Statistics and Probability Letters 126 (2017) 173–178

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Bounds on Kendall's tau for zero-inflated continuous variables

Michel M. Denuit [a], Mhamed Mesfioui [b,*]

[a] *Institut de statistique, biostatistique et sciences actuarielles (ISBA), Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium*

[b] *Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières, Trois-Rivières (Québec), Canada G9A 5H7*

ARTICLE INFO

Article history:
Received 5 October 2016
Received in revised form 25 January 2017
Accepted 3 March 2017
Available online 22 March 2017

Keywords:
Association measure
Bivariate zero-inflated data
Fréchet–Hoeffding bounds

ABSTRACT

In this short note, we derive the lower and upper bounds on the association measure for zero-inflated continuous random variables proposed by Pimentel et al. (2015). These bounds only involve the respective probability masses at the origin. This provides analysts with the set of values that can be attained, helping them to interpret the obtained results as shown in an example based on Danish fire insurance losses data.

1. Introduction

Zero-inflated (or zero-adjusted) distributions combine a continuous distribution on the positive real line and a point mass at zero. There are numerous examples of such distributions in different areas. For instance, in insurance studies, the total claim amount related to a given contract is often equal to 0, when no claims have been filed against the insurer, but may also by strictly positive when one or several accidents occurred. The total claim cost is thus decomposed into the product of an indicator for the event "the policy produces at least one claim during the reference period" and a positive random variable representing the total claim amount produced by the policy when at least one claim has been filed. See e.g. Section 5 for an example with fire insurance losses, as well as Klein et al. (2014) and the references therein.

When the focus is on the abundance of a species or resources, data that are positively skewed and contain a substantial proportion of zeros are often encountered in ecological applications. The probability mass at zero indicates whether or not the species or resource is present whereas a continuous probability density function describes the distribution of the abundance when it is present. See e.g. Fletcher et al. (2005) or Ancelet et al. (2010). Climatic data, such as rainfall data for instance, also exhibit many zero values. See e.g. Bruno et al. (2014).

Bivariate zero-inflated data are now under study in various areas. Coming back to insurance applications for example, multivariate models combining different insurance products subscribed by policyholders are now investigated by actuaries. The joint examination of the losses produced by policyholders on several covers accounts for the correlation effects, improving on marginal analyses. See for instance Shi (2014) and Thuring et al. (2012). Similarly, species abundance or climatic measures at two locations or at two points in time provide examples of bivariate zero-inflated data that are of interest in different applications.

* Corresponding author.
E-mail address: mhamed.mesfioui@uqtr.ca (M. Mesfioui).

http://dx.doi.org/10.1016/j.spl.2017.03.005
0167-7152/

Proper assessment of the strength of dependence existing in a pair of zero-inflated random variables is not an easy task, due to the possibly substantial probability mass at the origin. Pimentel et al. (2015) proposed an estimator of Kendall's tau for bivariate zero-inflated data. This estimator possesses a natural interpretation and improved statistical properties. As pointed out by these authors, their association measure cannot attain the values -1 and $+1$. In this paper, we derive the lower and upper bounds on this newly proposed association measure. As expected, these bounds only depend on the respective probability masses at zero, and not on the continuous conditional distributions on $\mathbb{R}^+ = (0, \infty)$.

2. Zero-inflated distributions

Let us adopt the same notation as in Pimentel et al. (2015). In this paper, we consider non-negative random variables X having a continuous distribution on $\mathbb{R}^+$ with probability density function h and a positive probability mass at 0, $\mathrm{P}[X = 0] = p_1 > 0$. The distribution function of X is given by

$$F(s) = \begin{cases} 0 & \text{if } s < 0 \\ p_1 & \text{if } s = 0 \\ F_h(s) = p_1 + (1 - p_1) \displaystyle\int_0^s h(x)dx & \text{if } s > 0. \end{cases} \tag{2.1}$$

Recall that the generalized inverse of F is defined by $F^{-1}(\alpha) = \inf\{x \geq 0 | F(x) \geq \alpha\}, \alpha \in [0, 1]$. In our setting, it can be explicitly expressed as

$$F^{-1}(\alpha) = \begin{cases} 0 & \text{if } 0 \leq \alpha \leq p \\ F_h^{-1}(\alpha) & \text{if } p < \alpha \leq 1 \end{cases} \tag{2.2}$$

where F_h^{-1} denotes the inverse of the function F_h.

3. Kendall's tau for bivariate zero-inflated data

Let (X, Y) be a pair of random variables with joint distribution function H, i.e. $H(x, y) = \mathrm{P}[X \leq x, Y \leq y]$. Suppose further that X and Y are zero-inflated continuous and non-negative with respective distribution functions F and G, and positive probability masses at 0. Specifically, the distribution function F of X is given in (2.1) and G is defined similarly, with $\mathrm{P}[Y = 0] = G(0) = p_2 > 0$ and

$$G_h(s) = p_2 + (1 - p_2) \int_0^s k(x)dx \quad \text{for } s > 0,$$

where k is the probability density function of Y given $Y > 0$.

Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed random pairs, each with joint distribution function H. The population version of Kendall's tau corresponding to H, denoted as τ_H, is then defined as the probability of concordance minus the probability of discordance, i.e. $\tau_H = \mathrm{P}\big[(X_1 - X_2)(Y_1 - Y_2) > 0\big] - \mathrm{P}\big[(X_1 - X_2)(Y_1 - Y_2) < 0\big]$. Denuit and Lambert (2005) and Mesfioui and Tajar (2005) studied general versions of the population Kendall's tau. In the context of zero-inflated continuous and non-negative random pairs (X, Y), Pimentel et al. (2015) have established an interesting and useful formula of the population Kendall's tau that is recalled next. Let us denote as X_{10} a random variable distributed as X given that $Y = 0$, as X_{11} a random variable distributed as X given that $Y > 0$, as Y_{01} a random variable distributed as Y given that $X = 0$, as Y_{11} a random variable distributed as Y given that $X > 0$, as X_1 a random variable distributed as X given $X > 0$, and as Y_1 a random variable distributed as Y given $Y > 0$. Now, define $p_{00} = \mathrm{P}[X = 0, Y = 0]$, $p_{10} = \mathrm{P}[X > 0, Y = 0]$, $p_{01} = \mathrm{P}[X = 0, Y > 0]$, $p_{11} = \mathrm{P}[X > 0, Y > 0]$, $p_1^* = \mathrm{P}[X_{10} > X_{11}]$, $p_2^* = \mathrm{P}[Y_{01} > Y_{11}]$, and denote as τ_{11} Kendall's tau of (X_1, Y_1). The association measure studied in Pimentel et al. (2015) is then given by

$$\tau_H = p_{11}^2 \tau_{11} + 2\left(p_{00}p_{11} - p_{01}p_{10}\right) + 2p_{11}\Big(p_{10}(1 - 2p_1^*) + p_{01}(1 - 2p_2^*)\Big). \tag{3.1}$$

An estimator for (3.1) has also been proposed in Pimentel et al. (2015), replacing all probabilities involved in (3.1) by the corresponding relative frequencies and computing $\widehat{\tau}_{11}$ on data points with both components positive.

4. Bounds on the association measure (3.1)

As pointed out by Pimentel et al. (2015), the association measure (3.1) cannot reach the values ± 1 when $p_1 > 0$ or $p_2 > 0$. Our aim is to derive the range of admissible values for the association measure (3.1) in terms of the marginal distributions of X and Y. This can be done by using the property of monotonicity of Kendall's tau with respect to the concordance order; see e.g. Mesfioui and Tajar (2005) and Denuit and Lambert (2005). This means that if H_1 and H_2 are two bivariate distribution functions such that $H_1(x, y) \leq H_2(x, y)$ for all x and y, then $\tau_{H_1} \leq \tau_{H_2}$. Hence, the sharp lower and upper bounds of Kendall's tau can be derived using Fréchet–Hoeffding bounds. In other words, the desired lower bound $\tau_{\min}$ and upper bound $\tau_{\max}$ on (3.1) can be computed by using the fact that $\tau_{\min} = \tau_W$ and $\tau_{\max} = \tau_M$ where W and M denote the Fréchet–Hoeffding bounds defined by $W(x, y) = \max\{F(x) + G(y) - 1, 0\}$ and $M(x, y) = \min\{F(x), G(y)\}$. The next result establishes the expression of these bounds in terms of the probabilities p_1 and p_2.

Proposition 4.1. *The lower and upper bounds on the association measure (3.1) are given by*

$$\tau_{\max} = 1 - \max\{p_1^2, p_2^2\} \quad \text{and} \quad \tau_{\min} = (1-p_1-p_2)_+^2 - 2(1-p_1)(1-p_2)$$

where $x_+ = \max\{x, 0\}$.

Proof. We know from Proposition 2.3 in Mesfioui and Tajar (2005) that the upper bound $\tau_{\max}$ on (3.1) is obtained when the random pair (X, Y) obeys the upper Fréchet–Hoeffding bound, that is, when $X = F^{-1}(U)$ and $Y = G^{-1}(U)$, where U is a random variable uniformly distributed over the unit interval $[0, 1]$. Clearly, (2.2) shows that

$$X = F^{-1}(U) = F_h^{-1}(U)\mathrm{I}[p_1 < U \le 1] \quad \text{and} \quad Y = G^{-1}(U) = G_k^{-1}(U)\mathrm{I}[p_2 < U \le 1],$$

where $\mathrm{I}[A]$ is the indicator of the event A, equal to 1 when A is realized and to 0 otherwise. Suppose without loss of generality that $p_1 \le p_2$ holds. Then,

$$\begin{aligned}
p_{00} &= \mathrm{P}[X = 0, Y = 0] = \mathrm{P}[U \le p_1] = p_1\\
p_{10} &= \mathrm{P}[X > 0, Y = 0] = \mathrm{P}[p_1 < U \le p_2] = p_2 - p_1\\
p_{01} &= \mathrm{P}[X = 0, Y > 0] = 0\\
p_{11} &= \mathrm{P}[X > 0, Y > 0] = \mathrm{P}[U > p_2] = 1 - p_2.
\end{aligned}$$

It remains to calculate p_1^*. To this end, we first derive the distribution function of the random variable X_{11}: X_{11} has no probability mass at zero as $\mathrm{P}[X_{11} = 0] = \mathrm{P}[X = 0|Y > 0] = 0$ and

$$\begin{aligned}
F_{X_{11}}(x) &= \mathrm{P}[X \le x|Y > 0] = \frac{\mathrm{P}[X \le x, Y > 0]}{1-p_2}\\
\mathrm{P}[X \le x, Y > 0] &= \mathrm{P}[X \le x, U > p_2] = \mathrm{P}[F_h^{-1}(U) \le x, U > p_2] = \mathrm{P}[p_2 < U \le F_h(x)]
\end{aligned}$$

so that

$$F_{X_{11}}(x) = \begin{cases} 0 & \text{if } x \le F_h^{-1}(p_2)\\ \dfrac{F_h(x) - p_2}{1-p_2} & \text{if } x > F_h^{-1}(p_2). \end{cases}$$

Therefore, the quantile function of X_{11} is

$$F_{X_{11}}^{-1}(\alpha) = F_h^{-1}\big((1-p_2)\alpha + p_2\big), \quad \alpha \in [0, 1]. \tag{4.1}$$

Similar arguments imply that the distribution function and the quantile function of X_{10} are respectively given by

$$F_{X_{10}}(x) = \begin{cases} 0 & \text{if } x < 0\\ \dfrac{F_h(x)}{p_2} & \text{if } 0 \le x \le F_h^{-1}(p_2)\\ 1 & \text{if } x > F_h^{-1}(p_2) \end{cases}$$

and

$$F_{X_{10}}^{-1}(\alpha) = \begin{cases} 0 & \text{if } \alpha \in \left[0, \dfrac{p_1}{p_2}\right]\\ F_h^{-1}(\alpha p_2) & \text{if } \alpha \in \left[\dfrac{p_1}{p_2}, 1\right]. \end{cases} \tag{4.2}$$

We see from (4.1) and (4.2) that $F_{X_{10}}^{-1}(\alpha) \le F_{X_{11}}^{-1}(\alpha)$ for all $\alpha \in [0, 1]$, so that $p_1^* = \mathrm{P}\left[F_{X_{10}}^{-1}(U) > F_{X_{11}}^{-1}(U)\right] = 0$. Moreover, $\tau_{11} = 1$ when the random variables X and Y are perfectly positively dependent. We then get from (3.1) that

$$\begin{aligned}
\tau_{\max} = p_{11}^2 + 2p_{00}p_{11} + 2p_{11}p_{10} &= (1-p_2)^2 + 2p_1(1-p_2) + 2(1-p_2)(p_2-p_1)\\
&= 1 - p_2^2,
\end{aligned}$$

as announced.

Let us now establish the expression for the lower bound $\tau_{\min}$. To that end, we first calculate the probabilities p_{00}, p_{10}, p_{01} and p_{11} when X and Y obey the lower Fréchet–Hoeffding bound, that is, $X = F^{-1}(U)$ and $Y = G^{-1}(1-U)$ where U is uniformly distributed over the unit interval [0, 1]. We distinguish two cases: Either $1-p_1-p_2<0 \Leftrightarrow p_1+p_2>1$ so that

$$\begin{aligned}
p_{00} &= \mathrm{P}[X=0, Y=0] = \mathrm{P}[U\le p_1, 1-U\le p_2]\\
&\qquad\qquad\qquad\quad = \mathrm{P}[1-p_2\le U\le p_1] = p_1+p_2-1\\
p_{10} &= \mathrm{P}[X>0, Y=0] = \mathrm{P}[U>p_1, 1-U\le p_2]\\
&\qquad\qquad\qquad\quad = \mathrm{P}[U>p_1] = 1-p_1\\
p_{01} &= \mathrm{P}[X=0, Y>0] = \mathrm{P}[U\le p_1, 1-U>p_2]\\
&\qquad\qquad\qquad\quad = \mathrm{P}[U\le 1-p_2] = 1-p_2\\
p_{11} &= \mathrm{P}[X>0, Y>0] = \mathrm{P}[U>p_1, 1-U>p_2] = 0
\end{aligned}$$

and it follows from (3.1) that

$$\tau_{\min} = -2p_{01}p_{10} = -2(1-p_1)(1-p_2). \tag{4.3}$$

Or $1-p_1-p_2\ge 0$ so that

$$\begin{aligned}
p_{00} &= \mathrm{P}[X=0, Y=0] = 0\\
p_{10} &= \mathrm{P}[X>0, Y=0] = p_2\\
p_{01} &= \mathrm{P}[X=0, Y>0] = p_1\\
p_{11} &= \mathrm{P}[X>0, Y>0] = 1-p_1-p_2.
\end{aligned}$$

It remains to calculate p_1^* and p_2^*. To do so, let us derive the distribution of X_{10} as follows:

$$F_{X_{10}}(x) = \mathrm{P}[X\le x|Y=0] = \frac{\mathrm{P}[X\le x, Y=0]}{p_2}$$

$$\begin{aligned}
\mathrm{P}[X\le x, Y=0] &= \mathrm{P}[X\le x, 1-U\le p_2]\\
&= \mathrm{P}[X\le x, U\ge 1-p_2]\\
&= \mathrm{P}[F_h^{-1}(U)\le x, U\ge 1-p_2]\\
&= \mathrm{P}[U\le F_h(x), U\ge 1-p_2]
\end{aligned}$$

so that

$$F_{X_{10}}(x) = \begin{cases} 0 & \text{if } x\le F_h^{-1}(1-p_2)\\ \dfrac{F_h(x)+p_2-1}{p_2} & \text{if } x> F_h^{-1}(1-p_2)\end{cases}$$

and

$$F_{X_{10}}^{-1}(\alpha) = F_h^{-1}(p_2\alpha+1-p_2), \quad \alpha\in[0,1]. \tag{4.4}$$

Likewise, the distribution of X_{11} is given by

$$F_{X_{11}}(x) = \mathrm{P}[X\le x|Y>0] = \frac{\mathrm{P}[X\le x, Y>0]}{1-p_2}$$

with

$$\begin{aligned}
\mathrm{P}[X\le x, Y>0] &= \mathrm{P}[X\le x, 1-U>p_2]\\
&= \mathrm{P}[X\le x, U<1-p_2]\\
&= \mathrm{P}[U\le p_1] + \mathrm{P}[U\le F_h(x), p_1<U<1-p_2]
\end{aligned}$$

so that

$$F_{X_{11}}(x) = \begin{cases} 0 & \text{if } x<0\\ \dfrac{F_h(x)}{1-p_2} & \text{if } 0\le x\le F_h^{-1}(1-p_2)\\ 1 & \text{if } x\ge F_h^{-1}(1-p_2)\end{cases}$$

and

$$F_{X_{11}}^{-1}(\alpha) = \begin{cases} 0 & \text{if } \alpha\in\left[0, \dfrac{p_1}{1-p_2}\right]\\ F_h^{-1}(\alpha(1-p_2)) & \text{if } \alpha\in\left[\dfrac{p_1}{1-p_2}, 1\right].\end{cases} \tag{4.5}$$

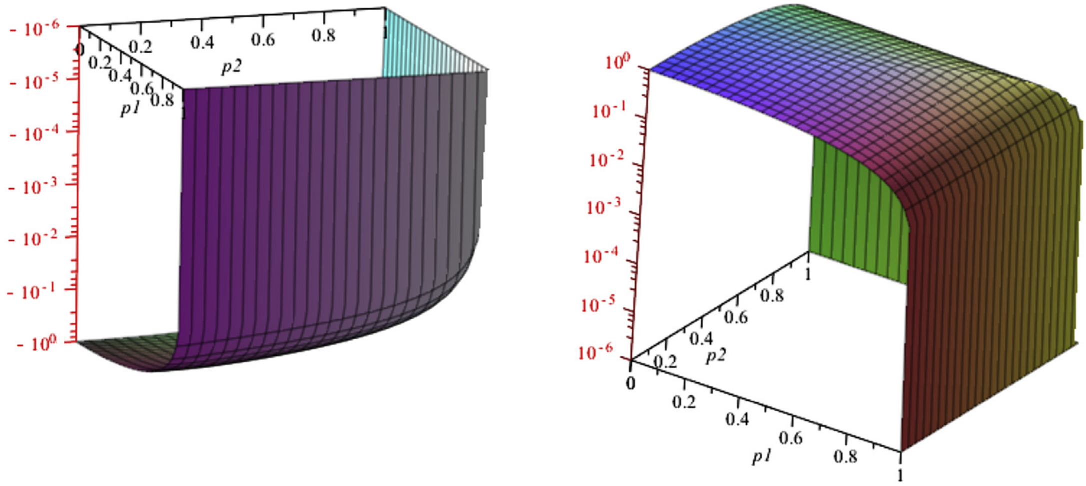

Fig. 4.1. Lower bound (left panel) and upper bound (right panel) on (3.1) in terms of $(p_1, p_2) \in [0, 1]^2$.

From (4.4) and (4.5), we see that $F^{-1}_{X_{11}}(\alpha) \leq F^{-1}_{X_{10}}(\alpha)$ for all $\alpha \in [0, 1]$. Thus, $p^*_1 = \mathrm{P}\left[F^{-1}_{X_{10}}(U) > F^{-1}_{X_{11}}(U)\right] = 1$. Similar arguments lead to $p^*_2 = 1$. Also, $\tau_{11} = -1$ when X and Y are perfectly negatively dependent. It then follows from (3.1) that

$$\begin{aligned}\tau_{\min} &= -p^2_{11} - 2p_{01}p_{10} - 2p_{11}(p_{10} + p_{01}) \\ &= -(1 - p_1 - p_2)^2 - 2p_1p_2 - 2(1 - p_1 - p_2)(p_1 + p_2) \\ &= p^2_1 + p^2_2 - 1 \\ &= (1 - p_1 - p_2)^2 - 2(1 - p_1)(1 - p_2). \end{aligned} \quad (4.6)$$

Finally, from (4.3) and (4.6), one deduces that

$$\tau_{\min} = \begin{cases} -2(1 - p_1)(1 - p_2) & \text{if } 1 - p_1 - p_2 < 0 \\ (1 - p_1 - p_2)^2 - 2(1 - p_1)(1 - p_2) & \text{if } 1 - p_1 - p_2 \geq 0 \end{cases}$$

which can be reduced to $\tau_{\min} = (1 - p_1 - p_2)^2_+ - 2(1 - p_1)(1 - p_2)$, and ends the proof. □

Fig. 4.1 displays the lower and upper bounds on (3.1) in terms of the probabilities p_1 and p_2. We see there that the lower bound increases in terms of p_1 and p_2. It reaches its minimum value -1 at $(p_1, p_2) = (0, 0)$ and its maximum value 0 when $p_1 = 1$ or $p_2 = 1$. The upper bound decreases in terms p_1 and p_2. It achieves its maximum value 1 at $(p_1, p_2) = (0, 0)$ and its minimum value 0 when $p_1 = 1$ or $p_2 = 1$. In addition, Fig. 4.2 displays the corresponding contour plots. We can clearly see there the different behavior of the lower and upper bounds as functions of p_1 and p_2.

5. Numerical illustration

Considering Proposition 4.1, the measure of association (3.1) can be compared to the lower and upper bounds $\tau_{\min}$ and $\tau_{\max}$ so that the analyst can properly interpret the values obtained from a given data sample. The following example based on Danish fire losses data available from the R package fitdistrplus illustrates what can be learned in this way.

These data were collected at Copenhagen Reinsurance and comprise 2167 fire losses over the period 1980 to 1990, adjusted for inflation. Here, we consider the pair with loss of contents and loss of profits (we do not include the building loss because there are only few zero values in this case). The estimated Kendall tau proposed by Pimentel et al. (2015) is equal to 0.262483 whereas the lower and upper bounds obtained in this paper are equal to -0.437009 and 0.487722. The estimated dependence measure should therefore not be compared to 1 but only to about 0.5.

Acknowledgments

Michel Denuit acknowledges the financial support from the contract “Projet d’Actions de Recherche Concertées” No. 12/17-045 of the “Communauté française de Belgique”, granted by the “Académie universitaire Louvain”. Mhamed Mesfioui acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada No. 261968-2013. Both authors thank two anonymous Referees and the Editor for their careful reading and for their constructive comments which helped them to improve the text.

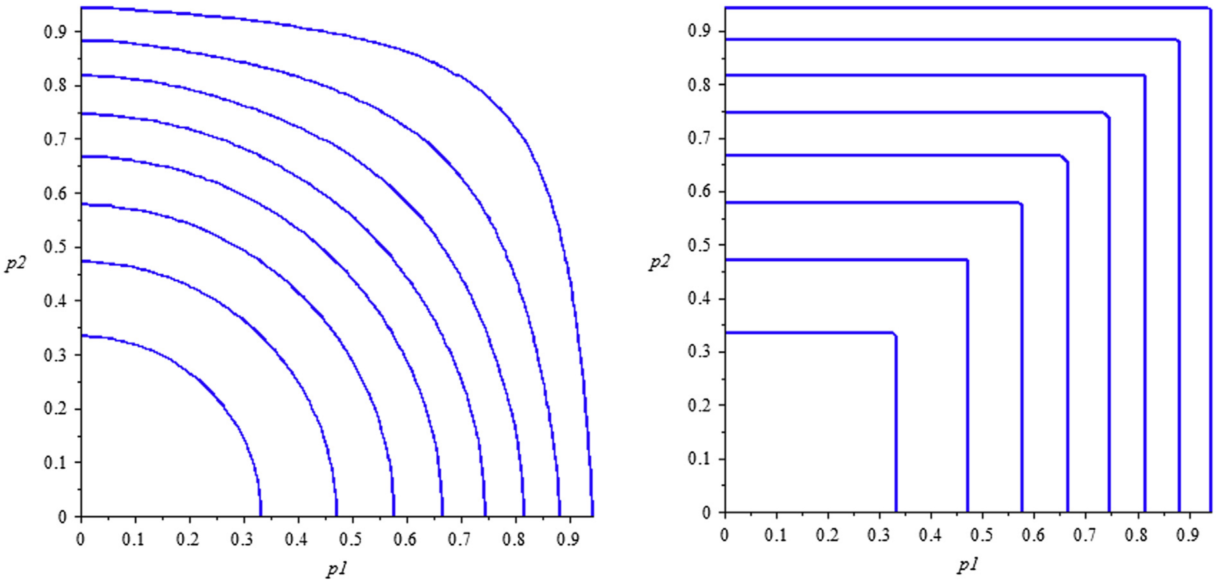

Fig. 4.2. Contour plots of the lower bound (left panel) and the upper bound (right panel) in terms of $(p_1, p_2) \in [0, 1]^2$.

References

Ancelet, S., Etienne, M.P., Benoît, H., Parent, E., 2010. Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process. Environ. Ecol. Stat. 17, 347–376.

Bruno, F., Cocchi, D., Greco, F., Scardovi, E., 2014. Spatial reconstruction of rainfall fields from rain gauge and radar data. Stoch. Environ. Res. Risk Assess. 28, 1235–1245.

Denuit, M., Lambert, P., 2005. Constraints on concordance measures in bivariate discrete data. J. Multivariate Anal. 93, 40–57.

Fletcher, D., MacKenzie, D., Villouta, E., 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. Environ. Ecol. Stat. 12, 45–54.

Klein, N., Denuit, M., Lang, S., Kneib, Th., 2014. Nonlife ratemaking and risk management with Bayesian additive models for location, scale and shape. Insurance Math. Econom. 55, 225–249.

Mesfioui, M., Tajar, A., 2005. On the properties of some nonparametric concordance measures in the discrete case. Nonparametr. Stat. 17, 541–554.

Pimentel, R.S., Niewiadomska-Bugaj, M., Wang, J.C., 2015. Association of zero-inflated continuous variables. Statist. Probab. Lett. 96, 61–67.

Shi, P., 2014. Insurance ratemaking using a copula-based multivariate Tweedie model. Scand. Actuar. J. 2014, 1–18.

Thuring, F., Nielsen, J.P., Guillen, M., Bolance, C., 2012. Selecting prospects for cross-selling financial products using multivariate credibility. Expert Syst. Appl. 39, 8809–8816.