left(\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{S}_i; \varpi, \hat{\varpi}_*) g(\boldsymbol{S}_i; \varpi, \hat{\varpi}_*)'\right)^{-1}$ with degrees of freedom determined by $\text{df} = (K-1)d_z$.

The test statistic is distributed as a $\chi^2_{df}$ with $df$ degrees of freedom and it, in fact, indexed $Q_{(s)}$ for each MCMC draw $s$ $(s = 1, \ldots, S)$. The test is quite stringent in our case as a large number of parameters is fixed so, it is not expected a priori that the test will indicate that the moment conditions are satisfied under these restrictions. To simplify the presentation, as $Q \xrightarrow{d} \chi^2_{(K-1)d_z}$ it follows that $\frac{Q}{\text{df}} \xrightarrow{d} \mathcal{G}(\text{df}, \text{df})$, a gamma distribution,[31] whose 95% (90%) critical value is approximately 1.17 (1.13).

FIGURE A.4. Diagnostic results via $R^2_*$ and $Q/\text{df}$.




The generalized R-squared ($R^2_*$) whose marginal posterior is reported in panel (a) of Figure A.4, is substantial for the nonlinear reduced form (NLRF) in (10) but, clearly, quite low for the linear reduced form (LRF) in (9). In terms of the Sargan-Hansen statistic ($Q/\text{df}$) whose marginal posteriors for LRF and NLRF are reported in panel (b), the LRF clearly fails the test while the NLRF clearly passes the test. This is surprising, as we have fixed a large number of parameters in

[31] This can be shown using a simple change of variables technique.

(10) or (A.9) to implement this test, which testifies to the fact that such semiparametric functional forms perform well.

## A.6. Aspects of the Least Favorable Priors

As discussed in the methodology section, a range of priors could be used to the inform the ANN that is deployed and one means to adjudged various priors is through implementation of a LFP. Here we do this by looking at the maximization of $\rho(\delta)$ over the $(J-1)$-dimensional simplex is performed using a Nelder-Mead algorithm. We use 10,000 draws for the Monte Carlo part. We set up a grid of 20 values for $a$ and $b$ in the interval $[-2, 2]$ and another 20 values for $h$ in the interval $(0, 5]$. After locating approximate minimax values we use a grid of 50 points for $a$, $b$, and $h$ in the neighborhood of the previously located minimax values. We refer to this as **LFP-I**.

Our prior for the scale parameter is $\frac{0.1}{\sigma_u^2} \sim \chi_1^2$. Our prior for $\boldsymbol{\gamma}$ is

$$\boldsymbol{\gamma}|a, b, h, \{\hat{u}_i\} \sim \mathcal{N}(a + b\hat{\boldsymbol{\gamma}},\ h^2\mathbf{S}), \tag{A.14}$$

where $\hat{\boldsymbol{\gamma}}$ is the least squares estimator of $\boldsymbol{\gamma}$ from a regression of $\ln \frac{1-e^{-\hat{u}_i}}{e^{-\hat{u}_i}}$ on $\boldsymbol{w}_i$, and $\mathbf{S}$ is its covariance matrix. We will call this **LFP-II**. The role of $a$, $b$, and $h$ is practically the same as before with the difference that now we adjust the least squares estimates of the parameter vector $\boldsymbol{\gamma}$ from DEA.

Aspects of LFP-I are presented in Figure A.5 (using interpolation).

The marginal posterior densities of $a$, $b$, and $h$ are reported in panels (g), (h), and (i) of Figure A.5 while in the other panels of Figure A.5 we report aspects of their joint bivariate densities (using interpolation). With the exception of the pair ($b$, $h$) the bivariate densities are multimodal and asymmetric with a positive relation between ($a$, $b$) and ($a$, $h$) near the dominant models. The relationship between $b$ and $h$ in panel (f) is positive and nearly symmetric.

Using the same minimax procedure as before, aspects of LFP-II for $a$, $b$, and $h$ and the results are reported in Figure A.6.

In Figure A.6 we report the same aspects for the second version, LFP-II. Pairs ($a$, $b$) and ($a$, $h$) seem to be nearly independent whereas there is a strong nonlinear relationship between $b$ and $h$ (panel (f)). As a result, the marginal posterior of $h$ in panel (i) is asymmetric and bimodal while marginal posteriors of $a$ and $b$ in panels (g) and (h) appear nearly symmetric.

From Figures A.5 and A.6, the location parameter $a$ is negative (averaging, approximately, -0.38 and -0.75 in the two LFPs), parameter $b$ is close to 0.5 and 2 respectively, while the scale

FIGURE A.5. Aspects of the least-favorable prior, LFP-I

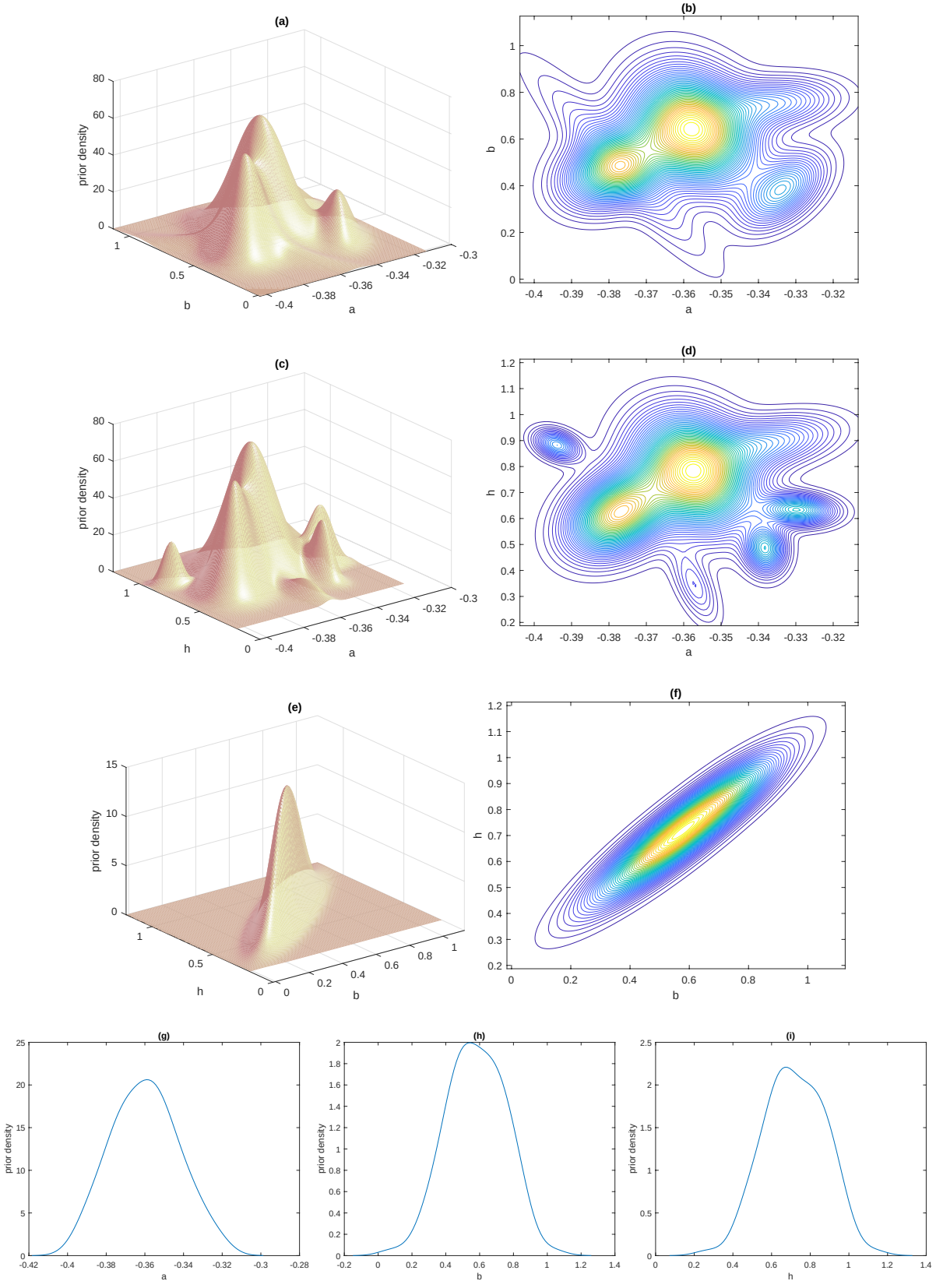
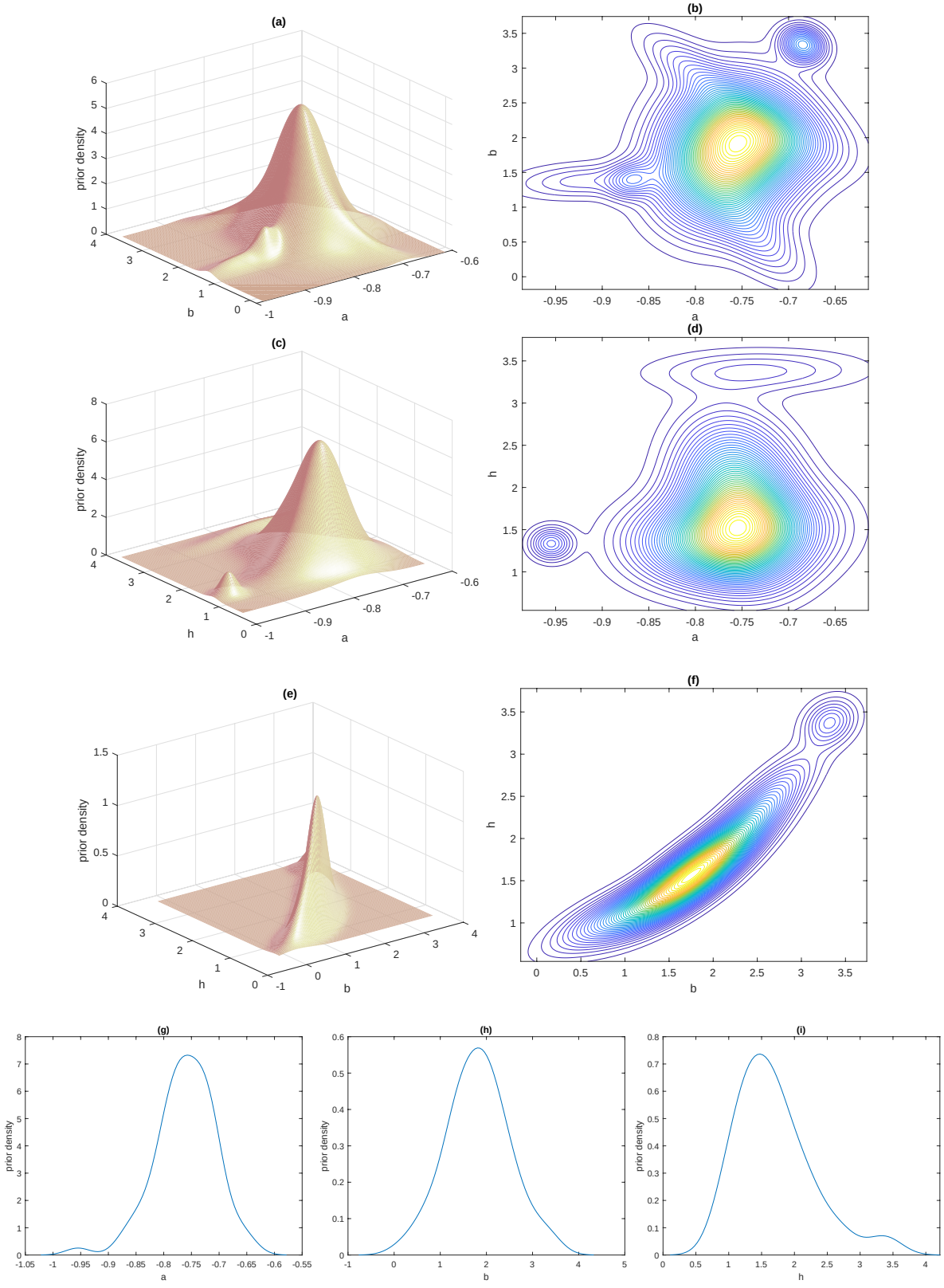


FIGURE A.6. Aspects of alternative least-favorable prior, LFP-II



parameter $h$ averages 0.8 and 2, respectively. In both cases, the rescaling of DEA scores is based on a scale parameter $h$ that exceeds one with significant probability. So, DEA scores are adjusted in a somewhat complicated manner which is not unexpected.