European Journal of Operational Research 000 (2017) 1–19

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

Dimension reduction in nonparametric models of production☆

Paul W. Wilson

Department of Economics and Division of Computer Science, School of Computing, Clemson University, Clemson, SC 29634-1309, USA

ARTICLE INFO

Article history:
Received 6 February 2017
Accepted 10 November 2017
Available online xxx

Keywords:
Dimensionality
Dimension reduction
DEA
FDH
Efficiency

ABSTRACT

It is well-known that the convergence rates of nonparametric efficiency estimators (e.g., free-disposal hull and data envelopment analysis estimators) become slower with increasing numbers of input and output quantities (i.e., dimensionality). Dimension reduction is often utilized in non-parametric density and regression where similar problems occur, but has been used in only a few instances in the context of efficiency estimation. This paper explains why the problem occurs in nonparametric models of production and proposes three diagnostics for when dimension reduction might lead to more accurate estimation of efficiency. Simulation results provide additional insight, and suggest that in many cases dimension reduction is advantageous in terms of reducing estimation error. The simulation results also suggest that when dimensionality is reduced, free-disposal hull estimators become an attractive, viable alternative to the more frequently used (and more restrictive) data envelopment analysis estimators. In the context of efficiency estimation, these results provide the first quantification of the tradeoff between information lost versus improvement in estimation error due to dimension reduction. Results from several papers in the literature are revisited to show what might be gained from reducing dimensionality and how interpretations might differ.

1. Introduction

Data envelopment analysis (DEA) and free-disposal hull (FDH) estimators are widely used to estimate technical efficiency, changes in productivity, returns to scale and other performance benchmarks. Farrell (1957) first introduced the DEA estimator which were subsequently popularized by Charnes, Cooper, and Rhodes (1978), while the FDH estimator was introduced by Deprins, Simar, and Tulkens (1984). The statistical properties of the estimators are developed in a number of papers; see the recent surveys by Simar and Wilson (2013, 2015) for details and discussion.

Practitioners using FDH or DEA estimators have long been aware that increasing the number of inputs or outputs causes (i) firms to appear increasingly efficient and (ii) increasing numbers of firms to lie on the estimated frontier. Because of this, a number of ad hoc "rules of thumb" for lower bounds on the sample size n in problems with p inputs and q outputs are proposed in the literature. Bowlin (1987, pp. 128–129), Golany and Roll (1989, p. 239), Vassiloglou and Giokas (1990, p. 592), and Homburg (2001, p. 56), propose $n \geq 2(p+q)$, while Banker, Charnes, Cooper, Swarts, and Thomas (1989, pp. 138–139), Bowlin (1998, p. 18), Friedman and Sinuany-Stern (1998, p. 783), and Raab and Lichty (2002, p. 589), propose $n \geq 3(p+q)$. Paradi, Vela, and Yang (2004, p. 359), split the difference by stating that n should be at least two to three times the total number of inputs plus outputs. Boussofiane, Dyson, and Thanassoulis (1991, p. 4), suggest $n \geq pq$, and Dyson et al. (2001, p. 148), suggest $n \geq 2pq$. Cooper, Seiford, and Tone (2000, pp. 106 and 272), Cooper, Li, Seiford, and Zhu (2004, p. 77), and Banker, Emrouznejad, Lopes, and de Almeida (2012, p. 231), suggest $n \geq \max(pq,\ 3(p+q))$. No theoretical justification is given for any of these rules, and as will be seen below, the suggested minimum sample sizes are too small to allow one to obtain meaningful estimates of technical efficiency.

In many cases, production data have substantial multicollinearity. This paper shows how multicollinearity among inputs or outputs can be exploited to reduce the dimensionality of nonparametric production models, resulting in (presumably) more accurate estimates of efficiency. While the methodology is not new, there are to date only a handful of instances where it has been adopted by practitioners working in the field of nonparametric efficiency estimation. Moreover, until now there has been almost no evidence on the specific nature of the tradeoff between information sacrificed when reducing dimensionality versus any

☆ Previous versions of this paper was presented at the 9th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2016) at the Higher Technical School of Engineering, University of Seville, Spain in December 2016 and at the 15th European Workshop on Efficiency and Productivity in London in June 2017. I am grateful to the Cyber Infrastructure Technology Integration group at Clemson University for operating the Palmetto cluster used for the computations in this paper.

E-mail address: pww@clemson.edu

https://doi.org/10.1016/j.ejor.2017.11.020
0377-2217/© 2017 Elsevier B.V. All rights reserved.

gains in terms of reduced estimation error that might result, nor are there useful guidelines for when dimension-reduction should be employed. Simulation results provided below show that substantial improvements in the accuracy of efficiency estimates are possible with dimension reduction and further suggest that these improvements are likely possible in many applications. Three diagnostics are provided to give the applied researcher insight regarding whether dimension reduction might be useful. The problems of dimensionality, as well as the diagnostics and the potential for useful dimension reduction are further illustrated with real data from several papers in the published literature.

The inverse relationship between dimensionality and convergence rates of estimators is well-known in nonparametric statistics and econometrics, and is often called the "curse of dimensionality" after Bellman (1957). Silverman (1986) and Scott (1992) discuss the problem in the context of nonparametric density estimation, while Härdle (1990), Pagan and Ullah (1999), Henderson and Parmeter (2015), Persson, Häggström, Waernbaum, and de Luna (2017) and Rekabdarkolaee, Boone, and Wang (2017) examine the problem in the context of regression. By now it is also well-known that the same curse of dimensionality affects the convergence rates of nonparametric DEA and FDH estimators used in the analysis of productive efficiency; see Simar and Wilson (2013, 2015) and the papers cited therein for details. It is perhaps less-well appreciated, but nonetheless true that dimensionality also affects partial frontier estimators such as the order-m estimators developed by Cazals, Florens, and Simar (2002), Wilson (2011) and Simar and Vanhems (2012) and the order-α estimators developed by Aragon, Daouia, and Thomas-Agnan (2005), Daouia and Simar (2007), Wheelock and Wilson (2008) and Simar and Vanhems (2012) (see Simar and Wilson, 2013, 2015, for recent surveys of order-m and order-α estimators). In the case of the order-m and order-α estimators, the convergence rate is not affected by the number of dimensions, but mean-square error increases with the number of dimensions (i.e., the number of inputs and outputs) for a given sample size. In extreme cases, all observations in a given sample may lie on the estimated frontier when full-frontier DEA or FDH estimators are used, or all observations may lie *above* the estimated partial frontier when order-m or order-α estimators are used, even if m is very large (but finite) or α is very close (but not equal) to one.

In the density and regression contexts, dimension-reduction methods are often used to mitigate the effects of dimensionality. Scott (1992) discusses several approaches in the density estimation context. In the regression context, Wheelock and Wilson (2001), Wheelock and Wilson (2012) and Wilson and Carey (2004) use eigensystem techniques to reduce dimensionality. In the context of nonparametric frontier estimation, eigensystem techniques for dimension reduction are proposed by Adler and Golany (2001, 2007), Mouchart and Simar (2002) and Daraio and Simar (2007, pp. 148–150) but have seen little use in the empirical literature.[1]

As discussed below in Section 2, the curse of dimensionality is a serious problem in nonparametric efficiency estimation, and there are numerous examples in the literature where it has been ignored. Where the problem exists—and where it is ignored—dubious results may be obtained. Section 2.1 presents a nonparametric model of production, and the nonparametric FDH and DEA efficiency estimators and their properties, including convergence rates, are presented and briefly discussed in Section 2.2. Section 2.3 gives some additional insight on why the curse of dimensionality arises. Section 3 discusses two methods for diagnosing when the curse of dimensionality is likely to substantially affect estimation error. The eigensystem methods proposed by Mouchart and Simar (2002) and Daraio and Simar (2007) are explained in Section 4. These methods, together with the simulation results presented later, provide a third diagnosis for the curse of dimensionality. Monte Carlo results are presented in Section 5, and some real-world examples are given in Section 6 to illustrate both the need for dimension-reduction and how results can change when dimensionality is reduced. Conclusions are discussed in Section 7.

2. The problem of too many dimensions

2.1. A nonparametric model of production

Let $x,\ X \in \mathbb{R}^p_+$ denote vectors of p input quantities, and let $y,\ Y \in \mathbb{R}^q_+$ denote vectors of q output quantities, with upper-case letters denoting random variables and lower-case denoting non-stochastic values. The production set is given by

$$\Psi := \{(x, y) \mid x \text{ can produce } y\}. \tag{1}$$

Various measures of technical efficiency are available; e.g., consider the Farrell (1957) input-oriented measure

$$\theta(x, y) := \inf\{\theta \mid (\theta x, y) \in \Psi, \quad \theta > 0\}. \tag{2}$$

Clearly, for $(x, y) \in \Psi$, $\theta(x, y) \le 1$. To conserve space, the presentation that follows is only in terms of the input orientation; all of the qualitative results obtained herein hold when efficiency is measured in other directions.

Of course, Ψ and hence $\theta(x, y)$ are unknown and must be estimated from a sample of observations $\mathcal{S}_n = \{(X_i, Y_i)\}_{i=1}^n$. But before anything can be estimated, and certainly before one can make inference, a statistical model must be specified. The following assumptions specify a statistical model similar to the one defined by Kneip, Simar, and Wilson (2008).

Assumption 2.1. Ψ is (a) closed and (b) convex.

Assumption 2.2. If $x = 0$, $y \ge 0$, $y \ne \mathbf{0}$, then $(x, y) \notin \Psi$, *i.e.*, all production requires use of some inputs.

Assumption 2.3. For $\widetilde{x} \ge x$, $\widetilde{y} \le y$, if $(x, y) \in \Psi$ then $(\widetilde{x}, y) \in \Psi$ and $(x, \widetilde{y}) \in \Psi$, i.e., both inputs and outputs are strongly disposable.[2]

Assumption 2.4. The n observations in $\mathcal{S}_n$ are identically, independently distributed (iid) random variables on the attainable set Ψ.

Assumption 2.5. (a) The (X, Y) possess a joint density f with support $\mathcal{D} \subseteq \Psi$; (b) f is continuous on $\mathcal{D}$; and (c) $f(\theta(x, y)x, y) > 0$ for all (x, y) in the interior of $\mathcal{D}$.

Assumption 2.6. The function $\theta(x, y)$ is twice continuously differentiable for all $(x, y) \in \mathcal{D}$.

Assumptions 2.1–2.3 are standard in the economic literature (e.g., see Shephard, 1970 or Färe, 1988). Assumption 2.1 ensures that the frontier

$$\Psi^\partial = \left\{(x, y) \mid (x, y) \in \Psi, \quad (\gamma^{-1}x, \gamma y) \notin \Psi \quad \forall \quad \gamma \in (1, \infty)\right\} \tag{3}$$

is included in Ψ, i.e., $\Psi^\partial \subset \Psi$. As noted below, DEA estimators require convexity of Ψ, but FDH estimators do not (the

[1] Adler and Golany (2001, 2007) decompose correlation matrices of either inputs or outputs (but not both), whereas Mouchart and Simar (2002) and Daraio and Simar (2007) decompose (raw) moment matrices of both inputs and outputs. Florens, Simar, and Van Keilegom (2014) and Daraio, Simar, and Wilson (2017) use these methods in applications, but their applications are only empirical illustrations in papers that otherwise are focused on statistical theory and methodology. Mouchart and Simar (2002) and Adler, Liebert, and Yazhemsky (2013) provide examples where dimension-reduction methods are used in studies driven by their empirical applications.

[2] Per convention, inequalities involving vectors are defined on an element-by-element basis.

Please cite this article as: P.W. Wilson, Dimension reduction in nonparametric models of production, European Journal of Operational Research (2017), https://doi.org/10.1016/j.ejor.2017.11.020