Nonparametric Estimation of Crop Yield Distributions: A Panel Data Approach

Ximing Wu
Texas A&M University
xwu@tamu.edu

Yu Yvette Zhang
Texas A&M University
yzhang@tamu.edu

Selected Paper prepared for presentation at the Agricultural & Applied Economics Associations 2012 AAEA Annual Meeting, Seattle, Washington, August 12-14, 2012

Nonparametric Estimation of Crop Yield Distributions: A Panel Data Approach

Ximing Wu* Yu Yvette Zhang†

June 3, 2012

Abstract

We propose a flexible nonparametric density estimator for panel data. One possible areas of application is estimation of crop yield distributions whose data tend to be short panels from many geographical units. Taking into account the panel structure of the data can likely improve the efficiency of the estimation when the crop distributions share some common futures over time and cross-sectionally. We apply this method to estimate annual average crop yields of 99 Iowa counties. The results demonstrate the usefulness of the proposed method to estimate simultaneously densities from a large number of cross-sectional units.

1 Introduction

Distributions of crop yield are of fundamental importance to farmers, policy makers and agricultural economists owing to the central role they play in the crop insurance markets and commodity prices. There exists a large body of literature on crop yields, falling into two broad categories: parametric methods and nonparametric methods. The former postulates parametric

*Department of Agricultural Economics, Texas A&M University; email: xwu@tamu.edu

†Department of Agricultural Economics, Texas A&M University; email: yzhang@tamu.edu

families for crop yield distributions. Common choices include, among others, the normal, log-normal, gamma, and beta families. Parametric estimations are asymptotically efficient when they are correctly specified, but run the risk of substantial biases under incorrect distributional assumptions. Free from rigid distributional assumptions, nonparametric methods allow flexible functional forms guided by data driven principles. The kernel, series and spline methods are commonly employed. Although consistent, nonparametric methods are generally less efficient than parametric methods, and therefore call for larger sample sizes. Successful implementation of nonparametric estimation depends crucially on the choice of smoothing parameters, which balances the goodness-of-fit and simplicity of modeling, trading off between bias and variation.

This study proposes a novel nonparametric estimator for crop yield distributions. Recognizing the similarity of crop distributions across regions and the relatively short length of individual time series, we opt to model county level crop yield distributions simultaneously using a panel data approach. A main contribution of this study is to propose a nonparametric density estimator that handles a large number of densities simultaneously.

2 Literature

There is a large literature on the estimation of crop yield distributions. For recent work on this important topic, see, e.g., Claassen and Just (2011), Koundouri and Kourogenis (2011), Woodard and Sherrick (2011), and references therein. The statistical methods employed in these studies can be categorized into two general groups: parametric methods and nonparametric methods (for simplicity, we view semiparametric methods as nonparametric in this discussion owing to that they share the common properties of infinite number of nuisance parameters asymptotically).

Parametric methods assume known functional forms (up to a finite number of unknown parameters) for crop yield distributions. The main benefits of parametric approach includes the simplicity of estimation and inference, and asymptotic efficiency when the parametric distributional assumptions

are correct. Popular parametric distributions entertained in the literature include the normal, log-normal, Beta, Gamma,and their generalizations, among others. In the absence of theoretical guidance, choice of parametric distributions is oftentimes based on convenience and other practical considerations. Consequently, assumed functional forms do not necessarily agree with unknown crop yield distributions and estimation results can be severely biased.

Nonparametric methods offer a flexible alternative. Instead of prescribing a specific distribution, nonparametric estimations let data determine a proper functional form and use data-driven methods to control the balance between fidelity to data and complexity of the model. Common methods of nonparametric distribution/density estimation include the kernel density estimation and series density estimation. The former is a 'local' average estimation while the latter is a 'global' one, using a basis function expansion to approximate an unknown distribution. Another possibility is the local maximum likelihood estimator, which combines the parametric maximum likelihood estimation and kernel density estimation. Although flexible, nonparametric estimations are generally less efficient than parametric methods (provided they are correctly specified) and thus require larger sample sizes.

One advantage of optimal series density estimator over kernel based methods is its automatic adaptiveness to the unknown smoothness of the underlying densities. This family of estimator suffers, however, the drawback of likely negative estimates. One remedy to this drawback is the the Exponential Series Estimator (ESE), which approximates the logarithm of an unknown density with a series estimator. Transforming the approximation back to its original scale results in a density estimator. Unlike the series estimator, the series estimator is strictly positive.

Most of existing nonparametric density estimators assume independently and identically distributed data. There is a small statistical literature on density estimation of inter-temporally dependent data. In this paper, we propose a novel nonparametric density estimator that combines the strength of exponential series estimator and the penalized spline method. The main contribution of this approach is that it provides a natural framework to

handle simultaneously data from multiple distributions. This is particularly useful for the estimation of crop yield distributions, whose data typically consist of short panels of a large number of geographic units.

3 Nonparametric Panel Density Estimations

In this section, we present a nonparametric density estimation for panel data.

3.1 The estimator

To fix idea, consider for now one single random variable x defined on a bounded support $\mathcal{X}$. Let $g_j, j = 1, ..., J$, be a series of real-valued linearly independent functions defined on $\mathcal{X}$. The Exponential Series Estimator (ESE) takes the form $f(x) = \exp(c_0 + c_1 g_1(x) + \cdots + c_J g_J(x))$, where c_0 is a constant that ensures f integrates to unity. This estimator has an appealing information theoretic interpretation and can be viewed as a maximum entropy density estimation subject to known moment conditions (Jaynes, 1957). Allowing the number of moment conditions (corresponding to the set of basis functions) to increase with sample size at a proper rate renders this density estimator a nonparametric one. Barron and Sheu (1991) establishes the large sample properties of this estimator, considering the power series, trigonometric and spline basis functions.

In this study, we adopt the spline basis functions because of their flexibility and numerical stability (compared to global power series). Splines are essentially piecewise polynomials constructed to have continuous derivatives up to certain order. For instance, an sth-degree spline parameterization of the ESE is given by

$$f(x) = \exp(\sum_{j=0}^{s} \alpha_j x^j + \sum_{j=1}^{K} \beta_j (x - z_j)_+^s), \tag{1}$$

with

$$\alpha_0 = \log \int_{\mathcal{X}} \exp(-\sum_{j=1}^{s} \alpha_j x^j - \sum_{j=1}^{K} \beta_j (x - z_j)_+^s) dx,$$

where $(x)_+ = \max(0, x)$, and $\min(\mathcal{X}) \leq z_1 < \cdots < z_K \leq \max(\mathcal{X})$ are spline knots. The global future of the density is shaped by the global polynomials in the exponent, while the splines terms modify the polynomials curve locally and smoothly. Given the order of the spline, the large is the number of knots (i.e., local polynomials), the more flexible the density is. Usual choice of s is 1,2, or 3, corresponding to the liner, quadratic and cubic splines. Thanks to the lower order of global polynomials, these splines do not suffer from oscillations typically associated with higher order polynomials. For a systematic treatment of spline estimations, see, e.g., Ruppert, et al. (2003).

Like the kernel density estimation, the ESE depends crucially on the degree of smoothing. For spline-based estimation, it is known that the smoothing depends on the degree of polynomial, the number and locations of knots. Conventionally, there exists two approaches to conduct spline estimation. The smoothing spline approach uses the same number of knots as the sample size, and uses a penalty to shrink spline coefficients towards zero. In contrast, the regression spline approach is economic in terms of knot selection; usually a small number of knots are selected and their coefficients are not penalized. These two methods have their own merits and limitations. The smoothing spline estimation can be computationally expensive, especially when sample size is large. As for the regression spline, the selection of a 'significant' set of spline functions from a large candidate set can be rather difficult, especially in multivariate case.

The penalized spline estimation provides a third alternative that combines the strength of the smoothing spline and regression spline. Like the smoothing spline, this approach entails penalizing spline coefficients. On the other hand, the number of knots is smaller than the sample size but usually larger than that selected by the regression spline estimation. Ruppert et al. (2003) demonstrate some theoretical and practical advantages of this approach. In particular, they show that with a modestly large number of splines, the degree of smoothing is largely controlled by the smoothing

parameter and little affected by the degree of splines or the number and locations of the knots. Consequently, model selection is reduced to the selection of smoothing parameters, simplifying the process considerably.

Suppose $X_1, \ldots, X_n$ is an iid sample from an unknown distribution whose density f we are to estimate. Given the spline basis function in (1), we denote the log-likelihood of the ith observation by $l_i(\theta)$, where $\theta = (\alpha, \beta)$ resides in a compact parameter space Θ. Our density estimation is given by the following optimization problem:

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} l_i(\theta) - \lambda \theta^T W \theta,$$

where W is a semi-positive definite weight matrix, and the smoothing parameter λ control the overall penalty on model complicity. This objective function strives for a balance between the goodness-of-fit and simplicity. Since the objective function is strictly convex in θ, there exists an unique solution to this optimization problem.

In practice, one has to specify the smoothing parameter. A commonly used method is the cross-validation, which can be computationally expensive especially for nonlinear models as is in our case. Alternatively, one can use a quasi-Bayesian approach, treating spline coefficients as a Gaussian process. This approach has an appealing mixed effect model interpretation, in which the spline coefficients are models as conditional means of random effects. See Gu and Qiu (2003) for the selection of smoothing parameters in nonlinear spline estimations. Wand (2002) presents a mixed effects model interpretation of penalized spline estimations. We adopt this approach in this study.

3.2 Extension to panel data

In this sequence, we present an extension to the spline ESE method for panel data. Let f_u be the density of x_u, $u = 1, \ldots, N$. In order to accommodate simultaneous estimation of multiple densities, we propose a random effect

estimator. Our random effect estimator takes the form

$$f_u(x) = \exp(\sum_{j=0}^{s}(\alpha_j + a_{j,u})x^j + \sum_{j=1}^{K}(\beta_j + b_{j,u})(x - z_j)_+^s)$$
$$= f(x)\exp(\sum_{j=0}^{s} a_{j,u}x^j + \sum_{j=1}^{K} b_{j,u}(x - z_j)_+^s), \quad (2)$$

where for identification, we assume $\sum_{u=1}^{N} a_{j,u} = 0$ and $\sum_{u=1}^{N} b_{j,u} = 0$ for each j. Thus all densities share a common baseline f as given in (1), while deviation from the baseline density is captured by a multiplicative 'individual effect'.

As is discussed above, we use the mixed effects model approach to select the smoothing parameter. This approach entails the following assumptions:

$$\alpha_{j,u} \sim N(0, \sigma_a^2), \beta_j \sim N(0, \sigma_b^2), \beta_{j,u} \sim N(0, \sigma_{c,u}^2), \quad (3)$$

for $j = 1, \ldots, J$ and $u = 1, \ldots, N$. We stress that although these assumptions appear to be similar, they reflect different kind of considerations. The first assumption concerns about the standard random effects of the global polynomial coefficients, which are not penalized in the spline estimation. The second assumption concerns about the common spline coefficients across all units. They are modeled as a Gaussian process whose variance is controlled by the smoothing parameter. The third coefficient captures deviation of individual spline coefficients from the common one. Note that these assumptions can be modified to accommodate alternative specifications. For instance, we can set $\sigma_{c,1}^2 = \cdots = \sigma_{c,N}^2$ such that individual deviations in the spline coefficients share a common variance, which simplifies the computation considerably.

4 Simultaneous Estimation of Crop Yield Distributions

In this section, we apply the proposed methods to estimating a large number of crop yield distributions simultaneously. This application is mainly illustrative. We look at the county level average corn yields of Iowa, the most important corn production state. Our data cover a sample period from 1926 through 2010 and include 99 counties in Iowa.

Figure 1 plots the mean and median annual crop yields across all counties in Iowa. One can see that these two series track each other closely. There is a clearly increasing trend during this period. To account for the time trend and heterogeneity across counties, we estimate a simple model with a linear time trend and a set of county dummies. For the rest of this section, our estimations are based on the residuals from this regression.[1]

Figure 2 plots the histogram of the crop yields with the normal distribution (of the same mean and variance) super-imposed. The distribution of crop yields across all counties deviates from the normal distribution considerably; the bulk of the distribution is more concentrated than the normal and right-skewed with an extended left tail. This overall shape is commonly observed in the crop yield literature. We also examine the distributions at the county level. As one can image, deviations from the normal distribution are more pronounced at the county level for many counties. To save space, the results are not reported.

We next proceed to estimate the county level distributions simultaneously using model (2). We set $s = 2$ such that our estimator nests the normal distribution as a special case (it is equivalent to the normal distribution when all spline coefficients are zero; this can be achieved by setting the smoothing parameter to infinity). We use 10 spline basis functions, i.e., $K = 10$. We experimented with larger number of knots; the final results are

[1] In the data pre-processing stage, we experimented with more complicated methods that account for inter-temporal correlation, spatial correlation and heteroskedasticity across counties and over time. We note that these alternatives do not lead to significantly different final density estimates based on the residuals. We therefore opt to use the simple linear model in the first stage.

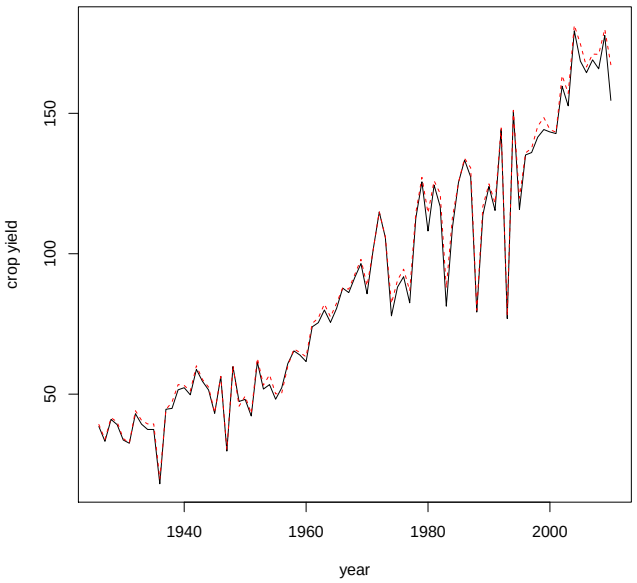

Figure 1: Historical annual crop yields of Iowa (solid: mean; dash: median)

almost identical to those from $K = 10$. Finally the smoothing parameter is selected according to the mixed effects model approach discussed above. Although our estimator has an analytical form, its coefficients cannot be solved analytically because of its nonlinearity. We use a Gauss-Jordan nonlinear optimization algorithm to solve for the problem.

We consider three estimators in our experiment. The first model is the 'pooled' model which ignores the panel structure of the data. This is obtained by setting $\sigma_a^2 = 0$ and $\sigma_{c,j}^2 = 0$ for $u = 1, \ldots, N$. The second model is the most general one, in which we allow functional deviations for each county from a common baseline. This corresponds to the smoothing assumption given in (3). Lastly we consider an intermediate case where all individual deviations share a common variance such that $\sigma_{j,1}^2 = \cdots = \sigma_{j,N}^2$. We denote these models by Model One, Two and Three respectively.

Figure 3 plots the pooled estimate, which tracks the histogram in Figure 2 closely. Figure 4 reports the panel estimate which allows individual curves

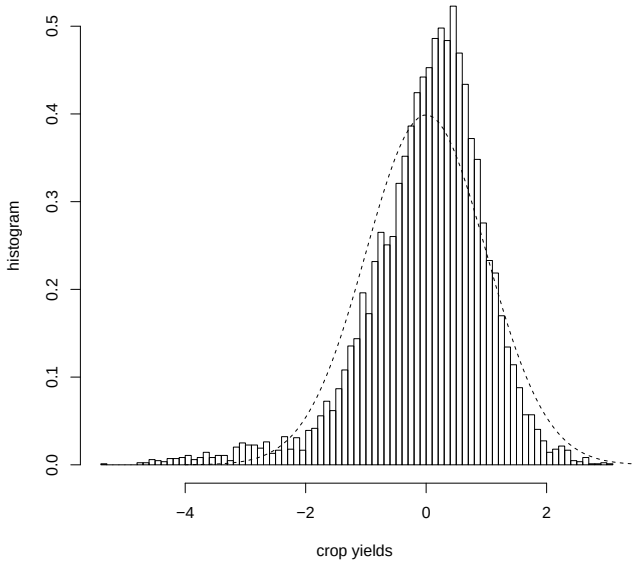

Figure 2: Histogram of crop yields with normal distribution super-imposed.

deviate from a common baseline. We observe significant variations across the counties (after detrending and demeaning in the first stage estimation). Lastly, Figure 5 reports the 'homogeneous' panel estimates, which assumes a common variance of spline coefficients across counties. It is seen that the variations across counties are subdued, but still quite considerable.

We hope the above examples illustrate the usefulness of the proposed method. We note that model (2) is flexible enough to accommodate other modifications or extensions. For instance, we can further impose restrictions on the structure of covariance among the units in our analysis to improve efficiency. We can use Bayesian method or Monte Carlo method for inference; the later is supposed to be more accurate, but at the expense of computation time. Since the estimations are conducted within the general framework of maximum likelihood estimation, one can use the likelihood ratio ratio for specification testing. Alternatively, one can use the information criterion, such as the AIC and BIC, for model selection.

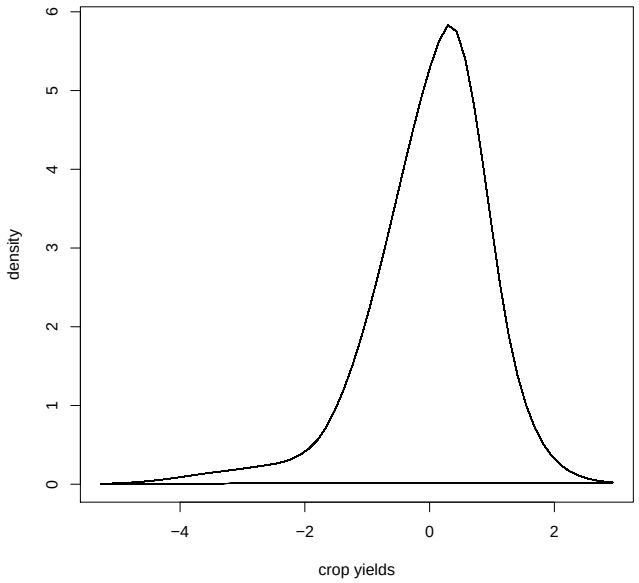

Figure 3: Estimated density: Model One (pooled estimation)

5 Concluding remarks

In this study, we propose a flexible nonparametric density estimator for panel data. We apply this method to estimate annual average crop yields of Iowa counties. The reported results demonstrate the usefulness of the proposed method to estimate simultaneously densities from a large number of cross-sectional units. One possible areas of application is estimation of crop yield distributions whose data tend to be short panels from many geographical units. Taking into account the panel structure of the data can likely improve the efficiency of the estimation when the crop distributions share some common futures over time and cross-sectionally.

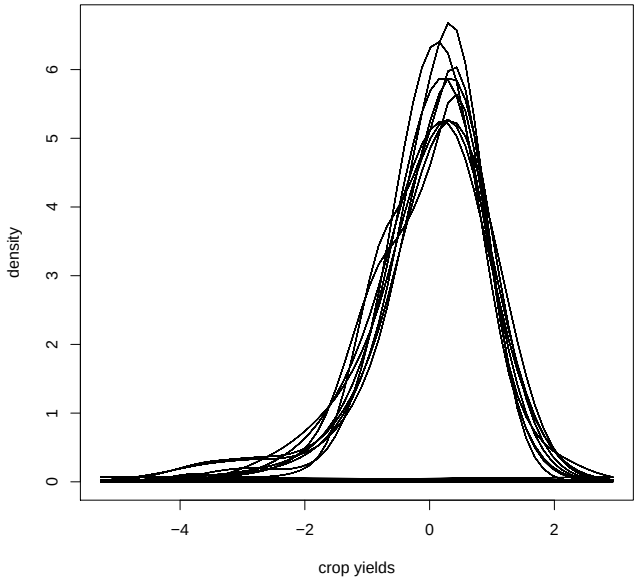

Figure 4: Estimated density: Model Two (panel model)

References

[1] Barron, A.R. and C.H. Sheu, 1991,Approximation of Density Functions by Sequences of Exponential Families, Annals of Statistics, **19**, 1347–1369.

[2] Claassen, R. and R. Just, 2011, Heterogeneity and distribuitonal form of farm-level yields, American Journal of Agricultural Economics, **93**, 144–160.

[3] Jaynes, E. E., 1957, Information Theory and Statistical Mechanics, Physical Review, **106**, 620–630.

[4] Gu, C. and C. Qiu (1993). Smoothing spline density estimation: Theory, Annals of Statistics, **21**, 217234.

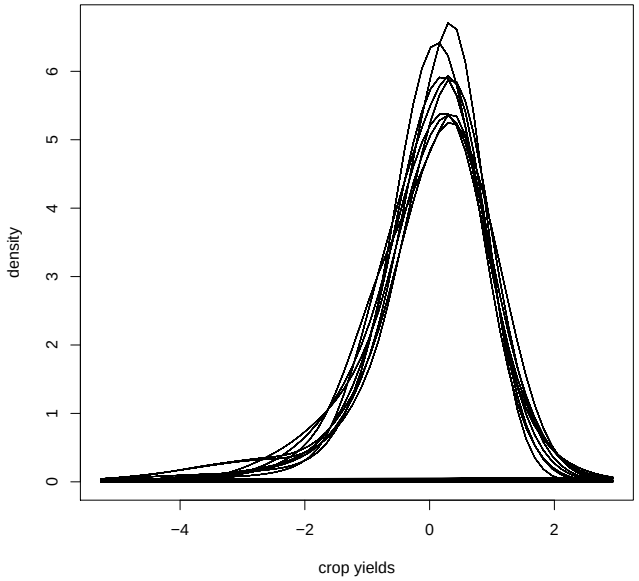

Figure 5: Estimated density: Model Three (homogeneous panel estimation)

[5] Koundouri, P. and N. Kourogenis, 2011, On the distribution of crop yields: does the central limit theorem apply? American Journal of Agricultural Economics, **93**, 1341–1357.

[6] Wand, M.P, 2002, Smoothing and mixed models. Memeo.

[7] Woodard, J. and B. Sherrick, 2011, Estimation of mixture models using cross-validation optimization: implicaions for crop yield distribution modeling, Amierican Journal of Agricultural Economics, **93**, 968–982.