Joint production in Stochastic Non-Parametric Envelopment of Data with firm-specific directions

Mike G. Tsionas*

Abstract

We propose a likelihood-based approach to Stochastic Non-Parametric Envelopment of Data (StoNED) estimator using a directional distance function with firm-specific directional vectors. Additionally, we show how to estimate firm-specific inefficiency estimates instead of focusing on their average only. Moreover, we propose models that are robust to misspecification in general and the use of unit-information-priors in this class of models. These priors control the amount of information to be exactly equal to one observation. In this context, we propose the use of Bayesian Bootstrapping to further mitigate possible misspecification. We also propose empirical tests for identification of the model. Monte Carlo experiments show the good performance of the new techniques and an empirical application to the technology of large U.S. banks shows the feasibility of the new techniques.

Key Words: Productivity and Competitiveness; Stochastic Non-Parametric Envelopment of Data; Convex Non-parametric Least Squares; Firm-Specific Directions; Misspecification.

Abbreviations: StoNED (Stochastic Non-Parametric Envelopment of Data), UIP (Unit Information Prior), F-UIP (full UIP), PM-UIP (product of marginals UIP), DDF (Directional Distance Function), DGP (Data Generating Process), CNLS (Convex Non-Parametric Least Squares).

*Montpellier Business School & Lancaster University Management School, LA1 4YX, U.K., m.tsionas@montpellier-bs.com

1 Introduction

Directional distance functions (DDFs) are standard tools in the arsenal of operations research and production economics (Chambers, Chung, and Färe, 1996, 1998). They are particularly useful in estimating inefficiency as different directions can be assumed, and the results can be compared. A major obstacle is the absence of a formal data generating process (DGP) for directional distance functions in non-parametric envelopment of data (i.e., Data Envelopment Analysis, DEA). This problem has been solved by Kuosmanen and Johnson (2017) who proposed a suitable DGP and extended Stochastic Non-parametric Envelopment of Data (StoNED) or Convex Nonparametric Least Squares (CNLS) to the multiple-output case. The proposed estimator does not rely on parametric assumptions about the technology as it uses StoNED (Kuosmanen, 2006, 2008; Kuosmanen and Johnson, 2010; Kuosmanen and Kortelainen, 2015; Kuosmanen et al., 2007, 2015). Kuosmanen and Johnson (2017) made a number of simplifying assumptions since, as they correctly argued, "having a formal statistical model with clearly defined assumptions is far better than having no model at all." (op. cit., p. 795). Moreover, in Kuosmanen and Johnson (2017) the direction vector of the directional distance function are assumed to be known in advance which is quite useful for a regulator (as in their empirical application to Finnish electricity distribution firms) but otherwise it is somewhat restrictive. Moreover, they are able to estimate average inefficiency but not firm-specific inefficiency. Other approaches to determine the directional vector have been proposed by Daraio and Simar (2016), Färe, Grosskopf, and Whittaker (2013), Zofio, Pastor, and Aparicio (2013) and Atkinson and Tsionas (2016). Our contributions in this paper are as follows. We allow for misspecification and we use the so-called Gibbs posterior to correct for it (see Jiang and Tanner, 2008; Yao, Jiang, and Tanner, 2011; Ghosh and Basu, 2016; Basu et al., 1998; Bissiri et al., 2016 which is the most representative way of defining Gibbs posteriors). In terms of prior information we would like to be as uninformative as possible so, we use the so-called *Unit-Information-Prior* (UIP; Kass and Wasserman, 1995) which provides information that is worth exactly one observation. Posteriors are robust in the sense that we control for the amount of information in the prior. Third, we show how to estimate firm-specific directions as well as firm-specific inefficiency estimates that are as robust as possible in the presence of model misspecification. In addition, we examine identification empirically,a tool that could be used in other studies as well. To implement Bayesian inference we use Markov Chain Monte Carlo (MCMC) procedures (e.g., Geweke, 1999). Fourth, analysis by CNLS and firm-specific directions is quite involved as the constraints are no longer linear. In Bayesian analysis this is not an issue, in principle, as general optimization problems can be handled without much difficulty (an early reference is

Pincus, 1970). The problem, however, is to choose "non-informative" priors not all for the directional vector but for the parameters as a whole. As "non-informative" is an ambiguous term, we use the "unit information prior" (UIP) to impose information that is worth exactly one observation. This prior depends on the data (as we use the empirical version of the Hessian rather than the expected-value form. It also depends on the data because the prior of directions is the posterior of common (as opposed to firm-specific) directions in the CNLS formulation of Johnson and Kuosmanen (2017). For all practical purposes, this is an empirical Bayes procedure. Moreover, Bayesian procedures for multivariate semiparametric Bayesian concave regression have been developed by Arreola et al. (2020). Layer, Johnson, Sickles and Ferrier (LJSF, 2020) assume a unimodal distribution for the data generation process, and find that a direction that aims towards the "center" of the data and is perpendicular to the true function at that point tends to outperform other directions. These directions are easy to compute in advance (see their Section 6). Relative to this study, the present paper is based on a completely data-driven approach for selecting the directions using two versions of an objective (unit information) prior. Arreola at al. (2020) propose a single-stage multivariate semiparametric Bayesian concave regression method. Relative to this important study, the present paper differs in that we use observation-specific directions. The paper by LJSF (2020) proposes firm-specific directions that are orthogonal to the true function. Here, we follow a different approach by crafting an objective prior based on the unit information concept. Two such priors are proposed and compared. The directions in LJSF (2020) are computed easily in advance whereas our directions are estimated based on the versions of the objective prior. Moreover, convexity is enforced via the Afriat inequalities and appropriate latent variables (denoted U). Although orthogonal directions as in Layer et al. (2020) seem reasonable our approach is also reasonable as directional vectors are posterior estimates resulting from a vague, unit information prior.

The plan of the paper is as follows. In Section 2, we present the directional distance function along with the contribution of Kuosmanen and Johnson (2017). In Section 3 we present the techniques that will be used. In Section 3.1.1 we present the concept of Gibbs posteriors. In Section 3.1.2 we present the concept of robust posteriors and in Section 3.1.3 the so-called objective priors where the amount of information is constrained to be equal to one observation. In Section 3.2 we present the Bayesian approach to the StoNED estimator. The new techniques are examined in Section 4 where we present a Monte Carlo experiment. Our empirical application to large U.S. banks is presented in Section 5. We conclude with a summary of the paper and suggestions for future research.

2 Directional Distance Function model

The directional distance function (DDF; Chambers et al., 1996, 1998) is defined as

$$\overrightarrow{D}(x, y, g^x, g^y) = \sup\{\lambda | (x - \lambda g^x, y + \lambda g^y) \in \mathcal{T}\}, \tag{1}$$

where $(g^x, g^y) \in \mathbb{R}_+^{m+s}$ is a direction vector, and $\mathcal{T}$ is the production possibilities set defined as

$$\mathcal{T} = \left\{(x, y) \in \mathbb{R}_+^{m+s} : \overrightarrow{D}(x, y, g^x, g^y) \geq 0, \forall (g^x, g^y) \in \mathbb{R}_+^{m+s}\right\}. \tag{2}$$

The properties of the distance function are as follows.

P1. Translation: $\overrightarrow{D}(x - ag^x, y + ag^y, g^x, g^y) = \overrightarrow{D}(x, y, g^x, g^y) - a$.

P2. Homogeneity of degree –1 in the directional vector: $\overrightarrow{D}(x, y, ag^x, ag^y) = a^{-1}\overrightarrow{D}(x, y, g^x, g^y)$, $\forall a$.

P3. The directional distance function is decreasing in (g^x, g^y) for all $x, y \in \mathcal{T}$ but increasing for all $x, y \notin \mathcal{T}$ (Proposition 1 in Kuosmanen and Johnson, 2017).

Moreover, if the technology exhibits constant returns to scale (CRS), then the DDF is linearly homogeneous in (x, y), and globally concave if $\mathcal{T}$ is convex. Building on Kuosmanen and Johnson (2017), we assume that the objective function is monotonically increasing in outputs y and monotonically decreasing in inputs x. We assume that the solution (x_i^*, y_i^*), which is unobserved, lies on the boundary of the production possibilities set, i.e.,

$$\overrightarrow{D}(x_i^*, y_i^*, g^x, g^y) = 0, \ \forall (g^x, g^y) > (\mathbf{0}, \mathbf{0}), \tag{3}$$

where i $(i = 1, ..., n)$ denotes the particular decision-making unit (DMU). The data generating process (DGP) is

$$\begin{aligned} x_i &= x_i^* + \varepsilon_i g^x, \\ y_i &= y_i^* - \varepsilon_i g^y, \ i = 1, \ldots, n, \end{aligned} \tag{4}$$

where $\varepsilon_i = v_i + u_i$, v_i is a two-sided error term, and $u_i \geq 0$ represents technical inefficiency, as in Stochastic Frontier Analysis (SFA). Therefore, we examine DGPs with directions in both the input and output directions to represent more general technologies. Moreover, we have

$$\overrightarrow{D}(\tilde{x}_i, \tilde{y}_i, g^x, g^y) = \varepsilon_i, \ i = 1, \ldots, n, \tag{5}$$

where

$$\begin{aligned} \tilde{x}_i &= x_i + (y_{1i}/g_1^y)g^x, \\ \tilde{y}_i &= y_i - (y_{1i}/g_1^y)g^y. \end{aligned} \tag{6}$$

In this formulation we assume g_1^y is strictly positive. If this not the case (as in our empirical

application where we have an input-oriented distance function which requires $g^y = \mathbf{0}$), we can use, for example g_1^x or any other input direction. However, as we have to normalize $g^x = 1$ it is easier to divide by g_1^x in (6). The results are invariant to which direction we choose for normalization. Kuosmanen and Johnson (2017) use Convex Non-Parametric Least Squares (CNLS) to solve the problem:

$$\begin{aligned} &\min_{a,\beta,\gamma,\varepsilon} \textstyle\sum_{i=1}^{n} \varepsilon_i^2, \\ &y_{1i}/g_1^y = a_i + \beta_i' \tilde{x}_i - \gamma_i' \tilde{y}_i + \varepsilon_i, i = 1, \ldots, n, \\ &a_i + \beta_i' \tilde{x}_i - \gamma_i' \tilde{y}_i \leq a_h + \beta_h' \tilde{x}_i - \gamma_h' \tilde{y}_i, \forall i, h, \\ &\beta_i' g^x + \gamma_i' g^y \leq 1, \forall i, \\ &\beta_i \geq 0, \gamma_i \geq 0, \forall i. \end{aligned} \tag{7}$$

Average efficiency is estimated in a second stage by estimating the minimum envelopment of the fitted values of the CNLS estimator (see also Kuosmanen and Kortelainen, 2012). Of course, another disadvantage of the approach is that the directional vector is the same across firms, but as Kuosmanen and Johnson (2017) write: "Our defense is that having a formal statistical model with clearly defined assumptions is far better than having no model at all." (op. cit., p. 795). On the positive side, Kuosmanen and Johnson (2017) do not use distributional assumptions about the components of the error ε_i. In turn, they can only estimate *average* inefficiency. Additionally, the directional vector is assumed to be fixed. This could be useful if there is a relevant direction desired by a regulator but otherwise it is difficult to be selected by the user. For example, Kuosmanen and Johnson (2017) fix $g_1^x = 1$ (this normalization is harmless) and then they select g_2^x (they have two inputs) by minimizing the value of White's test for heteroskedasticity or (as in their empirical application) they set it to zero. Assuming outputs are exogenous, they set the direction vectors for outputs equal to zero to have an input-oriented DDF. Finally, we have constant returns-to-scale provided $a_i = 0$, $\forall i$, which we assume throughout. As in Kuosmanen and Johnson (2017) we assume

$$\begin{aligned} x_i &= x_i^* + g_i^x \varepsilon_i, \\ y_i &= y_i^* - g_i^y \varepsilon_i, \; i = 1, \ldots, n, \end{aligned} \tag{8}$$

where, now, the directional vectors are random variables, and we have[1]

$$\tfrac{y_{1i}}{g_{1i}^y} = \overrightarrow{D}(x_i, y_i, g_i^x, g_i^y) - \varepsilon_i. \tag{9}$$

It appears, as Kuosmanen and Johnson (2017) write that the DGP in (4) is observationally

[1]For the treatment of x_i^* and y_i^*, see also the Online Supplement of Kuosmanen and Johnson (2017). Additionally, as these authors show, there is no endogeneity in the econometric sense, viz. correlation between the error components and explanatory variables.

equivalent to the errors-in-variables model in Kuosmanen et al. (2007). However, identification and non-reduction to Kuosmanen et al. (2007) follows from the fact that ε_i *is common across all equations in (4) whereas the directional vectors are non-negative.* So, we can think of (4) as a common factor model (where the common factor is ε_i) in multiplicative rather than additive form. Similarly, in stochastic frontier models, identification relies on the so-called signal-to-noise ratio. If this ratio is low, inefficiency would be hard to identify. Although identification *may* be weak, we propose to examine the matter empirically as in Section 5.3.

3 Model

3.1 Background material

3.1.1 Gibbs posterior

Suppose we have data from the vector random variable $\mathbb{Y} = (\mathbb{Y}_1, ..., \mathbb{Y}_n)$ (generated from a distribution P whose density is $f(\cdot;\theta)$), the random variables in $\mathbb{Y}$ are independent, and $\theta \in \mathbb{R}^d$ represents a vector of parameters of interest.The observed data is a realization from $\mathbb{Y}$, denoted by Y. If the model is correctly specified, updating of prior beliefs summarized in the prior $p(\theta)$, takes place according to Bayes' theorem

$$p(\theta|Y) \propto p(\theta)L(\theta;Y) = p(\theta)\prod_{i=1}^{n} f(y_i;\theta), \tag{10}$$

where $L(\theta;Y)$ is the likelihood function. If the model is not correctly specified then it is known that θ converges (asymptotically in the sample size) to a so-called "pseudo-true" parameter which minimizes the Kullback-Leibler divergence between the actual model and the parametric model $f(\cdot;\theta)$, e.g., Müller (2013). In many instances, we may not have a fully parametric model but instead we have a loss function, $\ell(Y,\theta)$ in the sense that the true parameter minimizes $\mathbb{E}_P\ell(Y,\theta)$ with respect to θ, where P denotes the true distribution of the data (see also Zellner, 1986). The empirical version of $\mathbb{E}_P\ell(Y,\theta)$ is $\mathcal{D}_n(\theta) = n^{-1}\sum_{i=1}^{N} l(Y_i,\theta)$. Given a prior $p(\theta)$, the Gibbs posterior is given as follows:

$$p_n(\theta|Y) = \frac{e^{-\omega n\mathcal{D}_n(\theta)}p(\theta)\,\mathrm{d}\theta}{\int e^{-\omega n\mathcal{D}_n(\overrightarrow{\theta})}p(\overrightarrow{\theta})\,\mathrm{d}\,\overrightarrow{\theta}}, \tag{11}$$

for some scale parameter $\omega > 0$. Parameter ω is the "learning rate" in the sense that as $\omega \to 0$ more weight is given to the prior (meaning that it is quite likely that there is misspecification, whereas as ω increases, more weight is placed on the likelihood, implying that there is no substantial misspecification.) This approach achieves better risk performance under model

misspecification as the Gibbs posterior (or α-posterior) is directly associated with a risk function (see Jiang and Tanner, 2008; Yao, Jiang, and Tanner, 2011).

3.1.2 Robust posteriors

Ghosh and Basu (2016), Basu et al. (1998) and Bissiri et al. (2016) developed the so-called density power divergence to robustify parameter inferences from a Bayesian model. A Gibbs posterior can be defined as

$$p(\theta|Y) = \frac{e^{-\mathcal{Q}_n(\theta)}p(\theta)}{\int e^{-\mathcal{Q}_n(\vec{\theta})}p(\vec{\theta})\,\mathrm{d}\,\vec{\theta}}, \tag{12}$$

where $\mathcal{Q}_n(\theta)$ is a loss function, for example as in (11). Miller and Dunson (2019) propose a robust posterior by raising the likelihood to the power $\frac{\alpha}{n+\alpha}$ and this approximation is good either when $n \gg \alpha$ or $n \ll \alpha$. The idea of using this "learning rate" has been used by many authors (Vovk, 1990; McAllester, 2003; Barron and Cover, 1991; Walker and Hjort, 2002; Zhang, 2006a). Specifically, Zhang (2006b) proposed the following robust posterior, also known as α-posterior:

$$p_\alpha(\theta|Y) = \frac{e^{-\alpha\mathcal{Q}_n(\theta)}p(\theta)}{\int e^{-\alpha\mathcal{Q}_n(\vec{\theta})}p(\vec{\theta})\,\mathrm{d}\,\vec{\theta}}, \tag{13}$$

where $\mathcal{Q}_n(\theta)$ is a loss function and $\alpha > 0$ is the learning rate; see also Jiang and Tanner (2008), and Bissiri et al. (2016). This form is similar to (11). Grünwald and van Ommen (2017) show that posteriors with different α (and α small enough) in fact coincide with posteriors *corresponding to different models*. Values of α close to zero indicate severe misspecification so, less weight is placed on the likelihood and more on the prior.

3.1.3 Objective priors

One particular instance of an objective prior is the so-called Unit Information Prior (UIP) which is crafted as to correspond to the information provided by a single observation (Kass and Wasserman, 1995). Suppose, for example, that the data Y arise from a normal distribution, $\mathcal{N}(\mu, \sigma^2)$. In large samples, the posterior can be approximated well by $\mathcal{N}(\bar{Y}, s^2/n)$, where $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ is the sample mean and $s^2 = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ is the sample standard deviation. The UIP in this case would have been $\mathcal{N}(\mu, \sigma^2)$ or, alternatively, $\mathcal{N}(\bar{Y}, s^2)$. More generally, the UIP is

$$p(\theta) \propto n^{-1/2}|\mathcal{I}(\theta)|^{1/2}, \tag{14}$$

where $\mathcal{I}(\theta)$ (also known as Fisher's information matrix) is equal to minus the *expected value (with respect to data given θ)* of the Hessian matrix of the log-likelihood:

$$\mathcal{I}(\theta) = -\left[\mathbb{E}_{\mathbb{Y}|\theta}\frac{\partial \ln L(\theta;\mathbb{Y})}{\partial\theta}\frac{\partial \ln L(\theta;\mathbb{Y})}{\partial\theta'}\right] \simeq -\left[\mathbb{E}_{\mathbb{Y}|\theta}\frac{\partial^2 \ln L(\theta;\mathbb{Y})}{\partial\theta\,\partial\theta'}\right]. \tag{15}$$

The two expressions in (15) are identical when the model is *correctly* specified, thus the

use of $\simeq$ in (15). If not, the two expressions will be different and it would be best to use the expectation of the Hessian of the log-likelihood. Our approach is to use the product of the score vectors (viz. $\frac{\partial \ln L(\theta;\mathbb{Y})}{\partial \theta}$) to give the model the "benefit of the doubt" in terms of specification. Of course, using the Hessian is not a problem from the computational point of view as it has to be used anyway in (24). A correction, however, is used in the selection of α as in (24) below. In this case Fisher's information matrix is divided by n to provide information given by a single observation.[2] There are alternatives; for example, the UIP can be crafted independently for each parameter, the parameter vector θ can be replaced by the maximum likelihood estimate or the posterior mean $\hat{\theta}$, etc. There are two ways to define the prior. The first option is to use (14) as is. The second option is to use the product of the diagonal elements of the covariance matrix in (14):

$$p(\theta) \propto \prod_{j=1}^{d_\theta} p_j(\theta_j),\ p_j(\theta_j) \propto n^{-1/2} \left[\mathcal{I}(\theta_j)\right]_{jj},\ j = 1, \ldots, d_\theta, \tag{16}$$

where $\left[\mathcal{I}(\theta_j)\right]_{jj}$ is the jth diagonal element of $\mathcal{I}(\theta)$, *and all other parameters are fixed at their median values from F-UIP.*[3] The second option is easy to compute as it requires the entire $\mathcal{I}(\theta)$ which, however, is available from the step leading to (14). Nevertheless, it is interesting to compare the performance of the two alternatives. The first approach is called Full UIP (F-UIP). The second, presented in (16), is called the product of marginals approach (PM-UIP) as in (16). All required derivatives in UIPs are computed numerically at each parameter value, using the approach of Dobmann et al. (1996). It should be noted that the PM-UIP also depends on all other parameters in vector θ but these are fixed at their median values from F-UIP.

3.2 StoNED approach and Bayesian inference

Our model is an extension of Kuosmanen and Johnson (2017) with firm-specific directions under normality of v_i ($v_i \sim \mathcal{N}(0, \sigma_v^2)$). We impose the harmless normalization $g_1^x = 1$.

[2]The UIPs are not necessarily proper (i.e. they do not necessarily integrate to one.) However, when restricted to bounded intervals (as the result of limits in computer precision) they are proper, for practical purposes.

[3]The problem of computing marginal priors in the F-UIP case is resolved as follows. Given the prior $p(\theta)$ we use a Metropolis-Hastings algorithm to draw from the prior conditional distributions of $\theta_j \sim \theta_j|\theta_{-j}$ ($j = 1, ..., d_\theta$) where θ_{-j} denotes all parameters with the exception of the jth element. We use 15,000 MCMC iterations omitting the first 5,000. Eventually, we produce a MCMC sample from $\theta_j^{(s)} \sim \theta_j|\theta_{-j}^{(s-1)}$ ($s = 1, \ldots, S$, with given initial conditions). In turn, the marginal prior of θ_j can be computed easily using kernel density estimation for $\left\{\theta_j^{(s)}, s = 1, \ldots, S\right\}$. Of course, these are empirical Bayes priors as they are defined based on the observed information matrix. These computations are necessary only for the purpose of presenting the marginal priors visually and finding the median values of the parameters in F-UIP. To draw a sample from the PM-UIP we generate $\theta_j^{(s)} \sim \theta_j|\tilde{\theta}_{-j}$, where $\tilde{\theta}_{-j}$ is the median of F-UIP parameters.

Posterior results are invariant to the selection of y_{1i}, i.e. we can select another output in place of y_{1i}. Moreover, (12) will correct or at least mitigate possible misspecification, an issue that we consider in the next section.[4]

Define $\tilde{x}_i = x_i - (y_{1i}/g^y_{1i})$ and $\tilde{y}_i = y_i + (y_{1i}/g^y_{1i})$. With firm-specific directions the CNLS procedure is:

$$\begin{gathered}\min_{a,\beta,\gamma} \sum_{i=1}^n \varepsilon_i^2, \\ y_{1i}/g^y_{1i} = a_i + \beta_i'\tilde{x}_i - \gamma_i'\tilde{y}_i + \varepsilon_i, i = 1, \ldots, n, \\ a_i + \beta_i'\tilde{x}_i - \gamma_i'\tilde{y}_i \leq a_h + \beta_h'\tilde{x}_i - \gamma_h'\tilde{y}_i, \forall i, h, \\ \beta_i' 1_m + \gamma_i' 1_s \leq 1, \forall i, \\ \beta_i \geq 0, \gamma_i \geq 0, \forall i,\end{gathered} \tag{17}$$

where 1_m denotes the vector of ones in $\mathbb{R}^m$. The third constraint is imposed to guarantee monotonicity with respect to the first output. The constraints are nonlinear in terms of β_i and the directions. Alternatively, we have the problem

$$\begin{gathered}\min_{a,\beta,\gamma} \sum_{i=1}^n \left(y_{1i}/g^y_{1i} - a_i - \beta_i'\tilde{x}_i + \gamma_i'\tilde{y}_i + u_i\right)^2, \\ a_i + \beta_i'\tilde{x}_i - \gamma_i'\tilde{y}_i \leq a_h + \beta_h'\tilde{x}_i - \gamma_h'\tilde{y}_i, \forall i, h, \\ \beta_i' g^x_i + \gamma_i' g^y_i \leq 1, \forall i, \\ \beta_i \geq 0, \gamma_i \geq 0, \forall i.\end{gathered} \tag{18}$$

In this formulation, u_i and the directional vector act as latent unobservable variables. The directional vectors affect the concavity restrictions in the Afriat inequalities (first restriction) but this is not a problem per se in MCMC. Other than that, the loss function is quadratic subject to a number of linear inequalities and non-negativity constraints, plus the definitions in (6). To avoid solving the problem in (18) as in the Monte Carlo experiment for each MCMC replication, we write the constraints as follows.

$$\begin{gathered}a_i + \beta_i'\tilde{x}_i - \gamma_i'\tilde{y}_i = a_h + \beta_h'\tilde{x}_i - \gamma_h'\tilde{y}_i - U_{ih}, \forall i, h, \\ \beta_i' g^x_i + \gamma_i' g^y_i = 1 - U_{i0}, \forall i, \\ \beta_i \geq 0, \gamma_i \geq 0, \forall i,\end{gathered} \tag{19}$$

where U_{ih} and U_{i0} are independent *non-negative* random variables following a half-normal distribution with scale parameter σ_U and σ_{U0}, respectively.[5] These scale parameter have a

[4] Using non-parametric techniques to separate the error term in a one-sided and a two-sided component is not innocuous particularly in small samples. A disadvantage is that one can only estimate average inefficiency in the sector as in Kuosmanen and Johnson (2017).

[5] This is a distributional assumption but results from a *gamma* and exponential specification did not yield substantial differences and they are available on request.

prior of the form $p(\sigma_U) \propto \sigma_U^{-1}$ and $p(\sigma_{U_0}) \propto \sigma_{U_0}^{-1}$. Their posteriors are updated as

$$\begin{array}{c} \frac{\sum_{(i,h;i\neq h)}(a_i+\beta_i'\tilde{x}_i-\gamma_i'\tilde{y}_i-a_h-\beta_h'\tilde{x}_i+\gamma_h'\tilde{y}_i+U_{ih})^2}{\sigma_U^2}|\theta, Y \sim \chi^2_{n(n-1)}, \\ \frac{\sum_{i=1}^n(\beta_i'g_i^x+\gamma_i'g_i^y-1+U_{i0})^2}{\sigma_{U_0}^2}|\theta, Y \sim \chi^2_n. \end{array} \tag{20}$$

where Y denotes the data and θ is defined below in (21). The Us are updated using their posterior conditional distributions as in standard SFMs. Our vector of parameters is

$$\begin{array}{c} \theta = \{\theta_i\}_{i=1}^n = \{a_i, \beta_i', \gamma_i', g_i^{x\prime}, g_i^{y\prime}, u_i\}_{i=1}^n, \\ \beta_i \geq \mathbf{0},\ \gamma_i \geq \mathbf{0},\ g_i^x \geq \mathbf{0},\ g_i^x \geq \mathbf{0},\ u_i \geq 0,\ \forall i = 1, \ldots, n. \end{array} \tag{21}$$

To define the UIP prior we proceed in two steps.

Step 1: We assume $g_i^y = g^y$, $g_i^x = g^x$, and $u_i = u$ are common $\forall i$. The prior on all parameters is a UIP as defined in (14). So, we can recover the posterior of g^x, g^y, and u.

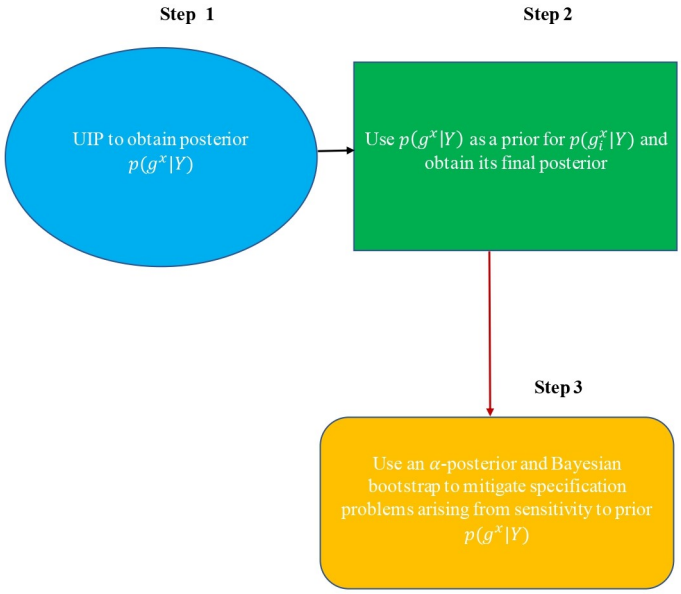
Step 2: As the prior of $\{g_i^x, g_i^y, u_i\}_{i=1}^n$ we adopt the posterior from Step 1.[6]

This is, essentially, a data-based prior where the UIP is used in the first step. In the presence of multiple latent variables in $\{g_i^x, g_i^y, u_i\}_{i=1}^n$ which account for heterogeneity for the directional vectors and inefficiency, it is best to avoid a UIP due to singularity problems of the *Hessian matrix. So, in the first step, these parameters are assumed to be common for all observations, and in the second step we correct for this restrictive assumption. The first step corresponds to the approach in Kuosmanen and Johnson (2017). The second step, assigns a prior to observation-specific directions and inefficiencies based on the posterior of common parameters (across i) in the first step.* Can this way of defining the prior bias the final posterior results? The question can be resolved on empirical grounds as there is no known theoretical results about this case. In principle, we accept, as a working hypothesis, that final results can be affected but we deal with this potential specification error: *First*, we use a Gibbs posterior whose parameter α mitigates the problem. *Second*, we use the Bayesian bootstrap to resolve the same problem in addition to corrections made by the Gibbs posterior. If α is close to zero then we know that the model is misspecified and more weight is placed on the prior. If α remains relatively large then the model is, most likely, well specified. Of course, this prior is data-based but the problem is not significant unless one is a Bayesian "purist". We illustrate the concepts in Figure 1. *First*, we use a UIP to produce an "objective" prior for the directions, assuming they are common across all observations. *Second*, we obtain a posterior $p(g^x|Y)$ (assuming, say, $g^y = \mathbf{0}$, so that we have input orientation of the DDF.) We will use this posterior as a prior when directions

[6]The joint posterior of $\{g^x, g^y, u\}$ is approximated using a multivariate kernel density technique so that it can evaluated point-wise in the second step.

are observation-specific to obtain a final posterior. Of course, this prior is informative and, therefore, it may bias the final results in various ways. *Third*, to eliminate or, at least, mitigate such problems we resort, in Step 3, to the Gibbs posterior or α-posterior. We can improve on these results by using, as in Step 3, a Bayesian bootstrapping procedure as specification errors should be evident in different bootstrapped samples of the data. It would be important to draw a clear distinction between the direction of the DGP (the usual approach is to set $g^x = 0$, $g^y = 1$) and the direction for the efficiency measurement (e.g., input or output orientation). Kuosmanen and Johnson (2017) assume that DGP has a fixed direction in the estimation stage (their Section 4.1), which is clearly somewhat restrictive, but they do allow for firm-specific directions in the efficiency analysis stage (Section 4.3) once the frontier has been estimated. The present study similarly starts with a fixed direction *to craft a UIP prior only* and proceeds to estimate the firm-specific directions in an iterative, data-driven fashion. This is an important distinction.

Figure 1: Model



We use Markov Chain Monte Carlo (MCMC) techniques to provide access to the Gibbs posterior

$$p_\alpha(\theta|Y) = \frac{e^{-\alpha Q_n(\theta)} p(\theta)}{\int e^{-\alpha Q_n(\overrightarrow{\theta})} p(\overrightarrow{\theta}) \mathrm{d}\overrightarrow{\theta}}, \tag{22}$$

where $Q_n(\theta)$ is the objective function of (18) as in (12), and $p(\theta)$ is one of the UIPs. Therefore, the Gibbs posterior is

$$p_\alpha(\theta|Y) \propto \exp\left\{-\alpha \sum_{i=1}^{n} \left(y_{1i}/g_{1i}^y - a_i - \beta_i' \tilde{x}_i + \gamma_i' \tilde{y}_i + u_i\right)^2\right\} p(\theta), \tag{23}$$

subject to the constraints of (18). MCMC is implemented using 150,000 iterations omitting the first 50,000 in the interest of mitigating possible start up effects. Specifically, we use the Girolami and Calderhead (2011) Hamiltonian Monte Carlo method, and Geweke's (1992) diagnostics to monitor convergence. To use the posterior in (22), in principle, we need to solve (18) for each MCMC iteration. However, this is avoided by using the non-negative terms in U,see (19). *First,* we condition on directions and draw all other parameters. *Second,* given all other parameters, we draw the directions. Implementation of the Girolami and Calderhead (2011) procedure in two steps is easy. An alternative is to solve the CNLS problem, using the "sweet spot" approach of Lee et al. (2013) but this is not necessary. *The selection of α is guided by Fisher's information (it is, therefore, data-driven if we omit the expectation in Fisher's information matrix)* and is given by

$$\alpha = \frac{\mathrm{tr}\{\mathcal{J}(\bar{\theta})\mathcal{I}(\bar{\theta})^{-1}\mathcal{J}(\bar{\theta})'\}}{\mathrm{tr}\{\mathcal{J}(\bar{\theta})\}}, \tag{24}$$

at a point $\bar{\theta}$, where $\mathcal{J}(\theta) = n^{-1}\sum_{i=1}^{n}\nabla^2\ell(Y_i,\theta)$, $\mathcal{I}(\theta) = n^{-1}\sum_{i=1}^{n}\{\nabla\ell(Y_i,\theta)\nabla\ell(Y_i,\theta)'\}$ where $\ell(Y_i,\theta)$ denotes the loss function, ∇ is the gradient operator with respect to θ, and $\bar{\theta}$ is the first-stage posterior mean estimate of θ (Lyddon et al., 2019). The expression in the numerator of (24) is the so-called *sandwich covariance matrix* (Müller, 2013) so, this method of selecting α matches information from the pseudo-true and misspecified model parameters (the pseudo-true value of the parameter minimizes the Kullback–Leibler divergence between the true but unknown DGP and the parametric DGP we actually use (e.g., Müller, 2013). Notice that this definition of α is unlike the definition of UIP priors where only first derivatives are used, effectively assuming that there is no misspecification, approximately, but a correction is made in (24). Before proceeding, we summarize our procedure in Algorithm 1. On an i9 Core 3.30 GHz desktop running compiled `fortran 77` code, the algorithm takes about five minutes of CPU time.

Table 1: Treatment of misspecification

1. Use the optimal value of α in (24) to get a posterior that is robust to specification problems.
2. Embed step 1 in a Bayesian bootstrap procedure so that different αs are generated for each bootstrap replication.
3. Based on the bootstrap distribution of α one can decide whether misspecification is present.
4. If α is small, then less weight is placed on the likelihood.
5. If α is large then less weight is placed on the prior.
6. Therefore, small values of α indicate misspecification.

Algorithm 1 Estimation procedure

1. Compute the robust prior $p(\theta)$ as in (14). Numerical derivatives are used to compute the prior. There are two priors, F-UIP and PM-UIP that correspond, respectively, to full and product of marginals approach. Both priors are based on the posterior distribution of common directions in the Kuosmanen and Johnson (2017) formulation.
2. Compute the robust posterior in (22) based on the CNLS approach in (18).
3. Combine the prior and the likelihood to obtain the posterior in (23). This involves the choice of parameter α as in (24) which helps in mitigating misspecification problems.
4. Use MCMC techniques to obtain a sample of parameters that converges to the posterior.
5. Further mitigation of misspecification in robust posteriors is performed via the bootstrap. We use B=500 replications throughout.
6. Testing of identification is possible in various ways, as in Section 5.3.

We consider it useful to summarize the way we deal with misspecification as in Table 1.

4 Monte Carlo experiment

In our Monte Carlo study, we have three inputs and three outputs generated as in (4), where the x_i^*s and y_i^*s are generated from a uniform distribution in the interval (1, 10). Noise v_i is generated from a Student-t distribution with five degrees of freedom, location parameter zero

and scale parameter 0.1. This is adopted to introduce misspecification as we use normality of v_i. The one-sided error is generated from a uniform distribution in the interval (0, 0.25). The directional vector g_i^x has $g_{1i}^x = 1$ $(\forall i)$ and its other two elements are generated as uniform in the interval (0, 1). We draw g_i^y from the same distribution but without normalization. As we introduce output-specific directional vectors as well, the DDF does not admit a representation as input- or output-oriented. However, we use this formulation in the interest of generality. Notice that directions are firm-specific. The distance function has the form

$$y_{1i}/g_{1i}^y = 0.1 + 0.4\tilde{x}_{1i} + 0.3\tilde{x}_{2i} + 0.2\tilde{x}_{3i} - 0.5y_{1i} - 0.7\tilde{y}_{2i} - 0.5\tilde{y}_{3i} - \varepsilon_i, \tag{25}$$

where $\varepsilon_i = v_i + u_i$, and $\tilde{x}_i$ and $\tilde{y}_i$ are defined below (7). This data generating process (DGP) differs from the one we use in several respects. Noise is non-normally distributed (Student-t with five degrees of freedom), inefficiency does not follow a half-normal distribution (but a uniform), and the directional vectors are generated from different distributions (viz., uniform) that the ones we assume (viz., UIP). We consider samples of various sizes (n) and we report our Monte Carlo results in Table 2, using 1000 Monte Carlo replications. Reported in Table 2 are mean absolute errors (MAE) between actual and estimated (posterior-mean based) measures, for inefficiency, u_i, and the directional vectors g_i^x and g_i^y. For the directional vectors we consider the mean across all inputs and outputs, respectively as well as their mean across firms. Using the wrong specification (normal-half error components, etc.) yields MAEs which increase with the sample size which is reasonable as biases accumulate with increasing n. The robust specification performs much better and its MAEs are much smaller declining sharply with n. This means that both firm-specific inefficiency and directions can be estimated accurately in moderately sized samples. In Table 3 we report means (with standard deviations in parentheses) of the optimal values of α when we use the correct specification (DGP) and when we use the misspecified model. For the correct DGP, the values of α across Monte Carlo simulations are fairly close to one and increase with the sample size, while for the misspecified case the values of α approach zero as the sample size increases. Moreover, we use F-UIP as the results from PM-UIP were, practically, the same. We embed the procedure within the bootstrap with B=500 replications to robustify further the posteriors.

Table 2: Monte Carlo results

	Wrong specification	Robust specification
u_i		
$n = 100$	0.042	0.013
$n = 250$	0.055	0.007
$n = 500$	0.061	0.005
$n = 1000$	0.088	0.003
g_i^x		
$n = 100$	1.00	1.00
$n = 250$	0.225	0.019
$n = 500$	0.315	0.007
$n = 1000$	0.436	0.005
g_i^y		
$n = 100$	0.226	0.009
$n = 250$	0.307	0.007
$n = 500$	0.421	0.005
$n = 1000$	0.497	0.003

Notes: Reported are mean absolute errors (MAE) between actual and estimated (posterior-mean based) measures.

Table 3: Optimal values of α

	Correct specification	Wrong specification
$n = 100$	0.895	0.015
	(0.013)	(0.004)
$n = 250$	0.970	0.007
	(0.010)	(0.002)
$n = 500$	0.955	0.004
	(0.012)	(0.001)
$n = 1000$	0.978	0.002
	(0.010)	(0.001)

Notes: Reported are means (across Monte Carlo replications) of optimal values of α, with standard deviations in parentheses.

5 Empirical application

5.1 Data and posterior results

For our empirical application we use an unbalanced panel with 9,807 bank-year observations for 285 large U.S. commercial banks operating in 2001-2019 (quarterly data), whose total assets were more than one billion dollars (in 2005 U.S. dollars) in the first three years of observation. The data for 2007-2010 serve to focus on the effects of the financial crisis and its aftermath. The data come from Call Reports available from the Federal Reserve Bank of Chicago. For the period 2001-2010 the data have been used in Malikov et al. (2016). For detailed description of the data construction, see Section 5 of Malikov et al. (2016). The list of included variables is as follows: y_1 Consumer Loans, y_2 Real Estate Loans, y_3 Commercial & Industrial Loans, y_4 Securities, y_5 Off-Balance Sheet Activities Income, x_1 Labor, number of full-time employees, x_2 Physical Capital (Fixed Assets), x_3 Purchased Funds, x_4 Interest-Bearing Transaction Accounts, x_5 Non-Transaction Accounts. We denote banks by $i \in \mathscr{I} = \{1, \ldots, n\}$ and time periods by $t \in \mathscr{T} = \{1, \ldots, T\}$.[7] Unlike the analysis in previous sections, we have panel data but this does not introduce any significant complications as we treat the data as cross-sectional (otherwise we would have to assume, for example, that directional vectors and inefficiency are time-invariant, which is restrictive). As in the Monte Carlo experiment we use the F-UIP formulation and we compare the results with PM-UIP. However, to have an input-oriented DDF, we set $g^y = \mathbf{0}$. Therefore, we normalize by $g_1^x = 1$ in (6). Marginal effects of inputs (β coefficients) and outputs (γ coefficients) from F-UIP are reported in Table 4. The median difference of posterior means across the sample is denoted $\mathscr{D}$. Evidently, the posterior results are fairly similar when we use PM-UIP instead of F-UIP. This result is not entirely unexpected as we control the amount of information in both versions of UIP.

[7] Although the panel is unbalanced we do not index T by i (viz., T_i) in the interest of simplifying the notation.

Table 4: Marginal effects of inputs (β) and outputs (γ)

	mean	s.d.	min	max	$\mathscr{D}$
β_1, Labor	0.215	0.035	0.012	1.000	0.0031
					(0.001)
β_2, Capital	0.344	0.029	0.010	0.758	0.0052
					(0.002)
β_3, Purchased Funds	0.255	0.040	0.000	0.585	–0.040
					(0.006)
β_4, Interest-Bearing Transaction Accounts	0.107	0.033	0.005	0.337	0.0032
					(0.0010)
β_5, Non-Transaction Accounts	0.086	0.015	0.015	0.128	–0.0030
					(0.0012)
γ_1, Consumer Loans	7.44	2.55	0.000	15.45	–0.0050
					(0.002)
γ_2, Real Estate Loans	12.55	4.81	0.005	21.15	0.0022
					(0.001)
γ_3, Commercial & Industrial Loans	65.73	15.30	0.020	90.45	–0.0041
					(0.001)
γ_4, Securities	55.18	7.86	0.018	110.4	–0.0032
					(0.0011)
γ_5, Off-Balance Sheet Activities Income	87.30	15.20	0.000	105.3	0.0052
					(0.0012)

Notes: All results are taken with respect to observation–specific posterior means. The median difference of posterior means across the sample (F-UIP minus PM-UIP) is denoted $\mathscr{D}$.

Table 5: Estimated directions and inefficiency

(a) Directional vectors

	mean	s.d.	min	max	$\mathscr{D}$
g_1^x, Labor	1.000	0.000	1.000	1.000	0.000
					(0.000)
g_2^x, Capital	0.252	0.087	0.00	0.81	0.0020
					(0.002)
g_3^x, Purchased Funds	0.178	0.055	0.00	0.50	–0.0032
					(0.0020)
g_4^x, Interest-Bearing Transaction Accounts	0.202	0.067	0.055	0.055	0.0010
					(0.0010)
g_5^x, Non-Transaction Accounts	0.083	0.032	0.000	0.001	–0.0021
					(0.001)

(b) Inefficiency

Period	sampling means and s.d
2001-2006	0.142
	(0.034)
2007-2010	0.209
	(0.048)
2011-2019	0.124
	(0.017)
2011-2019	0.158
	(0.020)

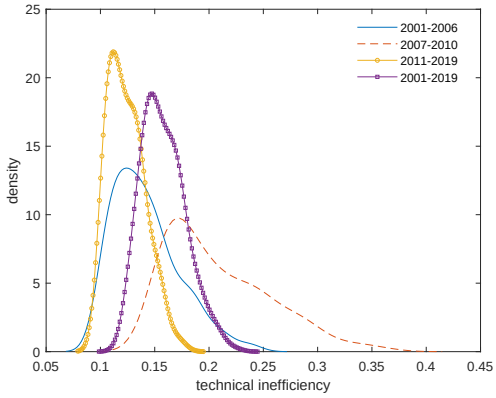
Notes: Reported are sample averages of posterior mean estimates. The median difference of posterior means across the sample between (F-UIP minus PM-UIP) is denoted $\mathscr{D}$. We set $g^y = \mathbf{0}$. Figures in parentheses are sample standard deviations.

In what follows we convert inefficiency loss to dollars as have input price data. The output coefficients γ have a price interpretation whereas input coefficients β are unit-invariant.[8] For

[8] Using multiple directions in CNLS may violate the global convexity of the production technology. LJSF (2020) derive the conditions under which multiple directions can be used in the CNLS problem while maintaining the axiomatic property of global convexity of the production technology. In our case, convexity is imposed via the Afriat inequalities and latent variables U.

example, increasing purchased funds by one dollar increases inefficiency by 17.8 cents on average, in this particular direction. If we weight by the directional vectors in all inputs, inefficiency averages 37.8% (s.d. 15.2%) which translates, approximately, to $36.4 million in 2005 prices. In Malikov et al. (2015), cost inefficiency is, approximately, 4% so the difference is quite large. This difference can be attributed to (i) the different method of estimation and inference; (ii) the estimation of bank-specific directional vectors, and (iii) that there is no need for distributional assumptions in our model as, effectively, their effect is mitigated through the robust Gibbs posterior. These calculations are based on parameters and directional posterior mean parameter vectors separately for the 2001-2006 period, 2007-2010, and 2011-2019. For the entire sample, inefficiency averages 15.8% (s.d. 2%) which translates, approximately, to $36.9 million at 2005 prices so, the two estimates are quite close. Inefficiency by sub-period is reported in the second part of Table 5. In terms of technical inefficiency (u_i) we present its sampling distribution in panel (a) of Figure 2 for all observations (i.e., we present the sampling distribution of posterior-mean-based inefficiency estimates). It appears that during the sub-prime crisis we have an increase in technical inefficiency. More specifically, during the 2001-2006 period, technical inefficiency averages 14.02% (standard deviation 3.4%). During 2007-2010, technical inefficiency averages 20.9% (standard deviation 4.8%), and during 2011-2019, it averages 12.4% (standard deviation 1.7%).

Figure 2: Sampling distributions



Notes: We report the sampling distribution of technical inefficiency (estimates using the posterior means).

Another parameter of interest is α, the learning rate estimated as in (24). The values of α are reported in Table 6 for the different sub-periods as well as the entire sample. Notice that for each MCMC draw, a different value of α is generated so, there is an overall

posterior distribution of α. We use B=500 bootstrap replications in this instance.The value of α is substantially smaller during the sub-prime crisis indicating severe misspecification but in the other sub-periods as well as in the entire sample, the values of α are relatively larger. To gain some insight in relatively large samples, from Table 3, the optimal value of α in misspecified models is (much less than) 0.002 (posterior standard deviation 0.001). Therefore, during 2007-2010, it seems that the model is correctly specified and using the entire sample or other sub-periods the values of α are quite large so, they place substantial weight on the likelihood, relative to the UIPs. Notice that posterior standard deviations of α (uncertainty is introduced as α is recomputed with each MCMC draw) are quite low so posterior mean estimates are quite precise. Moreover, we report median absolute differences of posterior means of parameters and directions in the column headed "differences". These are differences between posterior results from F-UIP and PM-UIP and they are presented graphically in panel (d) of Figure 2. Unsurprisingly, the results are nearly the same (although not identical) as the UIP controls the amount of information in the first-stage prior (i.e., equal to one observation.) Finally, as α seems not to change much in different sub-periods, we consider the full sample in the following discussion.

Table 6: Values of learning rate, α and differences relative to PM-UIP

Sub-period	α	differences
2001-2006	0.822	0.0005
	(0.041)	(0.0002)
2007-2010	0.817	0.0002
	(0.045)	(0.0001)
2011-2019	0.829	0.0001
	(0.047)	(0.002)
2001-2019	0.835	0.0002
	(0.043)	(0.0002)

Notes: Reported are sample means of posterior means with sample standard deviations in parentheses. Uncertainty arises as different values of α are computed for each MCMC draw. Moreover in the column labeled "differences" we report median absolute differences between posterior results from F-UIP and PM-UIP.

5.2 Further misspecification

An alternative way of eliminating or at least mitigating misspecification is using the Bayesian bootstrap; e.g., Huggins and Miller (2019). In the bootstrap, we draw random samples from

the data (preserving all blocks of observations for each bank) and we perform posterior analysis using the Gibbs posterior with the optimal values of α. An interesting benchmark against which we can compare the new techniques is to adopt half-normal or exponential distributions for the one-sided error component (in the second stage.) In other respects, the problem (18) remains the same but half-normal or exponential distributions are used as priors of u_i *explicitly*. The purpose of the comparison is to examine whether the Gibbs posterior can still be used with this type of misspecification of the prior. In panels (a) and (b) of Figure 3, we provide results for the half-normal case and in panels (c) and (d) for the exponential distribution. From these results, it is evident that the results for the half-normal and the exponential are approximately the same although from the densities for α we conclude that αs are several orders of magnitude smaller than results corresponding to UIP, indicating more misspecification. In other respects, the distributions of technical inefficiency are quite different compared to those delivered by UIP. Moreover, some technical inefficiency densities are asymmetric and multimodal. Asymmetry is not totally unexpected whereas bimodality reflects the presence of different groups in the data related to heterogeneity in performance. This may also be due to heterogeneity in estimated directions along which inefficiency is estimated.

Besides u_i, the directional vectors are also important. In Figure 4 we present the median absolute difference between the half-normal-based and exponential-based estimates relative to those delivered by bootstrapping and UIP. We report results for all directions g^x (except g_1^x which is normalized to be 1) in the interest of graphical simplicity in Figure 4 from which it is evident that the directions are also relatively robust relative to F-UIP as the differences in posterior means of differences are quantitatively small. In this sense, the Gibbs posterior can, indeed, at least mitigate misspecification problems arising from the prior or other aspects of the model.

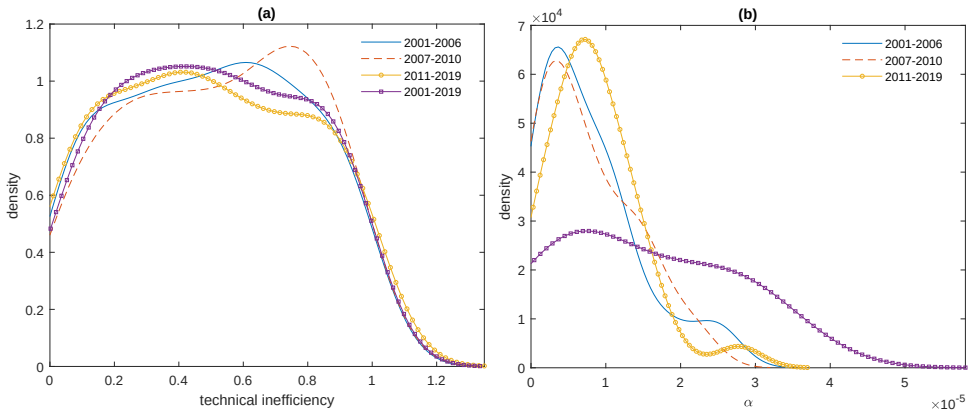
5.3 Identification

As identification may be tenuous for the DGP in (4) as both ε_i and g_i^x are random (we set $g^y = \mathbf{0}$ so that we have an input-oriented DDF interpretation), we propose to test identification in our empirical application using different priors. Comparing priors and posteriors to examine whether learning takes place (which means that a parameter of interest is identified) may not be enough. What is required is that

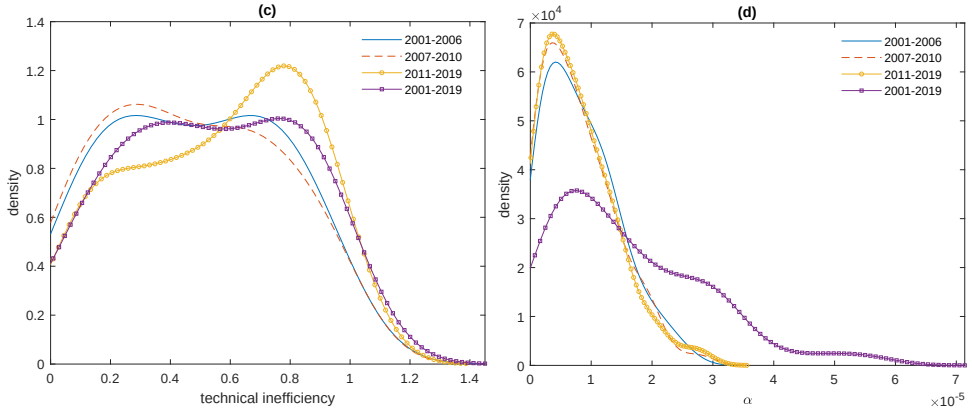
$$p(\theta_1|Y) \neq \mathbb{E}_{\theta_2|Y}[p(\theta_1|\theta_2)], \tag{26}$$

Figure 3: Sampling distributions, half-normal distribution

(a) Half-normal

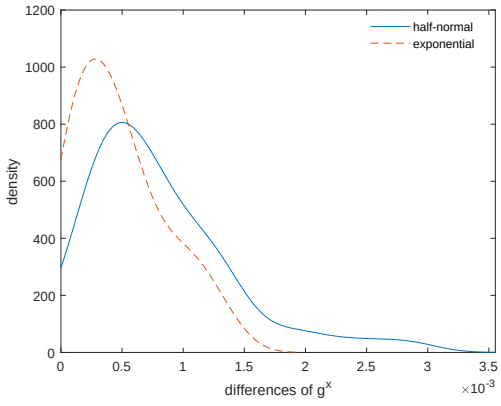


(b) Exponential



Notes: In panels (a) and (b) we report the sampling distribution of technical inefficiency and α (estimates using the posterior means) for the half-normal assumption for technical inefficiency. In panels (c) and (d) we present similar results for the exponential case. The results are presented over different periods (2001-2006, 2007-2010, 2011-2019, and the the entire sample).

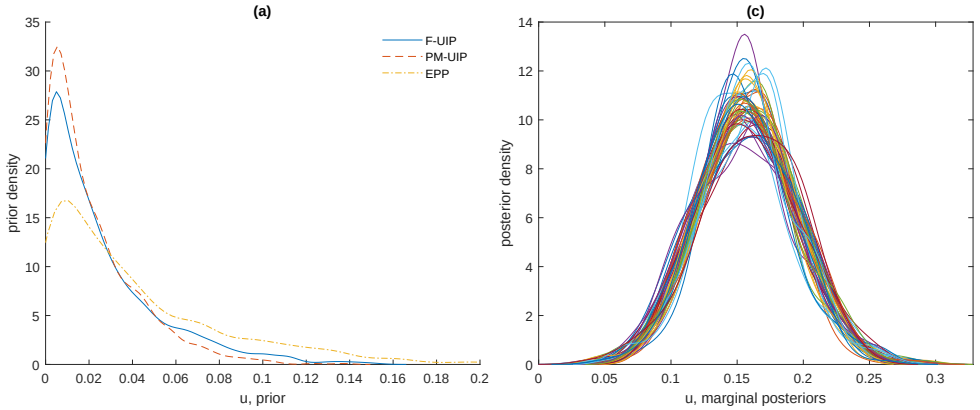
Figure 4: Median absolute differences of directions relative to F-UIP



Notes: Reported are sampling distributions of median absolute differences in directional vectors from the half-normal (straight line) and the exponential (dotted line) specifications relative to F-UIP. Differences are taken with respect to all directions for visual clarity. In both cases we use the bootstrap with B=500 replications. In panel (b) we do not. We set $g^y = \mathbf{0}$.

where θ_1 is a parameter of interest, θ_2 denotes the rest of the parameter vector, and $p(\theta_1|\theta_2)$ is the prior of θ_1 conditionally on θ_2. The term $\mathbb{E}_{\theta_2|Y}[p(\theta_1|\theta_2)]$ is known as the *expected posterior prior* (EPP, Poirier, 1998; Koop et al., 2013). So, comparing priors and posteriors is acceptable only when $p(\theta_1|\theta_2) = p(\theta_1)$ i.e., the parameters are independent. As this condition holds in our case we proceed to examine identification of certain parameters. In the left panel of Figure 5, we report the prior (thick line) and in the right panel we report marginal posteriors of u_i for fifty randomly selected banks. Again, the results show that the directional vectors are identified (we remind that $g^y = \mathbf{0}$). Moreover, evidence in Figure 5 shows that inefficiencies are identified as well. In panel (a) we present three priors viz., F-UIP, PM-UIP, and EPP previously defined in (26). This notion of identification is correct only when parameters are a priori independent which is not true for PM-UIP or for the F-UIP and EPP (see Footnote 3). Moreover, contrary to PM-UIP, the EPP is more concentrated around zero. We will investigate identification further in the following discussion.

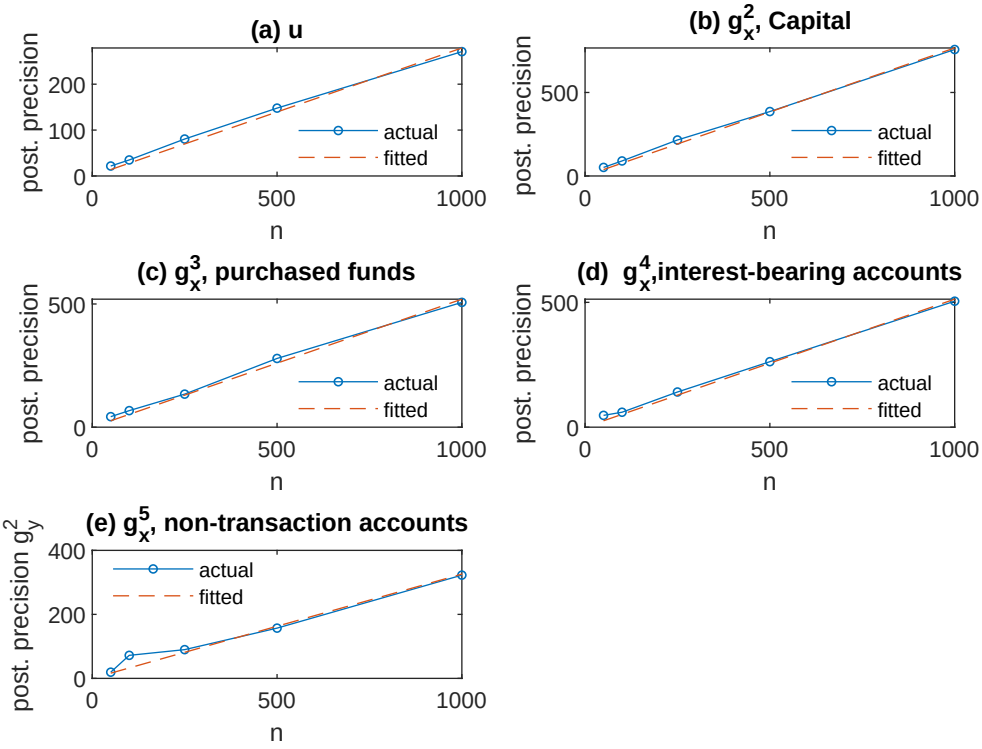
Figure 5: Comparison of prior and marginal posteriors of u_i



Notes: In the left panel we report the priors and in the right panel we report marginal posteriors of u_i for fifty randomly selected banks. EPP is the "expected posterior prior" in (26).

An alternative way to examine identification is to generate data sets of different size using the parameters estimated via the posterior means. If the model is identified, then posterior precision (measured by the inverse determinant of the sub-matrix of the posterior covariance matrix of the parameters, should increase proportionally with the sample size. As quite often we have relatively few observations in the time dimension, we select T=5. In turn, we vary n for samples of size 50, 100, 250, 500, and 1,000. Our results are reported in Figure 6, for median u_i and median of certain directions. The straight lines represent actual posterior precision and the dotted lines are fitted values from a regression of posterior precision on the sample sizes (without intercept.) In all cases (viz., all six panels of Figure 6) we consider medians of bank-specific inefficiencies or directions). As the two lines are quite similar, we conclude that the model is identified using the F-UIP or PM-UIP. This is not surprising as the second-stage prior is quite informative. This does not mean that prior and posterior results will be the same, or that the posterior is sensitive to the prior. The prior comes out as the result of assuming that directions are the same across all observations. Therefore, it is informative on the one hand, but it is corrected through α -posteriors and bootstrapping.

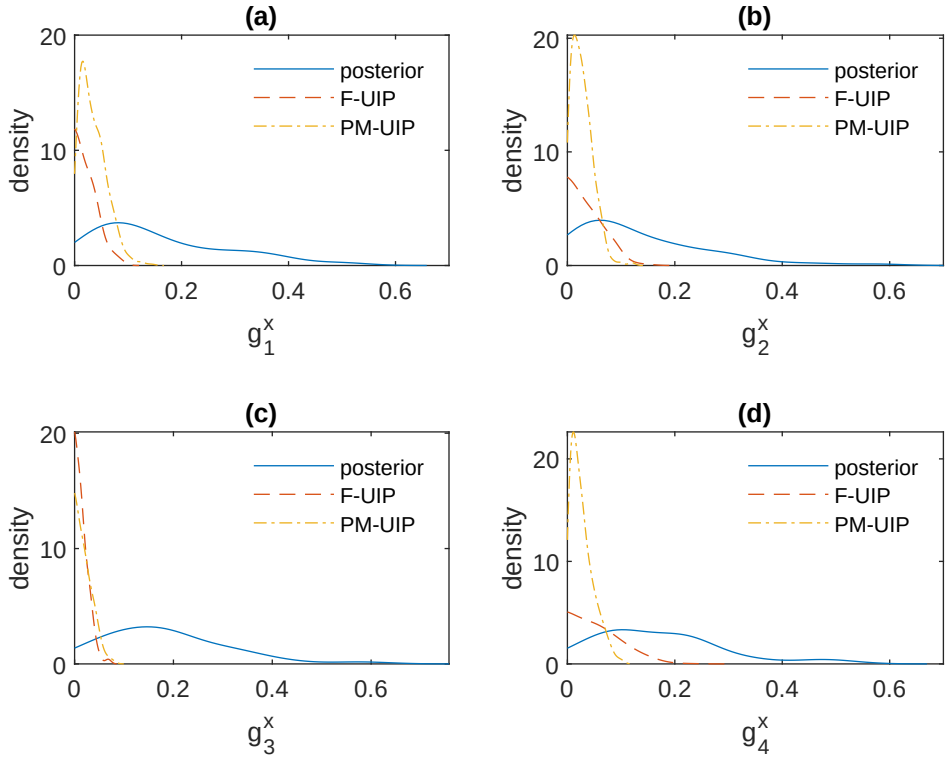
Figure 6: Posterior precision for different parameters



It would be interesting to know what the F-UIP or PM-UIP look like. We present these densities in Figure 7. The posterior and priors are for a randomly selected observation. Results for other observations were qualitatively the same, so they are not reported here to save space but they are available on request. The priors are truncated to the range of the posterior for visual clarity. The F-UIP and PM-UIP (applied to the first stage problem where g^x are common across firms) are quite different which implies that we should not assume the kind of prior parameter dependence in the PM-UIP. Unsurprisingly, the UIPs are quite informative despite the fact that UIPs are worth a single observation. The reason is that in (14) the priors are based on posteriors of the directions of (common across observations) g^x. Here, we use the *observed* Fisher information matrix whereas, normally, $\mathcal{I}(\theta)$ should be the *expected value* of the negative of the Hessian matrix of the log-likelihood.[9] In practice, the expectation is hard to compute so, in practice we use the observed information matrix. This is a way of making the prior depend on the data as in empirical Bayes procedures. Although, formally, this compromises "Bayesian purity" it is, arguably, a better procedure from the practical point of view. In turn, even when α is small giving more weight to the prior is giving *less* weight on a *data-dependent* prior, see Table 6.

[9]Any difference between the two implies specification problems.

Figure 7: Posteriors, F-UIP and PM-UIP

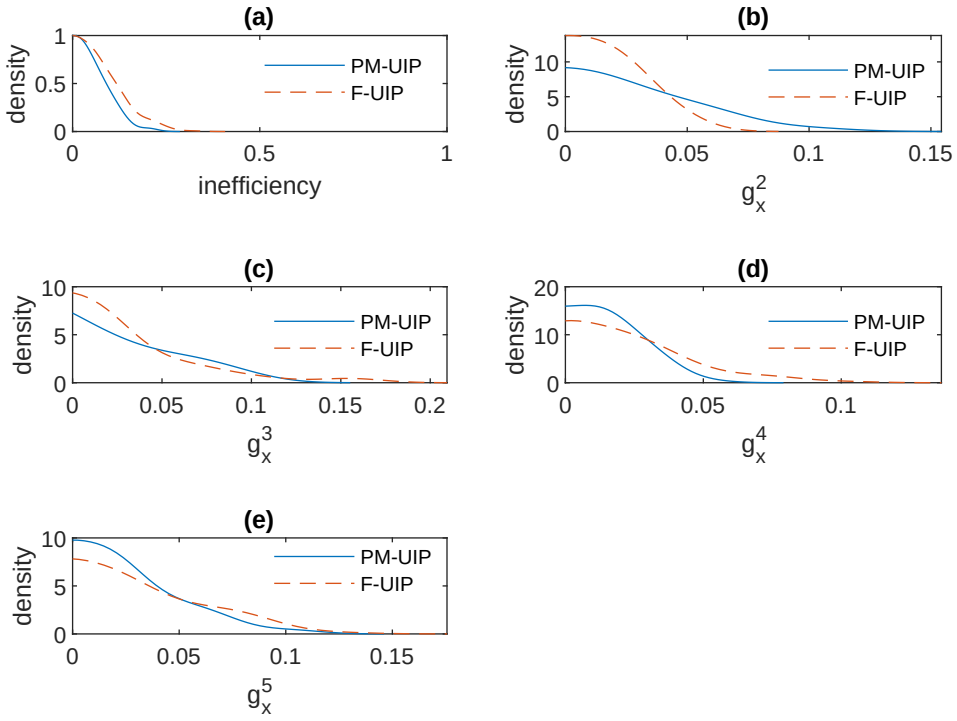


To examine in more detail the *expected* UIPs, we provide them in Figure 8 for the empirical application. The expectations are approximated using Monte Carlo with 10,000 replications in place of the observed data. In each case we report the marginal priors. In terms of inefficiency (u), the (second-stage) priors tend to allocate more prior probability towards zero and they are quite informative. In this sense, the prior parameter dependence in F-UIP does not seem to imply different results compared to PM-UIP (which incrporates a *different* kind of dependence).

6 Concluding remarks

In this paper, we extend the directional distance function approach of Kuosmanen and Johnson (2010) in several respects. First, we introduce firm-specific directions which are estimated rather than assumed a priori. Second, we use a UIP to specify explicitly the amount of information (which corresponds to one observation.) Two such priors are developed. Third, we use a robust posterior which accounts for potential misspecification and it is robust to

Figure 8: Different UIPs



Notes: We present marginal priors corresponding to F-UIP and priors corresponding to PM-UIP. The expectations are approximated using Monte Carlo with 10,000 replications in place of the observed data setting the parameters equal to their posterior means. For both the F-UIP and PM-UIP we report the marginal prior.

the prior (and other assumptions in the model that are often made for simplification). The extent of misspecification is reduced to a single parameter, the learning rate α. We showcase the new techniques in a Monte Carlo experiment as well as data for large U.S. banks. Potential misspecification, and the effect of UIP are discussed in detail. In terms of future research, it would be interesting to examine other techniques for mitigating misspecification, for example sub-sampling instead of the bootstrap in bagging (Politis and Romano, 1992, 1994). However, sub-sampling depends critically on the size of sub-samples which is, often, a difficult choice to make. Another interesting avenue for further research would be to propose more general data generating processes that are driven by more than one error component (ε_i). These data generating processes are likely to produce identification problems so, a resolution along Bayesian lines, would require crafting different "objective" priors

References

[1] Arreola, J. L. P., Johnson, A. L., Chen, X. C., Morita, H., 2020. Estimating stochastic production frontiers: A one-stage multivariate semiparametric Bayesian concave regression method. European Journal of Operational Research 287, 699–711.

[2] Atkinson, S. E., Tsionas, M. G. (2016). Directional distance functions: Optimal endogenous directions. Journal of Econometrics, 190 (2), 301–314 .

[3] Barron, A. R., Cover, T. M., 1991. Minimum Complexity Density Estimation. IEEE Transactions on Information Theory, 37 (4): 1034–1054.

[4] Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C., 1998). Robust and Efficient Estimation by Minimising a Density Power Divergence. Biometrika 85 (3), 549–559.

[5] Bissiri, P.G., Holmes, C.C., Walker, S.G., 2016. A general framework for updating belief distributions. Journal of the Royal Statistical Society Series B 78 (5), 1103–1130.

[6] Chambers, R. G. , Chung, Y. H., Färe, R., 1996. Benefit and distance functions. Journal of Economic Theory, 70 (2), 407–419.

[7] Chambers, R. G., Chung, Y. H., Färe, R., 1998. Profit, directional distance function, and Nerlovian efficiency. Journal of Optimization Theory and Applications, 98 (2), 351–364.

[8] Daraio, C., Simar, L., 2016. Efficiency and benchmarking with directional dis- tances: A data-driven approach. Journal of the Operational Research Society, 67 (7), 928–944.

[9] Dobmann, M., Liepelt, M., Schittkowski, K., Trassl, C., 1996. PCOMP: A FORTRAN Code for Automatic Differentiation. Unpublished Manuscript. https://www.schittkowski.de/downloads/numerical/PCOMPDOC.pdf

[10] Ghosh, A., Basu, A., 2016. Robust Bayes estimation using the density power divergence. Annals of the Institute of Statistical Mathematics 68, 413–437.

[11] Girolami, M., Calderhead, B., 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. Journal of the Royal Statistical Society Series B 73 (2), 123–214.

[12] Färe, R., Grosskopf, S. , Whittaker, G., 2013. Directional output distance functions: Endogenous directions based on exogenous normalization constraints. Journal of Productivity Analysis, 40 (3), 267–269.

[13] Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Clarendon Press, Oxford, UK, 169–193.

[14] Geweke, J., 1999. Using Simulation Methods for Bayesian Econometric Models. Computing in Economics and Finance 1999 832, Society for Computational Economics.

[15] Grünwald, P., van Ommen, T., 2017. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. Bayesian Analysis 12 (4), 1069–1103.

[16] Huggins, J.H, Miller, J.W., 2019. Robust inference and model criticism using bagged posteriors. arXiv 1912.07104.

[17] Jiang, W. Tanner, M.A., 2008. Gibbs posterior for variable selection in high dimensional classification and data mining. The Annals of Statistics 36, 2207–2231.

[18] Kass, R. E., Wasserman, L., 1995. A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. Journal of the American Statistical Association 90, 928–934.

[19] Koop, G., Pesaran, M.H., Smith, R.P., 2013. On Identification of Bayesian DSGE Models. Journal of Business & Economic Statistics 31 (3), 300–314.

[20] Kuosmanen, T., 2006. Stochastic nonparametric envelopment of data: Combining virtues of SFA and DEA in a unified framework. MTT Discussion paper no. 3/2006, MTT.

[21] Kuosmanen, T., 2008. Representation theorem for convex nonparametric least squares. Econometrics Journal, 11 , 308–325.

[22] Kuosmanen, T., Johnson, A. L., 2010. Data envelopment analysis as nonparametric least squares regression. Operations Research, 58 (1), 149–160

[23] Kuosmanen, T., Johnson, A., 2017. Modeling joint production of multiple outputs in StoNED: Directional distance function approach. European Journal of Operational Research 262, 792–801.

[24] Kuosmanen, T., Kortelainen, M., 2012. Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints. Journal of Productivity Analysis, 38 (1), 11–28.

[25] Kuosmanen, T., Johnson, A. L., Saastamoinen, A., 2015. Stochastic nonparametric approach to efficiency analysis: A unified framework. In J. Zhu (Ed.), Data envelopment analysis . Springer.

[26] Kuosmanen, T., Post, G. T., Scholtes, S., 2007. Testing for productive efficiency in case of errors-in-variables. Journal of Econometrics, 136 , 131–162.

[27] Layer, K, A.L. Johnson, R.C. Sickles, and G.D. Ferrier (2020) Direction selection in stochastic directional distance functions, European Journal of Operational Research 280(1), 351-364.

[28] Lee, C-Y., Johnson, A.L., Centeno, E.M., Kuosmanen, T., 2013. A more efficient algorithm for Convex Nonparametric Least Squares. European Journal of Operational Research 227 (2), 391–400.

[29] Lyddon, S. P., Holmes, C. C., Walker, S. G., 2019. General Bayesian updating and the loss-likelihood bootstrap. Biometrika 106 (2), 465–478.

[30] Malikov, E., Kumbhakar, S. C., Tsionas, M. G., 2016. A cost system approach to the stochastic directional technology distance function with undesirable outputs: the case of U.S. banks in 2001-2010. Journal of Applied Econometrics, 31 (7), 1407–1429.

[31] McAllester, D., 2003. PAC-Bayesian Stochastic Model Selection. Machine Learning, 51 (1), 5–21.

[32] Miller, J. W., Dunson, D., B., 2019. Robust Bayesian Inference via Coarsening. Journal of the American Statistical Association Volume 114 (527), 1113–1125.

[33] Müller, U. K., 2013. Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. Econometrica 81, 1805–49.

[34] Pincus, M., 1970. Letter to the Editor—A Monte Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems. Operations Research 18 (6), 1225–1228.

[35] Poirier, D. J.,1998. Revising Beliefs in Nonidentified Models. Econometric Theory 14, 483–509.

[36] Politis, D., Romano, J. 1992. A general theory for large sample confidence regions based on subsamples under minimal assumptions. Technical Report 399, Dept of Statistics, Stanford University.

[37] Politis, D., Romano, J., 1994. Large sample confidence regions based on subsamples under minimal assumptions. Annals of Statistics, 22, 203–2050.

[38] Yao, L.,Jiang, W., Tanner, M.A., 2011. Predicting panel data binary choice with the Gibbs posterior. Neural Computation 23, 2683–2712.

[39] Walker, S., Hjort, N. L., 2002. On Bayesian consistency. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(4): 811–821.

[40] Vovk, V. G., 1990. Aggregating strategies. In Proc. COLT' 90, 371–383.

[41] Zellner, A., 1986. Bayesian estimation and prediction using asymmetric loss functions. Journal of the American Statistical Association 81, 446–451.

[42] Zhang, T., 2006a. From ε-entropy to KL entropy: Analysis of minimum information complexity density estimation." Annals of Statistics, 34 (5), 2180–2210.

[43] Zhang, T., 2006b. Information Theoretical Upper and Lower Bounds for Statistical Estimation. IEEE Transactions on Information Theory, 52 (4), 1307–1321.

[44] Zofio, J. L., Pastor, J. T., Aparicio, J., 2013. The directional profit efficiency measure: On why profit efficiency is either technical or allocative. Journal of Productivity Analysis, 40 (3), 257–266.

c:\loss functions Bayes