ORNL/TM-13620

ANALYSIS AND MONITORING DESIGN FOR NETWORKS

V. Fedorov, D. Flanagan, T. Rowan
Computer Science and Mathematics Division

S. Batsell
Computing, Information and Networking Division

Date Published: June 1998

Prepared by the
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831
managed by
LOCKHEED MARTIN ENERGY RESEARCH CORP.
for the
U. S. DEPARTMENT OF ENERGY
under Contract No. DE-AC05-96OR22464

Preface

The primary aim of this text is to attract the attention of researchers and practitioners working with computer networks to possible gain in precision, speed, and economy of their experiments by employing advanced techniques of optimal experimental design. Unlike in many other experimental areas where the cost of measurements is dominant, in computer networks the emphasis is to maximization of useful information per given amount of either collected or stored information.

The main part of the results included in this text have been done in the framework of the Oak Ridge National Laboratory Director Research and Development Fund Project "Network Performance Understanding". The support of Bob Aiken and Fred Howes of the U.S. Department of Energy's Mathematical, Information, and Computational Sciences Division is gratefully acknowledged.

We thank Tom Dunigan and Max Morris for reading the text and making valuable comments and suggestions. We would like to thank Rachal McIntire for her help and expertise in LaTex.

Contents

1 INTRODUCTION **1**

1.1 MOTIVATION . . . 1

1.2 MEASURING NETWORK PERFORMANCES . . . 2

1.3 COMMENTS ON STATISTICAL METHODS OF EXPERIMENTAL DESIGN . . . 3

1.4 OUTLINE OF THE PAPER . . . 4

2 REGRESSION TYPE MODELS IN NETWORK MEASUREMENTS **7**

2.1 STANDARD LINEAR MODELS . . . 7

2.1.1 Example: Simple Network . . . 7

2.1.2 Linear Regression Models . . . 8

2.1.3 The Best Linear Unbiased Estimator and the Least Squares Method . . . 9

2.1.4 Properties of Information Matrices . . . 11

2.1.5 Presence of Prior Information . . . 12

2.2 CORRELATED OBSERVATIONS . . . 13

2.2.1 Example . . . 13

2.3 VARIANCE DEPENDING UPON UNKNOWN PARAMETERS . . . 14

2.3.1 Iterated Estimators Based on the Best Linear Unbiased Estimators . . . 14

2.3.2 Iterated and Maximum Likelihood Estimators . . . 16

2.3.3 Example: Estimation of Delay/Travel Time . . . 16

2.3.4 Estimating Source-Destination Network Traffic Intensities (Network Tomography) . . . 21

2.4 NONLINEAR MODELS . . . 23

3 REGRESSION MODELS IN OPTIMAL MONITORING DESIGN **25**

3.1 OPTIMALITY CRITERIA . . . 25

3.1.1 Optimality of Measurements on a Network . . . 25

3.1.2 Most Popular Criteria . . . 27

3.2 PROPERTIES OF OPTIMAL DESIGNS . . . 30

3.3 NUMERICAL METHODS . . . 32

3.3.1 The First Order Algorithms . . . 32

3.3.2 Practical Algorithms and Pilot Software . . . 33

3.4 EXAMPLES . . . 36

3.4.1 Optimal Designs for Simple Networks . . . 36

3.4.2 Multihost Experiments vs One-host Experiments 36
3.4.3 Measurement Errors Depending on the Route Length 37
3.5 ESNET EXAMPLE I 45
3.5.1 Model and Main Assumptions 45
3.5.2 Construction of Support Sets 47
3.5.3 D-optimal Designs for ESnet under Various Assumptions 50
3.6 NETWORK CHALLENGES: PROBLEMS TO EXPLORE 61
3.6.1 Multiresponse Models 61
3.6.2 Selection of the Response Component to Measure 63
3.6.3 Different Convergence Rates of Parameter Estimators 63
3.6.4 Other Types of the Information Matrix Normalization 64
3.6.5 Heavy Tailed Distributions 64

4 NONPARAMETRIC APPROACH IN OPTIMAL MONITORING 67
4.1 BEST LINEAR PREDICTOR 67
4.1.1 Introduction 67
4.1.2 Best Linear Predictor 69
4.2 DESIGNS WITH CONTINUOUS WEIGHTS 70
4.2.1 Properties of Optimal Designs 70
4.3 FIRST ORDER ALGORITHMS 72
4.4 ESNET EXAMPLE II 74
4.4.1 Covariance Matrix Estimation and Optimal Design 74
4.4.2 Modeling the Covaraince Matrix 79
4.5 SIMPLE HEURISTIC ALGORITHM 81
4.5.1 Short Survey of the Older Results 81
4.5.2 Approximate Duality of Two Approaches 82

REFERENCES 89

List of Figures

2.1 Network graph with 4 nodes and 5 edges. 8

2.2 Histograms for various number of edges included in a route. 18

2.3 The histogram of travel time for two neighbor nodes. 19

2.4 Normalized histograms for the travel time for two neighbor nodes. 20

2.5 The normalized histogram for the travel time and the simulated histogram (1000 trials) for the standard normal distribution. 20

3.1 Network graph with 8 nodes and 12 edges. 41

3.2 Model of ESnet backbone used for computational experiments. 46

List of Tables

2.1 Routes for the network problem with 4 nodes and 5 edges 14

3.1 Various optimality criteria $\Psi(\xi)$, their sensitivity functions $\phi(x, \xi)$ and majorization constants C. 28

3.2 Experimental designs for the single-host case of the network example with 4 nodes and 5 edges. 37

3.3 Experimental designs for the multi-host case of the network (4,5) example. 38

3.4 Experimental designs with $\sigma^2(x) = x^T x$ for single-host case network (4,5). 39

3.5 Experimental designs with $\sigma^2(x) = x^T x$ for multi-host case of network (4,5). 40

3.6 Routes for the network (8,12) with two hosts. 42

3.7 Experimental designs for the two-host case of network (8,12), $\sigma^2(x) \equiv 1$. 43

3.8 Experimental designs for the two-host case of network (8,12), $\sigma^2(x) = x^T x$. 44

3.9 List of ESnet edges used in computational experiments. 48

3.10 List of ESnet nodes used in computational experiments. 49

3.11 Comparison of results for the different final step lengths. 51

3.12 (Part 1 of 2) The computed design for $\alpha = 0.01$. 52

3.12 (Part 2 of 2) The computed design for $\alpha = 0.01$. 53

3.13 Comparison of parameter estimators and variances for different monitoring schemes. 54

3.14 Data for selecting node to partner with ORNL. 55

3.15 The best (CIT) and worst (ATM) nodes to partner with ORNL. 56

3.16 Optimal design for all hosts, $\sigma^2 = x^T x$, $\alpha = 0.001$. 57

3.17 Optimal design for one host, $\sigma^2 = x^T x$. 58

3.18 (Part 1 of 2) Optimal design for two hosts, $\sigma^2 = x^T x$. 59

3.18 (Part 2 of 2) Optimal design for two hosts, $\sigma^2 = x^T x$. 60

4.1 Site identifiers 75

4.2 Various designs to monitor ESNet sites ($N = 10, \sigma = 8.0ms$) 76

4.3 Dependence of the optimal design structure on σ^2/N. 77
4.4 Dependence of the optimal design structure on σ^2/N, k is modeled. 80

Chapter 1

INTRODUCTION

1.1 MOTIVATION

The idea of applying experimental design methodologies to develop monitoring systems for computer networks is relatively novel even though it was applied in other areas such as meteorology, seismology, and transportation [see Cressie (1991), Cheng (1991), Fedorov (1996)]. One objective of a monitoring system should always be to collect as little data as necessary to be able to monitor specific parameters of the system with respect to assigned targets and objectives. This implies a purposeful monitoring where each piece of data has a reason to be collected and stored for future use. When a computer network system as large and complex as the Internet is the monitoring subject, providing an optimal and parsimonious observing system becomes even more important.

A standardized and widely accepted set of metrics will help to (1) determine performance of routine maintenance and problem troubleshooting, (2) predict performance trends, and (3) develop simulations and the corresponding mathematical models that are necessary for operating system understanding and development planning. The metrics may characterize the operating system status, routes, capacity, resources, packet losses, response time, or intrusions, to name a few examples [cf. Claffy (1994), Paxson (1997), Cottrell et al. (1997), and Cottrell and Mathews (1998) for computer network metric examples and an extensive bibliography]. We do not discuss extensively an important problem of optimal selection of metrics (indicators, explanatory/response variables), which must be included in the monitoring. In practice, the choice is based on the trade-off between what is needed, what can be measured (given expenses), and what is allowed to be measured by various legal agreements.

Many data collection decisions must be made by the developers of a monitoring system. These decisions include but are not limited to the following:

1. The type of data collection hardware and software instruments to be used.
2. How to minimize interruption of regular network activities during data collection.
3. Quantification of the objectives and the formulation of optimality criteria.
4. The placement of data collection hardware and software devices.
5. The amount of data to be collected in a given time period, how large a subset of the available data to collect during the period, the length of the period, and the frequency of data collection.

6. The determination of the data to be collected (for instance, selection of response and explanatory variables).

7. Which data will be retained and how long (i.e., data storage and retention issues).

8. The cost analysis of experiments.

We would like to emphasize that this research is focused on the monitoring of networks through measurements that minimally disturb the networks themselves. We do not consider the very interesting and challenging "active" experiments in which a network's characteristics or regimes may be varied to gather information about its behavior at some specific conditions, hypotheses, assumptions, or test models. For instance, the capacity of some buffers may be intentionally reduced or some links (communications channels) may be blocked.

Mathematical statistics, and, in particular, optimal experimental design methods, may be used to address the majority of problems generated by $3-7$. In this study, we focus our efforts on topics $3-5$. Optimal design theory methodologies start with a candidate set of variables that have a potentially important impact on the response variables. In the most obvious cases, there are variables that are included in a model that must be fitted and used for prediction or simulation. The entire feasible set from which the levels of designated variables can be selected is called the design or operability region. The main task of experimental design is to define combinations of those levels that will provide the most cost effective information in the sense of a given criterion of optimality (for instance, variance of prediction or variance of the estimator of some specific parameter).

A few methods and techniques of standard experimental design theory can be used after adaptation similar to the approach that was applied for the monitoring of transportation networks [see Cheng (1991) who considered the optimal sampling problem for federal highway traffic data collection]. Computer network traffic, however, exhibits unusual patterns of stochastic behavior, including long range (in time and space) correlation, probability distributions with heavy tails, mixture of distributions, bursts, nonstationarity, etc. This fact necessitates more serious changes in the existing methods of experimental design theory and development of methods that address the needs of mathematical models used for the cases above.

Computer software for generating response surface design of experiments is currently available from various vendors such as the Statistical Analysis System Inc., SPSS, "Trial Run," etc., [see, for instance, SAS/QS (1995), SPSS (1997), and Wheeler(1994)]. While this software handles most standard linear model designs, it does not cover the optimal design theory methods needed for our research. In particular, even the most extensive package, SAS, has no features for correlated observations, does not allow the computation of optimal designs when prior information is available, and uses algorithms that do not guarantee global optimality. Thus, computer programs that work for the considered models are needed, and the development of their pilot versions is a part of this study.

1.2 MEASURING NETWORK PERFORMANCES

With the extensive growth of the Internet and its various components, only regular and well organized measurements and surveys enable one to understand ongoing tendencies and processes. Knowledge and correct diagnosis of short-term processes make it possible for various conditions to occur, such as connection establishment, retransmission, fragmentation, optimal routing, etc.

Understanding long term tendencies may be useful in network development strategies or technical improvements needed to satisfy the consumer needs.

There exists a rather intense flux of publications covering various aspects of measurements and some statistical techniques related to those measurements. Very comprehensive bibliographies may be found in Claffy (1994), Paxson (1997), Quarterman (1990), Willinger et al. (1995a,b), Willinger et al. (1996). It has been noted that the probability distributions of random variable and random processes observed on large networks have some unusual properties. Self-similarity, heavy tails, and long-memory processes are frequently discussed subjects in publications related to measurements and statistical analysis of large networks. Examples are Floyd (1996), Frost and Melamed (1994), Paxson (1995, 1996), Paxson and Floyd (1995), Willinger et al. (1995a), Willinger et al. (1995b), and Beran (1994), Chap. 1.

More practical aspects and, in particular, various "metrics" to measure are surveyed by Cottrell et al. (1997). See also Claffy (1994) and Paxson (1997). For quick, real-time evaluation of network state(s), simple graphical presentation and visualization of some basic statistics are essential components (see, for instance, Cottrell and Mathews (1998), Batsell et al. (1997)). There are a few Web sites where the reader can find the corresponding information; see *http://www.slac.stanford.edu/xorg/nmtf.html* and *http://www.epm.ornl.gov/~sgb/network.html* for references.

In all examples included in this study, we use software measurements, which are based on two popular programs; "*ping* (Packet Internet Groper)" and "*traceroute*". The detailed description of these programs may be found, for instance, in Paxson (1997) and Stevens (1994), Chaps. 7 and 8. Various scripts for the efficient use of the mentioned "tools" were written by T. Dunigan of Oak Ridge National Laboratory (ORNL).

1.3 COMMENTS ON STATISTICAL METHODS OF EXPERIMENTAL DESIGN

A number of statistical methodologies have been developed for constructing experimental designs. In general, experimental design methods help to select an appropriate sample size (number of design points, experimental units, treatments, experimental runs, etc.) and the most informative combinations of explanatory variables based on a prior knowledge about the uncertainties that might be expected and the functional relationships between various groups of variables that may be used to make inferences about the explored systems. Methodologies for linear models have a history going back to the beginning of the century [see Stigler (1974)]. Convex design theory was mainly initiated by Kiefer's celebrated results [see Kiefer (1959)] and was actively evolving afterwards [see details and references, for instance, in Atkinson and Donev (1992), Bandemer et al. (1977), Box and Draper (1987), Box, Hunter and Hunter (1978), Fedorov (1972), Fedorov and Hackl (1997), Karlin and Studden (1966), Pukelsheim (1993), and Silvey (1980)].

Standard experimental design methods for linear models (regression or response surfaces) are a subset of optimal design methodologies, which are used as a benchmark for determining how "good" designs could be under rigid assumptions. In practice, optimal design methodology is frequently applied to the situations in which those assumptions hold "very approximately". At that point a practitioner either believes that violation of those assumptions does not lead to a drastic divergence of a design proposed by idealized theory from actually optimal design or, together with statisticians, starts to look for more realistic approaches. Design for nonlinear response models, for models of finite validity range, or models with correlated observational errors are typical examples of the

latter situation.

There are a number of features that separate experiments on large computer networks from the "standard cases". In particular, the assumptions of homogeneous variances (or nonhomogeneous but known) and no correlation of observational errors are essential for many results in design theory. The assumption that there is no correlation between different observations can work as an admissible approximation for computer network experiments. For instance, the short-term measurements of activities on two remote sites (computers, servers) may be considered stochastically independent. But even in this case, the problem is still quite different from the ones analyzed in experimental design theory: most frequently both the response and its variance contain unknown parameters. A typical example is measuring counts (number of packages per time unit, train length, arrivals, etc.). In this case, a Poisson type distribution can be considered as a candidate to model the stochastic components of observed characteristics. For Poisson regression, the presence of unknown parameters in the variance moves the problem from the linear paradigm to the nonlinear one, which demands significantly more effort to build an optimal design.

As an example of another unusual model for experimental design theory, we can point to the use of compound distributions to model stochastic traffic of qualitatively different messages on the Internet. The "demixing" problem is still waiting for new approaches both in statistical analysis and experimental design.

Even when we use some simple linear models to describe network traffic, the size of a network can create tremendous difficulties in listing elements of the design space. These difficulties are amplified if, instead of deterministic control, an experimenter has weaker control in the selection of elements from this space. For instance, in selecting a particular route to communicate with a remote site, an experimenter cannot be sure that the message will travel along this selected route. With some probability it may be redirected by the network management to another route. So, some new approaches in experimental design taking into account that type of uncertainties must be elaborated.

The list can be continued, and we foresee a number of new studies addressing the above and related problems. Unlike many other areas we need on-line technologies that allow us to create optimal observational schemes in real time (consider, for instance, measurements in the optimal routing problem). The objective of this study is relatively modest: to adjust and complement existing experimental design techniques to make them useful in the construction of optimal monitoring networks, to attract the attention of practitioners to a possible increase of efficiency of measurements, and to post and discuss new statistical problems that are waiting for solutions.

1.4 OUTLINE OF THE PAPER

Most of the objectives of optimal experimental design are the natural extension of what is pursued in estimation theory. The aim of the latter is to obtain the best estimators given the data. The previous one pursues the further gain in precision through better allocation of the measuring resources. In this text, we are mainly working with regression or response type models. In other words, we analyze functional links between different groups of random variables. The stochastic behavior of communication networks requires that we consider those types of functional links that satisfactorily describe and incorporate the corresponding random variables. The regression models and corresponding analysis techniques are probably most widely applied when the average tendencies for one group of variables (responses) conditioned on the second groups of variables (predictors) is of interest. In Chapt. 2 we summarize the main results from regression analysis with the necessary

adaptations to handle the problems generated in the analysis of network performance. Sect. 2.1 contains a very short exposure of well known results that we need to make the text self-consistent and to introduce the reader to the most important facts about the information and covariance matrices which are the main elements in many design problems considered in this research. Additivity of information matrices is one of the most crucial facts and allows us to introduce a concept of optimal design (see Sect. 2.1). We include the Bayesian estimators but treat them exclusively in terms of the second moments: this allows us to easily adapt the existing methods to situations with prior information. The use of prior information frequently is the only resort from singularities in the design of optimal monitoring systems (see Sects. 3.4 and 3.5).

In Sects. 2.2 and 2.3 we discuss various generalizations that make it possible to analyze the results of correlated measurements or measurements with variances that depend upon unknown parameters. The latter is a necessity for many problems in network analysis in which the Poisson model or its various generalizations are popular and natural choices; see Sect.2.3.3. We propose "iterated" estimators as the main tool for data analysis. The use of this estimator provides compatibility between data analysis and existing optimal monitoring design techniques, especially in a multi-stage setting.

We briefly discuss models with nonlinear responses in Sect. 2.4. There exists vast literature on the subject. What is important for this study is to show one principal difference between linear and nonlinear response models: unlike the linear case, information matrices in the non-linear case depend on estimated parameters. This fact creates noticeable difficulties in experimental design and most frequently pushes a practitioner towards sequential design methods. We, however, reserve the extension of optimal monitoring design to the nonlinear models for our forthcoming studies.

Chapt. 3 starts with a trivial example that illuminates what kind of design problems may be encountered in network analysis. In the estimation problem, we can construct the best linear unbiased estimator, and it is the best one for any design in the sense of the covariance matrix ordering (i.e., in the sense of Loewner's ordering, see comments to (2.10)). One of the main conclusions of Sect. 3.1 is that, in general, Loewner's optimization cannot be done in the design world. Consequently, we must formulate some scalar optimality criterion and minimize it with respect to the experimental design.

We provide a short survey of experimental design theory, including only those facts that are needed in optimal monitoring design. The main reference for Chapt. 3 is a book on experimental design by Fedorov and Hackl (1997). Whenever it is possible, we provide some simple examples based on very simple networks to illustrate the techniques. Computational Sects. 3.3 and 3.4 are very practically oriented. The earlier versions of examples from Sect. 3.4 were mainly developed by Flanagan (1997) during her work in the LDRD project "Network Performance Understanding". In Sect. 3.5 we attempt to apply experimental design techniques to optimal monitoring of the Department of Energy's Energy Sciences Network (ESnet). We have made a number of assumptions which may be easily criticized by practitioners. It is obvious to us that there is a great distance between "on line" applications and what is described in that section. However, our objective is, to a very large extent, not only to produce something which can be of immediate use but to attract the attention of the network community to the opportunities. The elaboration of assumptions and bringing them closer to the real world is our nearest objective. In the framework of the selected model we discuss the following problems of optimal monitoring based on a simplified graph of the ESnet containing 34 site nodes and 39 edges and the D-criterion:

- Optimal monitoring design, in which all 34 nodes/sites can be considered as hosts, i.e., all of them are able to measure along the routes prescribed to them.

- Selection of the given number of partners (or "team" selection) and the corresponding destination nodes and routes to optimally monitor the ESnet.
- Introduction of measurement uncertainties depending on the selected routes and analysis of changes in optimal monitoring they can cause.

A number of practical and theoretical relatively straightforward generalizations, together with problems, which need principally new approaches, are discussed in the concluding section of Chapt. 3. We think that two problems (see Subsects. 3.6.4, *Other Types of the Information Matrix Normalization* and 3.6.5, *Heavy Tailed Distributions*) are most interesting for applications and are a real challenge from a mathematical point of view. We are not familiar with any publications on experimental design, or more specifically, in optimal monitoring design for models discussed in these sections.

In Chapt. 4 we abandon the parameterized response models and consider relationships between various groups of variables that can be modeled through covariance structures. We work with linear predictors and covariance matrices. The final objectives are close to what they are in Chapt. 3, but the emphasis is naturally on prediction of network characteristics at the prescribed set of sites (nodes, edges, etc.). Formally, the proposed approach is based on ideas that are close to the ideas of convex design theory. We describe main properties of optimal allocations (Sect. 4.2) and the first order algorithm (Sect. 4.3) and illustrate the applicability of the proposed methods in Sect. 4.4, in which we return to the monitoring problem for the ESnet. In particular, we analyze the changes that are observed when the ratio of the measurement variance to the number of available measurements varies. It is shown that the covariances between different interrogated sites are important only if this ratio is sufficiently large. Otherwise, the monitoring schemes adhere to this very simple principle: interrogate sites with the greatest variability of the measured characteristics.

In Sect. 4.5, we bridge some heuristic approaches developed in analysis of spatial fields with the developed theory and show that it is possible to introduce similar approaches in network monitoring design. The main idea is extremely simple and attractive. If you have data simultaneously collected for several sites, then project each of them on all others and delete what is well explained by measurements at other sites from your collection of sites to monitor. Repeat the action until an economically sound number of sites is left in your collection. Is this procedure optimal? We show that it may be very close to optimal under rather mild conditions.

Through the entire text, we try to make exact chapter references when results from books are cited. We skip the proofs of many results in hope that the concluding bibliography can help the reader find the necessary details.

Chapter 2

REGRESSION TYPE MODELS IN NETWORK MEASUREMENTS

2.1 STANDARD LINEAR MODELS

2.1.1 Example: Simple Network

Let us consider the simple network of Fig. 2.1 with a single host node 1. Let a test-signal (packet, file, etc.) be sent from that node to any of nodes 2-4 along a selected (i.e., the "host" may select it) route, and this signal returns along the same route. We assume that no significant changes occur in the network activity during an experiment. Let the expected travel time from the host node to the destination node along a specified i-th route and return by the same route be

$$E(y|x_i) = \theta^T x_i \quad , \tag{2.1}$$

where E stands for expectation; vectors x_i and θ have m components; m is the number of edges in the corresponding graph (in our example it equals 5). If the α-th edge is included in the i-th route, then $x_{i\alpha} = 1$, otherwise $x_{i\alpha} = 0$. The vector of expected individual travel times is $\tau = \theta/2$ in the case when a test signal is sent to a destination site and then without any additional delay is automatically bounced back to a host site, as assumed for the *ping* program.

Linear model (2.1) may describe more complicated and interesting situations than were described in the beginning of this example. For the network at Fig. 2.1, the vector θ may consist of components corresponding to the delay times on edges ($\theta_1 = 2\tau_1, \ldots, \theta_5 = 2\tau_5$) and to the processing times at nodes ($\theta_6 = \pi_1, \ldots, \theta_9 = \pi_4$). In this text, we will neglect processing times to simplify discussions.

To complete the formulation of model (2.1), we have to make assumptions about the probability distribution of y. Unless otherwise stated, our analysis and design methods are based on the first and second moments of this distribution. Thus, we have to complement (2.1) with some statement

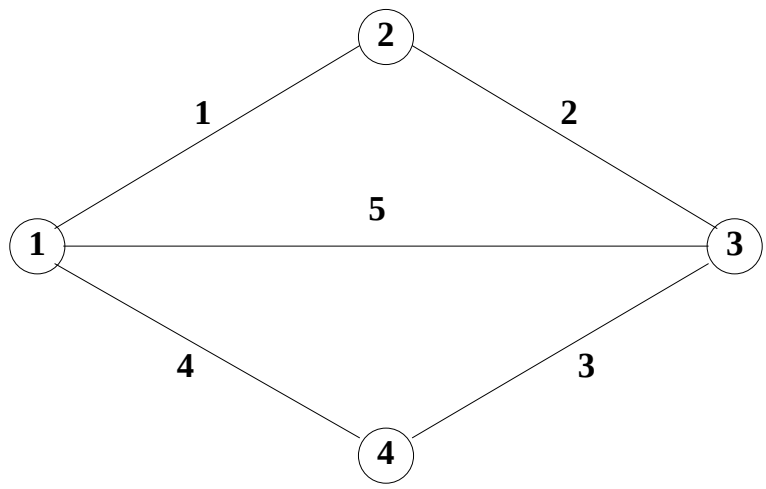

Figure 2.1: Network graph with 4 nodes and 5 edges.

about the second moments. For instance,

$$Var(y|x_i) = E(y - \theta^T x_i)^2 = \sigma^2(x_i), \tag{2.2}$$

where $\sigma^2(x_i)$ is given, and

$$Cov(y, y'|x_i, x_{i'}) = E\left[(y - \theta^T x_i)\ (y' - \theta^T x_{i'})\right] = 0, \quad i \neq i' \tag{2.3}$$

Note that x_i may be equal to $x_{i'}$ (i.e., repeated observations are also uncorrelated). To find a reasonable structure for the function $\sigma^2(x)$, information on the stochastic nature of the network and measurements must be thoroughly analyzed. The popular and simplest choice is that $\sigma^2(x) \equiv \sigma^2$. In the case of routes with a given or known number of edges, it may be reasonable or expedient to assume that $\sigma^2(x) = \sigma^2 \times$ (number of edges included in x).

2.1.2 Linear Regression Models

Model (2.1) - (2.3) is a particular case of the linear regression model. The methods of analysis and design of experiments are well developed for this class of models [cf. Fedorov and Hackl (1997), Pukelsheim (1993), Rao (1973)]. In this chapter, we consider a slightly more general model than (2.1) − (2.3). Namely, we assume that

$$E(y|x_i) = \theta^T f(x_i) \quad , \tag{2.4}$$

where $f(x)$ is a $(m \times 1)$ vector of given basis functions. Of course, $\theta^T f(x) = f^T(x)\theta$, and we use either presentation as needed without any additional comment. It may be that $f(x) = x$ like in the above example, but model (2.4) allows us to include more complicated cases. For instance, interaction $x_{i\alpha}x_{i\beta}$ that can occur if the behavior of different parts of the network depends upon each other.

In this section, we assume that

$$Var(y|x_i) = \sigma^2(x_i) \tag{2.5}$$

and all measurements are uncorrelated. It is a common practice to replace (2.5) by a simpler assumption

$$Var(y|x_i) \equiv 1 \quad , \tag{2.6}$$

because (2.5) can be transformed to (2.6) if one uses

$$y' = \sigma^{-1}(x_i)y, \quad f'(x_i) = \sigma^{-1}(x_i)f(x_i) \quad . \tag{2.7}$$

We prefer to work with (2.4) and (2.5) instead of (2.4) and (2.6) in spite of the longer formulae: it makes an easier transition to cases in which the variance-covariance structure is more complicated.

2.1.3 The Best Linear Unbiased Estimator and the Least Squares Method

The least squares and best linear unbiased type of estimator is defined either as

$$\hat{\theta} = \arg\min_{\tilde{\theta}} \; E\left[(\tilde{\theta} - \theta)\,(\tilde{\theta} - \theta)^T\right] \quad , \tag{2.8}$$

where $E(\tilde{\theta}) = \theta$, $\tilde{\theta} = L\mathcal{Y}$, $\mathcal{Y} = (y_1, \ldots, y_n)$, and L is a $(m \times n)$ matrix, or as

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \sigma^{-2}(x_i) \; \left[y_i - \theta^T f(x_i)\right]^2 \quad . \tag{2.9}$$

The Gauss-Markov theorem (c.f. Rao (1973), Chap. 4a) tells us that the solutions of (2.8) and (2.9) coincide if the covariance matrix

$$D = E\left[(\hat{\theta} - \theta)\,(\hat{\theta} - \theta)^T\right] \tag{2.10}$$

is regular. Note that *minimization in (2.8) is understood in the sense of non-negative definite matrix ordering* [i.e., Loewner's ordering (cf. Pukelsheim (1993), Chap. 4], so we say that $A \geq B$, if $A = B + C$ and C is non-negative definite ($C \geq 0$).

We use the same characters for random variables and their realizations. Usually the latter are marked by some subscripts. When it is necessary to emphasize that we are interested in analyzing the properties of $\hat{\theta}$ as a random vector, the term "estimator" is used. When the particular meaning of this vector is discussed we use the term "estimate".

Some exercises in matrix algebra show that

$$\hat{\theta} = \underline{M}^{-1}Y \quad , \tag{2.11}$$

where

$$\underline{M} = \sum_{i=1}^{n} \sigma^{-2}(x_i) f(x_i) f^T(x_i) \quad , \tag{2.12}$$

$$Y = \sum_{i=1}^{n} \sigma^{-2}(x_i) y_i f(x_i) \quad . \tag{2.13}$$

Matrix $\underline{M}$ is called "information matrix," and, for the sake of simplicity, it is assumed to be regular ($|\underline{M}| \neq 0$) when the inverse operation is used. Various discussions on singular cases may be found, for instance, in Fedorov and Hackl (1997), Chap. 1, and Pukelsheim (1993), Chap. 3. It must be emphasized that in any experiment the information matrix is nonnegative definite and additive (i.e., it is the sum of information matrices $m(x_i)$ describing the information gained at each measurement):

$$\underline{M} = \sum_{i=1}^{n} m(x_i) \quad , \tag{2.14}$$

where

$$m(x_i) = \sigma^{-2}(x_i) f(x_i) f^T(x_i) \quad . \tag{2.15}$$

If there are repeated measurements, then

$$\underline{M} = \sum_{i=1}^{n} r_i \sigma^{-2}(x_i) f(x_i) f^T(x_i) = \sum_{i=1}^{n} r_i m(x_i) \tag{2.16}$$

and

$$Y = \sum_{i=1}^{n} r_i \overline{y}_i f(x_i) \quad , \tag{2.17}$$

where r_i is a number of measurements made at point x_i, and $\overline{y}_i = r_i^{-1} \sum_{j=1}^{r_i} y_{ij}$.

The covariance matrix of the best linear unbiased estimator $\hat{\theta}$ is

$$\underline{D} = Var(\hat{\theta}) = \underline{M}^{-1} \quad . \tag{2.18}$$

The best linear unbiased estimator of any linear vector function $L\theta$ is $L\hat{\theta}$ with the covarience matrix

$$Var(L\hat{\theta}) = L\underline{D}L^T \quad . \tag{2.19}$$

For instance, to estimate a regression function $\theta^T f(x)$ we can use $\hat{\theta}^T f(x)$ and

$$\underline{d}(x) = Var[\hat{\theta}^T f(x)] \; = \; f^T(x) \underline{D} f(x) \quad . \tag{2.20}$$

In what follows, both matrix (2.19) and function (2.20) are essential for comparison of "quality" of experiments. Both are uniquely defined by the information matrix $\underline{M}$.

2.1.4 Properties of Information Matrices

Let us now summarize the main properties of matrices defined by (2.16). Some have been already mentioned; others directly follow from the definition (2.16).

1. The information matrix $\underline{M}$ is uniquely defined by the collection $\{x_i, r_i,\}_1^n$ and by the vector of the basis functions $f(x)$, and does not depend upon the results of measurements.
2. $\underline{M}$ is symmetric and non-negative definite.
3. $\underline{M}$ is the sum of the individual information matrices for each measurement.

Let the collection

$$\xi = \{x_i, p_i\}_1^n, \quad p_i = r_i/N, \quad N = \sum_{i=1}^{n} r_i$$

be called the design (sometimes the plan) of an experiment. To compare the quality of experiments with different numbers of observations, the normalized information matrix

$$M(\xi) = N^{-1}\underline{M} = \sum_{i=1}^{n} p_i \sigma^{-2}(x_i) f(x_i) f^T(x_i) = \sum_{i=1}^{n} p_i m(x_i) \tag{2.21}$$

can be useful.

If the set of designs is extended to the set of all possible probability measures, $\xi(dx)$, on a compact (for simplicity) set X, then the normalized information matrix is defined as

$$M(\xi) = \int_X \sigma^{-2}(x) f(x) f^T(x) \xi(dx), \quad \int_X \xi(dx) = 1 \quad . \tag{2.22}$$

Note that

$$M(\xi) = \int_X m(x)\xi(dx) \quad ,$$

where $m(x)$ is the information matrix of a measurement made at x; see (2.15). The normalized covariance matrix is defined as

$$D(\xi) = M^{-1}(\xi) \quad , \tag{2.23}$$

if $M(\xi)$ is regular. Unless it would otherwise be ambiguous, we omit the word "normalized".

Let $\underline{M}_1$ and $\underline{M}_2$ be the information matrices of two experiments made according to designs ξ_1 and ξ_2, respectively. Then, because of the additivity of the information matrix, we can write

$$\underline{M} = \underline{M}_1 + \underline{M}_2 \Leftrightarrow NM(\xi) = N_1 M(\xi_1) + N_2 M(\xi_2) \quad ,$$

and dividing both sides by N we have

$$M(\xi) = (1-\alpha)M(\xi_1) + \alpha M(\xi_2) \quad ,$$

where

$$\alpha = \frac{N_2}{N_1 + N_2} \quad \text{and} \quad \xi = (1-\alpha)\ \xi_1 + \alpha\ \xi_2 \quad .$$

Thus, the set of normalized information matrices is convex. Additivity is a property of information matrices only when observations are uncorrelated.

2.1.5 Presence of Prior Information

If prior information about the estimated parameters exists and can be expressed in terms of a prior distribution with mean θ_0 and variance-covariance matrix $\underline{D}_0$, then the combined information matrix is:

$$\underline{M}_{tot} = \underline{M} + \underline{M}_0 \quad , \tag{2.24}$$

where $\underline{M}_0 = \underline{D}_0^{-1}$. The covariance matrix of the parameter estimators is calculated as $\underline{D}_{tot} = \underline{M}_{tot}^{-1}$. The updated estimator for θ may be presented as

$$\hat{\theta}_{tot} = (\underline{M} + \underline{M}_0)^{-1}(\underline{M}\hat{\theta} + \underline{M}_0\theta_0) \quad , \tag{2.25}$$

where $\hat{\theta}$ is defined by (2.11). In the case of normally distributed measurements and a normal prior distribution, $\hat{\theta}_{tot}$ coincides with the Bayesian estimator [cf. Box and Tiao (1992), Chap. 2].

The latter formula contains a hint of how to regularize the estimation problem when the original information matrix is singular: one has to use a regularized matrix

$$\underline{M}(c) = \underline{M} + c\underline{M}_0,\ \underline{M}_0 > 0,\ \ c > 0$$

where frequently [for instance, in ridge regression exercises, Seber (1977), Chap. 3.10] $M_0 = I$, and the behavior of $\underline{M}^{-1}(c)$ as a function of c. This approach is convenient when a linear combination $\gamma = L\theta$ can be uniquely estimated for a given $\underline{M}$ even when the latter is singular. In this case [cf. Albert (1972), Chap. 3] for any estimable linear function $\gamma = L\theta$

$$\hat{\gamma} = \lim_{c\to 0} L(\underline{M} + c\underline{M}_0)^1 Y$$

and

$$Var\,\hat{\gamma} = \lim_{c\to 0} L(\underline{M} + c\underline{M}_0)^{-1} L^T \quad .$$

Details about the use of prior information and regularization can be found in Atkinson and Fedorov (1988) and Pilz (1991).

2.2 CORRELATED OBSERVATIONS

In the presence of correlation, it is more practical to rewrite model (2.4) in the matrix form:

$$E(\mathcal{Y}|F) = F^T\theta \quad , \tag{2.26}$$

where F is an $m \times n$ matrix and is defined as $F = (f(x_1), \ldots, f(x_n))$. Let the vector of observations have the following covariance matrix:

$$C = Var(\mathcal{Y}|F) = E\left[\left(\mathcal{Y} - F^T\theta\right)\left(\mathcal{Y} - F^T\theta\right)^T |F\right] \quad . \tag{2.27}$$

If C is known, then the best linear unbiased estimator is [cf. Rao (1993), Chap. 4a]:

$$\hat{\theta} = \underline{M}^{-1}Y \quad , \tag{2.28}$$

where

$$\underline{M} = FC^{-1}F^T \quad \text{and} \quad Y = \mathcal{Y}C^{-1}F^T \quad .$$

Note that

$$\hat{\theta} = \arg\min_{\theta} \left(\mathcal{Y} - F^T\theta\right)^T C^{-1} \left(\mathcal{Y} - F^T\theta\right) \quad .$$

Sometimes the matrix C may be constructed from an analysis of the stochastic behavior of the vector $\mathcal{Y}$. The most popular examples are various auto-regression models [cf. Anderson (1994)] and competition models [cf. Martin (1996)]. However, more frequently only some historical data are available to construct and estimate the covariance matrix C.

2.2.1 Example

To continue to analyze the model introduced in the example from Sect. 2.1.1, let the i-th observation be described as

$$y = \theta^T x_i + \mathcal{E}^T x_i \quad ,$$

where the random vector $\mathcal{E}$ has the same dimension as θ, $E(\mathcal{E}) = 0, Var(\mathcal{E}) = \sigma^2 I$, and $i = 1, \ldots, n$. For the sake of simplicity, we assume that there are no repeated measurements. One can imagine that all n observations are made almost instantaneously (i.e., nothing has been changed on the network during those n observations). If the vector x_i has k nonzero components, then

$$C_{ii} = Var(y|x_i) = x_i^T x_i \sigma^2 = k\sigma^2 \quad .$$

If the vectors x_i and $x_{i'}$ have ℓ common nonzero components (routes contain the same edges), then

$$C_{ii'} = Cov(y, y'|x_i, x_{i'}) = x_i^T x_{i'} \sigma^2 = \ell\sigma^2 \quad .$$

Table 2.1 lists all possible routes for the considered example (single-host and multihost cases are included). One can see, for instance, that observations made on routes 1 and 2 are not correlated (i.e. $C_{12} = 0$, while for routes 1 and 9, there are two common edges and hence $C_{19} = 2\sigma^2$).

Table 2.1: Routes for the network problem with 4 nodes and 5 edges

Host-Interrogated Site	Route Number	Edges 1	2	3	4	5
1 - 2	1	0	1	1	1	0
1 - 2	2	1	0	0	0	0
1 - 2	3	0	1	0	0	1
1 - 4	4	0	0	0	1	0
1 - 4	5	0	0	1	0	1
1 -4	6	1	1	1	0	0
1 - 3	7	0	0	0	0	1
1- 3	8	1	1	0	0	0
1 - 3	9	0	0	1	1	0
2 - 3	10	0	1	0	0	0
2 - 3	11	1	0	0	0	1
2 - 3	12	1	0	1	1	0
2 - 4	13	1	0	0	1	0
2 - 4	14	0	1	1	0	0
2 - 4	15	0	1	0	1	1
2 - 4	16	1	0	1	0	1
3 - 4	17	0	0	1	0	0
3 - 4	18	0	0	0	1	1
3 - 4	19	1	1	0	1	0

2.3 VARIANCE DEPENDING UPON UNKNOWN PARAMETERS

2.3.1 Iterated Estimators Based on the Best Linear Unbiased Estimators

In analyzing network data the variance of the observations may depend upon the same parameters θ as the regression function [see, for instance, Batsell et al. (1997), Vardi (1996)], that is

$$Var(y|x) = \sigma^2(x, \theta) \quad .$$

In this case "iterated estimators" can be used to estimate θ (see, for instance, Fedorov and Hackl (1997), Chap. 1). For the linear regression function and uncorrelated observations, the iterated estimator is

$$\hat{\theta} = \lim_{s \to \infty} \theta_s \quad , \tag{2.29}$$

where at each step s, we do the following calculations:

$$\begin{aligned}
\theta_s &= \underline{M}_s^{-1} Y_s \quad , \\
\underline{M}_s &= \sum_{i=1}^{n} r_i \sigma^{-2}(x_i, \theta_{s-1}) f(x_i) f^T(x_i) \quad , \\
Y_s &= \sum_{i=1}^{n} r_i \sigma^{-2}(x_i, \theta_{s-1}) \overline{y}_i f(x_i) \quad .
\end{aligned}$$

If $N = \sum_{i=1}^{n} r_i$ and P_N is the probability that limit (2.29) exists, then it can be shown that under rather mild assumptions

$$\lim_{N \to \infty} P_N = 1 \quad .$$

If there exists a regular matrix

$$M(\theta) = \lim_{N \to \infty} N^{-1} \underline{M}_N(\theta) \quad ,$$

where $\underline{M}_N(\theta) = \sum_{i=1}^{n} r_i \sigma^{-2}(x_i, \theta) f(x_i) f^T(x_i)$, and θ is the vector of true values of unknown parameters, then the iterated estimator is strongly consistent [see Rao (1973), Chap. 2c for a definition and compare with Batsell et al. (1997)] and asymptotically has the same efficiency as the best linear unbiased estimator for the case when variances $\sigma^2(x_i, \theta) = \sigma_i^2$ are given. The asymptotic normalized covariance matrix $D = \lim_{N \to \infty} N Var(\hat{\theta})$ coincides with $M^{-1}(\theta)$. Unlike the previous section, the covariance matrix depends on the vector θ, which is not known *a priori*.

Instead of the information matrix defined by (2.22), we have to use the matrix

$$M(\xi, \theta) = \int_X \sigma^{-2}(x, \theta) f(x) f^T(x) \xi(dx) \quad , \tag{2.30}$$

where $\xi(dx)$ is a limit design for the sequence $\{\xi_N\} = \{x_i, r_i | N\}$.

The vector θ_s also may be presented as

$$\theta_s = \arg\min_{\theta} \sum_{i=1}^{n} \sigma^{-2}(x_i, \theta_{s-1}) \left[\overline{y}_i - f^T(x_i)\theta \right]^2 \quad . \tag{2.31}$$

However, it should be emphasized that

$$\hat{\theta} \neq \tilde{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} r_i \sigma^{-2}(x_i, \theta) \left[\overline{y}_i - f^T(x_i)\theta \right]^2 \quad , \tag{2.32}$$

and, in general, the least square estimator $\tilde{\theta}$ is not consistent [cf. Fedorov (1974) and Malyutov (1987)].

2.3.2 Iterated and Maximum Likelihood Estimators

Construction of the estimator $\hat{\theta}$ is based on the concept of the best linear unbiased estimation and the least squares method. The maximum likelihood method may generate a different estimator. For example, in the case of normally distributed observations, the maximum likelihood method leads to the following iterated estimator:

$$\hat{\theta}_M = \lim_{s\to\infty} \theta_s \quad , \tag{2.33}$$

$$\theta_s = \arg\min_{\theta} \sum_{i=1}^{n} r_i \left\{ \sigma^{-2}(x_i, \theta_{s-1}) \left[\overline{y}_i - f^T(x_i)\theta \right]^2 \right.$$

$$\left. + \frac{1}{2}\sigma^{-4}(x_i, \theta_{s-1}) \left[\hat{\sigma}_i^2(\theta_{s-1}) - \sigma^2(x_i, \theta) \right]^2 \right\}$$

where $\hat{\sigma}_i^2(\theta) = \frac{1}{r_i} \sum_{j=1}^{r_i} \left[y_{ij} - f^T(x_i)\theta \right]^2$.

Whenever the sequence $\{\theta_s\}$ converges, the vector $\hat{\theta}_M$ coincides with the maximum likelihood estimator

$$\begin{aligned} \theta^* &= \arg\max_{\theta} \prod_{i=1}^{n} \prod_{j=1}^{r_i} \frac{1}{\sqrt{2}\pi\sigma(x_i, \theta)} e^{\frac{1}{2}[\frac{y_{ij} - f^T(x_i)\theta}{\sigma(x_i,\theta)}]^2} \\ &= \arg\min_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{r_i} \left\{ \ln \sigma^2(x_i, \theta) + \left[\frac{y_{ij} - f^T(x_i)\theta}{\sigma(x_i, \theta)} \right]^2 \right\} . \end{aligned} \tag{2.34}$$

If in estimator (2.33) we abandon the assumption on normality, then this estimator can be considered as a generalization of iterated estimator (2.31) when both functions $f^T(x)\theta$ and $\sigma^2(x, \theta)$ must be fitted to their observed values.

2.3.3 Example: Estimation of Delay/Travel Time

The travel time from the $(\alpha - 1)$-th node through the α-th node (i.e., the service time at the latter is included) may be characterized by a minimal delay time τ_α and by a mean time σ_α of various services provided at that element of a route. In other words, we can describe a transmission process by the random variable

$$z_\alpha = \tau_\alpha + u_\alpha \quad ,$$

where τ_α is the minimal delay time and u_α is a random component of service. We assume that in the case of repeated measurements, all the *pings* travel along the same route.

Under a rather natural set of assumptions, u_α is distributed exponentially with parameter σ_α, so the probability density function is exponential [cf. Snyder (1975), Chap. 2]:

$$p_\alpha(u) = \frac{1}{\sigma_\alpha} e^{-\frac{u}{\sigma_\alpha}}, \quad u \geq 0, \quad \text{or} \quad p_\alpha(z) = \frac{1}{\sigma_\alpha} e^{\frac{z-\tau_\alpha}{\sigma_\alpha}}, \quad z \geq \tau_\alpha \quad .$$

Probably the second simplest choice is a Weibull distribution, which may be derived from standard exponential distribution by setting $u = (z - \tau)^c, \quad z > \tau$. There are numerous publications related to network traffic analysis in which the applicability or non-applicability of the Poisson model (and consequently the exponential distribution for service, delay, or travel time) are discussed in depth. Examples are Jain and Routhier (1986), Leland et al. (1994), Frost and Melamed (1994), Paxson and Floyd (1995), and Paxson (1996, 1997). In this study, we prefer to be within the Poisson paradigm.

Let the *ping* message traverse k edges and let

$$Z_k = \sum_{\alpha=1}^{k} z_\alpha = \sum_{\alpha=1}^{k} \tau_\alpha + \sum_{\alpha=1}^{k} u_\alpha = \tau + U_k \quad . \tag{2.35}$$

It may be shown [see, for instance, Cox (1967), Chap.1.4 and Johnson et al. (1994), Chap.19] that for independent $u_\alpha, \alpha = 1, \ldots, k$, U_k has the general Erlangian distribution:

$$p(U) = \sum_{\alpha=1}^{k} A_\alpha \frac{1}{\sigma_\alpha} e^{-\frac{U}{\sigma_\alpha}} \quad , \tag{2.36}$$

where

$$A_\alpha = \prod_{\alpha' \neq \alpha}^{k} \frac{\sigma_\alpha}{\sigma_\alpha - \sigma_{\alpha'}} \quad ,$$

and all σ_α are different. The expected travel time is

$$E(Z_k) = \sum_{\alpha=1}^{k} (\tau_\alpha + \sigma_\alpha) = \tau + \underline{\mu}_k \quad ,$$

and the variance is

$$Var(Z_k) = Var(U_k) = \sum_{\alpha=1}^{k} \sigma_\alpha^2 \quad .$$

Let x be a vector with components that may be either 1 or 0. Similar to Sect. 2.1.1, the α-th component of the vector x indicates the α-th edge is included in the route on which the delay time is measured. Assuming that the minimal delay times $\tau^T = (\tau_1, \ldots, \tau_m)$, $m = \dim x$, are known and introducing $y = z - \tau^T x$, we have in terms of the regression model that

$$E(y|x) = \theta^T x \quad \text{and} \quad Var(y|x) = \sigma^2(x, \theta) \quad ,$$

where $\theta_\alpha = \sigma_\alpha$, $\sigma^2(x, \theta) = \sum_{\alpha=1}^{m} \theta_\alpha^2 x_\alpha^2$, and $x^T x = k$.

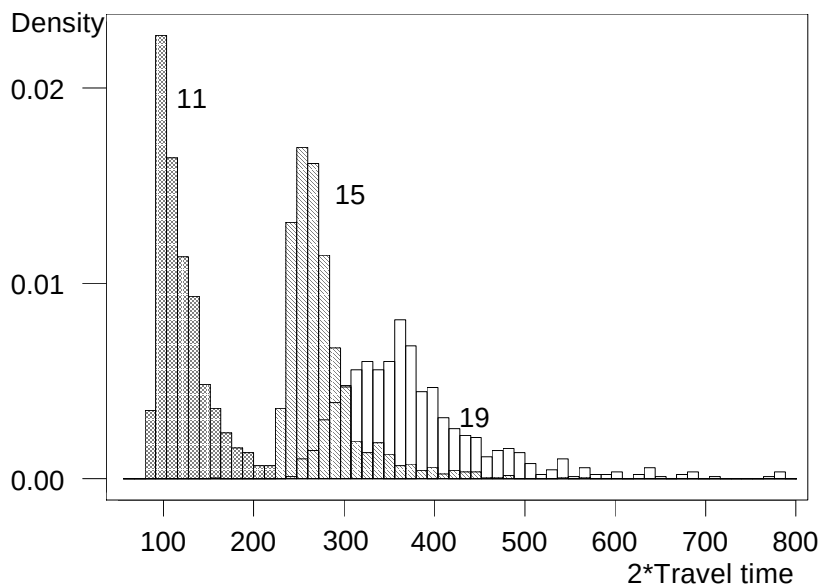

Figure 2.2: Histograms for various number of edges included in a route.

The above model is a perfect candidate for the iterated estimation based on (2.33). It must be noted that the maximum likelihood method may provide a better estimator for $\sigma_1, \ldots, \sigma_m$ being based on more detailed information described by a density function. The direct use of density function (2.36) leads to computational difficulties, which make the approach impractical. One can hope that for a large k [i.e., when a *ping* traverses sufficiently many edges] the probability distribution of U can be approximated by the normal distribution due to the Central Limit Theorem [cf. Rao (1973), Chap. 2c], and then the iterated estimator (2.31) can be used. The asymptotical ($k \to \infty$) relationship

$$\lim_{k\to\infty} \text{Prob} \left(\frac{U_k - \underline{\mu}_k}{\sqrt{Var(U_k)}} \leq u \right) = \Phi(u) \quad , \tag{2.37}$$

where $\Phi(u)$ stands for the standard normal distribution, $\underline{\mu}_k$ and $Var(U_k)$ were defined earlier, holds for random variables U_k if

$$\lim_{k\to\infty} \max_{1\leq\alpha\leq k} \frac{\sigma_\alpha}{\sqrt{Var(U_k)}} = 0 \quad ,$$

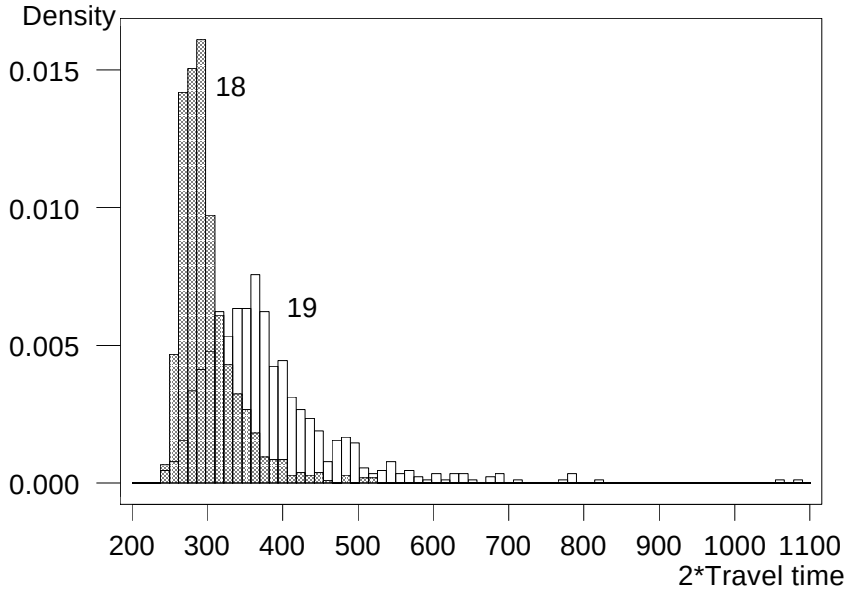

Figure 2.3: The histogram of travel time for two neighbor nodes.

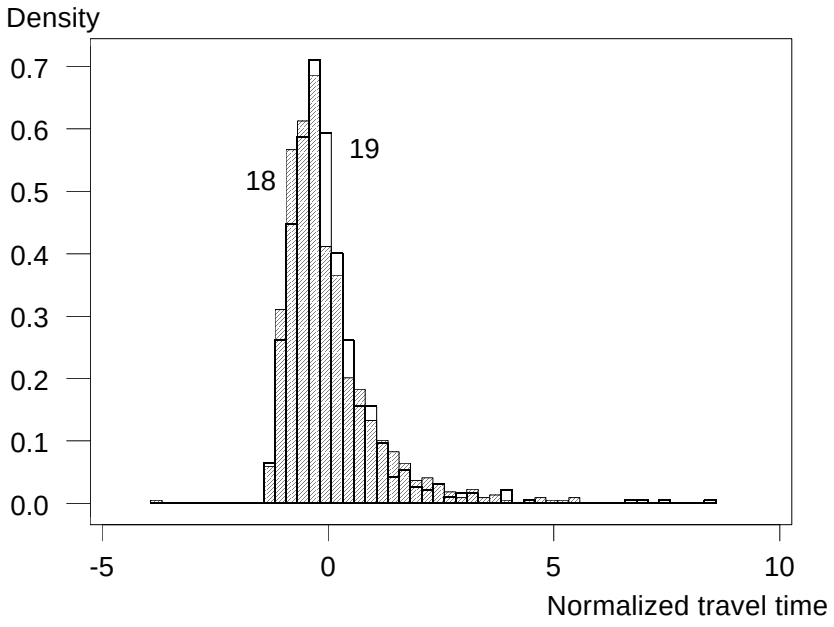

Figure 2.4: Normalized histograms for the travel time for two neighbor nodes.

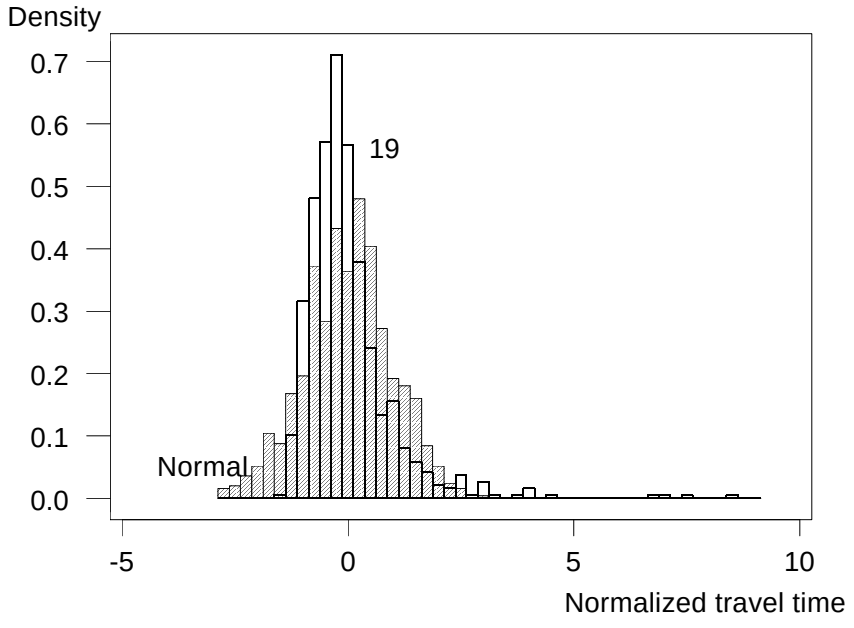

Figure 2.5: The normalized histogram for the travel time and the simulated histogram (1000 trials) for the standard normal distribution.

where only those $\alpha - \tau$ are included, for which $x_\alpha = 1$. The above statement means that the effect of individual terms in the sum becomes smaller and smaller with increasing k. Thus, to use the normal distribution as a reasonable approximation, one must be sure that there are no dominant terms in the sum (2.35).

We used the *traceroute* program [see details and further references in Stevens (1994), Chap. 8] to collect statistics about delay times Z_k. A few remote destination sites were observed. Results look very similar for different sites, and in what follows, the data (sample size is 1000) for the Vienna University of Economics (the destination site) are discussed. The *traceroute* program reports the round-trip travel times for every of k nodes that are included in the selected route.

The histograms for U_k with different k are presented in Fig. 2.2, and the reader can see that the significant increase of k does not make the histograms look closer to the normal distribution. One can notice that the distribution for node 19 has a very long right "tail". It can be explained by the presence of different traffic conditions, and that leads to the compoundness of the travel time distributions. Interestingly, the empirical distributions, even for the neighbor nodes look essentially different; see Fig. 2.3, in which the histograms for nodes 18 and 19 are presented. At first glance, the discrepancy between two consequent distributions does not provide too much hope for the convergence of the distributions, which is expected in (2.37). However, the picture (see Fig. 2.4) looks much better for the normalized random variables, that is, for

$$V_k = \frac{U_k - \underline{\hat{\mu}}_k}{\sqrt{\hat{\sigma}^2(U_k)}} \quad ,$$

where $\underline{\hat{\mu}}_k$ and $\hat{\sigma}^2(U_k)$ are the empirical average and variance. They replace $\underline{\mu}_k$ and $Var(U_k)$, respectively. Note that (2.37) takes place for the normalized random variables, when $\underline{\mu}_k$ and $Var(U_k)$ are given.

Still, the empirical distributions (see Fig. 2.5) are not very close to the normal distribution, and the heavy tails require some caution in the application of asymptotical normality; see (2.37). Consequently, estimator (2.34) can be considered an approximate maximum likelihood estimator with all the nice and readily available asymptotic properties only with very serious precaution. Estimator (2.33) is still a useful tool in the framework of estimation based on the first two moments. Some asymptotic properties, like consistency and asymptotical efficiency of the corresponding estimators, can be verified following ideas from Fedorov (1974) and Malyutov (1987). However, the details must be modified to take into account outliers generated by the presence in the compound distributions of "low weight" components responsible for the "tails."

2.3.4 Estimating Source-Destination Network Traffic Intensities (Network Tomography)

The problem of estimating the node-to-node traffic intensity from repeated measurements on the edges (links) of a network has been discussed by several authors. Vardi (1996) probably contains the most complete and updated information on the subject. In this subsection, we use Vardi's formulation of the problem but apply the iterated estimator instead of the EM (expected maximum likelihood) estimator; see Dempster, Laird, and Rubin (1977). We consider the derivation of the EM estimator unnecessarily complicated for the considered problem and much less transparent than the iterated estimator.

Let m denote the source-destination pairs, and let these pairs be numbered $\alpha = 1, \ldots, m$, where

$m \leq d(d-1)$, and d is a number of nodes. For the sake of simplicity, we assume that each source-destination pair uses a single route. In this case the total traffic on the i-th edge equals

$$u_i = \sum_{\alpha=1}^{m} G_\alpha x_{\alpha i} \quad ,$$

where $x_{\alpha i} = 1$, if the α-th source-destination pair uses the i-th edge, and $x_\alpha = 0$ otherwise. Following Vardi (1996), we assume that number of messages (communication units) G_α corresponding to the α-th pair has a Poisson distribution with the parameter θ_α, that is, $p_\alpha(g) = \theta_\alpha^g e^{-\theta_\alpha}/g!$. Hence, similar to (2.1) we can write that

$$E(y|x_i) = \sum_{\alpha=1}^{m} \theta_\alpha x_{\alpha i} \quad ,$$

where $x_i^T = (x_{1i}, \ldots, x_{mi})$. Noting that two different edges x_i and $x_{i'}$ may include traffic between the same source-destination pairs, we conclude that measurements on edges i and i' might be correlated. Therefore, we have to use a particular case of model (2.26), (2.27), that is,

$$E(\mathcal{Y}|X) = X^T\theta \quad , \tag{2.38}$$

$$C(\theta) = Var(\mathcal{Y}|X) = X^T\Theta X \quad , \tag{2.39}$$

where

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ \ldots & \ldots & \ldots & \ldots \\ x_{m1} & x_{m2} & \ldots & x_{mn} \end{pmatrix} \quad ,$$

Θ is diagonal and $\Theta_{\alpha\alpha} = \theta_\alpha$. Model (2.38), (2.39) gives us an example in which the results from Sects. 2.2 and 2.3 must be used together: we have correlated observations, and the covariance matrix depends on unknown parameters. Combining (2.28) and the multi-response version of (2.31) we come to the following iterated estimator:

$$\hat{\theta} = \lim_{s\to\infty} \theta_s, \quad \theta_s = \underline{M}_s^{-1} Y_s \tag{2.40}$$

where $\underline{M}_s = XC^{-1}(\theta_s)X^T$, and $Y_s = \mathcal{Y}C^{-1}(\theta_s)X^T$.

Note that to uniquely estimate all m components of the vector θ, we have to be sure that rank$X = m$. Obviously, rank$X \leq n$, where n is the number of edges included in the experiment, $n = dim\mathcal{Y}$. For the simple graphs presented at Fig. 2.1, we can measure the traffic only on 5 edges. At the same time, we have 12 ordered pairs. Thus, there exists a serious problem with the identification of θ, even if we can measure traffic in both directions separately for each edge (i.e., we have 10 measurements). It can be verified that if the limit in (2.40) exists, then $\hat{\theta}$ coincides with the maximum likelihood estimator as proposed by Vardi (1996) in the framework of the normal approximation without any recommendation about how to compute it. He assumed that the vector $\mathcal{Y}$ can be repeatedly measured sufficiently many times so the Central Limit Theorem can be applied, and reintroduced

the constraints $\theta_\alpha \leq 0$, $\alpha = 1, \ldots, m$, which are not very important in the asymptotical setting and, if needed, can be easily added to (2.40). The identification of θ is a recognized problem related to "demixing in Poisson mixture models" [see, for instance, Hengartner (1997)]. Even for well structured mixtures (in terms of our problem it means that there are some relationships between different components of θ), the estimation problem of θ belongs to the realm of ill-conditioned problems.

One of the opportunities to overcome the discussed difficulty may be the introduction of prior information. If this information can be described in terms of a prior vector $\theta_{(0)}$ and its variance matrix $D_{(0)}$, then we come to the interesting mixture of the Bayesian type estimation and the iterated estimation. Combining (2.25) and (2.40) we may derive the following iterated estimator

$$\hat{\theta}_B = \lim_{s \to \infty} \theta_s \quad , \theta_s = \underline{M}_{tot,s}^{-1} \left(Y_s + D_{(0)}^{-1} \theta_{(0)} \right) \quad , \tag{2.41}$$

where $\underline{M}_{tot,s} = \underline{M}_s + D_{(0)}^{-1}$; the matrix $\underline{M}_s$ and the vector Y_s are defined in the comments to (2.40).

2.4 NONLINEAR MODELS

If a linear (with respect to unknown parameters) regression function $f^T(x)\theta$ is replaced by a more general function $\eta(x, \theta)$ that assumes a possible nonlinearity, then we cannot use the best linear unbiased estimator and have to be confined to estimators that are not based on the concept of linearity. It may be, for instance, either the least squares method estimators or the maximum likelihood estimators. We briefly consider the first ones leaving to the reader some obvious generalizations to use the latter [cf. Fedorov and Hackl (1997), Chap. 1, and Seber and Wild (1989)]. Thus,

$$\hat{\theta}_N = \arg\min_{\theta} \sum_{i=1}^{n} r_i \sigma^{-2}(x_i) \left[\overline{y}_i - \eta(x_i, \theta) \right]^2 \quad . \tag{2.42}$$

The subscript N emphasizes that N observations are used to construct the estimator. It is known [see, for instance, Seber and Wild (1988), Chap. 12] that under rather mild conditions $\hat{\theta}_N$ is a strongly consistent estimator, that is, it almost surely converges [cf. Rao (1973), Chap. 2c.3] to the true vector of parameters θ_t. If the matrix

$$M(\theta) = \lim_{N \to \infty} M_N(\theta) \quad ,$$

where

$$M_N(\theta) = N^{-1} \sum_{i=1}^{n} r_i \sigma^{-2}(x_i) \frac{\partial \eta(x_i, \theta)}{\partial \theta} \frac{\partial \eta(x_i, \theta)}{\partial \theta^T}$$

exists and is regular in some vicinity of θ_t, then for sufficiently large N

$$N \, Var(\hat{\theta}_N) \cong M_N^{-1}(\hat{\theta}_N) \cong M^{-1}(\theta_t) \quad . \tag{2.43}$$

The latter follows from the strong consistency of $M_N^{-1}(\hat{\theta}_N)$ as an estimator of $NVar(\hat{\theta}_N)$. For experimental design, it is important to note that unlike the linear case, the information matrix $M(\theta_t)$ depends upon unknown parameters.

Most of the results discussed in Sects. 2.2 and 2.3 can be generalized in a straightforward manner. For instance, to estimate the unknown regression parameters θ in the case of the variance depending on θ, one has to use the iterated estimators based on (2.31) with the obvious replacement of $f^T(x)\theta$ by $\eta(x,\theta)$. This approach works well for "ad hoc" situations or pilot studies when a relatively small number of cases must be analyzed. Using macros for the nonlinear least squares method [see, for instance, SAS/STAT (1995), Chap. 29] one can almost entirely avoid programming work. Compare this approach with our earlier study Batsell et al. (1997). However, to develop more efficient and compact software, the algorithms combining the iterative nature of estimators and the stepwise Gauss-Newton updating ordinarily used in the nonlinear least squares [cf. Seber and Wild (1989), Chap. 2.1] must be implemented. For instance, the following algorithm can replace (2.31)

$$\begin{aligned} \theta_s &= \theta_{s+1} + \underline{M}^{-1}(\theta_{s-1})Y(\theta_{s-1}) \quad , \\ \underline{M}(\theta) &= \sum_{i=1}^{n} r_i \sigma^{-2}(x_i,\theta) \frac{\partial \eta(x_i,\theta)}{\partial \theta} \frac{\partial \eta(x_i,\theta)}{\partial \theta^T} \quad , \\ Y(\theta) &= \sum_{i=1}^{n} r_i \sigma^{-2}(x_i,\theta)\,(\overline{y}_i - \eta(x_i,\theta)) \quad . \end{aligned} \tag{2.44}$$

Note that some "standard" improvements of the Gauss-Newton iterative procedure that are popular in the least squares method software world do not work in the considered case. For instance, the introduction of

$$\theta_s = \theta_{s-1} + \gamma_s \underline{M}^{-1}(\theta_{s-1})Y(\theta_{s-1}) \quad ,$$

where γ_s provides the monotonic decrease of the sum

$$v^2(\theta_s) = \sum_{i=1}^{n} r_i \sigma^{-2}(x_i,\theta_s)\,(y_i - \eta(x_i,\theta_s))^2 \quad ,$$

may lead to the estimator $\hat{\theta} = \lim_{s\to\infty} \theta_s$, which is not consistent; see also comments to (2.32).

Chapter 3

REGRESSION MODELS IN OPTIMAL MONITORING DESIGN

3.1 OPTIMALITY CRITERIA

3.1.1 Optimality of Measurements on a Network

Let us continue to work with the example from Sect. 2.1.1 and with regression model (2.1) and (2.2), that is,

$$E(y|x) = \theta^T x \quad \text{and} \quad \text{Var}\ (y|x) = \sigma^2(x) \quad . \tag{3.1}$$

We have learned in Sect. 2.1 that given $x_1, \ldots, x_n$ and $r_1, \ldots, r_n$ the best estimator for the vector θ can be easily constructed, if selection is done among linear unbiased estimators. Now we can think about the next optimization step: the selection of sites in which measurements provide most of the information. The latter statement is rather vague and is to be justified. First note that statement (2.8) means that the best linear unbiased estimator has the covariance matrix, which is least in the sense of Loewner's ordering [see comment to (2.10)], that is,

$$\underline{D} = Var(\hat{\theta}) \leq Var(\tilde{\theta}) = \underline{D} + \Delta \quad , \tag{3.2}$$

where $\tilde{\theta}$ is any other linear unbiased estimator of θ, and Δ is a non-negative definite matrix.

The Loewner optimization with respect to possible experimental designs ξ is defined as

$$\xi^* = \arg\min_{\xi} \underline{D}(\xi) \quad , \tag{3.3}$$

where the total number of available measurements $N = \sum_{i=1}^{n} r_i$ is assumed to be given. Unfortunately, in general (3.3) does not have a solution. Therefore, instead of this optimization problem we have to pursue a more modest objective and minimize some function(s) of the covariance matrix $\underline{D}(\xi)$. For instance, it can be the average variance of estimated components of θ : $\text{tr}\underline{D}(\xi) = \sum_{\alpha=1}^{m} Var[\hat{\theta}_\alpha(\xi)]$ and in the considered simple graph $m = 5$ (see Fig. 2.1). Thus, we can define an optimal design as

$$\xi^* = \arg\min_{\xi} \mathrm{tr}\underline{D}(\xi) \quad . \tag{3.4}$$

If the total number of measurements is fixed (given), then (3.4) is equivalent to

$$\xi^* = \arg\min_{\xi} \mathrm{tr} N M^{-1}(\xi) \quad ,$$

and consequently to

$$\xi^* = \arg\min_{\xi} \mathrm{tr} M^{-1}(\xi) \quad . \tag{3.5}$$

The latter optimization problem does not depend upon N explicitly and it is possible to find optimal designs that work for different N. Introducing the information matrix into the minimized expression helps to overcome a few theoretical difficulties in cases when an optimal design is singular, that is, rank $M(\xi^*) < m$. It may happen, for instance, when a practitioner is interested only in a subset of unknown parameters. In these cases, "inversion" in (3.5) must be replaced by "pseudo-inversion."

The formulation of the design problem is not complete yet: we did not define the set Ξ of all possible designs. If no repeated measurements are allowed, then Ξ for the network in Fig. 2.1 consists of all possible combinations of N routes out of 9 in the single-host problem and N out 19 in the multihost problem (see Table 2.1). We assume that a link direction is not important, that is, measurements on route 1-2 (a *ping* makes 1-2-1 round trip) and on route 2-1 (a *ping* makes 2-1-2 round trip) provide exactly the same information. Obviously, $N \leq 9$ in the first case and $N \leq 19$ in the second case.

Thus, if N is fixed then C_N^9 of C_N^{19} different designs must be compared using as the "quality" measure (criterion of optimality) $\mathrm{tr} M^{-1}(\xi)$. Of course, with the modern computer power, the optimization can be done very easily for that size of a network. However, the volume of direct computation may be prohibitive for the larger networks.

The optimization problem becomes simpler if we expand the design space Ξ. For instance, we can admit the possibility of the repeated measurements. Thus, in the single-host setting during every experimental session we may send $r_1, \ldots, r_9$ *pings* to site $1, \ldots, 4$ using different routes. Of course, $r_1 + \ldots + r_9 = N$. Usually to emphasize that there are only N available observations, the subscript N is added to Ξ. Now we can write that

$$\xi^* = \arg\min_{\xi \in \Xi_N} \mathrm{tr} M^{-1}(\xi) \quad , \tag{3.6}$$

and $\xi = \{p_i, x_i\}_1^9$, where all 9 possible routes, x_i, are presented in Table 2.1. Note the weights $p_i = r_i/N$ are discrete and an optimal design ξ^* still depends upon N.

If N is large enough and the discreteness of weights can be neglected, then (3.6) becomes a continuous (in the sense of weights) optimization problem, which can be considered within convex design theory. Noting that zero weights mean that the corresponding routes are not included in the experiment, one can look at the design problem as the selection of the most informative routes and the corresponding weights or fractions of *pings* sent along a selected route. In other words, we have to find the best $x_i \in X$, where X is a collection of all feasible routes (see the first 9 lines from Table 2.1), and the best p_i. The set X is frequently called a design region.

As in the earlier examples, the vectors x_i consist of m components that are either 1 (a link is included in the route, x_i) or 0 (a link is not included in the route x_i). At first glance that type of optimization problem coincides with the classical "spring-balance weighing design" problem [see, for instance, Raghavarao (1971), Chap. 17]. However, unlike to the classical case, in which the vector x can be any of 2^m combinations of zeros (for items not included in weighing) and ones (for items included in weighing), in our problem we have fewer feasible combinations of zeros and ones. In the considered simple network, it is only 9 instead of 32. Therefore, it is not possible to use well established, mature, and elegant combinatorial techniques to build optimal designs, and we have to resort to the convex design theory, which rarely leads to analytical results but provides numerical routines for optimal design construction. At this point it is expedient to note that construction of a design region, X, for large networks (i.e., all feasible routes or feasible combinations of zeros and ones) may be a difficult problem on its own; see Sect.3.5.

The optimality criterion (objective function) that was used in (3.4) – (3.6) is only one of many possible alternatives. For instance, if a practitioner is in a pessimistic mood, then minimization of $\max_\alpha Var\left[\hat{\theta}_\alpha(\xi)\right]$ can assure him/her that even for the worst case the selected design provides a reasonable result. Another alternative may be the minimization of

$$\max_x Var\left[\hat{\theta}^T(\xi)x\right] \quad ,$$

that is, we try to select a design that guarantees that estimators of delay times for any route have the variance that does not exceed some level.

3.1.2 Most Popular Criteria

Popularity of any given criterion may be explained by at least two reasons:

- Is it practically useful and expedient?
- Can the corresponding design/optimization problem be solved at moderate expense?

For various collections of optimality criteria see Atkinson and Donev (1992), Atkinson and Fedorov (1988), Bandemer et al. (1977), Box and Draper (1987), Box et al. (1978), Fedorov (1972), Fedorov and Hackl (1997), Pazman (1986), and Pukelsheim (1993). Table 3.1 contains a few criteria and this collection covers all our needs. Note that the table, together with optimality criteria, contains some additional entries that will be explained later. In the framework of examples from Sects. 2.1.1 and 3.1.1 the listed criteria from Table 3.1 may be commented as follows:

The first criterion is probably most celebrated in the statistical literature and usually is called D-criterion. A D-optimal design minimizes determinant $|D(\xi)|$ of the covariance matrix $D(\xi)$ (or equivalently, $\ln|D(\xi)|$ or maximizes $|M(\xi)|$ or $\ln|M(\xi)|$). The determinant $|D(\xi)|$ is called "the normalized generalized variance". The generalized variance is $|\underline{D}(\xi)| = N^{-m}|D(\xi)|$. In the case of normally distributed measurements y, the logarithm $-\ln N^{-m}|D(\xi)|$ coincides with the Shennon information gained in an experiment based on ξ. It may be noted that D-optimal designs minimize the volume of the concentration ellipsoid [cf. Fedorov (1972), Chap. 1.8].

The second criterion listed in the table is a modification of D-criteria when an experimenter is interested only in a subset of parameters. For instance, in the above example, one may wish to estimate mean delay times only for edges 2 and 3. Then

Table 3.1: Various optimality criteria $\Psi(\xi)$, their sensitivity functions $\phi(x,\xi)$ and majorization constants C.

$\Psi(\xi)$	$\phi(x,\xi)$	C
$\log\lvert D\rvert$	$d(x) = f^T(x)Df(x)$	m
$\log\lvert D_\ell\rvert^*$	$d(x) - d_k(x)$ $d_k(x) = f_k^T(x)D_k f_k(x)$, $f^T(x) = (f_\ell^T(x), f_k^T(x))$	ℓ
$\mathrm{tr}AD,\ A \geq 0$	$f^T(x)DADf(x)$	$\mathrm{tr}AD$
$d(x_0)$	$d^2(x,x_0)$ $d(x,x_0) = f^T(x)Df(x_0)$	$d(x_0)$
$\int_Z d(x)dx$	$\int_Z d^2(x,z)dz$	$\int_Z d(x)dx$
$\lambda_{\max}(D)$	$\sum_{i=1}^{\alpha}\pi_i(f^T(x)P_i)^2$ $\lambda_{\min}P_i = MP_i$, α is a multiple of $\lambda_{\min}$, $\sum_{i=1}^{\alpha}\pi_i = 1,\ 0 \leq \pi_i \leq 1$	$\lambda_{\max}$
$\mathrm{tr}D^\gamma$	$f^T(x)D^{\gamma+1}f(x)$	$\mathrm{tr}D^\gamma$

* D_ℓ is a submatrix of D corresponding to ℓ parameters, $k = m - \ell$.

$$D_\ell = \begin{pmatrix} D_{22} & D_{23} \\ D_{32} & D_{33} \end{pmatrix} \quad .$$

The third line of the table describes the linear criteria. Matrix A is called elements of the utility matrix. For instance, if edges 2 and 3 are of interest, then

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad .$$

If an experimenter believes that the importance of various parameters can be described by some weights w_α, then $A_{\alpha\beta} = \delta_{\alpha\beta}w_\alpha$ may be a good choice for the utility matrix.

The next two entries deal with the response estimator variance and are the particular cases of the linear criteria. Indeed, from (2.20) we know that

$$d(x) = N^{-1}\underline{d}(x) = N^{-1}f^T(x)Df(x)$$

and consequently the variance of the function $\theta^T f(x)$ at the point of interest x_0 is

$$d(x_0) = f^T(x_0)Df(x_0) = \mathrm{tr}f(x_0)f^T(x_0)D = \mathrm{tr}AD \quad ,$$

where $A = f(x_0)f^T(x_0)$, $\mathrm{rank}A = 1$. For the average (over the set Z) variance of estimated response function we have

$$\int_Z d(x)dx = \int_Z f^T(x)Df(x)dx = \mathrm{tr}D \int_Z f(x)f^T(x)dx = \mathrm{tr}AD \quad ,$$

where $A = \int_Z f(x)f^T(x)dx$. In the latter case, usually rank $A = m$, whereas before m is a total number of unknown parameters. Whenever rank A is less than the total number of unknown parameters, some caution is necessary: an optimal design ξ^* could be singular [i.e., rank$\underline{M}(\xi^*) < m$]. To handle these cases, one either has to use pseudo-inverse matrices $M^-(\xi)$ or some kind of regularization. For instance, the first of the above utility matrices can be replaced by $A = f(x_0)f^T(x_0) + \gamma I$, where γ is a small positive constant and I is the identity matrix [cf. Fedorov and Hackl (1997), Chap. 2.6, and Pukelsheim (1993), Chap. 7].

In the network monitoring setting, one may wish to minimize the (normalized) variance $d(x_0)$ when, for instance, it is necessary to get the most precise information about a particular route x_0. The average variance criterion can be used if, for instance, the behavior of a response function over some time interval is of interest, and the "time" variable is included in the regression model. The popular spatial averaging must be replaced by summation over the selected set of routes, that is,

$$A = \sum_{i=1}^{k} f(x_i)f^T(x_i) \quad ,$$

and $x_1, \ldots, x_k$ is the set of routes or nodes about which we want to learn most.

The last two criteria listed in Table 3.1 are frequently used in theoretical exercises. Minimization (E-optimality) of the largest eigenvalue $\lambda_{\max}(D)$ of the dispersion matrix D results in the ellipsoid of concentration with the least maximal principal axis. Noting that [cf. Rao (1973), Chap. 1c]

$$\lambda_{\max} = \arg\max_{\lambda} \frac{\lambda^T D \lambda}{\lambda^T \lambda} \quad ,$$

we conclude that

$$\frac{f^T(x)Df(x)}{f^T(x)f(x)} < \lambda_{\max}$$

for any route x and, therefore, E-optimal designs assure a low upper bound on the variance of the response function at point x. In terms of Sects. 2.1.1 and 3.1.1 with $f(x) = x$ for every feasible route x

$$d(x) = x^T Dx \leq q\lambda_{\max} \quad ,$$

where $q = x^T x$ is the "length" of the route x or the total number of edges included in x.

The concluding criterion in Table 3.1 is popular because

$$(m^{-1}\mathrm{tr}D^\gamma)^{1/\gamma} = m^{-1}\mathrm{tr}D \text{ (average variance)}, \quad \gamma = 1;$$

$$\lim_{\gamma \to 0}(m^{-1}\mathrm{tr}D^\gamma)^{1/\gamma} = |D|^{1/m} \ (D - \text{criterion}) \quad ;$$

$$\lim_{\gamma\to\infty}(m^{-1}\mathrm{tr}D^{\gamma})^{1/\gamma} = \lambda_{\max}\ (E-\text{criterion}) \quad .$$

Thus, analyzing one criterion we get results for three most widely used criteria. In general, the optimal monitoring design problem can be formulated as

$$\xi = \arg\min_{\xi\in\Xi}\Psi(\xi) \quad ,$$

where the function Ψ may coincide with any of the discussed optimality criteria and Ξ is the set of all feasible designs.

3.2 PROPERTIES OF OPTIMAL DESIGNS

Optimality criteria, which were discussed in Sect. 3.1, may be used in various settings including nonlinear models and models with the correlated observations. However, the careful analysis of the design space Ξ is needed for every specific case. For instance, in regression problems with observational errors generated by some auto-regression schemes (see, for instance, Kiefer and Wynn (1984)) the concept of repeated observations is not very useful and Ξ contains only designs without repeated observations [i.e., $r_1 = \ldots = r_N = 1$]. In this section we are concerned with regression models with independent observation, i.e., with models described by (2.4) and (2.5). To make the notations and formulae simpler, we assume that $\sigma^2(x) \equiv 1$.

For all results of this section, it is essential that the information matrix $M(\xi)$ can be presented as a sum of information matrices of individual observations; see (2.21) and (2.22). Another important assumption is that the function $\Psi(\xi)$ introduced in (3.7) is convex, that is,

$$\Psi\left[(1-\alpha)\xi_1 + \alpha\xi_2\right] \le (1-\alpha)\Psi(\xi_1) + \alpha\Psi(\xi_2),\ \ 0 \le \alpha \le 1 \quad ,$$

and has a directional derivative $\psi(\xi^*,\xi)$ at ξ^* for any $\bar{\xi} = (1-\alpha)\xi^* + \alpha\xi,\ \ 0 \le \alpha < 1$.

If the above assumptions hold, then:

Theorem 3.1 *A necessary and sufficient condition for a design ξ^* to be optimal is fulfillment of the inequality*

$$\psi(\xi^*,\xi) \ge 0$$

for any feasible design $\xi \in \Xi$.

This result is widely used in experimental design theory [Cook and Fedorov (1995), Fedorov and Hackl (1997), Chap. 2.3] and is essential for the proof of Theorem 3.2. Note that all the criteria from Table 3.1 satisfy conditions of the above theorem.

To formulate one of the main theorems of experimental design theory we have to complement the above by the assumption about boundness of functions $f(x)$ (or individual matrices $m(x)$) on X and assume that Ξ consists of all possible probability distributions (measures) on X. With this additional assumption, we can state for all the criteria from Table 3.1 the following theorem holds:

Theorem 3.2 *1. A necessary and sufficient condition for ξ^* to be optimal is that*

$$\phi(x, \xi^*) \leq C(\xi^*)$$

where the functions ϕ and the constants C are defined in Table 3.1.

2. If the design ξ^* has nonzero measure on $X' \subset X$, then the function $\phi(x, \xi^*)$ reaches its upper bound $C(\xi^*)$.

3. The optimization problems,

$$\xi^* = \arg\min_{\xi} \Psi(\xi)$$

and

$$\xi^* = \arg\min_{\xi} \max_{x \epsilon X} \phi(x, \xi)$$

are equivalent (i.e., have identical sets of solutions).

4. There exists an optimal design with no more than $m(m+1)/2$ supporting points.

The proof of this theorem and its various modifications can be found in Fedorov and Hackl (1997), Chap. 2. We recommend this source only because it contains the closest formulation of the above results. There are a number of books on experimental design theory and applications that contain alternative formulations and proofs; among them Bandemer et al. (1977), Fedorov (1972), Pazman (1986), Pukelsheim (1993), and Silvey (1980). Note that the sensitivity function $\phi(x, \xi)$, constant $C(\xi^*)$, and the directional derivative $\psi(\xi^*, \xi)$ are related as

$$\psi(\xi^*, \xi) = C(\xi)^* - \int \phi(x, \xi^*)\xi(d\alpha) \quad . \tag{3.7}$$

So far, we have assumed that $\sigma^2(x) \equiv 1$. If it is not so, then in all formulae, $f(x)$ must be replaced by $\sigma^{-1}(x)f(x)$. In particular, every element in the second column [corresponding to $\phi(x, \xi)$] of Table 3.1 must be divided by $\sigma^2(x)$ to be used in Theorem 3.2.1 or in the numerical procedures considered later.

The second section of the theorem partly explains why the function $\phi(x, \xi^*)$ is called "sensitivity function". All observations are recommended to be placed at sites, in which the value of this function is maximal. Later, discussing the numerical procedure, we will see that a design ξ can be improved if observations from sites with the low $\phi(x, \xi)$ are relocated to sites with the higher values of $\phi(x, \xi)$.

The third section of the theorem leads to very useful results: it establishes the equivalence of some optimality criteria. For instance, the designs, which minimize the generalized variance $|D|$ of the parameter θ, also mimimize the maximal (over the design region X) variance of the regression function estimator. Hence, at least in some cases, a practitioner can avoid a painful process of optimality criterion selection. D-criterion is also equivalent to some criteria used in model testing [cf. Fedorov and Hackl (1997), Chap. 5, Fedorov and Khabarov (1986), and Kiefer (1958)].

The final section concludes that for any design problem with an optimality criterion included in Table 3.1, there exists a solution with a finite number of support points (sites). It means that a reader who is not comfortable with the Stiltjes integral, can replace all integrals in the above and following discussions by finite sums. Actually, the boundary $m(m+1)/2$ is too high for practical

needs and in many applications there exist optimal designs with the number of support points that is equal or moderately more than m.

Theorem 3.2 is very helpful in the construction of optimal designs in cases when the set of basis functions $f(x)$ contains some simple components, for instance, $f^T(x) = (1, x_1, x_2, ..., x_{m-1})$ in a multivariate case, or $f^T(x) = (1, x, x^2, ...x^{m-1})$ in a univariate case, or $f^T(x) = (1, \sin x,$ $\cos x,\ ...,\ \sin kx,\ \cos kx,\ \ m = 2k+1)$, and the design region, X, has a regular structure (interval, cube, or sphere). Most of those type design problems have been explored, optimal design have been constructed and tabulated. From our experience, there are not too many problems in optimal monitoring, which belong to the above simple realm.

Theorem 3.2 and correspondingly Table 3.1 do not assume the presence of prior information about estimated parameters. In our examples and studies, prior information expressed in terms of a prior covariance matrix $D_0 = \sigma^{-2} N \underline{D}_0$ may be available. It is not difficult to adapt Table 3.1 and Theorem 3.2.1 for cases with prior information. For instance, for D-criterion

$$\phi(x, \xi) = d_{tot}(x, \xi) = f^T(x) D_{tot}(\xi) f(x) \quad ,$$

$$C(\xi) = \mathrm{tr} D_{tot}(\xi) M(\xi) \quad ,$$

and for linear criteria

$$\phi(x, \xi) = f^T(x) D_{tot}(\xi) A D_{tot}(\xi) f(x) \quad ,$$

$$C(\xi) = \mathrm{tr} D_{tot}(\xi) A D_{tot}(\xi) M(\xi) \quad ,$$

where $D_{tot}^{-1}(\xi) = M_{tot}(\xi) = M(\xi) + D_o^{-1}$. The subscript "*tot*" emphasizes that the corresponding functions or matrices include prior information and information from the experiment based on the design ξ.

3.3 NUMERICAL METHODS

3.3.1 The First Order Algorithms

In this study, we use only the first order algorithms, which are based on the linear approximation of the criterion $\Psi(\xi)$ in the vicinity of any intermediate point ξ_s [cf. Bandemer et al. (1977), Cook and Nachtsheim (1980), Fedorov and Uspensky (1975), Nguyen and Miller (1992)]. The discussions of methods of higher orders can be found, for instance, in Fedorov and Hackl (1997), Chap. 3.2; Gaffke and Heiligers (1996); and Gaffke and Mathar (1992). These methods provide numerical results with higher precision, but in their present form look impractical for the large size monitoring problems.

The main idea of the first order algorithms is to find a correction design $\bar{\xi}$ that leads to the greatest decrement of the optimality criteria, that is, to find

$$\overline{\xi} = \arg\min_{\xi} \Psi\left[(1-\alpha)\xi_s + \alpha\xi\right] \quad . \tag{3.8}$$

Assuming that the step α is small enough to neglect terms of the higher order than $s(\alpha)$, we come to [see (3.7)]:

$$\overline{\xi} = \arg\min_{\xi\in\Xi}[\Psi(\xi_s) + \alpha\psi(\xi_s,\xi)] \;\cong\; \arg\min_{\xi\in\Xi}\left[C(\xi_s) - \int \phi(x,\xi_s)\xi(dx)\right] \quad . \tag{3.9}$$

Note that the optimization problem

$$\overline{\xi} = \arg\max_{\xi\in\Xi}\int \phi(x,\xi_s)\xi(dx)$$

has a solution that coincides with a design atomized at the point

$$x_s = \arg\max_{x\in X}\phi(x,\xi_s) \quad .$$

Therefore, one can construct the following simple iterative procedure:

1. Given ξ_s, find

$$x_s = \arg\max_{x\in X}\phi(x,\xi_s) \quad .$$

2. Construct

$$\xi_{s+1} = (1-\alpha_s)\xi_s + \alpha_s\xi(x_s) \quad ,$$

   where the design $\xi(x_s)$ has weight 1 at point x_s.

3. Compare $\phi(x_s,\xi_s)$ with $C(\xi_s)$, and if it is close enough (see Theorem 3.2), then stop computation. Otherwise go back to step 1.

The above procedure requires justification for theoretical analysis and for its implementation in practice.

It is interesting to note that for D-criterion at each step the iterative procedure recommends to add some additional weight to a point where the variance of an estimated response function is largest. In other words, place more observations at sites where you know less.

3.3.2 Practical Algorithms and Pilot Software

In our pilot studies, we used the following version of the iterative procedure for optimal monitoring design:

1. Given a design ξ_s, $|M(\xi_s)| > 0$. Find

$$x_s = \arg\max_{x\in X}\{\phi(x^+,\xi_s),\ \phi(x^-,\xi_s)\}$$

   where

$$x^+ = \arg\max_{x\in X}\phi(x,\xi_s) \quad ,$$

and

$$x^- = \arg\min_{x \in X_s} \phi(x, \xi_s) \quad .$$

The set X_s contains all support points of design ξ_s.

2. Given α_s, x_s and ξ_s construct

$$\xi_{s+1} = (1 - \alpha_s)\xi_s + \alpha_s \xi(x_s) \quad .$$

3. Compare $\Psi(\xi_s)$ and $\Psi(\xi_{s+1})$. If $\Psi(\xi_s) - \Psi(\xi_{s+1}) > 0$ select $\alpha_{\alpha+1} = \alpha_s$ and continue with step 4. Otherwise, select $\alpha_{s+1} = \gamma\alpha_s$, where $0 < \gamma < 1$ (usually $\gamma = \frac{1}{2}$), and continue with step 4.

4. If $\phi(x^+, \xi_s) - C(\xi_s) \leq \beta$, where β is some small preselected constant, stop calculations. Otherwise, continue with step 1, given ξ_{s+1} .

The above procedure converges for any criterion from Table 3.1 but the fourth one (E-criterion); see details in Atkinson and Donev (1992), Chap. 4; Fedorov (1972) Chaps. 3 and 4; Fedorov and Hackl (1997), Chap. 3; and Pilz (1991), Chap. 12. Note that the "practical" version of the algorithm contains opportunity for direct deleting of the "bad" points. It happens when $x_s = x^-$. In the cited publications, other choices of the sequence $\{\alpha_s\}$ are discussed. One of them is $\alpha_s = (N_0+s)^{-1}$, where N_0 is a member of support points in ξ_0. When there is no prior information, some simple recursions can be applied to helping avoid multiple inversions of information matrices and calculations of their determinants:

$$(1 - \alpha_s)D(\xi_{s+1}) = D(\xi_s) - \frac{\alpha\gamma(x, \xi_s)\gamma^T(x, \xi_s)}{1 - \alpha_s + \alpha d(x, \xi_s)} \quad , \tag{3.10}$$

$$|D(\xi_{s+1})| = \frac{|D(\xi_s)|}{(1 - \alpha)^{m-1}(1 - \alpha + \alpha d(x, \xi_s))} \quad , \tag{3.11}$$

where $d(x, \xi) = f^T(x)D(\xi)f(x)$, $D(\xi) = M^{-1}(\xi)$, and $\gamma(x, \xi) = D(\xi)f(x)$. In the presence of prior information, the analogues of (3.10) and (3.11) become tediously long and we did not use them in our software. Actually, we used a modified version of procedure 1−4, which is called the "exchange" algorithm [cf. Mitchell (1974), Nguyen and Miller (1992)]:

1. There is a design ξ_s, $|M(\xi_s)| > 0$. Find

$$x_s^+ = \arg\max_{x \in X} \phi(x, \xi_s) \quad .$$

Construct matrix

$$M_s^+ = M(\xi_s) + \alpha_s f(x_s^+) f^T(x_s^+) \quad ,$$

and find

$$x_s^- = \arg\min_{x \in X_s} \phi^+(x, \xi_s) \quad ,$$

where the function $\phi^+(x, \xi_s)$ is defined exactly as $\phi(x, \xi_s)$, but the matrix $M(\xi_s)$ must be replaced by M_s^+.

2. Add weight α_s to point x_s^+, and delete exactly the same weight from point x_s^-. Call new design ξ_{s+1}.

Steps 3 and 4 are identical to the original procedure.

The sequence $\{\alpha_s\}$ may be selected as it was done before with one exception. Whenever weight at the point to be deleted is less than α_s, the latter must be reset to be equal to that weight. The first stage of step 1 must be redone with the new α_s or all the weights must be rescaled to make their sum equal to 1. The recursion formulas can be easily derived and are

$$(M_s^+)^{-1} = M^{-1}(\xi_s) - \frac{\alpha_s \gamma(x^+, \xi_s)\gamma^T(x^+, \xi_s)}{1 + \alpha_s d(x^+, \xi_s)} \quad , \tag{3.12}$$

$$|M_s^+| = \frac{|M(\xi_s)|}{1 + \alpha_s d(x^+, \xi_s)} \quad . \tag{3.13}$$

For the deleting step in expressions for $\gamma(x^+, \xi_s)$ and $d(x^+, \xi_s)$, the matrix $D(\xi_s)$ must be replaced by $(M_s^+)^{-1}$. It is desirable to provide to a practitioner the capability of choosing the "length of excursions" (i.e., the number of consecutive additions and deletions of points). In the above presentation this length is equal to 1. More technical details on the pilot software developed at ORNL can be found in Flanagan (1997).

The above two procedures differ for the cases with prior information and lead in many cases to almost identical results (i.e., to the same final precision and computational time). Indeed, recursions (3.10) and (3.11) cannot be used if the matrix $D(\xi)$ is composed from prior and newly gained information, that is, $D(\xi_s) = [M_0 + M(\xi_s)]^{-1}$. The reason for that is technical: at each step we decrease all elements of the matrix $D(\xi_s)$. Applying (3.10) to $[M_0 + M(\xi_s)]^{-1}$, we change the structure of the combined matrix $M_0 + M(\xi_s)$ through multiplication of the latter by $(1 - \alpha_s)$ when we introduce the design $\xi_{s+1} = (1 - \alpha_s)\xi_s + \alpha\xi(x_s)$. At the same time, constructing

$$M_s^+ = M_{tot}(\xi) + \alpha_s f(x_s^+) f^T(x_x^+) = M_0 + M(\xi_s) + \alpha f(x_s^+) f^T(x_s^+)$$

and using (3.12), we leave the prior matrix M_0 unchanged. Thus, the exchange algorithm together with (3.12) provides opportunity to work with prior information or, what is even more important for us, to use some regularization techniques (see Sect.3.5), in which the original problem of minimization of $|D(\xi)| = |M(\xi)|^{-1}$ is replaced by minimization of $|\gamma I + M(\xi)|^{-1}$, where γ is some adjustable small positive number.

In the pilot Fortran program used in various examples for network monitoring, the following inputs are necessary:

a. *a priori* information (or covariance) matrix M_0 (or D_0);

b. initial design ξ_0 with $|M(\xi_0)| > 0$;

c. transformed design set

$$F = \{f : \ f(x), x \in X\} \quad ;$$

d. step reduction constant γ, initial α_0, stopping rule constant β and/or the maximal number of iterations.

The output includes:

a. final value of the optimality criterion $\Psi(\xi)$ and $\max_x \phi(x, \xi^*)$;

b. final design ξ^* and matrices $M(\xi^*)$, $D(\xi^*)$, $M_{tot}(\xi^*)$, $D_{tot}(\xi^*)$.

3.4 EXAMPLES

3.4.1 Optimal Designs for Simple Networks

Let us continue consideration of the example from Sect.2.1 in its simplest single host version and $\sigma^2(x_i) \equiv 1$. All feasible routes are given in Table 2.1 (first 9 rows). The use of the exchange algorithm with $M_0 = 0$ gives a D-optimal allocation of weights and sites to be "*pinged*", which is presented in Table 3.2 (second column). The calculations were stopped when

$$\max_x \phi(x, \xi_s) - C(\xi) = \max_x d(x, \xi_s) - m \leq 0.01.$$

If we consider weights as fractions of experimental time, then most of the available time must be spent at routes p_3 and p_5 while measurements at route p_7 are less informative, and, therefore, this route is not included in the optimal design at all. One can note that we have a very moderate improvement comparatively to the uniform design (first column, it was used as the initial design for the iterative procedure). An explanation is a rather simple. There are not very many feasible routes to choose from; we have 5 unknown parameters and only 9 feasible routes. From Table 3.2 it is easy to see that the deletion of routes with smaller weights and rounding of weights leads to a deterioration of the design characteristics (see columns 4 and 5). But the most drastic deterioration occurs when we leave only 5 routes. It is interesting to point out that the last two columns correspond to two different attempts made with the exchange algorithm (with different ξ_0). In both cases, $\alpha \equiv 1/5$. In general, the exchange algorithm does not converge to D-optimal discrete design. We recommend making several attempts to get the best design with weights proportional to α (which does not change!).

3.4.2 Multihost Experiments vs One-host Experiments

For multihost experiments and $\sigma^2(x) \equiv 1$ on the same network, the design set X contains 19 routes (i.e., we have a richer choice comparatively to the single host case). The value of $|D(\xi^*)|$ drops to ~ 100 compared to ~ 450 in the one host situation; compare Tables 3.2 and 3.3. Thus, the multihost experiment is better, and the results confirm the common sense conclusion that cooperation is good.

The maximal (among all 19 routes) value of the response variance $d(x, \xi^*)$ has the same value as in the single host case and equals ~ 5 (see Sect.1 of Theorem 3.2 and recollect that $C(\xi^*) = m = 5$), that is, to the number of estimated parameters. However, do not forget that this maximum is over the larger number of feasible routes: 19 vs 9.

The last three columns in Table 3.3 show the designs built with an exchange algorithm with fixed α. Starting from 7 support points this algorithm produces reasonable results both in terms of the

Table 3.2: Experimental designs for the single-host case of the network example with 4 nodes and 5 edges.

	I Uniform	II D-optimal Continuous	III Rounded 6 points	IV Exchange-Type 6 points	V Exchange-Type 5 points (a)	VI Exchange-Type 5 points (b)
N	9	9	9	9	9	9
n	9	8	6	6	5	5
p_1	1/9	.14	.15	1/6	0	1/5
p_2	1/9	.14	.15	1/6	1/5	1/5
p_3	1/9	.18	.20	1/6	0	1/5
p_4	1/9	.14	.15	1/6	1/5	1/5
p_5	1/9	.18	.20	1/6	0	1/5
p_6	1/9	.14	.15	1/6	0	0
p_7	1/9	0	0	0	1/5	0
p_8	1/9	.04	0	0	1/5	0
p_9	1/9	.04	0	0	1/5	0
$\|D\|$	563	450	463	486	3125	781
$\max_i d(x_i, \xi)$	5.91	5.01	5.83	6.00	15.0	15.0

determinant $|D|$ and response variance values. Interestingly, all the constructed designs "avoid" routes with a single edge (i.e., routes 2, 4, 7, and 10).

3.4.3 Measurement Errors Depending on the Route Length

In both considered cases we make a strong assumption that the variance of the observed variable is identical for all feasible routes, i.e., $\sigma^2(x) \equiv 1$. Selecting it to be equal to 1 is, of course, the matter of scaling. Let us explore situations when the variance of the observed variable depends upon a route. Similar to Sect.2.1.1, we assume that $\sigma^2(x) = x^T x = \ell$, where ℓ is a number of nonzero components of the vector x. Unlike that example, we assume that all observations are uncorrelated.

Our two choices of the measurement variances are quite extreme ones. The constant variance of measurements could be a good approximation of reality if all variations in delay times for *pings* occur due to various activities at the destination sites (nodes), and all sites can be considered similar. The cumulative model with $\sigma^2(x) = x^T x$ works if variations of delay times may be explained by activities on each edge (link), and all these edges have similar technical characteristics.

We repeated most calculations made for the model with the constant variance to verify how changes in the model can influence the structure of optimal designs. Note that in all calculations, function $f(x)$ must be replaced by $\sigma^{-1}(x)f(x)$. Previously, the $\max d(x, \xi)$ must be close to $m = 5$ for the computed optimal designs. Now, $\sigma^{-2}(x)d(x, \xi)$ must be close to the same number.

Tables 3.4 and 3.5 contain information similar to Tables 3.2 and 3.3.

There exist no remarkable changes for the single host problem. However, for the multiple host problem with a richer choice of optimal routes, the changes are dramatic. For $\sigma^2(x) \equiv 1$, all the "shortest" routes (e.g., 2, 4, 7, 10, 17) are not in the support set of the D-optimal design; for

Table 3.3: Experimental designs for the multi-host case of the network (4,5) example.

	I	II	III	IV	V	VI
	Uniform	D-optimal	Rounded	Exchange-Type		
		Continuous	14 points	5 points	7 points (a)	7 points (b)
n	9	14	14	5	7	7
p_1	1/19	.07	1/14	0	1/7	0
p_2	1/19	0	0	0	0	0
p_3	1/19	.05	1/14	0	0	1/7
p_4	1/19	0	0	0	0	0
p_5	1/19	.05	1/14	0	1/7	0
p_6	1/19	.07	1/14	0	1/7	1/7
p_7	1/19	0	0	0	0	0
p_8	1/19	.06	1/14	1/5	0	0
p_9	1/19	.06	1/14	1/5	1/7	1/7
p_{10}	1/19	0	0	0	0	0
p_{11}	1/19	.05	1/14	0	0	0
p_{12}	1/19	.07	1/14	0	0	0
p_{13}	1/19	.06	1/14	0	1/7	0
p_{14}	1/19	.06	1/14	1/5	0	1/7
p_{15}	1/19	.12	1/14	1/5	1/7	1/7
p_{16}	1/19	.12	1/14	1/5	1/7	1/7
p_{17}	1/19	0	0	0	0	0
p_{18}	1/19	.05	1/14	0	0	0
p_{19}	1/19	.07	1/14	0	0	1/7
$\|D\|$	193	100	105	195	127	127
$\max_i d(x_i, \xi)$	6.76	5.00	5.95	15.0	6.84	6.84

$\sigma^2(x) = x^T x$, the D-optimal design consists entirely of the "shortest" routes! A rather simple lesson might be learned by the accumulated experience. If the variance of observation does not differ too much for different routes, then select the longest (but not having too many edges in common) routes; if the variance increases noticeably with the increase of the route length, then select the shortest routes.

Table 3.4: Experimental designs with $\sigma^2(x) = x^T x$ for single-host case network (4,5).

	I Uniform	II D-optimal Continuous	III Rounded 6 points	IV Exchange-Type 5 points (a)	V Exchange-Type 5 points (b)
n	9	9	6	6	5
p_1	1/9	.07	0	1/5	0
p_2	1/9	.17	1/6	1/5	1/5
p_3	1/9	.14	1/6	1/5	1/5
p_4	1/9	.17	1/6	1/5	1/5
p_5	1/9	.14	1/6	1/5	1/5
p_6	1/9	.07	1/6	0	1/5
p_7	1/9	.12	1/6	0	0
p_8	1/9	.06	0	0	0
p_9	1/9	.06	0	0	0
$\lvert D \rvert$	7589	6214	7775	9375	9375
$\max_i \frac{d(x_i,\xi)}{\sigma^2(x_i)}$	6.66	5.00	8.00	10.00	10.00

Table 3.5: Experimental designs with $\sigma^2(x) = x^T x$ for multi-host case of network (4,5).

	I Uniform	II D-optimal Continuous	III Exchange-Type 6 points (a)	IV Exchange-Type 6 points (b)
n	19	5	6	6
p_1	1/19	0	0	0
p_2	1/19	1/5	1/6	1/6
p_3	1/19	0	0	0
p_4	1/19	1/5	1/6	1/6
p_5	1/19	0	0	0
p_6	1/19	0	1/6	0
p_7	1/19	1/5	0	1/6
p_8	1/19	0	0	0
p_9	1/19	0	1/6	0
p_{10}	1/19	1/5	0	1/6
p_{11}	1/19	0	0	0
p_{12}	1/19	0	0	0
p_{13}	1/19	0	0	0
p_{14}	1/19	0	1/6	0
p_{15}	1/19	0	0	0
p_{16}	1/19	0	0	0
p_{17}	1/19	1/5	1/6	1/6
p_{18}	1/19	0	0	0
p_{19}	1/19	0	0	1/6
$\lvert D \rvert$	5251	3125	3888	3888
$\max_i \frac{d(x_i,\xi)}{\sigma^2(x_i)}$	5.99	5.00	6.00	6.00

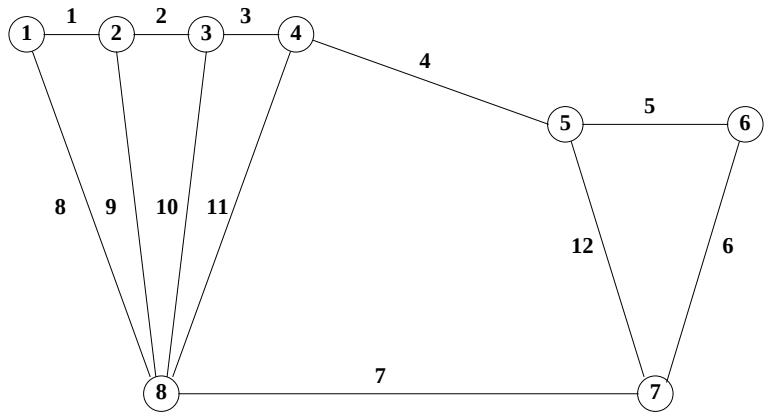

Figure 3.1: Network graph with 8 nodes and 12 edges.

The picture is less clear if the network to monitor is more complex and the setting is neither single-host nor multihost. To illuminate that, we considered the network presented in Fig. 3.1, which is still relatively simple. It was assumed that only two hosts can do measurements. The feasible routes for *pings* are listed in Table 3.6. The computed designs are presented in Tables 3.7 and 3.8 for cases with $\sigma^2(x) \equiv 1$ and $\sigma^2(x) = x^T x$, correspondingly.

Whenever we constructed discrete designs with the small number of support points, using exchange algorithm, we started with a "small" *a priori* information matrix M_0, $|M_0| > 0$. The reason for this was the high probability of getting a singular initial design if the latter is generated at random. After a few iterations, the calculations were terminated, and the design constructed at the latest iteration was used as the initial design for the original problem, in which $M_0 = 0$. That trick can be considered as some kind of regularization of possibly singular initial designs. In general, starting with $M_0 = \gamma I$, where γ some positive constant, we avoid singularities in the iteration procedure, which are frequent (at least in our experience) for even moderately complex networks. Computations may be repeated with diminishing γ and with the use of designs constructed with the larger γ-s as initial designs for the smaller γ-s.

Table 3.6: Routes for the network (8,12) with two hosts.

From-To Hosts	Route Number	Edges											
		1	2	3	4	5	6	7	8	9	10	11	12
1 - 3	1	1	1	0	0	0	0	0	0	0	0	0	0
1 - 3	2	0	0	0	0	0	0	0	1	0	1	0	0
1 - 3	3	0	1	0	0	0	0	0	1	1	0	0	0
1 - 3	4	0	0	1	0	0	0	0	1	0	0	1	0
1 - 6	5	0	0	0	0	0	1	1	1	0	0	0	0
1 - 6	6	1	1	1	1	1	0	0	0	0	0	0	0
1 - 6	7	0	0	0	0	1	0	1	1	0	0	0	1
1 - 2	8	1	0	0	0	0	0	0	0	0	0	0	0
1 - 2	9	0	0	0	0	0	0	0	1	1	0	0	0
1 - 2	10	0	1	0	0	0	0	0	1	1	1	0	0
1 - 2	11	0	1	1	0	0	0	0	1	0	0	1	0
1 - 4	12	0	0	0	0	0	0	0	1	0	0	1	0
1 - 4	13	1	1	1	0	0	0	0	0	0	0	0	0
1 - 4	14	1	0	0	0	0	0	0	0	1	0	1	0
1 - 4	15	1	0	0	0	0	0	0	0	0	1	1	0
1 - 4	16	0	1	1	0	0	0	0	1	0	1	0	0
1 - 4	17	0	0	0	1	0	0	1	1	0	0	0	1
1 - 5	18	1	0	0	1	0	0	0	0	1	0	1	0
1 - 5	19	1	1	1	1	0	0	0	0	0	0	0	0
1 - 5	20	0	0	0	1	0	0	0	1	0	0	1	0
1 - 5	21	0	0	0	0	0	0	1	1	0	0	0	1
1 - 5	22	0	0	0	0	1	1	1	1	0	0	0	0
1 - 5	23	1	0	1	1	0	0	0	0	1	1	0	0
1 - 5	24	1	0	0	0	1	1	1	0	1	0	0	0
1 - 7	25	0	0	0	0	0	0	1	1	0	0	0	0
1 - 7	26	0	0	0	1	0	0	0	1	0	0	1	1
1 - 7	27	1	0	0	0	0	0	1	0	1	0	0	0
1 - 8	28	0	0	0	0	0	0	0	1	0	0	0	0
1 - 8	29	1	0	0	0	0	0	0	0	1	0	0	0
1 - 8	30	1	1	0	0	0	0	0	0	0	1	0	0
1 - 6	31	0	0	0	1	1	0	0	1	0	0	1	0
1 - 6	32	0	0	0	1	0	0	0	1	0	0	1	1
1 - 7	33	1	1	0	0	0	0	1	0	0	1	0	0
1 - 7	34	1	1	1	1	0	0	0	0	0	0	0	1
1 - 7	35	1	1	1	0	0	0	1	0	0	0	1	0
1 - 7	36	1	1	1	1	1	1	0	0	0	0	0	0
1 - 8	37	1	1	1	0	0	0	0	0	0	0	1	0
1 - 8	38	1	1	1	1	0	0	1	0	0	0	0	1
4 - 2	39	0	1	1	0	0	0	0	0	0	0	0	0
4 - 2	40	0	0	0	0	0	0	0	0	1	0	1	0
4 - 2	41	0	1	0	0	0	0	0	0	0	1	1	0
4 - 2	42	1	0	0	0	0	0	0	1	0	0	1	0
4 - 2	43	0	0	1	0	0	0	0	0	1	1	0	0
4 - 2	44	1	0	1	0	0	0	0	1	0	1	0	0
4 - 3	45	0	0	1	0	0	0	0	0	0	0	0	0
4 - 3	46	0	0	0	0	0	0	0	0	0	1	1	0
4 - 3	47	0	1	0	0	0	0	0	0	1	0	1	0
4 - 3	48	0	0	0	1	0	0	1	0	0	1	0	1
4 - 3	49	0	1	0	1	0	0	1	0	1	0	0	1
4 - 3	50	1	1	0	0	0	0	0	1	0	0	1	0
4 - 5	51	0	0	0	1	0	0	0	0	0	0	0	0
4 - 5	52	0	0	0	0	0	0	1	0	0	0	1	1
4 - 5	53	0	0	1	0	0	0	1	0	0	1	0	1
4 - 5	54	0	1	1	0	0	0	1	0	1	0	0	1
4 - 5	55	0	0	0	0	1	1	1	0	0	0	1	0
4 - 6	56	0	0	0	1	1	0	0	0	0	0	0	0
4 - 6	57	0	0	0	0	0	1	1	0	0	0	1	0
4 - 6	58	0	0	0	0	1	0	1	0	0	0	1	1
4 - 6	59	0	0	1	0	0	1	1	0	0	1	0	0
4 - 6	60	0	0	0	1	0	1	0	0	0	0	0	1
4 - 7	61	0	0	0	1	0	0	0	0	0	0	0	1
4 - 7	62	0	0	0	0	0	0	1	0	0	0	1	0
4 - 7	63	0	0	0	1	1	1	0	0	0	0	0	0
4 - 7	64	0	0	1	0	0	0	1	0	0	1	0	0
4 - 7	65	0	1	1	0	0	0	1	0	1	0	0	0
4 - 7	66	1	1	1	0	0	0	1	1	0	0	0	0
4 - 8	67	0	0	0	0	0	0	0	0	0	0	1	0
4 - 8	68	0	0	1	0	0	0	0	0	0	1	0	0
4 - 8	69	0	0	0	1	0	0	1	0	0	0	0	1
4 - 8	70	0	1	1	0	0	0	0	0	1	0	0	0
4 - 8	71	1	1	1	0	0	0	0	1	0	0	0	0
4 - 8	72	0	0	0	1	1	1	1	0	0	0	0	0

Table 3.7: Experimental designs for the two-host case of network (8,12), $\sigma^2(x) \equiv 1$.

	I	II	III
	D-optimal	Exchange-Type	
	Continuous	15 point (a)	12 point (b)
n	32	15	12
p_4	.02	0	0
p_5	.03	0	0
p_6	.03	0	0
p_9	.06	1/15	1/12
p_{12}	.01	0	0
p_{22}	.05	1/15	0
p_{23}	.03	0	1/12
p_{24}	.05	1/15	1/12
p_{26}	.02	0	0
p_{28}	.02	0	1/12
p_{29}	.02	1/15	0
p_{31}	.04	1/15	0
p_{32}	.06	1/15	1/12
p_{33}	.04	0	1/12
p_{34}	.03	0	0
p_{35}	.02	0	0
p_{36}	.05	1/15	1/12
p_{41}	.03	0	0
p_{43}	.01	0	0
p_{44}	.04	1/15	1/12
p_{48}	.03	1/15	0
p_{49}	.03	0	0
p_{50}	.02	0	0
p_{53}	.01	0	0
p_{54}	.04	1/15	1/12
p_{57}	.03	1/15	0
p_{58}	.05	1/15	1/12
p_{59}	.04	1/15	1/12
p_{60}	.05	1/15	1/12
p_{65}	.03	0	0
p_{66}	.03	1/15	0
p_{72}	.03	0	0
$\lvert D\rvert$	$.221 \times 10^8$	$.613 \times 10^8$	$.206 \times 10^9$
$\max_i \frac{d(x_i,\xi)}{\sigma^2(x_i)}$	12.00	22.50	43.50

Table 3.8: Experimental designs for the two-host case of network (8,12), $\sigma^2(x) = x^T x$.

	I	II	III
	D-optimal	Exchange-Type	
	Continuous	15 points	12 points
n	39	15	12
p_1	.02	0	0
p_2	.04	1/15	1/12
p_3	.04	1/15	1/12
p_4	.03	1/15	0
p_6	.03	1/15	1/12
p_7	.05	1/15	1/12
p_8	.02	0	0
p_{10}	0.00	0	1/12
p_{13}	.01	0	0
p_{16}	.02	0	0
p_{18}	.02	0	0
p_{19}	.01	0	0
p_{20}	.01	0	0
p_{21}	.01	0	0
p_{26}	.03	0	0
p_{32}	.02	0	0
p_{33}	.01	1/15	0
p_{39}	.04	1/15	0
p_{40}	.04	1/15	1/12
p_{41}	.03	0	1/12
p_{42}	.02	0	0
p_{43}	.04	1/15	0
p_{45}	.04	1/15	1/12
p_{46}	.02	0	0
p_{48}	.01	0	0
p_{49}	.02	0	0
p_{51}	.05	1/15	1/12
p_{53}	.01	0	0
p_{56}	.05	0	0
p_{57}	.03	0	1/12
p_{58}	.04	0	0
p_{59}	.02	0	0
p_{60}	.05	1/15	1/12
p_{61}	.02	0	0
p_{62}	.02	0	0
p_{63}	.04	1/15	1/12
p_{64}	.01	0	0
p_{67}	.01	1/15	0
p_{69}	0.00	1/15	0
$\lvert D\rvert$	$.758 \times 10^{14}$	$.247 \times 10^{15}$	$.482 \times 10^{15}$
$\max_i \frac{d(x_i,\xi)}{\sigma^2(x_i)}$	12.02	21.63	30.67

3.5 ESNET EXAMPLE I

3.5.1 Model and Main Assumptions

After testing the proposed algorithm and software with simple examples, we can proceed with a more realistic problem. Let us consider a simplified graph of the Energy System Network (ESnet) backbone; see Fig. 3.2. For this graph, we have 34 nodes and 39 edges. Let us assume that we want to know delay times on all 39 edges.

Let "*pinging*" be our measurement tool. It is assumed that any route without loops is feasible in the planned monitoring. This assumption is the most vulnerable for a critique, because the standard "*ping*" software does not allow selection of a route. Nevertheless, we do make this assumption to show what can be gained if the route selection is possible.

Thus, we have $E(y|x) = \theta^T x$, where the vectors θ and x have 39 components. Components of the vector θ are, for instance, travel times, and components of the vector x equal 1 if the corresponding edge belongs to a route, and 0 otherwise. Concerning the variance of observation we handle two cases:

$$Var(y|x) \equiv \text{const} \quad \text{and} \quad Var(y|x) = \text{const} \times x^T x. \tag{3.14}$$

The first case may be a reasonable approximation if uncertainties in y are explained by measurement errors. The second choice tries to relate uncertainties in y to the number of edges included in the route x. We do not consider either of these models as a very practical choice. However, they correspond to rather divergent and extreme cases and may help to understand major changes in optimal routing, which occur due to different assumptions on the random behavior of measurements.

In the Poisson-type process setting, it is expedient to choose

$$Var(y|x) = \sum_{\alpha=1}^{m} \theta_\alpha^2 x_\alpha^2 + \sigma^2 \tag{3.15}$$

where θ_{true} are the true values of mean travel times and σ^2 describes various "interferences" on line. The problem with model (3.15) is that the information matrix, and correspondingly, covariance matrix depend on unknown values; see Sect. 2.3. In this case, optimal designs also depend (in general) on those values. There are at least three approaches to handling the design problem [see Fedorov and Hackl (1997), Chaps. 2.6 and 5.6]:

- Bayesian approach, in which we introduce a prior θ_0 and its covariance matrix D_0 and then solve the design problem with the averaged objective function.
- Minimax approach, in which the optimal design is built for the worst $\theta \in \Theta$, where the set Θ is assumed to be given.
- Sequential design, which assumes multi-stage experiments.

We spend most of our discussion with models (3.14) and only at the end of this section return to model (3.15). As before we use the D-criterion in all computations.

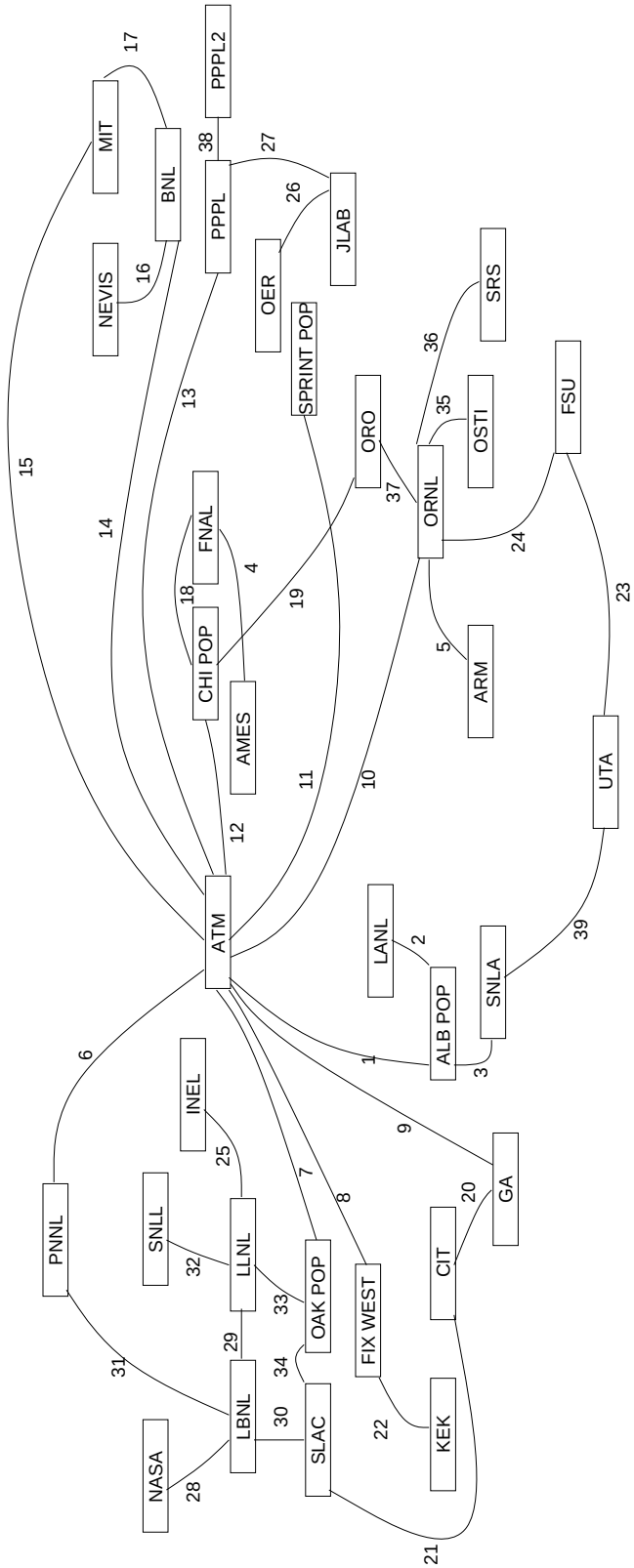

Figure 3.2: Model of ESnet backbone used for computational experiments.

3.5.2 Construction of Support Sets

To determine an optimal experimental design, we first need a mechanism for generating a set of routes that can be used as input to the Optimal Design Construction (ODC) software. To do this, we first generate all valid routes for a specified network, and then apply one or more filters to generate the subset that will be processed by the software as a design region (or set) X.

Separating the generation of all valid routes (when it is not prohibitive due to the size of the considered network) from the filtering (it is necessary to create the final input for ODC) has several advantages. One advantage is that the complete set of routes need only be generated once. The post-processing filtering step can be done as many times as needed to run ODC in different configurations or for the various optimality criteria. The separation also facilitates validating the correctness of the generated routes and the visualization of results through graphical representation. We have developed programs and scripts to assist in the validation and the visualization of the results.

The method for generating all routes through a network (see Fig. 3.2) starts with a list of edges as in Table 3.5.2, each edge being represented as a pair of nodes. Attaching node names to the node numbers, as in Table 3.5.2, is not required by the route generation code ROUTES but is useful for later visualization of results. In particular, generated routes can be overlayed on the graph of the ESnet network. From the list of edges, ROUTES creates an *incidence matrix* [cf. Reingold, Nievergelt and Deo (1977), Chap. 8]. This matrix is sufficient to represent the network as an undirected graph.

The routine ROUTES generates the set of routes incrementally, generating the set of length-$(k+1)$ routes from the set of length-k routes. This is done by extending a length-k route by one previously unused edge from one of the end nodes of that route. ROUTES checks that the added edge does not create a route previously generated and does not introduce a cycle. In addition to the validity checks built into the ROUTES program, we have run separate validity checking scripts on the output of various test problems and have visually inspected the graphical representation of the routes.

After ROUTES produces the complete set of valid routes as output, the results are filtered to produce a smaller set of routes for input to ODC. This filtering may include but is not limited to

- Select set of routes for a *given set of hosts.*
- Select routes whose lengths fall within a specified range. The *length* of a route is based on number of edges (unit weighting of edges) but arbitrary weighting of edges is also possible.
- Select routes such that the selected set has a *Hamming distance* no less than a specified number. In this context, the Hamming distance will be the minimum number of different edges for any two routes in the set.

Additionally, filtering ROUTES allows us to produce transformed inputs for ODC. For instance, for some models we need the normalized vectors x_i. So far we have used

$$x'_{\alpha i} = \frac{x_{\alpha i}}{\sqrt{x_i^T x_i}} \quad \text{and} \quad x'_{\alpha i} = \frac{x_{\alpha i}}{x_i^T x} \quad .$$

Table 3.9: List of ESnet edges used in computational experiments.

Edge	Node 1	Node 2	Edge	Node 1	Node 2
1	1	4	21	7	29
2	1	15	22	8	14
3	1	31	23	10	34
4	2	9	24	10	23
5	3	23	25	12	17
6	4	26	26	13	22
7	4	21	27	13	27
8	4	8	28	16	19
9	4	11	29	16	17
10	4	23	30	16	29
11	4	30	31	16	26
12	4	6	32	17	32
13	4	27	33	17	21
14	4	5	34	21	29
15	4	18	35	23	25
16	5	20	36	23	23
17	5	18	37	23	24
18	6	9	38	27	28
19	6	24	39	31	34
20	7	11			

Table 3.10: List of ESnet nodes used in computational experiments.

ID	Site Acronym	Site Name
1	ALB POP	Sprint POP (Albuquerque, NM)
2	AMES	Ames Laboratory, ISU (Ames, IA)
3	ARM	Atmospheric Radiation Measurement Project (Lamont, OK)
4	ATM	
5	BNL	Brookhaven National Laboratory (Upton, NY)
6	CHI POP	Sprint POP (Chicago, IL)
7	CIT	California Institute Of Technology (Pasadena, CA)
8	FIX WEST	NASA Ames Research Center (Mountain View, CA)
9	FNAL	Fermi National Accelerator Laboratory (Batavia, IL)
10	FSU	Florida State University SCRI (Tallahassee, FL)
11	GA	General Atomics (San Diego, CA)
12	INEL	Idaho National Engineering Lab (Idaho Falls, ID)
13	JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)
14	KEK	Japan-Tsukuba via link out of FIX-West (KEK) (Tsukuba, Japan)
15	LANL	Los Alamos National Laboratory (Los Alamos, NM)
16	LBNL	Lawrence Berkeley National Laboratory (Berkeley, CA)
17	LLNL	Lawrence Livermore National Laboratory (Livermore, CA)
18	MIT	Massachusetts Institute of Technology - LNS (Cambridge, MA)
19	NASA	NASA Ames Research Center (Mountain View, CA)
20	NEVIS	Columbia University Nevis Lab (Irvington, NY)
21	OAK POP	Sprint POP (Oakland, CA)
22	OER	DOE HQ - Office of Energy Research (Germantown, MD)
23	ORNL	Oak Ridge National Laboratory (Oak Ridge, TN)
24	ORO	Oak Ridge Operations (Oak Ridge, TN)
25	OSTI	Office of Scientific and Technical Information (Oak Ridge, TN)
26	PNNL	Pacific Northwest National Laboratory (Richland, WA)
27	PPPL	Princeton Plasma Physics Laboratory (Princeton, NJ)
28	PPPL(2)	Princeton Plasma Physics Laboratory (Princeton, NJ) (additional site)
29	SLAC	Stanford Linear Accelerator (Stanford, CA)
30	SPRINT POP	Sprint POP Connecticut Avenue (Washington, DC)
31	SNLA	Sandia National Laboratories Albuquerque (Albuquerque, NM)
32	SNLL	Sandia National Laboratories Livermore (Livermore, CA)
33	SRS	Savannah River Site (Aiken, SC)
34	UTA	University of Texas at Austin (Austin, TX)

3.5.3 D-optimal Designs for ESnet under Various Assumptions

All Host Case.

Let us assume that we have the maximum freedom of choice, i.e., any route listed by ROUTES can be included in the constructed design. In other words, we may select both host nodes and destination nodes without any constraints. In total ROUTES listed 3918 feasible routes. We started to run OCD with the initial design with weights uniformly distributed between all feasible routes similar to the above simple examples. Unfortunately, the subroutine (modified pivoting) did not invert the corresponding 39×39 information matrix $M(\xi_0)$ reliably. The replacement of this subroutine by more sophisticated one (for instance, LINPAC inversion routine) did not help. Selection of routes for the initial design at random led to the similarly disappointing results, and therefore, we proceeded with the regularized approach.

As was mentioned in the concluding part of Sect. 3.3.2, ODC software allows incorporation of prior information if the latter can be expressed in terms of a prior information matrix M_0; see also Sect. 2.1.5. In terms of computation, the incorporation of a prior information means that instead of the inversion of $M(\xi_0)$, which is needed only at the first iteration, the matrix $M_{tot}(\xi_0) = M(\xi_0) + M_0$ is inverted.

In the considered example we do not assume the presence of any prior information but use the ability of our software to handle cases with prior information to solve a regularized design problem, that is, to construct

$$\xi^*(\gamma) = \arg\max_{\xi} |M(\xi) + \gamma I| \tag{3.16}$$

where γ is a small positive number and I is the identity matrix. If γ is small enough, then there is a hope that $M[\xi^*(\gamma)]$ is well defined and can be reliably inverted.

In all our calculations, we selected $\gamma = 0.01$ and ran ODC software with the initial design either with 50 randomly selected routes or with 50 shortest routes. Lately, we found that the selection of 50 shortest routes (the possible choice is not unique) was effective and can be used in problems of smaller dimensions even without regularization.

We applied a stopping rule based on the value of the step length α_s. To secure the reliability of the final results, we also checked the value of the difference (deficiency)

$$\delta_s = \max_x d(x, \xi_s) - m \tag{3.17}$$

recollecting that [cf. Fedorov and Hackl (1997), Chap. 3.1]

$$\frac{|D(\xi_s)|}{|D(\xi^*)|} \leq e^{\delta_s}. \tag{3.18}$$

Note that m is the number of unknown parameter and equals 39 in the considered case.

To analyze dependence of the computed results on the final α_s we ran the program in the mode with a fixed step length. The final design for a larger α was used as the initial design for a smaller α. The results are presented in Table 3.11. In this table the "deficiency" is defined as $\max_x d(x, \xi_f) - 39$, where the maximum of the variance $d(x, \xi_f)$ of the predicted response is selected among all 3918 possible values; subscript f stands for the final design. The "determinant ratio" is a ratio of $|D(\xi_f)|$

for given α to the same determinant for the least α (i.e., for $\alpha = 0.0005$, when $|D(\xi_f)| = 0.22 \times 10^{41}$). The number of routes is the number of different routes included in the final design ξ_f. The weights of different routes are, in general, different. Comparing the second and third columns from the table with inequality (3.18), one may conclude that the latter inequality is too rough. In practice, we have the better results than what follows from inequality (3.18). The ratio of the determinants for α=0.01 and α=0.0005 looks large. However, in terms of variances of the estimated parameters the discrepancy is much less impressive. It is approximately $\sqrt[39]{7.73} \simeq 1.05$. Hence, for practical needs even computations with the relatively large α=0.005 lead to a decent result. Additionally, one can notice that the computed designs for the smaller step lengths contain more supporting routes. It may be a tedious task to distribute available time between 514 routes with different weights. Note that for any initial random design we never got $\max_x d(x, \xi_0)$ less than a few hundred, even for the regularized problem (3.16). However, further trials might have generated better results.

Table 3.11: Comparison of results for the different final step lengths.

Step length	Deficiency	Determinant ratio	Number of routes
0.01	13.7	7.73	98
0.005	5.0	1.9	167
0.0025	2.5	1.23	268
0.001	1.0	1.05	423
0.0005	0.4	1.0	514

We think that in most practical cases the "rougher" design constructed with α=0.01 can be used. In Table 3.5.3, we reproduce a typical printout for the final design ξ_f for that step length. Only two support route have weights different from 0.01. So it is very easy to schedule the corresponding measurements.

Computing optimal designs for the case, in which all host and all destination nodes are feasible, we select a "team" of experimenters and a set of nodes to be interrogated. Because the graphs are undirected, the rules can be changed: all host nodes can be considered as destination nodes and, corresponding, all destination nodes can be claimed as host nodes. From the convexity of the D-criterion, it follows that there exists an optimal design, which is symmetrical with respect to replacement of host nodes by destination needs [i.e., if the route (a,...,b) enters the optimal design, then the route (b,...,a) is also include in the same design]. With the decrease of the step length α, there is a tendency to include almost all nodes either as hosts or destinations. But a few nodes, for instance, 4 and 6, never appear on the list.

One- and Two-Hosts Cases

It is a rare opportunity when a large number of hosts can be involved in an experiment. Therefore, it is interesting to evaluate how much we lose working with one or two hosts compared to the unconstrained case.

Table 3.12: (Part 1 of 2) The computed design for $\alpha = 0.01$.

Route Number	Weight	Edges of Support Route						
62	0.010	6	9					
82	0.010	8	11					
117	0.020	14	15					
136	0.010	27	38					
209	0.010	6	9	20				
268	0.010	8	10	24				
273	0.010	8	12	18				
279	0.010	8	14	16				
331	0.010	11	12	18				
335	0.010	11	14	16				
378	0.020	26	27	38				
427	0.010	1	3	13	27			
466	0.010	5	8	10	22			
549	0.010	7	9	21	34			
663	0.010	10	12	19	24			
700	0.010	12	13	18	27			
781	0.010	1	3	8	23	39		
791	0.010	1	3	10	37	39		
797	0.010	1	3	12	19	39		
1176	0.010	10	12	18	23	24		
1258	0.010	1	2	10	23	24	39	
1260	0.010	1	2	5	12	19	37	
1301	0.010	1	3	10	23	37	39	
1310	0.010	1	3	12	19	23	39	
1356	0.010	4	7	12	18	25	33	
1372	0.010	4	10	15	18	19	37	
1399	0.010	6	7	28	31	32	33	
1400	0.010	6	7	21	28	31	34	
1443	0.010	6	10	25	29	31	36	
1548	0.010	7	9	20	29	31	33	
1584	0.010	7	10	20	21	34	36	
1595	0.010	7	10	30	31	34	35	
1657	0.010	7	14	16	20	21	34	
1673	0.010	7	15	16	17	32	33	
1765	0.010	10	13	23	24	27	39	
1791	0.010	12	14	17	19	35	37	
1808	0.010	1	2	7	25	29	30	34
1880	0.010	1	3	12	19	23	24	37
1881	0.010	1	3	12	19	24	37	39
1990	0.010	5	12	13	19	26	27	37
2098	0.010	6	14	16	30	31	33	34
2100	0.010	6	14	17	20	21	30	31
2119	0.010	7	8	22	29	30	32	34
2146	0.010	7	10	18	19	25	33	37
2168	0.010	7	11	20	21	29	30	33
2201	0.010	7	13	26	27	28	29	33
2204	0.010	7	13	20	21	26	27	34
2213	0.010	7	13	25	29	30	34	38
2215	0.010	7	14	16	21	29	30	33
2254	0.010	9	10	20	21	28	30	37
2304	0.010	9	13	20	21	27	30	31
2311	0.010	9	13	20	21	30	31	38
2359	0.010	12	15	16	17	19	35	37

In the one-host case, ORNL (node 23) was selected as a host node. Only 231 routes are feasible and can be used in computations (compared with 3918 for all-host case). With α=0.001, the computed design has the determinant $|D(\xi_f| = 0.36 \times 10^{52}$; compare with $|D(\xi_f| = 0.22 \times 10^{41}$ for the previous case. For the two-host case (ORNL and LBNL), there are 524 feasible routes. It was found that $|D(\xi_f| = 0.24 \times 10^{49}$. Table 3.13 contains information on the variance of all 39 estimated parameters for the all-hosts, one- and two-hosts cases. From comparison of determinants and variances for the different cases we may conclude that cooperation is very useful and allows accumulation of more information given the same total number of measurements.

The Partner Selection

After conclusion that cooperation is good, we may make another step and try to select the best partner. Theoretically, it looks rather simple. The COD software must be run for all possible 33 pairs, which include ORNL, if the latter is the initiator of "teaming". The results of these runs are shortly described in Table 3.5.3, in which the variances of the parameter estimators and the

Table 3.12: (Part 2 of 2) The computed design for $\alpha = 0.01$.

Route Number	Weight	Edges of Support Route												
2365	0.010	1	2	6	30	31	32	33	34					
2494	0.010	2	3	10	14	17	23	24	39					
2520	0.010	4	6	12	18	20	21	30	31					
2551	0.010	5	6	12	19	29	31	32	37					
2568	0.010	6	8	21	22	29	31	33	34					
2648	0.010	6	13	30	31	32	33	34	38					
2659	0.010	6	15	17	21	29	31	33	34					
2733	0.010	7	15	16	17	25	29	30	34					
2736	0.010	8	9	20	21	22	25	29	30					
2845	0.010	9	15	16	17	20	21	30	31					
2863	0.010	1	2	9	20	21	28	29	33	34				
2955	0.010	1	3	13	23	24	26	27	36	39				
2965	0.010	1	3	15	16	17	23	24	36	39				
2970	0.010	2	3	6	10	23	24	28	31	39				
3034	0.010	4	7	10	18	19	28	30	34	37				
3041	0.010	5	9	10	20	21	29	31	33	34				
3058	0.010	6	10	23	24	25	30	31	33	34				
3095	0.010	6	13	26	27	30	31	32	33	34				
3116	0.010	7	12	19	21	29	30	33	35	37				
3122	0.010	7	12	19	20	21	23	24	34	37				
3199	0.010	9	13	20	21	28	29	33	34	38				
3203	0.010	9	14	17	20	21	28	29	33	34				
3221	0.010	1	3	6	23	24	29	31	32	35	39			
3261	0.010	1	3	7	23	24	29	31	33	37	39			
3270	0.010	1	3	5	7	23	24	28	30	34	39			
3312	0.010	1	3	4	11	18	19	23	24	37	39			
3340	0.010	2	3	8	12	19	22	23	24	37	39			
3370	0.010	4	9	12	18	20	21	29	31	33	34			
3399	0.010	6	12	19	23	24	25	29	31	37	39			
3411	0.010	6	12	19	25	30	31	33	34	36	37			
3429	0.010	7	12	19	23	24	28	30	34	37	39			
3435	0.010	9	10	20	21	23	24	29	30	32	39			
3443	0.010	9	10	18	19	20	21	29	30	32	37			
3473	0.010	9	12	19	20	21	32	33	34	36	37			
3510	0.010	1	3	7	21	23	24	29	30	33	36	39		
3534	0.010	1	3	9	20	21	23	24	28	30	35	39		
3563	0.010	1	3	4	13	18	19	23	24	37	38	39		
3565	0.010	1	3	4	14	17	18	19	23	24	37	39		
3702	0.010	1	3	7	18	19	23	24	30	31	34	37	39	
3720	0.010	1	3	5	9	20	21	23	24	25	33	34	39	
3721	0.010	1	3	9	20	21	23	24	25	33	34	35	39	
3742	0.010	2	3	7	10	20	21	23	24	29	30	33	39	
3846	0.010	2	3	7	12	19	23	24	29	30	32	34	37	39

determinants of the covariances matrices for the computed designs are presented. It appears that CIT is the best partner for ORNL. In Table 3.5.3 the best and worst partners are compared.

Variance Depending on Routes

Intuitively it is clear that the increase of the measurement variance with the increase of the route length might result in selection of shorter routes for optimal monitoring designs. We experimented with $Var(Y|x) = x^T x$. For the all-hosts case, the optimal monitoring design looks natural and almost trivial. It includes all one-edge routes with relatively large weights (between 0.02 and 0.025) and distributes small weights (0.002, 0.003) between the longer routes. The longest route contains 10 edges. See more details in Table 3.16. Strangely enough the computed design has 2- and 5-edges routes but skips 3- and 4-edges routes. For more realistic one and two hosts cases, the optimal designs contains the larger portion of longer routes, and only the shortest ones can be identified on an intuitive level (see Tables 3.5.3 and 3.5.3).

Table 3.13: Comparison of parameter estimators and variances for different monitoring schemes.

Edge	All hosts	2 hosts	1 host
1	10.8506	18.4519	20.4076
2	11.1233	20.6538	26.5529
3	23.3553	43.9910	49.2563
4	23.3336	43.0290	50.6497
5	16.8525	34.3890	38.4615
6	15.5287	29.8379	44.1772
7	7.9504	8.6672	26.7796
8	19.6434	34.4523	43.3488
9	5.7673	29.0536	44.4625
10	6.4051	5.8516	10.5631
11	19.4606	34.4523	43.4883
12	8.4291	13.0099	14.1725
13	15.7592	34.4651	44.6929
14	12.6968	33.4043	43.3851
15	18.8220	33.9145	44.3778
16	17.4377	32.8024	36.7196
17	17.3032	32.8071	36.7196
18	14.3867	26.2262	32.9405
19	19.1655	38.7371	43.0290
20	17.6391	45.5270	35.1296
21	13.3079	25.1410	26.2582
22	38.4753	66.6800	72.7548
23	26.4606	55.5669	60.6493
24	16.3087	29.5349	39.4100
25	12.4533	24.6633	25.9346
26	32.5310	66.6910	74.0741
27	23.4582	66.7000	74.0741
28	13.3242	34.4810	27.2631
29	7.0935	11.7790	13.3515
30	6.7655	9.3346	12.9401
31	12.6046	29.1031	24.9268
32	12.6969	25.0673	25.9392
33	7.4444	12.1627	13.6535
34	6.8715	5.2495	12.6264
35	16.9016	34.4078	38.4615
36	16.7393	34.4571	38.4615
37	15.7152	29.4935	38.8207
38	15.3252	66.7000	74.0741
39	26.7182	55.5786	60.6493
Determinant	0.22E+41	0.24E+49	0.36E+52
Average of variances	16.2335	32.7311	38.0384

Table 3.14: Data for selecting node to partner with ORNL.

Node	Determinant	Average of variances	Maximal variance
1	0.94E+50	35.1554	39.6849
2	0.88E+49	34.3190	39.8497
3	0.14E+51	36.3375	39.8139
4	0.10E+52	36.5121	39.5497
5	0.16E+50	33.9359	39.7020
6	0.16E+51	35.9283	39.6957
7	0.25E+48	31.0723	39.7940
8	0.57E+50	33.6918	39.9377
9	0.87E+49	33.4465	39.8512
10	0.43E+50	34.5181	39.6482
11	0.25E+48	31.1211	39.7295
12	0.12E+48	30.8987	39.7371
13	0.16E+50	31.4065	40.0000
14	0.57E+50	34.5396	39.9724
15	0.52E+49	33.4762	39.8790
16	0.24E+49	32.7354	39.6367
17	0.65E+48	31.8312	39.7653
18	0.52E+49	32.5790	39.8984
19	0.35E+48	31.6782	39.8419
20	0.52E+49	33.3984	39.9093
21	0.17E+49	32.5455	39.7407
22	0.16E+50	32.3056	40.0000
24	0.34E+50	34.0752	39.8571
25	0.14E+51	36.3117	39.7863
26	0.28E+48	30.9673	39.7613
27	0.57E+50	32.8992	39.7912
28	0.98E+49	31.6014	39.8461
29	0.18E+49	32.4952	39.6027
30	0.57E+50	34.5609	39.9717
31	0.31E+50	33.9804	39.8702
32	0.12E+48	30.8987	39.7371
33	0.14E+51	36.2623	39.6947
34	0.36E+50	34.1160	39.9167

Table 3.15: The best (CIT) and worst (ATM) nodes to partner with ORNL.

Node	CIT	ATM
Determinant	0.249E+48	0.103E+52
Number of routes	233	155
Variances of parameter estimators	18.5	20.0
	21.0	22.7
	43.3	49.8
	42.6	50.0
	34.8	39.4
	28.0	39.0
	8.9	21.5
	34.6	39.4
	29.4	39.2
	6.1	4.9
	34.6	39.4
	13.0	13.7
	34.6	39.4
	33.5	37.3
	34.0	38.5
	32.8	36.4
	32.8	36.4
	26.1	29.3
	37.0	43.7
	28.4	35.4
	8.7	26.5
	65.6	74.1
	54.1	62.5
	28.9	32.6
	22.4	26.0
	64.5	74.1
	65.6	74.1
	28.1	27.3
	9.5	13.3
	8.6	13.0
	23.9	24.8
	22.3	25.7
	9.5	13.7
	8.8	12.8
	34.8	39.4
	34.8	39.4
	28.1	32.5
	65.6	74.1
	54.1	62.5

Table 3.16: Optimal design for all hosts, $\sigma^2 = x^T x$, $\alpha = 0.001$.

Route Number	Weight	Edges of Support Route										
1	0.022	1										
2	0.025	2										
3	0.025	3										
4	0.025	4										
5	0.022	5										
6	0.022	6										
7	0.020	7										
8	0.020	8										
9	0.025	9										
10	0.025	10										
11	0.025	11										
12	0.025	12										
13	0.020	13										
14	0.022	14										
15	0.022	15										
16	0.025	16										
17	0.025	17										
18	0.025	18										
19	0.025	19										
20	0.025	20										
21	0.025	21										
22	0.025	22										
23	0.025	23										
24	0.022	24										
25	0.025	25										
26	0.025	26										
27	0.025	27										
28	0.025	28										
29	0.025	29										
30	0.025	30										
31	0.025	31										
32	0.025	32										
33	0.025	33										
34	0.025	34										
35	0.022	35										
36	0.025	36										
37	0.022	37										
38	0.025	38										
39	0.025	39										
40	0.002	1	2									
49	0.002	1	13									
50	0.002	1	14									
56	0.002	5	24									
57	0.002	5	35									
59	0.003	5	37									
60	0.002	6	7									
61	0.002	6	8									
68	0.002	6	15									
70	0.002	7	8									
75	0.003	7	13									
77	0.002	7	15									
84	0.003	8	13									
86	0.002	8	15									
87	0.003	8	22									
92	0.003	9	14									
113	0.002	13	14									
131	0.002	24	37									
133	0.003	25	32									
149	0.002	35	37									
833	0.002	3	23	24	35	39						
847	0.003	4	12	13	18	38						
1034	0.003	7	10	28	30	34						
3417	0.002	6	13	20	21	26	34	27	29	31	33	34

Table 3.17: Optimal design for one host, $\sigma^2 = x^T x$.

Route Number	Weight	Edges of Support Route									
1	0.025	5									
2	0.005	10									
3	0.025	24									
4	0.025	35									
5	0.025	36									
6	0.025	37									
7	0.010	1	10								
8	0.020	6	10								
10	0.025	8	10								
11	0.020	9	10								
12	0.025	10	11								
14	0.025	10	13								
15	0.010	10	14								
16	0.025	10	15								
18	0.020	23	24								
19	0.020	1	2	10							
20	0.020	1	3	10							
24	0.025	8	10	22							
25	0.015	9	10	20							
26	0.020	10	12	18							
27	0.020	10	12	19							
28	0.025	10	13	27							
29	0.025	10	13	38							
30	0.015	10	14	16							
31	0.020	10	14	17							
34	0.020	18	19	37							
35	0.020	23	24	39							
36	0.010	1	3	10	39						
39	0.020	4	10	12	18						
40	0.020	4	18	19	37						
41	0.015	6	10	28	31						
45	0.015	7	10	25	33						
47	0.010	7	10	32	33						
48	0.015	7	10	21	34						
50	0.005	7	12	19	37						
54	0.025	10	13	26	27						
55	0.020	10	15	16	17						
56	0.005	11	12	19	37						
60	0.005	1	2	12	19	37					
64	0.020	2	3	23	24	39					
65	0.010	6	10	25	29	31					
66	0.015	6	10	29	31	32					
68	0.005	6	10	21	30	31					
71	0.010	7	10	28	29	33					
72	0.005	7	10	29	30	33					
73	0.010	7	10	29	31	33					
74	0.010	7	10	20	21	34					
75	0.010	7	10	28	30	34					
77	0.010	7	10	30	31	34					
86	0.010	12	14	16	19	37					
88	0.010	12	15	17	19	37					
92	0.005	1	3	9	23	24	39				
99	0.005	6	10	29	31	33	34				
101	0.005	6	10	30	31	33	34				
105	0.010	7	10	21	29	30	33				
106	0.005	7	10	25	29	30	34				
107	0.015	7	10	29	30	32	34				
110	0.005	7	12	19	32	33	37				
113	0.010	9	10	20	21	28	30				
115	0.005	9	10	20	21	30	31				
129	0.005	1	3	13	23	24	38	39			
133	0.005	6	10	21	29	31	33	34			
134	0.015	6	10	25	30	31	33	34			
139	0.005	6	12	19	21	30	31	37			
141	0.005	7	10	20	21	29	30	33			
149	0.010	9	10	20	21	25	29	30			
154	0.015	9	10	20	21	32	33	34			
206	0.005	1	3	6	23	24	29	31	33	34	39
207	0.010	1	3	6	20	21	23	24	30	31	39
210	0.005	1	3	7	23	24	25	29	30	34	39
217	0.010	9	12	19	20	21	28	29	33	34	37

Table 3.18: (Part 1 of 2) Optimal design for two hosts, $\sigma^2 = x^T x$.

Route Number	Weight	Edges of Support Route			
1	0.026	5			
3	0.026	24			
4	0.026	28			
5	0.009	29			
6	0.013	30			
7	0.022	31			
8	0.026	35			
9	0.026	36			
10	0.023	37			
11	0.007	1	10		
12	0.015	6	10		
15	0.018	8	10		
16	0.016	9	10		
17	0.018	10	11		
19	0.018	10	13		
20	0.003	10	14		
21	0.020	10	15		
23	0.016	21	30		
24	0.022	23	24		
25	0.020	25	29		
26	0.020	29	32		
27	0.002	29	33		
29	0.017	1	2	10	
30	0.014	1	3	10	
31	0.004	1	6	31	
33	0.005	6	8	31	
34	0.003	6	9	31	
36	0.005	6	11	31	
38	0.005	6	13	31	
39	0.006	6	14	31	
40	0.002	6	15	31	
45	0.018	8	10	22	
46	0.011	9	10	20	
47	0.016	10	12	18	
48	0.013	10	12	19	
49	0.019	10	13	27	
50	0.018	10	13	38	
51	0.018	10	14	16	
52	0.020	10	14	17	
55	0.020	18	19	37	
56	0.009	20	21	30	
57	0.018	23	24	39	
58	0.008	29	33	34	
59	0.001	30	33	34	
60	0.001	1	2	6	31
61	0.001	1	3	6	31
62	0.008	1	3	10	39
63	0.001	1	7	29	33
67	0.011	4	10	12	18
68	0.018	4	18	19	37
74	0.003	6	8	22	31
83	0.002	6	12	18	31
86	0.003	6	13	27	31
87	0.003	6	13	31	38
91	0.003	7	8	29	33
92	0.003	7	8	30	34
93	0.002	7	9	29	33
94	0.003	7	9	30	34
95	0.009	7	10	25	33
97	0.009	7	10	32	33
98	0.010	7	10	21	34
100	0.003	7	11	29	33
101	0.002	7	11	30	34
105	0.002	7	13	29	33
106	0.003	7	13	30	34
107	0.003	7	14	29	33
108	0.003	7	14	30	34
109	0.001	7	15	29	33
110	0.001	7	15	30	34
111	0.003	8	12	19	37
113	0.001	9	12	19	37
115	0.017	10	13	26	27
116	0.012	10	15	16	17
117	0.003	11	12	19	37
118	0.003	12	13	19	37
120	0.003	12	15	19	37
121	0.008	21	29	33	34
122	0.011	25	30	33	34
123	0.011	30	32	33	34

Table 3.18: (Part 2 of 2) Optimal design for two hosts, $\sigma^2 = x^T x$.

Route Number	Weight	Edges of Support Route										
124	0.002	1	2	7	29	33						
125	0.004	1	2	7	30	34						
126	0.003	1	2	12	19	37						
127	0.001	1	3	6	31	39						
128	0.002	1	3	7	29	33						
129	0.002	1	3	7	30	34						
131	0.002	1	3	12	19	37						
133	0.001	1	9	20	21	30						
134	0.021	2	3	23	24	39						
135	0.004	4	6	12	18	31						
138	0.001	6	7	25	31	33						
139	0.002	6	7	31	32	33						
140	0.001	6	7	21	31	34						
151	0.004	6	13	26	27	31						
152	0.006	6	15	16	17	31						
153	0.002	7	8	22	29	33						
154	0.003	7	8	22	30	34						
156	0.001	7	9	20	29	33						
157	0.002	7	9	20	30	34						
161	0.002	7	10	29	30	33						
166	0.009	7	10	20	21	34						
168	0.001	7	10	29	30	34						
173	0.002	7	12	18	29	33						
174	0.002	7	12	18	30	34						
179	0.002	7	13	27	29	33						
180	0.002	7	13	27	30	34						
181	0.003	7	13	29	33	38						
182	0.002	7	13	30	34	38						
183	0.002	7	14	16	29	33						
184	0.002	7	14	16	30	34						
187	0.002	7	15	17	29	33						
188	0.002	7	15	17	30	34						
189	0.002	8	9	20	21	30						
190	0.003	8	12	19	22	37						
193	0.002	9	11	20	21	30						
196	0.001	9	13	20	21	30						
197	0.002	9	14	20	21	30						
198	0.001	9	15	20	21	30						
199	0.003	12	13	19	27	37						
200	0.003	12	13	19	37	38						
201	0.003	12	14	16	19	37						
202	0.004	12	14	17	19	37						
204	0.004	20	21	29	33	34						
205	0.001	1	2	9	20	21	30					
208	0.001	1	3	7	29	33	39					
209	0.002	1	3	7	30	34	39					
214	0.001	1	3	11	23	24	39					
215	0.001	1	3	12	19	37	39					
217	0.001	1	3	13	23	24	39					
220	0.002	4	7	12	18	29	33					
221	0.002	4	7	12	18	30	34					
246	0.001	7	12	19	25	33	37					
248	0.001	7	12	19	32	33	37					
249	0.001	7	12	19	21	34	37					
251	0.003	7	13	26	27	29	33					
252	0.002	7	13	26	27	30	34					
253	0.002	7	15	16	17	29	33					
254	0.002	7	15	16	17	30	34					
255	0.002	8	9	20	21	22	30					
264	0.001	9	12	18	20	21	30					
267	0.001	9	13	20	21	27	30					
268	0.001	9	13	20	21	30	38					
273	0.002	12	13	19	26	27	37					
274	0.001	12	15	16	17	19	37					
280	0.001	1	3	8	22	23	24	39				
283	0.001	1	3	12	18	23	24	39				
286	0.001	1	3	13	23	24	27	39				
287	0.001	1	3	13	23	24	38	39				
288	0.002	1	3	14	16	23	24	39				
290	0.002	1	3	15	17	23	24	39				
293	0.002	4	9	12	18	20	21	30				
328	0.001	8	9	20	21	29	33	34				
337	0.001	9	11	20	21	29	33	34				
341	0.002	9	13	20	21	26	27	30				
342	0.002	9	13	20	21	29	33	34				
343	0.001	9	14	20	21	29	33	34				
344	0.002	9	15	16	17	20	21	30				
345	0.001	9	15	20	21	29	33	34				
497	0.001	1	3	9	20	21	23	24	25	33	34	39
499	0.001	1	3	9	20	21	23	24	32	33	34	39

3.6 NETWORK CHALLENGES: PROBLEMS TO EXPLORE

3.6.1 Multiresponse Models

Almost any study on network performance analysis stresses the fact that it is not difficult to measure a variety of technical characteristics, such as delays, jitter, loss of packets, traffic intensities, queue lengths, etc. [see, for instance, Brownlee (1995), (1996); Claffy (1994); and Paxson (1997)]. The first reference is an example of a particular collection of measurement tools. The other two references contain an extensive bibliography on measurement efforts, together with descriptions of the most popular metrics. Generally, a host site can measure, at every experimental session, a few response variables corresponding to a selected host-destination pair, route, or link/channel, server, etc.

Model (2.4) and (2.5) must be replaced now by the multiresponse model [cf. Fedorov and Hackl (1997), Chap. 1.3 and Fedorov (1972), Chap. 5]:

$$E(y|x) = F^T(x)\theta \quad \text{and} \quad Var(y|x) = \Sigma(x) \quad , \tag{3.19}$$

where y is a k-dimension random vector, $F(x) = (f^{(1)}(x), \ldots, f^{(k)}(x))$ is a given $(m \times k)$-matrix function, θ is a vector of unknown parameters, and $\Sigma(x)$ is a $k \times k$ matrix.

If all the elements of the matrix $\Sigma(x)$ are known, then the generalization of results previously discussed is straightforward [compare with (2.11)– (2.15)]:

$$\hat{\theta} = \underline{M}^{-1} Y \quad , \tag{3.20}$$

where

$$\underline{M} = \sum_{i=1}^{k} r_i F(x_i) \Sigma^{-1}(x_i) F^T(x_i), \tag{3.21}$$

$$Y = \sum_{i=1}^{k} r_i F(x_i) \Sigma^{-1}(x_i) y_i. \tag{3.22}$$

Some handy simplifications for (3.20)–(3.22) can be derived when different component responses (components of y) depend upon disjoint subsets of the vector θ, that is, when

$$E(y_{(j)}|x) = f_{(j)}^T(x)\theta_{(j)}$$

and

$$F(x) = \left\{ \begin{array}{cccc} f_{(1)}(x) & 0 & \ldots & 0 \\ 0 & f_{(2)}(x) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & f_{(k)}(x) \end{array} \right\} \quad .$$

The corresponding formulae can be found, for instance, in Muirhead (1982), Chap. 10 and Seber (1984), Chap. 8.

The information matrix (3.21) may be conveniently rewritten as

$$\underline{M} = \sum_{i=1}^{k} r_i m(x_i) \quad ,$$

where $m(x)$ is the information matrix of the vector measurement at the point x. Note that $\text{rank} m(x_s) \geq 1$ while previously $\text{rank} m(x_s) = 1$. Transition to the normalized information matrix gives

$$M(\xi) = \int_X m(x)\xi(dx) \tag{3.23}$$

and we can immediately apply all the results from Sects. 3.1–3.5 with some minor formal changes. For instance, the sensitivity function for the D-criterion is now

$$\phi(x,\xi) = \text{tr}\Sigma^{-1}(x)d(x,\xi) \quad , \tag{3.24}$$

where the $(k \times k)$ normalized variance matrix of the estimated vector response function is defined as $d(x,\xi) = F^T(x)M^{-1}(\xi)F(x)$. Thus, the generalization is obvious if the matrix $\Sigma(x)$ is given.

However, the situation worsens when this matrix is unknown. In single response cases, a solution of any design problem without prior information (i.e., $M_0 = 0$) does not depend upon σ^2 if the latter is constant. That is why we could assume that $\sigma^2 \equiv 1$ without loss of generality. For instance, a D-optimal design must be (see Theorem 3.2.1) a solution of the minimax problem

$$\xi^* = \arg\min_{\xi \in X} \max_{x \in X} \sigma^{-2} d(x,\xi) \quad ,$$

which obviously does not depend on σ^2. For multiresponse models, a D-optimal design coincides with one of the solutions of the similar minimax problem

$$\xi^* = \arg\min_{\xi \in \Xi} \max_{x \in X} \text{tr}\Sigma^{-1} d(x,\xi) \quad ,$$

which unfortunately depends, in general, on the structure of the matrix Σ.

Not much is known about the design problem with unknown Σ. One of the simplest approaches is to proceed with some initial experiment to estimate the matrix Σ and then to continue replacing in all calculations Σ by its estimate $\hat{\Sigma}$. The corresponding estimation problem is discussed, for instance, in Fedorov and Hackl (1997), Chap. 1.3. Another alternative approach can be based on the fact that

$$\lambda_{\max}^{-1}\mathcal{M}(\xi) \leq M(\xi) \leq \lambda_{\min}^{-1}\mathcal{M}(\xi) \tag{3.25}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the greatest and least eigenvalues of the matrix Σ and

$$\mathcal{M}(\xi) = \int_X F(x)F^T(x)\xi(dx)$$

[i.e., $\mathcal{M}(\xi)$ is the information matrix for the case in which all components of the observed vector are not correlated]. If it may be assumed that $\lambda_{\max}$ and $\lambda_{\min}$ are not far apart, then the design

$$\xi^0 = \arg\min_{\xi} \Psi[\mathcal{M}(\xi)]$$

may be a reasonable approximation for the optimal design ξ^*. Indeed, from (3.25) it follows that for any criterion from Table 3.1

$$\Psi\left[\lambda_{\max}^{-1}\mathcal{M}(\xi^0)\right] \leq \Psi\left[\lambda_{\max}^{-1}\mathcal{M}(\xi^*)\right] \leq \Psi\left[M(\xi^*)\right] \leq \Psi\left[M(\xi^0)\right] \leq \Psi\left[\lambda_{\min}^{-1}\mathcal{M}(\xi)\right] \quad .$$

If the covariance matrix of the response vector changes in X and is to be estimated, then even the estimation problem (to our knowledge) belongs to an unknown realm. One possible approach can be based on parameterization of $\Sigma(x)$ with the consequent use of the iterated estimator technique discussed in Sect. 2.3.

3.6.2 Selection of the Response Component to Measure

One interesting problem for the multiresponse case is of practical interest and was not explored in experimental design theory. Let us consider the simplest case when the covariance matrix Σ is known and constant. In all previous discussions, we were concerned only with an optimal choice of the pairs $\{p_i, x_i\}_1^n$ assuming that at every support point x_i, all components of the response vector y are measured. If the measurement of each component leads to some expense (time, storage space, for example), then we can consider the optimization problem in which, together with the choice of p_i and x_i, the most informative components of y must be selected to be measured.

One of the possibilities is to introduce a dummy variable x_0, which has k (i.e., number of y-s component) levels and to include it in the list of controlled variables. Another way to transform the problem into something that can be handled with convex design theory is the introduction of k dummy two-level variables. At present we are trying to combine ideas from Batsell et al. (1998), Fedorov (1996) and Fedorov and Flanagan (1998) to develop an algorithm, which allows building of optimal designs for the discussed problem.

3.6.3 Different Convergence Rates of Parameter Estimators

So far we have been interested in the "mean" behavior of the response function(s). If we return to the example from Sect. 2.2.1, then the parameters $\theta_\alpha = \sigma_\alpha$, $\alpha = 1, \ldots, m$ are of interest. For this parameter, the information matrix increases as $\sim N$ (or the covariance matrix decreases as $\sim N^{-1}$) and therefore it can be normalized by N^{-1}; see (2.34). This fact allows us to apply the results of convex design theory or at least the version with information matrix depending upon unknown parameters; see Atkinson and Fedorov (1988), Fedorov and Hackl (1997), Chap. 2.6.

A very interesting and unexplored problem arises when the parameters $\tau^T = (\tau_1, \ldots, \tau_m)$ (in notations of examples 2.3.1) are not given and must be estimated together with σ_α, $\alpha = 1, \ldots, m$. An example of such a problem can be found in Cottrell (1998). There exist estimators for the parameter τ with the variances decreasing faster than N^{-1}; see related results in Akahira and Takeuchi (1995), Chaps. 1 and 2, and Johnson, Kotz and Balakrishnan (1994), Chap. 7. For instance, for the single link case (i.e., τ is scalar), the maximum likelihood estimator for s is

$$\hat{\tau} = \min_{1 \le i \le N} y_i \quad ,$$

where N is a number of observations made on this single link. The variance of $\hat{\tau}$ is

$$Var(\hat{\tau}) = \frac{\sigma^2}{N^2} \quad .$$

The subscript α is skipped in both formulae.

On an intuitive level, it is clear that an optimal design must essentially depend on the total number of available measurements. Perhaps in the first and probably smaller part of the designed experiment, the efforts must be directed to estimate s, and then the large portion must be associated with the better estimation of $\theta_\alpha = \sigma_\alpha^2$, $\alpha = 1, \ldots, m$. We are not familiar with any attempt to design experiments for the above case.

3.6.4 Other Types of the Information Matrix Normalization

In (2.21) or in the more general case (2.22), we divide the information matrix by the total number of observations, and optimal designs maximize (in the sense of a selected criterion) the information (matrix) per observation. In terms of examples from Sect. 3.4, it means maximization of the gain per *ping*. In many cases it is a very reasonable approach. However, it is not difficult to imagine situations in which other normalizations would look more natural. For instance, if our "expenses" are proportional to the total number of edges included in observed routes, then, this normalization results in optimal designs that include shorter routes than in the examples of Sect. 3.4. If the generation of *pings* consumes most of the experimental time, then the basic (i.e., normalization by number *pings*) approach must be used. If the *ping* travel time is the main contributor then the second approach may be used. The two above ways of normalization are simple and are recommended for the practical use.

The direct and explicit normalizations of the information matrix by experimental time (or by cost) are also possible. For instance, one can introduce the total time

$$T = (\text{total time that is necessary to generate all } pings) + (\text{sum of travel times}).$$

Note the sum of travel times equals $\sum_i \tau_i \sum_{\alpha=1}^m \theta_\alpha x_{\alpha i}$ and depends upon unknown parameters θ. Thus, we have to find

$$\xi^* = \arg\min_{\xi} \Psi[\underline{M}(\xi)] \ , \quad T(\xi) \le T \quad . \tag{3.26}$$

We do not know how to transform (3.26) to the optimization problem similar to (3.8). Most probably, the techniques developed for the constrained design problem must be applied to (3.26) [cf. Cook and Fedorov (1995)].

3.6.5 Heavy Tailed Distributions

In a number of studies related to the network analysis, it was noted that the observed variables have the heavy tailed distributions; see, for instance, Willinger et al. (1995a,b); Willinger, Taqqu

and Erramilli (1996); and Samorodnitsky and Taqqu (1994), Chap. 1. There are a few possible explanations for that phenomenon. In this section we prefer to stay within the Poisson model paradigm which includes (or leads to) the heavy tail distributions in a very natural way.

To see that, we return again to the example from section 2.3.3 and assume that the previously fixed parameters σ_α (mean service/travel/arrival times) are randomly distributed. For instance, let us assume that $\mu = \sigma^{-1}$ (we skip the subscript α) has a gamma distribution, that is,

$$p(\mu) = \frac{\mu^{\gamma-1} e^{-\mu/s}}{s^\gamma \Gamma(\gamma)} \quad . \tag{3.27}$$

Then the resulting distribution density of the total traveling time Z is

$$\bar{p}(z) = \frac{\gamma s}{(1+sz)^{\gamma+1}} \quad ; \tag{3.28}$$

see, Johnson, Kotz, and Balakrishnan (1994), Chap. 20.2. Note that $E(\mu) = \gamma s$ and $Var(\mu) = \gamma s^2$. The expectation γs can be considered as the mean intensity of the Poisson process corresponding to an exponentially distributed random variable z with $p(z) = \frac{1}{\sigma} e^{-z/\sigma}$ and random s.

Density (3.28) is the density of the Pareto distribution, which is, probably, one of the most popular heavy tailed distributions. If $\gamma \leq 2$, then for the Pareto distribution the second moment and, consequently, the variance, does not exist.

However, all the ideas and results in most of Chap. 1 and Chap. 2 are essentially based on the existence of the first and second moments and, in particular, on the concept of the best linear unbiased estimation. Thus, we have to reconsider the whole approach, and the first thing to do is to replace the objective functions based on either dispersion matrix or information matrix with something else. For instance, it may be the volume or other quantitative characteristics of the confidences regions. To our knowledge, this area was never touched in the statistical literature related to experimental design.

Chapter 4

NONPARAMETRIC APPROACH IN OPTIMAL MONITORING

4.1 BEST LINEAR PREDICTOR

4.1.1 Introduction

In two previous chapters, data analysis and optimal monitoring methods were essentially based on the regression model concept. We have assumed that there exists a function that defines the response, given input/independent variables or predictors. For instance, in example 2.1.1, it was the linear function that connected the total travel time with delays at every given edge. Further deliberation allowed us to describe the distribution of "noise" term. As a result, we had a stochastic model of a known structure but containing unknown parameters that had to be estimated. The series of examples in Chaps. 2 and 3 shows find the collection of various assumptions. Linearity of the response function with respect to parameters, additiveness of observational errors, and independence of errors are most frequently used.

Better knowledge of an analyzed system allows better design of experiments to collect additional information. However, there exists a danger in which a practitioner may (involuntarily) replace the lack of knowledge by seemingly reasonable assumptions that lead to a design mathematically optimal but practically useless if the guess is wrong.

This chapter discusses methods that are based on a possibly minimal set of assumptions. In particular, we abandon the use of the response function; see (2.2) and (2.3). Instead of the latter one, we make a very modest assumption that the performance characteristics of different sites are correlated. However, in some cases, in which the probability distributions of observed characteristics are heavily tailed, the standard concept of "covariance" (or "correlation") cannot be used because the corresponding expectations may not exist; see, for instance, Athreya, Lahiri and Wu (1998).

Nevertheless, the existence of covariances is a much milder assumption than the assumption of a specific functional relationship. All of the following results are based on the covariance structure(s) and do not explicitly include unknown parameters. Hence, the approach is called here nonparametric. However, the terminology is slightly different from what is used in classical nonparametric statistics [cf. Conover (1980)]. Let S sites/nodes $X = (x_1, \ldots, x_S)$ be monitored by one host. To keep notation simple, we will mainly analyze the case in which only one variable (characteristic, metric) is measured by a host at all S sites. Unlike the previous chapters we do not assume that a route connecting the host site and a destination site can be chosen at our wish. We assume

that only a round trip time of a *ping* is available. In general, several variables can be measured, for instance, flow-rate delays of various types, queue lengths, etc. A few hosts may be involved in measurement and data collection. The ideas and techniques developed for the univariate single host case can be extended for multivariate and multi-host situations, and the possible generalization will be discussed later. In what follows, we use results by Fedorov and Flanagan (1997) and by Batsell et al. (1998) together with some new findings and examples. The simpler versions of the considered statistical problem had attracted the attention of statisticians a long time ago in areas related to statistical communication theory [cf. Ermakov (1983), Chap. 9, interrogation of parallel communication channels]. In these earlier studies, the considered problems are close to what is considered here when the covariance matrix is diagonal and the criterion of optimality is equivalent to our D-optimality criterion. Probably, the closest formulations of the problem (in a different setting) may be found in Fedorov and Mueller (1989) and Sacks and Schiller (1988). However, in the latter two publications, neither necessary and sufficient conditions of optimality nor convergence of the proposed numerical procedures was discussed.

Let the measurements be at a relatively short period, and time trends can be neglected. The following model may be applied

$$y_j(x_i) = u_j(x_i) + \varepsilon_j(x_i) \quad , \tag{4.1}$$

where $u_j(x_i)$ describes the i-th node at the j-th observation and $\varepsilon_j(x_i)$ is the corresponding observational error, $j = 1, \ldots, r_i$. Note that in (4.1) the definition of x_i is different from what has been used in the previous chapter. In this chapter it is just a "label" of the i-th site. All terms in (4.1) are assumed to be random variables. The first term, $u(x_i)$, describes the random behavior of the monitored network, while the second term is related to observational errors or short-time disturbances. As in the previous chapters, the same characters are used both for random variables and their realizations. The latter are standardly marked by additional indices [i.e., $u(x_i)$ stands for the random variable, and $u_j(x_i)$ is its realization].

Let the vector

$$U = (u(x_1), \ldots, u(x_S))^T$$

describe the network performance consisting of S nodes (sites to monitor), and let

$$E_u(U) = U_0, \;\; Var_u(U) = E_\alpha \left[(U - U_0)(U - U_0)^T \right] = K \quad ,$$

where the $S \times 1$ vector U_0 and the $S \times S$ covariance matrix K are given. The subscript u (or ε) means that expectation or variance is taken with respect to u (or ε); subscripts $\varepsilon|u$ (or $u|\varepsilon$) are used for conditional expectations. The transform $U \to U - U_0$ zeroes the expectation of U. Therefore, in what follows we assume that $E_u(U) = 0$. The observational errors $\varepsilon(x_i)$ are assumed to have zero means and to be uncorrelated:

$$E_{\varepsilon|u}\left(\varepsilon_j(x_i)\right) \equiv 0, \quad E_{\varepsilon|u}\left(\varepsilon_j(x_i)\varepsilon_{j'}(x_{i'})\right) \equiv \sigma^2 \delta_{ii'} \delta_{jj'}.$$

Introduction of σ^2 depending on x does not lead to any significant changes and is not considered here. The "label" x_i might be omitted (replaced by subscript "i") to simplify notation. However, we will continue to use it to make it easier to bridge our results with what was discussed in the previous chapters.

Note that we do not use any properties of the set X. For instance, we do not introduce a distance between two points x_i and $x_{i'}$ [cf. Fedorov (1996) and Sacks and Schiller (1988)]. The concept of a distance is much less natural in communication network measurements than in meteorology or seismology where the physical distance $\| x_i - x_{i'} \|$ between observing stations may define the behavior of elements $K(x_i, x_{i'})$ of the matrix K as functions of $\| x_i - x_{i'} \|$. Introduction of concepts similar to "distance," such as the number of switches or complexity of routes between x_i and $x_{i'}$, may lead to more efficient and realistic modeling of communication networks, but is beyond the scope of this paper.

We assume that designs are defined as:

$$\xi_n = \{p_i, x_i\}_1^n, \;\; p_i = r_i/N, \;\; N = \sum_{i=1}^{n} r_i, \;\; x_i \in X, \; n \leq S.$$

Frequently, p_i is called the weight of the node x_i. As before, nodes x_i are called support nodes/points of the design ξ_n.

Let $K(\xi_n)$ be a submatrix of K, which corresponds to the nodes $x_1, \ldots, x_n$, and let $K(x, \xi_n)$ be a column vector of covariances between $u(x)$ and $u(x_1), \ldots, u(x_n)$. We also introduce the matrices $K(Z, \xi_n) = (K(x_1, \xi_n), \ldots, K(x_q, \xi_n))$, where $x_1, \ldots, x_q \in Z \subset X$, and $K(Z)$ is a submatrix of K corresponding to these nodes, and the weight matrix $W(\xi_n)$ will be diagonal with the elements $W_{ii} = N\sigma^{-2} p_i \delta_{ii}$.

4.1.2 Best Linear Predictor

Let $Y(\xi_n)$ be the vector of averaged observations made according to ξ_n:

$$Y(\xi_n) = \begin{pmatrix} \frac{1}{r_1} \sum_{j=1}^{r_1} y_j(x_1) \\ \vdots \\ \frac{1}{r_n} \sum_{j=1}^{r_n} y_j(x_n) \end{pmatrix} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

It must be emphasized that $Y(\xi_n)$ is considered here as a random variable. One can verify by direct minimization that the predictor

$$\hat{U}(Z) = K^T(Z, \xi_n) \left(K(\xi_n) + W^{-1}(\xi_n) \right)^{-1} Y(\xi_n) \tag{4.2}$$

minimizes the matrix of expected squared residuals

$$D\left(\xi_n, \tilde{U}(Z)\right) = E_{u,\varepsilon} \left[\left(\tilde{U}(Z) - U(Z) \right) \left(\tilde{U}(Z) - U(Z) \right)^T \right]$$

among all linear estimators $\tilde{U}(Z) = LY(\xi_n)$ such that $E_{u,\varepsilon}\left[\tilde{U}(Z)\right] = 0$. The statement about unbiasness is trivial because $E\left[y(\xi_n)\right] = 0$. Thus,

$$D(\xi_n) = D\left(\xi_n, \hat{U}(Z)\right) \leq D\left(\xi_n, \tilde{U}(Z)\right) \quad , \tag{4.3}$$

where inequality must be understood in the sense of ordering of nonnegative definite matrices, see comments to inequality (3.2). From (4.1) and (4.2), it follows that

$$D(\xi_n) = K(Z) - K^T(Z, \xi_n) \left(K(\xi_n) + W^{-1}(\xi_n) \right)^{-1} K(Z, \xi_n) \quad . \tag{4.4}$$

4.2 DESIGNS WITH CONTINUOUS WEIGHTS

In what follows, we consider the methods that allow the minimization of some given functions of the matrix $D(\xi_n)$; for instance, $\mathrm{tr} D(\xi_n)$, $\ln |D(\xi_n)|$, $\max_i D_{ii}(\xi_n)$, etc., compare with Sect. 3.1. In other words, we explore the following optimization problem;

$$\xi_n^* = \arg\min_{\xi_n} \Psi \left[D(\xi_n) \right], \tag{4.5}$$

where Ψ is a selected objective function (criterion of optimality). In (4.5), the number of nodes (or supporting points) n is fixed, and the total number of available observations $N = \sum_{i=1}^n p_i$ is assumed to be given. In general, n may be optimized as well.

4.2.1 Properties of Optimal Designs

Two features of the optimization problem (4.5) may cause serious computational hurdles. Weights p_i are discrete, and the optimal number of supporting points must be found, in general. The problem is simplified both theoretically and numerically if we allow weights to be continuous so that $0 \le p_i \le 1$, $\sum_{i=1}^n p_i = 1$, and make $n = S$. If an optimal n is less than S, then some of the weights equl zero. In other words, similar to Chapter 3, instead of ξ_n^*, we are looking for approximate solutions that usually work well for larger N.

For $n = S$ and Z coinciding with X, it follows from (4.5) and the identity [see, for instance, Harville (1997), Chap. 18]: $(A + B)^{-1} = A^{-1} - A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1}$ that for any design ξ

$$D(\xi) = \left(K^{-1} + W(\xi) \right)^{-1} \quad , \tag{4.6}$$

where the subscript n is skipped to simplify notations. The matrix K is assumed regular. If Z is a subset of X, then the covariance matrix (4.4) is an obviously defined submatrix of (4.6). The subscript S will be skipped if it does not lead to ambiguity. Comparison of (4.6) with (2.24) and the consequent comments allow us to consider the covariance matrix K as a carrier of prior information about the monitored network and the matrix $W(\xi)$ as a carrier of newly accrued information.

Now we can reformulate the design problem as

$$\xi^* = \arg\min_{\xi} \left[\Psi \left(K^{-1} + W(\xi) \right)^{-1} \right], \tag{4.7}$$

where ξ can be any probability distribution with the support X.

If the function Ψ is a convex function of ξ and has a directional derivative $\psi(\xi^*, \xi)$ at ξ^* for any $\overline{\xi} = (1 - \alpha)\xi^* + \alpha\xi$ and $0 \le \alpha < 1$, then:

Theorem 4.1 *A necessary and sufficient condition for a design ξ^* to be optimal is fulfillment of the inequality*

$$\psi(\xi^*, \xi) \geq 0 \tag{4.8}$$

for any feasible design ξ.

This result is well known in optimization theory of convex functions and is widely used in experimental design theory [cf. Cook and Fedorov (1995), Fedorov and Hackl (1997), Chap. 2]. For the D-criterion, $\Psi(D) = \ln|D|$, we have found that

$$\psi(\xi^*, \xi) = \mathrm{tr} D(\xi^*) \left(W(\xi^*) - W(\xi)\right) \quad . \tag{4.9}$$

Noting that

$$\mathrm{tr}\, D(\xi)W(\xi) = \sigma^{-2} N \sum_{i=1}^{S} D_{ii}(\xi^*) p_i \quad ,$$

we derive that the following results.

Theorem 4.2 *A necessary and sufficient condition for a design ξ^* to be D-optimal (i.e., minimizing $|D(\xi)|$) is that*

$$\max_i D_{ii}(\xi^*) \leq \frac{\sigma^2}{N} \mathrm{tr} W(\xi^*) D(\xi^*) \tag{4.10}$$

and equality holds at all points where $p_i^* > 0$.

A D-optimal design also minimizes the maximal variance of prediction:

$$\xi^* = \arg\min_{\xi} \max_i D_{ii}(\xi).$$

In this theorem and in what follows $\max_i$ means maximization over all points from X (i.e., $1 \leq i \leq S$). Thus, observations in a D-optimal or minimax design must be placed at points (sites) where prediction might be the worst. It is illuminating to compare the concluding part of the above theorem with part 3 of Theorem 3.2.

For linear criteria $\Psi(D^{-1}) = \mathrm{tr} AD$, where $A \geq 0$ is the utility matrix, we have

$$\psi(\xi^*, \xi) = \mathrm{tr} D(\xi^*) A D(\xi^*) \left(W(\xi^*) - W(\xi)\right) \quad ,$$

and the following result holds.

Theorem 4.3 *A necessary and sufficient condition for a design ξ^* to be linear optimal is that*

$$\max_i \left\{D(\xi^*) A D(\xi^*)\right\}_{ii} \leq \frac{\sigma^2}{N} \mathrm{tr} W(\xi^*) D(\xi^*) A D(\xi^*) \quad ,$$

and the equality holds at all points where $p_i^* > 0$.

If $A = I$, i.e., the average variance of prediction must be minimized, then the theorem tells us that an optimal design ξ^* allocates observations at sites in which the predicted value $\hat{U}(x^*)$ of $U(x^*)$ might have the greatest average squared covariance with all other $\hat{U}(x), x \in X$. If the matrix A is diagonal and the elements A_{ii} describe the "importance" of node i, then the average squared variance is naturally replaced by the weighted average one.

Note that, unlike the results from Section 3.2 and in particular Theorem 3.2, the properties and characteristics of optimal designs considered in this section do depend on the ratio σ^2/N. For instance, in the trivial case of the diagonal covariance matrix K and minimax or $D-$ criteria, we must allocate measurements in the following manner. For the smaller N, all measurements must be done at the node with the largest variance $K_{i_1 i_1}$. As soon as

$$\frac{N}{\sigma^2} \geq \frac{1}{K_{i_2 i_2}} - \frac{1}{K_{i_1 i_1}} \quad ,$$

some part of measurements must be done at the node with the second largest variance $K_{i_2 i_2}$. With further increase in the available number of measurements, our efforts should be extended to a larger number of nodes to keep inequality (4.10) fulfilled. We have to remember that the fraction of measurements $p_i N$ at node i is treated in Theorems 4.1−4.3 as a continuous variable. With "*pinging*" or "*tracerouting*", the rounding procedure could be very simple as was discussed in the previous chapters.

4.3 FIRST ORDER ALGORITHMS

The above theorems help to develop and analyze various first order algorithms for construction of optimal designs. For computer networks the matrices processed during computations are large. It is therefore especially important to use recursions that are computationally simple and stable. The most convenient algorithms in this sense are first order exchange-type (compare with Sect. 3.3). The main idea is similar to what was proposed in that section: at each stage, add that new point that improves the current design most, and delete from the same (or just corrected) design the point that contributes least. We start with the simplest version of that kind of algorithm for D-criteria.

Let the initial design ξ_0 be such that all weights $p_{0i} = b_i \alpha_0$, where b_i is an integer and $\sum_{i=1}^{S} p_{0i} = 1$. For instance, we may choose $b_i \equiv 1$ and $\alpha_0 = 1/N$. That choice of initial weights and step length keeps the total weight equal to 1, helping to keep the computation simple.

1. Given ξ_t and $D(\xi_t)$ find

$$a = \arg\max_i D_{ii}(\xi_t). \tag{4.11}$$

   Add α_t to the weight of point x_a to construct the design ξ_t^+ and the matrix $D(\xi_t^+)$. Note that the sum of the weights in the design ξ_t^+ is greater than 1.

2. Find

$$d = \arg\min_{i \in I_t} D_{ii}(\xi_t^+) \quad , \tag{4.12}$$

where I_t is the set of all supporting points of ξ_t, i.e., points with nonzero weights at step t. Delete α_t from the weight of point x_d to modify ξ_t^+ and to construct the design ξ_{t+1} in which the sum of the weights is restored to 1 as it was in ξ_t.

3. If

$$|D(\xi_{t+1})|/|D(\xi_t)| < 1-\gamma \quad , \tag{4.13}$$

where γ is a small positive number less than 1, then put $\alpha_{t+1} = \alpha_t$ and go to (1). Otherwise make $\alpha_{t+1} = \alpha_t/2$ and then go to (1).

Computations may be stopped when α_t is sufficiently small. Another simple stopping rule may be based on the inequality that for any design ξ

$$\ln \frac{|D(\xi)|}{|D(\xi^*)|} \leq \frac{N}{\sigma^2} \max_i D_{ii}(\xi) - \mathrm{tr} W(\xi) D(\xi) \quad , \tag{4.14}$$

which is a direct corollary of the convexity of $\ln |D(\xi)|$ as a function of ξ [compare with Fedorov and Hackl (1997), Chap. 3.1].

The choice of p_{0i} and α_0 is a matter of convenience. For instance, the above choice guarantees that no more than α_t^{-1} observations are needed to avoid any "fractional" observation in design ξ_t, which is a frequent case in the continuous design theory setting. Actually, in the "classical" version of the exchange algorithm, α equals N^{-1} where N is a preselected number of observations. The algorithm with $\alpha_t \equiv N^{-1}$ was applied to the construction of spatial designs by Sacks and Schiller (1988) in the setting in which repeated observations were not allowed. Unfortunately, in this case, the limit design (if it exists) is generally not an optimal one. That is why we introduced the possibility of infinitely reducing the step length α.

The rule for adding and deleting weights becomes obvious if we note that

$$|D(\xi_t^+)| = \frac{|D(\xi_t)|}{1+\zeta_t D_{aa}(\xi_t)} \tag{4.15}$$

and

$$D(\xi_t^+) = D(\xi_t) - \frac{\zeta_t C^+(\xi_t)}{1+\zeta_t D_{aa}(\xi_t)} \quad , \tag{4.16}$$

where $C_{ij}^+(\xi_t) = D_{ai}(\xi_t) D_{aj}(\xi_t)$ and $\zeta_t = \sigma^{-2} N \alpha_t$. The above formulae may be derived using the fact that $D(\xi_t^+) = \left(K^{-1} + W(\xi_t) + \zeta_t \ell_a \ell_a^T\right)^{-1} = \left(D^{-1}(\xi_t) + \zeta_t \ell_a \ell_a^T\right)^{-1}$, where $\{\ell_a\}_i = \delta_{ia}$. In the versions of (4.15) and (4.16) for the deletion procedure ζ_t must be replaced by $-\zeta_t$ and ξ_t by ${\xi^+}_t$.

Similar to the classical results of experimental design theory, we established the following result.

Theorem 4.4 *The sequence* $\{|D(\xi_t)|\}$ *converges and*

$$\min_\xi |D(\xi)| \leq \lim_{t\to\infty} |D(\xi_s)| \leq (1-\gamma)^{-1} \min_\xi |D(\xi)|.$$

The proof is based on monotonicity of the iterative procedure and convexity of $\ln|D(\xi)|$ as a function of ξ and Theorem 4.2.

The iterative procedure 1-3 admits various improvements. For instance, the number of "forward" steps (i.e., the length of the forward excursion), may be selected by a user instead of being equal to 1 as in the original formulation. Consequently, the same number of "backward" steps must be done to keep the total weight of supporting points in ξ to be equal to 1. An alternative to the user defined length of the excursion can be the continuation of forward steps until the increase of the standardized determinant $(1+\ell\alpha)^{-S}|D(\xi_\ell^+)|$ is decreasing. Here, ℓ is the number of accomplished forward steps, and ξ_ℓ^+ is an "extended" design with the total weight $1+\ell\alpha$. Obviously, the backward excursion must return the total weight to 1. Note that formulae (4.15), (4.16), and their siblings for the deletion steps are convenient recursions for large size-problems.

4.4 ESNET EXAMPLE II

4.4.1 Covariance Matrix Estimation and Optimal Design

We have used the Department of Energy's ESnet backbone, a portion of the Internet, as a testbed for the proposed numerical procedure. Using a network host computer at ORNL, we interrogated 39 other sites (see Table 4.1) to construct a reasonable estimate $\hat{K}$ for the matrix K and to use $\hat{K}$ in the numerical procedure instead of K. We have used the script written by T. Dunigan (ORNL), which is based on the (*ping*) software (see Stevens(1994), Chap. 7) to measure the response time for each interrogation. Because of the lower priority that network routers may give to *ping* requests, the minimum response time (among three *ping* requests per interrogation) is used as the response variable. This also reduces the probability of "missing" observations (i.e., there is more hope that at least one *ping* out of three will result in a response). All 39 sites were interrogated 50 times in random order during approximately 3 hours on a weekday in mid-March 1997. We estimated the elements of matrix K for each pair of sites separately without imposing any conditions like positive-definiteness of $\hat{K}$, for instance. We have to abandon the simplest and traditional estimator (see notations in Sect. 4.1)

$$\hat{K} = \frac{1}{k}\sum_{\ell=1}^{k}(Y_\ell - \bar{Y})(Y_\ell - \bar{Y})^T, \ \bar{Y} = \frac{1}{k}\sum_{\ell=1}^{k} Y_\ell \quad , \tag{4.17}$$

which guarantees non-negativeness of the matrix $\hat{K}$, because of the total number of interrogations with all 39 components of the vector Y reported is much less than 39. If $K < 39$, then rank $\hat{K} < 39$. In addition, the fact that the matrix K is not very well estimated (and that is not important for a pure illustration example), the singularity of the matrix $\hat{K}$ causes the formal obstacle: we cannot compute the matrix $\hat{K}^{-1}$. The information on more sophisticated methods of estimation of covariance matrices can be found in Dixon (1992), Vol. 2, Chap. 8D; Little and Rubin (1987), Chap.. 1 and 3; Muirhead (1982) Chap. 4.3.

The value of the standard error σ was estimated by averaging differences between results of neighboring in time interrogations over the whole set of interrogations for all sites. We found that $\sigma \simeq 8.0ms$. This may not be the best estimator, especially if one takes into account heterogeneity of ESnet. However, for our illustrative purposes, it is not important. For the real applications, the use of more sophisticated estimators is crucial because the form of an optimal design depends on both K and σ^2/N.

Table 4.1: Site identifiers

ID	Site Acronym	Site Name
1	JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)
2	ARM	Atmospheric Radiation Measurement Project (Lamont, OK)
3	FNAL	Fermi National Accelerator Laboratory (Batavia, IL)
4	SNL	Sandia National Laboratories Albuquerque (Albuquerque, NM)
5	KEK	KEK, Japan
6	NYU*	New York University Courant Institute (New York, NY)
7	MSRI*	Mathematical Sciences Research Institute, Univ. CA (Berkeley, CA)
8	ANL-MR1	Argonne National Laboratory-Main Router 1
9	AMES	AMES Laboratory, Iowa State University (Ames, IA)
10	FSU*	Florida State University (Tallahassee, FL)
11	CIT	California Institute of Technology (Pasadena, CA)
12	MIT*	Massachusetts Institute of Technology (Cambridge, MA)
13	FNAL-MR1	Fermi National Accelerator Laboratory-Main Router 1
14	GAT	General Atomics (San Diego, CA)
15	UTA	University of Texas at Austin (Austin, TX)
16	SRS*	Savannah River Site (Aiken, SC)
17	SLAC	Stanford Linear Accelerator (Stanford, CA)
18	INEL*	Idaho National Engineering Laboratory (Idaho Falls, ID)
19	LLNL	Lawrence Livermore National Laboratory (Livermore, CA)
20	AUCK*	University of Auckland (Auckland, New Zealand)
21	DOE	Department of Energy (Washington, DC)
22	PPPL	Princeton Plasma Physics Laboratory (Princeton, NJ)
23	UTK	University of Tennessee (Knoxville, TN)
24	LANL-MR1	Los Alamos National Laboratory - Main Router 1 (Los Alamos, NM)
25	NASA*	AMES Research Center, NASA (San Francisco, CA)
26	BNL	Brookhaven National Laboratory (Upton, NY)
27	PPPL-local	additional PPPL site
28	CU	Columbia University Academic Information Systems (New York, NY)
29	ANL	Argonne National Laboratory (Argonne, IL)
30	Pro.PPPL	additional PPPL site
31	PNNL*	Pacific Northwest National Laboratory (Richland, WA)
32	OSTI	Office of Scientific and Technical Information (Oak Ridge, TN)
33	NEVIS*	Columbia University Nevis Laboratory (Irvington, NY)
34	LBNL-MR1	Lawrence Berkeley National Laboratory - Main Router 1
35	LLNL-MR2	Lawrence Livermore National Laboratory - Main Router 2
36	NERSC*	National Energy Research Scientific Computing, LBNL (Berkeley, CA)
37	LBNL	Lawrence Berkeley National Laboratory (Berkeley, CA)
38	SNL/LLNL	Sandia National Laboratories at LLNL (Livermore, CA)
39	YALE*	Yale University (New Haven, CT)

Table 4.2: Various designs to monitor ESNet sites ($N = 10, \sigma = 8.0ms$)

Site ID	Site	Site Variance	D-optimal Weight	Variance of the Prediction, D_{ii}		
				D-optimal Continuous	Rounded 10 points	Uniform All 39 points
1	JLAB	96.8	.0000	23.7	24.2	12.8
2	ARM	72.0	.0000	35.1	42.0	25.1
3	FNAL	90.7	.0000	21.2	23.2	8.2
4	SNL	8.8	.0000	2.8	2.8	1.9
5	KEK	56.2	.0000	31.1	31.5	18.4
6	NYU	1042.7	.0970	58.0	56.4	176.7
7	MSR1	289.2	.0128	57.9	61.2	39.8
8	ANL-MR1	100.9	.0000	32.7	32.6	19.8
9	AMES	83.5	.0000	26.0	27.9	11.7
10	FSU	1497.7	.1010	58.0	58.5	194.8
11	CIT	19.3	.0000	8.3	8.9	4.9
12	MIT	1002.6	.1000	57.8	57.9	187.6
13	FNAL-MR1	6.1	.0000	5.4	5.5	3.8
14	GAT	60.5	.0000	21.5	23.0	13.2
15	UTA	54.0	.0000	17.7	18.4	12.3
16	SRS	978.5	.0985	57.9	57.1	181.3
17	SLAC	93.0	.0000	26.0	31.8	14.9
18	INEL	1537.9	.1010	58.0	58.5	194.7
19	LLNL	57.8	.0000	27.7	28.2	15.5
20	AUCK	1967.8	.1036	57.9	60.0	210.8
21	DOE	25.3	.0000	12.4	12.7	6.9
22	PPPL	41.9	.0000	11.4	12.2	5.2
23	UTK	0.4	.0000	0.3	0.3	0.3
24	LANL-MR1	51.7	.0000	21.1	21.8	10.7
25	NASA	949.6	.0978	58.9	56.8	177.8
26	BNL	183.8	.0000	56.7	56.8	33.9
27	PPPL-local	126.0	.0000	48.2	51.3	23.1
28	CU	75.3	.0000	23.9	23.9	15.7
29	ANL	88.3	.0000	22.4	24.1	10.1
30	Pro.PPPL	35.3	.0000	15.6	15.6	9.4
31	PNL	365.1	.0853	57.9	51.0	130.1
32	OSTI	0.4	.0000	0.3	0.3	0.3
33	NEVIS	402.1	.0809	57.9	52.8	70.3
34	LBNL-MR1	62.4	.0000	16.4	17.4	7.0
35	LLNL-MR2	58.9	.0000	25.5	26.9	12.7
36	NERSC	121.2	.0236	58.0	74.7	47.2
37	LBNL	86.7	.0000	21.3	22.0	10.7
38	SNL/LLNL	131.8	.0000	43.1	46.3	20.9
39	YALE	1137.1	.0987	58.0	57.3	183.6

Table 4.3: Dependence of the optimal design structure on σ^2/N.

Site ID	Site	Site Variance	σ^2/N 500	10	1	0.1
1	JLAB	96.8				
2	ARM	72.0				.046
3	FNAL	90.7				
4	SNL	8.8				
5	KEK	56.2			.028	.046
6	NYU	1042.7	.024	.100	.080	.062
7	MSR1	289.2			.038	.052
8	ANL-MR1	100.9				
9	AMES	83.5				
10	FSU	1497.7	.218	.106	.080	.064
11	CIT	19.3				
12	MIT	1002.6	.052	.104	.080	.064
13	FNAL-MR1	6.1				
14	GAT	60.5				.008
15	UTA	54.0				.042
16	SRS	978.5	.026	.102	.080	.064
17	SLAC	93.0				
18	INEL	1537.9	.220	.106	.080	.064
19	LLNL	57.8				.006
20	AUCK	1967.8	.328	.110	.080	.064
21	DOE	25.3				
22	PPPL	41.9				
23	UTK	0.4				
24	LANL-MR1	51.7				
25	NASA	949.6		.100	.080	.064
26	BNL	183.8			.054	.056
27	PPPL-local	126.0			.038	.030
28	CU	75.3				
29	ANL	88.3				
30	Pro.PPPL	35.3				
31	PNL	365.1		.082	.076	.062
32	OSTI	0.4				
33	NEVIS	402.1		.088	.062	.060
34	LBNL-MR1	62.4				
35	LLNL-MR2	58.9				
36	NERSC	121.2			.064	.058
37	LBNL	86.7				
38	SNL/LLNL	131.8				.026
39	YALE	1137.1	.132	.102	.080	.062

There exists another problem with the estimator (4.17), which was not mentioned by Fedorov and Flanagan (1997), but deserves to be discussed. From the definition of vector Y [see (4.1)], it follows that

$$E(\hat{K}) = K + \Sigma \quad , \tag{4.18}$$

where Σ is a diagonal matrix and the diagonal elements Σ_{ii} describe the variance generated by the measurement error ε_i. If at each single measurement of the i-th component r_i repeated *ping* interrogations are performed, then $\Sigma_{ii} = \sigma^2/r_i$. Thus, the estimator (4.17) must be corrected by subtracting an independent estimator $\hat{\Sigma}$ of the matrix Σ. This correction may lead to non-positive definite estimators of K. The similar corrections must be done for the pairwise estimators of elements of the matrix K, and, of course, with the similar side effects. We postpone further discussion of the problem and for this example retain the simplest estimators

$$\hat{K}_{ij} = \frac{1}{a_{ij}} \sum_{\ell \in A_{ij}} (y_{i\ell} - \bar{y}_i)(y_{jl} - \bar{y}_j) \quad , \tag{4.19}$$

where A_{ij} is a set of all measurements in which both y_i and y_j are included and α_{ij} is the size of this set. The estimators of means $\overline{y}_i$ and $\overline{y}_j$ may be averages with respect to observations available for every component.

The first order algorithm was applied to the matrix constructed according to (4.19). The diagonal elements of matrix $\hat{K}$ appear in Table 4.2. The table also reports the optimal weights for each site and the variance of prediction (the diagonal elements D_{ii}) for the D-optimal continuous and rounded designs. Also reported is the variance of prediction for the uniform design containing all 39 sites. The optimal design is nearly four times more efficient than the 39-point uniform design. The number of *pings* sent to a particular site must be proportional to the corresponding weight. In practice, we use "rounded" weights. This rounding may lead to some increase in the maximal variance of prediction. For instance, when only 10 points with the largest weight are selected and all their weights are set to 0.1, then $\max_i D_{ii} = 74.0$, which is not significantly larger than $\max_i D_{ii}$ for the continuous D-optimal design.

We selected a relatively small number of available observations ($N = 10$) to emphasize the difference between continuous and discrete designs. In reality, it takes a few minutes to send hundreds of *pings* to different sites. Therefore, the approximation of reasonable weights is not a serious problem in that type of experiment. Actually, one may introduce the optimal partitioning of available time periods for a given experimental session for monitoring various sites instead of the selection of an optimal number of *pings*.

It seems that covariances between different sites do not play a very important role. All measurements should be at the sites with the largest variances. This fact agrees with our comments at the end of Sect. 3.3.

As we have mentioned before, the structure of optimal designs depends on the ratio σ^2/N. Table 4.3 contains the optimal designs for $\sigma^2/N = 500,\ 10,\ 1,\ 0.1$. The tendency is clear: for the larger measurement errors (or for the smaller number of available measurement), all efforts must be directed to measure at the sites having the largest variance. However, starting with $\sigma^2/N = 1$ some sites (see # 6 and # 36 for the design with $\sigma^2/N = 1$), have variances that are less than the variances at the sites not included in the optimal design (see # 38 for the design with $\sigma^2/N = 1$).

This means that the designs with the small ratio σ^2/N depend on the covariances between different sites.

The following observation can be useful for a practitioner. All support points from the optimal designs with the larger ratio σ^2/N have the support sets that are the subsets of the optimal designs with the smaller ratio σ^2/N. Therefore, it is reasonable to start measurements at sites with the large variances and to place the rest of the available measurements at other support points. This recommendation, to some extent, contradicts the "randomization" principle, which is popular in the experimental design theory. Randomization usually helps to avoid the adverse impact of spatial or longitudinal time trends. Perhaps, some compromising approaches like stratified randomization may help to follow our recommendation and still to avoid effects of time trends that are important only if measurement on networks take a relatively long period.

4.4.2 Modeling the Covaraince Matrix

Following the hint at the end of Sect. 4.4.1, let us try to model the covariance matrix k. To do this we can, for instance, use the information reported by the "*traceroute*" software [see Stevens (1994), Chap. 8], which reports all edges on the network graph traversed to reach a destination node. Delays for each edge also can be evaluated from the above mentioned information.

We assume that during the interrogation process (based on "*ping*" software) the host and the destination node are connected by a single route, which coincides with the route most frequently reported by the "*traceroute*" software. This assumption essentially simplifies modeling and is sufficiently accurate for our illustrative needs.

Each route i may be described by the vector ζ_i containing q components, where q is the number of different edges among all routes connecting the host site (ORNL) and all destination nodes (39 in the considered case). Similar to Sect. 2.1.1 $\zeta_{i\alpha} = 1$, if the α-th edge is included in the i-th route, and $\zeta_{i\alpha} = 0$ otherwise, $\alpha = 1, \ldots, q$.

Let the $(q \times q)$ matrix Λ describe our "theoretical" knowledge of characteristics of the q edges. In the simplest cases Λ may be diagonal and $\Lambda_{\alpha\alpha} = \lambda_\alpha^2$, where λ_α may be, for instance, the delay time at the α-th edge.

Assuming that the covariance between traveling times from the host site to nodes i and j is explained by the time intervals that interrogating "*pings*" spend on common edges, we may conclude that [compare with (2.19)]

$$K_{ij} = \zeta_i^T \Lambda \zeta_j \quad .$$

For our numerical example we use the rough approximation $\Lambda_{\alpha\alpha} \equiv 100$. The particular value of $\Lambda_{\alpha\alpha}$ is selected to make comparable the scales from this section and from the previous one. The results of computations are described in Table 4.4, which is similar to Table 4.3. As before, the support points of the optimal designs coincide with the nodes having the larger variances k_{ii}. The non-zero covariance influence is noticeable starting from $\sigma^2/N \leq 100$.

Table 4.4: Dependence of the optimal design structure on σ^2/N, k is modeled.

Site ID	Site	Site Variance	σ^2/N 500	100	10
1	JLAB	300			0.03
2	ARM	400			0.03
3	FNAL	500		0.06	0.04
4	SNL	400		0.01	0.03
5	KEK	700		0.10	0.04
6	NYU	300			0.03
7	MSR1	200			
8	ANL-MR1	200			
9	AMES	400		0.01	0.03
10	FSU	500		0.06	0.04
11	CIT	500		0.06	0.04
12	MIT	500		0.06	0.04
13	FNAL-MR1	200			0.01
14	GAT	300			0.03
15	UTA	800	0.11	0.13	0.05
16	SRS	200			0.01
17	SLAC	300			0.03
18	INEL	1300	0.35	0.18	0.05
19	LLNL	400		0.01	0.03
20	AUCK	1400	0.37	0.18	0.05
21	DOE	300			0.02
22	PPPL	300			0.01
23	UTK	300			0.03
24	LANL-MR1	200			
25	NASA	300			0.01
26	BNL	400			0.03
27	PPPL-local	300			0.02
28	CU	400			0.03
29	ANL	400			0.03
30	Pro.PPPL	300			0.02
31	PNL	400			0.03
32	OSTI	200			0.01
33	NEVIS	500			0.02
34	LBNL-MR1	200			
35	LLNL-MR2	200			
36	NERSC	300			0.02
37	LBNL	400			0.03
38	SNL/LLNL	400			0.03
39	YALE	900	0.17	0.14	0.05

4.5 SIMPLE HEURISTIC ALGORITHM

4.5.1 Short Survey of the Older Results

In the ESnet example, we use the "plug in" idea, that is, all unknown elements in the design procedure (in our case the covariance matrix K and the variance of measurement error σ^2) must be estimated using preliminary measurements, and then the corresponding estimates are to replace unknown values. A rigorous mathematician may argue with the applicability of that idea in general, but intuitively, one expects to get designs that are close to optimal if the preliminary data set was informative enough. In most cases the "plug in" approach is practical and effective. Interestingly enough, sometimes the two step design procedure (find estimates – compute design) can be replaced by a computationally more effective one-step procedure. Moreover, that replacement helps to illuminate some basic facts about optimal designs.

The procedure, which we intend to discuss was invented in studies related to meteorology [see Megreditchan (1979, 1989)]. It is intuitive and simple. Let $\{y_\ell(x_i)\}_{11}^{rn}$ be a data set accumulated by n observing stations (compare with nodes/sites) during $\ell = 1, \ldots, r$ observing session. In meteorology they observe precipitation, temperature, atmospheric pressure. In our case, there may be delays, packet loss, reachability, etc.

Now we want to reduce the total number of sites (stations in meteorology, but from now on we shall use network terminology) in the belief that data collected on all of them are redundant. The redundancy means that the behavior of some sites can be explained by measurements that are made at other sites. One of the simplest "explanatory" models is the linear regression:

$$y_\ell(x_{i'}) = \sum_{i \neq i'} \theta_i y_\ell(x_i) + \varepsilon_\ell \quad , \tag{4.20}$$

where ε_j comprises whatever is unexplained and uncertain. As soon as this model is selected, the sums of squared residuals

$$v(x_{i'}) = \min_\theta \sum_{l=1}^{r} \left[\sum_{i \neq i'} \theta_i y_\ell(x_i) - y_\ell(x_{i'}) \right]^2 \tag{4.21}$$

must be computed for all $i' = 1, \ldots, n$. The site i^* with the least $v(x_{i'})$ is considered "well explained" by other sites and is removed from the monitoring set and from the consequent computations. The procedure is repeated until either the number of sites is small enough to guarantee the cost range of future measurements or until the values of $u(x_{i'})$ increase dramatically compared to their initial values.

In practice (at least in meteorology), the approach worked very well. Fedorov and Mueller (1989) have found that its efficiency can be theoretically explained for a rather broad class of practical problems. In particular, the above procedure coincides with the backward excursion of the exchange type algorithm of optimal experimental design for the regression model with random coefficients. This model can be used to approximate the covariance structure of the observed random fields in meteorology. Here we show that the similar result is true for the problem of selection of optimal monitoring network.

4.5.2 Approximate Duality of Two Approaches

Let us assume that the estimator $\hat{K}$ defined by (4.2) can be used, that is, we assume that in all measurements, all components of the vector Y are successfully measured. Without loss of generality, we can select $i' = 1$ and present $\hat{K}$ as

$$\hat{K} = \frac{1}{r} \begin{pmatrix} z_1^T z_1 & z_1^T Z_{-1} \\ Z_{-1}^T z_1 & Z_{-1}^T Z_{-1} \end{pmatrix} = \begin{pmatrix} \hat{K}_{1,1} & \hat{K}_{1,-1} \\ \hat{K}_{-1,1} & \hat{K}_{-1,-1} \end{pmatrix} \quad , \tag{4.22}$$

where

$$z_1^T = (y_1(x_1), \ldots, y_r(x_1)) \quad ,$$

$$\underline{Z}_1^T = \begin{pmatrix} y_1(x_2) & \ldots & y_r(x_2) \\ \vdots & \vdots & \vdots \\ y_1(x_n) & \ldots & y_r(x_n) \end{pmatrix} = \begin{pmatrix} z_2^T \\ \vdots \\ z_n^T \end{pmatrix} \quad .$$

If at every interrogation session k *pings* are sent, than $y_\ell(x_i)$ is an average of k measured travel times.

Let us assume that, suggested by (4.20), (4.21), we "regress" z_1 on $z_2, \ldots, z_n$. It is known that the sum of squared residuals reaches its minimum if [compare with Section 2.1 and Rao (1973), Chap. 4a]

$$\hat{\theta}_1 = (Z_{-1}^T Z_{-1})^{-1} Z_{-1}^T z_1 \quad , \tag{4.23}$$

and that the minimum equals

$$v(x_1) = z_1^T z_1 - z_1^T Z_{-1} (Z_{-1}^T Z_{-1})^{-1} z_{-1} \quad . \tag{4.24}$$

Two presentations, (4.23) and (2.11) are identical because, for the linear response function $\sum_{i \neq 1} \theta_i y_i$ we have $\underline{M} = Z_{-1}^T Z_{-1}$ and $Y = Z_{-1}^T z_1$. To bridge the Megreditchan idea and the numerical method proposed in Sect. 4.4, let us note that

$$\frac{1}{r} v(x_1) = \hat{K}_{1,1} - \hat{K}_{1,-1} \hat{K}_{-1,-1} \hat{K}_{-1,1} \quad .$$

For the larger r due to the consistency of the estimator $\hat{K}$, we can state that

$$\hat{K} \simeq K + I\sigma^2/k \quad ,$$

where, as before, σ^2 is the variance of the measurement error, k is a number of repeated measurements at every ℓ-th session, and I is the identity matrix. Combining the two last expressions we come to the following approximation

$$\frac{1}{r}v(x_1) \simeq K_{1,1} + \sigma^2/k - K_{1,1}(K_{-1,-1} + I\sigma^2/k)^{-1}K_{-1,1} = \sigma^2/k + D_{11}(\xi_u) \quad , \tag{4.25}$$

where ξ_u is the design with uniformly distributed weight (i.e., $p_i = 1/n,\ N = nr$). To derive (4.25) we applied the formula for the inversion of partitioned matrices [see, for instance, Harville (1997) Chap. 8.5]. Using permutation we can easily verify the more general result:

$$\frac{1}{k}v(x_i) \simeq \sigma^2/k + D_{ii}(\xi_u) \quad . \tag{4.26}$$

From this approximation it follows that the minimization of $v(x_i)$ has approximately the same solution as the minimization of $D_{ii}(\xi_u)$, that is, we delete the site i^*, if

$$i^* = \arg\min_i D_{ii}(\xi_u) \quad . \tag{4.27}$$

Comparison of (4.26) and the deleting stage of the iterative procedure from Sect. 4.3 shows that deleting redundant sites accordingly to the Megreditchan method is nothing else but the backward excursion with $\alpha = \frac{1}{k}, \frac{1}{k-1}, \ldots$ for minimax or $D-$ criteria. That is probably why the Megreditchan method leads to very reasonable subsets. Thus, instead of computing $\hat{K}$ and consequent application of the "plug in" approach we may apply the least square method (software is widely available) to run the backward excursions and delete redundant sites.

Duality between these two approaches allows extension of the Megreditchan idea and inclusion of the forward excursions. Indeed, after deleting q_1 sites, we can think about adding $q_2 < q_1$ sites, then deleting q_3, then adding back $q_4 < q_3$ sites, etc. The site that must be added in each case is defined by finding

$$i^* = \max_i v(x_i) \quad , \tag{4.28}$$

where i belongs to set of the previously deleted sites.

The original and generalized Megreditchan methods are computationally simple and intuitively attractive, and we recommend their use. However, the following facts must be remembered:

1. The method cannot be used directly when measurements contain missing components.
2. The step length α does not diminish. Therefore theoretically, we cannot guarantee convergence to an optimal design.
3. The method improves the $D-$ or minimax criterion. For other criteria, different deleting/adding rules would be necessary. To our knowledge they have not been proposed so far.

References

Albert, A. (1972). *Regression and the Moore-Penrose Pseudoinverse*, Academic Press, New York.

Akahira M. and K. Takeuchi (1995). *Non-Regular Statistical Estimation*, Springer, New York.

Anderson, T. W. (1994). *The Statistical Analysis of Time Series*, Wiley Classics Library Edition, Wiley, New York.

Athreya, K. B., S. N. Lahiri, and Wei Wu (1998). "Inferences for Heavy Tailed Distributions," *JSPI*, **66**, 61-75.

Atkinson, A. C. and A. N. Donev (1992). *Optimal Experimental Design*, Clarendon Press, Oxford.

Atkinson, A. C. and V. V. Fedorov (1988). "Optimum Design of Experiments in Presence of Uncontrolled Variability and Prior Information," in *Optimal Design and Analysis of Experiments*, eds. Y. Dodge, V. Fedorov, and H. Wynn, North Holland, New York.

Bandemer M., A. Bellman, W. Jung, Le Anh Sor, J. Pilz, and K. Richter (1977). *Theorie und Anwendung der Optimalen Versuchplanung*, Akademie-Verlag, Berlin.

Batsell, S., D. Downing, T. Dunigan, and V. Fedorov (1997). "Poisson Type Models and Descriptive Statistics of Computer Network Information Flows," *ORNL/TM-1368*, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Batsell, S., V. Fedorov, and D. Flanagan (1998). "Multivariate Prediction: Selection of the Most Informative Components to Measure," accepted for *Proceedings of MODA-5 International Workshop*, Springer-Verlag, New York.

Beran, J. (1994). *Statistics for Long-Memory Process*, Chapman and Hall, New York.

Box, G. E. P. and N. R. Draper (1987). *Empirical Model-Building and Response Surfaces*, Wiley, New York.

Box, G. E. P., W. G. Hunter, and J. S. Hunter (1978). *Statistics for Experiments. An Introduction to Design, Analysis, and Model Building*, Wiley, New York.

Box, G. E. P. and G. C. Tiao (1992). *Bayesian Inference in Statistical Analysis*, Wiley Classics Library Edition, Wiley, New York.

Brownlee, N. (1995). "New Zealand Experiences with Network Traffic Charging," presented at the *MIT Workshop on Internet Economics*, March 1995.

Brownlee, N. (1996). *Reference Manual NeTraMet and NeMac*, University of Auckland, New Zealand.

Cheng, Chih-Hsu (1991). *Optimal Sampling for Traffic Volume Estimation*, Ph.D. Dissertation,

Carlson School of Management, University of Minnesota, December 1991.

Claffy, K. C. (1994). *Internet Traffic Characterization*, Ph.D. Dissertation, Computer Science and Engineering, University of California, San Diego.

Conover, W. J. (1980). *Practical Nonparametric Statistics*, Second Edition, Wiley, New York.

Cook, R. D. and C. J. Nachtsheim (1980). "A Comparison of Algorithms for Constructing Exact D-optimal Designs," *Technometrics*, **22**, 315-324.

Cook, R. D. and V. V. Fedorov (1995). "Constrained Optimization of Experimental Design," *Statistics*, **26**, 129-178.

Cottrell, L., G. Haney, T. Healy, L. Logg, D. Martin, L. White, and W. Wing (1997). "ESnet's Internet Monitoring Activities," *IEEE Communications*, January 1997, ????.

Cottrell, L., and W. Mathews (1998). *Report of the ICFANTF Monitoring Working Group*, http://www.slac. stanford.edu/xorg/icfal/ntf/mon-wg-report-may98.html.

Cox, D. R. (1967). *Renewal Theory*, Science Paperbacks and Methuen & Co. Ltd., London.

Cressie, N. A. (1991). *Statistics for Spatial Data*, Wiley, New York.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *JRSS (B)*, **39**, 1-38.

Dixon, W. J. (ed.) (1992). *BMDP Statistical Software Manual, Volume 2*, University Press of California.

Ermakov, S. M. (ed.) (1983). *Mathematical Theory of Experimental Design*, Nauka, Moscow (in Russian).

Fedorov, V. (1972). *Theory of Optimal Experiments*, Academic Press, New York.

Fedorov, V. (1974). "Regression Problems with Controllable Variables Subject to Error," *Biometrika*, **61**, 49-56.

Fedorov, V. (1996). "Design of Spatial Experiments: Model Fitting and Prediction," in *Handbook of Statistics, Volume 13*, eds. S. Ghosh and C. Rao, Elsevier Science, New York, 515-553.

Fedorov, V. and A. Atkinson (1988). "The Optimum Design of Experiments in the Presence of Uncontrolled Variability and Prior Information," in *Optimal Design and Analysis of Experiments*, eds. Y. Dodge, V. Fedorov, and H. Wynn, North-Holland, Amsterdam.

Fedorov, V. and D. Flanagan (1998). "Optimal Monitoring Network Design Based on Mercer's Expansion of Covariance Kernel," to appear in *Journal of Combinatorics and Information Science.*

Fedorov, V. and D. Flanagan (1997). "Optimal Monitoring of Computer Networks," presented at *1997 AMS/SIAM/IMS Joint Summer Research Conference in the Mathematical Sciences*, June 28, 1997.

Fedorov, V. and P. Hackl (1997). *Model-Oriented Design of Experiments*, Springer-Verlag, New York.

Fedorov, V. and V. Khabarov (1986). "Quality of Optimal Designs for Model Discrimination and Parameter Estimations," *Biometrika*, **73**, 183-190.

Fedorov, V. and W. Mueller (1989). "Comparison of Two Approaches in the Optimal Design of an Observation Network," *Statistics*, **20**, 339-351.

Fedorov, V. and A. Uspensky (1975). *Numerical Aspects of the Least Squares and Design of Experiments*, Moscow State University Publishing House, Moscow.

Flanagan, D. (1997). *Optimal Monitoring Systems*, Ph.D. Thesis, University of Tennessee, Knoxville.

Floyd, S. (1994). "Wide Area Traffic: The Failure of Poisson Modeling," *SIG COMM 1994 Conference*, Lawrence Berkeley National Laboratory, California, 325238, CONF-9408202-2.

Frost. V. and B. Melamed (1994). "Traffic Modeling for Telecommunications Networks," *IEEE Communications Magazine*, March, 70-81.

Gaffke, N. and R. Mathar (1992). "On a Class of Algorithms from Experimental Design Theory," *Optimization*, **24**, 91-126.

Gaffke, N. and B. Heiligers (1996). "Second Order Methods for Solving Extremum Problems from Optimal Linear Regression Design," *Optimization*, **36**, 41-57.

Harville, D. (1997). *Matrix Algebra from a Statistician's Perspective,* Springer-Verlag, New York.

Hengartner, N. W. (1997). "Adaptive Demixing in Poisson Mixture Models", *Ann. Statist.*, **25**, 917-928.

Jain R. and S. Routhier (1986). "Packet Trains - Measurements and New Models for Computer Network Traffic," *IEEE Journal of Selected Areas in Communications*, SAC **4**(6), 986-995.

Johnson N., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distribution, Volume 1, Second Edition*, Wiley, New York.

Karlin S. and W. Studden (1996). *Tchebysheff Systems: With Applications in Analysis and Statistics*, Wiley, New York.

Kiefer, J. (1958). "On the Nonrandomized Optimality and Randomized Nonoptimality of Symmetrical Designs," *Amer. Math. Statist.*, **29**, 675-699.

Kiefer, J. (1959). "Optimal Experimental Designs," *Journal of the Royal Statistical Society, Series B*, **21**, 272-319.

Leland, W., N. Taqqu, W. Willinger, and D. Wilson (1994). "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/AC Transactions of Networking*, **2**(1).

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Malyutov, M. (1987). "Design and Analysis in Generalized Regression Model F," in *Model-Oriented Date Analysis*, eds. V. Fedorov and M. Lauter, Springer-Verlag, New York.

Martin, R. J. (1996). "Spatial Experimental Design," in *Handbook of Statistics*, eds. S. Shosh and C. R. Rao, Elsevier Science, New York, **13**, 477-514.

Megreditchan, G. (1979). "Optimization des reseaux d'obseration des champs meteorologiques," *La Meteorologie*, **6**, 51-66.

Megreditchan, G. (1988). "Statistical Redundancy as a Criterion for Meteorological Network Optimization," *Österreich. Z. Statist. Informatik*, **19**, 18-29.

Mitchell, T. (1974). "An Algorithm for Construction of D-Optimal Design," *Technometrics*, **16**, 203-210.

Muirhead, R. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

Nguen, N. K. and A. J. Miller (1992). "A Review of Some Exchange Algorithms for Constructing Discrete D-Optimal Designs," *Computational Statistics and Data Analysis*, **14**, 489-498.

Paxson, V. (1995). "Fast Approximation of Self-Similar Network Traffic," Lawrence Berkeley National Laboratory, California, 36750, CONF-9508118-1.

Paxson, V. (1996). "End-to-End Routing Behavior in the Internet," *Proceedings of SIG COMM*, August 1996, 25-38.

Paxson, V. (1996). "Towards a Framework for Defining Internet Performance Metrics," *Proceedings of INET'96*, Montreal, June.

Paxson, V. (1997). "Measurement and Analysis of End-to-End Internet Dynamics," Ph.D. Thesis, University of California, Berkeley, UCB//CSD-97-945, Lawrence Berkeley National Laboratory, LBNL-40319.

Paxson, V. and S. Floyd (1995). "Wide-Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Transactions on Network*, **3**, 226-244.

Pazman, A. (1986). *Foundations of Optimum Experimental Design*, Reidel, Dordrecht.

Pilz, J. (1991). *Bayesian Estimation and Experimental Design in Linear Regression Models*, Wiley, New York.

Pukelsheim, F. (1993). *Optimal Design of Experiments*, Wiley, New York.

Quarterman, J (1990). *The Matrix: Computer Networks and Conferencing Systems Worldwide*, Digital Press, Bedford, MA.

Raghavarao, D. (1971). *Construction and Combinatorial Problems in Design of Experiments*, Wiley, New York.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications, Second Edition*, Wiley, New York.

Reingold, E. M., J. Nievergelt, and N. Deo (1977). "Combinational Algorithms: Theory and Practice," Prentice Hall, Engelwood Cliffs, New Jersey.

Sacks, J. and S. Schiller (1988). "Spatial Design" In *Statistical Decision and Related Topics IV* (Burger J. and S. Gupta, eds.), Volume 2, pp. 385-395, Academic Press, New York.

SAS Software (1995). *SAS/Stat User's Guide (1995). Volume 2*, SAS Institute Inc., Cary, North Carolina.

SAS/QC Software (1995). *Design of Experiments Tools*, SAS Institute Inc., Cary, North Carolina.

Samorodnitsky, G. and Taqqu M. S. (1994). *Stable Non-Gaussion Processes: Stochastic Models with Infinite Variance*, Chapman and Hall, New York.

Seber, G. A. F. and C. J. Wild (1989). *Nonlinear Regression*, Wiley, New York.

Silvey, D. S. (1980). *Optimal Design*, Chapman and Hall, London, England.

Snyder, D. (1975). *Random Point Processes*, Wiley, New York.

Stevens, W. (1994). *TCP/IP Illustrated. Volume 1*, Addison-Wesley, Reding, Massachusetts.

Stigler, S. (1974). "Gergonne's 1815 Paper on the Design and Analysis of Polynomial Regression Experiments," *Historia Mathematica*, **1**, 431-447.

SPSS/Trail Run 1.0 Software (1997). "Comprehensive Experimental Design and Analysis Made Easy," *Eureka*, **1**(3), 18.

Vardi, Y. (1996). "Network Tomography: Estimation Source - Destination Traffic From Link Data," *JASA*, **91**, 365-377.

Wheeler, B. (1994). *ECHIP: Version 6.0 for Windows*, ECHIP, Inc., Hockessin.

Willinger, W., M. S. Taqqu, W. E. Leland, and V. Wilson (1995). "Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements," *Statistical Science*, **10**, 67-85.

Willinger, W., M. S. Taqqu, R. Sherman, and D. Wilson (1995). "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," in *Proc. of the ACM/SIG COMM 95*.

Willinger, W., M. S. Taqqu, and A. Erramilli (1996). "A Bibliographical Guide to Self-similar Traffic and Performance Modeling for Modern High-speed Networks," in *Stochastic Networks Theory and Applications*, eds. F. P. Kelly, S. Zachary, and I. Ziedins, Clarendon Press, Oxford, 339-366.