1 Methods

For this analysis, we used STATA's cmp command which is a system of the seemingly unrelated equations program to assess the effects of diverse exogenous hog producer and operation characteristics on the adoption of any of the five technologies, the use of computers, managerial technologies, size of production, and labor demand. The exogenous hog features are the region, education, gender, and age. As a result, we obtained a system of equations that are developed on the *classical linear regression model* with *normally distributed errors* and these errors might be correlated, sharing a multidimensional distribution [?].

Further, the seemingly unrelated regression model assumes that hog farmers maximize their expected utility (farmers pick the most profitable technology bundle), farmer's characteristics are heterogeneous (different bundles are selected due to heterogeneous farmer and farm operation characteristics), and that there is underlying dynamic process of learning that affect the adoption of these technologies (Some of these were observed, while others were not observed).

$$max\ U = E\left(\pi_i\left(Technology\ Adoption,\ \ Operation\ Size,\ Labor\ Demand \mid Production\ Characteristics\right)\right)$$

$$= E\left(\pi_i\left(\overbrace{AI,\ SSF,\ PF,\ MSP,\ EW,\ PC,\ MT}, \overbrace{NS,\ NP}, \overbrace{FTE} \mid \overbrace{Region,\ Education,\ Gender,\ Age}\right)\right)$$

We obtained a system of equations (mixed model) that are developed on: The *classical probit regression model*, interval regression model, and truncated regression model; all of them with *Normally Distributed Errors. The Seemingly Unrelated Equations* (S.U.R.) system of 10 equations are define with the following latent variables:

$$Y^*_{i,j,t} = \beta_{0,j,t} + X_{i,j,t}\beta_{j,t} + \sum_{k \epsilon F} \ell_{j,k,t}\omega_{j,k,t} + \varepsilon_{i,j,t}$$

For $j \in \{AI,\ SSF,\ PF,\ MSP,\ EW,\ PC,\ MT,\ NP,\ NS,\ FTE\} = J$ where $\mathbf{X}_{i,t}$ is a vector of control variables for survey respondent i in year t (1995 or 2005); $\beta_{j,t}$ is a conformable parameter vector; $F = common\ unobservable\ characteristics$, $\ell_{j,k,t} = factors\ loadings$; $\varepsilon_{i,j,t}$ is a mean zero error with the vector of errors $\varepsilon_{i,t}$ having the correlation matrix R_t; **J** includes the technologies artificial insemination (AI), split sex feeding (SSF), phase feeding (PF), multiple site production (MSP), early weaning (EW), personal computer (PC), and management technologies (MT); and **J** also includes firm size measures in terms of the number of pigs produced annually (NP), average sow herd side (NS), and the number of full-time employees (FTE). These are defined as latent variables because the technology variables were recorded as indicator variable equal to one if the technology was used on the farm or 0 otherwise. Alternatively, measures for the number of pigs produced annually and average sow heard size were categorical, and the measure for the number of full-time employees was truncated from above and below. Given these latent variables, the observed survey responses are

$$y_{i,j,t} = \begin{cases} 1\ for & y^*_{i,j,t} \geq 0 \\ 0 & otherwise \end{cases} \quad for\ j \in \{AI,\ SSF,\ PF,\ MSP,\ EW,\ PC,\ MT\} \tag{1}$$

Equations 3.1 are probit regression models, this model implicates two normalizations. It normalize location by setting 0 as the cut point and because it is no longer possible to determine the scale of $y^*_{i,j,t}$, the model normalizes to $\sigma^2 = 1$ [?].

Equations 3.2 and 3.3 are interval regression models for the latent variables. It generalizes the interval regression models by slicing the continuum into a finite set of ranges, each corresponding to range of number of pigs produced annually (NP) and to range of the average number of sows in an operation (NS) [?]. Assume that $y_{i,NP,t}$ and $y_{i,NS,t}$ can achieve $J = 8^{th}$ outcomes, $O_1, ..., O_8$. Use the known ascending sequences of cut points $c_1, ..., c_{J-1}$ to define the regions into which $y^*_{i,NP,t}$ and $y^*_{i,NS,t}$ fall then the link function is:

$$y_{i,NP,t} = \begin{cases} 1\ for\ 1,000 & \geq y^*_{i,NP,t} & > 0 \\ 2\ for\ 2,000 & \geq y^*_{i,NP,t} & > 1,000 \\ 3\ for\ 3,000 & \geq y^*_{i,NP,t} & > 2,000 \\ 4\ for\ 5,000 & \geq y^*_{i,NP,t} & > 3,000 \\ 5\ for\ 10,000 & \geq y^*_{i,NP,t} & > 5,000 \\ 6\ for\ 15,000 & \geq y^*_{i,NP,t} & > 10,000 \\ 7\ for\ 25,000 & \geq y^*_{i,NP,t} & > 15,000 \\ 8\ for & \geq y^*_{i,NP,t} & > 25,000 \end{cases} \tag{2}$$

$$y_{i,NS,t} = \begin{cases} 1\ for\ 0 & \geq y^*_{i,NS,t} & \\ 2\ for\ 100 & \geq y^*_{i,NS,t} & > 0 \\ 3\ for\ 200 & \geq y^*_{i,NS,t} & > 100 \\ 4\ for\ 500 & \geq y^*_{i,NS,t} & > 200 \\ 5\ for\ 1,000 & \geq y^*_{i,NS,t} & > 500 \\ 6\ for\ 2,000 & \geq y^*_{i,NS,t} & > 1,000 \\ 7\ for\ 5,000 & \geq y^*_{i,NS,t} & > 2,000 \\ 8\ for & \geq y^*_{i,NS,t} & > 5,000 \end{cases} \tag{3}$$

Equations 3.4 is a regression that was truncated from above 99 and below 0 for measure for the number of full-time employees. The model posits lower and upper truncation $99 > y^*_{i,FTE,t} > 0$. For generality, they can take the value $-\infty$ or ∞ respectively, but the likelihood must be normalized by the total probability over the observable range. In this way, we exclude observations with dependent variables outside of the range of bellow 0 and above 99 number of full-time employees then the link function is:

$$y_{i,FTE,t} = \begin{cases} 0 & for\ 0 \geq & y^*_{i,FTE,t} & \\ 99 & for & y^*_{i,FTE,t} & \geq 99 \\ y^*_{i,FTE,t} & for\ 99 > & y^*_{i,FTE,t} & > 0 \end{cases} \tag{4}$$

Stata's cmp command uses simulated maximum likelihood to obtain estimates of $\beta_{j,t}$ and R_t for $j \in \mathbf{J}$ and $t \in \{1995, 2005\}$ assuming the vector of errors $\varepsilon_{i,t}$ are distributed multi-variate normal with mean zero and correlation R_t.

The likelihood ratio test was used to compare if there were variations between the coefficients and correlation error matrices for the pooled model (years 1995 and 2005 together) versus the two years separately. Our null and alternative hypotheses are:

$$H_0 : \begin{cases} \beta_{j,1995} = \beta_{j,2005} \\ R_{1995} = R_{2005} \end{cases} \quad for\ all\ j$$

$$H_1 : \begin{cases} \beta_{j,1995} \neq \beta_{j,2005} \\ R_{1995} \neq R_{2005} \end{cases} \quad for\ all\ j$$

Table 1: Likelihood Ratio Test

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
Year 1995 & 2005 Model	4055	-44782.04	-43858.28	228	88172.57	89610.72
Year 1995 Model	3363	-34912.5	-34455.84	138	69187.69	70032.33
Year 1995 Model	692	-7709.729	-7543.798	138	15363.6	15990.06

LR chi2(48) = 3717.28; Probability > chi2 = 0.0000

Table 3.1 gives the model choice summaries; the likelihood ratio statistic yields a χ^2 value of 3717.28 with 48 degrees of freedom for a $p-value$ that is less than 0.001, so we rejected the null hypothesis meaning that there is variations between the coefficients and correlation error matrices on the separated models, between years 1995 and 2005.

After controlling the exogenous characteristics (region, education, gender, and age) used on the seemingly unrelated regression (S.U.R) model we determine complementarity relationships among the latent variables $y^*_{i,j,t}$ (hog technology, labor demand, and farm operation size). These associated patterns allow us to understand the U.S. hog industry evolution throughout the last 20 years and to what degree these complementarity relationships have developed as these technologies matured. Therefore, we run a factor analysis on the sample correlation error matrix ($\mathbf{R}$) obtained from the two S.U.R models for data interpretation to find features that are important for explaining correlation error matrix.

We used STATA's factor analysis command "factormat" using the "pf" which specifies that the principal-factor method be used to analyze the correlation matrix (are computed using the squared multiple correlations as estimates), and "pcf" which specifies that the principal-component factor method be used to analyze the correlation matrix [?].

$$\underset{(\rho\ x\ 1)}{\mathbf{X}-\mu} = \underset{(\rho\ x\ m)}{\mathbf{L}}\underset{(m\ x\ 1)}{F} + \underset{(\rho\ x\ 1)}{\varepsilon}$$

Where $\mathbf{X} = (X_1, \ldots.., X_p)'$ is a random vector with mean μ and correlation matrix $\mathbf{R}$, $\mathbf{L}$ = denotes the matrix of factor loadings (loadings of common factor), F =denotes the vector of latent factor scores (scores of common factor), and ε = denotes the vector of latent error terms (specific factor).

For the principal-factor method, we applied Factor Analysis to the sample correlation matrix $\mathbf{R}$ where $\mathbf{R} - \psi = \mathbf{LL}'$. Where $\hat{\psi}_j$ [1] are the initial of the specific variance [?].

For the principal components solution for factor analysis, the parameters of interest are the factor loadings $\mathbf{L}$ and specific variances on the diagonal of ψ for $m < p$ common factors, the PCA solution estimates $\mathbf{L}$ and ψ as L and Ψ as $\hat{L} = \left[\lambda_1^{1/2}\nu_1, \lambda_2^{1/2}\nu_2, \ldots\ldots, \lambda_m^{1/2}\nu_m\right]$ and $\hat{\psi}_j = \sigma_{jj} - \hat{h}_j^2$. Where the eigenvalue decomposition of $\Sigma = \vee \wedge \vee'$, and $\hat{h}_j^2 = \sum_{k=1}^{m} \ell_{jk}^2$ is the estimated communality of the $j-th$ variables [?].

[1] $\hat{\psi}_j = \frac{1}{r^{jj}}$ $(where\ r^{jj}\ is\ the\ diagonal\ of\ \mathbf{R^{-1}})$