Estimation of Internet File-Access/Modification Rates from Incomplete Data

Norman Matloff
Department of Computer Science
University of California at Davis
Davis, CA 95616
USA
1-530-752-1953 (voice)
1-530-752-4767 (fax)
matloff@cs.ucdavis.edu

July 16, 2002

Abstract

Consider an Internet file for which last-time-of-access/modification (A/M) data is collected at periodic intervals, but for which full A/M data are not available. Methodology is developed here which enables estimation of the A/M rates, in spite of having only incomplete data of this nature.

Keywords: access rate; estimation; World Wide Web; renewal process; nonparametric density estimation

1 Introduction

One of the major functions of computer networks—ranging from databases on private local area networks (LANs) to World Wide Web sites—is the sharing of information. A question which then arises concerns the number of people who are sharing that information.

Consider for example an Internet site which distributes public-domain software, written by various authors, available World Wide Web. In order to justify the time and funding the authors devote these projects, it would be of interest to know how many users download the software, that is the mean number of downloads per unit time.

Another example arises with Web search engines. The user inputs one or more keywords, say "sailboats." The search engine will then produce a lengthy list of Web sites related to sailboats, ordered according to various criteria. One such criterion (possibly provided by the user as an option) might be frequency of modification; some users may be interested mainly in active sites which are frequently updated. In this case, the we are interested in modification rates instead of access rates.

If we had direct data on A/M transactions, estimation of these and other simple rates would be straightforward [5, Ch.12]. However, such data may either be difficult to collect or else simply unavailable to the public.

For example, collection of some data of this type may require modification of server software, and a server administrator may not have the time to make such a modification, or may not have the source code at all [11, Sec.4.1.5]. For instance, in our Web search engine example above, the Web server software probably does not collect data on file-modification times, even though it does log external accesses.

Even worse, many server administrators would not be willing to divulge A/M data to the public, or would not have the time to do so. And even if they did, the logistics of dealing with thousands of Web site administrators would make our task infeasible.

However, there is related, publicly-accessible information which is available, in the form of time of the last A/M transaction times for a file. On a UNIX system with a file named, say x, the shell command

```
ls -ul x
```

will produce the time of last access to a file, while

```
ls -l x
```

will yield the time of last modification. Typically, as external users without accounts on the server machine, we will not have access to shell commands in the contexts we are considering here, but standard Internet client-side software does have commands which are similar to these. Even basic ftp versions have commands like these (with similar syntax). Similarly, http Web servers can allow users to acquire last-modification times [3].

At first glance, last-A/M time data seems insufficient for estimating A/M rates, as there is no direct relation between the data and the rates. To our knowledge, this problem has not been addressed in the literature. However, we will develop methodology here with which one actually can estimate the A/M rates from last-A/M time data.

An interesting issue which will arise as a byproduct concerns a theoretical question of *conditional* versus *unconditional* statistical inference. We will present quantitative analyses comparing the two, with interpretations of the results.

2 Two Competing Estimators of an A/M Rate

Assume that A/M transactions to the given file occur as a Poisson process with intensity parameter λ. Suppose we sample the process at n intervals of length τ, in each case recording the time of the last file A/M transaction in the interval. By scaling time units, we may assume without loss of generality that $\tau = 1.0$.

Let L_i denote the (unobserved) number of file A/M transactions in the i^{th} interval, so that

$$P(L_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, ... \tag{1}$$

A problem which will become central to the issues addressed in this work is that some L_i may be 0. Let M denote the number of i for which $L_i > 0, i = 1, 2, ..., n$. In addition, let A_1 denote the value of the first nonzero L_i, A_2 the second one, and so on.

Define $T_{i1}, ..., T_{iA_i}$ to be the A/M transaction times to the file within the interval associated with A_i, mod 1.0; that is, these random variables consist of the noninteger portions of the A/M transaction times. We are able to observe only M and for i = 1,...,M the values of $W_i = T_{iA_i}$.

Estimation based on the W_i will be conditional on M, while estimation based on M itself will be unconditional.

2.1 Estimation of λ Via M

We begin with the simpler estimator, based only on M. It would seem more natural to estimate λ from the W_i, but if M is small, there will be too few W_i to get an accurate estimate from them, so we turn to using M itself.

Define

$$p = P(L_i > 0) = 1 - e^{-\lambda}, \tag{2}$$

so that

$$\lambda = -log(1 - p) \tag{3}$$

Maximum likelihood estimators (MLEs) are generally optimal, i.e. have minimum asymptotic variance [2]. The MLE of p based on M is well known to be $\check{p} = M/n$, so from (3) the MLE of λ based on M is

$$\check{\lambda} = -log(1 - \check{p}) = -log(1 - M/n) \tag{4}$$

2.2 Estimation of λ Via the W_i

Let $f_k(w)$ and $F_k(w)$ denote the p.d.f. and c.d.f., respectively, of W_i, conditioned on $A_i = k$. Then for k > 0,

$$\begin{aligned} F_k(w) &= P(W_i \leq w \,|\, A_i = k) & (5) \\ &= P(T_{i1} \leq w, ..., T_{iA_i} \leq w \,|\, A_i = k) & (6) \\ &= P(Y_{i1} \leq w, ..., Y_{iA_i} \leq w \,|\, A_i = k), \end{aligned}$$

where $Y_{i1}, ..., Y_{iA_i}$ are the *unordered* versions of the T_{ij} (the former variables can be thought of as a random permutation of the latter ones). The Y_{ij} are i.i.d. U(0,1) [8, Theorem 6.1] , so we have that $F_k(w) = w^k$ and $f_k(w) = kw^{k-1}$ for w in (0,1) and k > 0.

Now let $G(w) = P(W_i \leq w \,|\, A_i > 0)$, and set $g = G'$. Then

$$\begin{aligned} G(w) &= \frac{1}{P(A_i > 0)} \sum_{k=1}^{\infty} P(W_i \leq w \, and \, A_i = k) & (7) \\ &= \frac{1}{1 - e^{-\lambda}} \sum_{k=1}^{\infty} F_k(w) \cdot \frac{e^{-\lambda}\lambda^k}{k!} & (8) \\ &= \frac{e^{-\lambda}}{(1 - e^{-\lambda})e^{-\lambda w}} \sum_{k=1}^{\infty} \frac{e^{-\lambda w}(\lambda w)^k}{k!} & (9) \\ &= \frac{e^{\lambda w} - 1}{e^{\lambda} - 1}. & (10) \end{aligned}$$

Thus

$$g(w) = \frac{\lambda e^{\lambda w}}{e^{\lambda} - 1}. \tag{11}$$

We wish to find the MLE of λ based on the W_i, conditional on $M = m > 0$. The conditional likelihood function of $W_i = w_1, ...W_m = w_m$ is

$$L(w_1, ..., w_m) = g(w_1)g(w_2)...g(w_m) = \frac{\lambda^m}{(e^{\lambda} - 1)^m} \cdot e^{\lambda \sum_{i=1}^m w_i}. \tag{12}$$

Maximizing this yields that the conditional MLE, $\hat{\lambda}$, must satisfy the equation

$$r(\hat{\lambda}) = \bar{W}, \tag{13}$$

where $\bar{W} = (W_1 + ... + W_m)/m$ and

$$r(t) = \frac{1}{1 - e^{-t}} - \frac{1}{t}.$$

2.2.1 Solution of the Likelihood Equation

Equation (13) has no closed-form solution. Thus iterative numerical methods must be used.

However, we can at least establish that the root exists and is unique, as follows. It is easily verified that r(t) above is an increasing function of t, and varies continuously between 0.5 and 1. However, $\bar{W}$ can take on any value between 0 and 1. Thus the solution of the equation exists and is unique if $\bar{W} > 0.5$, and otherwise the MLE is 0.0.

Later, in determining the statistical accuracy of our estimator, we will again need to deal with the non-closed form of the MLE here, but will present a way to circumvent this problem.

3 Statistical Inference on λ

3.1 The “Delta Method”

The “delta method” [6] says, roughly, that a function of an asymptotically Gaussian-distributed sequence of random variables is itself an asymptotically Gaussian-distributed sequence. In our context here, the relevant formulation is as follows.

Suppose that a random variable U_k, calculated from k observations from a distribution family parameterized by θ, is used to estimate θ and is distributed approximately $N(\theta, \sigma^2(\theta)/k)$. Then if q is a differentiable function, the quantity $q(U_k)$ is approximately distributed as

$$N[q(\theta), q'(\theta)^2 \sigma^2(\theta)/k]$$

The quantity

$$q'(\theta)^2 \sigma^2(\theta)/k$$

is then the *asymptotic variance* (AVar) of $q(U_k)$.

The approximate square root of this quantity,

$$SE(q(U_k)) = |q'(U_k)|\sigma(U_k)/\sqrt{k}$$

is known as the *standard error* of U_k. The standard error can be used for statistical inference purposes. For instance, an approximate 95% confidence interval for $q(\theta)$ based on $q(U_k)$ is

$$q(U_k) \pm 1.96q'(U_k)\sigma(U_k)/\sqrt{k}$$

3.2 Inference on $\check{\lambda}$

In the notation above, take k to be n, take θ to be p, and take U_k to be $\check{p} = M/n$. Then from Equation (2), take q(t) to be -log(1-t). Since the Central Limit Theorem shows that $\check{p}$ is approximately distributed as N(p, p(1-p)/n), then by the delta method $\check{\lambda}$ has an approximately normal distribution which has mean λ and variance

$$\frac{1}{n} \cdot \frac{p}{1-p}$$

Inference can then be done by replacing p in this expression by $\check{p}$. In other words,

$$SE(\check{\lambda}) = \sqrt{\frac{1}{n} \cdot \frac{\check{p}}{1-\check{p}}} \tag{14}$$

3.3 Inference on $\hat{\lambda}$

Now, let us see what can be done in the case of $\hat{\lambda}$. Note first that since $\hat{\lambda}$ is a function of $\bar{W}$, we would ordinarily take q() to be this function. In other words, q() would be the functional inversion of Equation (13). However, as noted earlier, we do not have this latter function in closed form.

We could find the approximate value of that function (actually, its derivative) during our iterative procedure to find $\hat{\lambda}$, but there is an easier approach. Instead, we use the delta method on the function r() in Equation (13), "pretending" that we do not know the asymptotic variance of $\bar{W}$ but do know that of $\hat{\lambda}$. Since we actually do know the asymptotic variance of $\bar{W}$, we can solve for what we do want. Here are the details.

The quantity $\bar{W}$ has mean

$$E\bar{W} = EW = \frac{1}{1-e^{-\lambda}} - \frac{1}{\lambda} = r(\lambda), \tag{15}$$

and variance

$$\begin{aligned} Var(\bar{W}) &= \frac{1}{m} \cdot Var(W) \qquad (16) \\ &= \frac{1}{m}\left(\frac{2}{\lambda} - \frac{1}{1-e^{-\lambda}} - (EW)^2\right) \qquad (17) \\ &= \frac{1}{m}\left(\frac{1}{\lambda} - r(\lambda)[1+r(\lambda)]\right). \qquad (18) \end{aligned}$$

(By the way, note that by comparing Equations (13) and (15), we see that $\hat{\lambda}$ is not only the MLE of λ, but also the Method of Moments Estimator of that quantity .)

Due to the Central Limit Theorem, $\bar{W}$ has an approximately normal distribution with mean and variance as in (15) and (16). So, now considering $\bar{W}$ to be a function of $\hat{\lambda}$ in Equation (13), rather than vice versa, and thus "applying the delta method in reverse," we have that

$$Var(\bar{W} = AVar(\bar{W}) = r'(\lambda)^2 AVar(\hat{\lambda}),$$

so that the standard error of $\hat{\lambda}$ is

$$SE(\hat{\lambda}) = \frac{1}{r'(\hat{\lambda})}\sqrt{Var(\bar{W})} = \frac{1}{\sqrt{m} \cdot r'(\hat{\lambda})}\sqrt{\frac{1}{\hat{\lambda}} - r(\hat{\lambda})[1+r(\hat{\lambda})]}. \tag{19}$$

One can choose between $\check{\lambda}$ and $\hat{\lambda}$ by comparing the standard errors in Equations (14) and (19).

4 Evaluation

We performed a simulation study, calculating the mean squared errors (MSE) for $\hat{\lambda}$ and $\check{\lambda}$:

$$MSE(\hat{\lambda}) = E[(\hat{\lambda} - \lambda)^2],$$

$$MSE(\check{\lambda}) = E[(\check{\lambda} - \lambda)^2].$$

The results are shown in Figures 1 and 2, in the form of the square root of MSE, normalized by λ; in other words, what is plotted is

$$\frac{\sqrt{MSE}}{\lambda}$$

Intuitively, $\hat{\lambda}$ should typically be a superior estimator to $\check{\lambda}$, since the former is based on "richer" information than the latter (i.e. last-A/M times rather than counts of nonzero intervals). However, such intuition must be tempered by the fact that if λ is small, the quantity M might also be very small—in which case $\hat{\lambda}$ will be based on such a small sample that $\check{\lambda}$ may actually be the superior estimator.

This intuition is confirmed in Figures 1 (n = 50) and 2 (n = 500). For a sample size of 50, $\check{\lambda}$ performs better than $\hat{\lambda}$ for $\lambda < 3.7$, while for n = 500 the change point comes earlier, at approximately $\lambda = 3.0$. In other words, a sample size of n = 500 is large enough so that we will get a fairly large value of M even if λ is small.

Let d represent the ratio of the two MSEs:

$$d = \frac{MSE(\check{\lambda})}{MSE(\hat{\lambda})}$$

The quantity d is plotted in Figure 3, displaying the comparison between the two estimators more explicitly.

As a basis of comparison, we also plot the "ideal" MSE, that is the MSE we would attain if the full file A/M-time data were available to us, rather than merely the last-A/M times and M. This is defined as follows. Using the L_i from Section 2, let

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i$$

Then $\bar{K}$ has mean λ and variance λ/n, so $MSE(\bar{K}) = \lambda/n$.

Figures 1 and 2 show that $\bar{K}$ performs better than both $\check{\lambda}$ and $\hat{\lambda}$, as expected since $\bar{K}$ is based on complete information. However, if such information is not available, we see here that $\check{\lambda}$ and $\hat{\lambda}$ do a good job under the circumstances.

Indeed, the restriction to partial information is not really a limitation in the context of estimating file A/M rates, for the simple reason that in this context it is easy to generate very large samples. The collection of data can be typically automated via programs which can just as easily be run for months as for hours. One can run the programs long enough to achieve reasonable estimation accuracy.

Moreover, in the application contexts we have outlined earlier, there is typically no need for very fine levels of accuracy. All that is needed is methodology to derive some reasonable estimate of A/M rates from the last-access-time data, which $\check{\lambda}$ and $\hat{\lambda}$ now provide for us.

5 The Poisson Assumption and Alternatives

Up to this point, we have been assuming that the A/M transactions occur as a Poisson process. Let us now give this assumption closer examination.

5.1 Formulation as a Renewal Process

Let S_i denote the time between the $(i-1)^{st}$ and i^{th} A/M transactions. Under the Poisson assumption, these inter-transaction times are independent and identically distributed, with the common distribution being exponential. We will now continue to assume that the S_i are i.i.d., but not necessarily with an exponential distribution.

Let N(t) denote the total number of A/M transactions which have occurred on or before time t, i.e.

$$N(t) = max\{i : S_1 + ... + S_i \leq t\}$$

N(t) is known as a *renewal process* [8, Sec.6.5].

From here on, we will assume that the sampling interval width τ is large enough so that $P(L_i = 0)$ is negligible. It is important to keep in mind that we can arrange things this way, because we always have the option of increasing τ, even on an after-the-fact basis subsequent to collecting our data. (If we do this, we will pay a price for it, in that we will discard some of our data, but at least it will allow us to handle the non-Poisson case, as seen below.) After doing so, though, we still assume that we rescale time units so that $\tau = 1.0$, as before. We will now have that (except for negligible probability) $A_i = L_i$ for all i, and that M = n.

The random variable W_i is the i^{th} *forward recurrence time*, meaning the time to the next renewal, measured from the i^{th} observation epoch i-1. From renewal theory, the (asymptotic) density function of this quantity is

$$b(t) = \frac{1 - C(t)}{E(S)} \tag{20}$$

where C is the cumulative distribution function of S.

5.2 Assessing the Exponential Assumption

Since we do observe the values of the variables W_i, we can apply standard statistical goodness-of-fit assessment procedures [8, Sec.10.3.4]. It should be noted, though, that in our context there is typically a large amount of data, which means that if a goodness-of-fit hypothesis test is used, even slight departures from the exponential model will (misleadingly) result in rejection of they hypothesis at standard significance levels. Thus care should be used [4, Ch.7].

5.3 Alternative Approaches for Use in the Nonexponential Case

From renewal theory we know that the exponential assumption is necessary and sufficient for our renewal process to have *independent increments,* i.e. the property that events in disjoint intervals of time are independent. (This is related to the fact that the exponential density is the only one to be "memoryless.") Thus even if we were to find a suitable nonexponential parametric model, say a Γ distribution, we would have a problem with statistical inference procedures; that methodology assumes that the W_i are independent, which would not be true.

As an alternative to assuming that the density function c(t) for S is of exponential form, we present here a novel application to renewal theory of tools for nonparametric density estimation. Here Equation (20) will play the central role.

Nonparametric density estimation is a refinement of the usual histogram methods taught in elementary statistics courses. It is used primarily as a tool for exploratory data analysis, with the aim being to answer questions about the overall shape of the density function, such as: Is the density unimodal or multimodal? Where does the bulk of the distribution lie? By contrast, it is rare that nonparametric density estimation has been used to estimate a density at only one point, which is what we will do here.

From Equation (20), we see that

$$b(0) = \frac{1}{E(S)}$$

The right-hand side here is the reciprocal of the mean inter-access time. By standard renewal theory that is equal to the asymptotic mean number of A/M accesses per unit time—exactly what we need.

So, our job now is to estimate b(0) from our data W_i, without assuming a parametric family such as the exponential.

5.3.1 Kernel-Based Estimation

The classic kernel nonparametric density estimator, applied here to the function b(t), is

$$\widehat{b}(t) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{t - W_i}{h})$$

where h is a *smoothing parameter* and the *kernel* K is chosen to be a mean-0 density function in its own right. The choice of K is up to the user, provided K satisfies certain regularity conditions [7].

The smoothing parameter is similar to the bin width in histograms. A large body of mathematical theory exists on this point; here we will assume that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Here we will choose

$$K(t) = \begin{cases} 0.5, & -1 < t < 1 \\ 0, & \text{otherwise} \end{cases}$$

With this choice of K we would have

$$\widehat{b}(0) = \frac{0.5\#(0,h)}{nh} \tag{21}$$

where #(u,v) denotes the count of the number of W_i in the interval (u,v).

However, kernel estimators are subject to serious bias problems near the boundary of a density's nonzero region. For a kernel estimator $\widehat{f}$of a density f based on a kernel K which has the value 0 outside of (-1,1),

$$E[\widehat{f}(0)] = \int_{-1}^{0} K(u)du \text{ f}(0) + O(h)$$

[7].

So, in our case here,

$$E[\widehat{b}(0)] = 0.5b(0)$$

Thus we must redefine (21) to be

$$\widehat{b}(0) = \frac{\#(0,h)}{nh}$$

It is up to the user to choose the value of the smoothing parameter h. Though some methods have been proposed for choosing h, no fully practical method has yet been developed, so nonparametric density estimation is used typically has a data-exploratory tool rather than a means of formal statistical inference.

5.3.2 Use of Monotonicity

Equation (20) shows that b(t) is a nonincreasing function. This allows us to make use of nonparametric maximum likelihood estimators for unimodal densities [9].

Let B(t) denote the cumulative distribution function corresponding to b(t), and consider its estimator, the empirical distribution function of the W_i, $\widetilde{B}(t) = \frac{1}{n}\#\ (0, t)$. $\widetilde{B}$ will then be a step function with jumps of size 1/n, the i^{th} of which occurs at $W_{(i)}$, defined to be the i^{th}-smallest among $W_1, ..., W_n$. Then one defines a density estimator $\widetilde{b}$ to consist of the slopes of the upper envelope of $\widetilde{B}$, so that for t between $W_{(i-1)}$ and $W_{(i)}$

$$\widetilde{b}(t) = \min_{s \leq i-1} \max_{t \geq i} \frac{\widetilde{B}(W_{(t)}) - \widetilde{B}(W_{(s)})}{W_{(t)} - W_{(s)}}$$

where $W_{(0)}$ is taken to be 0 [1].

Ordinarily, an important advantage of this kind of estimator is that there is no smoothing parameter for the user to choose. This would be true in our case if we were estimating b(t) for some t $>$ 0, but unfortunately there is a problem at 0: Just as was the case with kernel-based density estimators at the boundary, the nonparametric maximum likelihood estimator also is *nonconsistent* at 0, meaning that the estimator $\widehat{b}(0)$ does not converge to b(0) as $n \to \infty$ [10].

There is no fully satisfactory method for dealing with this, other than to use the estimator $\widetilde{b}(\epsilon)$ instead of $\widetilde{b}(0)$, for some positive ϵ near 0. Thus the user still must make a choice, in this case the value of ϵ.

References

[1] Barlow, R.; Bartholomew, D.; Bremner, J.; and H. Brunk (1972). *Statistical inference under order restrictions.* New York: Wiley.

[2] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical statistics.* London: Chapman and Hall.

[3] Hethmon, P. (1997). *Illustrated Guide to HTTP.* Manning.

[4] Matloff, N. (1988). *Probability modeling and computer simulation, applied to engineering and computer science.* Boston: PWS-Kent.

[5] Menasce, D. and Almeida, V. (1998). *Capacity planning for Web performance: metrics, models, and methods.* Englewood Cliffs, New Jersey: Prentice-Hall.

[6] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

[7] Simonoff, J. (1996). *Smoothing methods in statistics.* New York: Springer-Verlag.

[8] Trivedi, K. (1982). *Probability and statistics with reliability, queuing and computer science applications.* Englewood Cliffs, New Jersey: Prentice-Hall.

[9] Van der Vaart, A. (1998). *Asymptotic statistics*, New York: Cambridge University Press.

[10] Wegman, E. (1975). Maximum likelihood estimation of a probability density, *Sankhya (A)*, 37, 211-224.

[11] Yeager, N. and McGrath, R. (1996). *Web server technology: the advanced guide for World Wide Web information providers.* Belmont, California: Morgan Kaufmann.

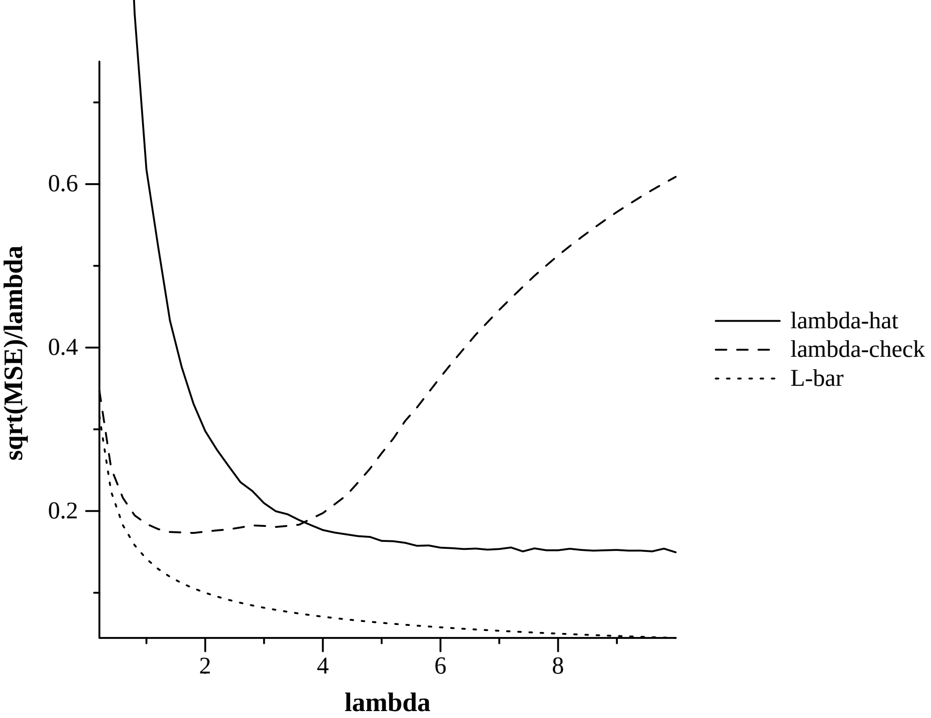
sqrt(MSE)/lambda
0.6
0.4
0.2
2
4
6
8
lambda
lambda-hat
lambda-check
L-bar

Figure 1 (n = 50)

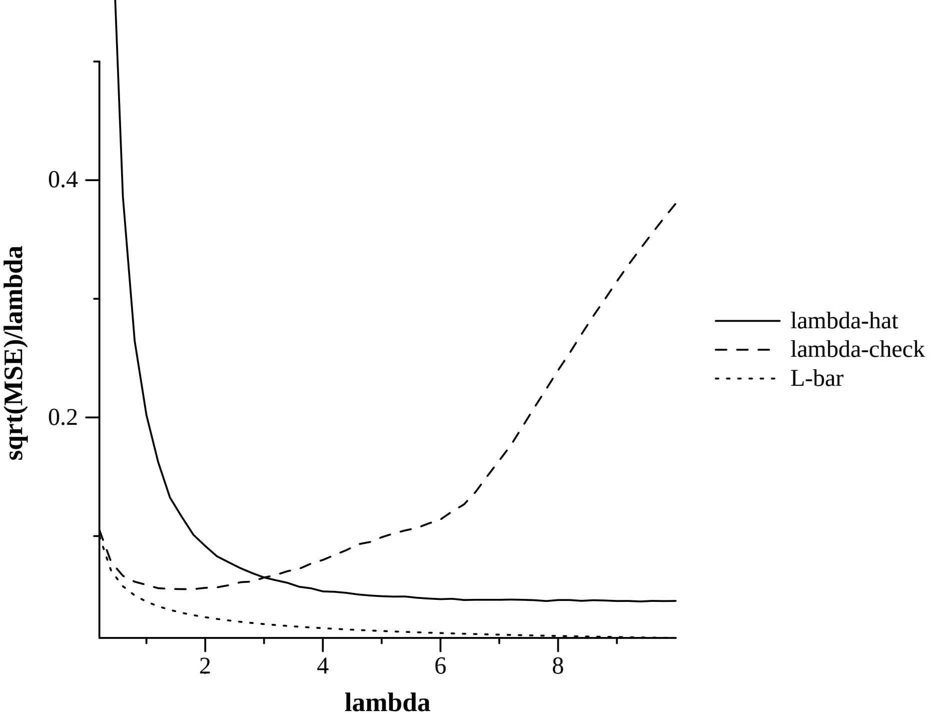

Figure 2 (n = 500)

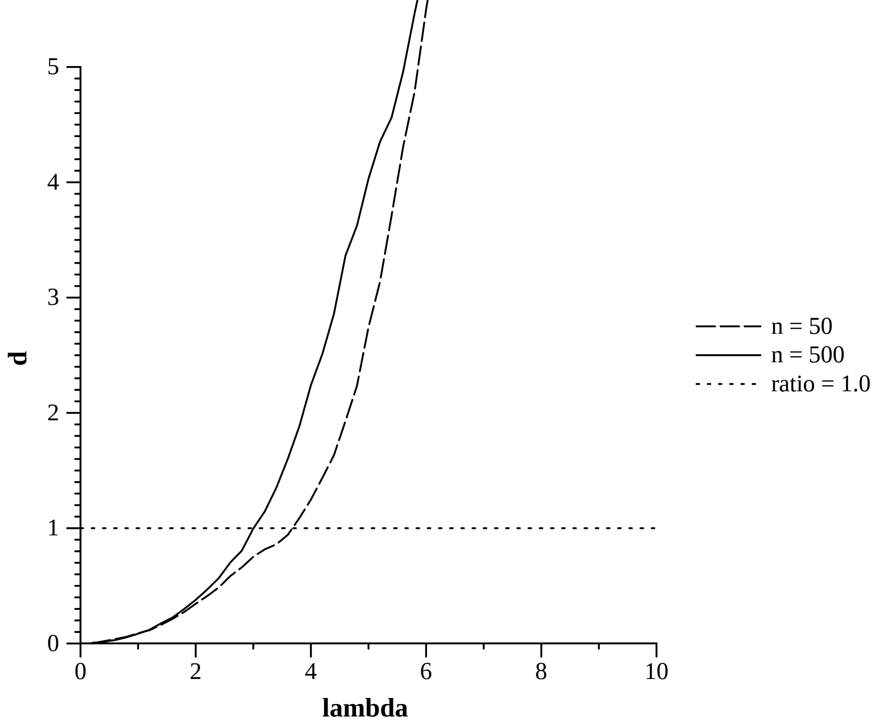

Figure 3