A closer look at stochastic frontier analysis in economics

Hung T. Nguyen
Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico, USA

Received 5 July 2020
Revised 7 July 2020
Accepted 8 July 2020

Abstract

Purpose – While there exist many surveys on the use stochastic frontier analysis (SFA), many important issues and techniques in SFA were not well elaborated in the previous surveys, namely, regular models, copula modeling, nonparametric estimation by Grenander's method of sieves, empirical likelihood and causality issues in SFA using regression discontinuity design (RDD) (sharp and fuzzy RDD). The purpose of this paper is to encourage more research in these directions.

Design/methodology/approach – A literature survey.

Findings – While there are many useful applications of SFA to econometrics, there are also many important open problems.

Originality/value – This is the first survey of SFA in econometrics that emphasizes important issues and techniques such as copulas.

Keywords Copulas, Fuzzy regression discontinuity, Empirical likelihood, Production efficiency, Regular models, Regularized regression

Paper type Research paper

1. Introduction

This paper is viewed as an addition to existing surveys on the state-of-the-art of stochastic frontier analysis (SFA) in econometrics, in the sense that it spells out and emphasizes not well-known research methodologies which could be useful for advancing the topic toward more credible and efficient results. Accepting that "all models are wrong, but some are useful," the stochastic frontier models (SFM) in microeconomics are useful models to investigate production efficiency and are interesting regression models among other types of regression models in econometrics. They present a typical example for developing advanced statistical methods to make results coming out from it rigorous and trusted. Without getting into the general setting of partial identification, we look at current efforts to handle point identification and estimation problems of SFM. The first lesson to learn in proposing models in empirical research is that we need to be careful with their validity and that should be based upon theories. The actual accepted SFM came out from the fact that a naive model is not regular, so that, among other things, the standard maximum likelihood estimation (MLE) method cannot be used. Afterward, when trying to make the statistical analysis more credible, by relaxing parametric assumptions, but still having the MLE

JEL classification – B21, B23, B41, C1, C5, C13, E23

method in mind, it seems useful to point out the sieves MLE method for, say, semi-nonparametric estimation in SFM. The SFM is ideal for introducing copulas into statistical modeling. While SFA is somewhat fully explored in the context of association inference, it is about time to move to causal inference, especially using observational data from regression discontinuity design (sharp and fuzzy RDD).

The paper is structured as follows. We emphasize in Section 2 the need to check regularity of proposed models before proceeding to statistical analysis, with an example from survival analysis. In Section 3, we elaborate on copula modeling in SFA. In Section 4, we call researchers to pay attention to a not very well known nonparametric estimation method (MLE based) which should be used in SFM to render SFA more credible. We illustrate Grenander's method of sieves with an example of identification and estimation of an infinitely dimensional parameter in a diffusion model. Finally, in Section 5, we touch upon causal inference in SFA with emphasis on using RDD.

2. Generalities on stochastic frontier analysis

We take a closer look at a typical topic in economics which seems to exhibit most of the "statistical concerns" that we focus in this paper to move forward to a credible and efficient econometrics.

The background on stochastic frontier analysis (SFA), up to 2000, is summarized in (Kumbhakar and Knox Lovell, 2000). We retrace the road leading to it to illustrate how to develop better statistical methods to study an economic issue.

We are concerned with the efficiency of firms' production in microeconomics. For that, we need to be able to measure it or at least estimate it. The quantity (or "parameter") of interest is the efficiency of a firm (when it uses its technology to produce outputs from available inputs). While the notion of "efficiency" is understood in common sense, it is fuzzy in nature. However, it is possible to quantify it (or in "technical terms," to defuzzify it) for application purposes. When viewing production efficiency as "degrees of success" (of producers), we are concerned with carrying out econometric analysis to estimate them.

Starting with (Cobb and Douglas, 1928), the economic analysis proceeded as follows. First, we formulate the problem to be investigated. For a given technology, a firm tries to produce a maximum output y from an available input x. Such a production process is expressed as a production function $\varphi(x) = y$. In fact, this production function could depend on some unknown factor θ (a vector of technological parameters), so that, in a simple form, we consider $y = \varphi(\theta, x)$, i.e. we consider a parametric form of a production function. Then, our task is to identify θ and estimate it, from which we could infer the technology efficiency of interest.

Remark. You could realize right away that this is a "traditional" modeling process, namely, "regression" practice!

To estimate θ, we need data. Let $I = \{1, 2,\ldots,k\}$ be a finite set of producers. Let x_i be the input, say, a d–vector, of producer i, to produce a, say, scalar output y, $i \in I$. Our data, say, consists of cross-section data (x_i, y_i), $i \in I$. Can you guess the next step? i.e. how to estimate θ from such data and model assumptions? Well, remember how Legendre invented his OLS method? We have k equations and d unknowns. To transform them into an "ordinary" situation, namely, d equations with d unknowns, to solve for the d components of θ, we use either the mean squared error (MSE) or the least absolute deviation error (LAD) concepts, i.e. estimating θ by minimizing either:

$$\sum_{i=1}^{k} [y_i - \varphi(\theta, x_i)]^2 \quad \text{or} \quad \sum_{i=1}^{k} |y_i - \varphi(\theta, x_i)|$$

To carry out the above minimization problems, in some simple way (!), we "assume" some simple form for the function, just like in linear regression, namely, that:

$$\varphi(\theta, x_i) = \theta x_i = \sum_{j=1}^{d} \theta_j x_{ij}$$

With these simplifications, the above optimization problems can be solved by quadratic and linear programming techniques, respectively.

Note, however, that, as $\varphi(\theta, x) = y$ represents the maximum produced output, the above optimization problems are subject to the same constraint $y_i \leq \varphi(\theta, x_i), i \in I$.

Just like OLS without random error term, the above ad-hoc estimation procedures lack statistical justifications. In fact, "noise" and technical inefficiencies of firms should be also parts of the model. Putting these two uncertainties together, the model becomes:

$$y_i = \varphi(\theta, x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, k$$

First, if ε is an "ordinary" noise, i.e. can be modeled as, say, a normal distribution, then the structural model should be in log-form, i.e.:

$$\log y_i = \log \varphi(\theta, x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, k$$

Moreover, by the nature of the problem itself, the noise ε must be nonpositive (one-sided disturbance), i.e. $\varepsilon \leq 0$ (almost surely) to make $\log y_i = \log \varphi(\theta, x_i)$, for all i. for example, $-\varepsilon$ could be modeled as a half-normal distribution. or an exponential distribution (leading to quadratic and linear programming, respectively). Thus, the estimator of θ in the above stochastic model is maximum likelihood estimators (MLE) under these special error distributions. Are we happy with that? You may say, why not, is this similar to OLS in general linear regression with Gaussian noise? Unfortunately, this is not similar to OLS in linear regression. Why?

It is true that, under special error distributions above, the estimator of θ (in the Cobb-Douglas log-linear model) is obtained as an MLE. However, what good to say that an estimator is an MLE? Well, to ascertain that the estimator has all the "good" statistical properties of an MLE, namely, (strong) consistency and asymptotically normal. However, recall that there are conditions for MLE to have such good properties. Specifically, MLEs are "good" if the model is *regular*.

Remark. Statistical procedures are "valid" under specified conditions. For example, AIC or BIC can be used only to select *regular models*. In other words, each procedure has its domain of applicability. These model selection criteria are established under the assumption that the model is regular. So, if a model is not regular, empirically you still can compute its AIC or BIC, but there is no rationale to use them to judge the model's quality (for explaining or for prediction). Using them, without checking their applicability conditions, is wrong.

Let's spell out, for ease of reference, the *regularity conditions* of a statistical model. These are conditions under which, MLE is proved to be consistent and asymptotically normal. A statistical model satisfying these conditions is called a *regular model.* A non-regular model might not have MLE! Specifically, a regular model is a *parametric* model satisfying the following conditions.

Let X be a random vector with probability density $f(x, \theta)$ with $\theta \in \Theta \subseteq \mathbb{R}^d$.

The model $\mathcal{F} = \{f(x, \theta)\colon \theta \in \Theta \subseteq \mathbb{R}^d\}$ is called a regular model if it satisfies the following regular conditions:

(1) The mapping $\theta \to P_\theta(dx) = f(x, \theta)dx$ is injective (one-to-one);

(2) The probability measure $P_\theta(dx)$, $\theta \in \Theta$, has common support;

(3) The parameter space Θ is an open set of $\mathbb{R}^d$;

(4) For almost all x, all third-order partial derivatives $\frac{\partial^3 f(x,\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}$ exist for all $\theta \in \Theta$;

(5) For all $\theta \in \Theta$, $E_\theta\left[\frac{\partial \log f(X,\theta)}{\partial\theta}\right] = 0$;

(6) $E_\theta\left[-\frac{\partial^2 \log f(X,\theta)}{\partial\theta\partial\theta'}\right]$ is finite and positive definite; and

(7) There exist functions M_{ijk} such that $E_{\theta_o}\left[M_{ijk}(X)\right] < \infty$ (θ_o is the true parameter).

The above regular conditions are used to obtain asymptotic properties of MLE. For proof, see e.g. (Nguyen and Rogers, 1989).

Back to our efficiency estimation. It is "interesting" that we run into a non-regular model, so that, having MLE does not help for justifying the estimation procedure.

The condition (2) is violated. Look at our model:

$$\log y_i = \log\varphi(\theta, x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, k$$

As $y_i \leq \varphi(\theta, x_i)$, $i \in I$, the range of the random variable Y depends on the parameter θ to be estimated. In other words, the probability measures $P_\theta(dy)$, $\theta \in \Theta$, do not have common support.

A situation such as this surfaces often in many areas of applications, so that either statistical properties of MLE (when they do exist) need to be proved (and not just based on established results in the regular case), or, if MLE does not exist (See an example shortly), we must look for other estimation methods.

Typically, the above condition (2) is violated when some components of the parameter vector θ lie in the support of the model, e.g. in *change-point models* of econometrics. Here is a situation in survival analysis.

It is observed that failures of lifetimes of subjects appear to occur at one rate and late failures appear to occur at another rate. If we denote by F(t),f(t) the population distribution and density functions of lifetimes, respectively, then the above observation is expressed in terms of the hazard rate function, as:

$$\frac{f(t)}{1 - F(t)} = a1_{[0,\tau]}(t) + b1_{(\tau,\infty)}(t)$$

This differential system gives the lifetime model (a special case of the extended Cox' proportional hazard model):

$$f(t|\theta) = a\exp\{-at\}1_{[0,\tau]}(t) + b\exp\{-a\tau - b(t-\tau)\}1_{(\tau,\infty)}(t)$$

where the parameter $\theta = (a,b,\tau) \in \Theta = ([0, \infty))^3$.

Now, if $T_1, T_2, \ldots, T_n$ is a random sample from the population with density $f(t|\theta)$, then the log-likelihood function is

$$L(\theta) = \sum_{i=1}^{n} (\log a) 1_{[0,\tau]}(T_i) - a \sum_{i=1}^{n} T_i 1_{[0,\tau]}(T_i)$$

$$+ \sum_{i=1}^{n} \{1 - 1_{[0,\tau]}(T_i)\} \{\log b - (a-b)\tau\} - b \sum_{i=1}^{n} T_i \{1 - 1_{[0,\tau]}(T_i)\}$$

Let $R(\tau) = r$ be the number of the $T_i \leq \tau$ and denote the order statistics as t_i. Then $L(\theta)$ is proportional to:

$$r \log a - a \sum_{i=1}^{r} t_i + (n-r)\{\log b - (a-b)\tau\} - b \sum_{i=r+1}^{r} t_i$$

For $t_{n-1} \leq \tau \leq t_n$, this reduces to:

$$(n-1)\log a - a \sum_{i=1}^{n-1} t_i - a\tau + \log b - b(t_n - \tau)$$

If we let $b = \frac{1}{t_n - \tau}$ and $\tau \to t_n$, then clearly $L(\theta)$ is unbounded. Note that, if $a > b$, then $L(\theta)$ is bounded, but, as Θ is unbounded, it is not clear that sup $L(\theta)$ can be attained.

As a consequence, we cannot estimate the unknown "change-point" τ by the standard method of maximum likelihood. So, how to estimate τ, of course, consistently? Note that, once τ is estimated (consistently), the hazard rates a, b can be estimated consistently.

A characterization of τ is this. Let:

$$M_n(\tau) = \sum_{i=r+1}^{n} \frac{t_i}{n-r} \ldots, \quad S_n^2(\tau) = \sum_{i=r+1}^{n} \frac{t_i^2}{n-r} - [M_n(\tau)]^2$$

and $\overline{T_n} = \frac{1}{n} \sum_{i=1}^{n} T_i$. Then applying the strong law of large numbers several times, we get:

$$X_n(\tau) = \frac{1}{n} \left\{ S_n(\tau) \left[(n-r) \log \left[\frac{n}{n-r} - r \right] \right] + r M_n(\tau) - \bar{T}_n \log \frac{n}{n-r} \right\}$$

converging strongly to zero, as $n \to \infty$.

Thus, as a familiar strategy, we consider the stochastic process $X_n(t)$, $t \geq 0$ and value $\hat{t_n}$ such that $X_n(\hat{t_n})$ is close to zero is taken as a candidate for an estimate of τ. Of course, it remains to prove that $\hat{t_n}$ converges strongly to τ as $n \to \infty$. It is indeed so, i.e. there is a strongly consistent estimator for the change-point τ, see the proof in (Nguyen *et al.*, 1984).

Remark. It is interesting to note that the above "change-point" problem is different than common models in structural changes in econometrics, as here the change-point τ is a point of discontinuity in the distribution function of a DGP. On the other hand, follow-up studies to complete this change-point hazard rate problem in survival analysis were in (Nguyen and Pham, 1982), (Nguyen and Pham, 1990).

Now, the above model for efficiency estimation also has another drawback, namely, it cannot isolate the effect of inefficiency from that of the random noise as these two types of effects are lumped together in one disturbance term ε.

The actual improved stochastic frontier model is as follows. As an input $x \in \mathbb{R}_+^d$ (e.g. labor, resource, capital, [. . .]) can produce various different, say, scalar outputs y, depending on how to "manage" the input, there is such thing as the "production frontier," namely, the maximum output an input can produce $\varphi(x) = \max\{y{:}x \to y\}$. If we know the frontier $\varphi(.)$ (of a given technology) then, when observing (x_i,y_i) from firm i, we can take $\frac{y_i}{\varphi(x_i)}$ as the (degree) of technology efficiency of firm i.

Now, an output y could be a result, not only of an input x and the production technology but also of some random "shock," resulting in a value, which could exceed the frontier value $\varphi(x)$. Thus, the concept of" frontier" should be extended to a "stochastic frontier": $x \to \varphi(x) + V$, with V being a symmetric random noise (variable), to cover this situation. For $\varphi(x) + V$ to be a frontier, we would have $Y \leq \varphi(X) + V$. Therefore, if we let $U = \varphi(X) + V - Y$, then the random variable U is nonnegative ($U \geq 0$, *a*.s.), representing the technical inefficiency. We arrive at the so-called stochastic frontier model (SFM):

$$Y_i = \varphi(X_i) + V_i - U_i$$

In particular, a (parametric) linear SFM is

$$Y_i = X_i'\theta + V_i - U_i$$

Remark. *As* U_i "captures" technology inefficiency of firm i, we would like to "estimate" it, from observations across firms (X_i, Y_i), $i \in I$. However, like V, U is not observable! However, if we could estimate θ, by some estimation method, then we can compute the estimated residuals $\varepsilon_i = v_i - u_i$ by $\hat{\varepsilon_i} = y_i - x_i'\hat{\theta}$ which are values of the random variable $V - U$ Thus, in MSE sense, the best predictor of U_i is $E\left(U_i | V_i - U_i = \hat{\varepsilon_i}\right)$.

3. Entering copulas

The (linear) stochastic frontier model is an "interesting" linear regression model $Y = X'\theta + V - U$, in which the random "error term" $V - U$ consists of two separate terms.

If we assume, say, a parametric form of the distribution of $V - U$, then we could estimate θ by MLE. However, it is very important to remember that, in a situation such as this, not only do we need to justify any model assumptions but also keep in mind that "while stronger assumptions lead to stronger results but also take us further away from realities!" Empirical research must be credible in the first place. Within credible statistics, we still need to do our best, i.e. strike to achieve "efficient" inference (not only via the search for best inference procedures but also via the way we collect data).

To estimate θ in the linear model $Y = X'\theta + V - U$, we should first examine the *identification* of the parameter of interest, before jumping into the estimation problem!.

Note that, in applications, as X, Y are non-negative quantities, we are considering a (Cobb-Douglas) log-linear model.

Now, by their nature, we can somewhat justify the forms of the distributions of the errors U and V, for example, the distribution F of U could be an exponential or half-normal distribution, whereas the distribution G of V could be a normal distribution.

The question is: Is θ identifiable, point or partial? from the data and model assumptions so far? The answer is no, as not only U, V are latent variables (no observations from them available) so that F, G cannot be estimated but also because we cannot even carry out MLE, as the distribution of $V - U$ is not known (in the form) even F, G are specified parametrically.

Remark. Even in a general situation where the marginal distributions F, G of two random variables U, V are estimable, say, when we have observations from U and V, the joint

distribution H of the random vector (U, V) is only *partially identified*, so that our "parameter of interest," namely, the distribution of $V-U$ is not point identified. However, in such a case, the distribution K of $V-U$ is partially identified, as, in view of Frechet's bounds (See a text on copulas, e.g. (Durante and Sempi, 2016), we have:

$$\max\{F(u)+G(v)-1,0\} \leq H(u,v) \leq \min\{F(u),G(v)\}$$

from which $K(.)$ can be partially identified.

As a "habit" in empirical practice, econometricians impose more assumptions to reach point identification so that they can estimate θ, without worrying about whether additional assumptions (even if justifiable) will take them further away from realities, let alone considering estimation in a partially identified model. For the topic of partial identification, see (Manski, 2003).

Let's look at model assumptions maintained in, say, "traditional" statistical practices, in a situation such as SFM. First, "assume" that V is $N(0,\sigma^2)$ and U is exponential with density $\left(\frac{1}{a}e^{-\frac{u}{a}}\right)1_{(u>0)}$ or half normal, i.e. U is distributed as $|N(0,\eta^2)|$ (with probability density of $\frac{\sqrt{2}}{\eta\sqrt{\pi}}\exp\{-\frac{u^2}{2\eta^2}\}1_{(u>0)}$).

As we need the distribution of the error term $\varepsilon = V-U$ in the regression model, the knowledge of the marginal distributions of U, V is not enough. We need the joint distribution H of (U, V) for it. Well, the simplest way to get H from F and G is to "assume" (or take) $H(u,v) = F(u)G(v)$, i.e. assuming that U and V are independent.

Remark. It is "interesting" to note that, starting with Frechet's work in 1951, Abe Sklar asked himself "what are all possible joint distribution functions H admitting given marginals F and G?" Leading to his PhD thesis in 1959, in which he "discovered" the concept of *copulas*. Prior to the discovery of copulas, and even after that, econometricians and statisticians still asked "are copulas useful?" Because, when facing the marginal problem, they rely on additional "conditional models" to get the joint distribution, avoiding even the partial identification issue. Among others, SFA is a perfect situation where we might not have additional conditional models.

It has to wait until 1990 for an explosion of copulas everywhere. Also, without being aware of Sklar's work in 1959, probabilists consider, as "difficult" an exercise of the form "find a joint distribution admitting two given marginals" (as given above in an SFM)!

Thus, without knowing the existence of copulas, it is not surprising that the only "feasible" additional assumption to get point identification is to assume independence of the error components in an SFM! Of course, with all assumptions so maintained, point identification is possible and estimation procedures are implemented.

Although copulas were known to econometricians in the 1990s, it did take some time for them to be applied to various fields. It was Smith (Smith, 2008), who was the first to consider, in 2008, the use of copulas in SFM, relaxing the "traditional" independence assumption on error components.

The "justifications" of the assumptions maintained in the literature of SFA prior to 2000, see (Kumbhakar and Knox Lovell, 2000), p. 74, namely, (i) V is $N(0,\sigma^2)$, (ii) U is half normal and (iii) U and V are independent of each other and of the regressors, are as follows:

- is "conventional!";
- "plausible"; and
- The independence of U, V seems "innocuous" (harmless, not controversial)!.

Recall that assumptions maintained in empirical research are responsible for the credibility of its results. Let's examine the assumption (3). Honestly speaking, is (3) innocuous? First, as stated above, we are interested in knowing the distribution of $V - U$ and from (1) and (2), all we need is the joint distribution of (U, V). Second, as also mentioned above, given the marginals F, G, is it a "difficult" exercise to find a joint distribution H of (U, V)? Yes, it is. In fact, there are lots of possible "solutions," as we know now from Sklar's theory of copulas. Assuming (3) provides one simple solution! But, for the credibility of "statistical science," any assumption must be justified. Smith in (Smith, 2008) did an excellent job by questioning (3) and replaced it by a copula model exhibiting correlated error components in SFM; see also (Amsler *et al.*, 2019).

How to "find out" whether an assumption like (3) is plausible or justifiable? Well, in general, a model assumption is about the "behavior" of the phenomenon under study. As such, we could use available data to check whether a proposed assumption "captures" the behavior it was proposed to capture! It is a data-driven approach, as it should be. Smith (Smith, 2008) said it precisely "allowing the data the opportunity to determine the adequacy of the independence assumption."

A general approach to SFM with correlated error components, e.g. in a linear model $Y = X'\theta + V - U$ in which U, V are not independent, is using copulas ((Durante and Sempi, 2016))) to model the dependence between U and V. Of course, the approach is data-driven. Again, see, e.g. (Smith, 2008) and (Amsler *et al.*, 2019), for illustrations.

Remark. As the main tool in econometrics seems to be regression, it should be pointed out that regression analysis cannot isolate cause and effect (i.e. correlation does not tell us how and why certain phenomena have occurred). We also need *a causal inference* theory.

4. Estimation by the method of sieves

The linear regression model in SFA is "interesting" as it presents a typical situation where, in one hand, *copula modeling* is a mandate and on the other hand, maximum likelihood estimation (MLE) by the *method of sieves* should be used, both are somewhat unfamiliar to statisticians and econometricians. Note that to make the econometric analysis more credible (i.e. fewer model assumptions maintained to make the statistical analysis closer to realities), classical parametric estimation is extended to semiparametric, to semi-non-parametric and to nonparametric estimation. The sieves method is a "non-standard" nonparametric estimation, invented by U. Grenander in 1981 (Grenander, 1981), followed immediately by (Geman and Hwang, 1982) and (Nguyen and Pham, 1982) in 1982. It took quite some time for econometricians to appreciate its usefulness, not only in general semi-nonparametric estimation (Chen, 2007; Chen and Pouzo, 2015)) but also in causal inference with regression discontinuity design (RDD), e.g. (Davezies and Le Barbanchon, 2017). Well, it sounds familiar: while the notion of copulas was discovered by abe Sklar in 1959, it was dormant until 1990 for econometricians to be aware of it and from that time on, "copulas are everywhere!" The same phenomenon happened to the concept of RDD, considered in (Thistlewaite and Campbell, 1960), which stayed dormant also until 1990 and from that year on, RDD is everywhere in causal inference!

Traditionally, when we wish to predict a random variable Y from a "chosen" collection of covariates (variables affecting Y), a vector $X = (X_1, X_2, \ldots, X_k)$, we seek the best predictor $\Psi(X)$ in the sense that its mean squared error (MSE) is the smallest, where $\text{MSE}(\Psi(X)) = E[\Psi(X) - Y]^2$. It is a theorem that our best predictor, in the MSE sense, is $E(Y|X)$. The process of estimating $E(Y|X)$ from data is termed *predictive modeling*.

Remark. If we are interested in asking another question, e.g. "if we intervene on the covariate X_j (say, increasing by an amount ∂X_j), what will happen to Y?" Then we are not

doing predictive modeling, as what we seek is the partial derivative $\frac{\partial E(Y|X)}{\partial X_j}$ and not $E(Y|X)$. We are doing *causal estimation*.

Traditionally (again), rather than addressing the estimation of $E(Y|X)$ nonparametrically (which would be more "credible"), people proceed parametrically by "considering" a linear model for $E(Y|X)$, namely, $E(Y|X) = X'\theta$ (justifying as a good approximation). Then the problem is the estimation of the (finitely dimensional) parameter $\theta \in \mathbb{R}^k$. The associated model for the Data Generating Process (DGP) is $Y = X'\theta + \varepsilon$ with the "ideal condition" $E(\varepsilon|X) = 0$ and a "natural" assumption that ε is a $N(0,\sigma^2)$ random variable. Doing so we have a point identified problem followed by the standard MLE estimation method.

Suppose we care about credible statistics (!)while keeping $E(\varepsilon|X) = 0$, we leave the density f of the error term ε unspecified. Then we are facing a *semiparametric* estimation problem, in which, the parameter of interest θ is finitely dimensional, whereas the infinitely dimensional f is a nuisance parameter. How to estimate θ in this setting? Is it still possible to use MLE? As we will see, the answer is yes if we apply Grenander's *method of sieves* to MLE (Grenander, 1981), see also (Geman and Hwang, 1982).

More generally, suppose we want to estimate $E(Y|X = x)$ by estimating the conditional distribution of Y given X (in a non-linear model). One framework for this problem is this. Let (X, Y) be a random vector. The joint distribution function H of (X, Y) is related to the marginal distributions F, G of X, Y, respectively, by $H(x,y) = C(F(x), G(y))$ where C is a copula. It boils down to estimating H which can be breaking down into two cases:

(1) Specifying parametrically F and G and leave C unspecified;

(2) Specifying parametrically C and leave F and G both unspecified.

The above estimation problems can be carried out by using the method of sieves; see (Chen *et al.*, 2004) for (2), (Panchenko and Prokhorov, 2016) for (1).

Remark. Another situation where copula modeling appears naturally is the (James Heckman) sample selection model. Sieve MLE is applicable too, see e.g. (Schwiebert, 2003).

Essentially, Grenander's method of sieves is a method for implementing standard MLE for estimating finitely dimensional parameters when facing infinitely dimensional parameters. It is known that MLE has difficulties in the infinite-dimensional case, as the maximum likelihood solution is generally not attained or is not consistent. To handle these difficulties, the method of sieves was invented by U. Grenander in 1981 in his "abstract inference." In this method, for each sample size, a sieve (a suitable subset of the parameter space) is chosen. The likelihood function is then maximized over the sieves yielding a sequence of estimators. The crucial point is the choice of appropriate sieves so that, as the sample sizes increase, the sequence of sieve estimators is consistent. When the infinitely dimensional parameter space is a Hilbert space (e.g. the Sobolev space of copulas), a sequence of sieves can be chosen simply as finitely dimensional subspaces of it. The restricted MLE on these sieves is consistent and asymptotically normal provided that the dimensions of the sieves grow not too fast with respect to the sample size. This is illustrated in (Nguyen and Pham, 1982) which we reproduce here a bit.

Consider a "general" form of a standard linear regression model with Gaussian noise (a nonstationary diffusion model):

$$dX(t) = \theta(t)X(t)dt + dW(t) \quad , \quad X(0) = x_o$$

where x_o is deterministic, $W(t)$ is a Brownian motion with $E[dW(t)]^2 = \sigma^2 dt$ and $\theta(.) \in L^2([0, T], dt)$, $[0,T]$ being the time interval of observation of the process.

The problem is the estimation of the function $\theta(.)$ on $[0,T]$, a functional parameter, where the data is a sample of $X_i(.)$, $i = 1, 2, \ldots, n$ of trajectories of $X(t)$ on $[0,T]$. This is a "functional data" (Ramsay and Silverman (2002), e.g. interpolated panel data, or, as in machine learning, a "data point" is not only high dimensional but also could be infinitely dimensional, such as a curve in the plane.

As sieves, we choose an increasing sequence S_n of subspaces of $L^2([0, T], dt)$, with finite dimension d_n, such that $\cup_{n\geq 1} S_n$ is dense in $L^2([0, T], dt)$, as follows. Let $f_j, j \geq 1$ be a sequence of independent elements of $L^2([0, T], dt)$, with $f_1, f_2, \ldots, f_n$ forming a basis of S_n, for all n. Then for $\theta(.) \in S_n$, we have $\theta(\cdot) = \sum_{j=1}^{n} \theta_j f_j(\cdot)$.

We maximize the log-likelihood function $L_n(\theta)$ on each S_n, where, for $\theta(\cdot) \in S_n$, $L_n(\theta) =$

$$\sum_{i=1}^{n} \left\{ \frac{1}{\sigma^2} \int_0^T \left[\sum_{j=1}^{d_n} \theta_j f_j(t) \right] X_i(t) dX_i(t) - \frac{1}{2\sigma^2} \int_0^T \left[\sum_{j=1}^{d_n} \theta_j f_j(t) \right]^2 X_i^2(t) dt \right\}$$

Then, under the conditions, as $n \to \infty$, $d_n \to \infty$ and $\frac{d_n^2}{n} \to 0$, the sequence of restricted MLE possesses desirable asymptotic properties. For details, see Nguyen and Pham (1982).

Remark. Note that the construction of sieves above is similar, in spirit, to the technique of projections (or of orthogonal functions) in nonparametric density estimation. Here, our parameter of interest is the infinitely dimensional $\theta(.)$ which is "nonparametric" and we estimate it "parametrically" in the sense that an estimator of $\theta(.)$ is a sequence of finitely dimensional (statistical) functions. Thus, by "nonparametric estimation" we mean the problem of estimating a nonparametric *parameter* (an infinitely dimensional object) and not necessarily "nonparametric estimators" *per se*! i.e. a nonparametric estimator of a nonparametric parameter could be parametric!

Now the situation in SFA is different. With the notation in previous Sections, the error components U, V are not observable, i.e. in searching for their marginals or copula, we are not really facing an estimation problem *per se*.

The copula model in SFA, as initiated by Smith in (Smith, 2008), aims at extending the independence assumption on error components, while keeping other seemingly "plausible" assumptions, namely, *parametric* forms of marginal distributions F_α, G_β. Given data from a linear regression model, the only thing to do (to study technology efficiency) is to choose an "appropriate" copula C to form the joint distribution $H(u,v) = C(F_\alpha(u), G_\beta(v))$ of (U, V), from which the distribution of the error term $\varepsilon = V - U$ can be derived. Thus, the crucial question is: How to choose C? The problem is somewhat delicate here, as (U, V) is not observable (otherwise, the choice problem can be treated as a copula estimation problem). The choice procedure in (Smith, 2008) is simple. On the one hand, C is chosen as a *parametric copula* C_γ and on the other hand, that choice is based upon dependence coverage considerations in the theory of copulas. Together with parametric forms of the marginals, the whole setting is a parametric estimation problem with parameters $(\theta, \alpha, \beta, \gamma)$ which can be estimated by MLE from the SFM $Y = X'\theta + V - U$.

To "improve" upon this fully parametric setting, say, while the marginals F_α, G_β are specified parametrically, the copula C is left unspecified, i.e. just a member the space of bivariate copulas $\mathcal{C}$. The finitely dimensional parameter of interest is $(\theta, \alpha, \beta) \in \Omega$ and the nuisance parameter is infinitely dimensional $C \in \mathcal{C}$. We are facing a semi-parametric estimation problem in the structural model $Y = X'\theta + V - U$, with data (X_i, Y_i), $i = 1, 2, \ldots,$

n. The likelihood function is maximized over $\Omega \times \mathcal{C}$. Now, as $\mathcal{C}$ is a subset of a Hilbert space (the Sobolev space $W^{I,2}(I^2, \mathbb{R})$, where $I = [0,1]$, see Siburg and Stoimenov (2008), sieve MLE can be used as in (Nguyen and Pham, 1982). as above.

5. Causality analysis in stochastic frontier models

Moving from association inference to causation inference is a natural path. In the context of SFA, this is not only because SFM is a regression model but also by its subject matter, it is useful to do so, for example, modeling determinants affecting production inefficiency. So far, the literature on causal inference for SFM seems limited, especially with respect to the use of regression discontinuity designs (RDD); see, however, (Johnes and Tsionas, 2019).

In this Section, we elaborate and emphasize the now popular RDD in general causal inference practices with obvious possible applications to SFA. On the other hand, as SFM are regression models, regularizations in various forms for covariate selection and appropriate inferences could be considered.

Essentially, causal inference is about detecting causal effects of, say, a "treatment" (intervention) on some variables of interest ("outcome"). It consists of choosing two groups in a population of units, with one group subjected to the treatment while the other group is not, for comparison. When the golden standard of random experiments (for choosing two "equivalent" groups) cannot be used, for various (social) reasons, researchers now turn to a popular design method, called regression discontinuity design (RDD).

The RDD was initiated in (Thistlewaite and Campbell, 1960) back in 1960. The idea is to replace a random experiment by a "design" which provides a group where units in that group are somewhat "similar" (equivalent), from which, a "tie-breaking" randomization assignment is applied to form two desired groups for comparison (see the Foreword of D.T. Campbell to Trochim's book (Trochim, 1984)). The design is illustrated by a real problem. To give scholarships to potentially good seniors so that they perform better in their graduate studies, a university gives an examination (a "pretest") to the seniors at the end of their senior year to determine which students deserve the awards. For this purpose, a cutoff score (threshold) is chosen: If X_i denotes the score of student i in this pretest and the chosen cutting point is x_o, then i will be awarded the scholarship iff $X_i \geq x_o$. Call X the score or "running variable" of the design. Each student (a unit in a population I) has a score (revealed after the pretest). To find out whether or not the scholarship award (a "treatment," intervention) has a positive effect, several years later, a "post-test" (with score Y) will be given to all units (in two groups). So a design consists of (X, x_o, Y). It is called a "regression," as we are going to "regress" Y on X (say, by linear regression) for comparison; there could have a "jump" (discontinuity) from the line on the left of the threshold x_o to the line on its right side. That was why the name of the design is "regression discontinuity." Finally, given the threshold x_o, if the assignment rule is applied "seriously," i.e. unit i is selected for an award if and only if $X_i \geq x_o$, then the design is termed "sharp," leading to a "sharp RDD." If the assignment rule is somewhat "elastic" (or flexible) in the sense that for scores X_i around x_o (a neighborhood $N(x_o)$ of the cutoff point), below or above, some additional procedures will be used to classify them, e.g. viewing them as equivalent, do a tie-break such as classify them into two groups at random or seeking extra information (e.g. interviews, recommendation letters, etc [. . .].). In this case, the RDD is termed *fuzzy* RDD. Thus, sharp or fuzzy RDD is about how we apply the assignment rule around the cutoff point.

Remark. Fuzzy RDD was first termed in (Campbell, 1969); see also, (Trochim, 1984), p. 55. The adjective "fuzzy" is used only to mean that the assignment rule is not sharp (in a specific way). It is not related to Zadeh's notion of fuzzy sets in 1965, recalling that a fuzzy set is a "generalized set" whose elements can have partial membership degrees, e.g. a coalition of

players in a (coalitional) game where partial memberships are allowed; see, e.g. (Nguyen *et al.*, 2019) for Zadeh's theory of fuzzy sets and logic.

However, this idea of "fuzzy assignments" of units near the cutoff point is similar to a situation in Statistical Quality Control where observations near the boundaries of a control chart should be treated with care, and fuzzy set techniques could be used.

Now, recall that causal inference concerns "what" would happen to an "outcome" (response) Y as a result of a "treatment" (or "intervention"). More specifically, it concerns the comparison of treatment with something else, say "no treatment" or "control." The question is: how to figure out whether there is a causal effect?

Suppose we have a sample (random or not) of size n of units from a population U to spit into two groups t (treatment) and c (control). Unlike associational inference, each unit i now have two "potential outcomes": Y_{1i} and Y_{0i}, representing the outcome on unit i if it is exposed to t and c, respectively.

The individual treatment effect is obviously the difference $Y_{1i} - Y_{0i}$. However, we cannot observe both Y_{1i} and Y_{0i}, but only observe one of them.

When $i \in t$, we may wish to substitute to the unobserved Y_{0i} by some $j \in c$ which is "similar" to i, say, in terms of other characteristics. This is possible if the observation study is can be conducted by a random process, which tends to "balance out" similarity so that counterfactuals can be found.

If we let D be the assignment rule: $D_i = 1$ or 0 according to $i \in t$ or $i \in c$, then the "regression" observed model is:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

RDD provides a design for observational studies, which allow us to view, at least, locally, the observational studies as the "golden standard design" of random experiments. Here the assignment rule D is not random and is not under the control of the evaluator: $D_i = 1_{(X_i \geq x_o)}$.

In such a situation, how to "identify" the treatment effect? and how to estimate it? The observed model is:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

If we plot the data, then we see two pictures:

(1) Plotting X_i versus D_i: there is a "sharp" jump of the assignment at the cutoff point x_o:$P(D_i = 1|X_i)$ jumps from 0 to 1 at x_o,

(2) Plotting X_i versus Y_i: there is a discontinuity of $E(Y_i|X_i)$ at x_o which could be used to determine a causal effect.

Specifically, the question is: How to estimate the causal effect $E(Y_1 - Y_0)$ from data $(D_i, Y_i; i = 1, \ldots, n)$?

Let $a_i = Y_{0i}$ and $b_i = Y_{1i} - Y_{0i}$, then our observed model is $Y_i = a_i + b_i D_i$.

Consider first the "sharp" design where the assignment rule $D = 1_{(X \geq x_o)}$ where X is a concomitant variable.

The population parameter b_i is said to be (nonparametrically) "(point) identifiable" if we can express it uniquely in an "estimable" fashion. For example, if the treatment effect is constant throughout the population, i.e. when $b_i = b$ for all i, then it can be shown that the condition "$x \to E(Y_{0i}|X_i = x)$ is continuous at x_o" is sufficient to identify b as $b = Y^+ - Y^-$ where

$$Y^+ = \lim_{x \to x_o +} E(Y_i | X_i = x) \ldots\ldots, \quad Y^- = \lim_{\mathrm{x} \to x_o -} E\left(Y_{\mathrm{i}} \mid X_{\mathrm{i}} = x\right)$$

When the treatment effect varies across units, additional conditions are needed for (point) identification. For example, if "$x \to E(Y_{0i}|X_i = x)$ is continuous at x_o," "$x \to E(Y_{1i} - Y_{0i} \mid X_i = x)$ is continuous at x_o," and "D_i is independent of $Y_{1i} - Y_{0i}$ conditional on X_i near x_o," then it can be shown that:

$$E(b_i | X_i = x_o) = Y^+ - Y^-$$

Note that, for fuzzy RDD, $\lim_{x \to x_o +} E(D_i | X_i = x_o) - \lim_{x \to x_o -} E(D_i | X_i = x_o)$ is different than zero and we have:

$$E(b_i | X_i = x_o) = (Y^+ - Y^-)/(D^+ - D^-)$$

Under appropriate conditions of the RDD, the treatment effect can be estimated, locally around the cutoff point from observed data, just like in a random experiment.

Clearly, the estimate of the treatment effect near the cutoff point is obtained as a plug-in estimator. Specifically, it suffices to estimate Y^+, Y^-, D^+, D^-. Now observe that these parameters are conditional means. As such, the nonparametric regression method is used for estimation. However, besides point estimators, the problem of variance estimation for confidence interval estimation is complicated. A novel nonparametric method for confidence intervals, known as *Empirical Likelihood* (EL) is, therefore, called for, as this method avoids variance estimation and provides confidence regions based solely on data.

This nonparametric method can be used in a variety of situations, especially for parameters in moment condition models.

Consider the simplest (standard) setting: let X_1, X_2, …, X_n be i.i.d. drawn from a population X with unknown distribution function F_o. As the (nonparametric) parameter space for F_o is the set (or a subset) $\mathcal{F}$ of all distribution functions, a likelihood of $F \in \mathcal{F}$, given the observations is:

$$L(F | X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} [F(X_i) - F(X_i -)] = \prod_{i=1}^{n} p_i$$

This likelihood is "consistent" with the fact that the empirical distribution function:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(X_i \le x)}$$

maximizes it. Note that:

$$L(F_n | X_1, X_2, \ldots, X_n) = \left(\frac{1}{n}\right)^n$$

so that the likelihood ratio:

$$r(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i$$

Suppose our parameter of interest is $\theta = T(F)$. Then the profile likelihood is:

$$R(\theta) = \sup\{r(F) : F \in \mathcal{F} \cap T^{-1}(\theta)\}$$

and the associated confidence interval for θ_o is of the form $\{\theta : R(\theta) \geq c\}$.

This concept of (nonparametric) likelihood is particularly useful for setting up natural confidence intervals in moment condition models (frequently encountered in econometrics) and in quantile regression.

References

Amsler, C., Prokhorov, A.B. and Schmidt, P. (2019), "A new family of copulas, with application to estimation of a production frontier system", BA Working paper No: BAWP-2019-04, *Univ. of Sydney*.

Campbell, D.T. (1969), "Reforms as experiments", *American Psychologist*, Vol. 24 No. 4, pp. 409-429.

Chen, X. (2007), "Large sample sieve estimation of semi-nonparametric models", *Handbook of Econometrics*, Elsevier, Vol 6, Part B, Chapter 76, pp. 5549-5632.

Chen, X. and Pouzo, D. (2015), "Sieve Wald and GLR inferences on semi/nonparametric conditional moment models", *Econometrica*, Vol. 83 No. 3, pp. 1013-1079.

Chen, X., Fan, Y. and Tsyrennikov, V. (2004), "Efficient estimation of semiparametric multivariate copula models", *Journal of the American Statistical Association*, Vol. 101 No. 475, pp. 1228-1240.

Cobb, C.W. and Douglas, P.H. (1928), "A theory of production", *Amer. Econ. Review*, Vol. 18, pp. 139-165.

Davezies, L. and Le Barbanchon, T. (2017), "Regression discontinuity design with continuous measurement error in the running variable", *Discussion Paper Series IZADP No 10801*.

Durante, F. and Sempi, C. (2016), *Principles of Copula Theory*, Chapman and Hall/CRC Press, Boca Raton, FL.

Geman, S. and Hwang, C.R. (1982), "Nonparametric maximum likelihood estimation by the method of sieves", *The Annals of Statistics*, Vol. 10 No. 2, pp. 401-414.

Grenander, U. (1981), *Abstract Inference*, Wiley, Hoboken, NJ.

Johnes, G. and Tsionas, M.G. (2019), "A regression discontinuity frontier model with an application to educational attainment", doi: 10.1002/sta4.242.

Kumbhakar, S.C. and Knox Lovell, C.A. (2000), *Stochastic Frontier Analysis*, *Cambridge Univ. Press*.

Manski, C.F. (2003), *Partial Identification of Probability Distributions*, *Springer*.

Nguyen, H.T. and Pham, T.D. (1982), "Identification of nonstationary diffusion model by the method of sieves", *SIAM Journal on Control and Optimization*, Vol. 20 No. 5, pp. 603-611.

Nguyen, H.T. and Pham, T.D. (1990), "Strong consistency of maximum likelihood estimator in a change-point hazard rate model", *Statistics*, Vol. 21 No. 2, pp. 203-216.

Nguyen, H.T. and Rogers, G.S. (1989), *Fundamentals of Mathematical Statistics (Volume II: Statistical Inference)*, *Springer*.

Nguyen, H.T., Rogers, G.S. and Walker, E.A. (1984), "Estimation in change- point hazard rate models", *Biometrika*, Vol. 71 No. 2, pp. 299-304.

Nguyen, H.T., Walker, C.L. and Walker, E.A. (2019), *A First Course in Fuzzy Logic*, 4th ed., Chapman and Hall/CRC Press, Boca Raton, FL.

Panchenko, V. and Prokhorov, A. (2016), *Efficient Estimation of Parameters in Marginals in Semiparametric Multivariate Models*, Google.

Ramsay, J.O. and Silverman, B.W. (2002), *Applied Functional Data Analysis*, Springer.

Schwiebert, J. (2003), *Sieve Maximum Likelihood Estimation of a Copula-Based Sample Selection Model*, Google.

Siburg, K.F. and Stoimenov, P.A. (2008), "A scalar product for copulas", *Journal of Mathematical Analysis and Applications*, Vol. 344 No. 1, pp. 429-434.

Smith, M.D. (2008), "Stochastic frontier models with dependent error components", *The Econometrics Journal*, Vol. 11 No. 1, pp. 172-192.

Thistlewaite, D. and Campbell, D. (1960), "Regression discontinuity analysis: an alternative to the ex post facto experiment", *Journal of Educational Psychology*, Vol. 51 No. 6, pp. 309-317.

Trochim, W.M.K. (1984), *Research Design for Program Evaluation: The Regression Discontinuity Approach*, Sage Publications.

Further reading

Hahn, J., Todd, P. and Van der Klaauw, W. (2001), "Identification and estimation of treatment effect with a regression discontinuity design", *Econometrica*, Vol. 69 No. 1, pp. 201-209.

Hardy, G.H., Littlewood, J.E. and Polya, G. (1967), *Inequalities*, Cambridge Univ. Press.

Nguyen, H.T. and Pham, T.D. (1993), "Bootstrapping the change-point in a hazard rate model", *Annals of the Institute of Statistical Mathematics*, Vol. 45 No. 2, pp. 331-340.

Owen, A.B. (2001), *Empirical Likelihood*, Chapman and Hall/CRC Press, Boca Raton, FL.

Appendix

Empirical copulas

A research report by R. Johnson, J. Evans and D. Green, of the USDA, published in 1999, entitled "some bivariate distributions for modeling the strength properties of lumber" [Research paper FPL-RP-575 (Google)] is a typical example of the need to have a bivariate distribution having specified marginals. This is so, as, single strength property can be easily modeled with univariate distributions, but not the overall strength properties of the lumber.

The goal of that paper is to "review major techniques for obtaining bivariate distributions" and "to pick a promising bivariate distribution" for applications in agriculture.

The review of "major techniques for obtaining bivariate distributions" is interesting on two counts. First, it reminds us of the efforts to solve this essential problem, from "transformation methods," "Farlie-Gumbel-Morgenstern-families" to Marshall-Olkin's mixed models (1988). The second count is amazing. Total ignorance of the existence of copulas or more precisely, Maurice Frechet's problem! When searching for a bivariate distribution, say with marginals as univariate Weibull distributions (for reliability), you see statements such as "because of the dearth of bivariate distributions with appropriate marginal distributions, no good candidates are available at this time," and "the only viable candidate distributions appear to be the bivariate Weibull."

We review here "statistical inference about copulas" by viewing the copula as a "parameter" of a joint distribution.

If we focus on the practical problem of obtaining a copula C from observed data to obtain the joint distribution H of (X, Y), as well as modeling the dependence between X and Y, then situations are as follows.

Depending on situations, C could be obtained by one of the following ways:

- by estimation;
- by some selection rules.

In any case, we need to figure out, first, how to relate our "parameter" C to the observed data? In other words, what the data tell us about C? Well:

$$C(u,v) = P\left(X \leq F^{-1}(u), Y \leq G^{-1}(v)\right) = H\left(F^{-1}(u), G^{-1}(v)\right)$$

The data (X_i,Y_i), $i = 1, 2, \ldots, n$ came from $H(.,.)$, so that H, F, G can be estimated empirically, say by H_n, F_n, G_n and F_n^{-1}, G_n^{-1}, so that, $C(u, v)$ can be estimated (pointwise) accordingly.

As far as *estimation* is concerned, procedures will depend on two main things:

(1) *Data:* Recall that C is the joint distribution of (U, V) with uniform marginals on [0,1]. As $U = F(X)$ and $V = G(Y)$, the "data" for C is $(F(X_i), G(Y_i))$, $i = 1, 2, \ldots, n$. However, F, G are unknown, so that these "data" are not observable. If F, G are estimated by F_n, G_n, then the observed "pseudo- data" is $(F_n(X_i), G_n(Y_i))$, $i = 1, 2, \ldots, n$. What are the properties of this pseudo-data?

(2) *Estimation approaches:*
   - *Parametric*: when it is "appropriate" to assume that both the marginals and the copula are parametric;
   - *Semi-parametric*: when only the copula is parametric and not the marginals; and
   - *Nonparametric*: when marginals and copula are arbitrary.

Modeling dependence structures. When X and Y are independent, their "dependence" structure is "independence" which is captured by the product copula $C_I(u,v) = uv$. When they are not independent, i.e. their copula $C \neq C_I$, we face an infinity of types of dependence between them! All we know is that each functional form of C represents one type of dependence structure. A dependence structure is a "way" that the variables depend on each other. For example, the dependence structure is linear when $Y = aX + b$ (a,*s*.), $a \neq 0$, which is expressed as a functional form on the variables. This functional form of dependence is represented by the copula $C^+(u,v) = \min\{u,v\} = u \wedge v$ if $a > 0$ and by $C^-(u,v) = \max\{u + v - 1,0\}$ if $a < 0$. In other words, the *linear* dependence structure is represented by the subsets of copulas $\{C^+, C^-\} \subseteq \mathcal{C}$ (the infinitely dimensional space of all bivariate copulas), noting that a copula is simply a (bivariate) distribution with uniform marginals on [0,1].

The copula C^+ by itself represents the specific form of dependence called *comonotonicity*: X and Y "depend" on each other in the sense that they move up or down simultaneously, i.e. the set $\{(X(\omega), Y(\omega)):\omega \in \Omega\}$ is a chain in $\mathbb{R}^2$.

The copula C^- by itself represents the dependence structure known a *counter-monotonicity:* $(X,Y) =_d (\varphi(Z),\psi(Z))$ with φ strictly increasing and ψ strictly decreasing (or vice versa).

Variables can exhibit some form of dependence such as they are more likely to be simultaneously large (or small) than if they were independent. This notion of dependence is termed *positive quadrant dependence (PQD)*. Translating this description of dependence into mathematics, we have that X and Y are PQD iff $H(x,y) \geq F(x)G(y)$ for all $(x,y) \in \mathbb{R}^2$, which is equivalent to $C \geq C_I$. Thus, the subset $\{C \in \mathcal{C} : C \geq C_I\}$ characterizes PQD.

Why we are worried about various types of dependence structures? Typically, say, in econometrics where we are often interested in studying relationships among variables, dependence plays an essential role. For example, premium calculations in actuarial science are based upon dependence structures of insurance claims; computations of financial risks of investment portfolios, say, via Tail Value-at-Risk, require the dependence structure of the assets in the portfolio under study.

Theoretically speaking, dependence structures can be specified if we know the joint distribution of variables involved. Typically, we only, at best, have information on marginal distributions. Thus, we face the modeling of the dependence structure by itself.

Parametric estimation

Maximum likelihood estimation. We consider first the simplest setting of estimation, namely, parametric setting in which the joint distribution is linked to its marginals and copula by:

$$H_\theta(x,y) = C_\gamma\big(F_\alpha(x), G_\beta(y)\big)$$

for all $(x,y) \in \mathbb{R}^2,\ \theta = (\alpha, \beta, \gamma) \in \Phi \subseteq \mathbb{R}^k$.

Note that the separation of copula and marginals from the joint distribution is very important for modeling: We can say, first model the marginals (without worrying about their copula), of course, based on evidence from observed data, for example:

$$F_\alpha(x) = \exp\{-x^\alpha\}1_{[0,\infty)}(x)$$

which is a Frechet distribution (heavy-tailed) with $\alpha < 0$ and:

$$G_\beta(y) = (1 - e^{-\beta y})1_{(0,\infty)}(y)$$

which is an exponential distribution, with $\beta > 0$. Then (independently of the marginals), the copula could be parametrized by, say:

$$C_\gamma(u,v) = uv + \gamma uv(1-u)(1-v)$$

for $\gamma \in [-1,1]$. Thus, $\theta = (\alpha, \beta, \gamma) \in \Theta = (-\infty, 0) \times (0, \infty) \times [-1, 1] \subseteq \mathbb{R}^3$.

Let (X_i,Y_i), $i = 1, 2, \ldots, n$ be a random sample drawn from (X, Y). We wish to use these observations to estimate the dependence structure of (X, Y), which is captured by its true, unknown copula $C \in \{C_\gamma\}$.

Remark. To relate C to the data, we note that:

$$C(u,v) = H\big(F^{-1}(u), G^{-1}(v)\big) = P\big(X \le F^{-1}(u), Y \le G^{-1}(v)\big)$$

$$= P\big(F(X) \le u, G(Y) \le v\big)$$

or, when F and G are continuous, $U = F(X)$, $V = G(Y)$ are uniformly distributed on [0,1], so that their joint distribution C. As such, the observed data (X_i,Y_i), $i = 1, 2, \ldots, n$ is "related" to C via the unobserved data $(F(X_i), G(Y_i))$, $i = 1, 2, \ldots, n$, like F, G are unknown.

However, by looking at the joint distribution as a whole, we can write down the *likelihood function of the parameter* θ when observing (X_i,Y_i), $i = 1, 2, \ldots, n$, assuming, of course, their copula C is absolutely continuous, namely:

$$L_n\big(\theta|(X_i, Y_i), i = 1, 2, \ldots, n\big) = L_n(\theta) = \prod_{i=1}^{n} h_\theta(X_i, Y_i)$$

where

$$h_\theta(x,y) = \frac{\partial^2 H_\theta(x,y)}{\partial x \partial y} = \frac{\partial^2 C_\gamma\left(F_\alpha(x), G_\beta(y)\right)}{\partial x \partial y}$$

$$= f_\alpha(x) g_\beta(y) c\left(F_\alpha(x), G_\beta(y)\right)$$

where $f_\alpha(x) = \frac{dF_\alpha(x)}{dx}$, $g_\alpha(x) = \frac{dG_\alpha(x)}{dx}$ and $c_\gamma(u,v) = \frac{\partial^2 C_\gamma(u,v)}{\partial u \partial v}$.

Then, formally, the MLE of θ is obtained by maximizing $L_n(\theta)$ over Θ. This "exact" MLE aims at *simultaneously* maximizing the parameter vector $\theta = (\alpha, \beta, \gamma)$.

Note that:

$$\log L_n(\theta) = \log \prod_{i=1}^{n} f_\alpha(X_i) g_\beta(Y_i) c\left(F_\alpha(X_i), G_\beta(Y_i)\right)$$
$$= \sum_{i=1}^{n} \log c\left(F_\alpha(X_i), G_\beta(Y_i)\right) + \sum_{i=1}^{n} \log f_\alpha(X_i) + \sum_{i=1}^{n} \log g_\beta(Y_i)$$

in which the first term is the log-likelihood due to the dependence structure, whereas the sum of the last two terms is the log-likelihood due to the marginals.

Remark.

(1) A copula C is a joint (bivariate) distribution function. As such, it generates a probability measure, denoted as dC, on $\mathcal{B}(\mathbb{R}^2)$. Saying that C is absolutely continuous, we simply mean that dC is absolutely continuous with respect to Lebesgue measure $dx \otimes dy$ on $\mathbb{R}^2$, i.e. $dC(u,v) = \left[\frac{\partial^2 C(u,v)}{\partial u \partial v}\right] dudv$.

(2) When F and G are continuous, C is the joint distribution of the uniform marginals $U = F(X)$, $V = G(Y)$. However, as F and G are unknown, the values $(F(X_i), G(Y_i))$, $i = 1, 2, \ldots, n$ are not observable, so that we do not have data from (U, V) to estimate their joint distribution C by standard methods. If F and G are estimated from their data X_i, $i = 1, 2, \ldots, n$ and Y_i, $i = 1, 2, \ldots n$ (which are separately random samples), say, by F_n and G_n, then $(F_n(X_i), G_n(Y_i))$, $i = 1, 2, \ldots, n$ are observable, called the *pseudo-observations*. Then, formally, we can consider estimating γ by MLE based on the copula density $c_\gamma(.,.)$ and the pseudo-observations: Maximizing:

$$\prod_{i=1}^{n} c_\gamma(F_n(X_i), G_n(Y_i))$$

over the parameter space of γ. Such an estimator of γ is referred to as an *omnibus estimator* (a bus for all!). Of course, the statistical properties of such estimators should be examined.

Important note on pseudo-observations. If F and G are continuous and *known*, then $(F(X_i), G(Y_i))$, $i = 1, 2, \ldots, n$ are observable and are i.i.d. $(F(X), G(Y))$ which has C as its joint distribution. In other words, in this case, $(F(X_i), G(Y_i))$, $i = 1,2, \ldots, n$ is viewed as a random sample drawn from the copula C.

When F and G are unknown, the pseudo-observations $(F_n(X_i), G_n(Y_i))$, $i = 1, 2, \ldots, n$ are *not* mutually independent and $(F_n(X_i)$, $i = 1, 2, \ldots, n$ and $G_n(Y_i))$, $i = 1, 2, \ldots, n$ are only *approximatively* uniform. Thus, any statistical inference based on pseudo-observations should take these features into account.

(3) Now observe that there are various measures of dependence, such as Kendall tau or Spearman rho, which are functions of C, say $\lambda = \delta(C)$. So we might attempt to estimate λ by, say, λ_n and then, formally, derive an estimate for C by $C_n = \delta^{-1}(\lambda_n)$.

For the Kandell tau:

$$\tau(C) = 1 - 4\int_0^1\int_0^1 \left[\frac{\partial C(u,v)}{\partial u}\frac{\partial C(u,v)}{\partial v}\right] dudv$$

and in particular, for Archimedean copulas:

$$\tau(C_\gamma) = 1 + 4\int_0^1\int_0^1 \frac{\varphi_\gamma(t)}{\varphi'_\gamma(t)}dt$$

Thus, in principle, we can (say, numerically!) solve this equation to obtain C_γ.

Note that this approach is in fact *nonparametric* in nature, as we do not postulate any parametric models. If we use Archimedean copulas, then we have a parametric model for the copula, but leave unspecified the marginals (i.e. nonparametric): a situation like this is called *semiparametric*.

Handy estimators of the Kendall tau are in fact available. As the (population) Kandell tau is:

$$\tau(X,Y) = \tau(C) = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

where $(X_1, Y_1)(X_2, Y_2)$ are i.i.d. (X, Y), its empirical counterpart τ_n is obtained as follows.

Let a denote the number of pairs$(X_i, Y_i)(X_j, Y_j)$ in the data set such that $(X_i - X_j)(Y_i - Y_j) > 0$ and let b be the number of observation pairs such that $(X_i - X_j)(Y_i - Y_j) < 0$. Then:

$$\tau_n = \frac{a-b}{\binom{n}{2}}$$

The above "exact" likelihood estimation procedure could be computationally difficult in view of parameter dimension. A two-stage MLE could help. The following estimation procedure is called *inference functions for margins* (IFM).

In view of an above remark, namely:

$$\log L_n(\theta) = \sum_{i=1}^n \log c\big(F_\alpha(X_i), G_\beta(Y_i)\big) + \sum_{i=1}^n \log f_\alpha(X_i) + \sum_{i=1}^n \log g_\beta(Y_i)$$

we could, in a first step, maximize $\sum_{i=1}^n \log f_\alpha(X_i)$ and $\sum_{i=1}^n \log g_\beta(Y_i)$, over parameter spaces for α and β, respectively, to obtain MLE α_n, β_n and then, in a second step, maximize $\sum_{i=1}^n \log c\big(F_{\alpha_n}(X_i), G_{\beta n}(Y_i)\big)$ over the parameter space of γ, to obtain an estimator γ_n and taking $\theta_n = (\alpha_n, \beta_n, \gamma_n)$ as the estimator of θ.

While in general, MLEs in the two above procedures are not the same, although they enjoy similar asymptotic statistical properties, under regular conditions, of course. The IFM method could simplify computations.

Semiparametric estimation

The setting of semiparametric estimation of copulas is this. The joint distribution H of (X, Y) is partially parametric. Specifically, the marginal distributions F and G are left unspecified (i.e. nonparametric), whereas the copula C is parametric, say, C_γ, $\gamma \in \Gamma \subseteq \mathbb{R}^k$, so that:

$$H(x,y) = C_\gamma\big(F(x), G(y)\big)$$

Note that the "parameter" of H is (γ, F, G) where F and G are univariate distribution functions.

As F and G are infinitely dimensional, MLE cannot be used. To use MLE, we need the likelihood, with observed data, in a finitely dimensional setting. As $(F(X_i), G(Y_i))$, $i = 1, 2, \ldots, n$ are not observable, we need their "estimates." That could be achieved by first estimating F, G, *nonparametrically*, say by F_n, G_n, then use the pseudo-observations $(F_n(X_i), G_n(Y_i))$, $i = 1, 2, \ldots, n$ to form an approximate likelihood function for the finitely dimensional parameter γ. More specifically, under the assumption that both F and G are continuous, the copula C_γ is a bivariate distribution with pseudo-observations $(F_n(X_i), G_n(Y_i))$, $i = 1, 2, \ldots, n$. Suppose, in addition, that C_γ is absolutely continuous, we can consider the *pseudo-log likelihood function* of γ when we observed $(F_n(X_i), G_n(Y_i))$, $i = 1, 2, \ldots, n$, as:

$$\log L_n(\gamma) = \sum_{i=1}^{n} \log c_\gamma(F_n(X_i), G_n(Y_i))$$

so that, an estimator of γ could be:

$$\gamma_n = \arg Max_\gamma \sum_{i=1}^{n} \log c_\gamma(F_n(X_i), G_n(Y_i))$$

Note that the nonparametric estimation of F, based on the random sample X_i, $i = 1, 2, \ldots, n$ is the usual empirical distribution:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_i \le x)}(x)$$

Nonparametric estimation

Generalities on nonparametric estimation. When we face an unknown population parameter, finitely or infinitely dimensional, we should try to estimate it in a setting as general as possible, to be as close as possible to *real situations* and should not "impose" conditions, just for the sake of easy computations! Of course, careful data analyzes and domain contexts sometimes do suggest simpler models, such as parametric or semiparametric ones.

In the absence of convincing evidence to consider the two above models, we should try to face the problem as it is, namely, a nonparametric setting.

Let (X_i, Y_i), $i = 1, 2, \ldots, n$ be i.i.d. (X, Y). Without any additional information, we could simply use the empirical counterparts as estimators for population parameters. For example, the value $F(x)$ of the (marginal) distribution function of X is: for fixed $x \in \mathbb{R}$, $F(x) = P(X \le x)$. Its empirical counterpart is the ratio of the observations X_i, $i = 1, 2, \ldots, n$ which are less than x over n, i.e.

$$F_n(x; X_i, i = 1, 2, \ldots n) = \frac{1}{n}\#\{i : X_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty, x]}(X_i)$$

In fact, the justification of using the empirical distribution function $F_n(.)$ to estimate the unknown distribution function $F(.)$, nonparametrically, is deeper than what we mentioned above. Specifically, as $n \to \infty$:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to_{a.s.} 0$$

in other words, $F_n(x)$ converges to $F(x)$ uniformly (in x), almost surely. This convergence means that, for any desired specified error level $\varepsilon > 0$, we have $|F_n(x) - F(x)| \leq \varepsilon$, with probability one, when $n \geq N(\varepsilon)$ (depending only on ε) and that for all $x \in \mathbb{R}$. This is important for applications, as the uniform convergence tells us that we can decide on how large the sample size should be, once for all x.

How large is large? It is a matter of *approximation accuracy*. We need the rate of convergence! For example, if we wish to approximate F by F_n within, say, $\varepsilon = 0.0001$, i.e. $|F_n(.) - F(.)| \leq \varepsilon$, then, knowing its rate of convergence, we infer the minimum n which is considered as large.

Note also that while $F_n(x)$ is a *pointwise* estimator of $F(x)$, the function $F_n(.)$ is a *global* estimator of F, i.e. an estimator of the function $F(.)$. To measure the closeness of F_n to F, we use the Kolmogorov-Smirnov (random) distance:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

The Glivenko-Cantelli theorem asserts that $F_n(.)$ is strongly consistent.

The situation is similar for the joint distribution function H of (X, Y). As $H(x,y) = P(X \leq x, Y \leq y)$, its empirical counterpart is:

$$H_n(x, y) = \frac{1}{n}\#\{i : X_i \leq x, Y_i \leq y\} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty, x] \times (-\infty, y]}(X_i, Y_i)$$

While empirical distribution functions provide reasonable estimators, they are not smooth ones. One way to obtain smooth estimators is to use the method of *kernels* that we will now outline, first as a method for estimating probability density functions.

The method of the kernel for density estimation. The model is $\mathcal{F}$, the class of all absolutely continuous distribution functions. Our random variable of interest X has a distribution function F known only to belong to $\mathcal{F}$. There are many different ways to estimate $f(x) = dF(x)/dx$. Here I will only discuss the most popular one, namely, the kernel method.

Specifically, while $f(x)$ does not have a probabilistic meaning, it is so "asymptotically." Indeed, we have:

$$f(x) = \lim_{h \to 0} \big(F(x + h) - F(x - h)\big)/2h$$

$$= \lim_{h \to 0} P(x - h < X \leq x + h)/2h$$

Now the sample counterpart of $P(x - h < X \leq x + h)/2h$ is:

$$f_n(x) = (1/2nh)\#\{X_i \in (x-h, x+h]\}$$

which is the proportion of the observations falling into the interval $(x-h, x+h]$.

Now, observe that $f_n(x)$ can be also written as:

$$(1/nh)\sum_{i=1}^{n} K\left[(x - X_i)/h\right]$$

where the *kernel* K is:

$$K(x) = \begin{cases} 1/2 & \text{if x} \in [-1, 1) \\ 0 & \text{otherwise} \end{cases}$$

This kernel estimator is called the *naive estimator* of $f(x)$

The kernel is also referred to as a "window" (open around the point x).

The above "naive" kernel is a uniform kernel with weights (degrees of importance) 1/2 assigned to every observation X_i in the window around x. This weight assignment is not sufficiently efficient in the sense that observations closer to x should receive higher weights than those far away. To achieve this, we could modify the kernel accordingly. For example, a kernel of the form:

$$K(x) = (3/4)(1 - x^2)1_{(|x| \le 1)}$$

could reflect an appropriate weight assignment.In any case, the general form for kernel density estimators should be:

$$f_n(x) = (1/nh)\sum_{i=1}^{n} K\left[(x - X_i)/h\right]$$

for some choice of kernel K, i.e. K is a probability density function, i.e. $K(.) \geq 0$ and $\int_{-\infty}^{\infty} K(x)dx = 1$. Note that the naive kernel satisfies the conditions of a probability density function.

Now we have a general form for density estimators, we could proceed ahead to *design* them to obtain "good" estimators. From the above form, we see clearly that estimators' performance depends on the design of the bandwidth h and kernel K.

Thus, from a practical viewpoint, the choice of the bandwidth h is crucial, as h controls the smoothness of the estimator (just like a histogram). Also, as we will see, the smoothness of kernel estimators depends on the properties of K.

Here is the analysis leading to the optimal design of density estimators.

1. Analysis of the bias. The bias of $f_n(x)$ is $b(f_n(x)) = Ef_n(x) - f(x)$. The mean squared error (MSE) is:

$$E\left[f_n(x) - f(x)\right]^2 = Var(f_n) - b^2\left(f_n(x)\right)$$

Assuming that f is sufficiently smooth, such as f'' exists, we get, as $h \to 0$:

$$b\left(f_n(x)\right) = \left(h^2/n\right)f''(x)\int_{-\infty}^{\infty} y^2 K(y)dy + o(h^2)$$

where $o(h^2)$ is a function such that $\lim_{h\to 0} o(h^2)/h^2 = 0$.

Thus, we need to choose K such that $\int_{-\infty}^{\infty} y^2 K(y)dy < \infty$. On the other hand, to make $h \to 0$, we choose h_n (the function of n) so that $h_n \to 0$ when $n \to \infty$.

Looking at the bias, we see that it is proportional to h^2. Thus, to reduce bias, we should choose h small. Also, the bias depends on $f''(x)$, i.e. the curvature of $f(.)$.

2. Analysis of the variance. We have, as $nh \to \infty$:

$$Var\left(f_n(x)\right) = (1/nh)\int_{-\infty}^{\infty} K^2(y)dy + o(1/nh)$$

The variance is proportional to $(nh)^{-1}$. Thus, to reduce the variance, we should choose h large! Also the variance increases with $\int_{-\infty}^{\infty} K^2(y)dy$: flat kernels should reduce the variance. Of course, increase the sample size n reduce the variance.

How to balance the choice of h for reducing bias and variance? Note that increasing h will lower the variance but increases the bias and vice versa.

A compromise is to minimize the MSE. Now:

$$MSE\left(f_n(x)\right) = \left(h^4/4\right)\left[f''(x)\right]^2\left[\int_{-\infty}^{\infty} y^2 K(y)dy\right]^2 + (1/nh)\left[\int_{-\infty}^{\infty} K^2(y)dy\right]f(x) + o(1/nh)$$

as $nh \to \infty$.

Thus, $MSE(f_n(x)) \to 0$, as $n \to \infty$ (i.e. $f_n(x)$ is MSE-consistent, and hence weakly consistent) when $h_n \to 0$ and $nh_n \to \infty$, as $n \to \infty$.

The optimal choice of bandwidth is $h_n = n^{-1/5}$ and the convergence rate is $n^{-4/5}$.

All of the above could give you a "flavor" of designing kernel estimators!

To summarize general results on asymptotic properties of kernel estimators, I list the following: Under suitable choices of K as indicated above and suppose f continuous:

- $f_n(x)$ is asymptotically unbiased when $h_n \to \infty$, as $n \to \infty$
- $f_n(x)$ is weakly consistent for $f(x)$ provided $h_n \to 0$ and $nh_n \to \infty$

Remark on methods of density estimation

The method of the kernel for density estimation is only one among a variety of other methods, such as orthogonal functions, excess mass. It is a popular approach. We illustrate briefly the recent method of *excess mass* which seems not to be well-known in econometrics.

This method has the advantage that we do not need to assume analytic conditions on f but only some information about its shapes.

For each $\alpha \geq 0$, a crossed section of f (say, in multivariate case) at level α is:

$$A_\alpha(f) = \{x \in \mathbb{R}^d : f(x) \geq \alpha\}$$

A piece of information about the shape of f could be $A_\alpha \in \mathcal{C}$, some specified class of geometric objects, such as ellipsoids (e.g. multivariate normal).

Now observe that f can be recovered from the $A_\alpha^{'}s$ as:

$$f(x) = \int_0^{\infty} 1_{A_\alpha}(x)d\alpha$$

so that it suffices to estimate the *sets* A_α, $\alpha \geq 0$ by some *set-statistics* $A_{\alpha,n}$ (i.e. some *random set statistics)* and use the plug-in estimator:

$$f_n(x) = \int_0^\infty 1_{A_{\alpha,n}}(x)d\alpha$$

to obtain an estimator for $f(x)$.

But of course, the question is how to obtain $A_{\alpha,n}$ from the sample $X_1, X_2, \ldots, X_n$?

Let $\lambda(dx)$ denote the Lebesgue measure on $\mathbb{R}^d$. Then:

$$(dF - \alpha\lambda)(A) = \int_A f(x)dx - \alpha\int_A dx = \mathcal{E}_\alpha(A)$$

is the excess mass of the set A at level α. Note that $(dF - \alpha\lambda)$ is a *signed measure.*

Theorem. A_α maximizes $\mathcal{E}_\alpha(A)$ over $A \in \mathcal{B}(\mathbb{R}^d)$.

Proof. For each $A \in \mathcal{B}(\mathbb{R}^d)$, write $A = (A \cap A_\alpha) \cup (A \cap A_\alpha^c)$, where A_α^c is the set complement of A_α. Then:

$$\mathcal{E}_\alpha(A) = \int_{A\cap A_\alpha} \big(f(x) - \alpha\big)dx + \int_{A\cap A_\alpha^c} \big(f(x) - \alpha\big)dx$$

Now, on $A \cap A_\alpha$, $f(x) - \alpha \geq 0$ so that:

$$\int_{A\cap A_\alpha} \big(f(x) - \alpha\big)dx \leq \int_{A_\alpha} \big(f(x) - \alpha\big)dx (A \cap A_\alpha \subseteq A_\alpha)$$

On $A \cap A_\alpha^c$, $f(x) - \alpha \leq 0$. Thus:

$$\int_{A\cap A_\alpha} \big(f(x) - \alpha\big)dx + \int_{A\cap A_\alpha^c} \big(f(x) - \alpha\big)dx \leq \int_{A\cap A_\alpha} \big(f(x) - \alpha\big)dx \leq \mathcal{E}_\alpha(A)$$

Just like MLE, this maximization result suggests a method for estimating $A_\alpha(f)$.

The empirical counterpart of $\mathcal{E}_\alpha(A)$ is:

$$\mathcal{E}_{\alpha,n}(A) = (dF_n - \alpha\lambda)(A)$$

Thus, a plausible estimator of the α-level set $A_\alpha(f)$ is the random set $A_{\alpha,n}$ maximizing $\mathcal{E}_{\alpha,n}(A)$ over $A \in \mathcal{C}$.

While the principle is simple, the rest is not!

Nonparametric regression

Let $(X_i, Y_i), i = 1, 2, \ldots, n$ be a random sample from a bivariate distribution with joint density $f(x, y)$.

The marginal density of X is:

$$f_X(x) = \int_{-\infty}^\infty f(x, y)dy$$

and the conditional density of Y given $X = x$ is:

$$f_{Y|X}(y|x) = f(x,y)/f_X(x)$$

The conditional mean (or regression of Y on X) is:

$$r(x) = E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} y f(y,x)/f_X(x)dy$$

Using kernel estimation for both densities $f(x, y)$ and $f(x)$, we arrive at the following.

The kernel estimator of $f(x, y)$ is of the form:

$$f_n(x,y) = \left(nh_n^2\right)^{-1} \sum_{i=1}^{n} K\left[(x - X_i)/h_n, (y - Y_i)/h_n\right]$$

and that of $f(x)$ is:

$$\hat{f}_n(x) = (nh_n)^{-1} \sum_{i=1}^{n} J\left[(x - X_i)/h_n\right]$$

Note that $\hat{f}_n(x) = \int_{-\infty}^{\infty} f_n(x,y)dy$.

From the above a kernel estimator of $r(x)$ is:

$$r_n(x) = \int_{-\infty}^{\infty} f_n(x,y)/\hat{f}_n(x)dy$$

For simplicity, we can take:

$$K(x,y) = J(x)J(y)$$

and arrive at:

$$r_n(x) = \sum_{i=1}^{n} Y_i J\left[(x - X_i)/h_n\right] / \sum_{i=1}^{n} J\left[(x - X_i)/h_n\right]$$

Nonparametric estimation of copulas

From a natural setting, nonparametric estimation of the copula C of (X, Y) from the observations $(X_i, Y_i), i = 1, 2, \ldots, n$ can be set up as follows.

As $C(u,v) = H(F^{-1}(u), G^{-1}(v))$, the empirical copula is:

$$C_n(u,v) = H_n\left(F_n^{-1}(u), G_n^{-1}(v)\right)$$

where H_n, F_n and G_n are empirical distributions and:

$$F_n^{-1}(u) = \inf\{x \in \mathbb{R} : F_n(x) \geq u\}$$

Of course, good statistical properties of this nonparametric estimator should be established, e.g. under regular conditions (i.e. analytic properties of the model giving rise to the observed data).

To obtain smooth estimators for C, one could use smooth versions of the above empirical distributions, using, say, the kernel method. For example, we replace the empirical H_n by:

$$H_n^*(x,y) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{a_n}, \frac{y - Y_i}{a_n}\right)$$

where K is some kernel and the sequence of bandwidths a_n properly chosen, such as $\lim_{n\to\infty} a_n = 0$.

When the copula C is absolutely continuous, we can consider the estimation of its density function $c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$.

Now $c(.,.)$ is a density with *finite support* $[0,1]^2$ in $\mathbb{R}^2$ and we intend to estimate it (pointwise) by using *pseudo-observations* $(F_n(X_i), G_n(Y_i))$, $i = 1, 2, \ldots, n$, where F_n, G_n are empirical distributions, some care should be taken to achieve consistency of its estimator when we use, say, kernel method for estimation (see remarks below).

First, we modify the empirical univariate distribution functions as:

$$F_n(x) = \frac{1}{n+1}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i)$$

$$G_n(y) = \frac{1}{n+1}\sum_{i=1}^{n} 1_{(-\infty,y]}(Y_i)$$

Note that $F_n(X_i)$ is the rank of X_i (among the $X_1, X_2, \ldots, X_n$) divided by $n + 1$, so that the pseudo-observations $F_n(X_i)$, $i = 1,2, \ldots, n$ are $\{\frac{i}{n+1} : i = 1, 2, \ldots, n\}$.

Thus, next, we could take:

$$c_n(u,v) = \frac{1}{na_n^2}\sum_{i=1}^{n} K\left(\frac{u - F_n(X_i)}{a_n}, \frac{v - G_n(Y_i)}{a_n}\right)$$

as a nonparametric estimator of $c(u, v)$, for $(u,v) \in [0,1]^2$.

Corresponding author
Hung T. Nguyen can be contacted at: hunguyen@nmsu.edu