An iterative algorithm for the Cramer-von Mises distance estimator

Jinhyo Kim [1] and Sangyeol Lee [2]

Abstract

This article concerns the numerical computing of the Cramer-von Mises distance estimator, which is known to have the desirable statistical properties such as the robustness and efficiency. Here, it is shown that the usual optimization algorithms, such as the Newton-Raphson method and the Bisection method, fail to find the estimator. As an alternative, a derivative-free grid-type algorithm, the Dichotomous Search method, is considered. The simulation results show that the Dichotomous Search method tends to find the estimator correctly.

Key Words: Cramer-von Mises distance, Minimum distance estimator, Unimodal, Derivative-free, Dichotomous Search, Convergence rate

1 Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with the distribution function F. Consider a parametric family $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$, where Θ is an open set in the real line. Define a minimum distance functional:

$$T(F) = \arg\min d_F(\theta) \tag{1}$$

[1]Researcher, The Research Institute for Basic Sciences, Seoul National University, Seoul, 151-742, Korea. This research was supported by KOSEF. JINKIM@STATS.SNU.AC.KR The author thanks to Professor R. G. Krutchkoff for his invaluable guide and suggestion of this paper's topic, during the entire PhD research.

[2]Assistant Professor, Department of Statistics, Seoul National University, Seoul, 151-742, Korea. SYLEE@STATS.SNU.AC.KR

where $d_F(\theta)$ is a criterion function, measuring the discrepancy between F and F_θ. If F belongs to $\mathcal{F}$ and the true parameter is θ_0, then $T(F) = \theta_0$. The true parameter θ_0 can be estimated by substituting the sample distribution function F_n for F. The estimator, defined as $T(F_n)$, is called a minimum distance estimator.

To specify the criterion function $d_F(\theta)$, we consider the set , consisting of all non-negative continuous functions G such that $G \geq 0$, $G(0) = 0$, $G^{'}(0) = 0$ and $G^{''}(0) = 0$. Define

$$d_F(\theta) = \int G[\delta_F(t,\theta)]w(t,\theta)dF(t) \tag{2}$$

where $\delta_F(t,\theta) = F(t) - F_\theta(t)$, $w(t,\theta) \geq 0$ is a weight function, and G is in the set , . The minimiser $\hat{\theta}$ of $d_{F_n}(\theta)$ is called the generalized weighted Cramer-von Mises distance estimator, which is presented in Öztürk and Hettmansperger (1997). It is well-known that the Cramer-von Mises estimator has the desirable statistical properties such as the robustness and efficiency and the asymptotic normality. As for G, one usually utilizes $G_1(t) \equiv t^2$, $G_2(t) \equiv t^2/(t+1)$ and $G_3(t) \equiv \left[(t+1)^{1/2} - 1\right]^2$. The functions of G_2 and G_3 are the Hellinger and Neyman distance functions, respectively, in the context of the minimum disparity function of Lindsay (1994).

Here, in order to obtain a Cramer-von Mises distance estimator, we employ $G_1 = t^2$ and $w(t,\theta) \equiv 1$. In this case, we have

$$\begin{aligned} d_{F_n}(\theta) &= \int_{-\infty}^{\infty} (F_n(t) - F_\theta(t))^2 dF_n(t) \\ &= \frac{1}{n}\sum_{i=1}^{n} (F_\theta(X_{[i]}) - \frac{i}{n})^2 \\ &\cong \frac{1}{n}\sum_{i=1}^{n} (F_\theta(X_{[i]}) - \frac{i}{n+1})^2. \end{aligned} \tag{3}$$

The Cramer-von Mises estimator $\hat{\theta}$ is defined to be θ that minimizes Formula (3). Usually, the estimator is obtained by solving the equation $\partial d_{F_n}(\theta)/\partial\theta = 0$. Since, in practice, the solution cannot be obtained using any analytical methods, one should resort to some numerical techniques for minimizing the objective function $d_{F_n}(\theta)$. In order to solve the equation $\partial d_{F_n}(\theta)/\partial\theta = 0$, the first choice among the popular numerical methods

will be the Newton-Raphson method, which is based on the first and second derivatives of the objective function. However, a solution of the equation is not necessarily a global optimum: it may be a local minimum, a local maximum or a saddle point. It is because despite the function $\theta \to d_{F_n}(\theta)$ has a unimodal feature, the function is not truly unimodal (see Definition 1 below for the unimodality) due to the noisy fluctuations as seen in Figure 4. Since the Newton-Raphson method does not perform well and even fails to find the solution that minimizes $d_{F_n}(\theta)$, here we consider using the Dichotomous Search method, which is a derivative-free grid-type optimization method.

Our goal of this paper is to demonstrate through simulation studies that the Dichotomous Search method finds the Cramer-von Mises estimator more correctly whereas the others including the Newton-Raphson method misbehave in searching for the estimator *numerically*. In Section 2, we provide a precise outline on the Dichotomous Search method for readers. In Section 3, the simulation results are reported for the normal random variables for comparing the performance of the Dichotomous Search method with other popular methods.

2 Dichotomous Search

In this section, we review the derivative-free Dichotomous Search method. Since the $d_{F_n}(\theta)$ has a unimodal shape, we start with a mathematical definition of unimodality.

(Bazarra, Sherali and Shetty, 1993)

A function f is unimodal iff for each x^1, x^2 with $f(x^1) \neq f(x^2)$ and for $0 < \lambda < 1$

$$f(\lambda x^1 + (1-\lambda)x^2) \leq \max\{f(x^1), f(x^2)\}. \tag{4}$$

In the literature, it is well-known that the unimodality is not a sufficient condition for the correct convergence of the Newton-Raphson method. It is easy to find out a counterexample. However, as will be shown below, the Dichotomous Search method *guarantees* the correct convergence.

With the unimodal curves in Figure 1, we illustrate the Dichotomous Search algorithm.

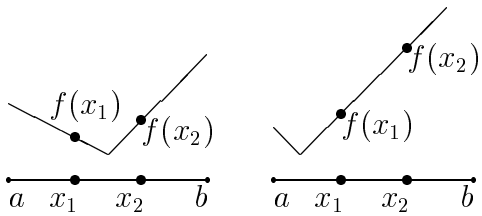

Figure 1: test points

Given an initial region of interest $\mathcal{I} = [a, b]$, as described in Figure 1, we evaluate the values at the two test points x_1 and x_2 with $x_1 < x_2$. If $f(x_1) \leq f(x_2)$, then the new interval of 'uncertainty' becomes $[a, x_2]$ since the optimum point cannot exist in $(x_2, b]$. Otherwise if $f(x_1) > f(x_2)$, then the new interval of uncertainty is $[x_1, b]$. Notice that depending on the value comparison of f at x_1 and x_2, the length of the new interval of uncertainty is either equal to $b - x_1$ or $x_2 - a$, which is less than $b - a$. In selecting x_1 and x_2, one usually takes them symmetrically around the midpoint $(b + a)/2$ of a and b with certain distance $\epsilon > 0$. Depending on the values of f at x_1 and x_2, as mentioned above, a new interval of uncertainty is determined. This procedure is repeated with placing two new observations x_1 and x_2 for the next iteration until it terminates. In fact, this procedure works for any $a < x_1 < x_2 < b$. However, a particular choice of $\epsilon = x_2 - (b - a)/2$ yields an optimal algorithm, so called the 'Golden Section Search', with the 'golden number' $\alpha \equiv (x_2 - a)/(b - a) = (\sqrt{5} - 1)/2 \cong 0.618$ (cf. Kim, 1997).

In the following, we summarize the Dichotomous Search method using the above iterative scheme.

Algorithm for the Dichotomous Search

Initialization step

Choose a constant α (=0.618) and an allowable final length of uncertainty, $l > 0$. Let $[a_1, b_1] = [a, b]$ be the initial interval of uncertainty, and note that the initial interval $\mathcal{I} = [a_1, b_1]$ includes the optimum point, and let $k = 1$ and go to the main step.

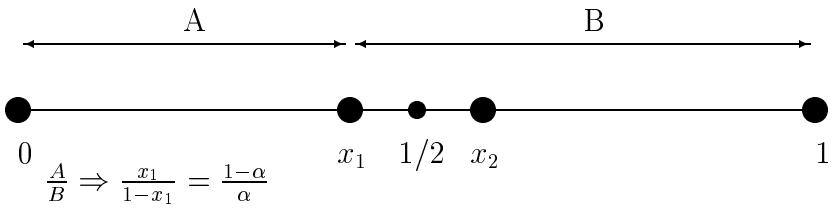

Figure 2: Using the fixed ratio α at kth iteration

Main Step

1. If $b_k - a_k < l$, then stop; the minimum point lies in $[a_k, b_k]$. Otherwise consider x_1 and x_2 defined in Formulae (5) and (6) ; go to step 2.

$$x_1 = a_k + \alpha(b_k - a_k) = (1-\alpha)a_k + \alpha b_k \tag{5}$$

$$x_2 = b_k - \alpha(b_k - a_k) = \alpha a_k + (1-\alpha)b_k \tag{6}$$

2. If $f(x_1) < f(x_2)$, let $a_{k+1} = a_k$ and $b_{k+1} = x_2$. Otherwise let $a_{k+1} = x_1$ and $b_{k+1} = b_k$. Replace k by $k+1$, go to step 1.

In Formulae (5) and (6), it should be noted that one of the values at the two test points in current iteration can be reused in the next iteration if the optimal constant $\alpha = 0.618\ldots$ is given, with which the the Golden Section Search is achieved as illustrated in Figures 2 and 3. It implies that the Golden Section Search requires *only one* additional test point in each iteration step. However, due to its floating-point representation in a digital computer (the digital computer cannot recognize irrational numbers), in practice one cannot use the above algorithm as precisely as above. Given $\alpha = 0.618\ldots$, the test point in next iteration step will be only *mathematically* coincided, but not *computationally*. Therefore, it is recommended to store 1)the locations x_1 and x_2 and 2)the values $f(x_1)$ and $f(x_2)$, and to reuse one of them without evaluating over again in the next iteration.

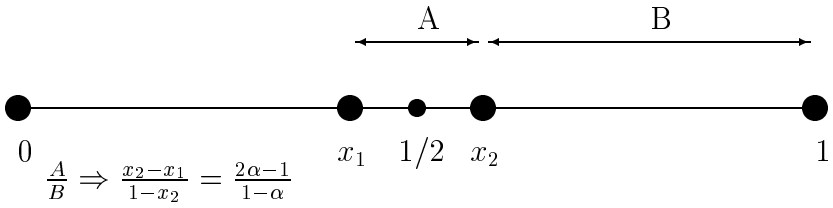

Figure 3: Using the fixed ratio α at *(k+1)*th iteration

The following two theorems, which can be found in Bazarra, Sherali and Shetty (1993), validate the convergence of the Dichotomous Search method.

Theorem 1 *Let $f : R^1 \longrightarrow R^1$ be unimodal over $[a, b]$ with $x_1 \leq x_2 \in [a, b]$. If $f(x_1) \leq f(x_2)$, then $f(z) \geq f(x_1)$ for all $z \in (x_2, b]$; and if $f(x_1) > f(x_2)$, then $f(z) \geq f(x_2)$ for all $z \in [a, x_1)$.*

From Theorem 1, we know that if $f(x_1) < f(x_2)$, then there must not exist an optimum point in $[x_2, b]$ since $f(z) \geq f(x_1)$ for all $z \in [x_2, b)$. Now we eliminate the region $[x_2, b)$ to get the new interval of uncertainty $[a, x_2]$ for the next iteration step. In this way, our region of interest will be reduced in each step until we reach the optimum point within an allowable final length of uncertainty. A similar argument follows for the case of $f(x_1) \geq f(x_2)$.

Theorem 2 *Consider the problem of minimizing a unimodal function $f(x)$ defined on an open set $\mathcal{S} \subset \mathcal{R}^p$. If x is a local optimal solution using the Dichotomous Search method, then x is also the global solution.*

By Theorem 2, we are convinced that if we find *one* local optimum point of a unimodal function $f(x)$, then it is necessarily *the* global optimum point. In the above two theorems, whenever a function to be minimized is unimodal, the Dichotomous Search *always* finds the optimum point. This is *definitely a great advantage* over the Newton-Raphson method, since Theorem 2 does not hold for the Newton-Raphson method.

derivative	quadratic	Newton-Raphson	2.0
derivative-free	superlinear	Muller's method	1.839
		Secant method	1.618
		Illinois method	1.442
	linear	Dichotomous Search	1.0
		Bisection	1.0

Table 1: (Comparison of the theoretical order p of convergence in Formula (7))

Given two competitive convergent algorithms, a theoretical comparison for convergence speed could be made on the basis of the mathematical order of convergence. Let the sequence $\{x_k, k = 1, 2, \ldots\} \subset \mathcal{R}^1$ converge to $\overline{x} \in \mathcal{R}^1$ in L_1-norm. The *order of convergence* of the sequence is defined as the supremum of $p \geq 0$ satisfying

$$\limsup_{k \to \infty} \frac{|x_{k+1} - \overline{x}|}{|x_k - \overline{x}|^p} = \beta < \infty \tag{7}$$

and the constant β is called the *convergence constant*. The sequence with $p = 1$ and $\beta < 1$ is said to have *linear* convergence implying geometric sequence; the sequence with $p > 1$ or with $p = 1$ and $\beta = 0$ is said to have *superlinear* convergence which means asymptotically faster than linear convergence; in particular, the sequence with $p = 2$ and $\beta < \infty$ is said to have a second-order or a *quadratic* convergence.

The comparison of convergence order of selected algorithms is provided in Table 1. It is simple to show that the Dichotomous Search method provides *linear convergence* with order $p = 1$. It implies that the Dichotomous Search method has a slower convergence rate than the popular derivative methods, most of which exhibit a better convergence rate than the Dichotomous Search method. Also, the Dichotomous Search method is slower than the Secant method as shown in Table 1 See Thisted(1988) for Secant method. This is a sort of drawback but we should recall that the derivative method does *not* converge in many cases whenever a chaotic behavior exists in the objective function (cf. Kim, 1997).

3 Simulation Study and Concluding Remarks

In this simulation study, 30 random variables are generated from $\mathcal{N}(\theta, \sigma^2)$ with $\theta = 1$ and $\sigma^2 = 4$. Using those random variables, the function $d_{F_n}(\theta)$ is drawn (see Figure 4) and the Cramer-von Mises estimate is computed by utilizing the Dichotomous Search method. This procedure is repeated 20 times, and the averages of the 20 estimates and their MSE are computed. From Figure 4, one can notice that $d_{F_n}(\theta)$, $-10 < \theta < 10$, have a feature of noisy unimodality. Figure 5 displays the magnified versions of the same graphs in Figure 4 on the smaller interval $0 \leq \theta \leq 2$. From Figure 5, one is able to watch the fluctuation feature more dramatically. According to Table 2, the averages of the Cramer-von Mises estimates with the initial guess intervals $[-10, 20]$ and $[-5, 10]$ are 0.94 and 0.95 and their MSE's are 0.074 and 0.463, respectively. Apparently, both average values are close to the true parameter $\theta = 1$ and the MSE's are reasonably small, which shows that the Dichotomous Search method finds the true parameter correctly regardless of the initial guess interval. These results suggest us that the Dichotomous Search method is suitable for finding the Cramer-von Mises estimator.

Here we also estimated θ via using the quasi-Newton method, which is a popular version of the Newton-Raphson methods. The results in Table 2 are obtained using the quasi-Newton method in the subroutine `nlmin` provided by the software SPLUS. From table 2, we can see that the estimates heavily depend on the preassigned initial guess point. The averages and the MSE's in Table 2 indicate this phenomenon clearly. Compared to the averages obtained by the Dichotomous Search method, the averages in this case deviate from the true parameter $\theta = 1$ severely. In fact, the average estimates of each 3 cases in quasi-Newton method are very close to the given initial guess points ; this implies that the quasi-Newton method tends to find only a local *optimum* close to the initial guess point. This result is not a surprise in view of the fluctuations observed in Figures 4 and 5. Recall that even the unimodality without fluctuations does not guarantee the correct convergence of the Newton-Raphson method. Meanwhile, note that the Newton-type method relies on the exact slope of *one point* at each iteration, and the Secant method uses the approximated slope based on the *two points* of $(x_1, f(x_1))$ and $(x_2, f(x_2))$. On the other hand, the Dichotomous Search method uses the *four points* consisting of the two test points inside

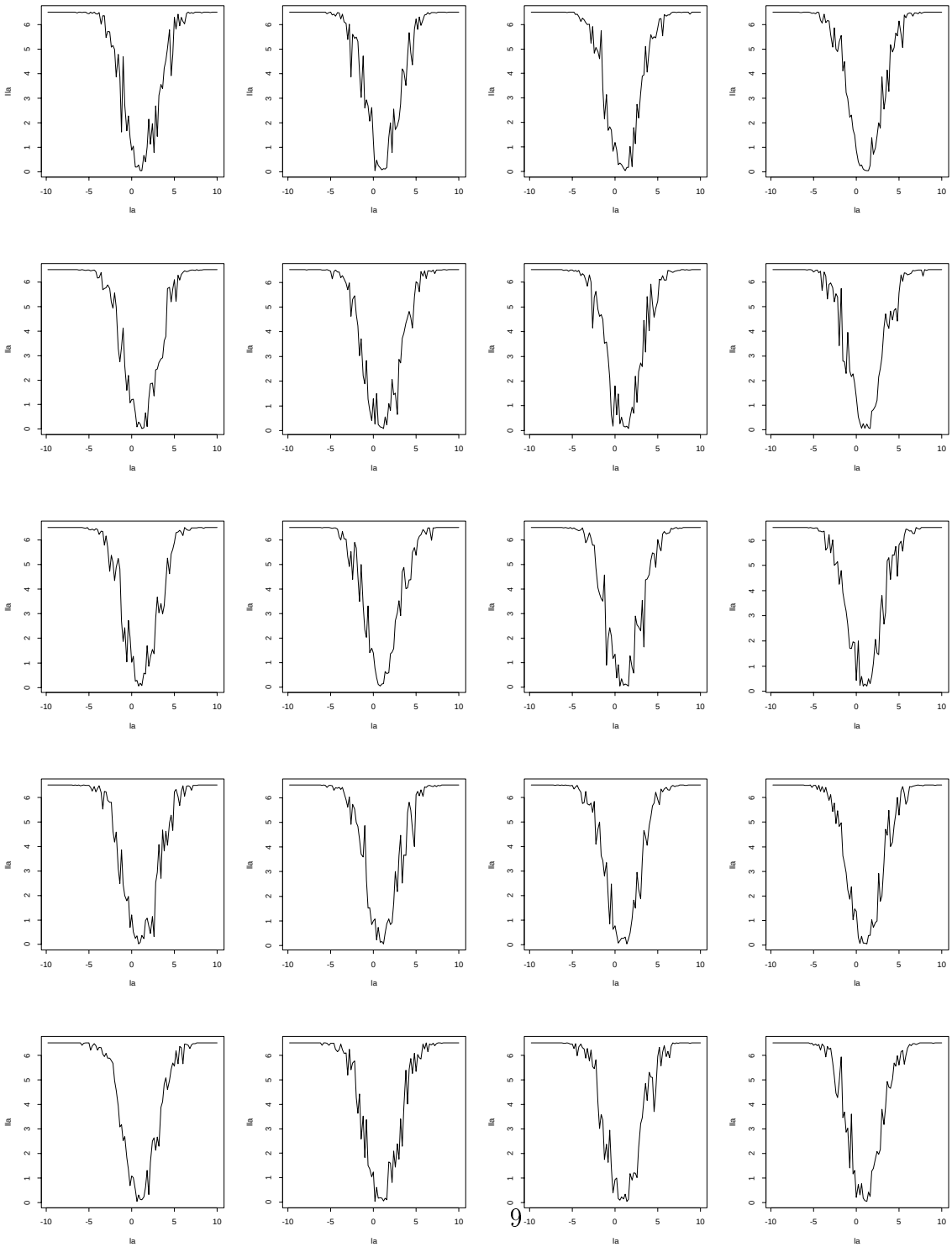

Figure 4: Replication of surfaces $d_{F_n}(\theta)$ for $-10 < \theta < 10$

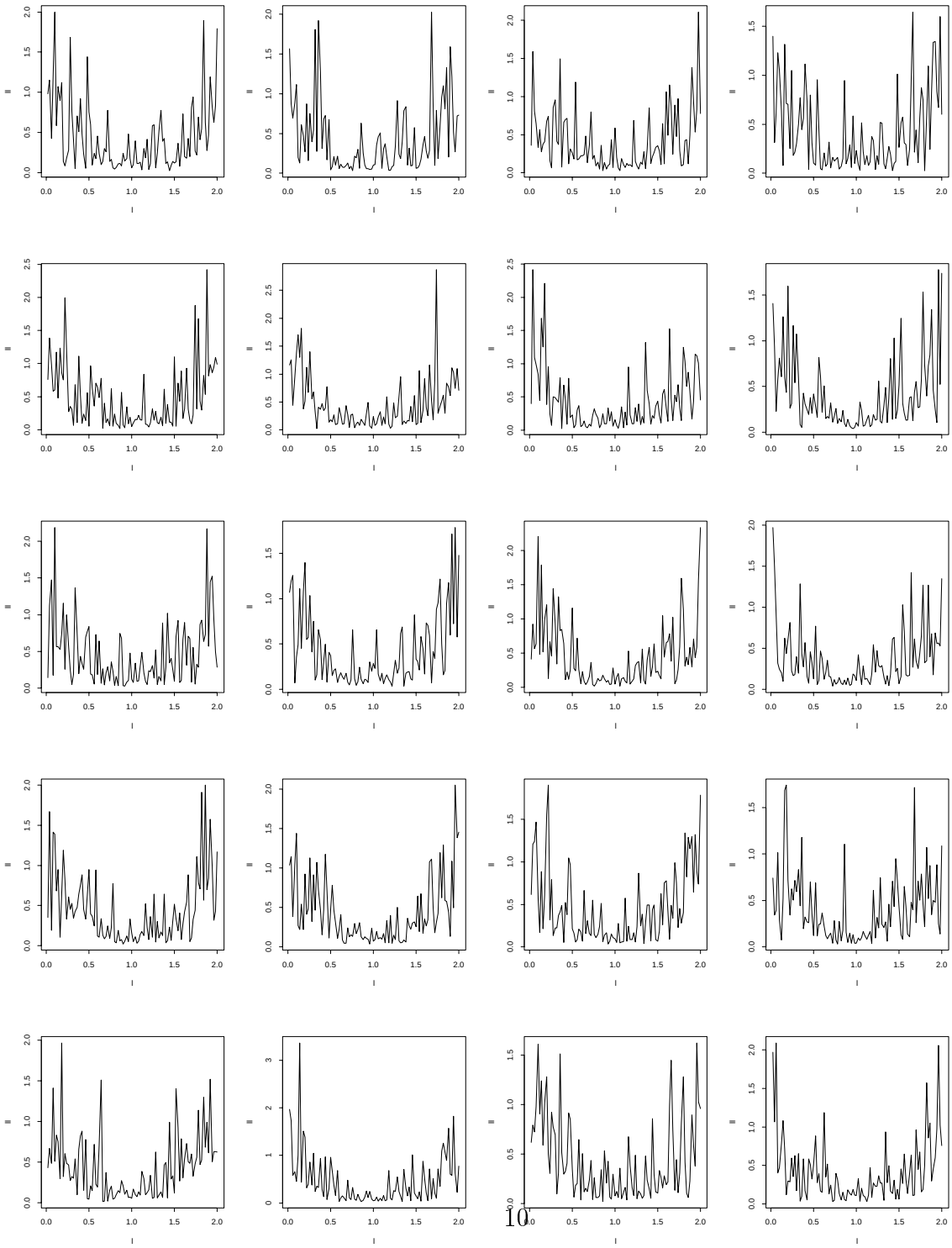

Figure 5: Replication of surfaces $d_{F_n}(\theta)$ for $0 < \theta < 2$

the current interval and its ending points. The set of these four points turns out to be sufficient for reflecting the unimodality. This is the reason why the Dichotomous Search method behaves more stably than the Newton-Raphson method.

Besides the Newton-Raphson and the Secant method, the Bisection method is often considered as an alternative on this problem. The Bisection method is also used for finding a zero of the equation $\partial d_{F_n}(\theta)/\partial\theta = 0$. However, the Bisection is not proper one, either, since the equation has numerous zeros as seen in Figure 4.

Through our simulation results, we could see that the Dichotomous Search method finds the Cramer-von Mises estimator correctly in the normal sample case, while the other derivative methods fail seriously. Therefore we conclude that the Dichotomous Search method is a suitable choice for finding the Cramer-von Mises estimator.

	methods	Initial Guess	estimate(s) of θ	mean	Var	MSE
Normal	Dichotomous Search	$[-10, 20]$	0.82 1.47 0.77 1.07 1.02 1.16 0.98 0.68 0.78 1.05 0.93 0.47 0.76 1.24 1.20 0.50 1.05 1.22 0.97 0.58	0.94	0.07	0.074
		$[-5, 10]$	0.99 0.53 1.09 1.03 -0.86 0.41 1.58 0.35 0.42 0.98 0.63 1.68 1.27 2.28 0.57 1.37 1.90 1.16 0.96 0.67	0.95	0.46	0.463
	`nlmin`	$\theta_0 = -3$	-3.00 -2.99 -3.00 -2.75 -2.99 -3.00 -3.00 -2.75 -2.5 -3.00 -3.00 -2.00 -3.00 -3.06 -2.00 -3.00 -3.12 -2.99 -2.50 -3.00	-2.83	0.11	14.78
		$\theta_0 = 3$	2.00 3.00 2.00 2.00 3.00 2.00 3.12 3.00 2.00 2.99 2.99 2.00 2.00 2.99 3.00 3.00 2.00 2.00 3.00 3.00	2.55	0.25	6.75
		$\theta_0 = 5$	4.99 5.00 5.00 4.99 5.00 5.00 4.99 4.00 4.00 4.00 5.00 4.00 5.00 4.50 4.00 5.00 5.00 4.99 4.00 5.00	4.67	0.21	22.02

Table 2: (Dichotomous Search vs. Quasi-Newton for normal sample)

References

Barzilai, J. & Dempster, M. (1993), "Measuring rates of convergence," *Journal of Optimization Theory and Application.* Vol. 78, No 1.

Bazarra, M., Sherali, H. and Shetty, C. (1993), *Nonlinear Programming: Theory and Algorithms.* Second edition. John Wiley & Sons. New York.

Kennedy, W. and Gentle, J. (1980), *Statistical Computing.* Marcel Dekker. New York and Basel.

Kim, J. (1997), *Iterated Grid Search Algorithm on Unimodal Criteria.* Doctoral dissertation, Department of Statistics, Virginia Tech.

Lindsay, B. (1994), "Efficiency versus robustness: the case for minimum Hellinger distance and related methods," *Annals of Statistics.* **22**:1081-1114

Öztürk, Ö. and Hettmansperger, T. (1997), "Generalized weighted Cramer-von Mises distance estimators," *Biometrika.* **84**:283-294

Parr, W. and Schucany, W. (1980), "Minimum distance and Robust estimator," *Journal of the American Statistical Association.* **75**:616-624

Thisted, R. (1988), *Elements of Statistical Computing -Numerical Computation.* Chapman and Hall. New York.