



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Efficient Estimation of Copula Mixture Models:

An Application to the Rating of Crop Revenue Insurance

Somali Ghosh

Department of Agricultural Economics
Texas A&M University
2124 TAMU
College Station, TX 77843-2124
sgosh@ag.tamu.edu

Joshua D. Woodard

Department of Agricultural Economics
Texas A&M University
2124 TAMU
College Station, TX 77843-2124
jdwoodard@tamu.edu

Dmitry Vedenov

Department of Agricultural Economics
Texas A&M University
2124 TAMU
College Station, TX 77843-2124
vedenov@tamu.edu

**Selected Paper prepared for presentation at the Agricultural & Applied Economics Association's
2011 AAEA & NAREA Joint Annual Meeting, Pittsburgh, Pennsylvania, July 24-26, 2011**

**Copyright 2011 by Somali Ghosh, Joshua D. Woodard, and Dmitry Vedenov. All rights reserved.
Readers may make verbatim copies of this document for non-commercial purposes by any means,
provided that this copyright notice appears on all such copies**

Efficient Estimation of Copula Mixture Models:

An Application to the Rating of Crop Revenue Insurance

Abstract

The association between prices and yields are of paramount importance to the crop insurance programs. Proper estimation of the association is highly desirable. Copulas are one such method to measure the dependence structure. Five single parametric copulas, a non- parametric copula and their fifteen different combinations taking a mixture of two different copulas at a time have been used in the crop insurance rating analysis. Using data of corn from 1973-2009 for 602 counties in the Mid-West area two different efficient methods have been proposed to generate the optimal mixtures using the cross validation approach. A resampling technique is used to check for the significance of the expected indemnities.

Key Words: Copulas, Crop Insurance, Cross-Validation, Empirical distribution, GRIP, Indemnities,

Out-Of-Sample Log-Likelihood

JEL Code: Q1,Q14

Introduction

Crop insurance is of critical importance in the farming business to properly address production or revenue shortcomings for farmers or crop producers. Correct estimation of the revenue distribution is of paramount importance in analysis of issues in crop insurance and risk management. The producers' revenue depends on the price and yield distributions and the correlations between the prices and yields. There are different ways to account for the correlation structure between prices and yields. The copula is one such method. It is quite obvious that if we do not take into account the correlation between any two random variables (in this case price and yield) a bias would result in the expected revenue estimation. The pricing of the indemnities thus would also give improper results. Indemnities for crop insurance refer to the payment of the insurance companies to the producers whose revenue falls short of a guaranteed revenue.

Thus, the objective of this paper is to use six different copulas like kernel, Gaussian, t, frank, gumbel, and clayton and also their fifteen different mixtures to model the joint distributions of prices and yields. In order to find out the optimal mixtures two different efficient methods has been used , the minimization objective and the loglikelihood approach both in the out of sample framework.

The association between prices and yield has been there in the agriculture economics literature for a long time. However modeling joint distributions using copula has been a recent addition to the agriculture economics literature. Previously different methods were used to check for association between two random variables. Under regressibility assumption with two random variables, one variable can be expressed as a linear function of the other variable. The beta coefficient then becomes

$$y = \alpha + \beta x + \varepsilon$$
$$\beta = \rho \frac{\sigma_y}{\sigma_x}$$

The ρ is just a linear association between the variables. It does not take into account non-linearities between the two random variables. It is a poor measure of dependence. Furthermore, if the uncorrelated random shocks are assumed normal, regressibility amounts to joint multivariate normality, an assumption crop yields typically violate (Ramirez, Misra and Field, 2003).

Several papers recognize the limitations of the above procedure and go for ad hoc procedures to model the dependence structure. It involves Choleski decomposition of the matrix or in other words they involve transformations of the multivariate normal distribution with the parameters estimated from the data (Ramirez, 1997). It allows greater flexibility in modeling joint distribution but relies on the assumption that the correlation matrix contains all necessary information about the dependence structure.

A multivariate empirical distribution is also used for modeling joint distributions without imposing any distributional assumption (Deng, Barnett and Vedenov, 2007). The disadvantage with this method is that they are limited to the observed realizations and results in discontinuous density functions which could be problematic in the insurance contracts.

Copulas

Copulas are just alternative ways to model joint distributions of random variables.

The advantage of the copula method is the flexibility of choosing the marginal distributions as well as the non-linearity between the random variables is captured rather than linear dependence structure in the regression models.

A two dimensional copula distribution $C(u,v)$ is defined as a function $C: [0,1]^2 \rightarrow [0,1]$. In other words, C is a bivariate distribution function with uniformly distributed marginals u and v . $C(\cdot)$ has the following properties:

$$C(u, 0) = C(0, v) = 0 \text{ for all } u, v \in [0,1]$$

$$C(u, 1) = u \text{ and } C(1, v) = v \text{ for all } u, v \in [0,1]$$

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \forall u_1 \leq u_2, v_1 \leq v_2$$

Copulas are related to joint bivariate distributions by virtue of Sklar's Theorem. The theorem states that any joint distribution function $H(x,y)$ with margins $F(x)$ and $G(y)$ can be represented as

$$H(x,y) = C(F(x), G(y)).$$

where $C(\cdot)$ is a uniquely determined copula function. If the random variables x and y have continuous distribution functions, $F(x)$ and $F(y)$, then by probability integral transformation $u=F(x)$, $v=F(y)$ are uniformly distributed on $[0,1]$. If the distribution functions and the copula are continuous Sklar's Theorem can be stated in terms of probability densities.

$$h(x,y) = c(F(x), G(y)) * f(x) * g(y)$$

where $h(x,y) = \frac{\partial^2 H(x,y)}{\partial x \partial y}$, $f(x) = F'(x)$, $g(y) = G'(y)$, and $c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$

There are different kinds of copulas in the literature. The most common are the Gaussian copula, student t copula and three copulas from the Archimedean family.

Gaussian Copula:

$$C_{\rho}^{Gauss}(u_1, u_2) = \Phi_{\rho}(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \quad (1)$$

where Φ_{ρ} is the bivariate normal distribution correlation matrix ρ , and $\Phi^{-1}(u_i)$ is the inverse of the normal cumulative distribution function.

Students 't Copula:

$$C_{\rho}^{Student's t}(u_1, u_2) = T_{\rho, \nu}(t_{\nu}^{-1}(u_1), t_{\nu}^{-1}(u_2)) \quad (2)$$

Where $T_{\rho, \nu}$ is the bivariate student-t distribution with correlation matrix ρ and degrees of freedom ν .

$t_{\nu}^{-1}(u_i)$ is the inverse of the *Student-t* cumulative distribution function.

Archimedean Copula:

$$C(u_1, u_2) = \varphi^{-1}\left(\sum_{m=1}^2 \varphi(u_m)\right) \quad (3)$$

The three different Archimedean families are Clayton, Frank and Gumbel. Each of the different copulas has different association parameters embedded within them, that is, not necessarily they are Pearson's correlation parameter *rho*.

Elliptical copulas (Normal or Student-t) are restricted to radial symmetry and don't have a closed form. For the Archimedean family lower tail dependence is captured by the Clayton family for $\theta > 0$ (parameter) > 0 . Upper tail dependence is captured in the Gumbel family for $\theta > 1$ which due to its parameter space holds only for positive co dependence. Yet, the proper shape of some true underlying copula may not be adequately described by a single parametric copula form. Along with these parametric copulas the non- parametric copulas (e-g kernel copulas) also exist. However, non-parametric copulas have the potential to overfit to observed data, potentially resulting in lower estimation efficiency.

To date, copulas have primarily been viewed in a single copula context, and only within “in-sample” fitting frameworks. Noting that a copula is no more than a multivariate distribution with uniform marginals, the potential thus exists to mix copulas of different forms in order to reduce estimation bias. Furthermore, just as the ranking of univariate distributions are subject to “in-sample” overfitting—which can lead to erroneous conclusions regarding the ranking of distributional forms in empirical settings (e.g., Norwood, Roberts, and Lusk, 2004)—the potential also exists to cast the ranking and fitting of copulas in “out-of-sample” efficient frameworks. In their study, NRL(2004) suggested the out-of –sample log

likelihood functions (OSLL), where OSLL realizations are constructed by successively estimating the yield distribution model with holdout observation(s) and then evaluating the predicted density value at the out of sample observation(s). The method is popularly known as cross validation method (CV). The candidate distributions are then evaluated based on their log likelihood values.

With these issues in mind, this study proposes a copula mixing procedure based on out-of-sample criteria, extending the recent work of Woodard and Sherrick (2010)—which develops an “out-of-sample” efficient method for the mixing of univariate distributional forms—to the copula case. The mixing procedure is carried out by assigning weights between 0 and 1 to each of the single copulas to obtain the copula mixture. The model that results from this procedure is a mixed distribution composed of multiple underlying distributions of different classes—in this case the copula. It results in a form which is out of sample efficient and can express higher degree of flexibility than any single underlying model alone.

Thus, the novelty of this paper is to develop a copula mixing procedure and numerically derive and compare generated insurance rate distributions under several different copula forms (Gaussian, Student-t, Gumbel, Clayton, Frank, Kernel) and their mixtures in order to assess the statistical and economic significance of differences among copula alternatives. A popular group risk insurance contract is used as the basis of the application for several major crops and regions. Preliminary results suggest that the choice of copula can have a large impact on the pricing of revenue insurance in some cases for this insurance product, and that the mixing procedure shows promise for ameliorating inefficiencies endemic to simple in-sample modeling frameworks. The results from our analysis (Table 7) shows a huge difference between the rates generated from the Gumbel copula and the rates generated from the optimal mixture between Gumbel-Clayton at all levels of coverage.

Group Risk Insurance Policy (GRIP)

Crop Revenue Insurance programs that deal with county yields and revenues are called Group Risk Insurance Policy (GRIP). GRIP is revenue insurance paying indemnities when county revenue is below a

revenue guarantee. GRIP has two options: GRIP without the harvest revenue option (GRIP-NoHR) has a guarantee that will not increase while GRIP with the harvest revenue option (GRIP-HR) has a guarantee that will increase if the harvest price is above the base price .

$$GRIP-HR\ Indemnity = \text{Max}(0, \left(\frac{\text{Max}(\text{BasePrice}, \text{HarvestPrice}) - \text{Ybar}}{\text{HarvestPrice} - \text{Ybar}} \right) * 1.5) \quad (4)$$

$$E(GRIP-HR\ Indemnity) = \iint \left(\frac{\text{Max}(\text{BasePrice}, \text{HarvestPrice}) - \text{Ybar}}{\text{HarvestPrice} - \text{Ybar}} \right) * 1.5 f(p, y) dp dy \quad (5)$$

where Ybar is the expected county yield, Base Price is the average of the December futures prices in February, Harvest Price is the average of the December futures prices in October, Cov is the coverage levels, y is the county yield and the leverage factor is 1.5

To estimate the joint density $f(p, y)$ in the GRIP model the different copulas Gaussian, student- t , three different copulas of the Archimedean family, kernel copula, and their fifteen different mixtures will be used.

Optimization of a Mixed Probability Model based on Out-of-sample procedures:

The two different objectives used in the optimization procedure in this study are:

- 1) The loss minimization objective, and
- 2) The out of sample log likelihood approach.

For optimal mixing of the copulas the weights between 0 and 1 will be assigned to the expected indemnities generated from the different parametric copulas and the kernel copula. The mixture combinations can be carried out to different component copula models. The optimization objective is a loss minimization function and is defined as $C_{mix} = \sum_i (r_{-i,mix} - r_i)^2$ where $r_{-i,mix}$ is the out of sample expected indemnity with hold out observation for model mix (mixture of two copula models). r_i is the actual indemnity for that year. It should be noted that this does not depend on any model. Hence the optimal set of weights, w^* , can be obtained as the solution to,

$$W^* = \arg_w \min(C_{mix})$$

The optimal w^* will then be plugged back into the copula densities and the optimal copula mixture will be generated. The optimal copula mixtures will then be used to obtain the insurance rates.

The out of sample log likelihood method for optimal mixing of two or more copula models, given sample data P and Y is carried out by optimizing the “leave-one-out” OSLL CV criterion.

Let $W = \{w_1, w_2, \dots, w_s : \sum_{i \in K} w_i = 1; w_i \geq 0 \forall i\}$ be the K component weights and

$f_{mix}(x|W, M, \eta) = \sum_{k \in K} (w_k \cdot f_k(x, \theta_k^*(\eta)))$ be the mixture distribution associated with weights W, the set

of candidate mixture model, M, and η , the data which contains both the prices and yields, x is the point where the density needs to be predicted. The out of sample likelihood function of f_{mix} is then

$$L_{mix}^{out}(W, M, \eta) = \prod_{i=1}^N [\sum_{k \in K} (w_k \cdot f_{k,i}(p_i y_i, \theta_{k,i}^*))].$$

To truly represent the out of sample measures, the

parameters must be functions of y_{-i} and not y_i . The above conditions suggest a well defined objective

that can be optimized to obtain the optimal mixing weights of models in f_{mix} . Therefore, the optimal w^* is obtained by maximizing the out of sample likelihood function,

$$W^*(M, \eta) = \arg_w \max(L_{mix}^{out}(W, M, \eta)). \quad (6)$$

The optimization of equation (6) is straightforward but is intensive. For many applications involving likelihood functions, it is more convenient to work in terms of the natural logarithm of the likelihood function, called the log-likelihood, than in terms of the likelihood function itself. Since the logarithm is a monotonically increasing function, the logarithm of a function achieves its maximum value at the same points as the function itself, and hence the log-likelihood can be used in place of the likelihood in maximum likelihood estimation and related techniques.

The last stage involves the employment of the full sample data P , and Y with optimal weights w^* to arrive at the final mixture model *pdf*

$$f_{mix}^*(x | w^*, M, \eta) = \sum_{k \in K} (w_k^* \cdot f_k(x, \theta_k^*(\eta))). \quad (7)$$

Data

The corn yield data from 1975-2009 has been used in this study. For corn yield, data has been collected from Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, Nebraska, and Wisconsin for 602 counties. The source of the data is from the USDA National Agricultural Statistics Service. The price for corn is a percentage change between harvest price and base price. The base price is the average of the December futures prices in February and the harvest price for corn is the average of the December futures prices in October. The prices are used in the Crop Revenue Coverage (CRC) and the Group Risk Income prices (GRIP) under Risk Management Agency (RMA) rules.

A feature of corn yield is that they have increased through time due to technological gains. To account for this technological change the yields have been detrended using a linear regression on time trend (Sherrick et al, 2004). This study uses a linear trend form at the county levels with homoscedasticity around the trend. Implicitly, this assumption assumes that yields are distributed independently across time. This homoscedasticity assumption has been supported for Mid-West corn yields in previous researches (Woodard et al, 2008; Yu and Babcock, 2009).

Empirical Approach

The estimation procedure consists of denoting the distribution of corn yield and prices. For the corn yields an **Empirical** distribution is assumed because of its ability to fit the data well. The distribution captures the skewness and kurtosis of the distribution of yields as yields do not always follow a normal distribution according to several literatures (Goodwin and Ker, 2000; Just and Weninger, 1999; Nelson and Preckel, 1989 to name a few). The prices are assumed to follow a **Lognormal** distribution with mean and variance of the distribution as the given base price and price volatility respectively. Since prices cannot be negative, a lognormal distribution is assumed. The worst case can be a price equal to zero which would give a non zero value to the lognormal distribution. Besides it is more skewed to the right suggesting more weight is given to the lower prices. However, it should be remembered that we are focusing on the copula estimation impacts on the rates independent of the marginals we specify for yields and prices. To carry forward the analysis, the rates are calculated using the historically available data of corn. The optimal mixing of two different copulas is carried out by optimizing the “leave-one-out” cross validation (CV) criteria using two different objectives mentioned previously. There are other CV criteria like leave-more-out (Zhang, 1992). But this criteria is a straight forward extension of the more complex ones.

For each county, the parametric copulas are then calibrated using *Canonical Maximum Likelihood Method* (CML). The CML method does not imply any *a priori* assumption on the distributional form of the

marginals and uses the empirical distribution for each of the n variables to convert each of the observed data $X_{T \times N}$ into uniform variates, $\widehat{u}_{n,t}$. The CML method is implemented with a two step procedure:

1. Transformation of the initial dataset $X = (X_{1t}, X_{2t}, \dots, X_{Nt})$ where $t = 1, 2, \dots, T$ into uniform variates, using the empirical marginal distribution. Thus we can write

$$\widehat{u}_t = (\widehat{u}_1^t, \widehat{u}_2^t, \dots, \widehat{u}_N^t) = [\widehat{F}_1(X_{1t}), \widehat{F}_2(X_{2t}), \dots, \widehat{F}_N(X_{Nt})] \text{ where } F \text{ is the marginal distribution.}$$

2. The copula parameters α can be estimated via the following relation,

$$\widehat{\alpha}_{CML} = \arg \max_{\alpha} \sum_{t=1}^T \ln c(\widehat{u}_1^t, \dots, \widehat{u}_N^t; \alpha).$$

In this paper six different copulas have been used. The Gaussian Copula, student's t -copula, and the Archimedean copula from three different families, the Frank family, Clayton Family and the Gumbel family and the non-parametric kernel copula. After estimating the parameters of the copula we generate 5,000 random numbers from the five different parametric copulas and the kernel copula. For the Kernel copula we use the normal kernel and rule of thumb for bandwidth selection. Different copulas have different imbedded dependence structures. So the parameter structures will be different. Hence we have correlated uniforms for each copula type. Using the inverse transformation of the specified marginals for corn prices and yields we get the simulated prices and yields for each copula.

Since we are considering county revenue we consider GRIP. GRIP pays an indemnity when county revenue is less than the guaranteed revenue. County revenue is calculated using county yield times

harvest price. The Indemnity payment of GRIP-HR can be expressed as ,

$$GRIP-HR\ Indemnity = \text{Max}(0, (\text{Max}(\text{BasePrice}, \text{HarvestPrice}) * \text{APH} - \text{HarvestPrice} * y / \text{Cov}) * 1.5.$$

The base price is the expected value of price which is given to be 4. The harvest price is the simulated price from the lognormal distribution. APH is the mean yield of each county. Four different coverage levels are considered in this study. They are at the levels of .65, .75, .85, and .90. The leverage factor is given to be 1.5. The expected indemnities are estimated using different copulas using the “leave-one-out” CV criteria.

For the out of sample loglikelihood method the copula probability density of each copula model was calculated at the holdout observation. The parameters of each copula were calculated using the rest thirty four observations. Then those parameters were used to calculate the density at the left over observation or the holdout observation. The copula densities are then summed over all the thirty five observations and stored separately after converting them into log likelihoods. In this way log likelihoods are calculated for the six different copula models. For the restrictions on the parameters of the Gumbel and Clayton copula, the Kendall’s rank correlation has to be always positive. To account for this restriction perfectly negatively correlated yields were generated to allow for a positive co dependence between prices and yields. The next step was to calculate the optimal weights. To maintain tractability, the combination models were restricted to two. The optimal weights were then calculated using the stored log likelihood values of the six copulas. Fifteen different models were created using those six copulas. The optimal weights were then plugged back into the mixture density to get to the optimal mixtures.

Results and Analysis:

The single copula distribution models are assessed first to compare the performances of the two different objectives in the out of sample optimization procedure. Results of the minimization and the log likelihood objectives are presented in Tables 1 and 2 respectively.

For the minimization case the four coverage levels are considered. Table 1 shows the frequency with which each copula distribution appeared. At the 65% coverage level Clayton copula ranked best based on the frequency with which it is appearing. The next best model is the t-copula model followed by Kernel, Frank, Gumbel and Gaussian. With the increase in the coverage levels, the family of Archimedean copulas explain the data well. It should be noted that the Gaussian copula is ranked the lowest in all coverages and the clayton copula is the best in all coverages emphasizing the fact about the tail dependences of the associated variables. Table 2 represents the best ranking model based on out of sample loglikelihood values. The two different objective functions produce different results. Based on the likelihood values the Gumbel copula is the best ranking model, followed by Gaussian, t, Clayton, Frank, and Kernel.

Tables 3 and 4 shows the optimal mixture models and the average weights derived from the loss minimization objective respectively. Table 3 all four panels shows the frequency with which the fifteen different optimal mixture models showed up at each coverage level. Table 3 panel 1 shows that at 65% coverage level the mixture of kernel- clayton appeared most of the time

(about 12%) followed by frank-clayton combination and t-clayton combination. In other words, the kernel-clayton mixture is the best optimal mixture at 65% coverage level in terms of the frequency with which it appears in the counties. Panels 2, 3, and 4 show that at 75%, 85%, and 90% the Gumbel – Clayton mixture is the best. From Table 4 we see that at 65% coverage level the Clayton copula had the maximum average weight with all single copula distribution. Clayton had 68% of the weight share with Kernel copula, 67% with Gumbel, 64% with t, 60% with Frank and 57% with Gaussian. Hence it is clear

that any optimal mixture with Clayton is the best model in terms of frequency distribution. It is also the case with other coverage levels where any mixture with clayton as a distribution is the best ranking model.

Tables 5 and 6 represents the best mixture models in terms of frequency with which it appeared and the average optimal weights respectively from the loglikelihood approach.

Table 5 ranks the best model in the mixture cases in terms of frequency with which the mixture models showed up in the counties. The choice of the best model is based on the loglikelihood values. The gumbel–clayton mixture is the best mixture model in terms of appearances in 39% of the counties followed by gaussian-gumbel with 13% of the counties. In the loglikelihood approach the optimal weights associated with gumbel are more relative to all other distributions. The worst performing model among the mixtures are the kernel mixtures with other distributions. The entire weights are taken by the single distributions mixed with kernel.

Table 7, panel 1 provides the average rates across counties relative to the best optimal mix, Gumbel-Clayton which is the best ranking model in both the objective functions. Panel 2 presents the root mean square error of rates for each copula generated rate relative to best optimal mixture model. At 90% coverage levels the average rate generated from Gumbel is 130.57 compared to 122.91 of the Gumbel-Clayton mixture. The rates generated from kernel, Gaussian, t, frank, gumbel are comparatively higher than the rate generated from the best optimal mixture.

Panel 2 table 7 represents the RMSE's. The results indicate that the potential efficiency gains are quite large for all individual models when compared with the best model. For example the rates generated from the gumbel copula tend to exhibit largest divergence between it and the mixture model at all coverage levels.

Jackknife re-sampling method for rating analysis

It is reasonable to investigate the differences in alternative distributional representations with the optimal mixture representation. To check whether the expected indemnities generated through different single copulas and the mixture copulas do generate similar expected indemnities or not, we used the jackknife resampling method.

The best optimal mixture model for both the minimization objective and the loglikelihood objective is the gumbel-clayton mixture. However, in addition we also used frank-clayton mixture as it was the second best model in the loss minimization objective. We used these two mixtures with the six copula distributions to check for the significance of the expected indemnities at all four coverage levels.

Resampling is a computationally intensive statistical technique in which multiple new samples are drawn (generated) from the data sample or from the population inferred by the data sample. The Jackknife procedure was used as a resampling method. For the Jackknife resampling procedure we generated the expected rates from different copulas using the corn yield and price data sample and calculated the standard deviation of the estimates to check for the significance of the expected indemnities.

Let θ be the estimates of expected indemnity function under Kernal, Gaussian, t, Frank, Gumbel, and Clayton, frank-clayton, gumbel-clayton respectively. After calculating the estimates we tested the equality of the true indemnity function under two different copulas. For instance, if μ_G , and μ_{FC} stand for true expected indemnity under Gaussian and frank-clayton copula then we tested

$H_0: \mu_G = \mu_{FC}$ against $H_A: \mu_G \neq \mu_{FC}$. For this purpose we used the jackknife re-sampling method to calculate the standard deviation of the estimates, and the jackknife method is described below.

Let $\hat{\theta}_G^{-i}$ and $\hat{\theta}_{FC}^{-i}$ be the estimates of the expected indemnity function under Gaussian and frank-clayton copula respectively, based on the dataset after removing the i^{th} observation, $i=1,2,\dots,n$. In our dataset $n=35$.

Again let, $\hat{\theta} = \hat{\theta}_G - \hat{\theta}_{Fc}$. and $\hat{\theta}_{-i} = \hat{\theta}_{-i}^G - \hat{\theta}_{-i}^{Fc} = d_i$. Hence we have $\hat{\theta}(\cdot) = \frac{1}{n} \sum_{i=1}^N \{\hat{\theta}_{-i}^G - \hat{\theta}_{-i}^{Fc}\}$.

The variance of θ can be written as, $V(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{-i}^G - \hat{\theta}_{-i}^{Fc})$.

Table 8 all the four panels shows the frequency with which the expected indemnities generated through different copulas are significant at 5% significance level in all coverage levels. Panel 1 shows that at 65% coverage levels, 28% of the counties showed significant differences in expected indemnities generated through kernel copula and the optimal mixture of Frank-Clayton. 30% of the counties showed significant differences between kernel and Gumbel-Clayton mixture. Table 8 panel, 2,3, and 4 shows the different frequencies with which the individual copula generated rates and the two best optimal mixture rates- the Gumbel –Clayton mixture and the Frank-Clayton mixture are significant across counties at the 5% significance levels.

Conclusion

The main purpose of this study is to propose a framework for optimal model mixing in a cross- validation context. Copulas are comparatively new in the agriculture economics literature. Although new, copulas have been used quite vividly in the insurance rating literature. But the mixing of copulas for efficiency gains in the rating process is relatively new. Using two objective functions in the optimization process for optimal mixing weights of copulas in an out of sample framework allows for defining and designing specifications that are both efficient and flexible compared to the single copula distributions.

The optimal mixture models from the two different objectives used in this paper indicate that the mixture between the Archimedean families rank best. Even in the single distribution cases the elliptical copulas perform poorly compared to the copulas like clayton, gumbel which shows much dependence in the tails. All these indicate the importance of having an efficient association between the price and yield variables used to generate expected indemnities. The other important thing noted in this study is the poor performance of the non-parametric methods in the log likelihood objective function used in the optimization procedure.

The jackknife resampling method indicated that at least in some of the counties the expected indemnities generated with different copulas did generate different indemnities, that is, they are significant at the 5% significance level. There are different types of copulas present in the literature. Hence there is the possibility of combining different copulas other than these six types for optimal mixing which in turn can show better performances in the rate generation. Also here the optimal mixing is restricted to combinations of two for computational simplicity. But there is always a possibility of mixing more than two copulas.

References

- Deng, X., Barnett, B.J., and Vedenov, D.V. "Is There a Viable Market for Area-Based Crop Insurance?" *American Journal of Agricultural Economics*, **89**, 2(2007): 508-19.
- Just, R.E., and Q. Weninger. "Are Crop Yields Normally Distributed?" *American Journal of Agricultural Economics*, 81(2):287–304 (1999).
- Ker, A.P., and B.K. Goodwin, "Nonparametric Estimation of Crop Insurance Rates Revisited", *American Journal of Agricultural Economics* 82 (2): 463-478 (2000). Norwood, B., Roberts, M., and Lusk, J. "Ranking crop yield models using out-of-sample likelihood functions." *American Journal of Agricultural Economics*, 86(4):1021-1043
- Nelson C.H., and P. Preckel. "The Conditional Beta Distribution as a Stochastic Production Function." *American Journal of Agricultural Economics*, 71(2):370–378 (1989). 3 (2004).
- Ramírez, O.A., "Estimation and Use of a Multivariate Parametric Model for Simulating Heteroskedastic, Correlated, Nonnormal Random Variables: The Case of Corn Belt Corn, Soybean, and Wheat Yields." *American Journal of Agricultural Economics*, 79(1):191-205 (1997).

- Ramirez, O.A., Misra, S., and Field, J. "Crop-Yield Distributions Revisited." *American Journal of Agricultural Economics*, **85**, 1(2003): 108-20.
- Sherrick, B.J., Zanini, F.C., Schnitkey, G.D., and Irwin, S.H. "Crop insurance valuation under alternative yield distributions." *American Journal of Agricultural Economics*, 86(2):406-419 (2004).
- Vedenov, D., Application of copulas to estimation of joint crop yield distributions, *American Agricultural Economics Association Annual Meeting*, Orlando, FL, July 27-29, 2008.
- Woodard, J.D., Sherrick, B.J., and Schnitkey, G.D., "Revenue Risk Reduction Impacts of Crop Insurance in a Multi Crop Framework." *Applied Commodity Price Analysis, Forecasting and Market Risk Management NCCC-134*
- Woodard, J.D., B.J. Sherrick, "Estimation of Mixture Models using Cross-Validation Optimization: Implications for Crop Yield Distribution Modeling", *Working Paper* (2010).
- Yu, T., B.A. Babcock, "Are U.S. Corn and Soybeans Becoming More Drought Tolerant?" Iowa State University CARD Working Paper, 09-WP 500 (October, 2009).
- Zhang, P, "Model Selection via Multifold Cross Validation." *Annals of Statistics* 21(1993):299–313.

Appendix

Table 1 Frequency Model Ranked Best (Individual Copulas)

Model	Coverage			
	65%	75%	85%	90%
Kernel	0.15	0.15	0.14	0.15
Gaussian	0.12	0.09	0.09	0.07
t	0.19	0.13	0.11	0.09
Frank	0.12	0.21	0.2	0.16
Gumbel	0.12	0.06	0.1	0.21
Clayton	0.29	0.36	0.35	0.33

*Table presents frequency(or percentages) with which the rates generated with different individual copulas ranked best.

Table 2 Frequency Model Ranked Best based on OSLL values

Model	Kernel	Gaussian	t	Frank	Gumbel	Clayton
—		0.19	0.13	0.09	0.47	0.12

*Table presents frequency with which the rates generated with different individual copulas ranked best based on OSLL values

Table 3 Frequency Model Ranked Best-Loss Minimization Case**Cov.Lvl: 65%**

	Kernel	Gaussian	t	Frank	Gumbel	Clayton
Kernel	–	4.31%	5.98%	4.48%	2.49%	12.90%
Gaussian		–	4.15%	3.48%	3.82%	6.81%
t			–	7.31%	8.13%	10.29%
Frank				–	6.31%	10.79%
Gumbel					–	8.63%
Clayton						–

Cov.Lvl: 75%

	Kernel	Gaussian	t	Frank	Gumbel	Clayton
Kernel	–	4.15%	6.14%	6.31%	3.48%	12.79%
Gaussian		–	3.98%	4.98%	2.65%	4.81%
t			–	7.81%	1.87%	8.14%
Frank				–	4.98%	13.12%
Gumbel					–	14.78%
Clayton						–

Cov.Lvl: 85%

	Kernel	Gaussian	t	Frank	Gumbel	Clayton
Kernel	–	4.65%	4.31%	5.15%	4.48%	9.63%
Gaussian		–	4.98%	3.48%	1.49%	5.15%
t			–	5.81%	1.99%	6.81%
Frank				–	5.81%	12.62%
Gumbel					–	23.58%
Clayton						–

Cov.Lvl: 90%

	Kernel	Gaussian	t	Frank	Gumbel	Clayton
Kernel	–	4.48%	4.81%	5.31%	5.32%	6.97%
Gaussian		–	3.65%	2.65%	3.15%	5.15%
t			–	3.82%	1.82%	7.31%
Frank				–	7.14%	9.80%
Gumbel					–	28.57%
Clayton						–

*Table 3 presents the frequency with which the mixture models based on loss minimization objective ranked best across counties

Table 4 Average Optimal Weights for the Mixtures-Loss Minimization Objective

Coverage level=.65

	Kernel	Gaussian	t	Frank	Gumbel	Clayton
Kernel	–					
Gaussian	0.59	–				
t	0.49	0.44	–			
Frank	0.64	0.54	0.65	–		
Gumbel	0.47	0.27	0.4	0.32	–	
Clayton	0.68	0.57	0.64	0.6	0.67	–

*Weights are for the distribution in each row. The distribution in the column is the companion distribution in the mixture model

Table 5 Frequency Model Ranked Best based on OSLL

	Kernel	Gaussian	T	Frank	Gumbel	Clayton
Kernel	–	–	–	–	–	0.03%
Gaussian			8.97%	2.49%	13.95%	10.29%
t				0.06%	3.65%	7.47%
Frank					2.90%	9.40%
Gumbel						39.70%
Clayton						

*Table presents frequency with which the different individual copulas forming mixture models ranked best based on OSLL values

Table 6 Average Optimal Weights for Mixture Distributions-OSLL

	Gaussian	t	Frank	Gumbel	Clayton
Gaussian	—				
t	0.25	—			
Frank	0.19	0.51	—		
Gumbel	0.6	0.69	0.77	—	
Clayton	0.21	0.26	0.26	0.21	—

*Kernal mixture has been dropped because all the weights of kernel are equal to zero

*Weights are for the distribution in each row. The distribution
in the column is the companion distribution in the mixture model

Table7 Estimated Insurance Rates, Average Across Counties

Cov.Lvl	Kernel	Gaussian	T	Frank	Gumbel	Clayton	Optmix
65%	32.68	31.64	32.61	31.58	33.44	30.88	30.52
75%	59.41	59.19	59.39	57.98	75.39	57.27	57.83
85%	98.54	99.89	99.792	97.59	110.45	96.73	97.71
90%	122.91	125.12	125.19	122.61	130.57	121.23	122.91

RMSE Relative to OptimalMix

Cov.Lvl	Kernel	Gaussian	T	Frank	Gumbel	Clayton	Optmix
65%	3.43	2.31	3.92	1.92	4.67	1.43	
75%	3.73	3.23	3.98	2.62	20.8	1.9	
85%	3.58	3.81	3.92	2.81	15.7	2.59	
90%	3.66	3.97	4.1	3.2	10.29	3.39	

*table represents estimated insurance rates for each copula distribution
averaged across all counties in the sample as well as root mean square error
of each model relative to the optimal mixture

Table 8 Frequency of Significance between Individual Copulas and Optimal Mixtures

Cov.Lvl	65%	
	Frank-Clayton	Gumbel-Clayton
Kernel	28.24%	30.89%
Gaussian	3.65%	3.98%
T	22.75%	24.41%
Frank	2.65%	4.48%
Gumbel	27.91%	29.24%
Clayton	3.32%	6.81%
Cov.Lvl	75%	
	Frank-Clayton	Gumbel-Clayton
Kernel	20.27%	18.10%
Gaussian	7.14%	8.47%
T	11.46%	15.94%
Frank	1.66%	5.31%
Gumbel	83.88%	82.55%
Clayton	2.15%	18.93%
Cov.Lvl	85%	
	Frank-Clayton	Gumbel-Clayton
Kernel	6.15%	8.97%
Gaussian	8.88%	7.14%
T	6.81%	7.97%
Frank	0.16%	2.99%
Gumbel	76.25%	74.41%
Clayton	0%	17.77%
Cov.Lvl	90%	
	Frank-Clayton	Gumbel-Clayton
Kernel	2.65%	4.65%
Gaussian	9.13%	5.64%
T	8.63%	6.16%
Frank	0.16%	3.98%
Gumbel	57.47%	55.81%
Clayton	0%	20.43%
<p>* Table represents the frequency with which the individual copula generated rates and the optimal mixture rate is significant at the 5% significance level at all coverages.</p>		