Predictiv ith Copulas for Bivariate Data

Noryanti Muhammad

A Thesis presented for the degree of
Doctor of Philosophy

Durham University

Statistics and Probability Research Group
Department of Mathematical Sciences
University of Durham
England

February 2016

Contents

Predictiv ith Copulas for Bivariate Data

Noryanti Muhammad

A Thesis presented for the degree of
Doctor of Philosophy

Durham University

Statistics and Probability Research Group
Department of Mathematical Sciences
University of Durham
England

February 2016

Predictive Inference with Copulas for Bivariate Data

Noryanti Muhammad

Submitted for the degree of Doctor of Philosophy
February 2016

Abstract

Nonparametric predictive inference (NPI) is a statistical approach with strong frequentist properties, with inferences explicitly in terms of one or more future observations. NPI is based on relatively few modelling assumptions, enabled by the use of lower and upper probabilities to quantify uncertainty. While NPI has been developed for a range of data types, and for a variety of applications, thus far it has not been developed for multivariate data. This thesis presents the first study in this direction. Restricting attention to bivariate data, a novel approach is presented which combines NPI for the marginals with copulas for representing the dependence between the two variables. It turns out that, by using a discretization of the copula, this combined method leads to relatively easy computations. The new method is introduced with use of an assumed parametric copula. The main idea is that NPI on the marginals provides a level of robustness which, for small to medium-sized data sets, allows some level of misspecification of the copula.

As parametric copulas have restrictions with regard to the kind of dependency they can model, we also consider the use of nonparametric copulas in combination with NPI for the marginals. As an example application of our new method, we consider accuracy of diagnostic tests with bivariate outcomes, where the weighted combination of both variables can lead to better diagnostic results than the use of either of the variables alone. The results of simulation studies are presented to provide initial insights into the performance of the new methods presented in this thesis, and examples using data from the literature are used to illustrate applications

of the methods. As this is the first research into developing NPI-based methods for multivariate data, there are many related research opportunities and challenges, which we briefly discuss.

Bibliography

[1] Augustin T. and Coolen F.P.A. (2004). Nonparametric Predictive Inference and Interval Probability. *Journal of Statistical Planning and Inference*, 124(2), 251–272.

[2] Augustin T., Coolen F.P.A., de Cooman G. and Troffaes M.C.M. (2014). *Introduction to Imprecise Probabilities.* Chichester: Wiley.

[3] Baker R.M. and Coolen F.P.A. (2010). Nonparametric predictive category selection for multinomial data. *Journal of Statistical Theory and Practice*, 4(3), 509–526.

[4] Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4), 387–415.

[5] Bansal A. and Sullivan Pepe M. (2013). When does combining markers improve classification performance and what are implications for practice? *Statistics in Medicine*, 32(11), 1877–1892.

[6] Barnett V. and Lewis T. (1994). *Outliers in Statistical Data.* Chichester: Wiley, 3rd edition.

[7] Bellotti T. and Crook J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36(2), 3302–3308.

[8] Bowman A.W. and Azzalini A. (2004). *Applied Smoothing Techniques for Data Analysis.* Oxford: Clarendon Press.

[9] Bradley A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145–1159.

[10] Breiman L., Meisel W. and Purcell E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, 19(2), 135–144.

[11] Cacoullos T. (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18(1), 179–189.

[12] Charpentier A., Fermanian J.D. and Scaillet O. (2007). The estimation of copulas: Theory and practice. In *Copulas: From theory to application in finance*, (Editor) R. Jörn, pp. 35–62. London: Risk Books.

[13] Chen X., Fan Y. and Tsyrennikov V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, 101(475), 1228–1240.

[14] Cherubini U., Luciano E. and Vecchiato W. (2004). *Copula Methods in Finance*. Chichester: John Wiley & Sons.

[15] Clayton D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65, 141–151.

[16] Coolen F.P.A. (1996). Comparing two populations based on low stochastic structure assumptions. *Statistics & Probability Letters*, 29(4), 297–305.

[17] Coolen F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, 36(4), 349–357.

[18] Coolen F.P.A. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15(1-2), 21–47.

[19] Coolen F.P.A. (2011). Nonparametric predictive inference. In *International Encyclopedia of Statistical Science*, (Editor) M. Lovric, pp. 968–970. Berlin, Heidelberg: Springer Berlin Heidelberg.

[20] Coolen F.P.A. and Augustin T. (2005). Learning from multinomial data: A nonparametric predictive alternative to the Imprecise Dirichlet Model. In *ISIPTA'05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications.*, (Editors) F. Cozman, R. Nau and T. Seidenfeld, volume 5, pp. 125–134. SIPTA.

[21] Coolen F.P.A. and Augustin T. (2009). A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories. *International Journal of Approximate Reasoning*, 50(2), 217–230.

[22] Coolen F.P.A., Coolen-Schrijner P. and Yan K.J. (2002). Nonparametric predictive inference in reliability. *Reliability Engineering & System Safety*, 78(2), 185–193.

[23] Coolen F.P.A., Troffaes M.C. and Augustin T. (2011). Imprecise probability. In *International Encyclopedia of Statistical Science*, (Editor) M. Lovric, pp. 645–648. Berlin, Heidelberg: Springer Berlin Heidelberg.

[24] Coolen F.P.A. and Yan K.J. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126(1), 25–54.

[25] Coolen-Maturi T., Coolen F.P. and Muhammad N. (2016). Predictive Inference for Bivariate Data: Combining Nonparametric Predictive Inference for Marginals with an Estimated Copula. *Journal of Statistical Theory and Practice*, (just-accepted).

[26] Coolen-Maturi T., Coolen-Schrijner P. and Coolen F.P.A. (2012). Nonparametric predictive inference for binary diagnostic tests. *Journal of Statistical Theory and Practice*, 6(4), 665–680.

[27] Coolen-Maturi T., Coolen-Schrijner P. and Coolen F.P.A. (2012). Nonparametric predictive inference for diagnostic accuracy. *Journal of Statistical Planning and Inference*, 142(5), 1141 – 1150.

[28] Copas J. and Corbett P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika*, 89(2), 315–331.

[29] Cortes C. and Vapnik V. (1995). Support-vector networks. *Machine learning*, 20(3), 273–297.

[30] De Finetti B. (1974). *Theory of Probability: A Critical Introductory Treatment.* London: Wiley.

[31] Deheuvels P. (1980). Non parametric tests of independence. In *Statistique non Paramétrique Asymptotique: Actes des Journées Statistiques, Rouen, France, Juin 1979*, (Editor) J.P. Raoult, pp. 95–107. Berlin, Heidelberg: Springer Berlin Heidelberg.

[32] Donoho D.L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pp. 1–32. Http://statweb.stanford.edu/ donoho/Lectures/CBMS/Curses.pdf.

[33] Efron B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382), 316–331.

[34] Elkhafifi F.F. (2012). *Nonparametric Predictive Inference for Ordinal Data and Accuracy of Diagnostic Tests.* Ph.D. thesis, Durham University, Durham, UK. Available from www.npi-statistics.com.

[35] Elkhafifi F.F. and Coolen F.P.A. (2012). Nonparametric Predictive Inference for accuracy of ordinal diagnostic tests. *Journal of Statistical Theory and Practice*, 6(4), 681–697.

[36] Embrechts P., Lindskog F. and Mcneil A. (2003). Modelling dependence with copulas and applications to risk management. In *Handbook of Heavy Tailed Distributions in Finance*, (Editor) S.T. Rachev, volume 1, pp. 329 – 384. Amsterdam: North-Holland.

[37] Esteban L.M., Sanz G. and Borque A. (2011). A step-by-step algorithm for combining diagnostic tests. *Journal of Applied Statistics*, 38(5), 899–911.

[38] Frank M.J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae*, 19, 194–226.

[39] Frees E.W. and Valdez E.A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2, 1–25.

[40] Geenens G., Charpentier A. and Paindaveine D. (2014). Probit transformation for nonparametric kernel estimation of the copula density. *arXiv preprint arXiv:1404.4414*. Http://arxiv.org/abs/1404.4414.

[41] Genest C. and Favre A.C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4), 347–368.

[42] Genest C., Ghoudi K. and Rivest L.P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3), 543–552.

[43] Ghosh D. (May 2004). Semiparametic models and estimation procedures for binormal ROC curves with multiple biomarkers. *The University of Michigan Department of Biostatistics Working Paper Series.* Http://biostats.bepress.com/umichbiostat/paper39.

[44] Gijbels I. and Mielniczuk J. (1990). Estimating the density of a copula function. *Communications in Statistics-Theory and Methods*, 19(2), 445–464.

[45] Gumbel E.J. (1960). Distributions des valeurs extremes en plusieurs dimensions. *Publications de l'Institute de Statistíque de l'Université de Paris*, 9, 171–173.

[46] Hand D.J., Daly F., Lunn A.D., McConway K.J. and Ostrowski E. (1994). *A Handbook of Small Data Sets.* London: Chapman & Hall.

[47] Hanley J.A. and McNeil B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3), 839–843.

[48] Hawkins D.M. (1980). *Identification of Outliers*, volume 11. London: Chapman & Hall.

[49] Hayfield T. and Racine J.S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 1–32. R package version 0.60-2 (https://cran.r-project.org/web/packages/np/np.pdf).

[50] Hill B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, pp. 677–691.

[51] Hill B.M. (1988). De Finetti's theorem, induction, and A_n, or Bayesian nonparametric predictive inference (with discussion). In *Bayesian Statistics 3*, (Editors) J.M. Bernando, M.H. DeGroot, D.V. Lindley and A. Smith, pp. 211–241. Oxford University Press.

[52] Hofmann T., Schölkopf B. and Smola A.J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, pp. 1171–1220.

[53] Huang X., Qin G. and Fang Y. (2011). Optimal combinations of diagnostic tests based on AUC. *Biometrics*, 67(2), 568–576.

[54] Jin H. and Lu Y. (2009). The optimal linear combination of multiple predictors under the generalized linear models. *Statistics & Probability Letters*, 79(22), 2321–2327.

[55] Joe H. (1997). *Multivariate Models and Multivariate Dependence Concepts*, volume 73. New Jersey: Chapman & Hall.

[56] Joe H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2), 401–419.

[57] Kang L., Liu A. and Tian L. (2013). Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Statistical Methods in Medical Research*. SAGE Publications.

[58] Kim G., Silvapulle M.J. and Silvapulle P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, 51(6), 2836 – 2850.

[59] Klugman S.A., Panjer H.H. and Willmot G.E. (2012). *Loss Models: From Data to Decisions*, volume 715. New Jersey: John Wiley & Sons.

[60] Koita A., Daucher D. and Fogli M. (2013). Multidimensional risk assessment for vehicle trajectories by using copulas. In *ICOSSAR*, p. 7. France. Https://hal.archives-ouvertes.fr/hal-00865839/document.

[61] Kojadinovic I. and Yan J. (2010). Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insurance: Mathematics and Economics*, 47(1), 52–63.

[62] Lawless J.F. and Fredette M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92, 529–542.

[63] Li Q. and Racine J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2), 266–292.

[64] Li Q. and Racine J.S. (2007). *Nonparametric Econometrics: Theory and Practice.* Princeton University Press.

[65] Li X., Mikusiński P., Sherwood H. and Taylor M.D. (1997). On Approximation of Copulas. In *Distributions with given Marginals and Moment Problems*, (Editors) V. Beneš and J. Štěpán, pp. 107–116. Dordrecht: Springer Netherlands.

[66] Lin J. and Wu X. (2015). Smooth tests of copula specifications. *Journal of Business & Economic Statistics*, 33(1), 128–143.

[67] Liu A., Schisterman E.F. and Zhu Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, 24(1), 37–47.

[68] Liu C., Liu A. and Halabi S. (2011). A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, 30(16), 2005–2014.

[69] Loader C.R. (1999).. Bandwidth selection: classical or plug-in? *Annals of Statistics*, pp. 415–438.

[70] Loftsgaarden D.O. and Quesenberry C.P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3), 1049–1051.

[71] Maturi T.A. (2010). *Nonparametric Predictive Inference for Multiple Comparisons.* Ph.D. thesis, Durham University, Durham, UK. Available from www.npi-statistics.com.

[72] Muhammad N., Coolen F.P.A. and Coolen-Maturi T. (2015). Predictive inference for bivariate data with nonparametric copula. *AIP Conference Proceedings.* To appear.

[73] Nelsen R.B. (2007). *An introduction to copulas.* New York: Springer Science & Business Media.

[74] Parzen E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3), 1065–1076.

[75] Pepe M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

[76] Pepe M.S., Cai T. and Longton G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1), 221–229.

[77] Pepe M.S. and Thompson M.L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2), 123–140.

[78] Powell M.J. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2), 155–162.

[79] Purcaru O. (2003). Semi-parametric Archimedean copula modelling in actuarial science. *Insurance, Mathematics and Economics*, 33, 419–420.

[80] Rank J. (2007). *Copulas: From Theory to Application in Finance.* London: Risk Books.

[81] Rudemo M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pp. 65–78.

[82] Scaillet O. and Fermanian J.D. (2002). Nonparametric estimation of copulas for time series. *FAME (Financial Asset Management and Engineering) Research Paper*, (57).

[83] Schepsmeier U., Stoeber J. and Brechmann E.C. (2013). *VineCopula: Statistical inference of vine copulas.* R package version 1.1-1 (http://www2.uaem.mx/r-mirror/web/packages/VineCopula/VineCopula.pdf).

[84] Schultz M., Eskin E., Zadok E. and Stolfo S. (2001). Data mining methods for detection of new malicious executables. In *Security and Privacy, 2001. S P 2001. Proceedings. 2001 IEEE Symposium on*, pp. 38–49.

[85] Scott D.W. (2009). *Multivariate Density Estimation: Theory, Practice, and Visualization*, volume 383. New York: John Wiley & Sons.

[86] Sen K.O.P.K. (2003). Copulas: Concepts and novel applications. *Metron*, 61(3), 323–353.

[87] Shih J.H. and Louis T.A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, pp. 1384–1399.

[88] Silverman B.W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26. London: CRC press.

[89] Sklar A.W. (1959). Fonctions de répartitionà n-dimension et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

[90] Staniswalis J., Messer K. and Finston D. (1991). Kernel estimators for multivariate smoothing. Technical report, Department of Statistics, Stanford University, Stanford California. Https://statistics.stanford.edu/sites/default/files/OLK

[91] Su J.Q. and Liu J.S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424), 1350–1355.

[92] Tang X.S., Li D.Q., Zhou C.B. and Zhang L.M. (2013). Bivariate distribution models using copulas for reliability analysis. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 227(5), 499–512.

[93] Trivedi P.K. and Zimmer D.M. (2005). Copula Modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1, 1–111.

[94] Tsukahara H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3), 357–375.

[95] Vexler A., Liu A. and Schisterman E.F. (2006). Efficient design and analysis of biospecimens with measurements subject to detection limit. *Biometrical Journal*, 48(5), 780–791.

[96] Wand M.P. and Jones M.C. (1994). *Kernel Smoothing*, volume 60. CRC Press.

[97] Wen K. and Wu X. (2015). Transformation-kernel estimation of the copula density. *Working Paper, Department of Agricultural Economics, Texas A&M University.* Http://agecon2.tamu.edu/people/faculty/wu-ximing/agecon2/public/copula.pdf.

[98] Wieand S., Gail M.H., James B.R. and James K.L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76(3), 585–592.

[99] Yan L., Tian L. and Liu S. (2015). Combining large number of weak biomarkers based on AUC. *Statistics in Medicine*, 34(29), 3811–3830.

[100] Zou K.H., Liu A., Bandos A.I., Ohno-Machado L. and Rockette H.E. (2011). *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis.* Boca Raton: CRC Press.