Nonparametric Estimation of Homothetic and Homothetically Separable Functions

Arthur Lewbel
Boston College

Oliver Linton
LSE

National University of Singapore
March, 2005

Outline

1. Statement of the Problem

2. Literature Review

3. Matching Idea

4. A more general approach

5. Discussion on Distribution Theory

6. Simulations and Applications

7. Extensions

The Problem

- A given function $r(x,w)$ has the following structure. There exist functions h and g such that

  $$r(x,w) = h[g(x),w],$$

  where
  - g is linearly homogeneous, i.e., $g(cx) = cg(x)$ for $c \in \mathbb{R}$
  - h is strictly monotonic in g, i.e., $\partial h/\partial g > 0$.

- Homotheticity:

  $$r(x) = h[g(x)]$$

- Linear homogeneity vs any other nonzero known degree (or any known monotonic transformation of g) is WLOG.

- Goal: consistent and asymptotically normal estimator of h and g based on some estimator $\widehat{r}(x,w)$ of $r(x,z)$ when h, g are unknown but continuous/smooth.

Literature Review

- Homothetic and homothetically separable functions are common in models of consumer preferences and firm production.

- $r(x, w)$ could be a utility, cost function or production function, either directly estimated or recovered from consumer or factor demand equations.

- Examples: Blackorby, Primont, and Russell (1978), Chiang (1984), Zellner and Ryu (1998), Matzkin (1994). Zellner and Revankar (1969)

$$Ye^{\theta Y} = AK^{\alpha(1-\delta)}L^{\alpha\delta}$$

- Linear index models like standard censored, truncated, or discrete response models are homothetic functions, with $g(x) = x^\top \beta$. Replacing $x^\top \beta$ with an arbitrary linearly homogeneous function $g(x)$ is a natural generalization for contexts like price indices or constant returns to scale technologies.

Other Homogeneity related estimators

- Matzkin (1992) consistent estimator for

  $$y = I[g(x) + \varepsilon \geq 0],$$

  $g(x)$ homogeneous, $\varepsilon \perp\!\!\!\perp x$. Newey and Matzkin (1993) similar to ours, no w, more steps, incomplete.

- Matzkin (2003)

  $$y = m(x, \varepsilon)$$

  with $\varepsilon \perp\!\!\!\perp x$ and e.g., m homogeneous in x, ε.

- Nonparametric homogeneous functions: Matzkin (1992), Tripathi and Kim (2001).

- Yatchew and Bos (1997) consider estimating some homothetic demand models by nonparametric least squares.

Other Separability related estimators

- Weak separability: $r(x,w) = h[g(x),w]$ without g homogeneous. Gorman (1959), Goldman and Uzawa (1964), Blackorby, Primont, and Russell (1978). Pinkse (2001) estimates g up to monotonic transformation.
- Strong or additive separability:

  $$r(x,w) = g(x) + t(w).$$

  Härdle, Kim, and Tripathi (2001), Friedman and Stutzle (1981), Breiman and Friedman (1985), Andrews (1991), Tjøstheim and Auestad (1994), Linton and Nielsen (1995), Stone (1986).
- Generalized additive separability:

  $$r(x,w) = H\left(g(x) + t(w)\right)$$

  for some known or unknown H. Hastie and Tibshirani (1990), Linton and Härdle (1996), Horowitz (2001).

Identification and Estimation

- Recall that we have an estimate of r, where

$$r(x, w) = h(g(x), w)$$

- Identification issue, must restrict either h or g further.
  - One possibility is

$$g(x_0) = 1$$

    which was adopted in previous version of paper.
  - Instead we assume that for some weighting function $\omega(x)$

$$\int g(x)\omega(x)dx = 1.$$

    This normalization has advantages in terms of the distribution theory etc.
- The main issue is estimation of g; once we have g, one can estimate h by generated nonparametric regression.

Matching Idea

- For a given x, w, find $u_{xx'w}$ such that
$$r\,(x,w) = r(u_{xx'w}x', w),$$
a match.
- Then by monotonicity of h we obtain
$$g(x) = u_{xx'w}g(x') \Longrightarrow \; u_{xx'w} = \frac{g(x)}{g(x')}.$$
- Under our current normalization we have
$$g(x') = \frac{1}{\int u_{xx'w}\omega(x)dx}.$$
- Also,
$$g(x) = \frac{\int u_{xx'w}\omega(x')dx'}{\int\int u_{xx'w}\omega(x')\omega(x)dx'dx}.$$

A More General Approach
Polar Coordinates

- Write x in polar coordinates as ρ, θ, where ρ is length and θ is direction, so θ contains the same information as $x/||x||$.
- Define the functions R and G that are just the functions r and g expressed in polar coordinates

$$R(\rho, \theta, w) = r(x, w) \text{ and } G(\theta) = g(x/||x||),$$

- Any function G automatically corresponds to a homogeneous function g, defined by

$$g(x) = \rho G(\theta) \Longrightarrow R(\rho, \theta, w) = h(\rho G(\theta), w)$$

- To identify $G_0(\theta)$ we shall assume that

$$EG_0(\theta) = \int G_0(\theta) f_\theta(\theta) d\theta = 1,$$

where $f_\theta(\theta)$ is the marginal density of θ.

- Define $U_0(\theta, \theta')$ as the value that solves

$$m_1(U; \theta, \theta') = 0,$$

where

$$m_1(U; \theta, \theta') = \int [R(\rho, \theta, w) - R(U\rho, \theta', w)]\pi(d\rho, dw \mid \theta, \theta')$$

for each θ, θ' for a given non-negative measure $\pi(d\rho, dw \mid \theta, \theta')$ that has support contained in $\Psi_{\rho,w|\theta,\theta'}$. The two leading cases here would be:

– $\pi(d\rho, dw \mid \theta, \theta') = \pi(\rho, w \mid \theta, \theta')d\rho dw$ for some conditional density function $\pi(\rho, w \mid \theta, \theta')$ with non-trivial support;

– $\pi(d\rho, dw \mid \theta, \theta')$ represents a point mass at points $\rho_0(\theta, \theta'), w_0(\theta, \theta')$.

- We concentrate on case (a) in our theoretical analysis, because this type of averaging can yield improved rates of convergence, see inter alia Linton and Nielsen (1995). On the other hand, case (b) can have some computational advantages.

- For simplicity we take $\pi(d\rho, dw \mid \theta, \theta')$ not to depend on θ, θ' and in particular,

$$\pi(\rho, w) = \begin{cases} f_{\rho,w}(\rho, w) & \text{if } (\rho, w) \in A \\ 0 & \text{else} \end{cases}$$

for some fixed set $A \subset \cap_{\theta,\theta' \in \Psi_\theta \times \Psi^*_\theta} \Psi_{\rho,w|\theta,\theta'}$, so the set A does not vary with θ, θ'. By collapsing A to a single point we would obtain case (b) above.

- We compute the sample moment function

$$\widehat{m}_1(U;\theta,\theta') = \frac{1}{n}\sum_{i=1}^{n}[\widehat{R}(\rho_i,\theta,W_i) - \widehat{R}(U\rho_i,\theta',W_i)]1((\rho_i,W_i) \in A),$$

Define the estimator $\widehat{U}(\theta,\theta')$ for each θ,θ' to be any value such that

$$|\widehat{m}_1(\widehat{U}(\theta,\theta');\theta,\theta')| \leq \inf_u |\widehat{m}_1(u;\theta,\theta')| + o_p(n^{-1/2}).$$

- Under our normalization, it follows that

$$G_0(\theta) = \frac{\int U_0(\theta, \theta')\varpi(d\theta')}{\int U_0(\theta, \theta')\varpi(d\theta')f_\theta(\theta)d\theta},$$

  where $\varpi(d\theta')$ is an arbitrary measure with support in Ψ_θ. Specifically, $\varpi(d\theta')d\theta'$ could be:
  – the point mass at some point θ_0; or

  – $\varpi(d\theta') = \varpi(\theta')d\theta'$ with ϖ a density function on some non-trivial interval.

- In our theoretical work we focus on the latter case, but the former case has some computational advantages.

- We then estimate $G_0(\theta)$ by

$$\widehat{G}(\theta) = \frac{\frac{1}{n}\sum_{i=1}^n \widehat{U}(\theta, \theta_i)\varpi_f(\theta_i)}{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \widehat{U}(\theta_i, \theta_j)\varpi_f(\theta_j)},$$

  where $\varpi_f(\theta)$ is a weighting function such that

$$E[g(\theta)\varpi_f(\theta)] = \int g(\theta)\varpi(\theta)d\theta$$

  for any measurable function g.

- The choice of ϖ_f/ϖ is arbitrary, but it is related to the set A chosen earlier.

Estimate of h

- Let $\gamma = \rho G_0(\theta)$, $c = (\gamma, w)$, $\widehat{\gamma} = \rho \widehat{G}(\theta)$, $\widehat{\gamma}_i = \rho_i \widehat{G}(\theta_i)$, and let $\widehat{c} = (\widehat{\gamma}, w)$ and $\widehat{C}_i = (\widehat{\gamma}_i, W_i)$. Define the sample moment function

$$\widehat{m}_3(h; c) = \frac{1}{n b_*^{d_W+1}} \sum_{i=1}^{n} \mathcal{K}_{b_*}\left(\frac{c - \widehat{C}_i}{b_*}\right) \psi\left(\widehat{R}_i - h\right),$$

  where $\mathcal{K}_{b_*}(.) = \mathcal{K}(./b_*)/b_*^{d_W+1}$ is a $d_W + 1$-dimensional kernel and b_* is some bandwidth sequence. Here, ψ is a smooth function with $\psi(0) = 0$ and $\psi'(0) \neq 0$. If Y_i is observed and $R_i = E[Y_i|\rho_i, \theta, W_i]$, then can replace $\widehat{R}_i$ by Y_i.

- Define the estimator of $h(c)$ to be any sequence $\widehat{h}(c)$ of approximate zeros i.e.,

$$|\widehat{m}_3(\widehat{h}(c); c)| \leq \inf_{h \in \mathcal{H}} |\widehat{m}_3(h; c)| + o_p(n^{-1/2})$$

- Monotonicity of $\widehat{h}$ can be imposed by applying the pool-adjacent-violators algorithm to the estimated function $\widehat{h}(.)$.

- We also define an estimator of $\partial h(c)/\partial\gamma$ by differentiating $\widehat{h}(c)$ with respect to γ when this is permissible (i.e., when ψ is differentiable) and denoting this by $\widehat{\partial h}(c)/\partial\gamma$. Alternatively, one can use a local linear method replacing $\psi(\widehat{R}_i - h)$ by $\psi(\widehat{R}_i - h - h_\gamma^\top(\gamma_i - \gamma) - h_w^\top(W_i - w))$ and taking $\widehat{h}_\gamma$ as the estimate of $\partial h(c)/\partial\gamma$.

Simultaneous Estimation of h and G

- Our strategy for improving the efficiency of the estimators we defined above is based on using a more general definition of the functions $h_0(.)$ and $G_0(.)$. They can be defined as minimizers of the functional

$$E\left[R(\rho,\theta,W)-h(\rho G(\theta),W)\right]^2=\int\left[R(\rho,\theta,w)-h(\rho G(\theta),w)\right]^2 f_Z(\rho,\theta,w)d\rho d\theta dw$$

subject to the restriction that

$$\int G(\theta)f_\theta(\theta)d\theta=1,$$

where $f_Z(\rho,\theta,w)$ is the joint density of the random variables (ρ,θ,W).

- To find a characterization of the solutions we follow Weinstock (1952, Chapter 4) in our treatment. Define the objective functional

$$\mathcal{L}(h,G,\lambda)=\int\left[R(\rho,\theta,w)-h\left(\rho G(\theta),w\right)\right]^2 f_{\rho,\theta,w}(\rho,\theta,w)d\rho d\theta dw+\lambda\left[\int G(\theta)f_\theta(\theta)d\theta-1\right]$$

for each h,G,λ. Letting $G(.)=G_0(.)+\epsilon\tau(.)$ and $h(.)=h_0(.)+\delta\eta(.)$ we find the following first order conditions:

$$\begin{aligned}
0 &= \left.\frac{\partial \mathcal{L}(h_0+\delta\eta, G_0+\epsilon\tau, \lambda)}{\partial \delta}\right\downarrow_{\epsilon=0,\delta=0} \\
&= -\int \left[R(\rho,\theta,w) - h_0\left(\rho G_0(\theta), w\right)\right] \eta\left(\rho G_0(\theta), w\right) f_{\rho,\theta,w}(\rho,\theta,w) d\rho d\theta dw \\
0 &= \left.\frac{\partial \mathcal{L}(h_0+\delta\eta, G_0+\epsilon\tau)}{\partial \epsilon}\right\downarrow_{\epsilon=0,\delta=0} \\
&= -\int \left[R(\rho,\theta,w) - h_0\left(\rho G_0(\theta), w\right)\right] \frac{\partial h_0}{\partial \gamma}\left(\rho G_0(\theta), w\right) \rho\tau(\theta) f_{\rho,\theta,w}(\rho,\theta,w) d\rho d\theta dw + \lambda\int \tau(\theta) f_\theta(\theta) d\theta,
\end{aligned}$$

which should hold for all test functions η, τ.

- By setting the directions to be the Dirac deltas [$\tau(\theta) = 1(\theta = t)$ and $\eta\left(\rho G_0(\theta), w\right) = 1\left(\rho G_0(\theta) = s, w = u\right)$] and using the law of iterated expectation we get the necessary condition

$$\mathcal{L}_h(h_0, G_0)(s,u) = -E\left[\{R(\rho,\theta,W) - h_0(\rho G_0(\theta), W)\} | \rho G_0(\theta) = s, W = u\right] f_{\rho G_0(\theta),w}(s,u) = 0$$

for the first equation, where $f_{\rho G_0(\theta),w}(s,u)$ is the density function of the random variable $\rho G_0(\theta), W$.

- For the second equation, we obtain the necessary condition that

$$E\left[\{R(\rho,\theta,W) - h_0\left(\rho G_0(\theta), W\right)\} \frac{\partial h_0}{\partial \gamma}\left(\rho G_0(\theta), W\right) \rho | \theta = t\right] f_\theta(t) = \lambda f_\theta(t)$$

for all t.

- Multiplying by $G_0(t)$ and integrating over t and using the law of iterated expectations we obtain

$$\lambda = E\left[\left[R(\rho,\theta,W) - h_0\left(\rho G_0(\theta), W\right)\right] \frac{\partial h_0}{\partial \gamma}\left(\rho G_0(\theta), W\right) \rho G_0(\theta) \right].$$

Then substituting back and dividing through by $f_\theta(t)$ we obtain the equation: for all t,

$$\begin{aligned} & E\left[\left[R(\rho,\theta,W) - h_0\left(\rho G_0(\theta), W\right)\right] \frac{\partial h_0}{\partial \gamma}\left(\rho G_0(\theta), W\right) \rho | \theta = t \right] \\ & -E\left[\left[R(\rho,\theta,W) - h_0\left(\rho G_0(\theta), W\right)\right] \frac{\partial h_0}{\partial \gamma}\left(\rho G_0(\theta), W\right) \rho G_0(\theta) \right] = 0. \end{aligned}$$

- Suppose that one has consistent estimators of G_0, h_0, and $\partial h_0/\partial\gamma$, denoted by $\widehat{G}$, $\widehat{h}$, and $\partial\widehat{h}/\partial g$ respectively. We suggest the following estimation method.
- Define the sample moment function

$$\widehat{m}_4(g;\theta) = \frac{1}{nb_0^{d_\theta}} \sum_{i=1}^{n} K\left(\frac{\theta-\theta_i}{b_0}\right)\left[\widehat{\zeta}_i(g) - \frac{1}{n}\sum_{i=1}^{n}\widehat{\zeta}_i(\widehat{G}(\theta_i))\widehat{G}(\theta_i)\right]$$

$$\widehat{\zeta}_i(g) = \left[\widehat{R}(\rho_i,\theta_i,W_i) - \widehat{h}\left(\rho_i g, W_i\right)\right]\frac{\partial\widehat{h}}{\partial\gamma}\left(\rho_{_i}\widehat{G}(\theta_i),W_i\right)\rho_i,$$

where K is kernel and b_0 is a bandwidth.
- The bandwidth b_0 does not play a big role in the sequel and we shall assume as above that it is small. Then define the estimator $\widetilde{G}(\theta)$ for each θ to be any value such that

$$|\widehat{m}_4(\widetilde{G}(\theta);\theta)| \le \inf_{g\in\mathcal{G}} |\widehat{m}_4(g;\theta)| + o_p(n^{-1/2}),$$

where the set $\mathcal{G}$ can be chosen to be a small neighborhood of $\widehat{G}(\theta)$.

- It may be more convenient to avoid this optimization altogether and use a ‘two-step’ estimator

$$\widetilde{G}^{2-step}(\theta) = \widehat{G}(\theta) - \left[\frac{\partial \widehat{m}_4}{\partial g}(\widehat{G}(\theta);\theta)\right] \widehat{m}_4(\widehat{G}(\theta);\theta), \text{ where}$$

$$\frac{\partial \widehat{m}_4}{\partial g}(\widehat{G}(\theta);\theta) = \frac{1}{nb_0^{d_\theta}} \sum_{i=1}^{n} K\left(\frac{\theta - \theta_i}{b_0}\right) \left[\frac{\partial \widehat{h}}{\partial \gamma}\left(\rho_i \widehat{G}(\theta_i), W_i\right)\right]^2 \rho_i^2.$$

  Following Fan and Chen (1997) we expect $\widetilde{G}^{2-step}(\theta)$ to be asymptotically equivalent to $\widetilde{G}(\theta)$.

- The estimated function $\widetilde{G}(\theta)$ does not satisfy exactly the empirical restriction $n^{-1}\sum_{i=1}^{n}\widetilde{G}(\theta_i) = 1$ so we further replace $\widetilde{G}(\theta)$ by $\widetilde{G}(\theta)/n^{-1}\sum_{i=1}^{n}\widetilde{G}(\theta_i)$.

- Then compute $\widetilde{h}(.)$ as the nonparametric regression of $\widehat{R}_i$ on $\rho_i\widetilde{G}(\theta_i)$. We can compute an estimator of $\partial h(.)/\partial\gamma$ either by differentiating $\widetilde{h}(.)$ with respect to γ or by using local polynomials and taking the corresponding slope coefficient.

Distribution Theory

$$I_1(\theta,\theta') = G_0(\theta')\int \frac{\partial h}{\partial \gamma}(\rho G_0(\theta), w)\rho\pi(\rho,w)d\rho dw.$$

$$V_1(\theta,\theta') = \vartheta_V \int \frac{\sigma^2(\rho,\theta,w)}{f_Z(\rho,\theta,w)}\pi^2(\rho,w)d\rho dw + \vartheta_V \frac{1}{U_0(\theta,\theta')}\int \frac{\sigma^2(U_0(\theta,\theta')\rho,\theta',w)}{f_Z(U_0(\theta,\theta')\rho,\theta',w)}\pi^2(\rho,w)d\rho dw.$$

$$\beta_U(\theta,\theta') = I_1^{-1}(\theta,\theta')\int \left[\beta_R(\rho,\theta,w) - \beta_R(U_0(\theta,\theta')\rho,\theta',w)\right]\pi(\rho,w)d\rho dw.$$

THEOREM 2. *Suppose that assumptions A and B hold, and let* $\Omega_1(\theta,\theta') = I_1^{-2}(\theta,\theta')V_1(\theta,\theta')$. *Then, as* $n \to \infty$

$$\sqrt{nb^{d_\theta}}\left(\widehat{U}(\theta,\theta') - U_0(\theta,\theta') - b^p\beta_U(\theta,\theta')\right) \Longrightarrow N(0,\Omega_1(\theta,\theta')).$$

- The asymptotic bias and variance both contain two terms due to the dependence of $\widehat{U}(\theta,\theta')$ on $\widehat{R}$ at two points. The quantity $I_1(\theta,\theta')$ is sort of an information, and is guaranteed to be positive when the support of $\pi(\rho,w)$ contains only non-negative ρ.

- Let

$$I_2(\theta) = \int \frac{\varpi(\theta')}{G_0(\theta')} d\theta'$$

$$V_2(\theta) = \vartheta_V \int \frac{\pi^2(\rho, w)\sigma^2(\rho, \theta, w)}{f_Z(\rho, \theta, w)} d\rho dw \left(\int \frac{\varpi(\theta')}{I_1(\theta, \theta')} d\theta' \right)^2$$

$$\beta_G(\theta) = I_2^{-1}(\theta) \left[\int \beta_U(\theta, \theta')\varpi(\theta')d\theta' - G_0(\theta) \int \beta_U(\theta, \theta')\varpi(\theta')f_\theta(\theta)d\theta' d\theta \right].$$

THEOREM 3. *Suppose that assumptions A and B hold, and let* $\Omega_2(\theta) = I_2^{-2}(\theta)V_2(\theta)$. *Then, as* $n \to \infty$

$$\sqrt{nb^{d_\theta}} \left(\widehat{G}(\theta) - G(\theta) - b^p \beta_G(\theta) \right) \Longrightarrow N(0, \Omega_2(\theta)).$$

- Note that $\widehat{G}(\theta)$ converges to $G_0(\theta)$ at the same rate as $\widehat{U}(\theta, \theta')$ converges to $U_0(\theta, \theta')$. Both estimators behave like d_θ-dimensional smoothers. Although the asymptotic variance of $\widehat{U}(\theta, \theta')$ contains two terms, the asymptotic variance of $\widehat{G}(\theta)$ contains only one term because the integration wipes out the second term.

- Since $\int(\varpi(\theta')/I_1(\theta,\theta'))d\theta' = I_2(\theta)/\int \frac{\partial h}{\partial\gamma}(\rho G_0(\theta),w)\rho\pi(\rho,w)d\rho dw$, the variance constant $I_2^{-2}(\theta)V_2(\theta)$ simplifies to

$$\Omega_2(\theta) = \frac{\vartheta_V \int \frac{\pi^2(\rho,w)\sigma^2(\rho,\theta,w)}{f_Z(\rho,\theta,w)}d\rho dw}{\left(\int \frac{\partial h}{\partial\gamma}(\rho G_0(\theta),w)\rho\pi(\rho,w)d\rho dw\right)^2},$$

  which does not depend on the weighting function ϖ. The asymptotic variance $\Omega_2(\theta)$ reflects the way the integration was done through the choice of π and A. In the special case of homoskedasticity $\sigma^2(\rho,\theta,w) = \sigma^2$ and independence $f_Z(\rho,\theta,w) = f_{\rho,w}(\rho,w)f_\theta(\theta)$, the numerator of $\Omega_2(\theta)$ is $\vartheta_V\sigma^2(\int_A f_{\rho,w}(\rho,w)d\rho dw)/f_\theta(\theta)$.

- The asymptotic bias is affected by the weighting function ϖ and is basically a weighted average of biases of the estimator $\widehat{R}$ along two different rays θ,θ' : if $\beta_R(z) = 0$ for all z, then $\beta_G(\theta) = 0$ for all θ.

- Suppose that the bandwidth sequence b_*, b satisfies

$$b_* = \lambda_* n^{-1/(2p+d_W+1)} \text{ and } b = \lambda n^{-1/(2p+d_\theta)}$$

  for some λ_*, λ with $0 < \lambda_* < \infty$.

- Then define

$$\begin{aligned}
I_3(c) &= \psi'(0) f_C(c), \\
V_{3G}(c) &= \lambda^{-d_\theta} \left[\frac{\partial h}{\partial \gamma}(c)\rho\right]^2 \Omega_2(\theta) \\
V_{3h}(c) &= \lambda_*^{-(d_W+1)} e_V \psi'(0) f_C(c) E\left[\sigma^2(Z)|C=c\right] \\
\Omega_3(c) &= V_{3G}(c)1(d_W+1 \geq d_\theta) + I_3^{-2}(c)V_{3h}(c)1(d_W+1 \leq d_\theta).
\end{aligned}$$

$$\begin{aligned}
\beta_{3G}(c) &= \gamma\psi'(0)\frac{\partial h}{\partial \gamma}(\gamma, w)\int \frac{\beta_G(\theta)}{G_0(\theta)} f_{\gamma,\theta,W}(\gamma, \theta, w)d\theta \\
\beta_{3h}(c) &= I_3^{-1}(c)\frac{1}{p!}\int k(t)t^p dt \sum_{j=1}^{d_W+1} \frac{\partial^p}{\partial u_j^p}\left[\psi\left(h(c+u) - h(c)\right) f_C(c+u)\right]_{u=0}.
\end{aligned}$$

THEOREM 4. *Suppose that assumptions A, B, and C hold. Then, as $n \to \infty$,*

$$n^{\min\{p/(2p+d_W+1),p/(2p+d_\theta)\}}\left(\widehat{h}(\widehat{c}) - h(c) - b^p\beta_{3G}(c) - b_*^p\beta_{3h}(c)\right) \Longrightarrow N(0,\Omega_3(c)).$$

- The variance contains two terms
  - The term $V_{3h}(c)$ can be recognized as the covariate dependent part of the asymptotic variance that would result were G_0 known, i.e., if $\widehat{C}_i = C_i$.
  - The term $V_{3G}(c)$ arises from the fact that we estimate at the point $\widehat{c} = \rho\widehat{G}(\theta)$ rather than $c = \rho G(\theta)$. See Ahn (1995, Theorem 2) for comparison.
- Note that if Y_i is available and used in place of $\widehat{R}_i$, then the asymptotic variance is the same.
- The performance of $\widehat{h}(\widehat{c})$ should be compared with that of the unrestricted estimator $\widehat{R}(\rho,\theta,w)$, which it dominates in terms of magnitude of MSE.

Simultaneous Estimators of h and G

Let $\Omega_4(\theta) = I_4^{-2}(\theta)V_4(\theta)$, where

$$I_4(\theta) = \int \left[\frac{\partial h}{\partial \gamma}(\rho G_0(\theta), w)\right]^2 \rho^2 f_{\rho,w,\theta}(\rho, w, \theta) d\rho dw,$$

$$V_4(\theta) = \vartheta_V \int \sigma^2(\rho, \theta, w) \left[\frac{\partial h}{\partial \gamma}(\rho G_0(\theta), w)\right]^2 \rho^2 f_{\rho,w,\theta}(\rho, w, \theta) d\rho dw$$

$$V_{5G}(c) = \left[\frac{\partial h}{\partial \gamma}(c)\rho\right]^2 \Omega_4(\theta)$$

THEOREM 5. *Suppose that assumptions A, B, C, and D hold. Then, there exists a bounded continuous function $\beta_4(\theta)$ such that as $n \to \infty$*

$$\sqrt{nb^{d_\theta}}\left(\widetilde{G}(\theta) - G_0(\theta) - b^p I_4^{-1}(\theta)\beta_4(\theta)\right) \Longrightarrow N(0, \Omega_4(\theta)).$$

Let $\Omega_5(c) = V_{5G}(c)1(d_W + 1 \geq d_\theta) + V_{5h}(c)1(d_W + 1 \leq d_\theta)$ with $V_{5h}(c) = V_{3h}(c)$. Then, there exist bounded continuous functions $\beta_{5G}(c), \beta_{5h}(c)$ such that as $n \to \infty$, with $\alpha_n^* = \min\{\sqrt{nb^{d_\theta}}, \sqrt{nb_*^{d_W+1}}\}$,

$$\alpha_n^*\left(\widetilde{h}(\widetilde{c}) - h(c) - b^p \beta_{5G}(c) - b_*^p \beta_{5h}(c)\right) \Longrightarrow N(0, \Omega_5(c)).$$

- Under homoskedasticity, i.e., $\sigma^2(\rho, \theta, w) = \sigma^2$, we have

$$\begin{aligned}
\Omega_4(\theta) &= \frac{\vartheta_V \sigma^2}{f_\theta(\theta)} \frac{1}{\int \left[\frac{\partial h}{\partial \gamma}(\rho G_0(\theta), w)\right]^2 \rho^2 f_{\rho,w|\theta}(\rho, w|\theta) d\rho dw} \\
\Omega_2(\theta) &= \frac{\vartheta_V \sigma^2}{f_\theta(\theta)} \frac{\int_A \frac{f^2_{\rho,w}(\rho,w)}{f_{\rho,w|\theta}(\rho,w|\theta)} d\rho dw}{\left(\int_A \frac{\partial h}{\partial \gamma}(\rho G_0(\theta), w) \rho f_{\rho,w}(\rho, w) d\rho dw\right)^2}.
\end{aligned}$$

  In this case, by the Cauchy-Schwarz inequality $\Omega_4(\theta) \leq \Omega_2(\theta)$ and so $\widetilde{G}(\theta)$ is more efficient than $\widehat{G}(\theta)$.

- Regarding the bias term, this contains many terms. By undersmoothing the first stages one can obtain a simple bias.

- The limiting distribution of $\widetilde{h}(\widetilde{c})$ differs from that of $\widehat{h}(\widehat{c})$ only because the estimation points $\widetilde{c}$ and $\widehat{c}$ are different; for any common evaluation point $c = \widehat{c}$ or $c = \widetilde{c}$, $\widetilde{h}(c)$ and $\widehat{h}(c)$ have the same asymptotic variance.

- Define the infeasible criterion function

$$\widetilde{m}_4(g;\theta) = \int \left[\widehat{R}(\rho,\theta,w) - h_0\left(\rho g, w\right)\right] \frac{\partial h_0}{\partial \gamma}\left(\rho G_0(\theta), w\right) \rho f_{\rho,w|\theta}(\rho, w|\theta) d\rho dw$$

  for any g, and let $\overline{G}(\theta)$ be the estimator that is any approximate zero of $\widetilde{m}_4(g;\theta)$. This is equivalent to the least squares estimator one would want to compute given knowledge of h_0. The distribution theory for this estimator is readily obtained: to first order, the distribution of $\widetilde{G}(\theta)$ is equivalent to the distribution of $\overline{G}(\theta)$. This is a sort of oracle efficiency property of our estimator.

- Furthermore, the theory of Stone (1980,1986) yields an optimal rate for estimation of $G_0(\theta)$ given knowledge of $h_0(.)$; this rate is achieved by $\overline{G}(\theta)$ and hence by our estimator.

- In the presence of heteroskedasticity, i.e., $\sigma^2(\rho,\theta,w) = \sigma^2$ for all ρ, θ, w, one should alter the criterion from least squares to weighted least squares and the resulting estimator will involve an additional weighting factor. However, the efficacy of this approach in practice may be limited and depends on the form of the heteroskedasticity.

Simulations

- We take supports

$$\Psi_\rho = [0,2] \text{ and } \Psi_\theta = \{\theta : 0 \leq \theta \leq \pi \text{ radians}\},$$

 and we take (ρ, θ) mutually independent and uniform on their supports, so that $f_\rho(\rho) = 1(\rho \in [0,2])/2$ and $f_\theta(\theta) = 1(\theta \in [0,\pi])/\pi$. We take

$$G(\theta) = 1 \text{ and } h(\gamma) = \exp(\gamma).$$

 In this case, $U_0(\theta, \theta') = G(\theta)/G(\theta') = 1$. Then take $A = [0.5, 1.5] \subset \cap_{\theta \in \Psi_\theta} \Psi_{\rho|\theta}$.

- The nonparametric functions used in each step of the estimation are constructed using ordinary kernel regressions with a Gaussian kernel.
- For $\widehat{G}(\theta)$ supposing we take $\varpi(\theta') = f_\theta(\theta')$, the variance constant is

$$\Omega_2(\theta) = ||K||_2^2 \sigma^2 \frac{\int_A f_\rho(\rho) d\rho}{f_\theta(\theta) \left(\int_A \frac{\partial h}{\partial \gamma}(\rho G_0(\theta)) \rho f_\rho(\rho) d\rho \right)^2} = ||K||_2^2 \sigma^2 \frac{\pi}{2 \left(\int_{0.5}^{1.5} \exp(\rho) \rho d\rho \right)^2} = ||K||_2^2 \sigma^2 \times 0.16719$$

- For $\widetilde{G}(\theta)$ we have

$$\Omega_4(\theta) = ||K||_2^2 \frac{\sigma^2}{f_\theta(\theta)} \frac{1}{\int \rho^2 \exp(2\rho G_0(\theta)) f_\rho(\rho) d\rho} = ||K||_2^2 \sigma^2 \times 9.2403 \times 10^{-2}.$$

- We report results for three different sample sizes and three different error variances, for a total of nine designs. Each design is estimated using three different bandwidths b_1, b_2, and b_3, where b_2 is given by Silverman's rule ($1.06n^{-1/5}$ times the square root of the average of the regressor variances), $b_1 = 0.5 * b_2$, and $b_3 = 1.5 * b_2$. These kernel and bandwidth choices are not likely to be optimal for our setting, but are chosen because they are commonly used in applications and are easy to calculate.
- For each estimated function G and h we calculate four criteria summarizing goodness of fit. These are integrated mean squared error IMSE, integrated mean absolute error IMAE, pointwise mean squared error PMSE, and pointwise mean absolute error PMAE. Results are based on a hundred simulations of each design and bandwidth. These are reported in Tables 1 and 2.

Table 1

	$\frac{\sigma_x^2}{\sigma_x^2+\sigma_\varepsilon^2}$	0.75			0.5			0.25		
	n	100	200	400	100	200	400	100	200	400
IMSE	h_1	0.2355	0.1632	0.1186	0.3671	0.2958	0.2344	0.4547	0.4158	0.3678
	h_2	0.1130	0.0805	0.0621	0.2119	0.1539	0.1178	0.3409	0.2860	0.2337
	h_3	0.0797	0.0601	0.0475	0.1410	0.1044	0.0828	0.2514	0.1971	0.1592
IMAE	h_1	0.1852	0.1287	0.0928	0.2942	0.2382	0.1884	0.3682	0.3350	0.2977
	h_2	0.0895	0.0637	0.0493	0.1708	0.1223	0.0934	0.2796	0.2326	0.1892
	h_3	0.0644	0.0486	0.0381	0.1140	0.0843	0.0665	0.2082	0.1612	0.1287
PMSE	h_1	0.3921	0.2388	0.1669	0.5210	0.4885	0.3525	0.6240	0.9085	0.6021
	h_2	0.1098	0.0745	0.0576	0.2542	0.1842	0.1330	0.5193	0.4423	0.3310
	h_3	0.0636	0.0506	0.0415	0.1266	0.0933	0.0731	0.2599	0.2214	0.1809
PMAE	h_1	0.1947	0.1357	0.0967	0.2952	0.2661	0.2073	0.3607	0.3929	0.3451
	h_2	0.0736	0.0553	0.0447	0.1527	0.1139	0.0872	0.2718	0.2370	0.1955
	h_3	0.0489	0.0397	0.0318	0.0879	0.0681	0.0549	0.1695	0.1372	0.1120

Notes: The results for $\widehat{G}$

Table 2

	$\frac{\sigma_x^2}{\sigma_x^2+\sigma_\varepsilon^2}$	0.75			0.5			0.25		
	n	100	200	400	100	200	400	100	200	400
IMSE	h_1	0.7208	0.5817	0.4689	0.9271	0.8386	0.7401	1.0703	1.0258	0.9641
	h_2	0.4770	0.3765	0.3072	0.7149	0.5878	0.4930	0.9543	0.8553	0.7593
	h_3	0.3902	0.3149	0.2589	0.5716	0.4675	0.3934	0.8247	0.7027	0.6117
IMAE	h_1	0.4996	0.3900	0.3052	0.6705	0.5915	0.5095	0.7944	0.7474	0.6916
	h_2	0.3248	0.2494	0.1987	0.5090	0.4057	0.3305	0.7046	0.6197	0.5357
	h_3	0.2753	0.2179	0.1754	0.4085	0.3253	0.2666	0.6100	0.5060	0.4280
PMSE	h_1	0.6969	0.5521	0.4299	0.8395	0.7860	0.7358	0.9281	0.9613	0.9391
	h_2	0.4154	0.3045	0.2259	0.6245	0.5523	0.4647	0.8147	0.7541	0.7383
	h_3	0.3426	0.2588	0.2031	0.4972	0.4103	0.3353	0.6724	0.6051	0.5520
PMAE	h_1	0.4917	0.3843	0.2870	0.6158	0.5805	0.5261	0.7099	0.7398	0.7045
	h_2	0.2938	0.2139	0.1666	0.4635	0.3934	0.3274	0.6228	0.5686	0.5394
	h_3	0.2718	0.2045	0.1612	0.3833	0.3072	0.2487	0.5301	0.4639	0.4136

Notes: The results for $\widehat{h}$

- In Figures 2 and 3 we show the Q-Q plots of these normalized (to have mean zero and variance one) estimators at a central point θ and γ.

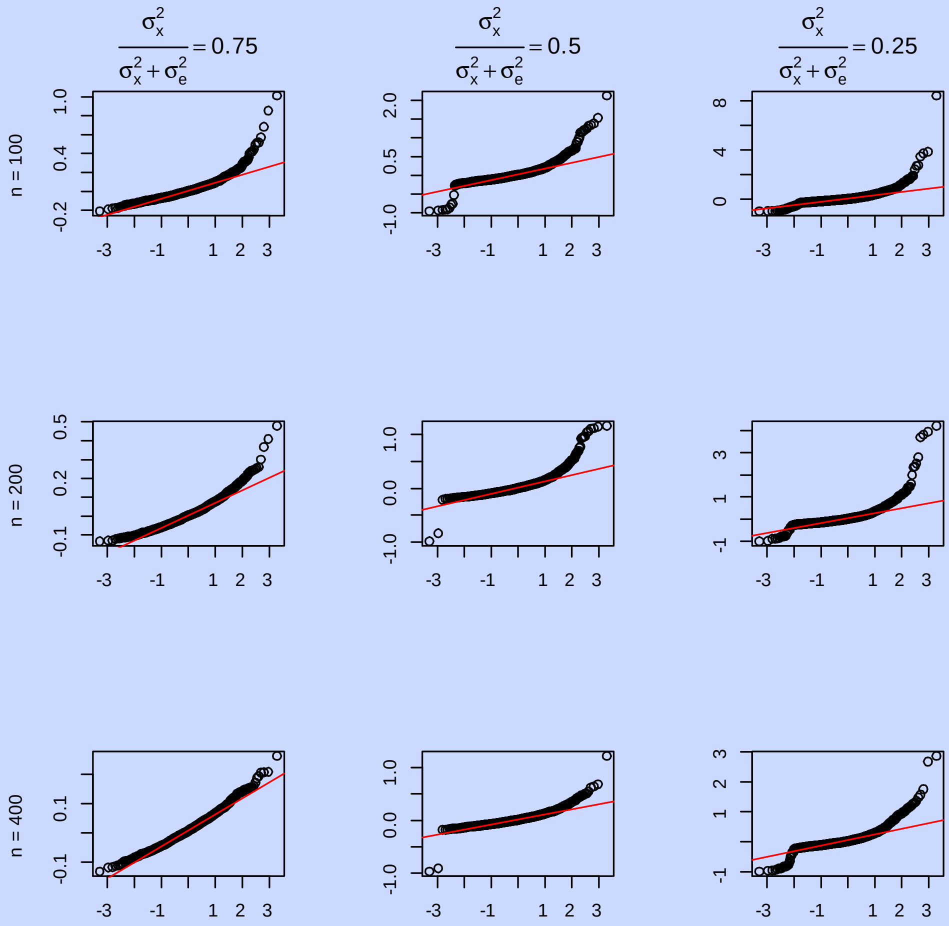


Figure 2. Q-Q plots for $\widehat{G}$

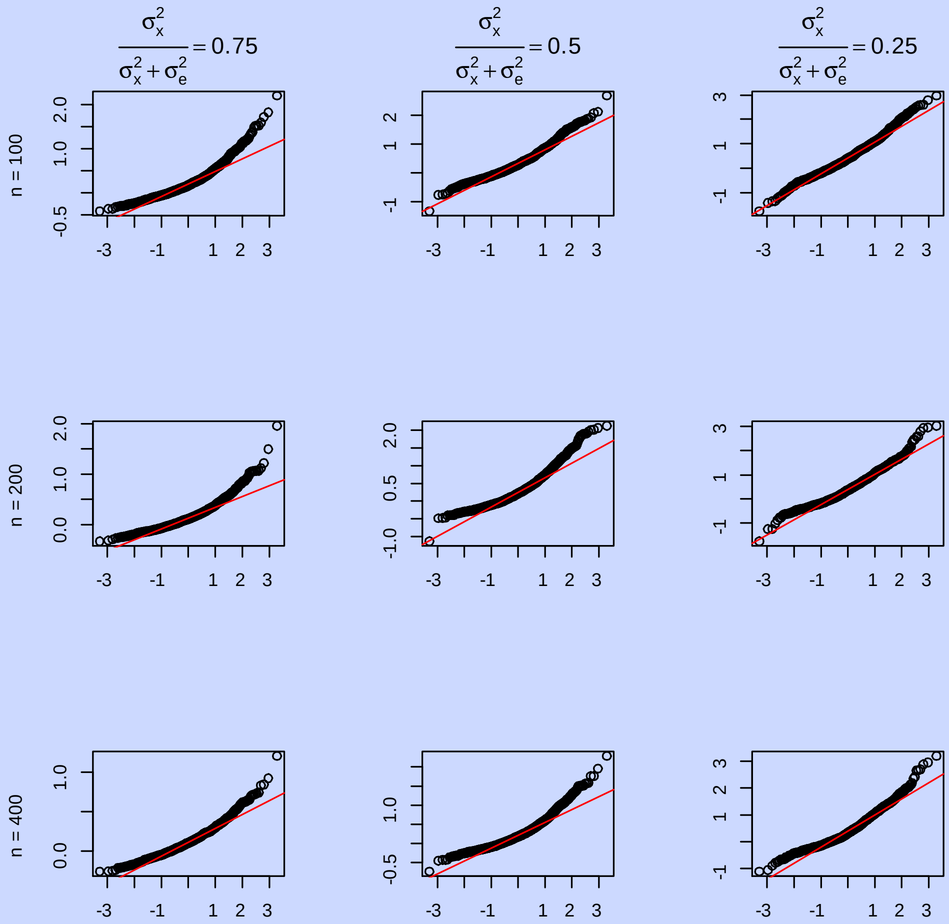


Figure 3. Q-Q plots for $\widehat{h}$

Application to Nonparametric Production Function Estimation

- Let y be the log output of a firm and x be a vector of inputs, and suppose that

$$E(y|x) = r(x) = h[g(x)]$$

  with linearly homogeneous g.

- A property of production that is empirically important is returns to scale, defined as

$$S(g) = \frac{\partial h(g)}{\partial \ln g}$$

- Other important properties are measures of substitutability of inputs, such as the technical rate of substitution and the elasticity of substitution. When x consists of just two elements, for example, capital K and labor L, then a simple measure of substitutability is

$$\alpha(K/L) = \frac{\partial \ln g\,(K/L, 1)}{\partial \ln(K/L)}$$

  Note in interpreting this measure that $g\,(K/L, 1) = g\,(K, L)\,/L$.

- The substitutability measure $\alpha(K/L)$ equals a constant α when $g(x) = K^{\alpha}L^{1-\alpha}$, that is, when the production function $r(x)$ is a monotonic transformation of a Cobb Douglas, which is a common specification for homothetic production.

- Observations of chemical manufacturing firms in mainland China in two time periods, 1995 and 2001. For each firm, we observe
  - the net value of real fixed assets K
  - the number of employees L
  - Y defined as the log of value-added real output.
- Output and capital are measured in thousands of Yuan converted to the base year 2000 using a general price deflator for the Chinese chemical industry. A total sample size of 1638 firms in 2001 and 1560 firms in 1995.

- We consider both nonparametric and parametric estimates of the production function $r(K, L)$. The parametric model we employ is a homothetic Translog production function, in which log output

$$Y = h[g(K, L)] + \epsilon$$

$$\begin{aligned} g(K, L) &= \left(\frac{K}{L}\right)^{\alpha} L \\ h(g) &= \beta_0 + \beta_1 \ln(g) + \beta_2 \ln(g)^2 \end{aligned}$$

- Fitting this model by nonlinear least squares in each of the years of data yields the parameter estimates reported in Table 3 (standard errors are in parentheses).

TABLE 3: Parametric Translog Estimates

	α	β_0	β_1	β_2
2001 Translog	0.696	9.815	0.783	0.036
	(0.043)	(0.031)	(0.028)	(0.012)
1995 Translog	0.478	9.585	0.961	0.045
	(0.046)	(0.024)	(0.041)	(0.017)

- Figures 2 and 3 show homothetic Translog and homothetic nonparametric estimates $\widehat{g}(K/L, 1)$ and $\widehat{h}(g)$ in 2001.
- Figure 3 also shows fits from a simple nonhomothetic kernel regression of Y on K, L, that is, the initial unconstrained estimator of the function r.
- For simplicity, at each nonparametric estimation step we used ordinary kernel regressions with a normal kernel and bandwidth given by Silverman's rule.

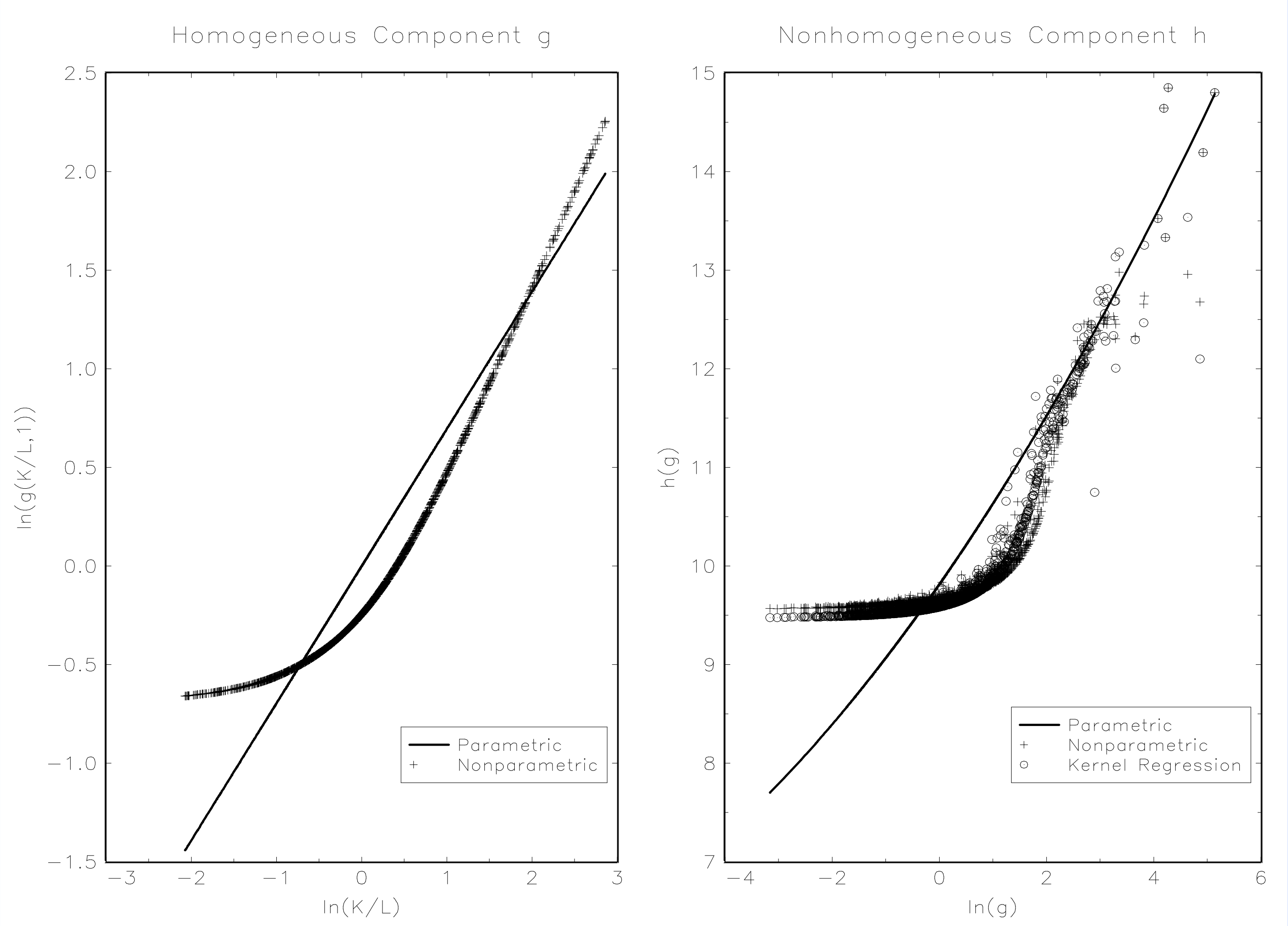
- The nonparametric fits of r and those of h shown in Figure are quite similar, indicating that the imposition of homotheticity is reasonable for this data set.
- The nonparametric estimates of the functions g and h are roughly similar to the parametric Translog model estimates, but show quite a bit more curvature, departing most markedly from the parametric model for g at low capital to labor ratios and from the model for $h(g)$ at low values of g.
- These differences are greatly magnified when one calculates the returns to scale $S(g)$ and the substitution measure $\alpha(K/L)$. For the Translog model,

  $$S(g) = \beta_1 + 2\beta_2 \ln(g) \text{ and } \alpha(K/L) = \alpha.$$

  For the nonparametric model we use the approximation

  $$\widehat{S}(\widehat{g}_i) \approx [\widehat{h}(\widehat{g}_{i+1}) - \widehat{h}(\widehat{g}_{i-1})]/(\widehat{g}_{i+1} - \widehat{g}_{i-1})$$

  after sorting the data by $\widehat{g}_i$ for each firm i, and similarly for $\widehat{\alpha}(K/L)$.
- Unlike the popular homothetic Translog model, which assumes α constant, the nonparametric estimates have α sharply increasing at low capital labor ratios and leveling off only at high levels. This result indicates likely inadequacies of the parametric model. The assumption of a constant α may be more reasonable for advanced economies like the United States, which tend to have higher capital labor.
- The models also differ in returns to scale $S(g)$. Both models imply similar returns to scale on


Homogeneous Component g
ln(g(K/L,1))
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3
ln(K/L)
Parametric
+ Nonparametric
Nonhomogeneous Component h
h(g)
15
14
13
12
11
10
9
8
7
−4
−2
0
2
4
6
ln(g)
Parametric
+ Nonparametric
○ Kernel Regression

average, but the parametric model has $S(g)$ mildly increasing, based on a small but statistically significant positive estimate of $\widehat{\beta}_2$. In contrast, the nonparametric estimates are roughly U shaped, with a majority of the data in the decreasing part. Given the substantial variability of the nonparametric $\widehat{S}$, it is difficult to draw conclusions about the dependence of S on g.

- The estimates based on 1995 data are broadly similar to 2001. The major difference between the two years is that average returns to scale appear to have declined over time, from approximately constant returns with average S near one in 1995, to decreasing returns with S near 0.8 in 2001.
- This finding could be an artifact of substantial ownership reform during this period. Many larger firms in the Chinese chemical industry may still be state-owned in 2001, while many smaller enterprises were privatized after 1995 and so could have substantially restructured, thereby enhancing their productivity. Combining these into a single cross section might then create the appearance of decreasing returns on average.
- This could explain the overall difference in mean S between the two years, but would explain the observed patterns in $S(g)$ within each year, though as noted above these departures of $S(g)$ from a constant are at best weakly estimated.
- Changes over time may more generally be due to changes in technology, demand, and other aspects of China's increasing economic liberalization and growth over this time period.

Some Extensions

Endogenous Regressors

- Assume

  $$y = H[g(x), w, \varepsilon],$$

  and elements of X, W are endogenous, correlated with ε. Let

  $$U = (X, W) - E(X, W \mid Z).$$

  Then $\varepsilon \mid X, W, Z \backsim \varepsilon \mid U, Z$. Define

  $$\Upsilon(X, W, U) = E(Y \mid X, W, U).$$

  Assume that $\varepsilon \mid U, Z \backsim \varepsilon \mid U$. Then

  $$\Upsilon(x, w, u) = h[g(x), w, u].$$

- Let $\widehat{U}$ be residuals from nonparametrically regressing X, W on Z. Let $\widehat{\Upsilon}$ be a nonparametric regression of Y on $X, W, \widehat{U}$. Then apply the homotheticity estimator to $\widehat{\Upsilon}$ to get $\widehat{g}$.
- Assumption $\varepsilon \mid U, Z \backsim \varepsilon \mid U$ is like control functions of Blundell,Powell (2000,1) nonparametric triangular system of Newey, Powell,Vella (1999), Imbens and Newey (2001), Chesher (2001).

Another Model

- Now consider

$$r(v, z, w) = h[m(z) + v, w]$$

  m need not be homogeneous. This is the same as above taking $m(z) = \ln(G(\theta))$, $v = \ln(\rho)$ and $h = \exp(h)$ from the polar notation of our old model.

- Example: partly linear index models, reservation price and willingness to pay models such as $y = I[-m(x) + \varepsilon \leq v]$ where v is the price.

- Also censored regression. Suppose that we observe Y, X where

$$\begin{aligned} Y^* &= g(X) - \varepsilon \\ Y &= \max\{Y^*, 0\}. \end{aligned}$$

  Then

$$\Pr(Y \leq y | X = x) = F_\varepsilon(y - g(x))$$

  for all $y \geq 0$ and all x.