مجلة العلوم البحثة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.ly ISSN 2521-9200
Received 05/01/2018 Revised 02/02/2018 Published online 28/03/2018

Probability Paper and Plotting Position of Extreme Value distribution For distribution selection and parameter estimation

*H. A. Alaswed & Alsaidi M. Altaher
Statistics Department, Faculty of Science, Sebha University-Libya
*Corresponding author: haf.alaswed@sebhau.edu.ly

Abstract The problem of estimation for tail index in extreme value distributions is very important in many applications. Statistical methods could be used to select the distribution based on the available data; where the initially specified distributions or family of distributions usually depend on unknown parameters and these parameters need to be estimated. This paper shows how probability papers plots (PPP) can be used to select the most appropriate distribution among the three types of maximum extreme value distribution (Gumbel, Weibull, Fréchet). Another objective is to use the PPP for parameter estimation using regression method. The last objective is to compare between PPP, MLE and PWM methods to estimate the parameters of the selected distribution using two criteria, the mean square error and correlation coefficient. Results of using daily maximum of temperature in weather data show that the Weibull distribution is the best distribution for each period of this data. Another result finding is that the PPP provide more efficient estimate for the shape parameter of Weibull distribution than MLE and PWM methods dependent on mean square error.

Keywords: Extreme Value distribution(EVD), Plotting Position(PP); Probability Paper Plot (PPP). Mean squared error(MSE).

رسمة ورقة الاحتمال لتوزيعات القيم المتطرفة لتحديد النموذج وتقدير المعالم

*حافظ ابوبكر امحمد و السعيدي المهدي الطاهر
قسم الإحصاء – كلية العلوم – جامعة سبها، ليبيا
haf.alaswed@sebhau.edu.ly:*للمراسلة

الملخص إن مشكلة تقدير معلمة الذيل لتوزيعات القيم المتطرفة له أهمية كبيرة في العديد من التطبيقات وأن الطرق الإحصائية يمكن استخدامها في اختيار النموذج المناسب للبيانات المتاحة حيث تعتمد هذه النماذج علي المعالم المجهولة التي يتم تقديرها بناء علي البيانات المتاحة باستخدام طرق التقدير. هذه الورقة تبين كيفية استخدام طريقة رسمة ورقة الاحتمال لاختيار التوزيع المناسب من بين توزيعات القيم المتطرفة الثلاثة (جامبل ، فريشت وويبل). بعد تحديد التوزيع المناسب تم استخدام طريقة رسمة ورقة الاحتمال في تقدير معالم النموذج المختار ومقارنتها مع طريقة الإمكان الأعظم وطريقة العزوم الموزونة وتحديد الأفضلية عن طريق متوسط مربع الخطأ. طبقت الطريقة على بيانات حقيقية وكان النموذج الافضل هو توزيع ويبل، كما ان طريقة رسمة ورقة الاحتمال كانت أكثر كفاءة في تقدير معالمه مقارنة بطريقة الامكان الاعظم وطريقة العزوم الاحتمالية الموزونة.

الكلمات المفتاحية: توزيعات القيم المتطرفة العظمي ، رسمة الموقع، رسمة ورقة الاحتمال، متوسط مربع الأخطاء.

1 Introduction

Extreme value theory (EVT) has emerged as one of the most important statistical disciplines for the applied sciences and widely used in many other disciplines over the last 50 years. In [1], EVT refers to branch of statistic that deals with extreme events. The asymptotic theory of sample extremes has been developed in parallel with the central limit theory (CLT), and in fact the two theories bear some resemblance. The CLT is concerned with the limit behaviour of the sums or average whereas the theory of sample extremes is concerned with the limit behavior of the sample extremes $\max(X_1, X_2, ..., X_n)$ or $\min(X_1, X_2, ..., X_n)$, for more details see,[2]. The main limiting results in EVT date back to the paper of [3]. The class of extreme value distributions (EVD) essentially involves three types of EVD that are needed to distribution the maximum of the collection of random observations from the same distribution as defined in section 2. This paper is organized as follows: In Section 2, we discuss the three types of maximum EVD. In Section 3, we describe the Plotting Positions (PP). Probability Paper Plot (PPP) for Gumbel, Fréchet and Weibull are presented in Section 4. Parameter estimation and distribution selection are presented in Section 5. Case study and application on real data is presented to evaluate the performance of the PPP in Section 6. Concluding remarks are given in Section 7.

2 Three Types of Maximum Extreme Value Distributions (MEVD)

EVD are usually considered to comprise the following three type families:

$$Type-I\ (Gumbel_M, \gamma \cong 0) = \exp[-e^{(x-\mu)/\sigma}] \quad -\infty < x < \infty \quad 1$$

$$Type-II\ (Frechet_M, \gamma>0)=\begin{cases}0, & x \prec \mu \\ \exp[-(\frac{x-\mu}{\sigma})^{-\gamma}], & x \geq \mu\end{cases} \quad 2$$

$$Type-III\ (Weibull_M, \gamma<0)=\begin{cases}\exp[-(\frac{\mu-x}{\sigma})^{\gamma}], & x \leq \mu \\ 0, & x \succ \mu\end{cases} \quad 3$$

In [1], these three types of distribution are termed the extreme value distributions, with types I, II and III widely known as the Gumbel, Fréchet and Weibull families respectively. Each distribution has a location $\mu \succ 0$ and scale parameter $\sigma \succ 0$; additionally, the Fréchet and Weibull distribution have a shape parameter γ. The previous three types are obviously with: $\gamma = 0$; $\gamma = \frac{1}{\alpha} > 0$ and $\gamma = -\frac{1}{\alpha} < 0$ respectively. In view of this, and the fact that distributions (2) and (3) can be transformed to type I distribution by the transformations $Z = \log(X-\mu)$ and $Z = -\log(\mu - X)$ respectively. Gumbel (Type I) distributions are also sometimes called, in earlier writings, doubly exponential distributions. The proof of the three types of maximum extreme value distributions (MEVD) can be found in [4]. [5], it was proposed a GEVD that includes the three types distributions Gumbel, Fréchet and Weibull can be combined into a single family of distribution having distribution functions of the form:

$$G_\gamma(x)=\begin{cases}\exp\left[-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right] & if\ \gamma \neq 0 \\ \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right] & if\ \gamma = 0\end{cases} \quad 4$$

Where $-\infty \prec \mu \prec \infty, \sigma \succ 0, -\infty \prec \gamma \prec \infty$

For $\gamma = 0\ (\gamma \to 0), \gamma \succ 0$ and $\gamma \prec 0$ we get Gumbel, Fréchet and Weibull respectively. The shape parameter γ, called the tail index (TI). The three limit types have different forms of tail index as shown in **Fig-1**

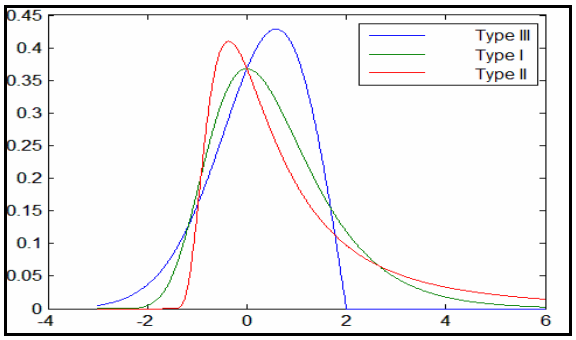

Fig-1: Tail index of three type of MEVD

3 Plotting Positions (PP)

Let $X_1, X_2, ..., X_n$ are independently and identically distributed (iid) random variables having a common parametric family of $f(x;\theta)$ and with unknown distribution function $F(x;\theta)$. Let $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ be the observed order statistics in a random sample drawn from $F(x;\theta)$. Define the so-called plotting positions:

$$P_{i:n} = \frac{i-\alpha}{n+\beta}, \quad i = 1,2,...,n \quad 5$$

For appropriate choices of $\alpha \geq 0$ and $\beta \geq 0$. In [6] and [7], plotting positions (α and , β values) can be chosen empirically (depending on the data, the type of distribution, the estimation method to be used, etc.). Here we use $\alpha = 0$ and $\beta = 1$; that is

$$P_1 = \frac{i}{n+1}, \quad i = 1,2,...,n \quad 6$$

Other alternative plotting positions include:

$$P_2 = \frac{i-0.375}{n+0.25}, \quad 7$$

$$P_3 = \frac{i-0.5}{n}, \quad 8$$

For justifications of these formulas see, for example,[8]. Other references for plotting positions include, [9] and [10].

4 Probability Paper Plots (PPP)

The Probability Paper Plot (PPP) is an important tool among the graphical data exploration techniques which can be used to distinguish visually between different tail index of distribution functions. We first need to introduce the basic idea of PPP of a two-parameter family of distributions consists of changing the random variable X to $\mu = h(X)$ and the probability P to $v = g(P)$ in such a manner that the cdf become a family of straight lines. In this way, when the cdf is drawn, a linear trend is an indication of the sample coming from the corresponding family. If the deviation from the straight line is too strong we conclude that the sample comes from a different distribution. For example, Let $X_1, X_2, ..., X_n$ be a sample drawn from $F(x;a,b)$ where a and b are the parameters, and let $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ be the corresponding order statistics and $P_{1:n}, P_{2:n}, ..., P_{n:;n}$ be plotting positions such as those given in (6-9). To obtain the probability plot, we look for a transformation that expresses the equation:

$$P = F(x;a,b) \quad 9$$

in the form of a straight line

$$g(P) = g(F(x;a,b)) = ah(x) + b \quad 10$$

or, equivalently,

$$v = au + b \quad 11$$

where $v = g(p)$ and $u = h(x)$ are called the PPP, where the ordinate and abscissa axes are

transformed by $v_i = g(p_{i:n})$ and $u_i = h(x_{i:n})$ as defined in equation (9 or 11). The scatter of points in the PPP would exhibit a straight line trend with positive slope. However, due to the random character of samples, even in the case of the sample drawn from the given family, one should not expect that the corresponding graph will be an exact straight line. Thus, if the trend is approximately linear, it can be used as an evidence that the sample did come from the assumed family. We first need to introduce the Probability Paper Plot of Gumbel distribution.

4.1 Gumbel Probability Paper Plot (GPPP)

One of the most commonly used graphical methods in statistics is the PPP. The basic idea of PPP of a two-parameter family of distributions consists of changing the random variable X to $u_i = h(x_{i:n})$ and the probability P to $v_i = g(p_{i:n})$ as defined in (10) in such a manner that the cdf become a family of straight lines. Now, we derive the maximal Gumbel probability paper plots(GPPP), see,[11] and [12]. When the cdf of maximal Gumbel defined in (1) is given by:

$$F(x;\mu,\sigma) = \exp[-e^{(x-\mu)/\sigma}] \qquad 12$$

Let $P = F(x;\mu,\sigma)$. Taking logarithms twice we get:

$$-\log[-\log(P)] = \frac{1}{\sigma}x - \frac{\mu}{\sigma}. \qquad 13$$

In the form of straight line comparing with (12), we get $v = g(p) = -\log[-\log(P)]$, $u = h(x) = x$, $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$. which shows that the transformation (14) transforms (13) to the family of straight lines

$$v = \frac{1}{\sigma}x - \frac{\mu}{\sigma} \qquad 14$$

Thus, in a maximal GPPP, the ordinate axis need to be transformed by $v = -\log[-\log(P)]$, whereas the abscissa axis need not be transformed. Estimation of the two parameters μ and $\mu+\sigma$ can be done by setting $v = 0$ and $v = 1$, and obtaining

$$v = 0 \Rightarrow x = \mu$$
$$v = 1 \Rightarrow x = \mu + \sigma \qquad 15$$

Once we have fitted a straight line to the data, the abscissas associated with the reduced variable, v, 0 *and* 1 are the values μ and $\mu+\sigma$, respectively. The scatter plot of v_i, versus u_i $i = 1,2,...,n$, where $v_i = g(p) = -\log[-\log(P_i)]$ and $u_i = h(x_i) = x_i$ are called the PPP. In this way, when the trend as approximately linear is an indication of the sample coming from the corresponding family (GPPP). In the following Table-1 we summarize The PPP for the maximum of Fréchet and Weibull derived in a similar manner.

Table -1: PPPs transformations of MEVD.

Distribution	Random Variable u=x	Probability Scale V=y
Weibull	$-\log(\lambda - x)$	$-\log(-\log p)$
Gumbel	x	$-\log(-\log p)$
Fréchet	$\log(x - \lambda)$	$-\log(-\log p)$

5 Parameter estimation

The parameters of a MEVD can be estimated with various methods. The maximum likelihood (ML) method, see,[2] and probability weighted moments (PWM) method see, [13],[14] and PPP can be used regression method, for more details see,[8] and the results of parameters estimation of three type of extreme distribution will be summarized in Table-2

Table 2: Estimation parameters of PPP by using regression methods

Distribution	Straight lines	Location
Weibull	$v = \beta\mu + \beta\log\sigma$	$\lambda > x_{n:n}$
Gumbel	$v = \frac{1}{\sigma}\mu - \frac{\lambda}{\sigma}$	$\lambda = 0$
Frechet	$v = \beta\mu - \beta\log\sigma$	$\lambda < x_{1:n}$

Where v *and* μ as defined in Table-1. The selection of a distribution to explain extreme events can be made by applying various goodness-of-fit tests. In this study the Formulas of mean square error (MSE) by [15], and correlation coefficient can be calculated as below:

$$MSE = \sum_{i=1}^{n} (\hat{F}(x_i) - F(P_i))^2 \qquad 16$$

Where $\hat{F}(x_i)$ is distribution and $\hat{F}(P_i)$ is empirical dependent on P_i and correlation coefficient (r) defined as follows:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}} \qquad 17$$

6 Case Study

In this section the PPP is used on real dataset application of the daily maximum of temperature in weather data from Fort Collins, Colorado, U.S.A. from 1900 to 1999. The data can be obtained from the NOAA/NCDC website. Data frame has 36524 observations. The sample of size n is divided into three different block sizes (monthly, quarterly, and yearly); because, we need to know the different selection periods when data are fitted have the same distribution or different as time increases, See [16] for more information and analyses. Tables from 3:5 summarize some estimation results of the intercept and slope parameter obtained via the PPP with three plot position formula. The first two columns gives the intercept and slope. The next two columns give the correlation coefficient and MSE while, the last two columns determine the best model in which high correlation and minimum MSE are gained.

Table 3 Parameter estimation for monthly data.

Distribution	Plot position	Intercept	Slope	Correlation coefficient	MSE	Best one
G		-6.60	0.09	9.24E-01	3.33E-01	
W	P1	-56.83	11.89	**9.50E-01**	**3.25E-01**	W
F		-53.04	11.11	9.12E-01	3.33E-01	
G		-6.62	0.09	9.22E-01	3.33E-01	
W	P2	-57.04	11.93	**9.48E-01**	**3.25E-01**	W
F		-53.22	11.15	9.10E-01	3.33E-01	
G		-6.63	0.09	9.21E-01	3.33E-01	
W	P3	-57.12	11.95	**9.48E-01**	**3.25E-01**	W
F		-53.29	11.16	9.09E-01	3.33E-01	

Table 4 Parameter estimation of quarterly data

Distribution	Plot position	Intercept	Slope	Correlation coefficient	MSE	Best one
G		-8.12	0.10	8.99E-01	3.33E-01	
W	P1	-63.50	13.40	**9.63E-01**	**3.26E-01**	W
F		-69.01	14.06	8.88E-01	3.33E-01	
G		-8.18	0.10	8.95E-01	3.33E-01	
W	P2	-64.09	13.53	**9.60E-01**	**3.25E-01**	W
F		-69.58	14.18	8.84E-01	3.33E-01	
G		-8.21	0.10	8.93E-01	3.33E-01	
W	P3	-64.32	13.58	**9.58E-01**	**3.24E-01**	W
F		-69.79	14.22	8.82E-01	3.33E-01	

Table 5 Parameter estimation for yearly data.

Distribution	Plot position	Intercept	Slope	Correlation coefficient	MSE	Best one
G		-20.37	0.22	7.28E-01	3.32E-01	
W	P1	-140.04	30.01	**8.55E-01**	**3.31E-01**	**W**
F		-156.83	31.03	6.91E-01	3.32E-01	
G		-21.06	0.23	7.27E-01	3.33E-01	
W	P2	-146.90	31.48	**8.67E-01**	**3.23E-01**	**W**
F		-162.09	32.07	6.91E-01	3.33E-01	
G		-21.34	0.23	7.27E-01	3.33E-01	
W	P3	-149.93	32.12	**8.73E-01**	**3.24E-01**	**W**
F		-164.26	32.50	6.91E-01	3.33E-01	

From Table 3, to Table 5, it can be seen that the P1, P2 and P3 show that all three selection period maximums converge to the Weibull distribution. The best distribution model is chosen to have high correlation and minimum MSE. Results give an indication that the P1, P2 and P3 work well. For the chosen Weibull distribution, the probability papers plot for monthly, quarterly and yearly data are depicted in Fig2, Fig3 and Fig 4 respectively.

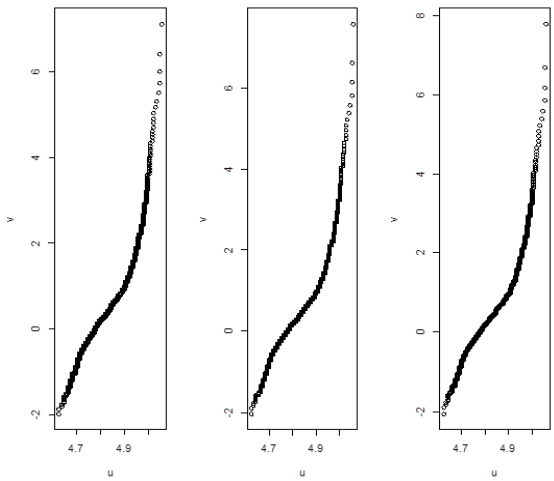

Fig 2: probability papers plots monthly data

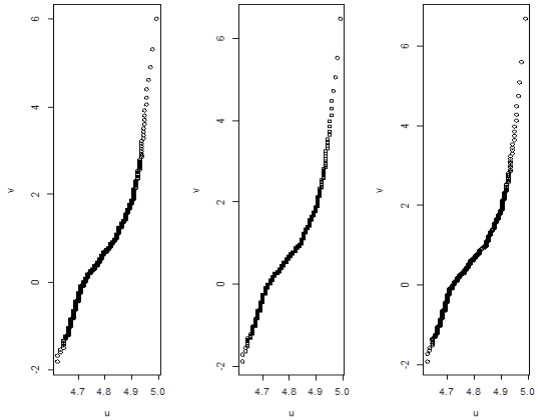

Fig 3:Probability papers plots for quarterly data

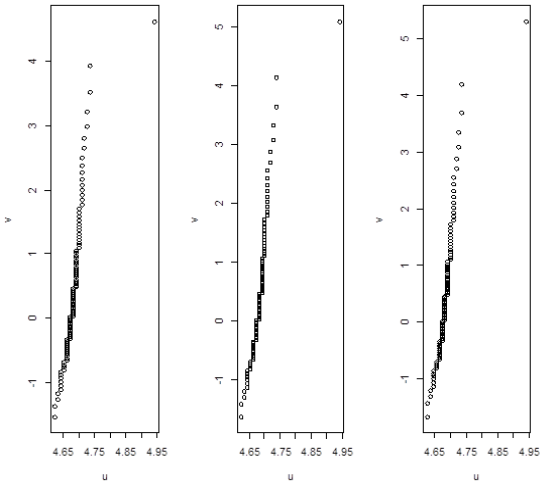

Fig 4: Probability papers plots for yearly data

The shape parameter estimates over different selection periods for the three methods of PPP, MLE and PWM are shown in Table 6. There are differences between the shape parameter estimates. In particular, the estimated values for shape parameters under PPP is smaller than MLE and PWM. In all, PPP estimates could probably be a better method with the smallest value of MSE.

Table6:Parameter estimates of PPP, MLE PWM

Period	Method	Shape	MSE	Best one
Monthly	PPP	-0.08	4.36E-01	PPP
	MLE	-0.50	5.29E-01	
	PWM	-0.39	4.90E-01	
Quarterly	PPP	-0.07	4.27E-01	PPP
	MLE	-0.59	5.78E-01	
	PWM	-0.56	5.61E-01	
Yearly	PPP	-0.03	3.93E-01	PPP
	MLE	-0.56	6.64E-01	
	PWM	-0.52	6.65E-01	

7 Conclusion

In this paper, we have presented probability paper plots (PPP) of Gumbel, Fréchet and Weibull distribution with different plotting position (PP). Extreme maximum temperature using 100 years of data from Fort Collins, Colorado is studied. Maximums of three different time periods data (monthly, quarterly and yearly) are fitted. Three analytical methods PPP, MLE and PWM are applied to select the best distribution which gives the smallest value of MSE. For each period when the data is fitted have the same distribution and results showed that the Weibull distribution whose shape parameter is obtained from PPP is the most appropriate for describing data. For further research it would be useful to derive PPP of minimum extreme distribution (Gumbel, Weibull and Fréchet).

References

[1]- S. Kotz and S. Nadarajah, *Extreme value distributions: theory and applications*. World Scientific, 2000.

[2]- S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*, vol. 208. Springer, 2001.

[3]- B. Gnedenko, "Sur la distribution limite du terme maximum d'une serie aleatoire," *Ann. Math.*, pp. 423–453, 1943.

[4]- F. L. Galeener, A. J. Leadbetter, and M. W. Stringfellow, "Comparison of the neutron, Raman, and infrared vibrational spectra of vitreous Si O 2, Ge O 2, and Be F 2," *Phys. Rev. B*, vol. 27, no. 2, p. 1052, 1983.

[5]- A. F. Jenkinson, "The frequency distribution of the annual maximum (or minimum) values of meteorological elements," *Q. J. R. Meteorol. Soc.*, vol. 81, no. 348, pp. 158–171, 1955.

[6]- O. B. Adeboye and M. O. Alatise, "Performance of probability distributions and plotting positions in estimating the flood of river Osun at Apoje Sub-basin, Nigeria," *Agric. Eng. Int. CIGR J.*, 2007.

[7]- S. L. Guo, "A discussion on unbiased plotting positions for the general extreme value distribution," *J. Hydrol.*, vol. 121, no. 1–4, pp. 33–44, 1990.

[8]- E. Castillo, A. S. Hadi, N. Balakrishnan, and J.-M. Sarabia, *Extreme value and related models with applications in engineering and science*. Wiley Hoboken, NJ, 2005.

[9]- M. Evans, N. Hastings, and B. Peacock, "Statistical distributions," 2000.

[10]- S. Harter and R. Pike, "The pictorial scale of perceived competence and social acceptance for young children," *Child Dev.*, pp. 1969–1982, 1984.

[11]- N. K. Goel and S. M. Seth, "Studies on plotting position formulae for Gumbel distribution," *Inst. Eng. Civ. Eng. Div. J.*, vol. 70, pp. 121–126, 1989.

[12]- M. De, "A new unbiased plotting position formula for Gumbel distribution," *Stoch. Environ. Res. Risk Assess.*, vol. 14, no. 1, pp. 1–7, 2000.

[13]- J. R. M. Hosking and J. R. Wallis, "Parameter and quantile estimation for the generalized Pareto distribution," *Technometrics*, vol. 29, no. 3, pp. 339–349, 1987.

[14]- T. Haktanira and A. Bozduman, "A study on sensitivity of the probability-weighted moments method on the choice of the plotting position formula," *J. Hydrol.*, vol. 168, no. 1–4, pp. 265–281, 1995.

[15]- Y. Lei, "Evaluation of three methods for estimating the Weibull distribution parameters of Chinese pine (Pinus tabulaeformis)," *J. For. Sci.*, vol. 54, no. 12, pp. 566–571, 2008.

[16]- Y. Yao, L. Song, Y. Katz, and G. Galili, "Cloning and characterization of Arabidopsis homologues of the animal CstF complex that regulates 3′ mRNA cleavage and polyadenylation," *J. Exp. Bot.*, vol. 53, no. 378, pp. 2277–2278, 2002.