A Copula Function Approach to Infer Correlation in Prediction Markets

Agostino Capponi[1] and Umberto Cherubini[2]

[1] California Institute of Technology,
Division of Engineering and Applied Sciences
Pasadena CA 91125, USA
acapponi@caltech.edu.edu

[2] University of Bologna,
Department of Mathematical Economics
Bologna 40126, Italy
umberto.cherubini@unibo.it

Abstract. We propose the use of copula methods to recover the dependence structure between prediction markets. Copula methods are flexible tools to measure associations among probabilities because they encompass both linear and non linear relationship among variables. We apply the proposed methodology to three actual prediction markets, the Saddam Security, the market of oil spot prices and the Saddameter. We find that the Saddam Security is nearly independent of the oil market, while being highly correlated to the Saddameter. The results obtained appear to suggest that the Saddam Security prediction market may be noisy or overlooking some political factors which are instead considered by Saddameter and the oil market.

1 Introduction

Prediction markets, also called information markets, are a mechanism to aggregate information from widely dispersed economic actors with the objective to produce prediction about future events [7] . The market price reflects a stable consensus of a large number of opinions about the likelihood of the event. This has also been analytically verified in [9], where it is shown that for a large class of models, prediction market prices correspond with mean beliefs of market actors.

The range of applications is quite broad, from helping businesses making better investment decisions, to helping governments making better policy decisions on health care, monetary policy, etc... [1]. There are three main types of contracts:

- "winner-take-all": such contract is linked to the realization of a specific event, it costs some amount p to enter and pays off 1 if and only if that event occurs.
- "index": the payoff of such contract varies continuously based on a quantity that fluctuates over time, like the percentage of the vote received by a candidate.

S. Das et al. (Eds.): Amma 2009, LNICST 14, pp. 4–12, 2009.

- "spread betting": traders bid on the cutoff that determine whether an event occurs, for example whether one football team will win by a number of points larger than a certain threshold.

The basic forms of these contracts will reveal the market consensus about the probability, mean or median of a specific event. However, by appropriate combinations of these markets, it is possible to evaluate additional statistics. For example, a family of winner-take-all contracts paying off if and only if a football team loses by $1, 2, \ldots, n$ points or wins by $1, 2 \ldots, n$ points would reveal the whole distribution on the outcome of the game.

The contracts described above depend on only one outcome. The same principles can be applied to contracts depending on the joint outcomes of multiple events. Such contracts are called contingent and provide insights into the correlation between different events. For instance, Wolfers and Zitzewitz [8] ran a market linked to whether George W. Bush would be re-elected, another market on whether Osama Bin Laden would be captured prior to the election, and a third market on whether both events would have occurred. Their findings were that there was a 91 % chance of Bush being reelected if Osama had been found, but a 67 % unconditional probability.

This paper proposes an approach to imply the correlation between two prediction markets, each linked to the realization of one specific event, without introducing a third market on the joint outcome of the two events. The rest of the paper is organized as follows. Section 2 recalls the basics of copula functions and proposes a methodology based on the Kendall function to recover the dependence structure of two prediction markets. Section 3 applies the methodology described in the previous section to infer the pairwise dependence structure of the prediction market of oil, the Saddam Security and the Saddameter. Section 4 concludes the paper.

2 Correlation Methods Based on Copulas

In this section we briefly recall the basics of copula functions and refer the reader to a textbook on copulas [5] and [2] for a more detailed treatment. Copula functions represent a general and flexible tool to measure association among probabilities. Association is a more general concept than correlation, encompassing both linear and non linear relationship among variables. In fact, while correlation measures linear relationships between variables, copula functions are invariant to whatever linear or non linear transformation of the variables. Applying linear correlation to the evaluation of non linear relationships may lead to outright blunders, as the following textbook example illustrates. Simply take x with standard normal distribution and defines $y = x^2$. Obviously x and y are associated, but computing covariance one gets

$$E(xy) - E(x)E(y) = E(x^3) = 0 \tag{1}$$

Copula functions enable to overcome these problems by studying the dependence among marginal distributions instead of variables.

2.1 Copula Functions

The basic idea is based on the *principle of probability integral transformation.* Take variables X and Y with marginal distributions F_X and F_Y. The principle states that transformations $u \equiv F_X(X)$ and $v \equiv F_Y(Y)$ have uniform distribution in $[0, 1]$. Marginal distributions can be inverted, and if they are continuous such inverse is unique. Take now the joint distribution $H(X, Y)$. This can be written as

$$H(X, Y) = H(F_X^{-1}(u), F_Y^{-1}(v)) \equiv C(u, v) \tag{2}$$

where $C(u, v)$ is called the copula function representing association between X and Y. So, every joint distribution can be written as a function taking the marginal distributions as arguments. On the contrary, one can prove that if $C(u, v)$ has suitable properties, then plugging univariate distributions into it generates a joint distribution. The requirements for $C(u, v)$ are summarized in what is known as Sklar's theorem, and are reported below.

Definition 1. *A copula function $C(u, v)$ has domain in the 2-dimension unit hypercube and range in the unit interval and satisfies the following requirements:*

1. ***Groundedness:*** $C(u, 0) = C(0, v) = 0$
2. ***Identity of marginals:*** $C(u, 1) = v$, $C(1, v) = u$
3. ***2-increasing:*** $C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2) \geq 0$, *with* $u_1 > u_2$,$v_1 > v_2$.

In case when the random variables X and Y are perfectly positively associated, i.e. $Y = T(X)$, with T being a monotonically increasing transformation, we have that

$$C(u, v) = \min(u, v)$$

The value $\min(u, v)$ is also the maximum value that a copula can achieve and is commonly referred as *upper Fréchet bounds.* In case when the random variables X and Y are uncorrelated, we obtain the *independence copula* given by

$$C(u, v) = uv$$

In case when X and Y are perfectly negative associated, $Y = L(X)$, with L being a monotonically decreasing transformation then

$$C(u, v) = \max(u + v - 1, 0)$$

The value $\max(u + v - 1, 0)$ is commonly referred to as *lower Fréchet bounds.* Copula functions are non-parametric representations of the dependence structure and as such they are linked to non parametric association measures such as *Spearman's* ρ_S (also called rank correlation) or *Kendall's* τ (also called concordance measure). We next discuss the relation to the *Kendall's* τ coefficient which will be used in this paper. Let (X_1, X_2),(Y_1, Y_2) be independent random pairs with distribution F and marginal distribution F_1 and F_2. Then the Kendall's τ is defined as

$$\tau = \mathbb{P}((X_1 - Y_1)(X_2 - Y_2) > 0) - \mathbb{P}((X_1 - Y_1)(X_2 - Y_2) < 0) \tag{3}$$

The condition $(X_1 - Y_1)(X_2 - Y_2) > 0$ corresponds to (X_1, X_2), (Y_1, Y_2) being two concordant pairs, i.e. one of the two pairs has the larger value for both components, while the condition $(X_1-Y_1)(X_2-Y_2) < 0$ corresponds to (X_1, X_2), (Y_1, Y_2) being two discordant pairs in that for each pair one component is larger than the corresponding component of the other pair and the other component is smaller. Therefore, Kendall's τ is the difference between the probability of two random concordant pairs and the probability of two random discordant pairs. Eq. (3) can be developed further to illustrate the relation with the associated copula as follows

$$\begin{aligned}\tau &= \Pr((X_1 - Y_1)(X_2 - Y_2) > 0) - \mathbb{P}((X_1 - Y_1)(X_2 - Y_2) < 0) \\ &= 4\int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1 \end{aligned} \tag{4}$$

This follows because

$$\begin{aligned}&\mathbb{P}((X_1 - Y_1)(X_2 - Y_2) > 0) - \mathbb{P}((X_1 - Y_1)(X_2 - Y_2) < 0) &= \\ &2\mathbb{P}((X_1 - Y_1)(X_2 - Y_2) > 0) - 1 &= \\ &2(\mathbb{P}(Y_1 \leq X_1, Y_2 \leq X_2) + \mathbb{P}(X_1 \leq Y_1, X_2 \leq Y_2)) - 1 & \end{aligned} \tag{5}$$

which can be developed further as

$$\begin{aligned}\mathbb{P}(Y_1 \leq X_1, Y_2 \leq X_2) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{P}(Y_1 \leq x_1, Y_2 \leq y_2) dC(F_1(x_1), F_2(x_2)) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} C(F_1(x_1), F_2(x_2)) dC(F_1(x_1), F_2(x_2)) \\ &= \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2)\end{aligned}$$

The same argument applies to $\mathbb{P}(X_1 \leq Y_1, X_2 \leq Y_2)$ since (X_1, X_2) and (Y_1, Y_2) are two independent pairs from the same distribution, thus implying Eq. (4).

Let $(V_1, W_1), (V_2, W_2), \ldots, (V_n, W_n)$ available samples from the joint distribution H. Let us define the empirical measures

$$\begin{aligned}K_i^- &= \sharp\{(X_j, Y_j) : X_j < X_i, Y_j < Y_i\} \\ K_i^+ &= \sharp\{(X_j, Y_j) : X_j > X_i, Y_j > Y_i\} \\ K_i &= K_i^- + K_i^+ \\ \bar{K}_i &= \sharp\{(X_j, Y_j) : X_j > X_i, Y_j < Y_i\} + \sharp\{(X_j, Y_j) : X_j < X_i, Y_j > Y_i\}\end{aligned} \tag{6}$$

where $\sharp$ is the cardinality of the set $\{.\}$. The population version of the association measure τ in Eq. (4) is given by

$$\tau_{pop} = \frac{\sum_{i=1}^n K_i - \bar{K}_i}{\sum_{i=1}^n K_i + \bar{K}_i} \tag{7}$$

The class of copulas used in this paper to represent cross-section dependence is that of *Archimedean copulas.* Each copula of this class is generated using a function $\phi(x)$ as

$$C(u, v) \equiv \phi^{[-1]} \left(\phi(u) + \phi(v)\right) \tag{8}$$

where $\phi(s)$ is called the *generator function* of the copula C. Here ϕ is a continuous, strictly decreasing function from $[0,1]$ to $[0,\infty]$ such that $\phi(1)=0$. The function $\phi^{[-1]}$ is called the pseudo-inverse of ϕ, it maps from the domain $[0,\infty]$ to $[0,1]$ and is given by

$$\phi^{[-1]}(t)\begin{cases}\phi^{-1}(t) & 0\le t\le\phi(0)\\ 0 & \phi(0)\le t\le\infty\end{cases} \tag{9}$$

2.2 Kendall Function

The principle of probability integral transformation is intrinsically univariate. Even if we know that $u=F_X(X)$ and $v=F_Y(Y)$ are uniformly distributed, we cannot say what is the joint distribution of $C(u,v)$. For the class of Archimedean copulas we may provide an interesting bivariate extension. For such class of copulas, the multivariate probability integral transformation is known as *Kendall function*, or *Kendall distribution.* The distribution of $C(u,v)$ is given by

$$\begin{aligned}KC(t)&=\Pr(C(u,v)\le t)\\&=t-\frac{\phi(t)}{\phi'^{+}(t)}\end{aligned} \tag{10}$$

where $\phi'^{+}(t)$ denotes the right derivative of the copula generator.

We can compute the Kendall function for the copula $C(u,v)=\min(u,v)$ of perfect positive dependence as follows. We have

$$\begin{aligned}\Pr(\min(u,v)\le t)&=\Pr(\min(F_X(X),F_Y(Y))\le t)\\&=\Pr(F_X(X)\le t)+\Pr(F_Y(Y)\le t)-\Pr(F_X(X)\le t,F_Y(Y)\le t)\\&=t+t-\Pr(\min(F_X(X),F_Y(Y))\le t)\end{aligned} \tag{11}$$

which implies that $\Pr(\min(u,v)\le t)=t$, thus the Kendall function is given by

$$KC(t)=t \tag{12}$$

and consequently the distribution is uniform as in the univariate case. The generator $\phi(t)=-\log(t)$ generates the independence copula $C(u,v)=uv$, as it is next shown

$$\begin{aligned}C(u,v)&=\phi^{[-1]}(\phi(u)+\phi(v))\\&=e^{-(-\log(u)-\log(v))}\\&=e^{\log(uv)}\\&=uv\end{aligned} \tag{13}$$

and, using Eq. (10) we have that the Kendall function is given by

$$KC(t)=t-t\ln(t) \tag{14}$$

The generator $\phi(t) = 1 - t$ generates the copula $C(u,v) = \max(u + v - 1, 0)$ of perfect negative dependence, as it is next shown. It can be checked that $\phi^{-1}(t) = \max(1 - t, 0)$, which in turn implies

$$\begin{aligned} C(u,v) &= \phi^{[-1]}(\phi(u) + \phi(v)) \\ &= \max(1 - [(1-u) + (1-v)], 0) \\ &= \max(u + v - 1, 0) \end{aligned} \tag{15}$$

and, using Eq. (10) we have that the Kendall function is given by

$$KC(t) = 1 \tag{16}$$

Kendall functions can be used to evaluate the dependence functions, and to gauge how far or close they are to the case of perfect dependence or independence. This can be done by computing the sample equivalent of the Kendall function which is

$$KC_i = \frac{K_i^-}{N-1} \tag{17}$$

Plotting the sample equivalent of the Kendall function can give a graphical idea of the dependence structure, and how it changes with levels of the marginal distributions. We will use this tool to calculate the correlation between three actual prediction markets in the next section.

3 An Empirical Analysis of Saddam Market and Oil

We apply the methodology described in earlier sections to infer correlation between the following three prediction markets:

- "Saddam Security". This was a contract offered on TradeSports paying $100 if Saddam Hussein were ousted from power by the end of June 2003.
- "Saddameter". This was an expert journalists estimate of the probability of the United States going to war with Iraq.
- Spot oil price. These are the spot prices of futures on oil.

For each pair of individual markets, the data are processed as follows. We first align the time series, so that they all start and end at the same date. Let n be length of the aligned time series. We construct an equally spaced grid of points in the interval $[0,1]$ where the distance between two consecutive points is $\Delta = 0.05$. We then calculate a frequency distribution such that the frequency associated with the i-th point in the grid is the number of indices $j \in \{1, \ldots, n\}$ such that $KC_j \leq \Delta i$, where KC_j has been defined in Eq. (17). We refer to this frequency distribution as the empirical copula.

Figures 1, 2 and 3 display the empirical Kendall function KC_i relating respectively the Saddam security with the oil price, the Saddam security with the Saddameter and the Saddameter with the oil price.

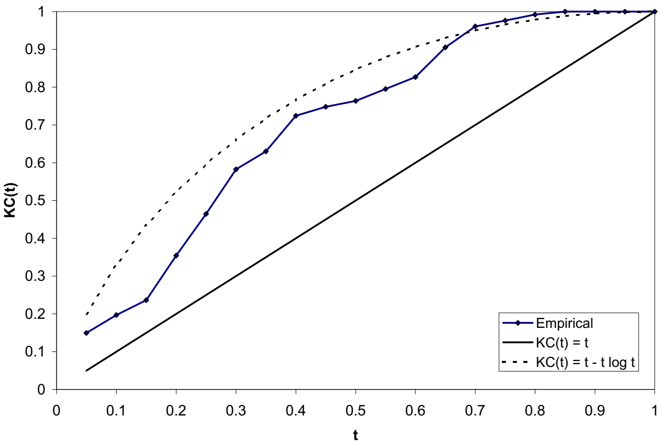

Fig. 1. Empirical kendall function $KC(t)$ between Saddam Security and oil price

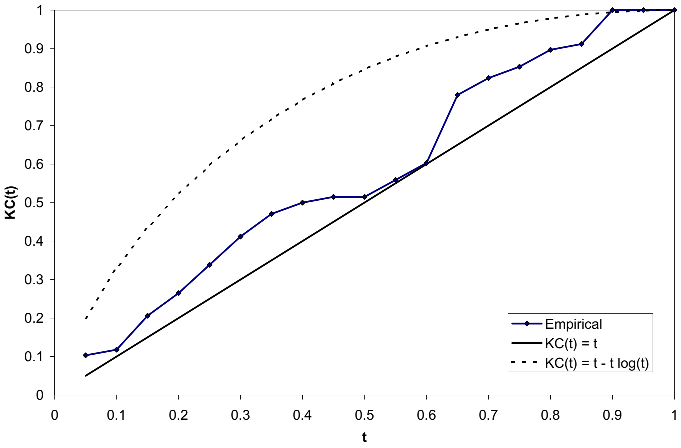

Fig. 2. Empirical kendall function $KC(t)$ between Saddam Security and Saddameter

It appears from Figure 1 that the prediction market evaluation of the Iraq war estimated by the Saddam security is nearly independent from the price of oil. This can be deduced by looking at the empirical copula which tends to lie closer to the curve $KC(t) = t - t\ln(t)$, representing the case of independence. The population value of the concordance measure τ_{pop} which evaluates to 25% reinforces the claim of independence. A closer look at Figure 1 shows that the independence is especially evidenced towards the tails (*tail independence*); when the Saddam security attributes large probability to Saddam being ousted, the oil market does not respond with increasing quotes of the oil price, and viceversa.

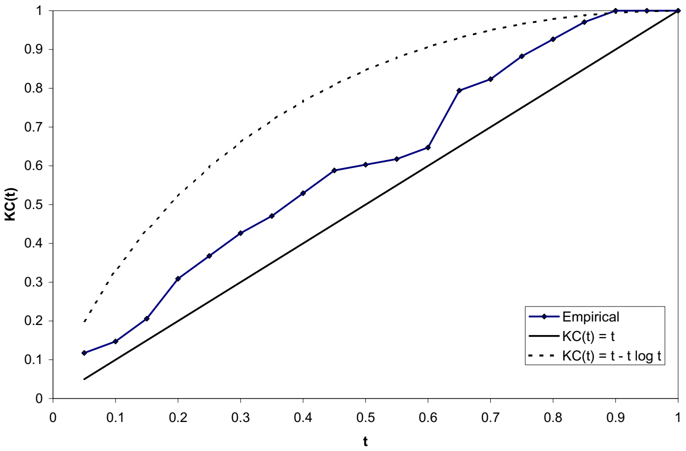

Fig. 3. Empirical kendall function $KC(t)$ between oil price and Saddameter

After repeating the same calculation for the case of the Saddam security and the Saddameter, we can see that an opposite argument takes place, and both expert evaluations and the Saddam security are in agreement with respect to the Iraq war. Figure 2, infact, shows that the empirical KC_i function tends to lie closer to the curve of perfect dependence $KC(t) = t$. For values of t closer to 0.5, the empirical curve lies on the curve of perfect dependence, whereas for extreme values of t, the empirical curve K_i tends to divert towards the curve of independence. The general behavior of positive correlation is also evidenced by the population value of the τ statistic which is given by $\tau_{pop} = 81\%$.

Figure 3 shows the empirical kendall function relating the Saddameter and the oil price. Even in this case, the empirical KC_i function tends to lie closer to the curve $KC(t) = t$, although not as close as for the case of the Saddam security and the Saddameter. Even in this case there is a tendency to move towards the curve of perfect independence for large values of t. The population value of the τ statistic which is given by $\tau_{pop} = 65\%$ confirms the graphical illustration of correlation.

The high correlation between Saddameter and oil price and the low correlation between the Saddam security and the oil price lead us to formulate the following hypotheses:

- *Hidden factors.* Journalist experts and the oil market may have looked at factors affecting the war in Iraq which were not considered by the prediction market of the Saddam security. In other words, there might have been a common political factor which drove both expert evaluations and traders in the oil market which was ignored by traders of the Saddam security.
- *Market Noise.* The prediction market of the Saddam security is noisy. Such noise makes the event of Iraq war independent of increases in the oil price. This noise may be caused by the presence of a sufficient number of “liquidity” traders who act like gamblers, and generate speculative bubbles by betting on unlikely

events in the hope of generating high returns. Those traders contrast with rational traders whose only motivation is expected return. If the market were only composed by the rational traders, then all information would be aggregated and fully reflected in prices and the No Trade Theorem would apply [4], [3].

4 Conclusions

In this paper, we have introduced an approach based on copula methods to infer dependence among prediction markets. Such methods are flexible enough to deal with non-linearity in the data. They are non-parametric, and thus do not require to impose any structure before performing the analysis. We have applied the methodology to infer dependence among three markets, the prediction market of the Saddam Security, the Saddameter, and the market of the spot future prices on oil. We found that the Saddam security is nearly independent of the oil market, while being highly correlated with the journalist expert evaluations provided by the Saddameter. We plan to perform a more detailed analysis on the correlation of those three markets as well as applying the proposed approach to other prediction markets and evaluate its robustness and predictive power.

Ackowledgments

Agostino Capponi would like to thank Prof. Justin Wolfers for sharing with him market quotes of the Saddam security, Saddameter, and spot oil prices.

References

1. Arrow, K.J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J., Levmore, S., Litan, R., Milgrom, P., Nelson, F.D., Neumann, G.R., Ottaviani, M., Schelling, T.C., Shiller, R.J., Smith, V.L., Snowberg, E., Sunstein, C.R., Tetlock, P.C., Tetlock, P.E., Varian, H.R., Wolfers, J., Zitzewitz, E.: The promise of Prediction Markets. Science 16, 320(5878), 877–878 (2008)
2. Cherubini, U., Luciano, E., Vecchiato, W.: Copula Methods in Finance. Wiley Finance, Chichester (2004)
3. Grossman, S., Stiglitz, J.: Information and Competitive Price Systems. The American Economic Review 66(2), 246–253 (1976)
4. Milgrom, P., Stokey, N.: Information, trade and common knowledge. Journal of Economic Theory 26(1), 17–27 (1982)
5. Nelsen, R.: An Introduction to Copulas. Springer, Heidelberg (2006)
6. Wolfers, J., Zitzewitz, E.: Prediction Markets. Journal of Economic Perspectives 18(2) (2004)
7. Wolfers, J., Zitzewitz, E.: Experimental Political Betting Markets and the 2004 Election. The Economists' Voice 1(2) (2004)
8. Wolfers, J., Zitzewitz, E.: Information Markets: A New Way of Making Decisions in the Public and Private Sectors. In: Hahn, R., Tetlock, P. (eds.). AEI-Brookings Press
9. Wolfers, J., Zitzewitz, E.: Intrepreting Prediction Market Prices as Probabilities. Under Review in Review of Economics and Statistics