An Introduction to Stochastic Nonparametric Envelopment of Data (StoNED)

Kuosmanen, Johnson, and Saastamoinen (2014)
Kuosmanen and Kortelainen (2012)

Dr. Chia-Yen Lee (李家岩 博士)

Institute of Manufacturing Information and Systems (製造資訊與系統研究所)
Research Center for Energy Technology and Strategy (能源科技與策略研究中心)
Department of Computer Science and Information Engineering (資訊工程學系)
National Cheng Kung University, Taiwan (國立成功大學)

Outline

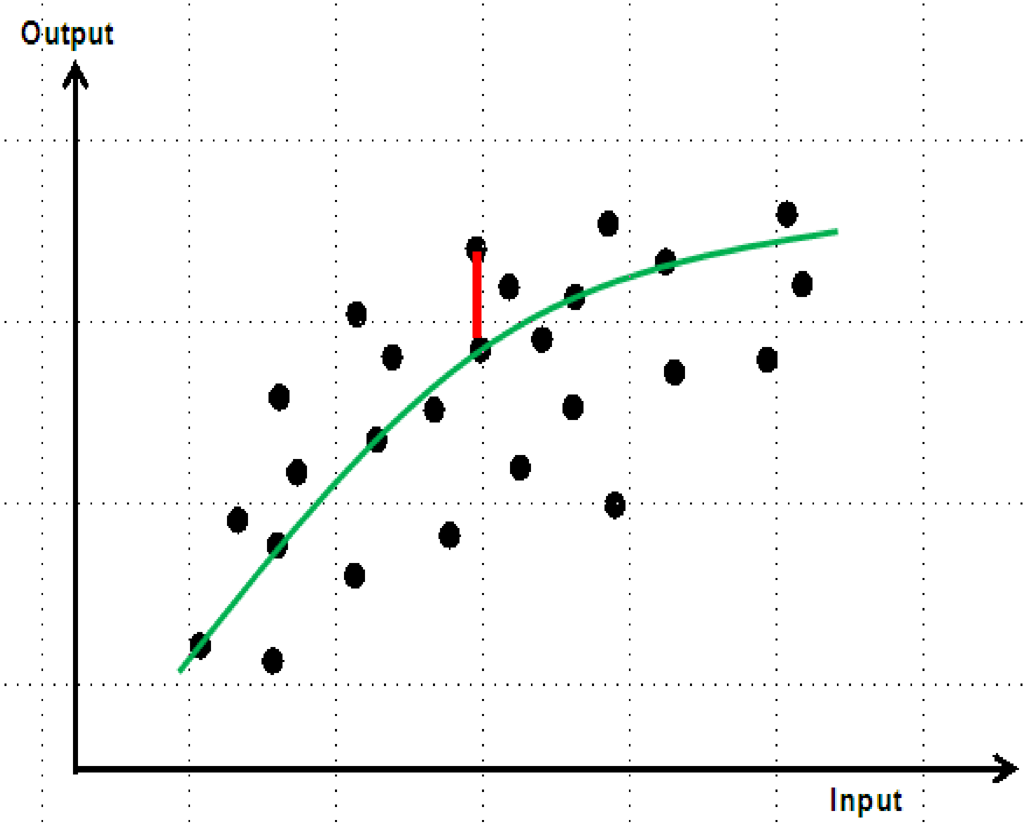
- Introduction
  - Stochastic Frontier Analysis (SFA) vs. Data Envelopment Analysis (DEA)
- Convex Nonparametric Least Squares (CNLS)
  - DEA as Sign-Constrained CNLS
  - Corrected Convex Nonparametric Least Squares (C^2NLS)
  - Relaxed CNLS
- Stochastic Nonparametric Envelopment of Data (StoNED)
- Conclusion

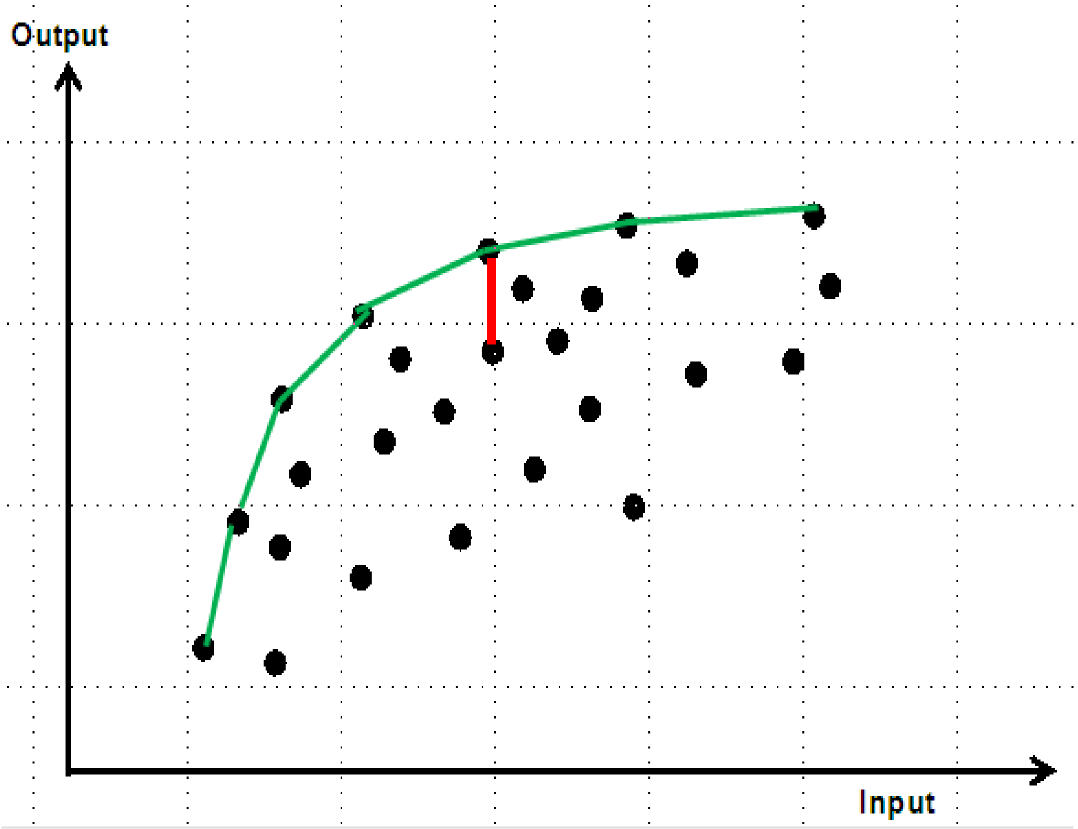
SFA vs. DEA

- Deviations from the regression line are considered unobserved effects
- Deviations from the DEA frontier are assumed to be systematic inefficiency
- Actually, there might be a mix of both. This is the motivation for stochastic frontier analysis.

SFA vs. DEA

- SFA vs. DEA





Modeling Tools and Assumptions

Comparison	DEA	SFA
Noise	No	Every observation is influenced by noise
Model Specification (production form, noise distribution, inefficiency distribution)	No, nonparametric	Yes, parametric functional form (linear)
Estimated Production Function	Piece-wise linear	Smooth
Principle	Minimum extrapolation	Composed error term
Outlier	Sensitive	Not sensitive

$$y_i = f(\mathbf{x}_i) - u_i + v_i$$

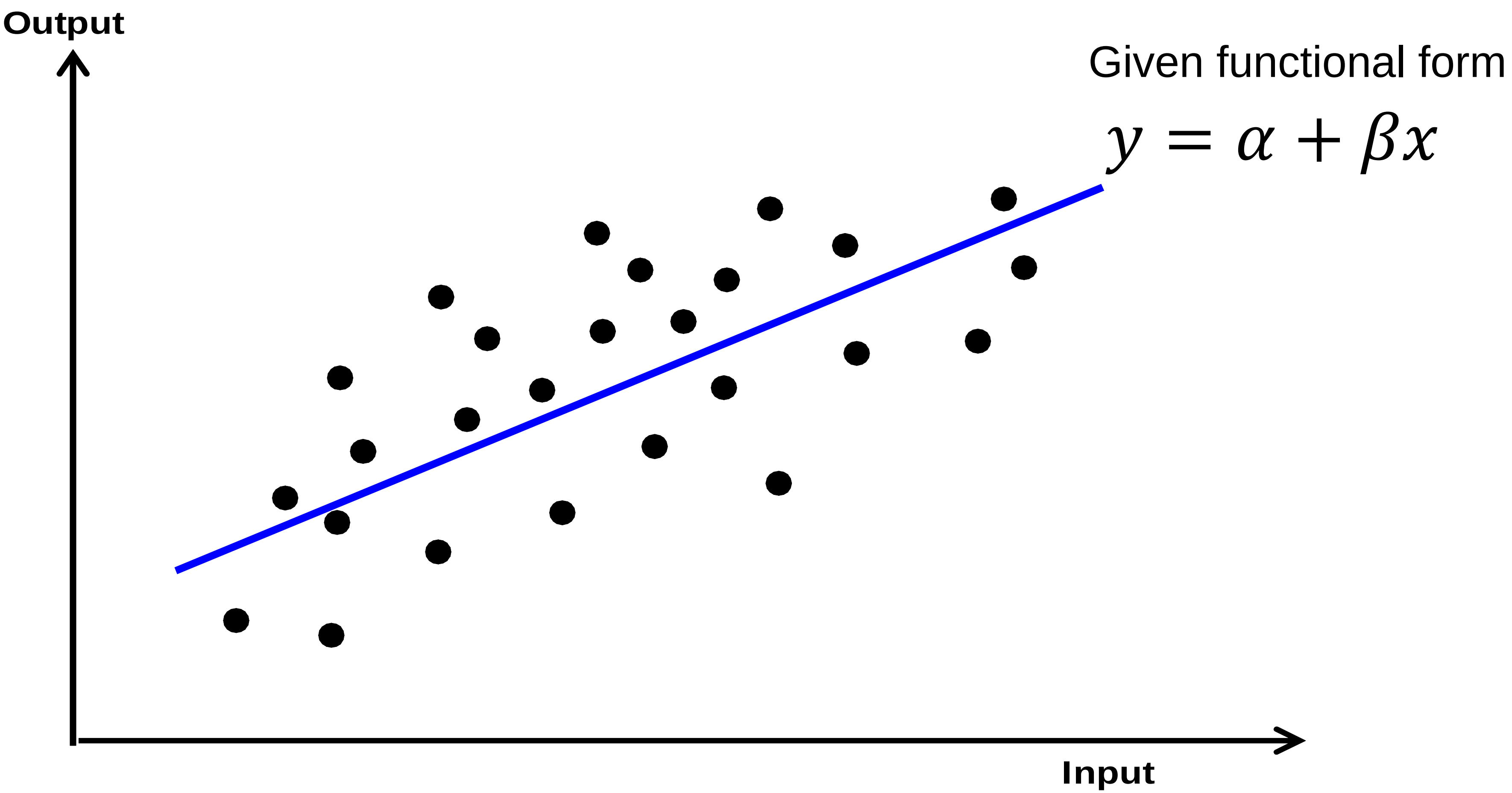
Parallel Development of Productivity Models

		Parametric	Nonparametric
Central tendency		*OLS* Cobb and Douglas (1928)	*CNLS* Hildreth (1954) Hanson and Pledger (1976)
Deterministic frontier	**Sign constraints**	*PP* Aigner and Chu (1968) Timmer (1971)	*DEA* Farrell (1957) Charnes et al. (1978)
	2-step estimation	*COLS* Winsten (1957) Greene (1980)	*C^2NLS* Kuosmanen and Johnson (2010)
Stochastic frontier		*SFA* Aigner et al. (1977) Meeusen and Vanden Broeck (1977)	*StoNED* Kuosmanen and Kortelainen (2012)

Kuosmanen et al. (2014)

Parametric Models

- Central Tendency
  - Regression-based: Ordinary Least Square (OLS)



A production function is a function that represents "maximum outputs" that can be achieved using input vector $\mathbb{x}$.

Parametric Models

- ## Deterministic Frontier
  - Parametric programming (PP)

$$\min_{\alpha,\beta,\varepsilon}\left\{\sum_{i=1}^{n}\varepsilon_i^2 \middle| \varepsilon_i \le 0\ \forall i=1,\dots,n; y_i=\alpha+\boldsymbol{\beta}'\mathbf{x}_i+\varepsilon_i\ \forall i=1,\dots,n\right\}$$

  - both shifts the OLS regression line upwards to the frontier and influences the coefficients.
  - the estimated intercept and slope coefficients obtained by PP model generally differ from the OLS estimates
- ## Quadratic Objective Function Linearization

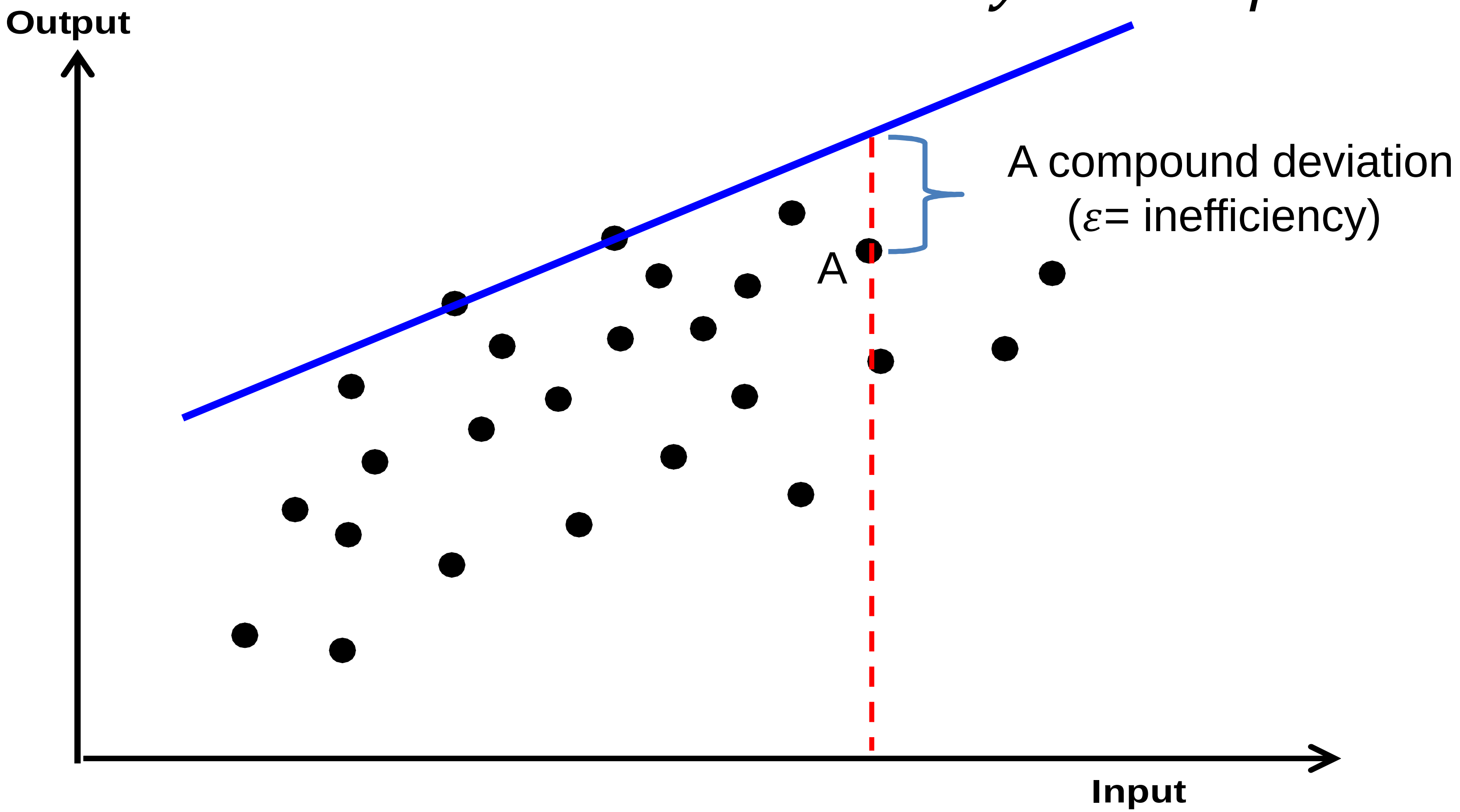
$$\min\sum_{i=1}^{n}-\varepsilon_i$$

  - However, this linearization will generally change the PP problem.
  - Schmidt (1976)
    - the linearized PP → the maximum likelihood estimator (MLE) for exponentially distributed inefficiency terms
    - the quadratic PP → the MLE for the half-normal inefficiency terms.

Corrected Ordinary Least Square (COLS)

- Deterministic Frontier
  - Corrected Ordinary Least Square (COLS)
    - Winsten(1957), Greene (1980)

$$y = \alpha + \beta x + \max(\varepsilon)$$



Parallel Development of Productivity Models

		Parametric	Nonparametric
Central tendency		*OLS* Cobb and Douglas (1928)	*CNLS* (Section 3) Hildreth (1954) Hanson and Pledger (1976)
Deterministic frontier	**Sign constraints**	*PP* Aigner and Chu (1968) Timmer (1971)	*DEA* (Section 4.1) Farrell (1957) Charnes et al. (1978)
	2-step estimation	*COLS* Winsten (1957) Greene (1980)	*C^2NLS* (Section 4.2) Kuosmanen and Johnson (2010)
Stochastic frontier		*SFA* Aigner et al. (1977) Meeusen and Vanden Broeck (1977)	*StoNED* (Section 5) Kuosmanen and Kortelainen (2012)

Kuosmanen et al. (2014)

Nonparametric Models

- Convex Nonparametric Least Squares (CNLS) Estimator

$$\min_{\varepsilon} \sum_{i=1}^{n} \varepsilon_i^2$$

s.t.

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \forall i$$

f is monotonic increasing and concave

CNLS

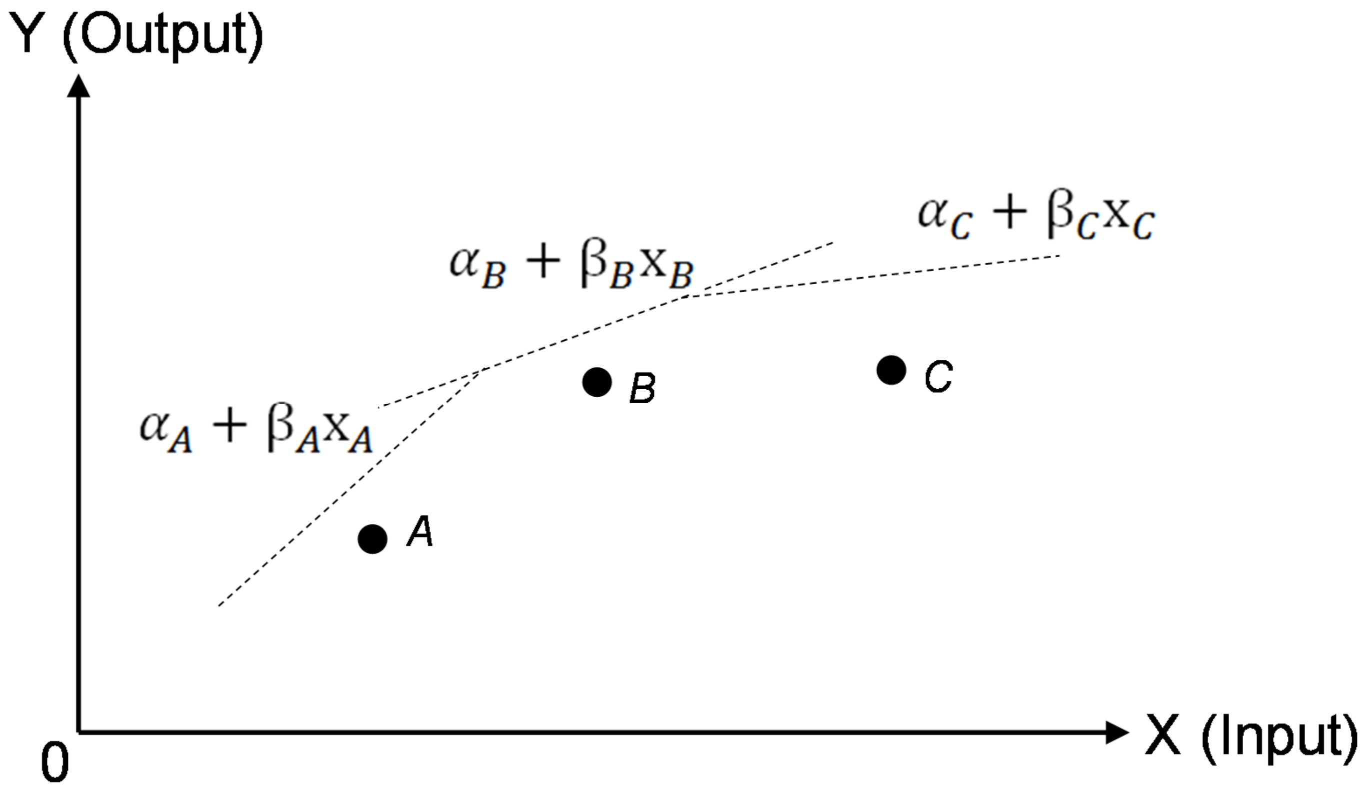
- CNLS can be traced to the seminal work of Hildreth (1954) and was popularized by Kuosmanen (2008) as a powerful tool for describing the average behavior of observations.
  - ➢ the convexity constraint can be modeled by the Afrait inequalities

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \left\{ \sum_{i=1}^{n} \varepsilon_i^2 \;\middle|\; \begin{array}{l} \overbrace{y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i \;\; \forall i = 1,\ldots,n;}^{\text{Linear Regression}} \\ \underbrace{\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i = 1,\ldots,n;}_{\text{Convexity}} \\ \underbrace{\boldsymbol{\beta}_i \ge \mathbf{0} \;\; \forall i = 1,\ldots,n}_{\text{Monotonicity}} \end{array} \right\}$$

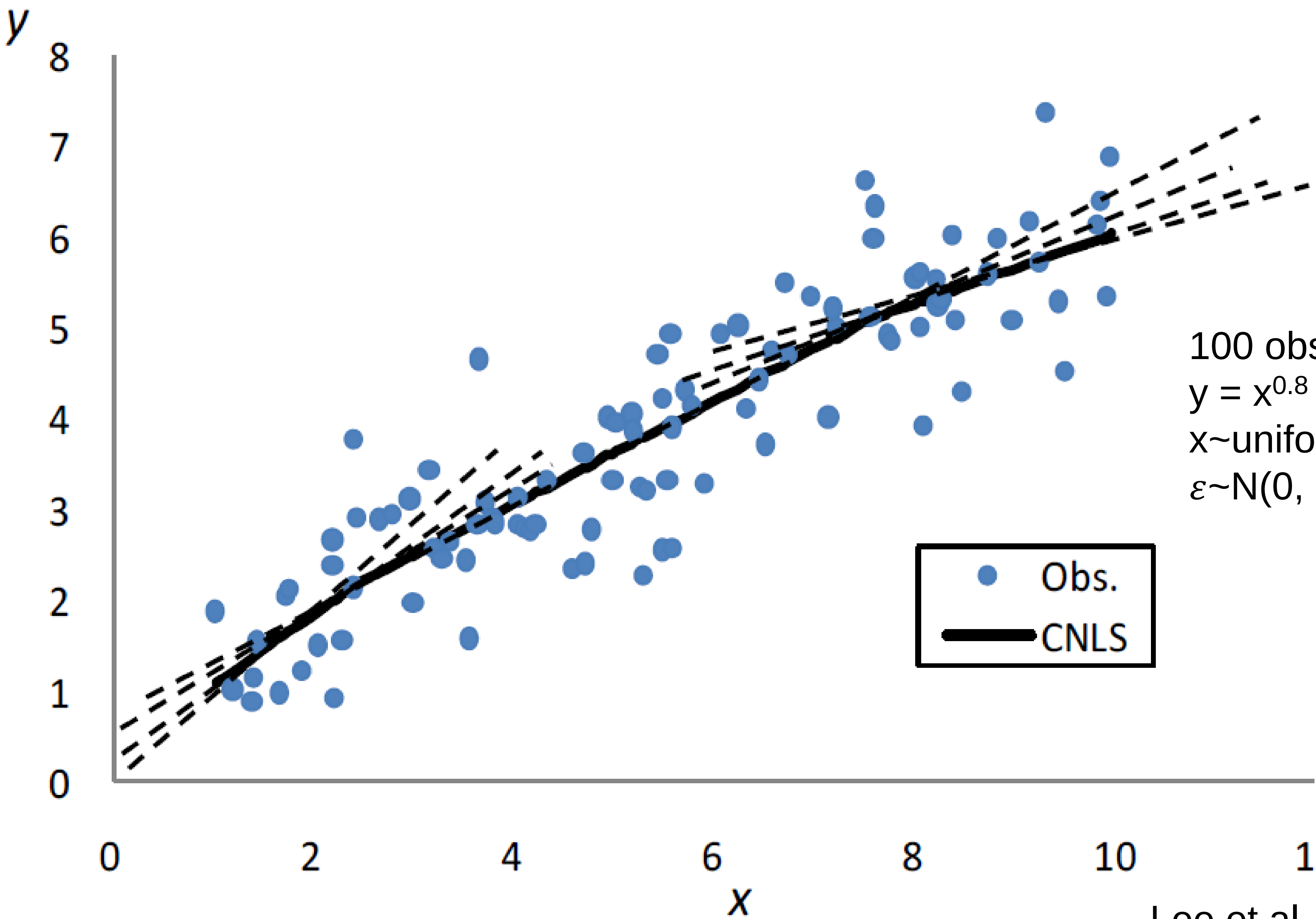
- Shortcomings
  - ➢ Multiple solution (Kuosmanen and Kortelainen, 2012)
    - Minimum extrapolation principle
  - ➢ Computational burden (Lee et al., 2013)
    - 2nd constraint will generate $n(n-1)$ constraints

Graphical Illustration of CNLS

- Single-Input Single-Output
  - Each observation has its own corresponding regression line.



Graphical Illustration of CNLS



100 observations
$y = x^{0.8} + \varepsilon$
x~uniform [1,10]
ε~N(0, 0.7^2)

Lee et al. (2013)

DEA as Sign-Constrained CNLS

- Additive DEA

$$\varepsilon_i^{DEA} = \min_{\lambda,\varepsilon}\left\{\varepsilon \,\middle|\, y_i = \sum_{h=1}^{n}\lambda_h y_h + \varepsilon;\ \mathbf{x}_i \geqslant \sum_{h=1}^{n}\lambda_h \mathbf{x}_h;\ \sum_{h=1}^{n}\lambda_h = 1;\ \lambda_h \geqslant 0\ \forall h = 1,\ldots,n\right\}$$

  - Radial (multiplicative) DEA measure

$$\theta_i^{DEA} = 1 - \varepsilon_i^{DEA}/y_i \quad \forall i = 1,\ldots,n.$$

- DEA as Sign-Constrained CNLS (Kuosmanen and Johnson, 2010)

$$\min_{\alpha,\beta,\varepsilon}\left\{\sum_{I=1}^{n}\varepsilon_i^2 \,\middle|\, \begin{array}{l} \boxed{\varepsilon_i \leqslant 0 \quad \forall i = 1,\ldots,n;} \\ y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i \quad \forall i = 1,\ldots,n; \\ \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \leqslant \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \quad \forall h, i = 1,\ldots,n; \\ \boldsymbol{\beta}_i \geqslant \mathbf{0} \quad \forall i = 1,\ldots,n \end{array}\right\}$$

Corrected Convex Nonparametric Least Squares (C2NLS)

- CNLS can be used in a two-stage shifting method. This method is a nonparametric variant of the Corrected Ordinary Least Squares (COLS) Winsten (1957); Greene (1980) model in which CNLS replaces the first-stage parametric OLS regression

Step1:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \left\{ \sum_{l=1}^{n} \varepsilon_i^2 \;\middle|\; \begin{array}{l} y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i \;\; \forall i = 1,\ldots,n; \\ \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i = 1,\ldots,n; \\ \boldsymbol{\beta}_i \ge \mathbf{0} \;\; \forall i = 1,\ldots,n \end{array} \right\}$$

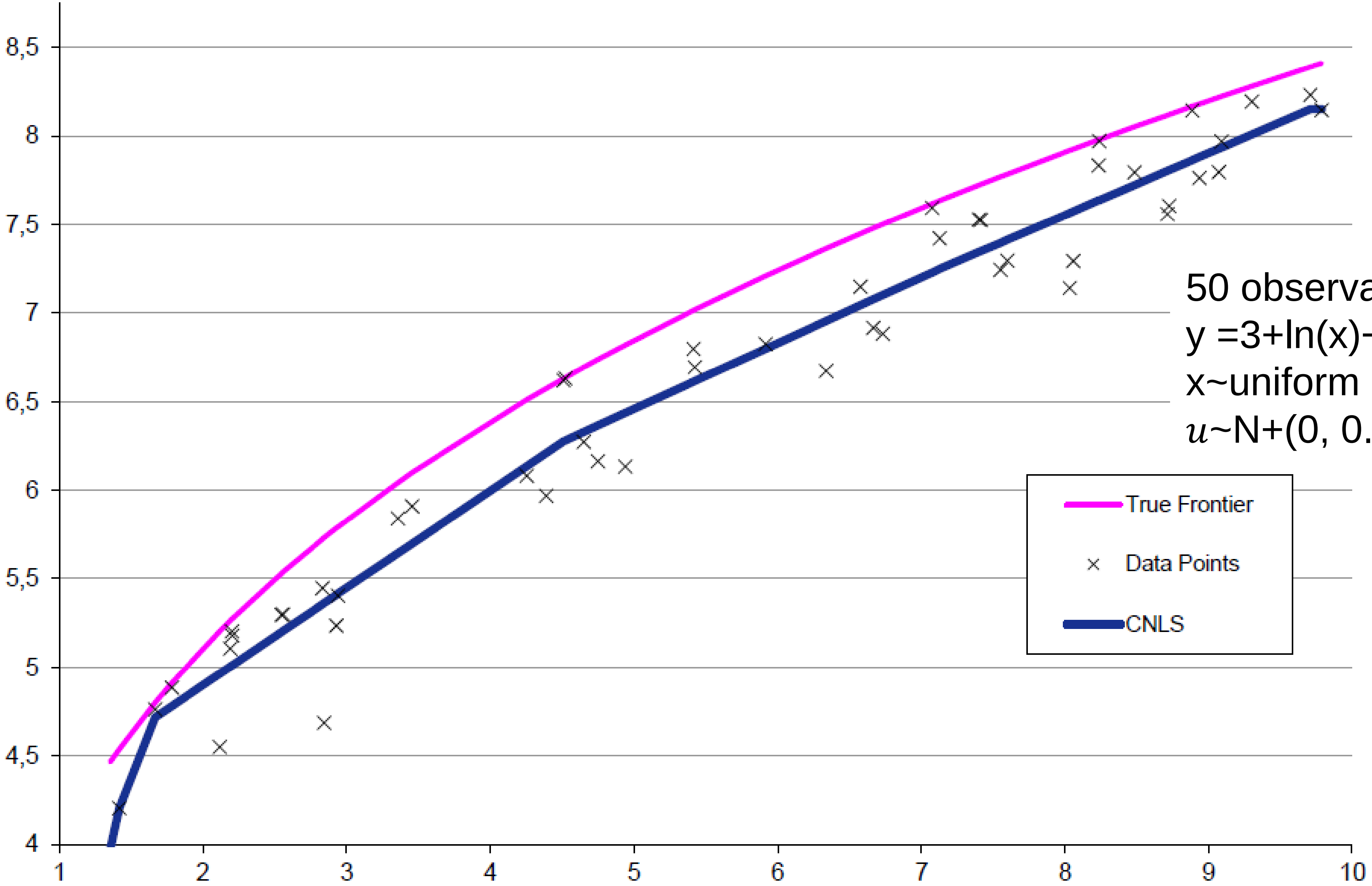
Step2:

$$\hat{\varepsilon}_i^{C2NLS} = \varepsilon_i^{CNLS} - \max_h \varepsilon_h^{CNLS}$$

$\hat{\varepsilon}_i^{C2NLS}$ range from $[0, -\infty]$

$$\hat{\alpha}_i^{C2NLS} = \alpha_i^{CNLS} + \max_h \varepsilon_h^{CNLS}$$

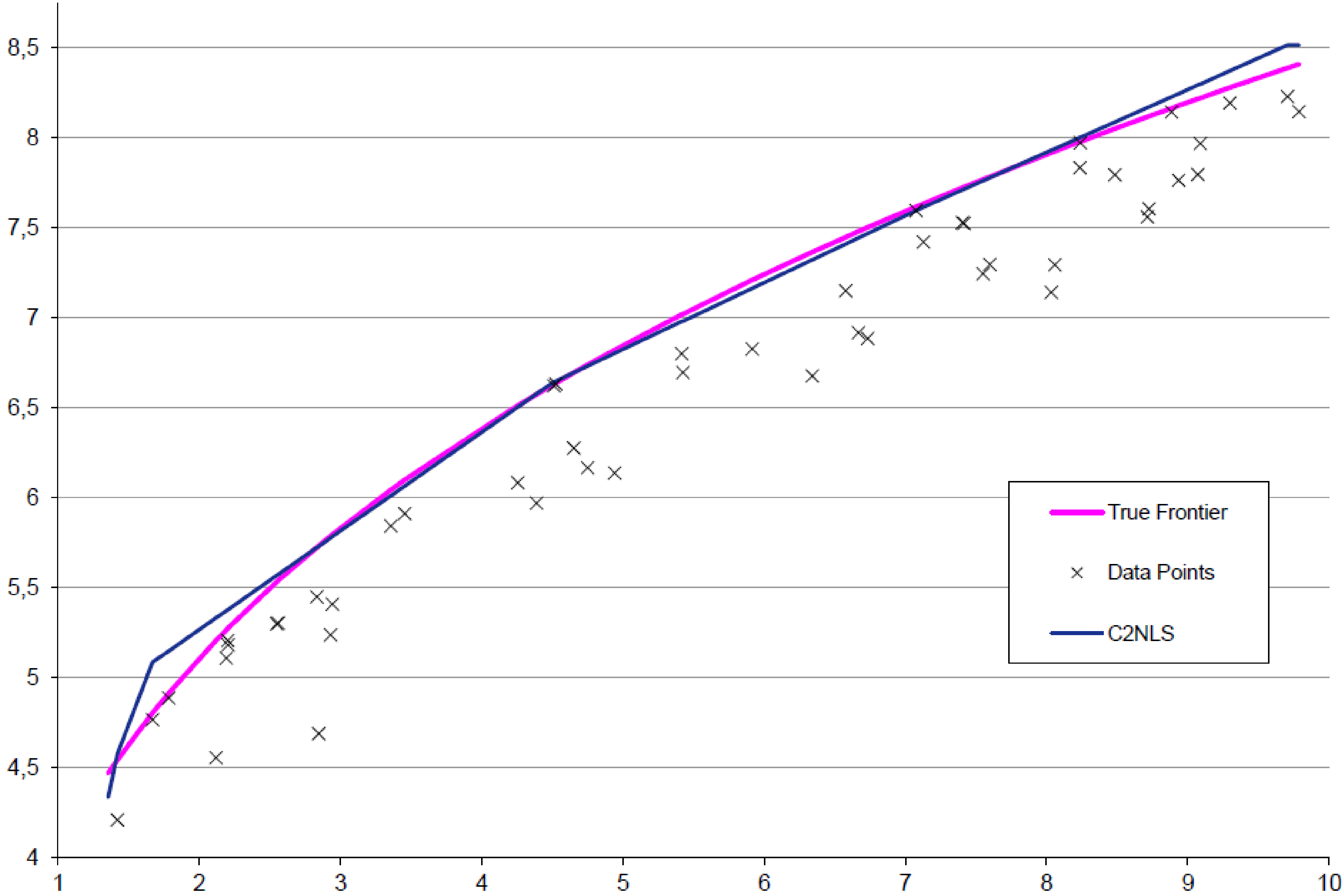
Graphical Illustration of CNLS



50 observations
y =3+ln(x)−u
x~uniform [1,10]
u~N+(0, 0.7^2)

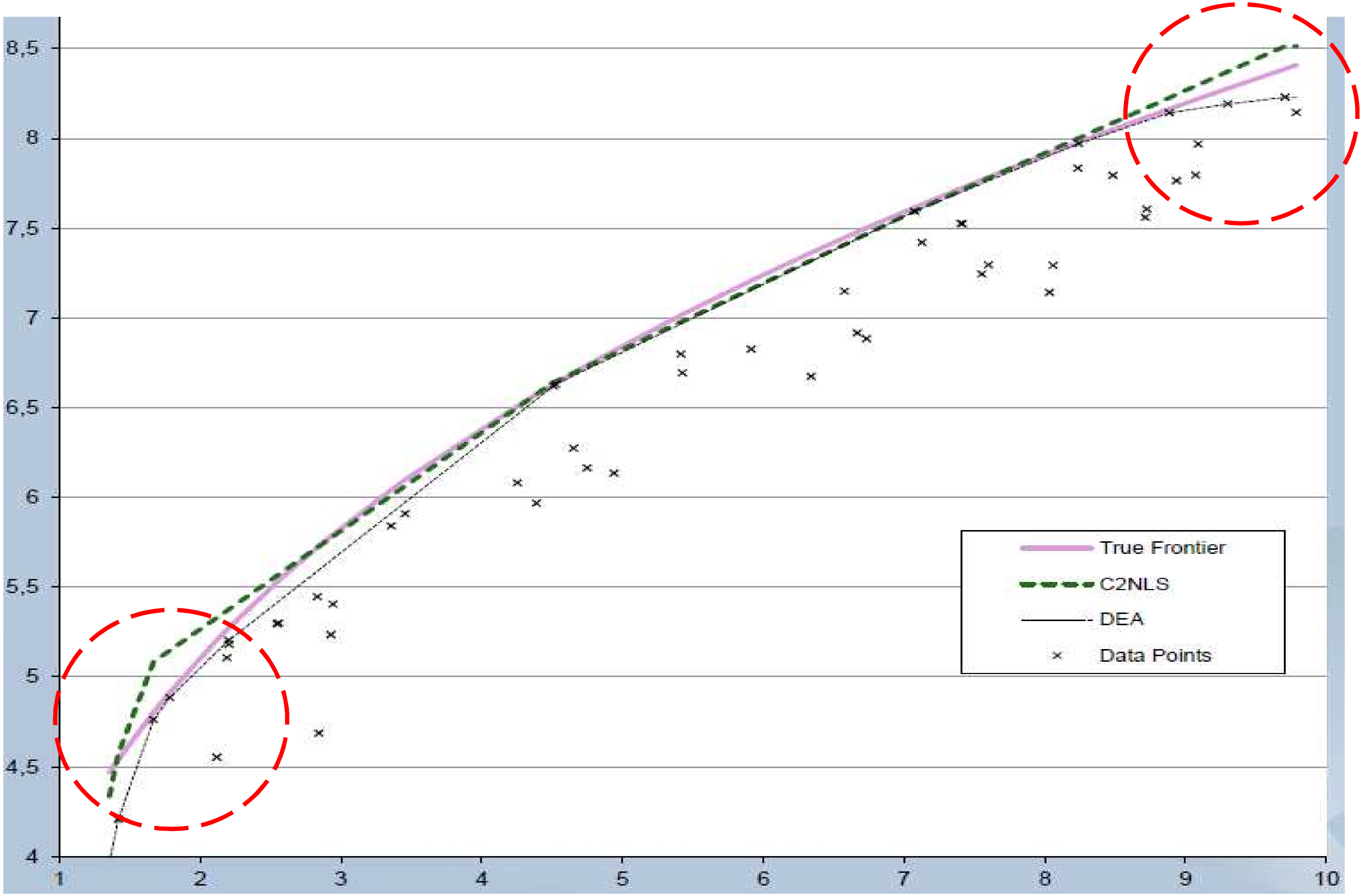
Kuosmanen and Johnson (2010)

Graphical Illustration of C²NLS



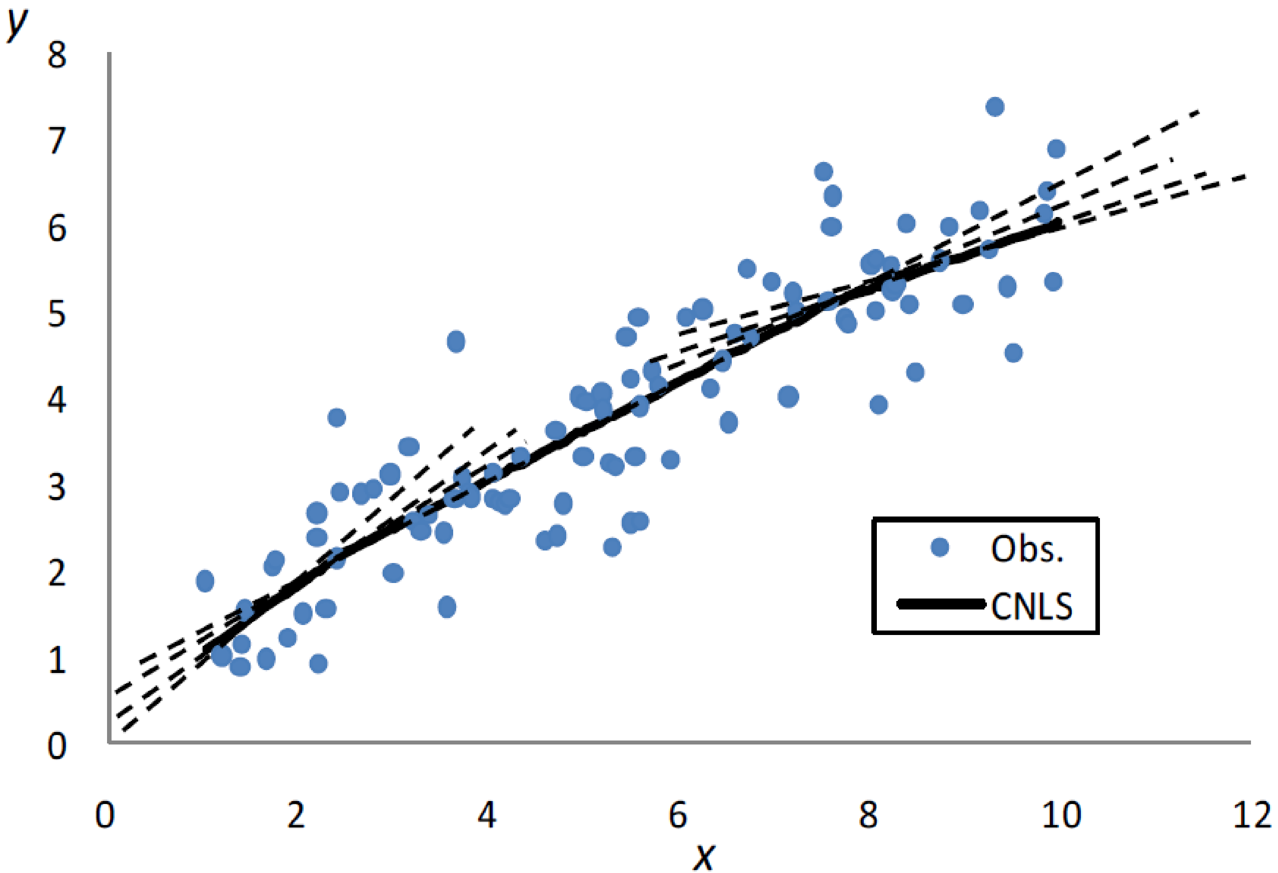
Kuosmanen and Johnson (2010)

C²NLS and DEA



Relaxed CNLS

- CNLS
  - Large-Scale Optimization Problem
  - Computational burden: 2nd constraint generate $n(n-1)$ constraints, where n is number of observations. (out-of-memory when n is large)
  - the number of hyperplanes to construct the function is generally much smaller than n.



Lee et al. (2013)

Relaxed CNLS

- Relaxed CNLS (Lee, et al, 2013)
  - Predict the relevant concavity constraints

Initial Solution

$$\min_{\alpha,\beta,\varepsilon}\left\{\sum_{i=1}^{n}\varepsilon_i^2 \;\middle|\; \begin{array}{c} y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i \quad \forall i = 1,\dots,n; \\ \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_{i+1} + \boldsymbol{\beta}_{i+1}'\mathbf{x}_i \quad \forall i = 1,\dots,n-1; \\ \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \quad \forall (i,h) \in V; \\ \boldsymbol{\beta}_i \ge 0 \quad \forall i = 1,\dots,n; \end{array}\right\}$$

1. Solve a relaxed model
2. Initial solution identification
3. Iteratively add violated “complicating” constraints
4. Stop when the optimal solution to the relaxed model is feasible

- Dantzig et al. (1954; 1959) used for solving the large-scale traveling-salesman problems (TSP)
- Average running time save around 70%

Parallel Development of Productivity Models

		Parametric	Nonparametric
Central tendency		*OLS* Cobb and Douglas (1928)	*CNLS* (Section 3) Hildreth (1954) Hanson and Pledger (1976)
Deterministic frontier	**Sign constraints**	*PP* Aigner and Chu (1968) Timmer (1971)	*DEA* (Section 4.1) Farrell (1957) Charnes et al. (1978)
	2-step estimation	*COLS* Winsten (1957) Greene (1980)	*C^2NLS* (Section 4.2) Kuosmanen and Johnson (2010)
Stochastic frontier		*SFA* Aigner et al. (1977) Meeusen and Vanden Broeck (1977)	$\boxed{\textit{StoNED}\text{ (Section 5) Kuosmanen and Kortelainen (2012)}}$

Kuosmanen et al. (2014)

Stochastic Nonparametric Envelopment of Data (StoNED)

- StoNED (Kuosmanen and Kortelainen, 2012)
  - ➢ StoNED uses a composed error term to model both inefficiency and noise without assuming a functional form and assuming only monotonicity and convavity.

- Step1: CNLS estimates $E(y_i|\mathbf{x}_i)$

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \left\{ \sum_{i=1}^{n} \varepsilon_i^2 \;\middle|\; \begin{array}{l} y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i \;\; \forall i = 1,\ldots,n; \\ \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i = 1,\ldots,n; \\ \boldsymbol{\beta}_i \ge \mathbf{0} \;\; \forall i = 1,\ldots,n \end{array} \right\} \Rightarrow \hat{f}^{CNLS}(\mathbf{x}_i) = \hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i'\mathbf{x}_i = y_i - \hat{\varepsilon}_i^{CNLS}$$

- Step2: Estimation of the expected inefficiency

- Step3: Estimating the StoNED frontier production function

- Step4: Estimating firm-specific inefficiencies

StoNED

- Step2: Estimation of the expected inefficiency
  - Apply the method of moments to the CNLS residual $\varepsilon_i^{CNLS} = v_i - u_i$ to estimate the expected value of inefficiency μ. (Aigner et al., 1977)
  - We know $\sum_{i=1}^{n} \hat{\varepsilon}_i^{CNLS} = 0$, and the second and the third central moment

$$\hat{M}_2 = \sum_{i=1}^{n} (\hat{\varepsilon}_i^{CNLS})^2 / (n-1)$$

$$\hat{M}_3 = \sum_{i=1}^{n} (\hat{\varepsilon}_i^{CNLS})^3 / (n-1)$$

  - We assume $u_i \sim N^{+}(0, \sigma_u^2)$ and $v_i \sim N(0, \sigma_v^2)$, then they are

$$M_2 = \left[\frac{\pi - 2}{\pi}\right]\sigma_u^2 + \sigma_v^2$$

$$M_3 = \left(\sqrt{\frac{2}{\pi}}\right)\left[1 - \frac{4}{\pi}\right]\sigma_u^3$$

$$\Rightarrow \quad \hat{\sigma}_u = \sqrt[3]{\frac{\hat{M}_3}{\left(\sqrt{\frac{2}{\pi}}\right)\left[1 - \frac{4}{\pi}\right]}} \quad \Rightarrow \quad \hat{\sigma}_v = \sqrt{\hat{M}_2 - \left[\frac{\pi - 2}{\pi}\right]\hat{\sigma}_u^2}$$

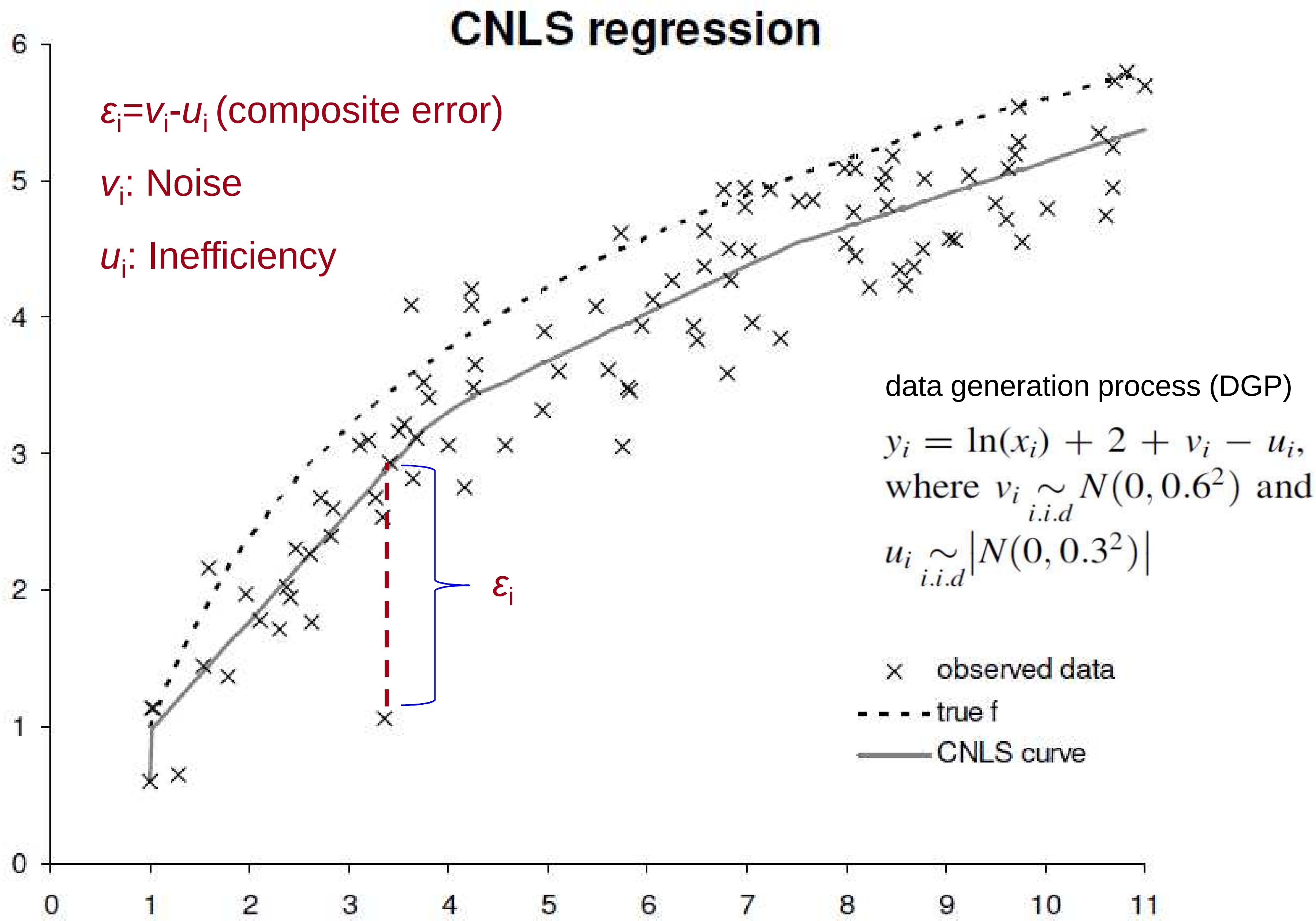
- Step3: Estimating the StoNED frontier production function
  - Shift the estimated curve upward by expected inefficiency μ.
  - Due to the multiple solutions of CNLS, estimate the minimum function (i.e., Minimum extrapolation)

$$\hat{g}_{\min}^{CNLS}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}} \left\{ \alpha + \boldsymbol{\beta}'\mathbf{x} \middle| \alpha + \boldsymbol{\beta}'\mathbf{x}_i \geq \hat{g}^{CNLS}(\mathbf{x}_i) \ \forall i = 1,\dots,n \right\}$$

  - Adjust the minimum function by adding the expected inefficiency μ to estimate the frontier using

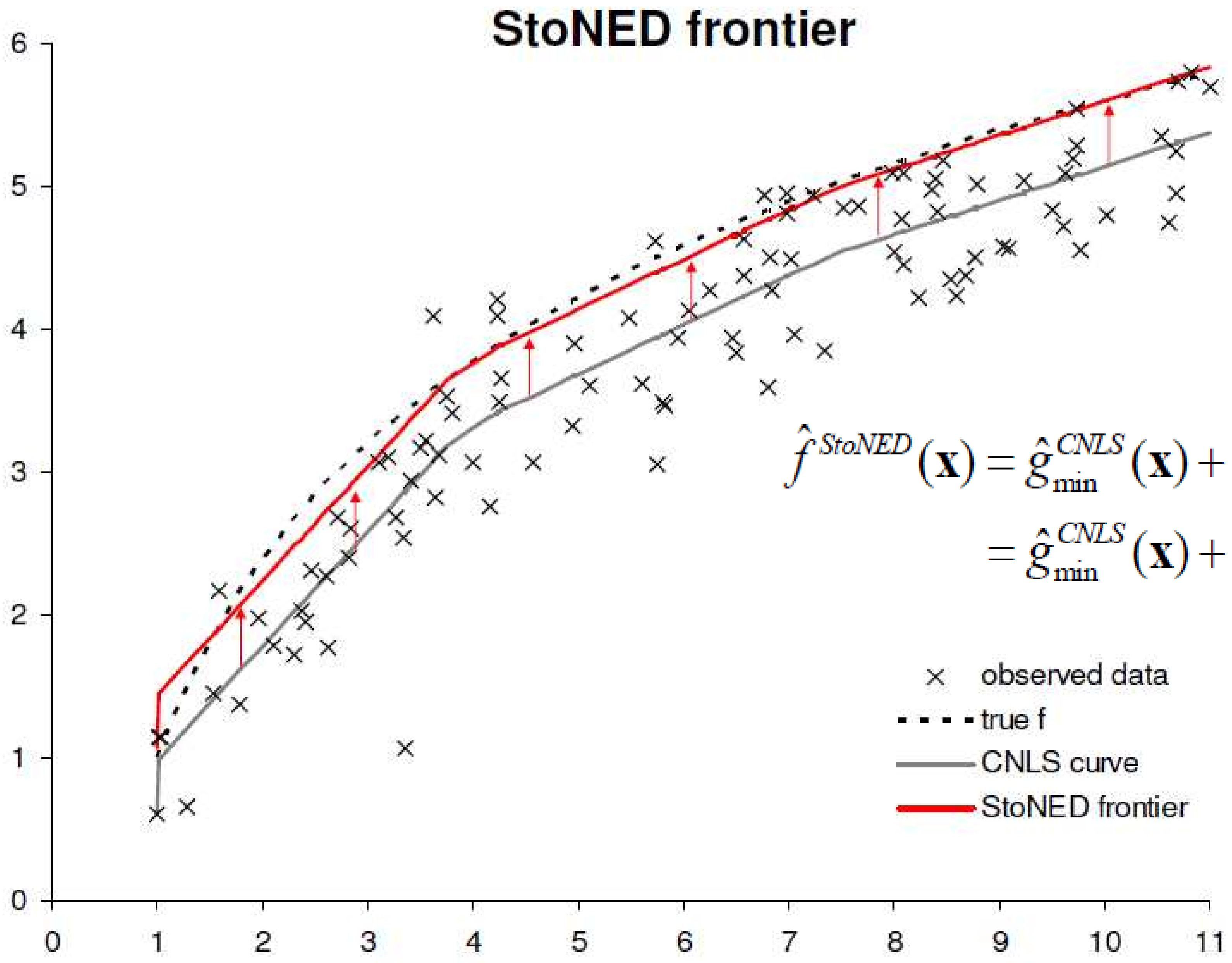
$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) + \hat{\mu} = \hat{g}_{\min}^{CNLS}(\mathbf{x}) + \hat{\sigma}_u \sqrt{2/\pi}.$$

Graphical Illustration of CNLS



Kuosmanen and Kortelainen (2012)

Graphical Illustration of StoNED



$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}^{CNLS}_{\min}(\mathbf{x}) + \hat{\mu}$$
$$= \hat{g}^{CNLS}_{\min}(\mathbf{x}) + \hat{\sigma}_u\sqrt{2/\pi}.$$

StoNED

- Step4: Estimating firm-specific inefficiencies
  - In the normal – half-normal case, Jondrow, Lovell, Materov and Schmidt (1982)

$$E(u_i \mid \hat{\varepsilon}_i) = \frac{\hat{\sigma}_u \hat{\sigma}_v}{\sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \left[\frac{\phi\left(\dfrac{\hat{\varepsilon}_i \hat{\sigma}_u}{\hat{\sigma}_v \sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \right)}{1 - \Phi\left(\dfrac{\hat{\varepsilon}_i \hat{\sigma}_u}{\hat{\sigma}_v \sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \right)} - \frac{\hat{\varepsilon}_i \hat{\sigma}_u}{\hat{\sigma}_v \sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \right]$$

$$= -\frac{\hat{\varepsilon}_i \hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_v^2} + \frac{\hat{\sigma}_u^2 \hat{\sigma}_v^2}{\hat{\sigma}_u^2 + \hat{\sigma}_v^2} \left[\frac{\phi(\hat{\varepsilon}_i / \hat{\sigma}_v^2)}{1 - \Phi(\hat{\varepsilon}_i / \hat{\sigma}_v^2)} \right]$$

where ϕ is the density function of the standard normal distribution $N(0,1)$

Φ is the corresponding cumulative distribution function.

$$\hat{\varepsilon}_i = \hat{\varepsilon}_i^{CNLS} - \hat{\sigma}_u \sqrt{2/\pi}$$

Conclusion

- StoNED
  - ➢ Kuosmanen (2006) " Combining Virtues of SFA and DEA…"
    - Consider the noise, without the specific functional form, ...etc.

- Extensions
  - ➢ Multiplicative composite error term

$$y_i = f(\mathbf{x}_i) \cdot \exp(\varepsilon_i) = f(\mathbf{x}_i) \cdot \exp(v_i - u_i)$$

  - ➢ Multiple outputs
    - DDF formulation
  - ➢ Contextual variable
    - Johnson and Kuosmanen (2011, JPA): stochastic (semi-) nonparametric envelopment of z-variables data (StoNEZD)

Thanks for your attention!

References

- Aigner, D., Lovell, C.A.K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6: 21-37.
- Hildreth, C. (1954). Point estimates of ordinates of concave functions. Journal of the American Statistical Association 49: 598-619.
- Jondrow, J., C.A.K. Lovell, I.S. Materov, and P. Schmidt (1982). On estimation of technical inefficiency in the stochastic frontier production function model. Journal of Econometrics 19, 233-238.
- Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. Econometrics Journal 11, 308-325.
- Kuosmanen, T. (2012). Stochastic Semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. Energy Economics, 34, 2189-2199.
- Kuosmanen, T., A.L. Johnson, and A. Saastamoinen, 2014. “Stochastic nonparametric approach to efficiency analysis: A unified framework”, in J. Zhu (Eds) Handbook on Data Envelopment Analysis Vol II, Springer.
- Kuosmanen, T. and M. Kortelainen (2012). Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints. Journal of Productivity Analysis 38(1), 11-28.
- Kuosmanen, T., A. Saastamoinen, and T. Sipiläinen (2013). What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. Energy Policy, 61, 740-750,
- Lee, C-Y, A.L. Johnson, E. Moreno-Centeno, and T. Kuosmanen (2013). A more efficient algorithm for convex nonparametric least squares. European Journal of Operational Research 227(2):391-400.