A Nonparametric Way of Distribution Testing

Ekrem Kilic
Istanbul Bilgi University

June 30, 2005

Abstract

Testing the distribution of a random sample can be considered, indeed, as a goodness-of-fit problem. If we use the nonparametric density estimation of the sample as a consistent estimate of exact distribution, the problem reduces, more specifically, a distance between two functions. This paper considers the distribution testing problem from this point of view and suggests a nonparametric procedure. Although the procedure is applicable for all distributions, paper emphasizes on normality test.The critical values for this normality test are generated by using Monte Carlo techniques.

1 Introduction

The distribution of a random variable is one of the most important question to be answered by many econometric studies. Econometricians needs to assume or know the distribution of a random variable to be able to make inference and, sometimes, efficient estimation as in the classical linear regression model where the unobserved disturbance vector , ε , is assumed to be normally distributed. Of course , then, these assumptions on the distribution must be tested. Therefore the literature on distribution testing is quite deep, there exists many studies on this topic. Since normality is the most common assumption, many of these studies examined the normality. These normality tests can be divided into two classes; parametric and nonparametric tests.

Let us first consider some parametric tests. One of the most common parametric normality test is ,of course, the Jarque-Bera test. This test based on the moment properties of the normal distribution. Jarque-Bera statistic is, simply, a function of skewness and kurtosis (see Jarque-Bera [5]) and asymptotically chi-squared distributed. Therefore Jarque-Bera test is very usefull, becuase it requires no special table for the critical values. Another parametric test is Shapiro-Wilk test. This test also uses moments of the distribution, however it uses a weighted sum of squared random variables. Shapiro and Wilk [9] provided the critical values for the test. Another class of parametric distribution tests those exploits a feature of the normal distribution is proposed by Vasicek [12]. As described by Prescott [7];

> Among all distributions that posses a density function f and have a given variance σ^2, the entropy $H(f)$, defined as,
>
> $$H(f) = \int_{-\infty}^{\infty} f(x) \log[f(x)] dx \tag{1}$$
>
> is maximized by the normal distribution.

Using this feature Vasicek defined a sample entropy test statistic. Vasicek's sample entropy test is a distribution free test.

There exists also some nonparametric tests. One important test is proposed by Kolmogorov and Smirnov. Kolmogorov and Smirnov's test is based on the empirical distribution function of the sample. This test statistic is maximum of the absolute difference between empirical distribution function and cumulative normal distribution. Kolmogorov and Smirnov test is strong in the sense that distribution of the test statistic itself does not depend on the underlying distribution that is tested. Therefore the critical values for this test do not change for non-normal distributions. Another nonparametric test

that uses the empirical distribution function is Cramer-von Mises test. The notion of maximum difference which is used by Kolmogorov - Smirnov test, is replaced with the integrated squared differences. This test is more powerful than the Kolmogorov-Smirnov's test because it considers the whole distribution by integrating the squared differences whereas the Kolmogorov-Smirnov test uses just the maximum of the distance at data points. Approximate critical values can be found in [1] by Andersen and Darling.

This paper will introduce a nonparametric distribution test that is based on the kernel density estimation and simple euclidian measure of distance between functions. The organization of the paper will be as follows. In the next section the general framework will be presented and the following section will consider the power of the test by using Monte Carlo simulations. Finally, in the conclusion section, the results will be discussed.

2 Test Procedure

Parametric statistics defines the form of a distribution with functions like $f(x, \theta)$. The functional form of normal distribution is as follows;

$$f(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{2}$$

where θ is the parameter vector that includes μ and σ.

If the parameters, μ and σ , are set to 0 and 1 respectively, the distribution can be written as;

$$f(x;0,1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{3}$$

this form is called as standard normal density and again it is a function of x. This kind of a setup is very restrictive and must be tested.

Nonparametric methods, on the other hand, suggest histograms or smoothed histograms for density estimation problem. One important feature of the histograms is it integrates to unity. Histograms , however , produces rough and discontinues density estimates. Therefore Kernel estimators are more useful because of their continuity and smoothness. Kernel density estimators are just smoothed histograms. One can formally write a histogram function as follows;

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} I\left(-1/2 \leq \frac{x_i - x}{h} \leq 1/2\right) \tag{4}$$

where $I(.)$ is the indicator function and h is called as bandwidth or smoothing parameter. In the smoothed version of the histograms, we need a smoother

function, called kernel. Rosenblatt [8] describes the kernel estimator as;

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\psi_i) \tag{5}$$

where $\int_{-\infty}^{\infty} K(\psi)\, d\psi = 1$ and $\psi_i = \frac{x_i - x}{h}$. Obviously if one choose the indicator function as the kernel, 5 will be exactly same with 4. Hence by replacing the indicator function with smoothing functions that satisfies the condition of unit integral, we can define different kernel estimators those are , indeed, just smoothed histograms.One of the most common kernel function is standard normal density function (as 3), generally called as gaussian kernel. For this kernel bandwidth can be chosen $h = n^{-\frac{1}{5}}\sigma$(see Pagan and Ullah [6]).

As described kernel density estimation methods provides a continuous distribution function for every random sample. The problem of goodness-of-fit, then, becomes a problem of distance. We can simply measure the distance , in an Euclidian fashion, between kernel estimate of density and underlying parametric distribution as follows;

$$D = \sqrt{\int_{-\infty}^{\infty} [\hat{f}(x) - f(x)]^2 dx} \tag{6}$$

where $\hat{f}(x)$ is kernel estimation of the density, $f(x)$ is the underlying distribution's probability density function. To integrate this function, numerical integration methods can be used. Numerical integration approach in this paper is a piecewise method that integrates the function using Newton-Cotes formulas (see Burden and Faires [2]). The method is called Composite Trapezoidal rule and can be written as follows;

$$\int_b^a f(x)\, dx = \frac{h}{2}[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)] - \frac{b-a}{12} h^2 f''(\mu). \tag{7}$$

where $h = (b-a)/n$, and $x_j = a + jh$ for each $j = 0, 1, ..., n$. In this formula, $\frac{h}{2}[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)]$ is the numerical integrated result for the integral at the left hand side and $\frac{b-a}{12} h^2 f''(\mu)$ is the error of the numerical integration. When the error is equal to zero, the numerical and analytical integral will be the same. By this formula it is clear that as h goes to zero, the error part of the equation tends to zero. And also as n goes to infinity, h tends to zero. Then if n, number of integrated pieces, is large enough, the numerical integration will be approximately the same with analytical one. In

this paper n is taken 200[1], a and b, upper and lower limits of the integral, are chosen 10 and -10 respectively.

The question, then, what is the critical values for the test? Critical values found using Monte Carlo simulations. The method is quite simple, I generated a random sample coming from normal distribution containing certain number of elements, and applied the test and stored the values that test produced. Repeating this process built up a probability distribution for the test statistics for the given sample size. I repeated the experiment for different sample sizes. Since the test is a distance test, it has only positive values and therefore I only dealt with the right tail of distribution. Table 1 obtained by repeating this process 100000 times for each sample size. Convergence graphs of the test statistics are shown at Figure 1.

3 Power and Size of the Test

After defining the test and obtaining critical values, we need to check power and size of the test. I used Monte Carlo simulations for this purpose too, and followed the methodology of Stengos and Wu [10].

The power can be defined as one minus probability of TypeII error or equivalently the probability that test rejects the null hypothesis when it should be rejected. I examined power of the distance to normality test against t distribution , mixture of two normals , lognormal distribution, chi distribution, exponential distribution and 4 different version of generalized lambda distribution.The normal distribution is also considered for the size of the test. For each of the distributions, 100000 random samples are generated at size $n = 20, 50, 100, 200, 500$ and 1000. Simulated distributions are listed below;

[1]I tried higher values of n for sample sizes 20, 50, 100, 200, 500 and 1000. It provided negligible improvements.

Norm	$N(0,1)$
t5	t_5
M2Norm	$z_1 I(p \leq 0.5) + z_2 I(p > 0.5)$ $where\ z_1 \sim N(-1,1), z_2 \sim N(1,1)\ \ and\ p \sim N(-1,1)$
LNorm	$\exp(z)\ \ where\ z \sim N(0,1)$
Chi3	χ_3^2
Exp	$-\ln(u)\ \ where\ u \sim U(0,1)$
Lam1	$\lambda_1 + \left(u^{\lambda_3} - \left(1-u^{\lambda_4}\right)\right)/\lambda_2$ $where\ parameters\ \Lambda = [0, 0.19754, 0.134915, 0.134915]$
Lam2	$\lambda_1 + \left(u^{\lambda_3} - \left(1-u^{\lambda_4}\right)\right)/\lambda_2$ $where\ parameters\ \Lambda = [0, -1, -0.8, -0.8]$
Lam3	$\lambda_1 + \left(u^{\lambda_3} - \left(1-u^{\lambda_4}\right)\right)/\lambda_2$ $where\ parameters\ \Lambda = [0, 1, 1.4, 0.25]$
Lam4	$\lambda_1 + \left(u^{\lambda_3} - \left(1-u^{\lambda_4}\right)\right)/\lambda_2$ $where\ parameters\ \Lambda = [0, -1, -0.0075, -0.03]$

Generalized Lambda distribution is an extended version of Tukey's Lambda Distribution. Inverse of the cumulative density function of generalized lambda distribution can be written as follows;

$$F^{-1}(u) = \lambda_1 + \frac{\left(u^{\lambda_3} - \left(1-u^{\lambda_4}\right)\right)}{\lambda_2} \tag{8}$$

Generalized lambda distribution can provide a wide range of symmetric and asymmetric distributions. Lam1 is approximately the same with normal distribution(see Figure 2,Figure 3,Figure 4,Figure 5).

Table 2 reports power and size tests. The simulations show that the distance to normality is less powerful with symmetric distributions like t5 and M2Norm. On the other hand, the test performs very good against asymmetric distributions so that it has quite high power even in small samples.For all asymmetric distributions having larger sample size than 200, the test's power is equal to one with 95% confidence level.

Size of the test is trivial for our case, because the critical values have already derived with Monte Carlo simulations. Table 2 shows that for all samples, size is equal to the significance level of the test. An interesting result is Lam1, an approximation to the normal distribution, have nearly same outputs with normal distribution. While sample size increases , the approximation starts to differ from exact normal distribution and test produces different results.

4 Conclusion

Distribution testing is very important for many studies,in this paper, a nonparametric density testing procedure is described. The test uses very simple euclidian definition of distance. Using this definition, a distance defined between the nonparametric estimation of density and the parametric form of underlying density (In this paper normality is considered). After describing the testing procedure and Monte Carlo simulations for critical values, paper analyzed the power and size of the test. Monte Carlo simulations pointed out that the test is very powerful against asymmetric distributions and even in the small samples, it produces very good results. However, the power is very limited against t distribution. For further studies I tend to answer two questions. First, for other distributions can this critical values be used or will they change for different distributions? Second, for other distributions will it produce same kind of power and size properties?

References

[1] Anderson, T., D. Darling, 1952. *"Asymptotic theory of goodness of fit criteria based on stochastic processes"*. Annals of Mathematical Statistics 23, 193-212.

[2] Burden, R.L., J.D. Faires, 2005. Numerical Analysis. Eigth Edition, Chp: 4, P:196-203, Thomson Higher Education, Belmont, CA,USA

[3] D'Agostino,Ralph D., B. Rossman, 1974. *"The power of geary's test of normality"*. Vol:64, No:1, P:181-184, Biometrika

[4] Gastwirth,Joseph L., M.E.B. Owens, 1977. *"On classical test of normality"*. Vol:64, No:1,P:135-139, Biometrika

[5] Jarque, C., A. Bera, 1980. *"Efficient tests for normality homoscedasticity and serial independence of regression residuals"*. Econometric Letters 6, 255-259.

[6] Pagan, A., A. Ullah, 1999. Nonparametric Econometrics. First Edition, Chp:1, P:71-77, Cambridge University Press

[7] Prescott, P., 1976. *"On a test for normality based on sample entropy"*.Journal of the Royal Statistical Society, Series B, Vol. 38, No.3, 254-56.

[8] Rosenblatt, M. 1956. *"Remarks on some nonparametric estimates of a density function"*. Annals of Mathematical Statistics, 27(3):832-837.

[9] Shapiro, S., M. Wilk, 1965. *"An analysis of variance test for normality (complete samples)"*. Biometrika 52, 591-611.

[10] Stengos, T., X. Wu, 2005.*"Information-Theoratic distribution tests with application to symmetry and normality"*.

[11] Thadewald, T., H. Büning, 2004. *"Jarque-Bera test and its competitors for tsting normality- A power comparison"*. Volkswirtschaftliche Reihe, Free University Berlin

[12] Vasicek, O., 1976. *"A test for normality based on sample entropy"*, Journal of the Royal Statistical Society, Series B, 38, 54-59.

Table 1: The critical values for the distance to normality test.

	α						
Sample Size	**0.5**	**0.75**	**0.85**	**0.9**	**0.95**	**0.975**	**0.99**
10	0.105	0.126	0.136	0.143	0.154	0.162	0.172
11	0.103	0.124	0.135	0.142	0.152	0.161	0.171
12	0.101	0.122	0.133	0.141	0.151	0.160	0.171
13	0.099	0.120	0.131	0.139	0.150	0.159	0.170
14	0.097	0.119	0.130	0.138	0.149	0.159	0.169
15	0.095	0.117	0.128	0.136	0.147	0.157	0.168
16	0.094	0.115	0.126	0.134	0.146	0.155	0.167
17	0.092	0.113	0.125	0.133	0.145	0.155	0.165
18	0.091	0.112	0.123	0.131	0.143	0.153	0.164
19	0.089	0.110	0.122	0.130	0.141	0.151	0.163
20	0.088	0.109	0.120	0.128	0.140	0.150	0.162
21	0.087	0.108	0.119	0.127	0.139	0.149	0.160
22	0.086	0.106	0.118	0.125	0.137	0.148	0.158
23	0.085	0.105	0.117	0.124	0.136	0.146	0.157
24	0.084	0.104	0.116	0.123	0.134	0.144	0.156
25	0.083	0.103	0.115	0.122	0.134	0.144	0.156
26	0.082	0.102	0.113	0.121	0.132	0.142	0.154
27	0.081	0.101	0.112	0.120	0.131	0.141	0.153
28	0.081	0.100	0.111	0.119	0.130	0.140	0.151
29	0.080	0.099	0.110	0.118	0.129	0.139	0.150
30	0.079	0.098	0.109	0.117	0.128	0.138	0.149
31	0.078	0.097	0.108	0.116	0.127	0.137	0.148
32	0.078	0.097	0.107	0.115	0.126	0.136	0.147
33	0.077	0.095	0.106	0.114	0.125	0.135	0.146
34	0.076	0.094	0.105	0.113	0.124	0.133	0.145
35	0.076	0.094	0.105	0.112	0.123	0.133	0.145
36	0.075	0.093	0.104	0.111	0.122	0.132	0.144
37	0.075	0.093	0.103	0.110	0.122	0.131	0.142
38	0.074	0.092	0.102	0.110	0.121	0.131	0.142
39	0.073	0.091	0.102	0.109	0.120	0.129	0.140
40	0.073	0.091	0.101	0.108	0.119	0.129	0.140
41	0.073	0.090	0.100	0.107	0.118	0.127	0.138
42	0.072	0.089	0.099	0.107	0.117	0.126	0.138
43	0.071	0.089	0.099	0.106	0.116	0.126	0.137

continues...	α						
Sample Size	**0.5**	**0.75**	**0.85**	**0.9**	**0.95**	**0.975**	**0.99**
44	0.071	0.088	0.099	0.106	0.116	0.125	0.137
45	0.071	0.088	0.098	0.104	0.115	0.124	0.135
46	0.070	0.087	0.097	0.104	0.115	0.124	0.135
47	0.070	0.086	0.097	0.104	0.114	0.124	0.135
48	0.069	0.086	0.096	0.103	0.113	0.122	0.133
49	0.069	0.085	0.095	0.102	0.113	0.122	0.132
50	0.069	0.085	0.095	0.102	0.112	0.121	0.132
51	0.068	0.085	0.094	0.101	0.111	0.120	0.131
52	0.068	0.084	0.094	0.100	0.111	0.120	0.130
53	0.067	0.084	0.093	0.100	0.110	0.119	0.130
54	0.067	0.083	0.093	0.099	0.109	0.119	0.129
55	0.067	0.083	0.092	0.099	0.109	0.118	0.129
56	0.066	0.082	0.092	0.098	0.108	0.117	0.128
57	0.066	0.082	0.091	0.098	0.108	0.117	0.127
58	0.066	0.081	0.091	0.097	0.107	0.116	0.126
59	0.065	0.081	0.090	0.097	0.107	0.116	0.126
60	0.065	0.080	0.090	0.096	0.106	0.115	0.125
61	0.065	0.080	0.089	0.096	0.106	0.114	0.125
62	0.064	0.080	0.089	0.095	0.105	0.114	0.124
63	0.064	0.079	0.089	0.095	0.104	0.113	0.123
64	0.064	0.079	0.088	0.094	0.104	0.113	0.122
65	0.064	0.079	0.088	0.094	0.103	0.112	0.122
66	0.063	0.078	0.087	0.094	0.103	0.112	0.122
67	0.063	0.078	0.087	0.093	0.103	0.111	0.122
68	0.063	0.078	0.087	0.093	0.102	0.111	0.120
69	0.063	0.077	0.086	0.092	0.102	0.110	0.120
70	0.062	0.077	0.086	0.092	0.101	0.110	0.119
71	0.062	0.077	0.085	0.092	0.101	0.109	0.119
72	0.062	0.077	0.085	0.091	0.101	0.109	0.119
73	0.061	0.076	0.085	0.091	0.101	0.109	0.119
74	0.061	0.076	0.085	0.091	0.100	0.108	0.118
75	0.061	0.075	0.084	0.090	0.099	0.107	0.117
76	0.061	0.075	0.084	0.090	0.099	0.108	0.117
77	0.061	0.075	0.083	0.089	0.098	0.106	0.116

continues...	α						
Sample Size	**0.5**	**0.75**	**0.85**	**0.9**	**0.95**	**0.975**	**0.99**
78	0.060	0.075	0.083	0.089	0.098	0.106	0.116
79	0.060	0.074	0.083	0.089	0.098	0.106	0.116
80	0.060	0.074	0.083	0.089	0.098	0.106	0.115
81	0.060	0.074	0.082	0.088	0.097	0.105	0.115
82	0.060	0.074	0.082	0.088	0.097	0.105	0.115
83	0.059	0.073	0.081	0.087	0.096	0.104	0.113
84	0.059	0.073	0.081	0.087	0.096	0.104	0.113
85	0.059	0.072	0.081	0.087	0.096	0.104	0.113
86	0.059	0.072	0.081	0.087	0.095	0.103	0.112
87	0.058	0.072	0.080	0.086	0.095	0.103	0.113
88	0.058	0.072	0.080	0.086	0.095	0.103	0.112
89	0.058	0.072	0.080	0.086	0.094	0.102	0.112
90	0.058	0.071	0.080	0.085	0.094	0.102	0.111
91	0.058	0.071	0.080	0.085	0.094	0.102	0.111
92	0.058	0.071	0.079	0.085	0.093	0.101	0.111
93	0.057	0.071	0.079	0.085	0.093	0.101	0.110
94	0.057	0.070	0.078	0.084	0.093	0.100	0.109
95	0.057	0.070	0.078	0.084	0.092	0.100	0.109
96	0.057	0.070	0.078	0.083	0.092	0.100	0.109
97	0.056	0.070	0.078	0.083	0.092	0.100	0.109
98	0.056	0.070	0.078	0.083	0.092	0.099	0.108
99	0.056	0.069	0.077	0.083	0.091	0.099	0.108
100	0.056	0.069	0.077	0.083	0.091	0.099	0.107
200	0.045	0.056	0.062	0.066	0.073	0.079	0.086
500	0.034	0.041	0.045	0.048	0.053	0.057	0.062
1000	0.027	0.032	0.035	0.038	0.041	0.044	0.048

Table 2: Power and size of distance to normality test

	n=20		n=50		n=100		n=200		n=500		n=1000	
	0.95	0.99	0.95	0.99	0.95	0.99	0.95	0.99	0.95	0.99	0.95	0.99
Norm	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
t5	0.07	0.02	0.08	0.03	0.14	0.07	0.29	0.16	0.73	0.55	0.97	0.92
M2Norm	0.09	0.02	0.17	0.05	0.28	0.11	0.48	0.23	0.83	0.62	0.98	0.93
LNorm	0.83	0.66	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Chi3	0.47	0.24	0.89	0.72	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00
Exp	0.65	0.41	0.98	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Lam1	0.05	0.01	0.05	0.01	0.05	0.01	0.04	0.01	0.04	0.01	0.04	0.01
Lam2	0.65	0.55	0.97	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Lam3	0.41	0.17	0.86	0.63	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00
Lam4	0.28	0.13	0.30	0.14	0.90	0.77	1.00	0.99	1.00	1.00	1.00	1.00

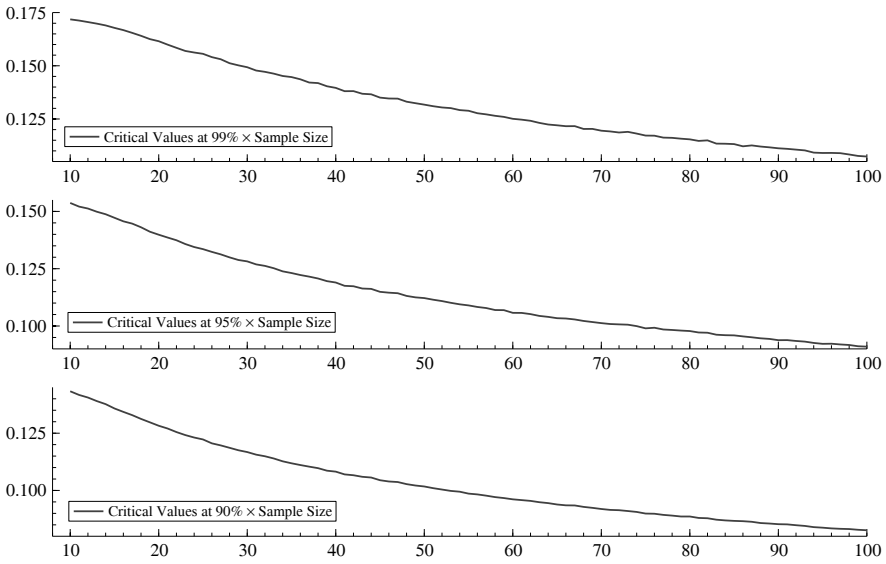

Figure 1: Convergence graphs of the test statistics at 0.99, 0.95, and 0.9 significance levels.

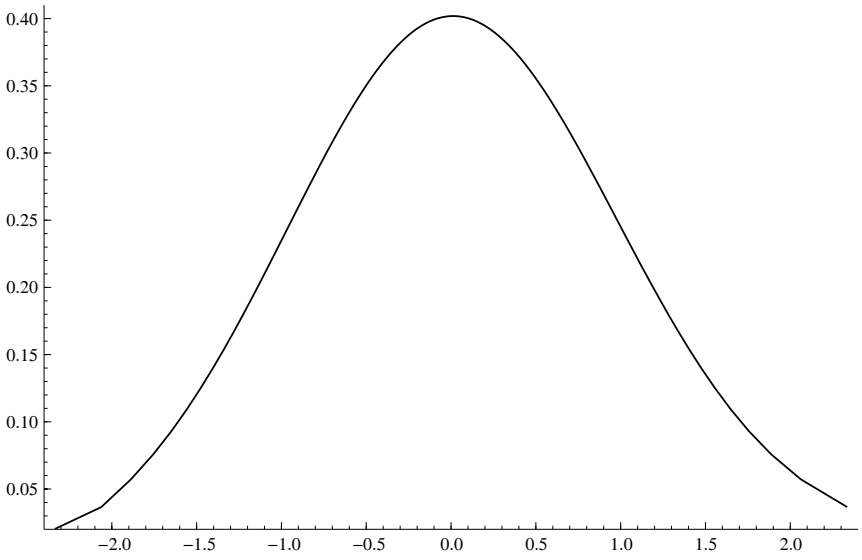

Figure 2: Graph of Lam1

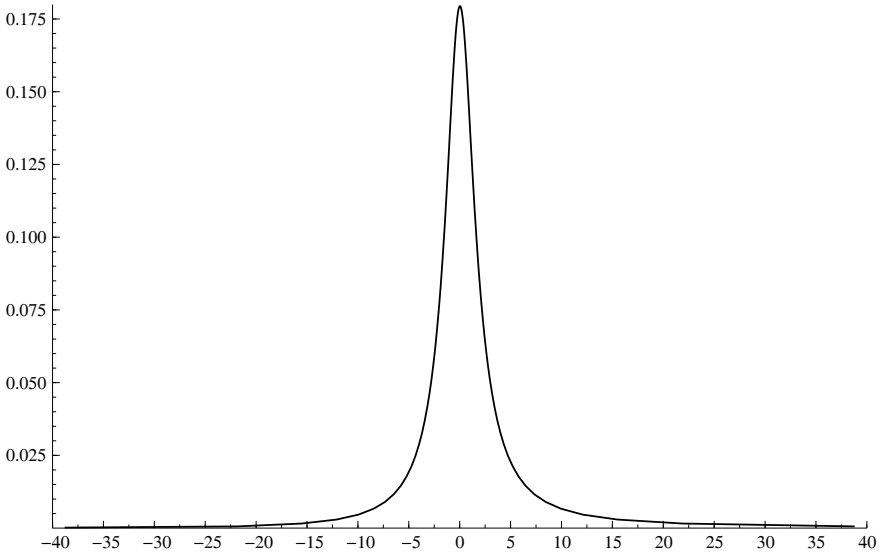

Figure 3: Graph of Lam2

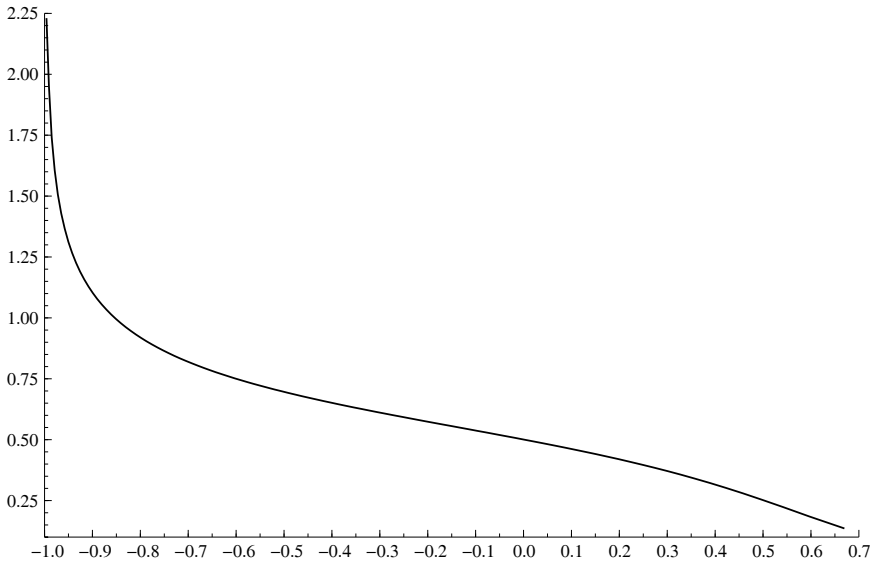

Figure 4: Graph of Lam3

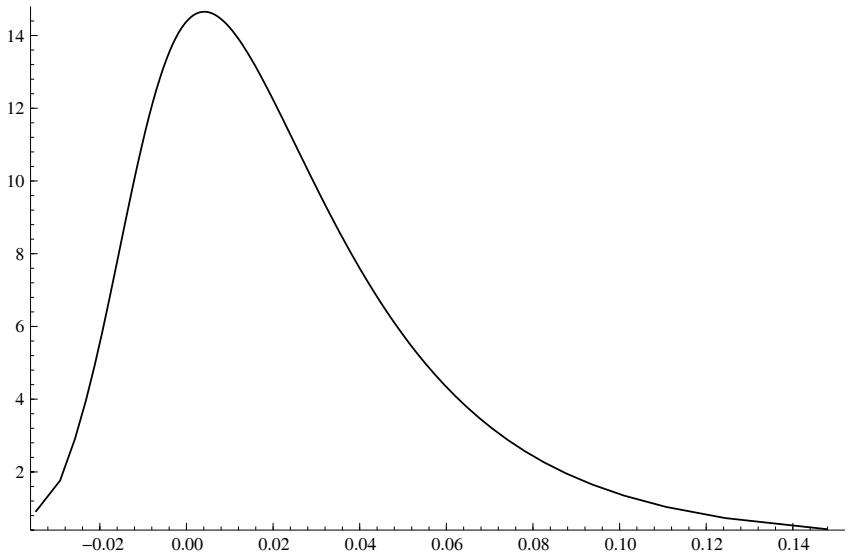

Figure 5: Graph of Lam4