COMPUTATIONAL METHODS FOR DATA EVALUATION AND ASSIMILATION

DAN GABRIEL CACUCI
IONEL MICHAEL NAVON
MIHAELA IONESCU-BUJOR

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

COMPUTATIONAL METHODS
FOR
DATA EVALUATION AND ASSIMILATION

COMPUTATIONAL METHODS FOR DATA EVALUATION AND ASSIMILATION

DAN GABRIEL CACUCI
IONEL MICHAEL NAVON
MIHAELA IONESCU-BUJOR

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20130701

International Standard Book Number-13: 978-1-58488-736-2 (eBook - PDF)

Contributors

Dan Gabriel Cacuci
University of South Carolina
Columbia, South Carolina

Mihaela Ionescu-Bujor
Karlsruher Institute of Technology
Karlsruhe, Campus North

Ionel Michael Navon
Florida State University
Tallahassee, Florida

Preface

This book is addressed to graduate and postgraduate students and researchers in the interdisciplinary methods of data assimilation, which refers to the integration of experimental and computational information. Since experiments and corresponding computations are encountered in many fields of scientific and engineering endeavors, the concepts presented in this book are illustrated using paradigm examples that range from the geophysical sciences to nuclear physics. In an attempt to keep the book as self-contained as possible, the mathematical concepts mostly from probability theory and functional analysis needed to follow the material presented in the book's five chapters, are summarized in the book's three appendices.

This book was finalized at the University of South Carolina. The authors wish to acknowledge the outstanding professional assistance of Dr. Madalina Corina Badea of the University of South Carolina, who has thoroughly reviewed the final version of the book, providing very valuable suggestions while improving its readability. Also acknowledged are the services of Dr. Erkan Arslan for his typing the word-version of this book into Latex. Last but not least, this book would have not have appeared without the continued patience, guidance, and understanding of Bob Stern (Executive Editor, Taylor and Francis Group), whom the authors appreciate immensely.

List of Figures

List of Tables

Contents

Introduction

Experience shows that it is practically impossible to measure exactly the true value of a physical quantity. This is because various imperfections occur at various stages involved in a measurement, including uncontrollable experimental errors, inaccurate standards, and other uncertainties arising in the data measurement and interpretation (reduction) process. Around any reported experimental value, therefore, there always exists a certain range of similar, more or less plausible, values that may also be true. In turn, this means that all inferences, predictions, engineering computations, and other applications of measured data are necessarily founded on weighted averages over all the possibly true values, with weights indicating the degree of plausibility of each value. These weights and weighted averages are what we call *probabilities* and *expectation values*. Consequently, the evaluation of scientific data is intrinsically intertwined with probability theory. The basic concepts underlying the evaluation of experimental data are presented in Chapter 1, which commences with a discussion, in Section 1.1, of the basic types of errors and probability distributions usually associated with them.

Section 1.2 presents the basic concepts of probability theory involved in the evaluation of uncertainty-afflicted scientific data. Since probabilities cannot be measured directly, they are either inferred from the results of observations or they are postulated and (partially) verified through accumulated experience. In scientific data evaluation, probabilities encode incomplete information. Persons possessing different information or knowledge assign different probabilities; furthermore, these probabilities are updated whenever new relevant information becomes available. It follows that probabilities cannot be considered as measurable physical properties. They are subjective in the sense that they depend on a person's knowledge. However, this does not mean that probabilities are arbitrary. They must obey the rules of logic, which de-

mand, for instance, that rational persons with the same knowledge assign the same probabilities. The elementary conditions for logical consistency imply two fundamental rules from which all other mathematical relationships between probabilities can be derived. These two fundamental rules are the sum and the product rules, respectively. A powerful logical tool that follows immediately from the two forms of the product rule is Bayes' theorem (1763). The customary way to express Bayes' theorem in words is: "the posterior is proportional to the likelihood times the prior" where the "prior" distribution summarizes our knowledge extant prior to observing the (new) data, the "likelihood" distribution conveys the impact of the new information brought by the (new) data, while the "posterior" distribution contains the full information available for further inferences, predictions, and decision making. Thus, Bayes' theorem is a formal model for updating, or learning from observations.

A "measurable" or "physical" quantity is a property of phenomena, bodies, or substances that can be determined qualitatively and can be expressed quantitatively. Measurement is the process of experimentally finding the value of a physical quantity, with the help of devices called *measuring instruments*. It is important to note that: (1) the purpose of a measurement is to represent a property of an object by a number, so the result of a measurement must always be a number expressed in sanctioned units of measurements; (2) a measurement is always performed with the help of a measuring instrument; and (3) a measurement is always an experimental procedure. If it were known, the true value of a measurable quantity would ideally reflect, both qualitatively and quantitatively, the corresponding property of the object. The theory of measurement relies on the following postulates: (a) the true value of the measurable quantity exists; (b) the true value of the measurable quantity is constant relative to the conditions of the measurement; and (c) the true value cannot be found.

Since measuring instruments are imperfect, and since every measurement is an experimental procedure, the results of measurements cannot be accurate. This unavoidable imperfection of measurements is quantitatively characterized by measurement uncertainty or measurement error, which can be expressed in absolute or relative form. Consequently, repeated measurements of the same physical quantity can never yield identical results; even the most carefully measured scientific data in data banks will inevitably differ from

the true values of the measured quantities. Consequently, nominal values for data, by themselves, are insufficient for applications. Quantitative uncertainties are also needed, along with the respective nominal values. Since the use of uncertain data may necessitate costly safety margins (in medicine, weather and climate prediction, or in the chemical, automotive, aerospace, or nuclear industries), working groups of the International Standards Organization have been developing uniform rules for reporting data uncertainties.

Combination of data from different sources involves a weighted propagation (via sensitivities, as will be seen subsequently) of all input uncertainties to uncertainties in the output values. Hence, data evaluation is intrinsically intertwined with uncertainty analysis, requiring reasoning from incomplete information, using probability theory for extracting "best estimate" values together with "best estimate" uncertainties from often sparse, incomplete, error-afflicted, and occasionally discrepant experimental data. A wide range of probability theory concepts and tools are employed in data evaluation and combination, from deductive statistics involving mainly frequencies and sample tallies to inductive inference for assimilating non-frequency data and a *priori knowledge*. Although grossly erroneous procedures and unintended mistakes (e.g., overlooking or miscalculating important corrections, equipment failure or improper calibration, bugs in computer codes, etc.) can produce defective data, such defective data will not be treated as "uncertainties" in this book. Nevertheless, data points that exhibit atypical behavior, which cannot be explained, need to be carefully scrutinized since outright rejection may not necessarily be appropriate. The terms "error" and "uncertainty" are interpreted in this book as being equivalent to the standard deviation of the probability distribution associated with the measurement process. This interpretation is consistent with the usual convention of considering "error" or "uncertainty" as an inherently positive number that quantifies the measurement dispersion of a specific observable parameter.

Legitimate errors are categorized either as random errors or as systematic errors. If the results of separate measurements of the same quantity differ from one another, and the respective differences cannot be predicted individually, then the error owing to this scatter of the results is called *random error*. Random errors can be identified by repeatedly measuring the same quantity under the same conditions. The scatter in results cannot be always tested in

practice, particularly in large-scale modern experiments, where it may be impractical to provide sufficient repetition in order to satisfy the explicit needs for quantifying the random errors based on strict statistical requirements. Nevertheless, reasonable estimates of random errors can often be made, particularly when the nature of the underlying probability distribution can be inferred from previous experience. Furthermore, due to the influence of the central limit theorem, many sources of random error tend to be normally distributed. A significant feature of random errors is that repeated measurements (under fixed conditions) not only permit these errors to be better determined, but they also lead to error reduction, as assured by the law of large numbers. This feature is particularly important when high precision (i.e., small random errors) is required. A subtle issue regarding random errors stems, from the fact that such errors may contain correlated components: whether an error component is correlated or not within a particular data set depends upon the role that the associated random variable plays in the respective physical problem.

In contradistinction to a random error, a systematic error is defined as a measurement error that remains constant or changes in a regular fashion when the measurements of that quantity are repeated. Such errors arise because of inherent laws in the investigative process itself, and they lead to bias. Although systematic errors are difficult to distinguish from blunders, particularly when the impact of a blunder is small, the most reliable way to uncover systematic errors is by using a more accurate measuring instrument and/or by comparing a given result with a measurement of the same quantity, but performed by a different method. Each distinct approach leads to results that differ somewhat from those obtained in other ways. These differences exhibit a pattern (i.e., are "systematic") no matter how many times each approach is repeated, because the inherent systematic deficiencies of each method cannot be avoided by mere repetition. When the errors are truly systematic, statistical regularity will emerge from the ensemble of all measurements. Such a statistical regularity will not emerge when the data sets are afflicted with blunders, since blunders are generally one-time occurrences that can be detected if a particular procedure is repeated. Consequently, redundancy within a given investigative procedure is desirable not only to improve precision, but also to purge the results of blunders.

Probability theory is a branch of mathematical sciences that provides a model for describing the process of observation. The need for probability theory arises from the fact that most observations of natural phenomena do not lead to uniquely predictable results. Probability theory provides the tools for dealing with actual variations in the outcome of realistic observations and measurements. The challenging pursuit to develop a probability theory which is mathematically rigorous and also describes many phenomena observable in nature has generated over the years notable disputes over conceptual and logical issues. Modern probability theory is based on postulates constructed from three axioms attributed to Kolmogorov (1933), all of which are consistent with the notion of frequency of occurrence of events. The alternative approach, traceable to Laplace (1812), is based on the concept that probability is simply a way of providing a numerical scale to quantify our reasonable beliefs about a situation which we know only incompletely; this approach is consistent with Bayes' theorem, conditional probabilities, and inductive reasoning. Either approach to probability theory would completely describe a natural phenomenon if sufficient information were available to determine the underlying probability distribution exactly. In practice, though, such exact knowledge is seldom, if ever, available so the features of the probability distribution underlying the physical phenomenon under consideration must be estimated. Such estimations form the study object of statistics, which is defined as the branch of mathematical sciences that uses the results of observations and measurements to estimate, in a mathematically well-defined manner, the essential features of probability distributions. Both statistics and probability theory use certain generic terms for defining the objects or phenomena under study. A *system* is the object or phenomena under study. It represents the largest unit being considered. A system can refer to a nuclear reactor, corporation, chemical process, mechanical device, biological mechanism, society, economy, or any other conceivable object that is under study. The output or *response* of a system is a result that can be measured quantitatively or enumerated. The power of a nuclear reactor, the yield of a process, life span of cell, the atmospheric temperature and pressure, are all examples of system responses. A *model* is a mathematical idealization that is used as an approximation to represent the system and its output. Models can be quite simple or highly complex; regardless of its complexity, though, the model is an idealization of the sys-

tem, so it cannot be exact. Usually, the more complex the system, the less exact the model, particularly since the ability to solve exactly mathematically highly complex expressions diminishes with increasing complexity. In other words, the simpler the model, the easier it is to analyze but the less precise the results.

Probabilities cannot be measured directly; they can be inferred from the results of observations or they can be postulated and (partially) verified through accumulated experience. In practice, though, certain random vectors tend to be more probable, so that most probability functions of practical interest tend to be localized. Therefore, the essential features regarding probability distributions of practical interest are measures of location and of dispersion of the observed results. Practice indicates that location is best described by the mean value, while dispersion of observed results appears to be best described by the variance, which is a second-order moment. In particular, the mean value can be interpreted as a locator of the center of gravity, while the variance is analogous to the moment of inertia (which linearly relates applied torque to induced angular acceleration in mechanics). For multivariate probability distributions, the collection of all second-order moments forms the so-called variance-covariance matrix, or, simply, the covariance matrix. If the probability function is known, then these moments can be calculated directly, through a process called *statistical deduction*. Otherwise, if the probability function is not known, then the respective moments must be estimated from experiments, through a process called *statistical inference*. The definitions, interpretations, and quantifications of the moments of a distribution, particularly the means and covariances, are discussed in Section 1.3. Particularly important is the method of propagation of errors or propagation of moments, which can be used to compute the error in a systems response (which can be either the result of an indirect measurement or the result of a computation), by propagating the uncertainties of the component system parameters using a form Taylor-series expansion of the response as a function of the underlying model parameters.

In practice, users of measured data seldom require knowledge of the complete posterior distribution, but usually request a "recommended value" for the respective quantity, accompanied by "error bars" or some suitably equivalent summary of the posterior distribution. Decision theory can provide such

a summary, since it describes the penalty for bad estimates by a loss function. Since the true value is never known in practice, it is not possible to avoid a loss completely, but it is possible to minimize the expected loss, which is what an optimal estimate must accomplish. As will be shown in the first section of Chapter 2, in the practically most important case of "quadratic loss" involving a multivariate posterior distribution, the "recommended value" turns out to be the vector of mean values, while the "error bars" are provided by the corresponding covariance matrix. Thus, Section 2.1 also presents simple examples of estimating covariances and confidence intervals from experimental data.

Practical applications require not only mathematical relations between probabilities, but also rules for assigning numerical values to probabilities. As is well known, Bayesian statistics provides no fundamental rule for assigning the prior probability to a theory. The choice of the "most appropriate" prior distribution lies at the heart of applying Bayes' theorem to practical problems, and has caused considerable debates in the past, lasting over a century. Section 2.2 discusses the assignment of prior probability distributions under incomplete information. Of course, when complete prior information related to the problem under consideration is available and can be expressed in the form of a probability distribution, this information should certainly be used. In such cases, the repeated application of Bayes' theorem will serve to refine the knowledge about the respective problem. At the other extreme, when no specific information is available, it may be possible to construct prior distributions using concepts of group theory to reflect the possible invariance and/or symmetry properties of the problem under consideration, as discussed in Section 2.2.1. On the other hand, if repeatable trials are not feasible, but some information could nevertheless be inferred by some other means, information theory can be used in conjunction with the maximum entropy principle (the modern generalization of Bernoulli's principle of insufficient reason) to assign numerical values to probabilities, thus constructing a prior distribution, as will be shown in Section 2.2.2. The material presented in Section 2.2 is certainly not exhaustive regarding the use of group theory and symmetries for assigning priors (which continues to remain an area of active research), but is limited to presenting only the most commonly encountered priors in practice, and which are also encountered throughout this book.

Section 2.3 presents methods for evaluating unknown parameters from data which is consistent "within error bars," and is afflicted solely by random errors. Three common situations are considered, as follows: (i) evaluation of a location parameter when the scale parameters are known; (ii) both the scale and location parameters are unknown but need to be evaluated; and (iii) evaluation of a counting rate (a scale parameter) in the presence of background (noise). It is probably not too unfair to say that, although measurements without systematic errors are the exception rather than the rule, conventional (frequentist) sampling theory has not much to offer to practitioners in science and technology who are confronted with systematic errors and correlations. This is in marked contrast to the wealth of material on statistical errors, for which satisfactory techniques are available, based on counting (Poisson) statistics or on Gaussian models for the scatter of repeatedly measured data. Using Bayesian parameter estimation under quadratic loss, group-theoretical least informative priors, and probability assignment by entropy maximization, Section 2.4 addresses the practical situation of evaluating means and covariances from measurements affected by both random (uncorrelated) and systematic (correlated) errors issues. It is explained how common errors, the most frequent type of systematic error, invariably induce correlations, and how correlations are described by nondiagonal covariance matrices. As will be mathematically shown in this section, the random errors can be reduced by repeated measurements of the same quantity, but the systematic errors cannot be reduced this way. They remain as a "residual" uncertainty that could be reduced only by additional measurements, using different techniques and instrumentation, geometry, and so on. Eventually, additional measurements using different techniques would reduce the correlated (systematic) error just as repetitions of the same measurement using the same technique reduce the uncorrelated statistical uncertainty.

As discussed in Chapter 1, unrecognized or ill-corrected experimental effects, including background, dead time of the counting electronics, instrumental resolution, sample impurities, and calibration errors usually yield inconsistent experimental data. Although legitimately discrepant data may occur with a nonzero probability (e.g., for a Gaussian distribution, the probability that two equally precise measurements are outside of two standard deviations is about 15.7%), it is much more likely that apparently discrepant experiments

actually indicate the presence of unrecognized errors. Section 2.5 illustrates the basic principles for evaluating discrepant data fraught by unrecognized, including systematic (common), errors. The marginal distributions for both recognized and unrecognized errors are obtained for both the Jeffreys least informative prior and for an exponential prior (when the unrecognized errors can be characterized by a known scale factor). This treatment of unrecognized systematic errors is an example of a two-stage "hierarchical" Bayesian method, involving a twofold application of Bayes' theorem to the sampling distribution that depends on parameters having a Gaussian prior, which in turn depended on a so-called "hyper-parameter," which had itself a "hyper-prior" distribution. Also the first-order expressions obtained for the various quantities are similar to the James-Stein estimators, which have sometimes lower risk than the estimates resulting from Bayesian estimation under quadratic loss (that minimize the square error averaged over all possible parameters, for the sample at hand). It is important to note, though, that the two-stage method Bayesian used in Section 2.5 yields results that are superior to the James-Stein estimators, especially for small samples. Moreover, the results presented in this section yield further improvements in a systematic and unambiguous way, without the discontinuities, questions of interpretation and restrictions associated with James-Stein estimators. This fact is particularly valuable and relevant for scientific data evaluation, where best values must often be inferred (for quadratic or any other loss) from just a single available sample.

Chapter 3 presents minimization algorithms, which are best suited for unconstrained and constrained minimization of large-scale systems such as time-dependent variational data assimilation in weather prediction and similar applications in the geophysical sciences. The operational implementation of "four-dimensional variational" data assimilation (customarily called 4–D VAR) hinges crucially upon the fast convergence of efficient gradient-based large-scale unconstrained minimization algorithms which are called to minimize a cost function that attempts to quantify the discrepancies between forecast and observations in a window of assimilation, subject to constraints imposed by the geophysical model. Data assimilation problems in oceanography and meteorology contain many typically of the order of ten million degrees of freedom. Consequently, conjugate-gradient (CG) methods and limited-

memory quasi-Newton (LMQN) methods come into consideration since they typically require storage of only a few vectors, containing information from a few iterations, converge to local minima even from remote starting points, and can be efficiently implemented on multiprocessor machines.

Section 3.2 highlights the common as well as distinctive salient features of the minimization algorithms called (acronyms) CONMIN, E04DGF, and of the Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method, and Bert-Buckley-Variable-Storage-Conjugate-Gradient (BBVSCG) method. These methods all fall under the category of Limited-Memory Quasi-Newton (LMNQ) methods, combining the advantages of CG-algorithms (low storage requirements) with the advantages of quasi-Newton of (QN) methods (computational efficiency stemming from their superlinear convergence). The LMQN algorithms build several rank one or rank two matrix updates to the Hessian matrix, thereby avoiding the need to store the approximate Hessian matrix, as required by full QN methods. Like the CG-methods, the LMQN methods require only modest amount storage for generating the search directions. Currently, the L-BFGS algorithm is the widest used minimization algorithm at the operational numerical weather prediction centers which rely on large-scale 4–D VAR assimilation and prediction methodologies.

Section 3.3 presents Truncated Newton (T-N) methods, which attempt to retain the rapid (quadratic) convergence rate of classic Newton methods, while requiring evaluations of functions and gradients only, thereby reducing the storage and computational requirements to sizes that are feasible for large-scale minimization applications. The T-N methods are also called *Hessian-free methods*. When used together with a finite difference approximation to the Hessian-vector products, the T-N methods achieve a quasi-quadratic convergence rate. In recent implementation, these Hessian-vector products are computed most efficiently using adjoint methods. Thus, the T-N methods require forward and backward adjoint inner iterations within a CG-formalism. Therefore, although the T-N methods offer competitive alternatives for two-dimensional problems, they are not yet competitive for 3-D operational problems. This is because the high cost of the CG inner iterations offsets the advantage of its almost quadratic convergence rate.

Section 3.4 discusses the use of information provided by the Hessian matrix for large-scale optimization, highlighting, in particular, the way in which the

eigenvalues of the Hessian matrix determine the convergence rate for unconstrained minimization. Section 3.5 discusses issues related to nonsmooth and nondifferentiable optimization, in view of the fact that precipitation and radiation parameterizations involve on/off processes. Methods of nondifferentiable optimization are needed to minimize nonsmooth functionals. Nonsmooth optimization methods are based on the assumptions that: (i) the functional to be minimized is locally Lipschitz continuous, and (ii) the functional and its arbitrary subgradients can be evaluated at each point. Nonsmooth optimization methods can be divided into two main classes: subgradient methods and bundle methods. The guiding principle underlying bundle methods is to gather the subgradient information from previous iterations into a bundle of subgradients. Although the additional computational cost for building the sub-gradient is several times larger than that for L-BFGS, bundle nonsmooth optimization methods may work advantageously for problems with discontinuities, where L-BFGS methods usually fail. Bundle nonsmooth optimization methods have not been tested yet on operational 4–D VAR systems, but investigations are in progress to assess the applicability of such methods to realistic large-scale problems.

Section 3.6 addresses two fundamental issues related to step-size searches in conjugate-gradient type methods: (i) How good is the search direction?; and (ii) What is the best choice for the length of the step along the search direction? Section 3.7 highlights the salient features of trust-region methods, which also seek global convergence while retaining fast local convergence of optimization algorithms. It is noted that the trust region methods follow a reverse sequence of operations, by first choosing a trial step length, and subsequently using a quadratic model to select the best step length.

Section 3.8 discusses scaling and preconditioning for linear and nonlinear problems. The goal of preconditioning is to improve the performance of conjugate gradient–type minimization methods, by reducing the number of iterations required to achieve a prescribed accuracy. Scaling can substantially improve the performance of minimization algorithms. An effective automatic scaling could also improve the condition number of the Hessian matrix for well-scaled problems, thus facilitating their solution. Scaling by variable transformations converts the variables from units that reflect the physical nature of the problem to units that display desirable properties for improving the

efficiency of the minimization algorithms. On the other hand, badly scaled nonlinear problems can become extremely difficult to solve.

Several popular methods for performing nonlinear constrained optimization are discussed in Section 3.9, namely: (i) the penalty method; (ii) barrier methods; (iii) augmented Lagrangian methods; and (iv) sequential quadratic programming (SQP) methods. The penalty method replaces a constrained optimization problem by a series of unconstrained problems whose solutions should converge to the solution of the original constrained problem. The unconstrained problems minimize an objective function which is constructed by adding to the original objective function a term that comprises a penalty parameter multiplying a measure of the violation of the constraints. The measure of violation is nonzero when the constraints are not satisfied, and is zero in the region where the constraints are satisfied. The original problem can thus be solved by formulating a sequence of unconstrained subproblems. Barrier methods are an alternative class of algorithms for constrained optimization. These methods also use a penalty-like term added to the objective function, but the results of iterations within the barrier methods are forced by the barrier to remain interior to and away from the boundary of the feasible solution domain. Augmented Lagrangian methods turn a constrained minimization problem into the unconstrained minimization The SQP solves a sequence of sub-problems designed to minimize a quadratic model of the objective functional subject to linearization of the constraints. The SQP method can be used within either a linesearch or a trust region framework, and is very efficient for solving both small and large problems. To be practical, a SQP method must be able to converge on nonconvex problems, starting from remote points.

Nonlinear optimization may involve cost functions characterized by the presence of multiple minima. The aim of global optimization is to determine all of the critical points of a function, particularly if several local optima exist where the corresponding function values differ substantially from one another. Global optimization methods can be classified into two major categories, namely deterministic and stochastic methods. Deterministic methods attempt to compute all of the critical points with probability one (i.e., with absolute success). On the other hand, stochastic methods sacrifice the possibility of an absolute guarantee of success, attempting to minimize the function under consideration in a random sample of points from a set, which is assumed

to be convex, compact, and to contain the global minimum as an interior point. Section 3.10 briefly discusses two stochastic global minimization methods: simulated annealing and genetic algorithms, which have recently been implemented in variational data assimilation in geophysical sciences applications. The simulated annealing algorithm exploits the analogy between the search for a global minimum and the annealing process (i.e., the way in which a metal cools and freezes) into a minimum energy crystalline structure. The genetic algorithms attempt to simulate the phenomenon of natural evolution, as each species searches for beneficial adaptations in an ever-changing environment. As species evolve, new attributes are encoded in the chromosomes of individual members. This information changes by random mutation, but the actual driving force behind evolutionary development is the combination and exchange of chromosomal material during breeding.

Simulated annealing algorithms are intrinsically sequential. On the other hand, genetic algorithms are particularly well suited for implementation on parallel computers. Evaluation of the objective function and constraints can be done simultaneously for the entire population; the production of the new population by mutation and crossover can also be parallelized. On highly parallel machines, therefore, a genetical algorithm (GA) can be expected to run nearly N times faster than on non-parallel machines, where N is the population size. Currently, however, the convergence rate of these global minimization methods does not outperform the convergence rate of LMQN methods for large-scale operational 4–D VAR models.

Chapter 4 discusses several basic principles of four-dimensional variational assimilation (4–D VAR). Initially, data assimilation methods were referred to as "objective analyses," in contradistinction to "subjective analyses," in which numerical weather predictions (NWP) forecasts were adjusted "by hand" by meteorologists, using their professional expertise. Subsequently, methods called "nudging" were introduced based on the simple idea of Newtonian relaxation. In nudging, the rightside of the model's dynamical equations is augmented with a term which is proportional to the difference between the calculated meteorological variable and the observation value. This term keeps the calculated state vector closer to the observations. Nudging can be interpreted as a simplified Kalman-Bucy filter with the gain matrix being prescribed rather than obtained from covariances. The nudging method is

used in simple operational global-scale and meso-scale models for assimilating small-scale observations when lacking statistical data. The recent advances in nudging methods are briefly presented in Section 4.1.

Section 4.2 briefly mentions the "optimal interpolation" (OI) method, "three-dimensional variational data assimilation" (3–D VAR), and the physical space statistical analysis (PSAS) methods. These methods were introduced independently, but were shown to be formally equivalent; in particular, PSAS is a dual formulation of 3–D VAR.

Data assimilation requires the explicit specification of the error statistics for model forecast and the current observations, which are the primary quantities needed for producing an analysis. A correct specification of observation and background error covariances are essential for ensuring the quality of the analysis, because these covariances determine to what extent background fields will be corrected to match the observations. The essential parameters are the variances, but the correlations are also very important because they specify the manner in which the observed information will be smoothed in the model space if the resolution of the model does not match the density of the observations. Section 4.3 briefly outlines the prevailing operational practices employed for the practical estimation of observation error covariance matrices and background error covariance matrices.

The goal of the 4–D VAR formalism is to find the solution of a numerical forecast or numerical weather prediction (NWP) model that best fits sequences of observational fields distributed in space over a finite time interval. Section 4.4 discusses the basic framework of "four-dimensional variational" data assimilation (4–D VAR) methods utilizing optimal control theory (variational approach). The advance brought by the variational approaches is that the meteorological fields satisfy the dynamical equations of the forecast model while simultaneously minimizing a cost functional, which measures the differences between the computed and the observed fields, by solving a constrained minimization problem. The 4–D VAR formalism is first presented without taking the modeling errors into account; subsequently, the functional to be minimized is extended to include model errors. This section concludes with a discussion of the consistent optimality and transferable optimality properties of the 4–D VAR procedure.

Section 4.5 presents results of numerical experiments with unconstrained

minimization methods for 4–D VAR using the shallow water equations, which are widely used in meteorology and oceanography for testing new algorithms since they contain most of the physical degrees of freedom (including gravity waves) present in the more sophisticated operational models. The numerical experiments were performed with four limited-memory quasi-Newton (LMQN) methods (CONMIN-CG, E04DGF, L-BFGS, and BBVSCG) and two truncated Newton (T-N) methods. The CONMIN-CG and BBVSCG algorithms failed after the first iteration, even when both gradient scaling and non-dimensional scaling were applied. The L-BFGS algorithm was successful only with gradient scaling. On the other hand, the E04DGF algorithm worked only with the non-dimensional shallow water equations model. This indicates that using additional scaling is essential for the success of LMQN minimization algorithms when applied to large-scale minimization problems. On the other hand, T-N methods appear to perform best for large-scale minimization problems, especially in conjunction with a suitable preconditioner. The importance of preconditioning increases with increasing dimensionality of the minimization problem under consideration. Furthermore, for the Shallow-Water Equations (SWE) numerical experiments, the T-N methods required far fewer iterations and function calls than the LMQN methods.

In the so-called strong constraint Variational Data Assimilation (VDA), or classical VDA, it is assumed that the forecast model perfectly represents the evolution of the actual atmosphere. The best fit model trajectory is obtained by adjusting only the initial conditions via the minimization of a cost functional that is subject to the model equations as strong constrains. However, numerical weather prediction (NWP) models are imperfect since subgrid processes are not included. Furthermore numerical discretizations produce additional dissipative and dispersion errors. Modeling errors also arise from the incomplete mathematical modeling of the boundary conditions and forcing terms, and from the simplified representation of physical processes and their interactions in the atmosphere. Usually, all of these modeling imperfections are collectively called *model error* (ME). Model error is formally introduced as a correction to the time derivatives of model variables. Section 4.6 highlights the treatment of MEs in VDA, taking into account numerical errors explicitly as additional terms in the cost functional to be minimized. However, taking MEs into account doubles (and can even triple) the size of the system to be

optimized by comparison to minimizing the cost functional when the model errors are neglected.

Chapter 5 highlights specific difficulties in applying 4–D VAR to large-scale operational numerical weather prediction models. Recall that the objective of 4–D VAR is to find the optimal set of control variables, usually the initial conditions and/or boundary conditions, such that a cost function, comprising a weighted least square norm that quantifies the misfit between model forecast and observations, is minimized subject to the constraints of satisfying the geophysical model. In order to minimize this cost function, we need to know the gradient of this function with respect to the control variables. A straightforward way of computing this gradient is to perturb each control variable in turn and estimate the change in the cost function. But this method is impractical when the number of control variables is large a typical meteorological model comprises $O(10^7)$ control variables. Furthermore, the iterative minimization of the cost function requires several gradient estimations on the way to finding a local minimum. Often, these gradient estimations are insufficiently accurate to guarantee convergence of the minimization process. As discussed in Section 5.1, the convergence of the minimization process in 4–D VAR is particularly affected (negatively) by strong nonlinearities and on/off physical processes such as precipitation and clouds.

Section 5.2 highlights the highly efficient adjoint method for computing exactly the gradient of the cost function with respect to the control variables by integrating once the adjoint model backwards in time. Such a backward integration is of similar complexity to a single integration of the forward model. Another key advantage of adjoint variational data assimilation is the possibility to minimize the cost function using standard unconstrained minimization algorithms (usually iterative descent methods). However, for precipitation observations, highly nonlinear parameterization schemes must be linearized for developing the adjoint version of the model required by the minimization procedure for the cost function. Alternatively, simpler physics schemes, which are not a direct linearization of the full model physics (i.e., the "linear model" is not tangent linear to the nonlinear full model), can be coded for the linear model. Of course, it is desirable to have a linearized model that approximates as closely as possible the sensitivity of the full nonlinear model; otherwise the forecast model may not be in balance with its own analysis, producing

so-called "model spin-up." Furthermore, multi-incremental approaches can exhibit discrete transitions affecting the stability of the overall minimization process.

On the one hand, nonlinear models have steadily evolved in complexity in order to improve forecast skill. For example, the prognostic cloud scheme introduced into the European Centre for Medium Range Weather Forecasts (ECMWF) model includes many highly nonlinear processes that are often controlled by threshold switches. On the other hand, even if it were possible to construct the tangent linear and, respectively, adjoint models corresponding to this complex cloud scheme, the validity of these models would be restricted due to these thresholds and their value would be questionable. Issues related to using non-smooth optimization methods to address discontinuities is an ongoing research topic. Standard solvers of elliptic equation often perform "fast Fourier transform" (FFT) and inverse FFT operations. Section 5.3 summarizes the adjoint coding of the FFT and of the inverse FFT, showing that the adjoint of the FFT is obtained by calling the inverse FFT routine and multiplying the output by the number of data sample points. Conversely, the adjoint of the inverse FFT is obtained by calling the FFT routine and dividing the output by the number of data sample points.

Section 5.4 indicates that the correctness of the adjoint code for interpolations and on/off processes can be verified by performing an additional integration of the nonlinear model with added bit vectors in order to determine the routes for the IF statements included in the physical processes. This additional integration of the nonlinear model with added bit vectors is needed for the verification of both the tangent linear model and the adjoint model. Section 5.5 discusses the construction of background covariance matrices, which play the following important roles: (i) spreading the information from the observations to neighboring domains; (ii) providing statistically consistent increments at the neighboring grid points and levels of the model; and (iii) ensuring that observations of one model variable (e.g., temperature) produce dynamically consistent increments in the other model variables (e.g., vorticity and divergence).

Section 5.6 revisits the characterization of model errors specifically for the 4–D VAR data assimilation procedure. Such errors are attributable to the dynamical model (e.g., poor representation of processes, omissions, or incorrect

formulations of key processes, numerical approximations) and observations or measurements (e.g., sensor design, performance, noise, sample averaging, aliasing). For dynamically evolving systems, the model errors are expected to depend on time and, possibly, on the model state variables. Controlling modeling errors in addition to the model's initial conditions in the weak constraint 4–D VAR doubles the size of the optimization problem by comparison to the strong constraint 4–D VAR. Furthermore, if the stochastic component is included in the model error formulation, then the random realization would need to be saved at each model time step. Consequently, the size of the optimization problem would be tripled. The size of the model error control vector can be reduced by projecting it onto the subspace of eigenvectors corresponding to the leading eigenvalues of the adjoint-tangent linear operators.

Most data assimilation systems are not equipped to handle large, systematic corrections; they were designed to make small adjustments to the background fields that are consistent with the presumed multivariate and spatial structures of random errors. Statistics of "observed-minus-background" residuals provide a different, sometimes more informative, view on systematic errors afflicting the model or observations. Operational NWP centers routinely monitor time- and space-averaged background residuals associated with different components of the observing system, providing information on the quality of the input data as well as on the performance of the assimilation system. In general, small root-mean-square residuals imply that the system is able to accurately predict future observations. Nonzero mean residuals, however, indicate the presence of biases in the observations and/or their model-predicted equivalents. It is paramount to develop physically meaningful representations of model errors that can be clearly distinguished from possible observation errors. This issue is the subject of intensive ongoing research.

Section 5.7 discusses the incremental 4–D VAR algorithm, which was formulated in the mid-1990s, and decisively facilitated the adoption, application, and implementation of 4–D VAR data assimilation at major operational centers, thereby advancing the timely state of weather prediction. Prior to the development of the "incremental 4–D VAR algorithm," implementation of the full 4–D VAR algorithm in operational models was impractical, since a typical minimization requires between 10 and 100 evaluations of the gradient. The cost of the adjoint model is typically 3 times that of the forward model,

and the analysis window in a typical operational model such as the ECMWF system is 12-hours. Thus, the cost of a 12 hour analysis was roughly equivalent to between 20 and 200 days of model integration (with 108 variables), making it computationally prohibitive for NWP centers which had to deliver timely forecasts to the public. In addition, the nonlinearity of the model and/or of the observation operator could produce multiple minima in the cost function, which impacted the convergence of the minimization algorithm.

The incremental 4–D VAR algorithm reduces the resolution of the model and eliminates most of the time-consuming physical packages, thereby enabling the 4–D VAR method to become computationally feasible. Furthermore, the incremental 4–D VAR algorithm removes the nonlinearities in the cost minimization by using a forward integration of the linear model instead of a nonlinear one. The minimization procedure is identical to the usual 4–D VAR algorithm except that the increment trajectory is obtained by integration of the linear model. The reference trajectory (which is needed for integrating the linear and adjoint models and which starts from the background integration) is not updated at every iteration. This simplified iterative procedure for minimizing the incremental cost function is called the *inner loop*, and is much cheaper computationally to implement by comparison to the full 4–D VAR algorithm. However, when the quadratic cost function is approximated in this way, the incremental 4–D VAR algorithm no longer converges to the solution of the original problem. Furthermore, the analysis increments are calculated at reduced resolution and must be interpolated to conform to the high-resolution model's grid. Consequently, after performing a user-defined number of inner loops, one outer loop is performed to update the high-resolution reference trajectory and the observation departures. After each outer loop update, it is possible to use progressively higher resolutions for the inner loop.

However, experiments show that the current implementations of the incremental 4–D VAR algorithm lead to divergent computational results after four outer loop iterations (e.g., this was the case when the incremental 4–D VAR was initially implemented into the ECMWF weather prediction system). Various numerical experiments indicate that convergence can be attained when the inner and outer loops use the same resolution and/or use the same time step. This feature is explained by the presence of gravity waves which propagate at different speeds in the linear and nonlinear models. These gravity waves are

related to the shape of the leading eigenvector of the Hessian of the 4–D VAR cost function; this eigenvector is determined by the surface pressure observation and controls the convergence of the minimization algorithm. Chapter 5 concludes with a short discussion, in Section 5.8, of current research issues.

This book also comprises three appendices. Appendix A (Chapter 6) is intended to provide a quick reference to selected properties of distributions commonly used for data analysis, evaluation, and assimilation. Appendix B (Chapter 7) introduces and summarizes the most important properties of adjoint operators in conjunction with differential calculus in vector spaces, as used for data assimilation. Appendix C (Chapter 8) highlights the main issues arising when identifying and estimating model parameters from experimental data. The problem of parameter identification can be formulated mathematically as follows: "an unknown parameter is *identifiable* if it can be determined uniquely at all points of its domain by using the input-output relation of the system and the input-output data." Such a mapping is generally known as an "inverse problem," in contradistinction with the forward mapping, which maps the space of parameters to the space of outputs. The uniqueness of inverse mappings is difficult to establish. Appendix C also discusses briefly three methods for estimating parameters: the maximum likelihood method, the maximum total variation L_1-regularization method for estimating parameters with discontinuities, and the "extended Kalman filter" method.

1

Experimental Data Evaluation: Basic Concepts

CONTENTS

Experience shows that it is practically impossible to measure exactly the true value of a physical quantity. This is because various imperfections occur at various stages involved in a measurement, including uncontrollable experimental errors, inaccurate standards, and other uncertainties arising in the data measurement and interpretation (reduction) process. Around any reported experimental value, therefore, there always exists a certain range of similar, more or less plausible, values that may also be true. In turn, this means that all inferences, predictions, engineering computations, and other applications of measured data are necessarily founded on weighted averages over all the possibly true values, with weights indicating the degree of plausibility of each value. These weights and weighted averages are what we call *probabilities* and *expectation values*. Consequently, the evaluation of scientific data is intrinsically intertwined with probability theory. The basic types of errors and probability distributions usually associated with them will be presented in Section 1.1.

The interpretation of probabilities as degrees of plausibility or rational expectation, on a numerical scale ranging from 0 (impossibility) to 1 (certainty), dates back at least to Bernoulli [16] and Laplace [111]. In scientific

data evaluation, probabilities encode incomplete information. Persons possessing different information or knowledge assign different probabilities; furthermore, these probabilities are updated whenever new relevant information becomes available. It follows that probabilities cannot be considered as measurable physical properties. They are subjective in the sense that they depend on a person's knowledge. However, this does not mean that probabilities are arbitrary. They must obey the rules of logic, which demand, for instance, that rational persons with the same knowledge assign the same probabilities. The elementary conditions for logical consistency imply two fundamental rules from which all other mathematical relationships between probabilities can be derived. These two fundamental rules are the sum and the product rules, respectively (see, e.g., refs. [39] and [202]). A powerful logical tool that follows immediately from the two forms of the product rule is Bayes' theorem (1763). In data analysis and evaluation, Bayes' theorem [10] shows how, under given experimental conditions C, the "direct" probabilities of experimental errors (i.e., likelihoods of possible observed data B given the true values A) are related to the "inverse" probabilities of possible true values A given the observed data B. The customary way to express Bayes' theorem in words is: "*posterior* $\propto$ *likelihood* $\times$ *prior*" where the "prior" distribution summarizes our knowledge extant prior to observing the (new) data, the "likelihood" distribution conveys the impact of the new information brought by the (new) data, while the "posterior" distribution contains the full information available for further inference, prediction, and decision making. Thus, Bayes' theorem is a formal model for updating, or learning from observations. Section 1.2 presents the basic concepts of probability theory involved in the evaluation of uncertainty-afflicted scientific data.

Since probabilities cannot be measured directly, they are either inferred from the results of observations or they are postulated and (partially) verified through accumulated experience. In practice, though, certain random vectors tend to be more probable, so that most probability functions of practical interest tend to be localized. Therefore, the essential features regarding probability distributions of practical interest are measures of *location* and of *dispersion*. These measures are provided by the *expectation* and *moments* of the respective probability function. If the probability function is known, then these moments can be calculated directly, through a process called *statisti-*

cal deduction. Otherwise, if the probability function is not known, then the respective moments must be estimated from experiments, through a process called *statistical inference*. Since measurements do not yield true values, it is necessary to introduce surrogate parameters to describe the "location" and "dispersion" for the observed results. Practice indicates that *location is best described by the mean value*, while *dispersion* of observed results appears to be *best described* by the *variance*, which is a second-order moment. In particular, the mean value can be interpreted as a locator of the center of gravity, while the variance is analogous to the moment of inertia (which linearly relates applied torque to induced angular acceleration in mechanics). For multivariate probability distributions involving n random variables, the collection of all second-order moments forms an $n \times n$ matrix, called the *variance-covariance matrix*, or, simply, the covariance matrix. Section 1.3 highlights the important properties of, and methods for evaluating covariances.

1.1 Experimental Data Uncertainties

A "measurable" or "physical" quantity is a property of phenomena, bodies, or substances that can be defined qualitatively and can be expressed quantitatively. *Measurement* is the process of experimentally finding the value of a physical quantity, with the help of devices called *measuring instruments*. A measurement has three features:

1. The result of a measurement must always be a number expressed in sanctioned units of measurements. The purpose of a measurement is to represent a property of an object by a number.

2. A measurement is always performed with the help of a measuring instrument; measurement is impossible without measuring instruments.

3. A measurement is always an experimental procedure.

If it were known, the true value of a measurable quantity would ideally reflect, both qualitatively and quantitatively, the corresponding property of the object. The theory of measurement relies on the following postulates:

1. The true value of the measurable quantity exists;
2. The true value of the measurable quantity is constant relative to the conditions of the measurement; and
3. The true value cannot be found.

Since measuring instruments are imperfect, and since every measurement is an experimental procedure, the results of measurements cannot be absolutely accurate. This unavoidable imperfection of measurements is quantitatively characterized by *measurement uncertainty* or *measurement error*, which can be expressed in absolute or relative form. Consequently, repeated measurements of the same physical quantity can never yield identical results. An indication of the probability of measurements to bracket the true value can be inferred by considering the "Principle of Insufficient Reason," already used by Bernoulli [16] and Laplace [111]. For two measurements, for example, this principle would assign equal probabilities to the two alternatives, namely, "measured value too low" and "measured value too high," unless there is a reason to consider one or the other as more likely. Hence, if the two measurements were independent, the four alternatives "low-low," "low-high," "high-low," and "high-high" would be equally probable, so the two measured values would bracket the true value with a probability of 1/2. Even after a third measurement, the probability would still be (as high as) 1/4 that the true value is below or above all of the measured values.

Thus, since even the most carefully measured scientific data in data banks will inevitably differ from the true values of the measured quantities, it is apparent that nominal values for data, by themselves, are insufficient for applications; the quantitative uncertainties are also needed, along with the respective nominal values. Therefore, in response to user requests (e.g., from nuclear science and technology, fission and fusion reactor development, astrophysics, radiotherapy), nuclear data evaluators, for example, are now also including "covariance files" comprising information about uncertainties and correlations when updating the extensive computer libraries of recommended

nuclear data. Since the use of uncertain data may necessitate costly safety margins (in medicine, weather and climate prediction, or in the chemical, automotive, aerospace, or nuclear industries), working groups of the International Standards Organization have been developing uniform rules for reporting data uncertainties.

Combination of data from different sources involves a weighted propagation (via sensitivities, as will be seen subsequently) of all input uncertainties to uncertainties in the output values finally recommended in data banks. Hence, data evaluation is intrinsically intertwined with uncertainty analysis, requiring reasoning from incomplete information, using probability theory for extracting "best" values together with "best" uncertainties from often sparse, incomplete, error-afflicted, and occasionally discrepant experimental data. A wide range of probability-theory concepts and tools are employed in data evaluation and combination, from deductive statistics involving mainly frequencies and sample tallies to inductive inference for assimilating non-frequency data and *a priori* knowledge.

In practice, grossly erroneous procedures and unintended mistakes (e.g., overlooking or miscalculating important corrections, equipment failure or improper calibration, bugs in computer codes, etc.) can produce defective data. In this book, such defective data will *not* be treated as "uncertainties." Nevertheless, data points that exhibit atypical behavior, which cannot be explained, need to be carefully scrutinized since outright rejection may not necessarily be appropriate: for example, a differential cross section point on an excitation curve might appear to be anomalous, but could be nevertheless correct due to the effect of a previously unidentified resonance. Such data points need to be well understood before subjecting the data to extensive statistical analysis for quantifying the associated uncertainties.

The terms "error" and "uncertainty" are interpreted in this book as being equivalent to the standard deviation of the probability distribution associated with the measurement process. This interpretation is consistent with the usual convention of considering "error" or "uncertainty" as an inherently positive number that quantifies the measurement dispersion of a specific observable parameter. Legitimate errors are categorized either as *random errors* or as *systematic errors*. If the results of separate measurements of the same quantity differ from one another, and the respective differences cannot be

predicted individually, then the error owing to this scatter of the results is called *random error*. Random errors can be identified by repeatedly measuring the same quantity under the same conditions. The scatter in results cannot be always tested in practice, particularly in large-scale modern experiments, where it may be impractical to provide sufficient repetition in order to satisfy the explicit needs for quantifying the random errors based on strict statistical requirements. Nevertheless, reasonable estimates of random errors can often be made, particularly when the nature of the underlying probability distribution can be inferred from previous experience. Furthermore, due to the influence of the central limit theorem, many sources of random error tend to be normally distributed. A significant feature of random errors is that repeated measurements (under fixed conditions) not only permit these errors to be better determined, but they also lead to error reduction, as assured by the law of large numbers. This feature is particularly important when high precision (i.e., small random errors) is required. A subtle issue regarding random errors stems from the fact that such errors may contain correlated components: whether an error component is correlated or not within a particular data set depends upon the role that the associated random variable plays in the respective physical problem.

In contradistinction to a random error, a *systematic error* is defined as a measurement error that remains constant or changes in a regular fashion when the measurements of that quantity are repeated. Such errors arise because of inherent flaws in the investigative process itself, and they lead to bias. Although systematic errors are difficult to distinguish from blunders, particularly when the impact of a blunder is small, the most reliable way to uncover systematic errors is by using a more accurate measuring instrument and/or by comparing a given result with a measurement of the same quantity, but performed by a different method. Each distinct approach leads to results that differ somewhat from those obtained in other ways. These differences exhibit a pattern (i.e., are "systematic") no matter how many times each approach is repeated, because the inherent systematic deficiencies of each method cannot be avoided by mere repetition. When the errors are truly systematic, statistical regularity will emerge from the ensemble of all measurements. Such a statistical regularity will not emerge when the data sets are afflicted with blunders, since blunders are generally one-time occurrences that can be de-

tected if a particular procedure is repeated. Consequently, redundancy within a given investigative procedure is desirable not only to improve precision, but also to purge the results of blunders.

Systematic errors can stem from many sources; typical systematic errors in the nuclear field, for example, can arise from cross sections used for neutron fluence determination, sample material standards, detector calibrations, shortcomings in deriving corrections (e.g., neutron multiple scattering), and nuclear decay properties. Once the sources of systematic error have been identified, it is necessary to estimate their respective magnitudes, corresponding to a consistent level of confidence. This is a very difficult task, since the applicable probability distribution laws are often unknown, and only an estimate of the ranges of possibilities for the variables in question may be available. The issue of confidence is important because the various error components must ultimately be combined to generate covariance matrices, and if the specific errors conform to widely different confidence levels, their combination may lead to misleading results.

Ideally, systematic errors should be estimated by theoretical analysis of the measurement conditions, based on the known properties of the measuring instruments and the quantity being measured. Although the systematic errors are reduced by introducing appropriate corrections, it is impossible to eliminate them completely. Ultimately, a residual error will always remain, and this residual error will then constitute the new systematic component of the measurement error. Good accuracy (i.e., small systematic error) can be obtained only by successfully reducing the magnitude of significant systematic errors. Important mathematical means for dealing with random and systematic errors will be presented in Chapter 2.

The quality of measurements that reflects the closeness of the results of measurements of the same quantity performed under the same conditions is called the *repeatability* of measurements. Good repeatability indicates that the random errors are small. On the other hand, the quality of measurements that reflects the closeness of the results of measurements of the same quantity performed under different conditions (e.g., in different laboratories, at different locations, and/or using different equipment) is called the *reproducibility* of measurements. Good reproducibility indicates that both the random and systematic errors are small.

Since the true value of the measured quantity is unknown, a measurement error cannot be found by using its definition as an algorithm. The measurement error must be evaluated by identifying its underlying sources and reasons, and by performing computations based on the estimates of all components of the respective measurement inaccuracy. The smallest of the measurement errors are customarily referred to as *elementary errors* (of a measurement), and are defined as those components of the overall measurement error that are associated with a single source of inaccuracy for the respective measurement. In turn, the total measurement error is computed by using the estimates of the component elementary errors. Even though it is sometimes possible to correct, partially, certain elementary errors (e.g., systematic ones), no amount or combination of corrections can produce an accurate measurement result; there always remains a residual error. In particular, the corrections themselves cannot be absolutely accurate, and, even after they are implemented, there remain residuals of the corresponding errors which cannot be eliminated and which later assume the role of elementary errors.

Since a measurement error can only be calculated indirectly, based on models and experimental data, it is important to identify and classify the underlying elementary errors. This identification and classification is subsequently used to develop mathematical models for the respective elementary errors. Finally, the resulting *overall measurement error* is obtained by synthesizing the mathematical models of the underlying elementary errors.

In the course of developing mathematical models for elementary errors, it has become customary to distinguish four types of elementary errors, namely absolutely constant errors, conditionally constant errors, purely random errors, and quasi-random errors. Thus, *absolutely constant errors* are defined as elementary errors that remain the same (i.e., are constant) in repeated measurements performed under the same conditions, for all measuring instruments of the same type. For example, an absolutely constant error arises from inaccuracies in the formula used to determine the quantity being measured, once the limits of the respective inaccuracies have been established. Typical situations of this kind arise in indirect measurements of quantities determined by linearized or truncated simplifications of nonlinear formulas (e.g., analog/digital instruments where the effects of electro-motive forces are linearized). Based on their properties, absolutely constant elementary errors are purely system-

atic errors, since each such error has a constant value in every measurement, but this constant is nevertheless unknown. Only the limits of these errors are known. Therefore, absolutely constant errors are often modeled mathematically by a determinate (as opposed to random) quantity whose magnitude lies within an interval of known limits.

Conditionally constant errors are, by definition, elementary errors that have definite limits (just like the absolutely constant errors) but (as opposed to the absolutely constant errors) such errors can vary within their limits due both to the non-repeatability and the non-reproducibility of the results. A typical example of such an error is the measurement error due to the intrinsic error of the measuring instrument, which can vary randomly between fixed limits. Usually, the conditionally constant error is mathematically modeled by a random quantity with a uniform probability distribution within prescribed limits. This mathematical model is chosen because the uniform distribution has the highest uncertainty (in the sense of information theory) among distributions with fixed limits. Note, in this regard, that the round-off error also has known limits, and this error has traditionally been regarded in mathematics as a random quantity with a uniform probability distribution.

Purely random errors appear in measurements due to noise or other random phenomena produced by the measuring device. In principle, the form of the distribution function for random errors can be found using data from each multiple measurement. In practice, however, the number of measurements performed in each experiment is insufficient for determining the actual form of the distribution function. Therefore, a purely random error is usually modeled mathematically by using a normal distribution characterized by a standard deviation that is computed from the experimental data.

Quasi-random errors occur when measuring a quantity defined as the average of nonrandom quantities that differ from one another such that their aggregate behavior can be regarded as a collection of random quantities. In contrast to the case of purely random errors, though, the parameters of the probability distribution for quasi-random errors cannot be unequivocally determined from experimental data. Therefore, a quasi-random error is modeled by a probability distribution with parameters (e.g., standard deviation) determined by expert opinion.

The term *uncertainty* is customarily used to express the inaccuracy of measurement results when the numerical value of the respective inaccuracy is accompanied by a corresponding confidence probability. In this respect, we also note that the second edition of the International Vocabulary of Basic and General Terms in Metrology (2nd edition, ISO 1993) defines the term *uncertainty* as "*an interval having a stated level of confidence.*"

The term *error* is customarily used for all components of uncertainty, while the term *limits of error* is used in cases in which the measurement inaccuracy is caused by the intrinsic error of the measuring instrument, and when a corresponding level of confidence cannot be stated. Measurements must be reproducible, since otherwise they lose their objective character and become meaningless. The *limits of measurement error or uncertainty* estimated by the experimenter provide a measure of the *nonreproducibility of a measurement permitted by the experimentalist.* The validity of the uncertainty computed for every measurement is based on the validity of the estimates of errors underlying the respective measurement. A correctly estimated measurement uncertainty permits comparisons of the respective result with results obtained by other experimenters.

When Q_m is the result of a measurement and Δ_U and Δ_L are the upper and lower limits of the error in the measurement, then *the result of a measurement and the respective measurement error* can be written in the form

$$(Q_m, \Delta_U, \Delta_L) \ , \quad \text{or} \quad (Q_m \pm \Delta) \tag{1.1}$$

when $|\Delta_L| = |\Delta_U| = \Delta$. When the inaccuracy of a measurement is expressed as uncertainty, then the corresponding confidence probability must also be given, usually in parentheses following the value of the uncertainty. For example, if a temperature measurement yields the value 316.24 K and the uncertainty for this result, say $\pm$ 0.2 K, was calculated for a confidence probability of 0.95, then this result should be written in the form

$$T(0.95) = (316.2 \pm 0.2)K \tag{1.2}$$

If the confidence probability is not indicated in the measurement result, e.g., if the result is written as

$$T = (316.2 \pm 0.2)K \tag{1.3}$$

then the inaccuracy is assumed to have been estimated without the use of probability methods. Although an error estimate obtained without the use of probability methods can be very reliable, it cannot be associated with a probability of one or some other value. In other words, when a probabilistic model was not employed to obtain the error estimate, the respective probability cannot be estimated and, therefore, should not be indicated.

In many cases, it is of interest to know not only the limiting values of the total measurement error but also the characteristics of the random and systematic error components separately, in order to analyze discrepancies between results of measurements of the same quantity performed under different conditions. Knowing the error components separately is particularly important when the result of a measurement is to be used for computations together with other data that are not absolutely precise. Furthermore, the main sources of errors should also be described together with estimates of their contributions to the total measurement uncertainty. For a random error, for example, it is of interest to indicate the form and parameters of the distribution function underlying the observations, the method employed for testing the hypothesis regarding the form of the distribution function, the significance level used in the testing, and so on.

The number of significant figures retained in the number expressing the result of a measurement must correspond to the accuracy of the measurement. This means that the uncertainty of a measurement can be equal to one or two units (and should not exceed 5 units) in the last figure of the number expressing the result of the measurement. Since measurement uncertainty determines only the vagueness of the results, the uncertainty is customarily expressed in its final form by a number with one or two significant figures. Two figures are retained for the most precise measurements. At least two additional significant digits should be retained during intermediate computations, in order to keep the round-off error below the value of the final error. The numerical value of the result of a measurement must be represented such that the last decimal

digit is of the same rank as its uncertainty. Including a larger number of digits will not reduce the uncertainty of the result; however, using a smaller number of digits (by further rounding off the number) would increase the uncertainty and would make the result less accurate, thereby offsetting the care and effort invested in the measurement. The rules for rounding off and for recording the results of measurements have been established by convention, and are listed below:

1. If the decimal fraction in the numerical value of the result of a measurement terminates in 0's, then the 0's are dropped only up to the digit that corresponds to the rank of the numerical value of the error. If the digit being discarded is equal to 5 and the digits to its right are unknown or are equal to 0, then the last retained digit is not changed if it is even and it is increased by 1 if it is odd. (Example: if three significant digits are retained, the number 100.5 is rounded off to 100.0 and the number 101.5 is rounded off to 102.0).
2. The last digit retained is not changed if the adjacent digit being discarded is less than 5. Extra digits in integers are replaced by 0's, while extra digits in decimal fractions are dropped. (Example: the numerical value of the result of a measurement 1.5333 with an error in the limits ± 0.04 should be rounded off to 1.53; however, if the error limits are ± 0.001, the same number should be rounded off to 1.533.)
3. The last digit retained is increased by 1 if the adjacent digit being discarded is greater than 5, or if it is 5 and there are digits other than 0 to its right. (Example: if three significant digits are retained, the number 2.6351 is rounded off to 2.64.)

1.2 Uncertainties and Probabilities

1.2.1 Axiomatic, Frequency, and Subjective Probability

Probability theory is a branch of mathematical sciences that provides a model for describing the process of observation. The need for probability theory arises from the fact that most observations of natural phenomena do not lead to uniquely predictable results. Probability theory provides the tools for dealing with actual variations in the outcome of realistic observations and measurements. The challenging pursuit to develop a probability theory which is mathematically rigorous and also describes many phenomena observable in nature has generated over the years notable disputes over conceptual and logical issues. Modern probability theory is based on postulates constructed from three axioms attributed to Kolmogorov [108], all of which are consistent with the notion of frequency of occurrence of events. The alternative approach, traceable to Laplace [111], is based on the concept that probability is simply a way of providing a numerical scale to quantify our reasonable beliefs about a situation which we know only incompletely. This approach is consistent with Bayes' theorem, conditional probabilities, and inductive reasoning. Either approach to probability theory would completely describe a natural phenomenon if sufficient information were available to determine the underlying probability distribution exactly. In practice, though, such exact knowledge is seldom, if ever, available so the features of the probability distribution underlying the physical phenomenon under consideration must be *estimated*. Such estimations form the study object of *statistics*, which is defined as the branch of mathematical sciences that uses the results of observations and measurements to estimate, in a mathematically well-defined manner, the essential features of probability distributions.

Both statistics and probability theory use certain generic terms for defining the objects or phenomena under study. A *system* is the object or phenomena under study. It represents the largest unit being considered. A system can refer to a nuclear reactor, corporation, chemical process, mechanical device, biological mechanism, society, economy, or any other conceivable object that is under study. The *output* or *response* of a system is a result that can be

measured quantitatively or enumerated. The power of a nuclear reactor, the yield of a process, life span of cell, the atmospheric temperature and pressure, are all examples of system outputs. A *model* is a mathematical idealization that is used as an approximation to represent the system and its output. Models can be quite simple or highly complex; they can be expressed in terms of a single variable, many variables, or sets of nonlinear integro-differential equations. Regardless of its complexity, the model is an idealization of the system, so it cannot be exact: usually, the more complex the system, the less exact the model, particularly since the ability to solve exactly mathematically highly complex expressions diminishes with increasing complexity. In other words, the simpler the model, the easier it is to analyze but the less precise the results.

A *statistical model* comprises mathematical formulations that express the various outputs of a system in terms of probabilities. Usually, a statistical model is used when the system's output cannot be expressed as a fixed function of the input variables. Statistical models are particularly useful for representing the behavior of a system based on a limited number of measurements, and for summarizing and/or analyzing a set of data obtained experimentally or numerically. For example, there exist families of probability distributions, covering a wide range of shapes, which can be used for representing experimentally obtained data sets. The resulting statistical model can then be used to perform extrapolations or interpolations, for determining the probability of occurrence of some event in some specific interval of values, and so on. Statistical models are constructed based on the physical properties of the system. When the physical properties and, hence, the principles of operation of the system are well understood, the model will be derived from these underlying principles. Most often, though, the basic physical properties of the system under consideration are known incompletely. In simple cases, and when sufficient data are available, it is possible to select a general yet flexible family of statistical models and fit one member of this family to the observations. The validity of such fits is then verified by performing a sensitivity analysis to determine the sensitivity of the system's output to the selection of a specific model to represent the system. In situations when the physical properties of the system under investigation are ill understood, but measurements are nevertheless available, the model for the system is selected on a trial-and-error

basis, usually based not only on objective criteria but also on subjective expert opinion. Finally, when no data are available, it is possible to study the conceptual model on a computer to generate numerically synthetic observations. The previously mentioned statistical methods can then be applied to analyze the synthetic observations generated numerically, just as if they had been physical observations.

The group study to be measured or counted, or the conceptual entity for which predictions are to be made, is called *population*. A *model parameter* is a quantity that expresses a characteristic of the system; a parameter could be constant or variable. A *sample* is a subset of the population selected for study. An *experiment* is a sequence of a limited (or, occasionally, unlimited) number of trials. An *event* is an outcome of an experiment. The set of all events (i.e., the set that represents all outcomes of an experiment) is called the *event space* or sample space, of the experiment; the respective events are also referred to as *sample points*.

When a process operates on an event space such that the individual outcomes of the observations cannot be controlled, the respective process is called a *random process* and the respective events are called *random events*. Although the exact outcome of any single random trial is unpredictable, a sample of random trials of reasonable size is expected to yield a pattern of outcomes. In other words, randomness implies lack of deterministic regularity but existence of statistical regularity. Thus, random phenomena must be distinguished from totally unpredictable phenomena, for which no pattern can be construed, even for exceedingly large samples. Observations of physical phenomena tend to fall somewhere between total unpredictability, at the one extreme, and statistically well-behaved processes, at the other extreme. For this reason, experimentalists must invest a considerable effort to identify and eliminate, as much as possible, statistically unpredictable effects (e.g., malfunctioning equipment), in order to perform experiments under controlled conditions conducive to generating statistically meaningful results.

The quantitative analysis of statistical models relies on concepts of probability theory. The basic concepts of probability theory can be introduced in several ways, ranging from the intuitive notion of frequency of occurrence to the *axiomatic development* initiated by Kolmogorov [108], and including the subjective *inductive reasoning* ideas based on "degree of belief" as originally

formulated by Laplace and Bayes. For the purposes of this book (namely data evaluation, assimilation, model adjustment/calibration, sensitivity and uncertainty analysis of models and data), all three interpretations of probability will be employed in order to take advantage of their respective strengths. From a mathematical point of view, the concepts of probability theory are optimally introduced by using Kolmogorov's axiomatic approach, in which probability is postulated in terms of abstract functions operating on well-defined event spaces. This axiomatic approach avoids both the mathematical ambiguities inherent to the concept of relative frequencies and the pitfalls of inadvertently misusing the concept of inductive reasoning. Thus, consider that S is the sample space consisting of a certain number of events, the interpretation of which is momentarily left unspecified. Assigned to each subset A of S, there exists a real number $P(A)$, called a *probability*, defined by the following three axioms (Kolmogorov [108]):

AXIOM I (EXISTENCE): For every subset A in S, the respective probability exists and is nonnegative, i.e., $P(A) \geq 0$.

AXIOM II (ADDITIVITY): For any two subsets A and B that are disjoint (i.e., $A \cap B = \emptyset$), the probability assigned to the union of A and B is the sum of the two corresponding probabilities, i.e., $P(A \cup B) = P(A) + P(B)$.

AXIOM III (NORMALIZATION): The probability assigned to the entire sample space is one, i.e., $P(S) = 1$; in other words, the certain event has unit probability.

Note that the statements of the three axioms mentioned above are not entirely rigorous from the standpoint of pure mathematics, but have been deliberately simplified somewhat, in order to suit the scope of this book. A mathematically more precise definition of probability requires that the set of subsets to which probabilities are assigned constitute a so-called σ-field.

Several useful properties of probabilities can be readily derived from the

axiomatic definition introduced above:

$$
\begin{aligned}
&a)\ P\left(\overline{A}\right) = 1 - P\left(A\right), \text{ where } \bar{A} \text{ is the complement of } A;\\
&b)\ 0 \leq P\left(A\right) \leq 1;\\
&c)\ P\left(\emptyset\right) = 0,\ \textit{but}\ P\left(A\right) = 0\ \textit{does NOT mean that}\ A = \emptyset;\\
&d)\ \text{if } A \subset B,\ \text{then } P\left(A\right) \leq P\left(B\right);\\
&e)\ P\left(A \cup B\right) = P\left(A\right) + P\left(B\right) - P\left(A \cap B\right)\ .
\end{aligned}
\tag{1.4}
$$

The last relation above can be extended to any *finite sequence* of events $(E_1, E_2, \ldots, E_k)$ in a sample space S, in which case it becomes

$$
\begin{aligned}
P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P\left(E_i\right) - \sum_{i \neq j} P\left(E_i \cap E_j\right) + \sum_{i \neq j \neq k} P\left(E_i \cap E_j \cap E_k\right)\\
+ \ldots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} E_i\right)\ .
\end{aligned}
\tag{1.5}
$$

When the events $(E_1, E_2, \ldots, E_n)$ form a *finite sequence of mutually exclusive* events, Eq. (1.5) reduces to

$$
P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P\left(E_i\right). \tag{1.6}
$$

When extended to the infinite case $\to \infty$, Eq. (1.6) actually expresses one of the defining properties of the probability measure, which makes it possible to introduce the concept of *probability function* defined on an event space S. Specifically, a function P is a probability function defined on S if and only if it possesses the following properties:

$$
\begin{aligned}
&a)\ P\left(\emptyset\right) = 0,\ \text{ where } \emptyset \text{ is the null set.}\\
&b)\ P\left(S\right) = 1.\\
&c)\ 0 \leq P\left(E\right) \leq 1\ \text{ where } E \text{ is any event in } S.\\
&d)\ P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P\left(E_i\right), \text{where } \left(E_1, E_2, \ldots\right) \text{is any sequence of}\\
&\text{mutually exclusive events in S.}
\end{aligned}
$$

The concept of probability defined thus far cannot address conditions (implied or explicit) such as "What is the probability that event E_2 occurs if it is known that event E_1 has actually occurred?" In order to address such conditions, an additional concept, namely the concept of *conditional probability*, must be introduced. For this purpose, consider that E_1 and E_2 are any two events in an event space S, with $P(E_1) > 0$. Then, the conditional probability $P(E_2 | E_1)$ that event E_2 occurs given the occurrence of E_1 is defined as

$$P(E_2 | E_1) \equiv \frac{P(E_1 \cap E_2)}{P(E_1)} . \tag{1.7}$$

From the definition introduced in Eq. (1.7), it follows that

$$P(E_1 \cap E_2) = P(E_2 | E_1) P(E_1) . \tag{1.8}$$

Conditional probabilities also satisfy the axioms of probability, as can be seen from the definition introduced in Eq. (1.7). In particular, since $P(E_1) = P(E_1 | S)$, the foregoing definition implies that the usual probability $P(E_1)$ can itself be regarded as a conditional probability, i.e., $P(E_1)$ is the conditional probability for E_1 given S.

The definition of conditional probability can be readily extended to more than two events by partitioning the sample space S into k mutually exclusive events $(E_1, E_2, \ldots, E_k)$, and by considering that $E \subset S$ is an arbitrary event in S with $P(E) > 0$. Then the probability, $P(E_i | E)$, that event E_i occurs given the occurrence of E is expressed as

$$P(E_i | E) = \frac{P(E | E_i) P(E_i)}{\sum_{j=1}^{k} P(E | E_j) P(E_j)}, \quad (i = 1, 2, \ldots, k) , \tag{1.9}$$

where $P(E | E_i)$ denotes the probability that event E occurs, given the occurrence of E_i. Equation (1.9) is known as *Bayes' theorem*, and is of fundamental importance to practical applications of probability theory to the evaluation of scientific data. Two events, E_1 and E_2, are said to be *statistically* independent if $P(E_2 | E_1) = P(E_2)$. This means that the occurrence (or nonoccurrence) of E_1 does not affect the occurrence of E_2. Note that if E_1 and E_2 are statistically independent, then $P(E_1 E_2) = P(E_1) P(E_2)$, $P(E_1 | E_2) = P(E_1)$, and

also conversely. Actually, the following statements are equivalent:

$$E_1 \text{ and } E_2 \text{ are statistically independent; or } P(E_1 | E_2) = P(E_1). \quad (1.10)$$

The numerical representation of the elementary events E in a set S is accomplished by introducing a function, say, X, which operates on all events $E \subset S$ in such a way as to establish a particular correspondence between E and the real number $x = X(E)$. The function X defined in this way is called a *random variable*. It is very important to note that X itself is the random variable, rather than the individual values $x = X(E)$, which X generates by operating on E. Thus, although X has the characteristics of a "function" rather than those of a "variable," the usual convention in probability theory and statistics is to call X a "random variable" rather than a "random function." There also exist "functions of random variables," which may, or may not, be random themselves.

Random events $E \subset S$ that can be completely characterized by a *single-dimensional random variable* X are called *single-variable events*. The qualifier "single-dimensional" is omitted when it is apparent from the respective context. The concept of single-dimensional random variable, introduced to represent numerically a single-variable event, can be extended to a multivariable event $E \subset S$ by considering that S is a sample space in k-dimensions. If each random variable X_i $(i = 1, 2, \ldots, k)$ is a real-valued function defined on a domain N_i (representing the i^{th}dimension of S), then $(X_1, \ldots, X_k)$ is a *multivariate random variable* or *random vector*. Furthermore, consider that each domain N_i $(i = 1, 2, \ldots, k)$ is a discrete set, either finite or denumerably infinite (N_i is usually the set of nonnegative integers or a subset thereof). Then, a *probability function*, $p(x_1, \ldots, x_k)$, of the discrete random vector $(X_1, \ldots, X_k)$ is defined by requiring $p(x_1, \ldots, x_k)$ to satisfy, for each value x_i taken on by X_i $(i = 1, 2, \ldots, k)$, the following properties:

$$\text{(i)} \quad p(x_1, x_2, \ldots, x_k) = P\{X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\} \quad (1.11)$$

and

$$\text{(ii)} \quad P\{A\} = \sum_{(x_1, \ldots, x_k) \in A} p(x_1, \ldots, x_k), \quad (1.12)$$

for any subset A of N, where N is the k-dimensional set whose i^{th}component is $N_i \ \ (i = 1, 2, \ldots, k)$ with $P\{N\} = 1$. Consider that A is the set of all random vectors $(X_1, \ldots, X_k)$ such that $\mathbf{X}_i \leq x_i \ \ (i = 1, 2, \ldots, k)$. Then

$$P\{A\} = P\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k\}, \tag{1.13}$$

is called the *cumulative distribution function* (CDF) of $(X_1, \ldots, X_k)$. The usual notation for the CDF of $(X_1, \ldots, X_k)$ is $F(x_1, \ldots, x_k)$. Note that the cumulative distribution function is not a random variable; rather, it is a real numerical-valued function whose arguments represent compound events.

To define the probability density function of a *continuous random vector* $(X_1, \ldots, X_k)$, consider that $(X_1, \ldots, X_k)$ is a random vector whose i^{th}component, X_i, is defined on the real line $(-\infty, \infty)$ or on a subset thereof. Suppose $p(x_1, \ldots, x_k) > 0$ is a function such that for all $x_i \in [a_i, b_i], \ (i = 1, 2, \ldots, k)$, the following properties hold:

$$\text{(i)} \ \ P\{a_1 < X_1 < b_1, \ldots, a_k < X_k < b_k\} = \int_{a_k}^{b_k} \ldots \int_{a_1}^{b_1} p(x_1, \ldots, x_k)\, dx_1 \ldots dx_k \tag{1.14}$$

and if A is any subset of k-dimensional intervals,

$$\text{(ii)} \ \ P\{A\} = \int\limits_{(x_1, \ldots, x_k) \in A} \ldots \int p(x_1, x_2, \ldots, x_k)\, dx_1\, dx_2 \ldots dx_k, \tag{1.15}$$

then $p(x_1, \ldots, x_k)$ is said to be a *joint probability density function* (PDF) of the continuous random vector $(X_1, \ldots, X_k)$, if it is normalized to unity over its domain.

Consider that the set $(x_1, \ldots, x_k)$ represents a collection of random variables with a multivariate joint probability density $p(x_1, \ldots, x_k) > 0$; then, the *marginal probability density* of x_i, denoted by $p_i(x_i)$, is defined as

$$p_i(x_i) = \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_{i-1} \int_{-\infty}^{\infty} dx_{i+1} \ldots \int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_k)\, dx_k \tag{1.16}$$

In addition, the *conditional probability density function* (PDF) $p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k | x_i)$ can be defined whenever $p_i(x_i) \neq 0$, by means

of the expression

$$p\left(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k|x_i\right)=p\left(x_1,\ldots,x_k\right)/p_i\left(x_i\right)\ . \tag{1.17}$$

The meaning of marginal and conditional probability can be illustrated by considering bivariate distributions. Thus, if (X,Y) is a discrete random vector whose joint probability function is $p\left(x,y\right)$, where X is defined over N_1 and Y is defined over N_2, then the *marginal probability distribution function* (PDF) of X, $p_x\left(x\right)$, is defined as

$$p_x\left(x\right)=\sum_{y\in N_2}p\left(x,y\right). \tag{1.18}$$

On the other hand, if (X,Y) is a continuous random vector whose joint PDF is $p\left(x,y\right)$, where X and Y are each defined over the domain $(-\infty,\infty)$, then the marginal PDF of X, $p_x\left(x\right)$, is defined by the integral

$$p_x\left(x\right)=\int_{-\infty}^{\infty}p\left(x,y\right)\ dy\,. \tag{1.19}$$

Consider that (X,Y) is a random vector (continuous or discrete) whose joint PDF is $p\left(x,y\right)$. The *conditional PDF* of y given $X=x$ (fixed), denoted by $h\left(y|\,x\right)$, is defined as

$$h\left(y\,|x\,\right)=\frac{p\left(x,y\right)}{p_x\left(x\right)}\,, \tag{1.20}$$

where the domain of y may depend on x, and where $p_x\left(x\right)$ is the marginal PDF of X, with $p_x\left(x\right)>0$.

Similarly, the conditional PDF, say $g\left(x\,|y\,\right)$, for x given y is

$$g\left(x\,|y\,\right)=\frac{p\left(x,y\right)}{p_y\left(y\right)}=\frac{p\left(x,y\right)}{\int p\left(x',y\right)\ dx'}\,. \tag{1.21}$$

Combining equations (1.20) and (1.21) gives the following relationship between $g\left(x\,|y\,\right)$ and $h\left(y\,|x\,\right)$,

$$g\left(x\,|y\,\right)=\frac{h\left(y\,|x\,\right)\ p_x\left(x\right)}{p_y\left(y\right)}\,, \tag{1.22}$$

which expresses *Bayes' theorem for the case of continuous variables*.

Consider that (X, Y) is a random vector whose joint PDF is $p(x, y)$. The random variables X and Y are called *stochastically independent* if and only if

$$p(x, y) = p_x(x)\ p_y(y), \tag{1.23}$$

over the entire domain of (X, Y) (i.e., for all x and y). From this definition and from Eq. (1.21), it follows that X and Y are independent if and only if $g(x\,|y) = p_x(x)$ over the entire domain of (X, Y); this definition of stochastic independence can be generalized to random vectors.

Consider that $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are two distinct vector random variables that describe the same events. Consider, further, that the respective multivariate probability densities $p_x(x)$ and $p_y(y)$ are such that the mappings $y_i = y_i(x_1, \ldots, x_n)$, $(i = 1, 2, \ldots, n)$, are continuous, one-to-one, and all partial derivatives $\partial y_i/\partial x_j$, $(i, j = 1, \ldots, n)$, exist. Then, the transformation from one PDF to the other is given by the relationship

$$p_y(y)\,|dy| = p_x(x)\,|dx|, \quad \text{or} \quad p_x(x) = |J|\ p_y(y)\ , \tag{1.24}$$

where $|J| \equiv \det\,|\partial y_i/\partial x_j\,|$, $(i, j = 1, \ldots, n)$, is the Jacobian of the respective transformation. An alternative way of writing the above relation is

$$p_y(y) = p_x\left(f^{-1}(y)\right)\left|\frac{\partial f^{-1}(y)}{\partial y}\right|; \quad \frac{\partial f^{-1}(y)}{\partial y} \triangleq \left[\left\{\frac{\partial f(x)}{\partial x}\right\}_{x=f^{-1}(y)}\right]^{-1}. \tag{1.25}$$

If the inverse function $f^{-1}(y)$ is not unique, i.e., then the above relation becomes, more generally:

$$p_y(y) = \sum_{i=1}^{m} p_x\left(f_i^{-1}(y)\right)\left|\frac{\partial f_i^{-1}(y)}{\partial y}\right|; \quad f_i^{-1}(y) = {x_i}^{-1}. \tag{1.26}$$

For example, the distribution $p_y(y)$ of a quadratic function $y = ax^2 + bx + c$, where a, b, c are known constants and the distribution $p_x(x)$ of x is known, is found by using Eq. (1.26) to obtain

$$p_y(y) = \left\{ p_x\left[\frac{-b + \left(b^2 - 4ac + 4ay\right)^{1/2}}{2a}\right] + p_x\left[\frac{-b - \left(b^2 - 4ac + 4ay\right)^{1/2}}{2a}\right]\right\} \times \left(b^2 - 4ac + 4ay\right)^{-1/2}.$$

The dimensionality of a random number is defined by the dimensionality of the particular set of values $x = X(E)$, generated by the function X when operating on the events E. Although the cardinalities (i.e., dimensions) of event sets and corresponding random variables are usually equal, it is important to note that the explicit dimensionality of E is irrelevant in this context. By definition, *finite random variables* are those for which the set of values x obtainable from X operating on E is in one-to-one correspondence with a finite set of integers. *Infinite but discrete random variables* are those for which the set of values x obtainable from X operating on E is in one-to-one correspondence with the infinite set of all integers. *Unaccountable or nondenumerable random variables* are those for which the set of values x obtainable from X operating on E is in one-to-one correspondence with the infinite set of all real numbers.

As discussed above, random variables are actually well-behaved functions that operate on event spaces to yield numerical values that characterize, in turn, the respective events. Random variables can also serve as arguments of other functions, whenever these functions are well behaved in the sense that such functions are bounded and are devoid of singularities except, perhaps, for a finite number of jump-discontinuities.

The main interpretations of probability commonly encountered in data and model analysis are that of *relative frequency* (which is used, in particular, for assigning statistical errors to measurements) and that of *subjective probability* (which is used, in particular, to quantify systematic uncertainties). These two interpretations will be discussed in more detail below.

When probability is interpreted as a *limiting relative frequency*, the elements of the set S correspond to the possible outcomes of a measurement, assumed to be (at least hypothetically) repeatable. A subset E of S corresponds to the occurrence of any of the outcomes in the subset. Such a subset is called an *event*, which is said to occur if the outcome of a measurement is

in the subset. A subset of S consisting of only one element denotes an *elementary outcome*. In turn, the probability of an elementary outcome E is defined as the fraction of times that E occurs in the limit when the measurement is repeated infinitely many of times, namely:

$$P(E) = \lim_{n\to\infty} \frac{\text{number of occurences of outcome } E \text{ in } n \text{ measurements}}{n}. \tag{1.27}$$

The probability for the occurrence of any one of several outcomes (i.e., for a non-elementary subset E) is determined from the probabilities for individual elementary outcomes by using the addition rule provided by the axioms of probability. These individual probabilities correspond, in turn, to relative frequencies of occurrence. The "relative frequency" interpretation is consistent with the axioms of probability since the fraction of occurrence is always greater than or equal to zero, the frequency of any outcome is the sum of the individual frequencies of the individual outcomes (as long as the set of individual outcomes is disjoint), and the measurement must, by definition, eventually yield some outcome [i.e., $P(S) = 1$]. Correspondingly, the conditional probability $P(E_2\,|E_1)$ represents the number of cases where both E_2 and E_1 occur divided by the number of cases in which E_1 occurs, regardless of whether E_2 occurs. In other words, $P(E_2\,|E_1)$ gives the frequency of E_2 with the subset E_1 taken as the sample space.

The interpretation of probability as a relative frequency is straightforward when studying physical laws, since such laws are assumed to act the same way in repeated experiments, implying that the validity of the assigned probability values can be tested experimentally. This point of view is appropriate, for example, in nuclear and particle physics, where repeated collisions of particles constitute repetitions of an experiment. Note, though, that the probabilities based on such an interpretation can never be determined experimentally with perfect precision. Hence, the fundamental tasks of classical statistics are to interpret the experimental data by: (a) estimating the probabilities (assumed to have some definite but unknown values) of occurrence of events of interest, given a finite amount of experimental data, and (b) testing the extent to which a particular model or theory that predicts the probabilities estimated in (a) is compatible with the observed data.

The concept of probability as a relative frequency becomes questionable

when attempting to assign probabilities for very rare (or even uniquely occurring) phenomena such as a core meltdown in a nuclear reactor or the big bang. For such rare events, the frequency interpretation of probabilities might perhaps be rescued by imagining a large number of similar universes, in some fraction of which the rare event under consideration would occur. However, such a scenario is pure utopia, even in principle; therefore, the frequency interpretation of probability must be abandoned in practice when discussing extremely rare events. In such cases, probability must be considered as a mental construct to assist us in expressing a degree of belief about the single universe in which we live; this mental construct provides the premises of the Bayesian interpretation of probability, which will be discussed next.

Complementary to the frequency interpretation of probability is the so-called *subjective* (also called *Bayesian*) *probability*. In this interpretation, the elements of the sample space are considered to correspond to *hypotheses* or *propositions*, i.e., statements that are either true or false; the sample space is often called the *hypothesis space*. Then, the probability associated with a cause or hypothesis A is interpreted as a measure of degree of belief, namely:

$$P(A) \equiv \text{ a priori measure of the rational degree of belief that } A \text{ is the correct cause or hypothesis.} \tag{1.28}$$

The sample space S must be constructed so that the elementary hypotheses are mutually exclusive, i.e., only one of them is true. A subset consisting of more than one hypothesis is true if any of the hypotheses in the subset is true. This means that the union of sets corresponds to the Boolean OR operation, while the intersection of sets corresponds to the Boolean AND operation. One of the hypotheses must necessarily be true, implying that $P(S) = 1$.

Since the statement "a measurement will yield a given outcome for a certain fraction of the time" can be regarded as a hypothesis, it follows that the framework of subjective probability includes the relative frequency interpretation. Furthermore, subjective probability can be associated with (for example) the value of an unknown constant; this association reflects one's confidence that the value of the respective probability is contained within a certain fixed interval. This is in contrast with the frequency interpretation of probability, where the "probability for an unknown constant" is not meaningful, since if we

repeat an experiment depending on a physical parameter whose exact value is not certain, then its value is either never or always in a given fixed interval. Thus, in the frequency interpretation, the "probability for an unknown constant" would be either zero or one, but we do not know which. For example, the mass of a physical quantity (e.g., neutron) may not be known exactly, but there is considerable evidence, in practice, that it lays between some upper and lower limits of a given interval. In the frequency interpretation, the statement "the probability that the mass of the neutron lies within a given interval" is meaningless. By contrast, though, a subjective probability of 90% that the neutron mass is contained within the given interval is a meaningful reflection of one's state of knowledge.

The use of subjective probability is closely related to Bayes' theorem and forms the basis of *Bayesian* (as opposed to classical) *statistics*. For example, in Bayesian statistics for the particular case of two subsets, the subset E_2 appearing in the definition of conditional probability is interpreted as the hypothesis that "a certain theory is true," while the subset E_1 designates the hypothesis that "an experiment will yield a particular result (i.e., data)." In this interpretation, Bayes' theorem takes on the form

$$P\,(theory\,|data\,) \propto P\,(data\,|theory\,) \cdot P\,(theory)\ . \tag{1.29}$$

In the above expression of proportionality, $P\,(theory)$ represents the *prior probability* that the theory is true, while the *likelihood* $P\,(data|\,theory)$ expresses the probability of observing the data that were actually obtained under the assumption that the theory is true. The *posterior probability*, that the theory is correct after seeing the result of the experiment, is given by $P\,(theory\,|data\,)$. Note that the prior probability for the data, $P\,(data)$, does not appear explicitly, so the above relation expresses a proportionality rather than an equality. Furthermore, Bayesian statistics provides no fundamental rule for assigning the prior probability to a theory. However, once a prior probability has been assigned, Bayesian statistics indicates how one's degree of belief should change after obtaining additional information (e.g., experimental data).

The choice of the "most appropriate" prior distribution lies at the heart of applying Bayes' theorem to practical problems, and has caused considerable

debates over the years. Thus, when prior information related to the problem under consideration is available and can be expressed in the form of a probability distribution, this information should certainly be used. In such cases, the repeated application of Bayes' theorem will serve to refine the knowledge about the respective problem. When only scant information is available, the maximum entropy principle (as described in statistical mechanics and information theory) is the recommended choice for constructing a prior distribution. Finally, in the extreme case when no information is available, the general recommendation is to use a continuous uniform distribution as the prior. In any case, the proper repeated use of Bayes' theorem ensures that the impact of the choice of priors on the final result diminishes as additional information (e.g., measurements) containing consistent data is successively incorporated. These issues will be revisited and set in a proper mathematical framework in Chapter 2.

As can be deduced from the previously mentioned *"axioms of probability,"* the fundamental relationships of probability theory are the sum and product rules

$$P\left(A|B\right) + P\left(\overline{A}|B\right) = 1. \tag{1.30}$$

$$P\left(AB|C\right) = P\left(A|BC\right)P\left(B|C\right) = P\left(B|AC\right)P\left(A|C\right). \tag{1.31}$$

where A, B, and C represent propositions (e.g., "the dice shows six"); AB signifies "both A and B are true," $\overline{A}$ signifies "A is false," and $P\left(A|B\right)$ denotes the "probability of A given B." The above notation indicates that all probability assignments are conditional, based on assumptions, on empirical and/or theoretical information. The two forms of the product rule reflect the symmetry $AB = BA$. As argued already by Bernoulli [16] and Laplace [111], probabilities can be interpreted as degrees of plausibility or rational expectation on a numerical scale ranging from 0 (impossibility) to 1 (certainty); intermediate values reflect intermediate degrees of plausibility. The sum rule indicates that, under all circumstances B, the more likely is A the less likely is $\overline{A}$; the unit sum of both probabilities indicates that one of these alternatives must be certainly true. The product rule indicates that, under all circumstances C, the probability that both A and B are true is equal to the probability of A given B, times the probability that, in fact, B is true. Since A and B enter symmetrically one can also take the probability of B given A and multiply it

by the probability of A. The interpretation of the Ps as degrees of plausibility has been occasionally criticized in the past based on the following arguments: (i) probability must only mean "relative frequency in a random experiment" (e.g., coin tossing, in the limit of very many repetitions); and (ii) if causes (stochastic laws and their parameters) are given, then probabilities of effects (observations) can be assigned, but not vice versa (i.e., probabilities cannot be assigned to various possible causes if observations are given); (iii) since physical constants are not random variables that take on given values with certain frequencies, they should not have probability distributions associated with them.

The above arguments would be too restrictive for data evaluation, since they would not permit statements like "according to measured data, the value of a physical constant has the probability P of lying between given limits." The practical problem of estimating values of physical quantities (e.g., natural constants, half-lives, reaction cross sections) from error-affected, uncertain and incomplete data is not a random experiment that can be repeated at will, but is a problem to be addressed by inductive inference (i.e., reasoning in the face of uncertainty). Therefore, the Bernoulli-Laplace concept of probability is more appropriate for data analysis and evaluation (as opposed to a purely statistical viewpoint), since, for example, a probability distribution associated with a measured value for a physical constant does not imply that the constant varies; it merely indicates how plausible various possible values are for the true but unknown value of the respective physical constant. Under the general assumptions that $P\left(\overline{A}|B\right)$ depends on $P\left(A|B\right)$, and that $P\left(AB|C\right)$ depends on $P\left(A|BC\right)$ and $P\left(B|C\right)$, with probabilities interpreted as degrees of plausibility between 0 and 1, Cox [39] proved that any consistent scheme of logical inference must be equivalent to probability theory as derived from the basic sum and product rules expressed by Eqs. (1.30) and (1.31). Schrödinger [183] arrived independently at the same conclusions, and Rényi [168] proved these conclusions under the most general conditions, without needing Cox's assumptions of (twice-) differentiability of probability functions. These works have paved the way to the current consensus that:

1. Probabilities are not frequencies. They can be applied equally well to non-repetitive situations as to repeated trials.

2. All internally consistent schemes of logical inference (including fuzzy logic or artificial intelligence) must be equivalent to probability theory.

1.2.2 Bayes' Theorem for Assimilating New Information

Equation (1.31) can be written in the form

$$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|C)}. \tag{1.32}$$

which expresses *Bayes' theorem* (1763) in its simplest form. In the course of evaluating experimental data, Bayes' theorem is used to compute *the updated* or *a posteriori probability* ("posterior"), $P(A|BC)$, which is proportional to the product $P(B|AC)P(A|C)$, where $P(B|AC)$ is the *likelihood function* and $P(A|C)$ is the *a priori probability* (or "prior," for short). In practice, the experimental data, B, to be evaluated and/or assimilated depends on the value of an unknown physical A and on other circumstances C. The likelihood function $P(B|AC)$ is provided by the theoretical model employed by the evaluator to assess how likely the data B would be under the circumstances C if the unknown quantity were in fact A. The prior reflects the information available about the data A before it actually became available. The likelihood function reflects the impact of the data, and the posterior contains the complete information available for further inference and prediction. For continuous variates A and B, the finite probabilities $P(B|AC)$ and $P(A|C)$ are replaced by infinitesimal probabilities $p(B|AC)dB$ and $p(A|C)dA$, where $p(B|AC)$ and $p(A|C)$ denote the respective probability densities.

The generalization of Eq. (1.32) to N distinct, mutually exclusive alternatives $A_\nu, \quad \nu = 1, 2, \ldots, N$, takes the form

$$P(A_\nu|BC) = \frac{P(B|A_\nu C)P(A_\nu|C)}{\sum\limits_{\nu} P(B|A_\nu C)P(A_\nu|C)}, \quad \nu = 1, 2, \ldots, N. \tag{1.33}$$

The above probability is normalized to unity as demanded by the sum rule.

For continuous variates, Eq. (1.33) takes the form

$$p(A|BC)dA = \frac{p(B|AC)p(A|C)dA}{\int p(B|AC)p(A|C)dA}, \quad A_{\min} \leq A \leq A_{\max}. \tag{1.34}$$

The above forms of Bayes' theorem provide the foundation for data evaluation and assimilation, indicating how prior knowledge (e.g., a data file) is to be updated with new evidence (new data). Since the denominator in Bayes' theorem is simply a normalization constant, the formal rule for updating observations can be briefly stated as "*the posterior is proportional to the likelihood times the prior.*" It should be understood that the terms "posterior" and "prior" have a logical rather than a temporal connotation, implying "with" and, respectively, "without" having assimilated the new data.

As an illustration of the use of Bayes' theorem for data evaluation and assimilation, consider the determination of the decay constant λ of some short-lived radioisotope from decays registered at times $t_1, t_2, \ldots, t_n$. In this case, the unknown value of the decay constant λ corresponds to A, the data $t_1, \ldots, t_n$ corresponds to B, while all other information about the situation such as applicability of the exponential decay law, purity of the sample, reliability of the recording apparatus, sufficiently long observation time for all observable decays to be recorded, and so on, corresponds to C. The statistical model for the experiment is represented by the so-called "sampling distribution," i.e., by the probability with which various alternatives can be "reasonably" expected after sampling once, given the parameters of the model. In this example, the statistical model is the probability that, given λ, one particular decay (e.g., the i^{th} one), is recorded in the time interval dt_i at t_i, namely

$$p\left(t_i|\lambda\right) dt_i = \exp\left(-\lambda t_i\right) \lambda dt_i, \qquad 0 < t_i < \infty. \tag{1.35}$$

Writing probability distributions in the above form, with the probability density p multiplied by the corresponding differential (i.e., as an infinitesimal probability), and with the range of the random variable explicitly stated, emphasizes the fact that all probability distributions are ultimately used for computing expectation values, as integrands subject to change of variables. Furthermore, the above notation omits explicit reference to the background information C.

Using the above models together with the product rule, it follows that the joint probability of observing the mutually independent data $t_1, \ldots, t_n$, given λ, is

$$p\,(t_1, \ldots, t_n|\lambda)\; dt_1, \ldots, dt_n = \exp\left(-\lambda \sum_{i=1}^{n} t_i\right) \lambda^n\, dt_1, \ldots, dt_n. \qquad (1.36)$$

The above expression corresponds to the likelihood $p\,(A|B)\,dB$. In this example, the likelihood function does not depend on the individual sample values, since they appear only in the form $\sum_i t_i \equiv n\bar{t}$. Hence, for a given n, the sample average $\bar{t}$ carries all the information contained in the data; customarily, $\bar{t}$ is called a "sufficient statistic," or an "ancillary statistic."

The prior is $p\,(\lambda)\,d\lambda$, and its expression will be discussed and specified later. According to Bayes' theorem, the posterior probability is proportional to the product of the likelihood function and the prior probability. Hence, multiplying the likelihood function $p\,(t_1, \ldots, t_n|\lambda)$ (which conveys the impact of the data) by the prior $p\,(\lambda)\,d\lambda$ (which summarizes the information that was already known without the data) leads to

$$p\,(\lambda|t_1, \ldots, t_n)\; d\lambda \;\; \propto \;\; \exp\left(-\lambda \sum_{i=1}^{n} t_i\right) \lambda^n\, p\,(\lambda)\, d\lambda \;\; , \qquad 0 < \lambda < \infty. \;\; (1.37)$$

where $\propto$ denotes proportionality. When all values of λ between 0 and ∞ are considered as being equally probable a priori, the prior becomes $p\,(\lambda)\,d\lambda \;\propto\; d\lambda$, and the above expression simplifies to

$$p\left(\lambda|n\bar{t}\right) d\lambda \;\sim\; e^{-\lambda\, n\, \bar{t}}\lambda^n d\lambda \;\; , \qquad 0 < \lambda < \infty. \qquad (1.38)$$

Normalizing the above distribution to unity leads to the following properly normalized posterior distribution:

$$p\left(\lambda|n\bar{t}\right) d\lambda = \Gamma(n+1)^{-1} e^{-\,x} x^n dx \;\; , \qquad 0 < x \equiv \lambda n\bar{t} < \infty, \qquad (1.39)$$

where $\Gamma\,(n+1) \equiv \int_0^\infty e^{-\,x} x^n dx$ is the customary gamma function. The above distribution is the "gamma distribution" or the "chi-square distribution with $\nu \equiv 2n+2$ degrees of freedom"; it represents the complete information about λ that can be obtained from the data and the assumed prior. As the sample size

n increases, the posterior becomes more and more concentrated, indicating that the more data are available, the better defined is λ.

1.3 Moments, Means, and Covariances

Probabilities cannot be measured directly; they can be inferred from the results of observations or they can be postulated and (partially) verified through accumulated experience. In practice, though, certain random vectors tend to be more probable, so that most probability functions of practical interest tend to be localized. Therefore, the essential features regarding probability distributions of practical interest are measures of *location* and of *dispersion*. These measures are provided by the *expectation* and *moments* of the respective probability function. If the probability function is known, then these moments can be calculated directly, through a process called *statistical deduction*. Otherwise, if the probability function is not known, then the respective moments must be estimated from experiments, through a process called *statistical inference*.

Consider that $\mathbf{x} = (x_1, \ldots, x_n)$ is a collection of random variables that represent the events in a space E, and consider that S_x represents the n-dimensional space formed by all possible values of $\mathbf{x}$. The space S_x may encompass the entire range of real numbers (i.e., $-\infty < x_i < \infty,\ i = 1, \ldots, n$) or a subset thereof. Furthermore, consider a real-valued function, $g(\mathbf{x})$, and a probability density, $p(\mathbf{x})$, both defined on S_x. Then, the *expectation* of $g(\mathbf{x})$, denoted as $E[g(\mathbf{x}])$, is defined as:

$$E[g(\mathbf{x})] \equiv \int_{S_x} g(\mathbf{x})\ p(\mathbf{x})\ d\mathbf{x}, \tag{1.40}$$

if the condition of *absolute convergence*, namely

$$E(|g|) = \int_{S_x} |g(\mathbf{x})|\ p(\mathbf{x})\ d\mathbf{x} < \infty, \tag{1.41}$$

is satisfied. When $\mathbf{x}$ is discrete with the domain N, the expectation of g is

defined as

$$E\left[g\left(\mathbf{x}\right)\right] = \sum_{\mathbf{x}\in N} g\left(\mathbf{x}\right)\, p\left(\mathbf{x}\right), \tag{1.42}$$

provided that $E\left(\left|g\left(\mathbf{x}\right)\right|\right) < \infty$.

In particular, the *moment of order* k about a point c is defined for a univariate probability function as

$$E\left[\left(x-c\right)^{k}\right] \equiv \int\limits_{S_x} \left(x-c\right)^{k} p\left(x\right)\, dx\,, \tag{1.43}$$

where S_x denotes the set of values of x for which $p\left(x\right)$ is defined, and the integral above is absolutely convergent.

For a multivariate probability function, given a collection of n random variables $\left(x_1,\ldots,x_n\right)$ and a set of constants $\left(c_1,\ldots,c_n\right)$, the *mixed moment of order* k is defined as

$$\begin{aligned} &E\left[\left(x_1-c_1\right)^{k_1},\ldots,\left(x_n-c_n\right)^{k_n}\right] \equiv \\ &\int\limits_{S_{x_1}} dx_1 \ldots \int\limits_{S_{x_n}} dx_n \left(x_1-c_1\right)^{k_1} \ldots \left(x_n-c_n\right)^{k_n} p\left(x_1 \ldots x_n\right)\,. \end{aligned} \tag{1.44}$$

The *zeroth-order moment* is obtained by setting $k=0$ (for univariate probability) in Eq. (1.43) or by setting $k_1=\ldots=k_n=0$ (for multivariate probability) in Eq. (1.44), respectively. *Since probability functions are required to be normalized, the zeroth moment is always equal to unity.*

In particular, when $c=0$ (for a univariate probability) or when $c_1=\ldots=c_n=0$ (for a multivariate probability), the quantities defined as $\nu_k \equiv E\left(x^k\right)$, and, respectively, $\nu_{k_1\ldots k_n} \equiv E\left(x_1^{k_1}\ldots x_n^{k_n}\right)$ are called the *moments about the origin* (often also called *raw* or *crude moments*). If $\sum\limits_{i=1,n} k_i = k$, then the moments of the form $\nu_{k_1\ldots k_n}$ are called the *mixed raw moments of order* k. For $k=1$, these moments are called *mean values*, and are denoted as $m_o = \nu_1 = E\left(x\right)$ for univariate probability, and $m_{oi} \equiv \nu_{o\ldots1\ldots o} \equiv E\left(x_i\right)$, $\left(i=1,\ldots,n\right)$, for multivariate probability, respectively. Note that a "0" in the j^{th} subscript position signifies that $k_j=0$ for the particular raw moment in question, while a "1" in the i^{th} subscript position indicates that $k_i=1$ for the respective moment.

The moments about the mean or *central moments* are defined as

$$\mu_k \equiv E\left[(x - m_o)^k\right], \quad \text{for univariate probability,} \tag{1.45}$$

and

$$\mu_{k_1 \ldots k_n} \equiv E\left[(x_1 - m_{o1})^{k_1} \ldots (x_n - m_{on})^{k_n}\right], \quad \text{for multivariate probability.} \tag{1.46}$$

Furthermore, if $\sum_{i=1,n} k_i = k$, then the above moments are called the *mixed central moments* of order k. Note that the central moments vanish whenever one particular $k_i = 1$ and all other $k_j = 0$, i.e., $\mu_{o \ldots 1 \ldots o} = 0$. Note also that all even-power central moments of univariate probability functions (i.e., μ_k, for $k = even$) are nonnegative.

1.3.1 Means and Covariances

The central moments for $k = 2$ play very important roles in statistical theory, and are therefore assigned special names. Thus, for univariate probability, the second moment, $\mu_2 \equiv E\left[(x - m_o)^2\right]$, is called *variance*, and is usually denoted as $\text{var}\,(x)$ or σ^2. The positive square root of the variance is called the *standard deviation*, denoted as σ, and defined as

$$\sigma \equiv [\text{var}\,(x)]^{1/2} \equiv \mu_2^{1/2} \equiv \left\{E\left[(x - m_o)^2\right]\right\}^{1/2}. \tag{1.47}$$

The terminology and notation used for univariate probability are also used for multivariate probability. Thus, for example, the standard deviation of the i^{th} component is defined as:

$$\mu_{o \ldots 2 \ldots o} \equiv \text{var}\,(x_i) \equiv \sigma_i^2 \equiv E\left[(x_i - m_o)^2\right]. \tag{1.48}$$

To simplify the notation, the subscripts accompanying the moments ν and μ for multivariate probability functions are usually dropped in favor of a simpler alternative notation. For example, μ_{ii} is often employed to denote $\text{var}\,(x_i)$, and μ_{ij} signifies $E\left[(x_i - m_{oi})\,(x_j - m_{oj})\right]$, $(i, j = 1, \ldots, n)$.

The raw and central moments are related to each other through the im-

portant relationship

$$\mu_k = \sum_{i=0}^{k} C_i^k (-1)^i \nu_{k-i} \nu_1^i \quad (k \geq 1)\,, \tag{1.49}$$

where $C_i^k = k!/[(k-i)\,!\,i\,!]$ is the *binomial coefficient*. This formula is very useful for estimating central moments from sampling data, since, in practice, it is more convenient to estimate the raw moments directly from the data, and then derive the central moments by using the above equation. Since measurements rarely yield true values, it is necessary to introduce surrogate parameters to measure location and dispersion for the observed results. Practice indicates that *location* is *best described by the mean value*, while *dispersion* of observed results appears to be *best described by the variance, or standard deviation*. In particular, the mean value can be interpreted as a locator of the center of gravity, while the variance is analogous to the moment of inertia (which linearly relates applied torque to induced angular acceleration in mechanics). Also very useful for the study of errors is the *Minimum Variance Theorem*, which states that: *if c is a real constant and x is a random variable, then* $\mathrm{var}\,(x) \leq E\left[(x-c)^2\right]$.

Henceforth, when we speak of *errors in physical observations*, they *are to be interpreted as standard deviations*, unless explicitly indicated otherwise. In short, errors are simply the measures of dispersion in the underlying probability functions that govern observational processes. *The fractional relative error or coefficient of variation*, f_x, is defined by $f_x = \sigma/|E\,(x)|$, when $E\,(x) \neq 0$. The reciprocal, $(1/f_x\,)$, is commonly called (particularly in engineering applications) the *signal-to-noise ratio*. Finally, the term percent error refers to the quantity $100 f_x$.

When the probability function is known and the respective mean and variance (or standard deviation) exist, they can be computed directly from their definitions. However, when the actual distribution is not known, it is considerably more difficult to interpret the knowledge of the mean and standard deviation in terms of confidence that they are representative of the distribution of measurements. The difficulty can be illustrated by considering a *confidence indicator associated with the probability function p*, $C_p\,(k\sigma)$, defined by means

of the integral

$$C_p(k\sigma) \equiv \int_{m_o-k\sigma}^{m_o+k\sigma} p(x)\, dx\,, \tag{1.50}$$

where σ is the standard deviation and $k \geq 1$ is an integer. Since the probability density integrated over the entire underlying domain is normalized to unity, it follows that $C_p(k\sigma) < 1$ for all k. However, $C_p(k\sigma) \approx 1$ whenever $k >> 1$. Thus, $C_p(k\sigma)$ can vary substantially in magnitude for different types of probability functions p, even for fixed values of σ and k. This result indicates that although the variance or standard deviation are useful parameters for measuring dispersion (error), knowledge of them alone does not provide an unambiguous measure of confidence in a result, unless the probability family to which the distribution in question belongs is *a priori* known. Consequently, when an experiment involves several observational processes, each governed by a distinct law of probability, it is difficult to interpret overall errors (which consist of several components) in terms of confidence. In practice, though, the consequences are mitigated by a very important theorem of statistics, called the *Central Limit Theorem*, which will be discussed in the sequel, following Eq. (1.58).

For multivariate probability, the second-order central moments comprise not only the variances $\mu_{ii} = \mathrm{var}(x_i) = E\left[(x_i - m_{oi})^2\right]$, $(i = 1, \ldots, n)$, but also the moments $\mu_{ij} = E\left[(x_i - m_{oi})(x_j - m_{oj})\right]$, $(i, j = 1, \ldots, n)$. These moments are called *covariances*, and the notation $\mathrm{cov}(x_i, x_j) \equiv \mu_{ij}$ is often used. The collection of all second-order moments of a multivariate probability function involving n random variables forms an $n \times n$ matrix, denoted in this section as $\mathbf{V}_x$, and called the *variance-covariance matrix*, or, simply, the *covariance matrix*. Since $\mu_{ij} = \mu_{ji}$ for all i and j, covariance matrices are symmetric.

When $\mu_{ii} > 0$, $(i = 1, \ldots, n)$, it is often convenient to use the quantities ρ_{ij} defined by the relationship

$$\rho_{ij} \equiv \mu_{ij}/(\mu_{ii}\,\mu_{jj})^{1/2}\,, \quad (i, j = 1, \ldots, n)\,, \tag{1.51}$$

and called *correlation parameters* or, simply, *correlations*. The matrix ob-

tained by using the correlations, ρ_{ij}, is called the *correlation matrix*, and will be denoted as $\mathbf{C}_x$ in this section.

Using the *Cauchy-Schwartz inequality*, it can be shown that the elements of $\mathbf{V}_x$ always satisfy the relationship

$$|\mu_{ij}| \leq (\mu_{ii}\mu_{jj})^{1/2} \,, \quad (i,j = 1,n) \,, \tag{1.52}$$

while the elements ρ_{ij} of the correlation matrix $\mathbf{C}_x$ satisfy the relationship

$$-1 \leq \rho_{ij} \leq 1 \,. \tag{1.53}$$

In the context of covariance matrices, the *Cauchy-Schwartz inequality* provides an *indicator of data consistency* that is very useful to verify practical procedures for processing experimental information. Occasionally, practical procedures may generate covariance matrices with negative eigenvalues (thus violating the condition of positive-definitiveness), or with coefficients that would violate the Cauchy-Schwartz inequality; such matrices would, of course, be unsuitable for representing physical uncertainty. Although the mathematical definition of the variance only indicates that it must be nonnegative, the variance for physical quantities should in practice be positive, because it provides a mathematical basis for the representation of physical uncertainty. Since zero variance means no error, probability functions for which some of the random variables have zero variance are not realistic choices for the representation of physical phenomena, since such probability functions would indicate that some parameters were without error, which is never the case in practice. Furthermore, $\mu_{ii} < 0$ would imply an imaginary standard deviation (since $\sigma_i = \mu_{ii}^{1/2}$), which is clearly unacceptable. The reason for mentioning these points here is because, in practice, the elements of covariance matrices are very seldom obtained from direct evaluation of expectations, but are obtained by a variety of other methods, many of them ad hoc. Practical considerations also lead to the requirement that $|\rho_{ij}| < 1$, for $i \neq j$, but a presentation of the arguments underlying this requirement is beyond the purpose of this book. These and other constraints on covariance and correlation matrices lead to the conclusion *that matrices which properly represent physical uncertainties are positive definite.*

Since covariance matrices are symmetric, the $n \times n$ covariance matrix contains no more than $n+[n\,(n-1)/2\,]$ distinct elements, namely the off-diagonal covariances and the n variances along the diagonal. Often, therefore, only the diagonal and upper or lower triangular part of covariance and correlation matrices are listed in the literature. A formula often used in practical computations of covariances is obtained by rewriting the respective definition in the form

$$\operatorname{cov}(x_i, x_j) = E(x_i x_j) - m_{oi} m_{oj}\,. \tag{1.54}$$

Note that if any two random variables, x_i and x_j, in a collection of n random variables are independent, then $\operatorname{cov}(x_i, x_j) = 0$. Note also that the converse of this statement is false: $\operatorname{cov}(x_i, x_j) = 0$ *does not necessarily imply that* x_i *and* x_j *are independent.*

A very useful tool for practical applications is the so-called *scaling and translation theorem*, which states that if x_i and x_j are any two members of a collection of n random variables, then the following relations hold for the random variables $y_i = a_i x_i + b_i$ and $y_j = a_j x_j + b_j$:

$$E(y_i) = a_i E(x_i) + b_i, \operatorname{var}(y_i) = a_i^2 \operatorname{var}(x_i)\,, \quad (i = 1, \ldots, n)\,;$$
$$\operatorname{cov}(y_i, y_j) = a_i a_j \operatorname{cov}(x_i, x_j)\,, \quad (i, j = 1, \ldots, n,\ i \neq j)\,.$$

The constants a_i and a_j are called *scaling parameters*, while the constants b_i and b_j are called *translation parameters*. The above relationships show that mean values are affected by both scaling and translation, while the variances and covariances are only affected by scaling. In particular, the above relationships can be used to establish the following theorem regarding the relationship between ordinary random variables x_i and their standard random variable counterparts $u_i = (x_i - m_{oi})/\sigma_i$, $(i = 1, \ldots, n)$: *the covariance matrix for the standard random variables* $u_i = (x_i - m_{oi})/\sigma_i$ *is the same as the correlation matrix for the random variables* x_i.

The *determinant,* $\det(\mathbf{V}_x)$, of the variance matrix is often referred to as the *generalized variance*, since it degenerates to a simple variance for univariate distributions. The probability distribution is called *nondegenerate* when $\det(\mathbf{V}_x) \neq 0$; when $\det(\mathbf{V}_x) = 0$, however, the distribution is called *degenerate.* Degeneracy is an indication that the information content of the set $\mathbf{x}$ of

random variable is less than rank n, or that the probability function is confined to a hyperspace of dimension lower than n. Of course, the determinant of a covariance matrix vanishes if and only if there exist (one or more) linear relationships among the random variables of the set $\mathbf{x}$.

Due to the above mentioned properties of the positive definite matrix $\mathbf{V}_x$, it also follows that

$$\det(\mathbf{V}_x) \leq \prod_{i=1,n} \mathrm{var}(x_i) = \prod_{i=1,n} \sigma_i^2. \tag{1.55}$$

The equality in the above relation is reached only when $\mathbf{V}_x$ is diagonal, i.e., when $\mathrm{cov}(x_i, x_j) = 0,\ (i, j = 1, n,\ i \neq j)$; in this case, $\det(\mathbf{V}_x)$ attains its maximum value, equal to the product of the respective variances. The determinant $\det(\mathbf{V}_x)$ is related to the determinant of the correlation matrix, $\det(\mathbf{C}_x)$, by the relationship

$$\det(\mathbf{V}_x) = \det(\mathbf{C}_x) \prod_{i=1,n} \sigma_i^2 . \tag{1.56}$$

From Eqs. (1.55) and (1.56), it follows that $\det(\mathbf{C}_x) \leq 1$. It further follows that $\det(\mathbf{C}_x)$ attains its maximum value of unity only when $\mathrm{cov}(x_i, x_j) = 0$, $(i, j = 1, n,\ i \neq j)$. In practice, $\det(\mathbf{C}_x)$ is used as a *measure of degeneracy* of the multivariate probability function. In particular, the quantity $[\det(\mathbf{C}_x)]^{1/2}$ is called the *scatter coefficient for the probability function*. Note that $\det(\mathbf{C}_x) = 0$ when $\rho_{ij} = \rho_{ji} = 1$ for at least one pair (x_i, x_j), with $i \neq j$.

Two random variables, x_i and x_j, with $i \neq j$, are called *fully correlated* if $\mathrm{cor}(x_i, x_j) = 1$; this situation arises if and only if the corresponding standard random variables u_i and u_j are identical, i.e., $u_i = u_j$. On the other hand, if $\mathrm{cor}(x_i, x_j) = -1$, then x_i and x_j are fully *anti-correlated*, which can happen if and only if $u_i = -u_j$. Therefore, the statistical properties of fully correlated or fully anti-correlated random variables are identical, so that only one of them needs to be considered, a fact reflected by the practice of referring to such random variables as being redundant.

In addition to covariance matrices, $\mathbf{V}_x$, and their corresponding correlation matrices, $\mathbf{C}_x$, a third matrix, called the *relative covariance matrix or fractional error matrix*, can also be defined when the elements of the covari-

ance matrix satisfy the condition that $m_{oi} \neq 0$, $(i = 1, \ldots, n)$. This matrix is usually denoted as $\mathbf{R}_x$, and its elements $(\mathbf{R}_x)_{ij} = \eta_{ij}$ are defined as

$$\eta_{ij} = \mu_{ij}/(m_{oi}m_{oj}) \quad . \tag{1.57}$$

Moments of first- and second-order (i.e., means and covariance matrices) provide information only regarding the location and dispersion of probability distributions. Additional information on the nature of probability distributions is carried by the higher-order moments, although moments beyond fourth-order are seldom examined in practice. The nature of such information can be intuitively understood by considering the third- and fourth-order moments of univariate probability functions. For this purpose, it is easiest to consider the *respective reduced central moments*, α_k, defined in terms of central moments and the standard deviation by the relationship $\alpha_k \equiv \mu_k/\sigma^k$.

The reduced central moment α_3 is called the *skewness* of the probability distribution, because it measures quantitatively the departure of the probability distribution from symmetry (a symmetric distribution is characterized by the value $\alpha_3 = 0$). Thus, if $\alpha_3 < 0$, the distribution is skewed toward the left (i.e., it favors lower values of x relative to the mean), while $\alpha_3 > 0$ indicates a distribution skewed toward the right (i.e., it favors higher values of x relative to the mean). The reduced central moment α_4 measures the degree of sharpness in the peaking of a probability distribution and it is called *kurtosis*. Kurtosis is always nonnegative. The standard for comparison of kurtosis is the normal distribution (Gaussian) for which $\alpha_4 = 3$. Distributions with $\alpha_4 < 3$ are called *platykurtic distributions*. Those with $\alpha_4 = 3$ are called *mesokurtic distributions*. Finally, distributions with $\alpha_4 > 3$ are called *leptokurtic distributions*.

Very often in practice, the details of the distribution are unknown, and only the mean and standard deviations can be estimated from the limited amount of information available. Even under such circumstances, it is still possible to make statements regarding confidence by relying on *Chebyshev's theorem*, which can be stated as follows: consider that m_o and $\sigma > 0$ denote the mean value and standard deviation, respectively, of an otherwise unknown multivariate probability density p involving the random variable x. Furthermore, consider that P represents cumulative probability, C_p represents confidence,

and $k \geq 1$ is a real constant (not necessarily an integer). Then, *Chebyshev's theorem states that the following relationship holds*:

$$C_p(k\sigma) = P(|x - m_o| \leq k\sigma) \geq 1 - \left(1/k^2\right) . \tag{1.58}$$

Chebyshev's theorem is a weak law of statistics since it *provides an upper bound on the probability of a particular deviation* ε. The actual probability of such a deviation (if the probability function were known in detail so that it could be precisely calculated) would always be smaller (implying greater confidence) than Chebyshev's limit. This important point is illustrated in Table (1.1), which compares probabilities for observing particular deviations (denoted as ε) from the respective means. This table clearly underscores the fact that normally distributed random variables are much more sharply localized with respect to the mean than indicated by Chebyshev's theorem.

TABLE 1.1
Probability of Occurrence

Deviation ε	**Normal distribution**	**Chebyshev's limit**
$> 1\sigma$	< 0.3173	< 1.0
$> 2\sigma$	< 0.0455	< 0.25
$> 3\sigma$	< 0.00270	< 0.1111
$> 4\sigma$	< 0.0000634	< 0.0625
$> 5\sigma$	$< 5.73 * 10 - 7$	< 0.04
$> 6\sigma$	$< 2.0 * 10 - 9$	< 0.02778

Central Limit Theorem: Consider that $(x_1, x_2, \ldots, x_n)$ is a random sample of the parent random variable x, with mean value m_o, var $(x) = \sigma^2$, and sample average $\xi_n = (\sum_{i=1}^{n} x_i)/n$. Furthermore, define $z_n \equiv (\xi_n - m_o)/\left(\sigma/n^{1/2}\right)$ to be the reduced random-variable equivalent of ξ_n. Then, the central limit theorem states that z_n, ξ_n, and $n\xi_n$ are all asymptotically normal in the limit as $n \to \infty$. The least restrictive necessary and sufficient condition for the validity of the central limit theorem is the *Lindeberg condition*, which states that if the sequence of random variables $(x_1, x_2, \ldots, x_n)$ is uniformly bounded (i.e., if there exists a positive real constant C such that $|x_i| < C$ for each x_i and all possible n) and the sequence is not degenerate, then the central limit theorem holds. *In practice, the Lindeberg condition requires that the mean values and variances exist for each of these variables, and that the overall variance in the sum* ξ_n *of these random variables be not dominated by just a*

few of the components. Application of the central limit theorem to correlated random variables is still an open field of research in mathematical statistics.

Rather than specify the conditions under which the central limit theorem holds exactly in the limit $n \to \infty$, *in practice it is more important to know the extent to which the Gaussian approximation is valid for finite* n. This is difficult to quantify exactly, but the rule of thumb is that the central limit theorem holds as long as the sum is made of a large number of small contributions. Discrepancies arise if, for example, the distributions of the individual terms have long tails, so that occasional large values make up a large part of the sum. Such contributions lead to "non-Gaussian" tails in the sum, which can significantly alter the probability to find values with large departures from the mean. In such cases, the main assumption underlying the central limit theorem, namely the assumption that the measured value of a quantity is a normally distributed variable centered about the mean value, breaks down. Since this assumption is often used when constructing a confidence interval, such intervals can be significantly underestimated if non-Gaussian tails are present. In particular, the relationship between the confidence level and the size of the interval will differ from the Gaussian prescription (i.e., 68.3% for a "1σ" interval, 95.4% for "2 σ," etc.). A better understanding of the non-Gaussian tails can often be obtained from a detailed Monte Carlo simulation of the individual variables making up the sum.

For example, the central limit theorem cannot be used for calculating the angle by which a charged particle is deflected upon traversing a layer of matter. Although the total angle can be regarded as the sum of a small number of deflections caused by multiple Coulomb scattering collisions with nuclei in the substance being traversed, and although there are many such collisions, the total angle cannot be calculated by using the central limit theorem. This is because the distribution for individual deflections has a long tail extending to large angles, which invalidates the main assumption underlying the central limit theorem.

1.3.2 A Geometric Model for Covariance Matrices

For a set of n variates, $x_1, x_2, \ldots, x_n$, the square symmetrical matrix $\mathbf{C}$ of order $n \times n$, with elements

$$\mathbf{C} \triangleq (C_{ij})_{n\times n} \triangleq \operatorname{cov}(x_i, x_j); \quad C_{ii} \triangleq \operatorname{var}(x_i), \tag{1.59}$$

is, by definition, the *variance-covariance* (or, in short, *covariance*) matrix associated with the variates $x_1, x_2, \ldots, x_n$. The role of the covariances C_{ij} can be further illustrated by considering a joint distribution of n variates, $x_1, x_2, \ldots, x_n$, normalized to unity, of the form

$$f(\mathbf{x}) = A \exp[-1/2\ \mathbf{Q}\ (\mathbf{x})], \quad \mathbf{x} \triangleq (x_1, x_2, \ldots, x_n), \tag{1.60}$$

where $\mathbf{Q}(\mathbf{x})$ is a positive quadratic form expressed as

$$\mathbf{Q}(\mathbf{x}) = (\mathbf{x} - \mathbf{c})^T \mathbf{B} (\mathbf{x} - \mathbf{c}), \tag{1.61}$$

where $\mathbf{c}$ is a n-dimensional vector of constants, $\mathbf{B}$ is a $n \times n$ (positive) symmetric matrix with constant coefficients, the "T" denotes "transposition," and the normalization coefficient A is defined as

$$A = \sqrt{|\mathbf{B}|/(2\pi)^n}. \tag{1.62}$$

By differentiating the normalization relation $\int f(\mathbf{x})\, d^n x = 1$ with respect to $\mathbf{c}$ it follows that

$$0 = \frac{\partial}{\partial \mathbf{c}} \int f(\mathbf{x})\, d^n x = -\frac{1}{2} \int \frac{\partial \mathbf{Q}(\mathbf{x})}{\partial \mathbf{c}} f(\mathbf{x})\, d^n x = \mathbf{B} \langle \mathbf{x} - \mathbf{c} \rangle,$$

where the angular brackets $\langle , \rangle$ denote integration of the respective quantitiy over the distribution $f(x)$, i.e., $\langle g(\mathbf{x}) \rangle \equiv \int g(\mathbf{x}) f(\mathbf{x}) d^n x$.

Since $|\mathbf{B}| > 0$, it follows that $\mathbf{c} = \langle \mathbf{x} \rangle$, i.e., the components

$$c_i = \mu_{x_i}, \qquad i = 1, 2, \ldots, n, \tag{1.63}$$

of the vector $\mathbf{c}$ are the means of the variates $x_i,\ i = 1, \ldots, n$. Differentiating the normalization relation $\int f(\mathbf{x})\, d^n x = 1$ with respect to the elements of $\mathbf{B}$

yields

$$0 = \frac{\partial}{\partial B_{ij}} \int f(\mathbf{x})\, d^n x = \frac{1}{2} \left(\frac{1}{|\mathbf{B}|} \frac{\partial |\mathbf{B}|}{\partial B_{ij}} - \int \frac{\partial \mathbf{Q}(\mathbf{x})}{\partial B_{ij}} f(\mathbf{x})\, d^n x \right).$$

Since the derivative of the determinant of a matrix with respect to an element of this matrix is the cofactor of the respective element, the above equation implies that

$$\frac{cofac(B_{ij})}{|\mathbf{B}|} = \int (x_i - \mu_i)(x_j - \mu_j) f(\mathbf{x})\, d^n x. \tag{1.64}$$

Furthermore, since $\mathbf{B}$ is symmetric, the left side of the above equation is just the element (i, j) of the inverse of $\mathbf{B}$ or, in matrix notation

$$\mathbf{B}^{-1} = \left\langle (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \right\rangle \triangleq \mathbf{C}. \tag{1.65}$$

The right side of Eq. (1.65) is the average of the outer product of the vector $(\mathbf{x} - \mu)$ with itself, which is by definition the covariance matrix $\mathbf{C}$ associated with the variates $\mathbf{x}$. Summarizing the information contained in Eqs. (1.60) through Eq. (1.65) indicates that n variates, $x_1, x_2, \ldots, x_n$, characterized by the vector of mean-values μ and the covariance matrix $\mathbf{C}$, are effectively distributed according to the multivariate normal distribution

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{C}|}} \exp\left[-1/2\, (\mathbf{x} - \mu)^T \mathbf{C}^{-1} (\mathbf{x} - \mu) \right] \tag{1.66}$$

Thus, the variance σ^2 of a single variate generalizes to the uncertainty (variance) matrix $\mathbf{C}$ containing the variances of the individual variates along its diagonal elements, and the covariances of every pair of variates as its off diagonal elements.

Since variances (or standard deviations) provide a measure for the dispersion of the probability distribution with respect to its mean, they can be visualized as providing a measure of the region of (n-dimensional) random-variable space where most of the probability is concentrated. The intuitive understanding of the meaning of correlation is less straightforward. Perhaps the simplest way to appreciate intuitively the meaning of correlation is to consider a bivariate distribution for the random variables x_1 and x_2. For sim-

plicity, suppose that $-\infty \leq x_1 \leq \infty$ and $-\infty \leq x_2 \leq \infty$, and that p is a probability density for x_1 and x_2. Considering, without loss of generality, that x_1 and x_2 correspond to orthogonal coordinates of a two-dimensional Cartesian plane. Then, the surface $x_3 = p(x_1, x_2)$ will appear as a "hill" in the third (Cartesian) coordinate x_3; note that x_3 is not a random variable. The surface $x_3 = p(x_1, x_2)$ is "centered," in the (x_1, x_2)-plane, on the point (m_{o1}, m_{o2}), while the lateral extents of this surface are measured by the standard deviations σ_1 and σ_2; furthermore, the surface $x_3 = p(x_1, x_2)$ would be symmetric when $\sigma_1 = \sigma_2$. Since p is normalized and since $p(x_1, x_2) \geq 0$ for all (x_1, x_2), the surface $p(x_1, x_2)$ will have at least one maximum in the direction x_3, say $(x_3)_{\max}$. Slicing through the surface $x_3 = p(x_1, x_2)$ with a horizontal plane $x_3 = h$, where $0 < h < (x_3)_{\max}$, and projecting the resulting planar figure onto the (x_1, x_2)-plane yields elliptical shapes that can be inscribed in the rectangle $(x_1 = 2l\sigma_1,\ x_2 = 2l\sigma_2)$, with $0 < l < \infty$. The covariance $\operatorname{cov}(x_i, x_j)$, or equivalently, *the correlation parameter* $\rho_{12} = \rho$ *indicates the orientation of the probability distribution* in the random variable space (x_1, x_2). This is particularly clear when $\sigma_1 \neq \sigma_2$, in which case the surface $x_3 = p(x_1, x_2)$ is not symmetric. Thus, a nonzero correlation implies that the surface $x_3 = p(x_1, x_2)$ is *tilted* relative to the x_1 and x_2 axes. On the other hand, a zero correlation implies that the surface $x_3 = p(x_1, x_2)$ is somehow aligned with respect to these axes. Since the intrinsic shape of the surface $x_3 = p(x_1, x_2)$ is governed by the underlying probability but not by the choice of variables, it is possible to perform an orthogonal transformation such as to align the surface relative to a new, transformed coordinate system. The transformed random variables generated by an orthogonal transformation are called *orthogonal random variables*, and they are independent from one another. Thus, consider a region Ω of the three-dimensional Euclidean space. Points in this space can be described by Cartesian coordinates $(x_1,\ x_2,\ x_3)$ or, equivalently, by the vector representation

$$\mathbf{r} = \sum_{i=1}^{3} x_i \mathbf{u}_i, \tag{1.67}$$

where the $\mathbf{u}_i$ are fixed, orthogonal, unit basis vectors obeying the relations

$$\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}\,, \quad where \;\; \delta_{ii} = 1, \quad \delta_{ij} = 0, \quad i \neq j. \tag{1.68}$$

The difference vector between two points, $\mathbf{r}$ and $\mathbf{r}'$, on a trajectory in Ω is

$$d\mathbf{r} = \mathbf{r}' - \mathbf{r} = \sum_{i=1}^{3} dx_i \mathbf{u}_i, \tag{1.69}$$

since the basis vectors are fixed in the orthogonal coordinate system representation. The *distance, dL*, between the points $\mathbf{r}$ and $\mathbf{r}'$ is given by

$$(dL)^2 = d\mathbf{r}^T d\mathbf{r} = \sum_{i=1}^{3}\sum_{j=1}^{3} dx_i \left(\mathbf{u}_i^T \mathbf{u}_j\right) dx_j = \sum_{i=1}^{3} (dx_i)^2. \tag{1.70}$$

Now consider *curvilinear coordinates* (y_1, y_2, y_3) as an alternative coordinate system representation for the region Ω. A typical example of an alternative coordinate system would be spherical coordinates. The curvilinear coordinates are related to the Cartesian coordinates by the transformation equations $y_i = y_i(\mathbf{x})$, where each function y_i is continuous and differentiable with respect to each element of $\mathbf{x}$, for $i = 1, 2, 3$. In terms of the curvilinear coordinates (y_1, y_2, y_3), the difference vector $d\mathbf{r}$ becomes

$$d\mathbf{r} = \sum_{i=1}^{3} dy_i \mathbf{g}_i, \quad \mathbf{g}_i \triangleq \partial \mathbf{r}/\partial y_i \;\;, \quad i = 1, 2, 3. \tag{1.71}$$

The vectors $\mathbf{g}_i$ are *effective basis vectors* for the alternative curvilinear coordinate system; they are neither orthogonal nor fixed unit vectors (as the vectors $\mathbf{u}_i$ in the Cartesian system). The distance dL is a scalar and is therefore an invariant quantity, independent of the coordinate representation. In terms of these alternative coordinates, the expression of dL becomes

$$(dL)^2 = d\mathbf{r}^T d\mathbf{r} = \sum_{i=1}^{3}\sum_{j=1}^{3} dy_i \left(\mathbf{g}_i^T \mathbf{g}_j\right) dy_j. \tag{1.72}$$

Equations (1.70) and (1.72) are examples of *differential quadratic forms*,

which can be generally represented in the unified form

$$(dL)^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} dq_i G_{ij} dq_j = d\mathbf{q}^T \mathbf{G} d\mathbf{q}, \tag{1.73}$$

where $\mathbf{G}$ is a symmetric matrix and q_i represents either x_i or y_i. Although $\mathbf{G}$ is diagonal for Cartesian coordinates (since $G_{ij} = \delta_{ij}$), it is not diagonal for general curvilinear coordinates. The coefficients G_{ij} are called *metric coefficients*, and the matrix $\mathbf{G}$ is usually referred to simply as the *metric matrix* for the coordinate representation used to describe points in the space Ω. The metric matrix enables the computation of the invariant quantity "distance" in terms of a particular coordinate representation for the respective space.

Since the vectors $\mathbf{g}_i$ need not be unit vectors, they can be expressed in terms of a set of unit vectors $\mathbf{v}_i$, parallel to the corresponding $\mathbf{g}_i$ in the form

$$\mathbf{g}_i = s_i \mathbf{v}_i\,, \quad i = 1, 2, 3. \tag{1.74}$$

Inserting Eq. (1.74) into Eq. (1.73) recasts the latter into the form

$$(dL)^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} (s_i dq_i)\, C_{ij}\, (s_j dq_j) = (\mathbf{S} d\mathbf{q})^T \mathbf{C}\, (\mathbf{S} d\mathbf{q})\,, \tag{1.75}$$

where $\mathbf{S}$ is a diagonal matrix (with elements $s_i \delta_{ij}$), called the *dispersion matrix*, and $\mathbf{C}$ is a matrix with elements C_{ij} satisfying the conditions

$$C_{ij} \triangleq \mathbf{v}_i^T \mathbf{v}_j; \quad -1 \le C_{ij} \le 1, \quad i,\, j = 1, 2, 3. \tag{1.76}$$

Equation (1.75) indicates that $\mathbf{C}$ plays the role of a correlation matrix. Since the vectors $\mathbf{v}_i$ are unit vectors, it follows that

$$\mathbf{v}_i^T \mathbf{v}_j = \cos\theta_{ij}, \tag{1.77}$$

where θ_{ij} is the angle between the vectors. Hence, the partial covariances C_{ij} and the angles $\cos\theta_{ij}$ are equivalent, which provides a geometric interpretation of correlation; full correlation (anti-correlation) can be interpreted in terms of parallel (anti-parallel) vectors, while the lack of correlation can be visualized in

terms of perpendicular vectors. This geometrical analogy can be summarized as follows:

1. The parameters of a physical problem correspond to the coordinates in a vector space Ω;
2. The space Ω represents the range of possibilities for these parameters;
3. Correlations arise because of the basic nonorthogonality of the set of parameters used to represent the physical problem;
4. Points along a curve in Ω correspond to various experimental outcomes for different values of the parameters (assuming, for simplicity, that the outcome of the experiment is a scalar quantity).
5. The "distance" between two nearby points, which differ only due to small positive or negative increments of the coordinates (deviations in the parameters) corresponds to the uncertainty in the experimental results; this uncertainty is computed using a differential quadratic form, with knowledge of the dispersion matrix and the correlation matrices.

The geometric model of error presented in the foregoing can be extended to include a wider range of possibilities, by considering that the set $\mathbf{x} = (x_1, \ldots, x_n)$ of random variables represents fundamental parameters of a physical problem, and that the components of the vector $\mathbf{y} = (y_1, \ldots, y_m)$ represent derived quantities, i.e., $y_k = y_k(\mathbf{x})$, for $k = 1, \ldots, m$. For example, in a typical nuclear experiment, the components of $\mathbf{x}$ might represent measured quantities such as reaction yields, sample masses, neutron fluences, neutron energies, etc., while the quantities $\mathbf{y}$ could represent computed reaction cross sections. Suppose that we know the covariance matrix, $\mathbf{C}_x$, of the parameters $\mathbf{x}$ and need to determine the covariance matrix, $\mathbf{C}_y$, of the derived quantities $\mathbf{y}$. For this purpose, we construct vectors $\mathbf{e}_{xi}$ in the error space corresponding to the variable x_i, such that

$$\mathbf{e}_{x\,i}^T \mathbf{e}_{xj} = C_{xij}, \quad i, j = 1, \ldots, n\,. \tag{1.78}$$

In particular, the amplitude $|\mathbf{e}_{xi}|$ equals the standard deviation, $\sigma_{x\,i}$, of x_i,

so that

$$\mathbf{e}_{xi}^T \mathbf{e}_{xj} = \sigma_{xi}\sigma_{xj}\cos\alpha_{ij}\,, \quad i,\, j = 1, \ldots,\, n\,. \tag{1.79}$$

The parameter α_{ij} is the angle between $\mathbf{e}_{xi}$ and $\mathbf{e}_{xj}$ and, as was discussed in the foregoing, there is a one-to-one correspondence between the error correlation factor C_{xij} and $\cos\alpha_{ij}$.

The error in y_k can be represented as a vector by assuming that errors add like vectors, according to the law of linear superposition, i.e.,

$$\mathbf{e}_{yk} = \sum_{i=1}^{n} s_{ik}\mathbf{e}_{xi}, \quad k = 1, \ldots, m, \tag{1.80}$$

where the quantities s_{ik} remain to be determined by requiring that the treatment of the uncertainties of $\mathbf{y}$ should be consistent with the treatment for the uncertainties of $\mathbf{x}$. This requirement implies that

$$\mathbf{e}_{yk}^T \mathbf{e}_{yq} = \sigma_{yk}\sigma_{yq}\cos\beta_{kq} = C_{ykq}, \quad k,\, q = 1, \ldots, m\,, \tag{1.81}$$

so that the standard deviation σ_{yk} equals $|\mathbf{e}_{yk}|$, β_{kq} represents the angle between $\mathbf{e}_{yk}$ and $\mathbf{e}_{yq}$, while $\cos\beta_{kq}$ is equivalent to the correlation factor C_{ykq}. Comparing Eqs. (1.78) through (1.81) indicates that Eq. (1.80) is valid if

$$s_{ik} = \left(\partial y_k/\partial x_i\,\right)_{\mu}\,, \quad \mu = \left(\langle x_1\rangle\,,\, \ldots,\, \langle x_n\rangle\right), \quad i = 1,\, \ldots, n; \quad k = 1,\, \ldots, m \tag{1.82}$$

The quantity $s_{i\,k}$ is customarily called the *sensitivity* of (the derived quantity) y_k to the parameter x_i. In terms of the covariance matrix, $\mathbf{C}_x$, of the parameters $\mathbf{x}$ the explicit expression of the covariance matrix, $\mathbf{C}_y$, associated with the derived quantities $\mathbf{y}$ can be explicitly written in matrix form as

$$\mathbf{C}_y = \mathbf{S}\mathbf{C}_x\mathbf{S}^T, \quad \mathbf{S} \triangleq (s_{ik})_{n\times m}. \tag{1.83}$$

The above expression is colloquially known as the "*sandwich rule*" for "propagation of input parameter uncertainties" to compute the uncertainties in "responses" (i.e., derived quantities). The above "sandwich rule" holds rigorously only if the components of the derived quantity $\mathbf{y}$ are linear functions

of the components of the parameters $\mathbf{x}$, i.e.,

$$y_k(\mu + \delta\mathbf{x}) = y_k(\mu) + (\partial y_k/\partial x_i)_{\mu}\delta\mathbf{x}. \tag{1.84}$$

The "geometrical model of errors" provides a useful mechanism for employing graphical analysis in the examination of errors, since error correlations can be visualized in terms of the cosines of angles between the various error vectors. The sandwich rule given in Eq. (1.83) is often used in practice, particularly for large-scale models involving many parameters, since information beyond first-order terms is often unavailable or impractical to use, which imposes (de facto) the assumption that responses are linearly dependent on parameters.

When information beyond first order is available, the respective response is expanded in a multivariate Taylor series, retaining derivatives beyond the first order sensitivities, by means of the *method of propagation of errors or propagation of moments* (see, e.g., ref. [29]), which can be used when the response is either the result of an indirect measurement or the result of a computation. The response, denoted by R, is considered to be a real-valued function of k system parameters, denoted as $(\alpha_1, \ldots, \alpha_k)$, with mean values $\left(\alpha_1^0, \ldots, \alpha_k^0\right)$, i.e.,

$$R = R(\alpha_1, \ldots, \alpha_k) = R\left(\alpha_1^0 + \delta\alpha_1, \ldots, \alpha_k^0 + \delta\alpha_k\right). \tag{1.85}$$

Expanding $R\left(\alpha_1^0 + \delta\alpha_1, \ldots, \alpha_k^0 + \delta\alpha_k\right)$ in a Taylor series around the nominal values $\alpha^0 = \left(\alpha_1^0, \ldots, \alpha_k^0\right)$ and retaining the terms up to the n^{th} order in

the variations $\delta\alpha_i \equiv \left(\alpha_i - \alpha_i^0\right)$ around α_i^0 gives:

$$\begin{aligned} R\left(\alpha_1,\ldots,\alpha_k\right) &\equiv R\left(\alpha_1^0+\delta\alpha_1,\ldots,\alpha_k^0+\delta\alpha_k\right) \\ &= R\left(\alpha^0\right) + \sum_{i_1=1}^{k}\left(\frac{\partial R}{\partial\alpha_{i_1}}\right)_{\alpha^0}\delta\alpha_{i_1} \\ &+ \frac{1}{2}\sum_{i_1,i_2=1}^{k}\left(\frac{\partial^2 R}{\partial\alpha_{i_1}\partial\alpha_{i_2}}\right)_{\alpha^0}\delta\alpha_{i_1}\delta\alpha_{i_2} \\ &+ \frac{1}{3!}\sum_{i_1,i_2,i_3=1}^{k}\left(\frac{\partial^3 R}{\partial\alpha_{i_1}\partial\alpha_{i_2}\partial\alpha_{i_3}}\right)_{\alpha^0}\delta\alpha_{i_1}\delta\alpha_{i_2}\delta\alpha_{i_3}+\cdots \\ &+ \frac{1}{n!}\sum_{i_1,i_2,\ldots,i_n=1}^{k}\left(\frac{\partial^n R}{\partial\alpha_{i_1}\partial\alpha_{i_2}\ldots\partial\alpha_{i_n}}\right)_{\alpha^0}\delta\alpha_{i_1}\ldots\delta\alpha_{i_n}\,. \end{aligned} \tag{1.86}$$

Recalling the definition of the moments for multivariate probability, cf. Eq. (1.46), the above Taylor-series expansion can now be used to determine the mean (expectation), variance, skewness, and kurtosis, of $R\left(\alpha_1,\ldots,\alpha_k\right)$ by formal integrations over the (unknown) joint probability density function $p\left(\alpha_1,\ldots,\alpha_k\right)$ of the parameters $\left(\alpha_1,\ldots,\alpha_k\right)$. For *uncorrelated parameters* $\left(\alpha_1,\ldots,\alpha_k\right)$, the expressions of the various moments of the response $R\left(\alpha_1,\ldots,\alpha_k\right)$ are as follows:

$$\begin{aligned} E\left(R\right) &= R\left(\alpha_1^0,\ldots,\alpha_k^0\right) + \frac{1}{2}\sum_{i=1}^{k}\left\{\frac{\partial^2 R}{\partial\alpha_i^2}\right\}_{\alpha^o}\mu_2\left(\alpha_i\right) \\ &+ \frac{1}{6}\sum_{i=1}^{k}\left\{\frac{\partial^3 R}{\partial\alpha_i^3}\right\}_{\alpha^o}\mu_3\left(\alpha_i\right) + \frac{1}{24}\sum_{i=1}^{k}\left\{\frac{\partial^4 R}{\partial\alpha_i^4}\right\}_{\alpha^o}\mu_4\left(\alpha_i\right) \\ &+ \frac{1}{24}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\left\{\frac{\partial^4 R}{\partial\alpha_i^2\partial\alpha_j^2}\right\}_{\alpha^o}\mu_2\left(\alpha_i\right)\mu_2\left(\alpha_j\right)\,; \end{aligned} \tag{1.87}$$

$$\mu_2(R) = \sum_{i=1}^{k} \left\{ \left(\frac{\partial R}{\partial \alpha_i} \right)^2 \right\}_{\alpha^o} \mu_2(\alpha_i) + \sum_{i=1}^{k} \left\{ \frac{\partial R}{\partial \alpha_i} \frac{\partial^2 R}{\partial \alpha_i^2} \right\}_{\alpha^o} \mu_3(\alpha_i)$$
$$+ \frac{1}{3} \sum_{i=1}^{k} \left\{ \frac{\partial R}{\partial \alpha_i} \frac{\partial^3 R}{\partial \alpha_i^3} \right\}_{\alpha^o} \mu_4(\alpha_i)$$
$$+ \frac{1}{4} \sum_{i=1}^{k} \left\{ \left(\frac{\partial^2 R}{\partial \alpha_i^2} \right)^2 \right\}_{\alpha^o} \left[\mu_4(\alpha_i) - (\mu_2(\alpha_i))^2 \right] ; \tag{1.88}$$

$$\mu_3(R) = \sum_{i=1}^{k} \left\{ \left(\frac{\partial R}{\partial \alpha_i} \right)^3 \right\}_{\alpha^o} \mu_3(\alpha_i)$$
$$+ \frac{3}{2} \sum_{i=1}^{k} \left\{ \left(\frac{\partial R}{\partial \alpha_i} \right)^2 \frac{\partial^2 R}{\partial \alpha_i^2} \right\}_{\alpha^o} \left[\mu_4(\alpha_i) - (\mu_2(\alpha_i))^2 \right] ; \tag{1.89}$$

$$\mu_4(R) = \sum_{i=1}^{k} \left\{ \left(\frac{\partial R}{\partial \alpha_i} \right)^4 \right\}_{\alpha^o} \left[\mu_4(\alpha_i) - 3(\mu_2(\alpha_i))^2 \right] + 3[\mu_2(R)]^2 . \tag{1.90}$$

In Eqs. (1.87) through (1.90), the quantities $\mu_l(R)$, $(l = 1, \ldots, 4)$, denote the respective central moments of the response $R(\alpha_1, \ldots, \alpha_k)$, while the quantities $\mu_k(\alpha_i)$, $(i = 1, \ldots, k;\ k = 1, \ldots, 4)$ denote the respective central moments of the parameters $(\alpha_1, \ldots, \alpha_k)$. Note that $E(R) \neq R^0$ when the response $R(\alpha_1, \ldots, \alpha_k)$ is a nonlinear function of the parameters $(\alpha_1, \ldots, \alpha_k)$.

It is important to emphasize that the "propagation of moments" equations, i.e., Eqs. (1.87) through (1.90), are used not only for processing experimental data obtained from indirect measurements, but are also used for performing statistical analysis of computational models. In the latter case, the "propagation of errors" equations provide a systematic way of obtaining the uncertainties in computed results, arising not only from uncertainties in the parameters that enter the respective computational model but also from the numerical approximations themselves. Clearly, the propagation of errors method can be used only if the exact response derivatives to parameters are also available, in addition to the respective parameter uncertainties.

1.3.3 Computing Covariances: Simple Examples

Consider the computation of the joint distribution of two variates u and v, defined as $u = x + z$ and $v = y + z$, where x, y, and z are three independent, normally distributed variates with zero means. The variances, and covariance of u and v can be readily obtained as

$$\begin{aligned}\langle \delta u \delta v \rangle &= \langle (\delta x + \delta z)(\delta y + \delta z) \rangle = \ldots = \left\langle (\delta z)^2 \right\rangle, \left\langle (\delta u)^2 \right\rangle \\ &= \left\langle (\delta x)^2 \right\rangle \left\langle (\delta z)^2 \right\rangle, \textit{and} \\ \left\langle (\delta v)^2 \right\rangle &= \left\langle (\delta y)^2 \right\rangle \left\langle (\delta z)^2 \right\rangle .\end{aligned}$$

Hence, the uncertainty matrix associated with u and v is

$$\mathbf{C} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} + \sigma_z^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The determinant of this uncertainty matrix is $|\mathbf{C}| = \sigma_x^2\sigma_y^2 + \sigma_y^2\sigma_z^2 + \sigma_z^2\sigma_x^2$. Hence, the inverse of $\mathbf{C}$ is

$$\mathbf{C}^{-1} = \frac{1}{|\mathbf{C}|} \left[\begin{pmatrix} \sigma_y^2 & 0 \\ 0 & \sigma_x^2 \end{pmatrix} + \sigma_z^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right].$$

It follows that the joint distribution is

$$f(\mathbf{w}) = \frac{1}{2\pi |\mathbf{C}|} \exp \left[-1/2 \, \mathbf{w}^T \mathbf{C}^{-1} \mathbf{w} \right],$$

where $\mathbf{w}$ denotes the column vector (u, v).

As the next example, consider a variate $y(x) = ax + b$ which depends linearly on a parameter x, but the coefficients a and b are unknown and need to be evaluated, together with their associated uncertainties, by using two measurements of y at the values x_1 and x_2 of x, $0 < x_1 < x_2$, such that

$$y(x_1) = y_1 \pm \sigma_1 , \;\; y(x_2) = y_2 \pm \sigma_2 , \quad \operatorname{cov}[y(x_1), y(x_2)] = \sigma_1 \sigma_2 \rho.$$

Using these two measurements, it follows that

$$y(x) = \frac{1}{\Delta x}\left[(y_2 - y_1)\, x + (x_2 y_1 - x_1 y_2)\right], \Delta x \triangleq x_2 - x_1.$$

From the above expression, the uncertainties in the coefficients a and b are computed as

$$\begin{aligned}
\mathrm{var}(a) &= \frac{1}{(\Delta x)^2}\left(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho\right), \\
\mathrm{var}(b) &= \frac{1}{(\Delta x)^2}\left(x_2^2\sigma_1^2 + x_1^2\sigma_2^2 - 2x_1x_2\sigma_1\sigma_2\rho\right), \\
\mathrm{cov}(a,\, b) &= \frac{1}{(\Delta x)^2}\left[(x_1 + x_2)\,\sigma_1\sigma_2\rho - x_2\sigma_1^2 - x_1\sigma_2^2\right].
\end{aligned}$$

The significance of correlations with respect to combinations of correlated data becomes evident: if, for example, the two measurements were fully correlated ($\rho = 1$), then the above expressions would reduce to

$$\sigma_a = \frac{1}{\Delta x}\left|\sigma_1 - \sigma_2\right| \quad , \qquad \sigma_b = \frac{1}{\Delta x}\left|x_2\sigma_1 - x_1\sigma_2\right|$$

and

$$\rho(a,\, b) = -\frac{(\sigma_1 - \sigma_2)\,(x_2\sigma_1 - x_1\sigma_2)}{\left|(\sigma_1 - \sigma_2)\,(x_2\sigma_1 - x_1\sigma_2)\right|}.$$

Since $x_2 > x_1 > 0$, it follows that, if $\rho = 1$, then

$$\rho(a,\, b) = \begin{cases} +1, & \sigma_1 < \sigma_2 < x_2\sigma_1/x_1 \\ -1, & \textit{otherwise}. \end{cases},$$

Following the same procedure, it can be shown that the coefficients a and b are partially anti-correlated when $\rho = 0$; finally, the coefficients a and b are fully anti-correlated if $\rho = -1$.

If the variate $y(x) = ax + b$ is evaluated at two other points, x' and x'', the corresponding variances and covariances are

$$\begin{aligned}
\mathrm{cov}\left[y(x'),\, y(x'')\right] &= x'x''\,\mathrm{var}(a) + \mathrm{var}(b) + (x' + x'')\,\mathrm{cov}(a,\, b) \\
\mathrm{var}\left[y(x)\right] &= x^2\,\mathrm{var}(a) + \mathrm{var}(b) + 2x\ \mathrm{cov}(a,\, b).
\end{aligned}$$

As a function of x, the quantity $\mathrm{var}\,[y\,(x)]$ takes on the minimal value $\mathrm{var}\,(b)\left[1-\rho^2\,(a,b)\right]$ at $x = -\mathrm{cov}\,(a,\,b)/\mathrm{var}\,(a)$. If the coefficients a and b are either fully correlated or fully anti-correlated (i.e., if $|\rho| = 1$), then $\mathrm{var}\,[y\,(x_0)] = 0$ at $x_0 = (x_2\sigma_1 - x_1\sigma_2\rho)/(\sigma_1 - \sigma_2\rho)$.

When evaluating uncertainties associated with products and quotients of variates, it is more convenient to perform the respective computations using relative rather than absolute variances and covariances. Using *capital letters* to denote *relative variances and covariances*, and noting that $\delta a/a$ represents the relative variation of a variate a, it follows that the relative variance is related to the absolute variance through the formula

$$Var\,(a) \triangleq \left\langle \left(\frac{\delta a}{a}\right)^2 \right\rangle = \frac{1}{a^2}\left\langle (\delta a)^2 \right\rangle = \frac{\mathrm{var}\,(a)}{a^2}.$$

Similarly, the relative covariance of two variates a and b is related to their absolute covariance through the formula

$$Cov\,(a,\,b) \triangleq \left\langle \frac{\delta a}{a}\frac{\delta b}{b} \right\rangle = \frac{\mathrm{cov}\,(a,\,b)}{ab}.$$

Hence, the absolute variance of a product $c = a\,b$ of two variates a and b is

$$\mathrm{var}\,(c) = b^2\,\mathrm{var}\,(a) + a^2\,\mathrm{var}\,(b) + 2ab\;\mathrm{cov}\,(a,\,b)\,, \tag{1.91}$$

while the relative variance of c is readily obtained as

$$Var\,(c) = Var\,(a) + Var\,(b) + 2Cov\,(a,\,b)\,. \tag{1.92}$$

Note, as an aside, that the above expression has the same form as the expression of the absolute variance of $c = a+b$. Similarly, for the ratio $c = a/b$ of two variates a and b, the corresponding relative variance is

$$Var\,(c) = Var\,(a) + Var\,(b) - 2Cov\,(a,\,b)\,.$$

Generalizing the above considerations to a variate of the form

$$y = \frac{x_1 x_2 ... x_m}{x_{m+1} ... x_n}$$

leads to the following expression for the relative variance $Var\,(y)$:

$$Var\,(y) = \sum_{i=1}^{n} Var\,(x_i) + 2\left\{\sum_{i=1}^{m}\left[\sum_{j=i+1}^{m} Cov\,(x_i,\, x_j) - \sum_{j=m+1}^{n} Cov\,(x_i,\, x_j)\right] + \sum_{i=m+1}^{n}\sum_{j=i+1}^{n} Cov\,(x_i,\, x_j)\right\}.$$

As another example of evaluating a covariance matrix, consider a typical measurement of a neutron activation cross section, for which the uncertainty analysis is performed using representative numerical data. A typical experimental setup for measuring a neutron activation cross involves the irradiation of a sample, which comprises N nuclei of an isotope that can be activated by neutrons, in a given constant and uniform neutron field, characterized by the neutron flux φ. The activated nuclei decay by emitting gamma-ray photons of a definite energy E. After the sample is irradiated to saturation, its activity (i.e., the rate of disintegration) equals the production rate of activated nuclei, which is $N\sigma\varphi$, where σ is the unknown microscopic activation cross section (averaged over the energy distribution of the given flux) to be determined/evaluated. Consider also that the efficiency of the gamma-ray detector for photons of energy E is ε. The unknown activation cross section is thus obtained from the simple relation

$$\sigma = \frac{A}{\varepsilon N \varphi}, \tag{1.93}$$

where A is the actual counting rate registered by the detector. Suppose that the irradiated sample contains three isotopes which are activated simultaneously in one irradiation field; in other words, the same neutron flux applies to all of the three different activities: $\varphi_1 = \varphi_2 = \varphi_3 = \varphi$. Hence, these "three" fluxes (in reality, the same flux φ) are fully correlated, i.e., $\rho\,(\varphi_i\,,\, \varphi_j) = 1$ for all i and j. Consider also that the following additional information is available:

1. The uncertainty in N is negligible by comparison to all other uncertainties;

2. The relative standard deviation of φ is 2%;
3. The measured activities A_1, A_2, and A_3, are all uncorrelated, and their respective standard deviations are 0.5%, 1.0%, and 0.3%, respectively;
4. The uncertainties in the detector efficiencies corresponding to the three photon energies are correlated, as given in Table (1.2).

Since the uncertainty in N is negligible, it follows that the relative variation of the cross-section σ_i, computed to first order using Eq. (1.93), is

$$\frac{\delta\sigma_i}{\sigma_i} = \frac{\delta A_i}{A_i} - \frac{\delta\varepsilon_i}{\varepsilon_i} - \frac{\delta\varphi_i}{\varphi_i},\ i = 1, 2, 3. \tag{1.94}$$

TABLE 1.2
Uncertainties in the detector efficiencies corresponding to the three correlated photon energies

	Relative Standard Deviation (in %)	**Correlation Matrix**		
ε_1	1.6	1.0		
ε_2	2.2	0.8	1.0	
ε_3	1.3	0.5	0.9	1.0

The three sources of uncertainty for A, ε, and φ, respectively, are uncorrelated. Hence, the relative variance of σ, and the relative covariances $Cov\,(\sigma_i,\, \sigma_j)$ can be computed to first order from Eq. (1.94) to obtain

$$Var\,(\sigma_i) = Var\,(A_i) + Var\,(\varepsilon_i) + Var\,(\varphi_i)\,,$$

$$Cov\,(\sigma_i,\, \sigma_j) = \left\langle \left(\frac{\delta A_i}{A_i} - \frac{\delta\varepsilon_i}{\varepsilon_i} - \frac{\delta\varphi_i}{\varphi_i} \right) \left(\frac{\delta A_j}{A_j} - \frac{\delta\varepsilon_j}{\varepsilon_j} - \frac{\delta\varphi_j}{\varphi_j} \right) \right\rangle =$$

$$= \left\langle \frac{\delta\varepsilon_i}{\varepsilon_i} \ \frac{\delta\varepsilon_j}{\varepsilon_j} \right\rangle + \left\langle \frac{\delta\varphi_i}{\varphi_i} \ \frac{\delta\varphi_j}{\varphi_j} \right\rangle = Cov\,(\varepsilon_i,\, \varepsilon_j) + Cov\,(\varphi_i,\, \varphi_j)\,.$$

Inserting in the above expressions the numerical information provided in

the foregoing yields $Var\,(\sigma_1) = \left(0.5^2 + 1.6^2 + 2.0^2\right)\% = 6.81\%$, so that the relative standard deviation is $\sigma_1 = 2.610\%$. A similar computation yields $Var\,(\sigma_2) = 9.84\%$, or $\sigma_2 = 3.137\%$. The relative covariance of σ_1 and σ_2 is $Cov\,(\sigma_1\,,\,\sigma_2) = 1.6 \times 2.2 \times 0.8 \;+\; 2.0 \times 2.0 \times 1.0 = 6.816$, yielding the following value for corresponding correlation coefficient:

$$\rho\,(\sigma_1\,,\,\sigma_2) = \frac{6.816}{2.610 \times 3.137} = 0.832\ .$$

The remaining computations are left as an exercise for the reader; the complete uncertainty information pertaining to the three measured cross sections is given in Table (1.3):

TABLE 1.3
Uncertainty information pertaining to the three measured cross sections

	Relative Standard Deviation (in %)	**Correlation Matrix**	
ε_1	2.610	1.000	
ε_2	3.137	0.832	1.000
ε_3	2.404	0.803	0.872 1.000

Bibliography

[1] E. Aarts and J. Korst. Simulated annealing and Boltzmann machines. New York, NY; John Wiley and Sons Inc., 1988.

[2] A.K. Alekseev and I. M. Navon. The analysis of an ill-posed problem using multi-scale resolution and second-order adjoint techniques. *Computer Methods in Applied Mechanics and Engineering*, 190(15):1937–1953, 2001.

[3] A.K. Alekseev and I.M. Navon. On estimation of temperature uncertainty using the second order adjoint problem. *International Journal of Computational Fluid Dynamics*, 16(2):113–117, 2002.

[4] R.A. Anthes. Data assimilation and initialization of hurricane prediction models. *Journal of Atmospheric Sciences*, 31:702–719, 1974.

[5] D. Auroux, J. Blum, et al. A nudging-based data assimilation method: the back and forth nudging (BFN) algorithm. *Nonlinear Processes in Geophysics*, 15:305–319, 2008.

[6] F. Baer. Adjustment of initial conditions required to suppress gravity oscillations in nonlinear flows. *Beitraege zur Physik der Atmosphaere*, 50(3):350–366, 1977.

[7] N.L. Baker and R. Daley. Observation and background adjoint sensitivity in the adaptive observation-targeting problem. *Quarterly Journal of the Royal Meteorological Society*, 126(565):1431–1454, 2000.

[8] H.T. Banks and K. Kunisch. *Estimation techniques for distributed parameter systems*. Birkhauser Boston, 1989.

[9] J.W. Bao and Y.H. Kuo. On-off switches in the adjoint method: step functions. *Monthly Weather Review*, 123(5):1589–1594, 1995.

[10] G.A. Barnard. Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3/4):293–315, 1958.

[11] S.L. Barnes. A technique for maximizing details in numerical weather map analysis. *Journal of Applied Meteorology*, 3(4):396–409, 1964.

[12] E.M.L. Beale. A derivation of conjugate gradients. *Numerical methods for nonlinear optimization*, pages 39–43, Academic Press, London, 1972.

[13] R.E. Bellman and R.E. Kalaba. Quasilinearization and nonlinear boundary-value problems. *Elsevier, New York*, 1965.

[14] A.F. Bennett. *Inverse methods in physical oceanography.* Cambridge University Press, 1992.

[15] J.O. Berger. *Statistical decision theory and Bayesian analysis,* 2nd edition. *Springer, New York*, 1985.

[16] J. Bernoulli. *Die Werke von Jakob Bernoulli: Bd. 3: Wahrscheinlichkeitsrechnung*, volume 2. Birkhäuser Basel, 1975.

[17] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Optimisation Numérique: aspects théoriques et pratiques.* Springer, 1997.

[18] D.G. Bounds. New optimization methods from physics and biology. *Nature*, 329:215–219, 1987.

[19] F. Bouttier and P. Courtier. Data assimilation concepts and methods march 1999. *Meteorological training course lecture series. ECMWF*, 2002.

[20] C.G. Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[21] A. Buckley and A. Lenir. QN-like variable storage conjugate gradients. *Mathematical programming*, 27(2):155–175, 1983.

[22] D.G. Cacuci. Sensitivity theory for nonlinear systems. i. nonlinear functional analysis approach. *Journal of Mathematical Physics*, 22:2794, 1981.

[23] D.G. Cacuci. Sensitivity theory for nonlinear systems. II. extensions to additional classes of responses. *Journal of Mathematical Physics*, 22:2803, 1981.

[24] D.G. Cacuci. Global optimization and sensitivity analysis. *Nuclear Science and Engineering*, 104(1):78–88, 1990.

[25] D.G. Cacuci and M.C.G. Hall. Physical interpretation of the adjoint functions for sensitivity analysis of atmospheric models. *Journal of the Atmospheric Sciences*, 40(10):2537–2546, 1983.

[26] D.G. Cacuci and M.C.G. Hall. Systematic analysis of climatic model sensitivity to parameters and processes. *Geophysical Monograph Series*, 29:171–179, 1984.

[27] D.G. Cacuci, M.C.G. Hall, and M.E. Schlesinger. Sensitivity analysis of a radiative-convective model by the adjoint method. *Journal of the Atmospheric Sciences*, 39(9):2038–2050, 1982.

[28] D.G. Cacuci and M. Ionescu-Bujor. On the evaluation of discrepant scientific data with unrecognized errors. *Nuclear Science and Engineering*, 165(1):1–17, 2010.

[29] D.G. Cacuci, M. Ionescu-Bujor, and I.M. Navon. *Sensitivity and Uncertainty Analysis:Applications to Large-Scale Systems*, volume II. CRC, 2005.

[30] M.R. Celis, J.E. Dennis, and Tapia. A trust region strategy for nonlinear equality constrained optimization. In *Proceedings of the SIAM Conference on Numerical Optimization, Edited by Paul T. Boggs, Richard H. Byrd and Robert B. Schnabel*, pages 71–82. SIAM, 1985.

[31] G. Chavent and P. Lemonnier. Identification de la non-linéarité d'une équation parabolique quasilinéaire. *Applied Mathematics & Optimization*, 1(2):121–162, 1974.

[32] X. Chen and I.M. Navon. Optimal control of a finite-element limited-area shallow-water equations model. *Studies in Informatics and Control*, 18(1):41–62, 2009.

[33] S.E. Cohn. An introduction to estimation theory. *Journal of the Meteorological Society of Japan SERIES 2*, 75:147–178, 1997.

[34] A.R. Conn, G.I.M. Gould, and P.L. Toint. *LANCELOT: a Fortran package for large-scale nonlinear optimization (Release A)*. Springer Publishing Company, Inc., 1992.

[35] A.R. Conn, N.I.M. Gould, and P.L. Toint. Trust-region methods: MPS-SIAM series on optimization. *Society of Industrial and Applied Mathematics*, Philadelphia, 2000.

[36] R. Courant and D. Hilbert. *Methods of mathematical physics. Vol. II: Partial differential equations*. Interscience, New York, 1962.

[37] P. Courtier and O. Talagrand. Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus A*, 42(5):531–549, 1990.

[38] P. Courtier, J.N. Thépaut, and A. Hollingsworth. A strategy for operational implementation of 4D-VAR, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519):1367–1387, 1994.

[39] R.T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13, 1946.

[40] G.P. Cressman. An operational objective analysis system. *Monthly Weather Review*, 87(10):367–374, 1959.

[41] A. Da Silva, J. Pfaendtner, J. Guo, M. Sienkiewicz, and S.E. Cohn. Assessing the effects of data selection with the daos physical-space statistical analysis system, 1995. In *Proceedings of the second international symposium on the assimilation of observations in meteorology and oceanography, Tokyo, Japan, WMO and JMA*.

[42] D.N. Daescu and I.M. Navon. An analysis of a hybrid optimization method for variational data assimilation. *International Journal of Computational Fluid Dynamics*, 17(4):299–306, 2003.

[43] R. Daley. Atmospheric data analysis, Cambridge atmospheric and space science series. *Cambridge University Press*, page 457, 1991.

[44] R. Daley. The effect of serially correlated observation and model error on atmospheric data assimilation. *Monthly Weather Review*, 120(1):164–177, 1992.

[45] C. Darwin. On the origins of species by means of natural selection. London: Murray, 1859.

[46] B. Das, H. Meirovitch, and I.M. Navon. Performance of enriched methods for large scale unconstrained optimization as applied to models of proteins. *Journal of Computational Chemistry*, 24(10):1222, 2003.

[47] W.C. Davidon. Variable metric method for minimization. *SIAM Journal on Optimization*, 1(1):1–17, 1991.

[48] R. Dawkins. *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. W.W. Norton & Company, 1986.

[49] K.A. De Jong. Analysis of the behavior of a class of genetic adaptive systems. *Doctoral dissertation, University of Michigan*, 1975.

[50] K. Deb. An efficient constraint handling method for genetic algorithms. *Computer Methods in Applied Mechanics and Engineering*, 186(2):311–338, 2000.

[51] D.P. Dee. Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3323–3343, 2005.

[52] R.S. Dembo and T. Steihaug. Truncated-Newton algorithms for large-scale unconstrained optimization. *Mathematical Programming*, 26(2):190–212, 1983.

[53] J.E. Dennis, M. Heinkenschloss, and L.N. Vicente. Trust-region interior-point SQP algorithms for a class of nonlinear programming problems. *SIAM Journal on Control and Optimization*, 36(5):1750–1794, 1998.

[54] J.C. Derber. The variational four-dimensional assimilation of analyses using filtered models as constraints. *Ph.D. Thesis, University of Wisconsin-Madison*, 1985.

[55] J.C. Derber. A variational continuous assimilation technique. *Monthly Weather Review*, 117:2437–2446, 1989.

[56] J.C. Derber, D.F. Parrish, and S.J. Lord. The new global operational analysis system at the national meteorological center. *Weather and Forecasting*, 6(4):538–547, 1991.

[57] J.C. Derber, D.F. Parrish, and J.G. Sela. The SSI analysis system and extensions to 4–D. In *Proceedings of ECMWF Workshop on variational assimilation with emphasis on three-dimensional aspects, Shinfield Park, Reading RG2 9AX, UK*, pages 15–29, 1992.

[58] A. Eliassen. Provisional report on calculation of spatial covariance and autocorrelation of the pressure field. inst. *Weather and Climate Research, Academy of Science*, Oslo, Report No5, 1954.

[59] M. Fisher. Background error covariance modelling. In *Seminar on Recent Development in Data Assimilation for Atmosphere and Ocean*, pages 45–63. European Centre for Medium-Range Weather Forecasts, 2003.

[60] M. Fisher. *Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems.* European Centre for Medium-Range Weather Forecasts, London, 2003.

[61] M. Fisher. Assimilation algorithms lecture 3: 4–D VAR. *European Centre for Medium-Range Weather Forecasts*, London, 2010.

[62] M. Fisher, J. Nocedal, Y. Trémolet, and S.J. Wright. Data assimilation in weather forecasting: a case study in PDE-constrained optimization. *Optimization and Engineering*, 10(3):409–426, 2009.

[63] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, 1970.

[64] R. Fletcher. A model algorithm for composite nondifferentiable optimization problems. *Nondifferential and Variational Techniques in Optimization, Mathematical Programming Studies*, 17:67–76, Springer,1982.

[65] R. Fletcher, S. Leyffer, and P.L. Toint. On the global convergence of a filter–SQP algorithm. *SIAM Journal on Optimization*, 13(1):44–59, 2002.

[66] R. Fletcher and C.M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154, 1963.

[67] W. Freeden and U. Windheuser. Spherical wavelet transform and its discretization. *Advances in Computational Mathematics*, 5(1):51–94, 1996.

[68] F.H. Fröhner. Assigning uncertainties to scientific data. *Nuclear Science and Engineering*, 126(1):1–18, 1997.

[69] F.H. Fröhner. Evaluation of data with systematic errors. *Nuclear Science and Engineering*, 145(3):342–353, 2003.

[70] L.S. Gandin. *Objective analysis of meteorological fields,Translated from the Russian.* Jerusalem (Israel Program for Scientific Translations), 1965.

[71] G. Gaspari and S.E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757, 2006.

[72] D.M. Gay. Computing optimal locally constrained steps. *SIAM Journal on Scientific and Statistical Computing*, 2(2):186–197, 1981.

[73] J.K. Gibson, P. Kallberg, S. Uppala, A. Nomura, E. Serrano, and A. Hernandez. ERA description. ECMWF reanalysis project report 1: project organisation. Technical report, ECMWF, European Centre of Medium-Range Weather Forecasts, Reading, 1997.

[74] J.C. Gilbert and C. Lemaréchal. Some numerical experiments with variable-storage Quasi-Newton algorithms. *Mathematical Programming*, 45(1):407–435, 1989.

[75] P.E. Gill and W. Murray. Conjugate-gradient methods for large-scale nonlinear optimization. *Defense Technical Information Center*, page 64, 1979.

[76] P.E. Gill, W. Murray, and M.H. Wright. *Practical optimization.* Addison-Wesley Professional Academic Press, 1981.

[77] D.E. Goldberg. Genetic algorithms in search, optimization, and machine learning. *Addison-Wesley Professional*, 1st edition:432, 1989.

[78] D. Goldfarb. A family of variable metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[79] G.H. Golub and C.F. Van Loan. *Matrix computations.* John Hopkins University Press, 1989.

[80] A. Grammeltvedt. A survey of finite-difference schemes for the primitive equations for a barotropic fluid. *Monthly Weather Review*, 97(5):384–404, 1969.

[81] S. Gratton, A.S. Lawless, and N.K. Nichols. Approximate Gauss-Newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 18(1):106–132, 2007.

[82] J.J. Grefenstette. Optimization of control parameters for genetic algorithms. *Systems, IEEE Transactions on Systems, Man and Cybernetics*, 16, 1986.

[83] A. Griewank and P.L. Toint. Partitioned variable metric updates for large structured optimization problems. *Numerische Mathematik*, 39(1):119–137, 1982.

[84] A. Griewank and A. Walther. Evaluating derivatives: principles and techniques of algorithmic differentiation. *Society for Industrial and Applied Mathematics (SIAM)*, page 460, 2008.

[85] A.K. Griffith and N.K. Nichols. Accounting for model error in data assimilation using adjoint methods. *Computational Differentiation: Techniques, Applications and Tools*, pages 195–204, 1996.

[86] A.K. Griffith and N.K. Nichols. Adjoint methods in data assimilation for estimating model error. *Flow, Turbulence and Combustion*, 65(3):469–488, 2000.

[87] M.R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

[88] R.N. Hoffman. A four-dimensional analysis exactly satisfying equations of motion. *Monthly Weather Review*, 114(2):388–397, 1986.

[89] J.H. Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence*. University of Michigan Press, 1975.

[90] A. Hollingsworth and P. Lönnberg. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: the wind field. *Tellus A*, 38(2):111–136, 1986.

[91] R.B. Hollstein. Artificial genetic adaptation in computer control systems. Doctoral dissertation, University of Michigan, Dissertation Abstracts International 32.(3), 1510B, 1971.

[92] C. Homescu and I.M. Navon. Optimal control of flow with discontinuities. *Journal of Computational Physics*, 187(2):660–682, 2003.

[93] B. Ingleby. The statistical structure of forecast errors and its representation in the meteorological office global model. *Quarterly Journal of the Royal Meteorological Society*, 124:1783–1807, 2001.

[94] C. Jakob. Ice clouds in numerical weather prediction models–progress, problems and prospects. pages 327–345, In *Cirrus* edited by David K. Lynch et al., Oxford University Press, 327-345, 2002.

[95] M. Janisková, J. Mahfouf, J. Morcrette, and F. Chevallier. Linearized radiation and cloud schemes in the ECMWF model: development and evaluation. *Quarterly Journal of the Royal Meteorological Society*, 128(583):1505–1527, 2002.

[96] M. Janisková, J.N. Thépaut, and J.F. Geleyn. Simplified and regular physical parameterizations for incremental four-dimensional variational assimilation. *Monthly Weather Review*, 127(1):26–45, 1999.

[97] T. Janjic and S.E. Cohn. Treatment of observation error due to unresolved scales in atmospheric data assimilation. *Monthly Weather Review*, 134(10):2900–2915, 2006.

[98] E.T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, 1968.

[99] E.T. Jaynes. Probability, statistics and statistical physics. *Dordrecht: Reidel*, 1983.

[100] A.H. Jazwinski. Stochastic processes and filtering theory. *Mathematics in Science and Engineering*, 64:376, 1970.

[101] E. Kalnay. *Atmospheric modeling, data assimilation and predictability.* Cambridge University Press, 2003.

[102] N.M.S. Karmitsa and M.M. Mäkelä. Adaptive limited memory bundle method for bound constrained large-scale nonsmooth optimization. *Optimization*, 59(6):945–962, 2009.

[103] N.M.S. Karmitsa, M.M. Mäkelä, and M.M. Ali. Limited memory interior point bundle method for large inequality constrained nonsmooth minimization. *Applied Mathematics and Computation*, 198(1):382–400, 2008.

[104] S. Kirkpatrick. Optimization by simulated annealing: quantitative studies. *Journal of Statistical Physics*, 34, Issues 5–6:975–986, 1984.

[105] S. Kirkpatrick, M.P. Vecchi, et al. *Optimization by simulated annealing.* Research Report RC 9355, IBM (Yorktown Heights, NY), 1982.

[106] K.C. Kiwiel. *Methods of descent for nondifferentiable optimization, lecture Notes in Mathematics.* Number 1133. Springer-Verlag Berlin, 1985.

[107] K.C. Kiwiel. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, 46(1):105–122, 1990.

[108] A.N. Kolmogorov, W.L. Doyle, and I. Selin. Interpolation and extrapolation of stationary random sequences. *RAND Reports and Bookstore, Research Memoranda*, RM(3090), 1962.

[109] D.G. Krige. *A statistical approach to some mine valuation and allied problems on the Witwatersrand.* University of the Witwatersrand, 1951.

[110] T.N. Krishnamurti, J. Xue, H.S. Bedi, K. Ingles, and D. Oosterhof. Physical initialization for numerical weather prediction over the tropics. *Tellus B*, 43(4):53–81, 1991.

[111] P.S. Laplace. *Théorie analytique des probabilités*. Paris, V. Courcier, 1812.

[112] S. Laroche and P. Gauthier. A validation of the incremental formulation of 4D variational data assimilation in a nonlinear barotropic flow. *Tellus A*, 50(5):557–572, 1998.

[113] A.S. Lawless. A note on the analysis error associated with 3D-FGAT. *Quarterly Journal of the Royal Meteorological Society*, 136(649):1094–1098, 2010.

[114] A.S. Lawless, S. Gratton, and N.K. Nichols. An investigation of incremental 4D-VAR using non-tangent linear models. *Quarterly Journal of the Royal Meteorological Society*, 131(606):459–476, 2005.

[115] A.S. Lawless and N.K. Nichols. Inner-loop stopping criteria for incremental four-dimensional variational data assimilation. *Monthly Weather Review*, 134(11):3425–3435, 2006.

[116] F.X. Le Dimet. Une application des méthodes de contrôle, optimal 'a l'analyse varationelle. *Report, LAMP, Université Blaise-Pascal*, 63170, 1981.

[117] F.X. Le Dimet, I.M. Navon, and D.N. Daescu. Second-order information in data assimilation. *Monthly Weather Review*, 130(3):629–648, 2002.

[118] F.X. Le Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38(2):97–110, 1986.

[119] D. Legler and I.M. Navon. VARIATM - A Fortran program for objective analysis of pseudo-stress wind fields using large-scale conjugate-gradient minimization. *Computers and Geosciences*, 17(1):1–21, 1991.

[120] D. Legler, I.M. Navon, and J.J. O'Brien. Objective analysis of pseudostress over the Indian Ocean using a direct-minimization approach. *Monthly Weather Review*, 117(8):709–720, 1989.

[121] C. Lemaréchal. Nondifferentiable optimization, subgradient and subgradient methods. *Optimization and Operations Research, Lecture Notes in Economics and Mathematical Systems*, 117:191, 1976.

[122] C. Lemaréchal. Nondifferentiable optimization. *Handbooks in Operations Research and Management Science*, 1:529–572, 1989.

[123] J.M. Lewis, S. Lakshmivarahan, and S. Dhall. *Dynamic data assimilation: a least squares approach*, volume 104. Cambridge University Press, 2006.

[124] Y. Li, I.M. Navon, P. Courtier, and P. Gauthier. Variational data assimilation with a semi-lagrangian semi-implicit global shallow-water equation model and its adjoint. *Monthly Weather Review*, 121(6):1759–1769, 1993.

[125] Z. Li, I.M. Navon, and Y. Zhu. Performance of 4–D VAR with different strategies for the use of adjoint physics with the FSU global spectral model. *Monthly Weather Review*, 128(3):668–688, 2000.

[126] D.C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

[127] A.C. Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194, 1986.

[128] D.G. Luenberger. *Linear and nonlinear programming*. Addison-Wesley Publishing Company, 1984.

[129] L. Lukšan and J. Vlček. NDA: Algorithms for nondifferentiable optimization. *Technical Report*, V-797, ICS AS CR, 2000.

[130] P. Lynch and X. Huang. Initialization of the HIRLAM model using a digital filter. *Monthly Weather Review*, 120(6):1019–1034, 1992.

[131] P. Lynch and A. McDonald. A multi-level limited-area slow-equation model: application to initialization. *Quarterly Journal of the Royal Meteorological Society*, 116(493):595–609, 1990.

[132] B. Machenhauer. On the dynamics of gravity oscillations in a shallow water model with applications to normal mode initialization. *Beiträge zur Physik Atmosphäre*, 50:253–271, 1977.

[133] J.F. Mahfouf and F. Rabier. The ECMWF operational implementation of four-dimensional variational assimilation. II: experimental results with improved physics. *Quarterly Journal of the Royal Meteorological Society*, 126(564):1171–1190, 2000.

[134] M.M. Makela and P. Neittaanmäki. *Nonsmooth optimization: analysis and algorithms with applications to optimal control*. World Scientific Singapore, 1992.

[135] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087, 1953.

[136] K. Miyakoda and R.W. Moyer. A method of initialization for dynamical weather forecasting. *Tellus*, 20(1):115–128, 1968.

[137] J.L. Morales and J. Nocedal. Automatic preconditioning by limited memory quasi-Newton updating. *SIAM Journal on Optimization*, 10(4):1079–1096, 1999.

[138] J.L. Morales and J. Nocedal. Algorithm 809: Preqn: Fortran 77 subroutines for preconditioning the conjugate gradient method. *ACM Transactions on Mathematical Software (TOMS)*, 27(1):83–91, 2001.

[139] S.G. Nash. Truncated-Newton methods. *Ph.D. Thesis, Computer Science Departement, Stanford University*, 1982.

[140] S.G. Nash. Newton-type minimization via the Lanczos method. *SIAM Journal on Numerical Analysis*, 21(4):770–788, 1984.

[141] S.G. Nash. Solving nonlinear programming problems using Truncated-Newton techniques. *Numerical Optimization*, pages 119–136, 1984.

[142] S.G. Nash. Truncated-Newton methods for large-scale function minimization. *Application of Nonlinear Programming to Optimization and Control*, page 91, 1984.

[143] S.G. Nash. User's guide for TN. *TNBC: Fortran Routines for Nonlinear Optimization Report*, 397, 1984.

[144] S.G. Nash. Preconditioning of Truncated-Newton methods. *SIAM Journal on Scientific and Statistical Computing*, 6(3):599–616, 1985.

[145] S.G. Nash and J. Nocedal. A numerical study of the limited memory BFGS method and the Truncated-Newton method for large scale optimization. *SIAM Journal on Optimization*, 1(3):358–372, 1991.

[146] S.G. Nash and A. Sofer. Block Truncated-Newton methods for parallel optimization. *Mathematical Programming*, 45(1):529–546, 1989.

[147] S.G. Nash and A. Sofer. Assessing a search direction within a Truncated-Newton method. *Operations Research Letters*, 9(4):219–221, 1990.

[148] S.G. Nash and A. Sofer. *Linear and nonlinear programming.* McGraw-Hill Science/Engineering/Math, 1996.

[149] I.M. Navon, F.B. Brown, and D.H. Robertson. A combined simulated annealing and quasi-Newton-like conjugate-gradient method for determining the structure of mixed argon-xenon clusters. *Computers & Chemistry*, 14(4):305–311, 1990.

[150] I.M. Navon and R. De Villiers. Combined penalty multiplier optimization methods to enforce integral invariants conservation. *Monthly Weather Review*, 111(6):1228–1243, 1983.

[151] I.M. Navon and D. Legler. Conjugate-gradient methods for large-scale minimization in meteorology. *Monthly Weather Review*, 115(4):1479–1502, 1987.

[152] I.M. Navon, P.K.H. Phua, and M. Ramamurthy. Vectorization of conjugate-gradient methods for large-scale minimization in meteorology. *Journal of Optimization Theory and Applications*, 66(1):71–93, 1990.

[153] I.M. Navon, X. Zou, J. Derber, and J. Sela. Variational data assimilation with an adiabatic version of the NMC spectral model. 120(7):1443–1446, 1992.

[154] J. Nocedal. Theory of algorithms for unconstrained optimization. *Acta Numerica*, 1:199–242, 1992.

[155] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer Verlag, 2006.

[156] J.M. Ortega and W.C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, New York, 1970.

[157] D.F. Parrish and J.C. Derber. The national meteorological center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, 120(8):1747–1763, 1992.

[158] V. Penenko and N.N. Obraztsov. A variational initialization method for the fields of the meteorological elements. *Meteorologiya i Gidrologiya*, 11:1–11, 1976.

[159] N.A. Phillips. On the problem of initial data for the primitive equations. *Tellus*, 12:121–126, 1960.

[160] M. Pincus. Letter to the editor: A Monte Carlo Method for the approximate solution of certain types of constrained optimization problems. *Operations Research*, 18(6):1225–1228, 1970.

[161] E. Polak and G. Ribiere. Note sur la convergence de méthodes de directions conjuguées. *Revue française d'informatique et de recherche opérationnelle*, 16:35–43, 1969.

[162] S. Polavarapu, M. Tanguay, and L. Fillion. Four-dimensional variational data assimilation with digital filter initialization. *Monthly Weather Review*, 128(7):2491–2510, 2000.

[163] M.J.D. Powell. Rank one methods for unconstrained optimization. Technical report, Atomic Energy Research Establishment, Harwell (England), 1969.

[164] M.J.D. Powell. Restart procedures for the conjugate gradient method. *Mathematical Programming*, 12(1):241–254, 1977.

[165] F. Rabier, H. Järvinen, E. Klinker, J.F. Mahfouf, and A. Simmons. The ECMWF operational implementation of four-dimensional variational assimilation. i: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, 126(564):1143–1170, 2000.

[166] M.K. Ramamurthy and F.H. Carr. Four-dimensional data assimilation in the monsoon region. i: experiments with wind data. *Monthly Weather Review*, 115(8):1678–1706, 1987.

[167] M.K. Ramamurthy and I.M. Navon. The conjugate-gradient variational analysis and initialization method: an application to MONEX SOP-2 data. *Monthly Weather Review*, 120(10):2360–2377, 1992.

[168] A. Renyi. Probability calculus, Tankonyvkiado, Budapest, 1954; German transl. *Hochschulbucher fur Mathematik, VEB Deutscher Verlag*, 54, 1962.

[169] A.H.G. Rinnooy Kan and G.T. Timmer. Chapter IX. Global Optimization. *Handbooks in Operations Research and Management Science*, 1:631–662, 1989.

[170] C. Robert, E. Blayo, and J. Verron. Comparison of reduced-order, sequential and variational data assimilation methods in the tropical Pacific Ocean. *Ocean Dynamics*, 56(5):624–633, 2006.

[171] C. Robert, E. Blayo, and J. Verron. Reduced-order 4–D VAR: A preconditioner for the incremental 4–D VAR data assimilation method. *Geophysical Research Letters*, 33:L18609, 2006.

[172] Kirkpatrick S., C.D. Jr. Gerlatt, and M.P. Vecchi. Optimization by simulated annealing. *IBM Research Report*, (9355), 1982.

[173] Y. K. Sasaki. Proposed inclusion of time variation terms, observational and theoretical, in numerical variational objective analysis. *Journal Meteorological Society Japan*, 47(2):115–124, 1969.

[174] Y. K. Sasaki. Numerical variational analysis formulated under the constraints as determined by longwave equations and a low-pass filter. *Monthly Weather Review*, 98(12):884–898, 1970.

[175] Y. K. Sasaki. Numerical variational analysis with weak constraint and application to surface analysis of severe storm gust. *Monthly Weather Review*, 98(12):899–910, 1970.

[176] Y. K. Sasaki. Some basic formalisms in numerical variational analysis. *Monthly Weather Review*, 98(12):875–883, 1970.

[177] Y. K.K. Sasaki. A fundamental study of the numerical prediction based on the variational principle. *Journal Meteorological Society Japan*, 33(6):262–275, 1955.

[178] T. Schlick. Modified Cholesky factorizations for sparse preconditioners. *SIAM Journal on Scientific Computing*, 14(2):424–445, 1993.

[179] T. Schlick and A. Fogelson. Algorithm 702: TNPACK–a Truncated Newton minimization package for large-scale problems: I. algorithm and usage. *ACM Transactions on Mathematical Software (TOMS)*, 18(2):141, 1992.

[180] T. Schlick and A. Fogelson. TNPACK: A Truncated Newton minimization package for large-scale problems: II. implementation examples. *ACM Transactions on Mathematical Software (TOMS)*, 18(1):71–111, 1992.

[181] T. Schlick and M. Overton. A powerful Truncated Newton method for potential energy minimization. *Journal of Computational Chemistry*, 8(7):1025–1039, 1987.

[182] H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2(1):121–152, 1992.

[183] E. Schrödinger. The foundation of the theory of probability: I. In *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences*, volume 51, pages 51–66, 1945.

[184] F.H.M. Semazzi and I.M. Navon. A comparison of the bounded derivative and the normal-mode initialization methods using real data. *Monthly Weather Review*, 114:2106–2121, 1986.

[185] D.F. Shanno. Conjugate gradient methods with inexact searches. *Mathematics of Operations Research*, 3(3):244–256, 1978.

[186] D.F. Shanno. On the convergence of a new conjugate gradient algorithm. *SIAM Journal on Numerical Analysis*, 15(6):1247–1257, 1978.

[187] D.F. Shanno et al. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

[188] D.F. Shanno and K.H. Phua. Remark on algorithm 500-a variable method subroutine for unconstrained nonlinear minimization. *ACM Transactions on Mathematical Software*, 6(4):618–622, 1980.

[189] E. Shannon. A mathematical theory of evidence:. *Bell System Technical Journal*, 27:379–423, 1948.

[190] N.Z. Shor, K.C. Kiwiel (translator), and R. Ruszczynski (translator). Minimization methods for non-differentiable functions. *Springer series in computational mathematics, 3, Berlin*, page 172, 1985.

[191] D.C. Sorensen. Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2):409–426, 1982.

[192] J.L. Steward, I.M. Navon, M. Zupanski, and N. Karmitsa. Impact of nonsmooth observation operators on variational and sequential data assimilation for a limited-area shallow-water equation model. *Quarterly Journal of the Royal Meteorological Society*, 138(663):323–339, Part B 2012.

[193] O. Talagrand and P. Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113(478):1311–1328, 1987.

[194] J.N. Thépaut and P. Courtier. Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model. *Quarterly Journal of the Royal Meteorological Society*, 117(502):1225–1254, 1991.

[195] P.D. Thompson. Reduction of analysis error through constraints of dynamical consistency. *Journal of Applied Meteorology*, 8:738–742, 1969.

[196] M. Tiedtke. Representation of clouds in large-scale models. *Monthly Weather Review*, 121(11):3040–3061, 1993.

[197] A.N. Tikhonov. Regularization of non-linear ill-posed problems. *Doklady Akademii Nauk*, 49(4), 1963.

[198] A.M. Tompkins and M. Janisková. A cloud scheme for data assimilation: description and initial tests. *Quarterly Journal of the Royal Meteorological Society*, 130(602):2495–2517, 2004.

[199] Y. Trémolet. Diagnostics of linear and incremental approximations in 4-D VAR. *Quarterly Journal of the Royal Meteorological Society*, 130(601):2233–2251, 2004.

[200] Y. Trémolet. Incremental 4-D VAR convergence study. *Tellus A*, 59(5):706–718, 2007.

[201] Y. Trémolet. Computation of observation sensitivity and observation impact in incremental variational data assimilation. *Tellus A*, 60(5):964–978, 2008.

[202] M. Tribus. Rational descriptions, decisions, and designs... *Elsevier Science and Technology xix*, page 478, 1969.

[203] J. Tshimanga, S. Gratton, A.T. Weaver, and A. Sartenaer. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 134(632):751–769, 2008.

[204] T. Tsuyuki. Variational data assimilation in the tropics using precipitation data. part III: Assimilation of SSM/I precipitation rates. *Monthly Weather Review*, 125(7):1447–1464, 1997.

[205] S.P. Uryasev. New variable-metric algorithms for nondifferentiable optimization problems. *Journal of Optimization Theory and Applications*, 71(2):359–388, 1991.

[206] F. Veerse and J.N. Thépaut. Multiple-truncation incremental approach for four-dimensional variational data assimilations. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1889–1908, 1998.

[207] P.A. Vidard, E. Blayo, F.X. Le Dimet, and A. Piacentini. 4–D variational data analysis with imperfect model. *Flow, Turbulence and Combustion*, 65(3):489–504, 2000.

[208] P.A. Vidard, A. Piacentini, and F.X. Le Dimet. Variational data analysis with control of the forecast bias. *Tellus A*, 56(3):177–188, 2004.

[209] J. Vlček and L. Lukšan. Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization. *Journal of Optimization Theory and Applications*, 111(2):407–430, 2001.

[210] T. Vukićević and R.M. Errico. Linearization and adjoint of parameterized moist diabatic processes. *NCAR*, 1030(92), 1992.

[211] A. Wald. *Statistical decision functions*. New York, John Wiley, 1950.

[212] Z. Wang, K.K. Droegemeier, L. White, and I.M. Navon. Application of a new adjoint newton algorithm to the 3–D ARPS storm-scale model using simulated data. *Monthly Weather Review*, 125(10):2460–2478, 1997.

[213] Z. Wang, K.K. Droegemeier, M. Xue, S.K. Park, J.G. Michalakes, and X. Zou. Sensitivity analysis of a 3–D compressible storm-scale model to input parameters. In *Preprints, Proceedings of the International Symposium on Assimilation of Observations in Meteorology and Oceanography, Tokyo, Japan*, 1995.

[214] Z. Wang, I.M. Navon, X. Zou, and F.X. Le Dimet. A Truncated Newton optimization algorithm in meteorology applications with analytic hessian/vector products. *Computational Optimization and Applications*, 4(3):241–262, 1995.

[215] A. Weaver and P. Courtier. Correlation modelling on the sphere using a generalized diffusion equation. *Quarterly Journal of the Royal Meteorological Society*, 127(575):1815–1846, 2001.

[216] N. Weiner. Extrapolation, interpolation, and smoothing of stationary time series With Engineering Applications. *MIT Press*, page 166, 1949.

[217] P. Wolfe. Convergence conditions for ascent methods. *SIAM Review*, 11(2):226–235, 1969.

[218] P. Wolfe. Convergence conditions for ascent methods. II: some corrections. *SIAM Review*, 13(2):185–188, 1971.

[219] Q. Xu. Generalized adjoint for physical processes with parameterized discontinuities. part I: basic issues and heuristic examples. *Journal of the Atmospheric Sciences*, 53(8):1123–1142, 1996.

[220] Q. Xu. Generalized adjoint for physical processes with parameterized discontinuities. part IV: problems in time discretization. *Journal of the Atmospheric Sciences*, 54(23):2722–2728, 1997.

[221] Q. Xu, J. Gao, and W. Gu. Generalized adjoint for physical processes with parameterized discontinuities. part V: Coarse-grain adjoint and problems in gradient check. *Journal of the Atmospheric Sciences*, 55(11):2130–2135, 1998.

[222] A.M. Yaglom. *Introduction to the theory of stationary random functions*. Dover, 1962.

[223] S. Zhang, X. Zou, J. Ahlquist, I.M. Navon, and J.G. Sela. Use of differentiable and nondifferentiable optimization algorithms for variational data assimilation with discontinuous cost functions. *Monthly Weather Review*, 128(12):4031–4044, 2000.

[224] K. Zhu, I.M. Navon, and X. Zou. Variational data assimilation with a variable resolution finite-element shallow-water equations model. *Monthly Weather Review*, 122(5):946–965, 1994.

[225] Y. Zhu and I.M. Navon. Impact of parameter estimation on the performance of the FSU global Spectral Model using its full-physics adjoint. *Monthly Weather Review*, 127(7):1497–1517, 1999.

[226] X. Zou. Tangent linear and adjoint of "on-off" processes and their feasibility for use in 4-dimensional variational data assimilation. *Tellus A*, 49(1):3–31, 1997.

[227] X. Zou, A. Barcilon, I.M. Navon, J. Whitaker, and D.G. Cacuci. An adjoint sensitivity study of blocking in a two-layer isentropic model. *Monthly Weather Review*, 121(10):2833–2857, 1993.

[228] X. Zou and Y.H. Kuo. Rainfall assimilation through an optimal control of initial and boundary conditions in a limited-area mesoscale model. *Monthly Weather Review*, 124(12):2859–2882, 1996.

[229] X. Zou, H Liu, J. Derber, J.G. Sela, R. Treadon, I.M. Navon, and B. Wang. Four–dimensional variational data assimilation with a full–physics version of the NCEP spectral model: system development and preliminary results. *Quarterly Journal of the Royal Meteorological Society*, 127(573):1095–1122, 2001.

[230] X. Zou, I.M. Navon, M. Berger, K.H. Phua, T. Schlick, and F.X. Le Dimet. Numerical experience with limited-memory Quasi-Newton and Truncated Newton methods. *SIAM Journal on Optimization*, 3(3):582–608, 1993.

[231] X. Zou, I.M. Navon, and F.X. Le Dimet. Incomplete observations and control of gravity waves in variational data assimilation. *Tellus A*, 44(4):273–296, 1992.

[232] X. Zou, I.M. Navon, and F.X. Le Dimet. An optimal nudging data assimilation scheme using parameter estimation. *Quarterly Journal of the Royal Meteorological Society*, 118(508):1163–1186, 1992.

[233] X. Zou, I.M. Navon, and J.G. Sela. Control of gravitational oscillations in variational data assimilation. *Monthly Weather Review*, 121:272–289, 1993.

[234] X. Zou, I.M. Navon, and J.G. Sela. Variational data assimilation with moist threshold processes using the NMC spectral model. *Tellus A*, 45(5):370–387, 1993.

[235] D. Zupanski. The effects of discontinuities in the Betts–Miller cumulus convection scheme on four-dimensional variational data assimilation. *Tellus A*, 45(5):511–524, 1993.

[236] D. Zupanski. A general weak constraint applicable to operational 4D-VAR data assimilation systems. *Monthly Weather Review*, 125(9):2274–2292, 1997.

[237] M. Zupanski. Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Monthly Weather Review*, 121(8):2396–2408, 1993.

[238] M. Zupanski. Maximum likelihood ensemble filter: theoretical aspects. *Monthly Weather Review*, 133(6):1710–1726, 2005.

[239] M. Zupanski, I.M. Navon, and D. Zupanski. The maximum likelihood ensemble filter as a non-differentiable minimization algorithm. *Quarterly Journal of the Royal Meteorological Society*, 134(633):1039–1050, 2008.

[240] M. Zupanski, D. Zupanski, T. Vukicevic, K. Eis, and T.V. Haar. CIRA CSU four–dimensional variational data assimilation system. *Monthly Weather Review*, 133(4):829–843, 2005.