

# Chapter 2

## Some Statistical Perspectives of Growth Models in Health Care Plans

Pranab K. Sen

**Abstract** Growth (and wear) curve models, having genesis in epidemiology and system biology, have cropped up in every walk of life and science. In statistics, such growth curve models have led to an evolution of multivariate analysis with better performance characteristics and enhanced scope of applications in many interdisciplinary field of research. Recent advances in bioinformatics and genomic science have opened the Pandora's box with high-dimensional data models, often with relatively smaller sample sizes. Growth curve models are especially useful in such contexts. There are also other areas where growth curve model-based analyses are in high demand. In this vein, the scope and perspectives of growth models are appraised with special emphasis on some health care and health study plans.

### 2.1 Introduction

In exploratory studies, especially in experimental biology, developmental biology, medicine, epidemiology, socio-economics, psychology, and more recently, in biotechnology, information technology, toxico-genomics and bioinformatics, such growth models have been systematically studied under the terminology *longitudinal data models* and *repeated measurement models*; classical *growth curve models* (GCM) in simple parametric setups are regarded as precursors. Box (1950) initiated the study of *growth and wear* curves in simple biometric setups. C.R. Rao (1958, 1965) made significant contributions to GCM while Potthoff and Roy (1964) systematically integrated GCMs in the main stream of *multivariate analysis of variance* (MANOVA) and linked it to *multivariate analysis of covariance* (MANOCOVA). Rao (1959) developed procedures for parameter estimation and

---

P.K. Sen (✉)

Departments of Biostatistics, and Statistics & Operations Research, University of North Carolina, Chapel Hill, NC 27599-7420, USA

e-mail: [pksen@bios.unc.edu](mailto:pksen@bios.unc.edu)

estimation of confidence bands for the response curve. [Cole and Grizzle \(1966\)](#) and [Grizzle and Allen \(1969\)](#) formulated some innovative statistical analysis of growth and dose response curves. [Geisser \(1970, 1981\)](#) annexed Bayesian methodology in GCMs. [Khatri \(1966\)](#) elaborated the connection of GCM and MANO(CO)VA. [Timm \(1981\)](#) and [Zerbe and Walker \(1977\)](#) studied further MANO(CO)VA of repeated measurement designs incorporating the basics of parametric GCM. This also laid down the foundation of *random-effects* and *mixed-effects* models in MANOCOVA. In some medical (and dental) research problems, often, *area under the curve* (AUC) has been used. It is also possible to relate AUC to GCM and obtain better performance of statistical tests and estimates ([Preisser et al. 2011](#)).

In most of these developments, as has been systematically accounted in [Gnanadesikan et al. \(1971\)](#), it has been tacitly assumed that (i) the underlying response variables are continuous, (ii) additivity of the effects hold, and (iii) the underlying probability distributions are all multivariate normal. The latter assumption is accompanied by the homogeneity of their dispersion matrices, the so-called, *homoscedasticity* condition in a general multivariate setup. Both the linearity of the model and multinormality of the errors have been critically appraised in the past 50 years, raising concern of the scope of adaptability of normal MANOCOVA models in various applications. Nonparametric (mostly based on marginal ranks) MANO(CO)VA evolved during the 1960s and reported in [Puri and Sen \(1971, 1985\)](#). For GCM, such nonparametrics have been incorporated by [Ghosh et al. \(1973\)](#) and [Sen \(1973, 1985\)](#), among others. The past three decades have witnessed the development of semi-parametric GCM and longitudinal data models. Both spatial and temporal variations are accounted in such models.

In epidemiology, epidemic models are earlier examples of growth models. The growth of a disease or disorder (in a population) follows another track of discrete GCM where typically the response variable is the number of infected people or their proportion in the target population. In population dynamics, such discrete GCM are commonly perceived wherein various demographic features account for explanatory or design variables. For example, for the HIV afflicted population in a *spatiotemporal* setups, discrete GCM are quite appealing, albeit the multi-normality or the linearity of effects assumption may not be reasonable. In system biology, for example, the growth of a tumor or spread of cancerous cells, growth (curve) models are very appealing, albeit they come under high-dimensional or functional data clouds. The classical fMRI models also pertain to growth models, although the commonly assumed multi-normality condition may not be generally tenable in such contexts. In many stochastic models, such as the diffusion process, birth and death process, and morbidity (illness) process, such GCM may appear, not only with some longitudinal or temporal features of the expectation parameters but also with subtle change in the shape or dispersion parameters. For example, the drift versus dispersion in generalized random walk models. From white noise to signal detection in high dimension (as related to chaos theory) is another example of this sort. Markov processes have also been attuned to GCM with appropriate growth condition on the failure rate or reliability functions. Nonhomogeneous Poisson processes and

their natural extension to *doubly stochastic Poisson processes* bear growth features in a stochastic mode. Also, GCM in HIV (AIDS) models come in a completely different setup. It is therefore perceived that conventional MANO(CO)VA-related GCM may not be universally adaptable in many other fields of applications. *Beyond parametrics* in GCM is therefore a natural avenue to traverse.

There is a class of scenarios of growth (or decay) models which are characterized by evolutionary growth or decay but subject to extraneous restraints. For example, in a *branching process* model the outcome variable may lead to an extinction if it reaches the absorbing state (0). On the other hand it can explode to an infinite state under plausible conditions on the branching parameters. In some toxicological models, such growth patterns may have similarity with GCMs but are subject to suitable upper bounds due to experimental constraints. For example, when the output variable attains an upper threshold level, the system moves to a different stage, and a new process model comes in the picture. This may be regarded as a GCM annexation to the classical *change-point model* which typically relates to either a change in location (regression) or scale parameter at an unknown time-point. The problems is much more complex in this general growth model. A typical example is the growth of HIV-AIDS afflicted population following some break-through medical intervention. For growth models with a finite upper bound, the classical [Gompertz \(1825\)](#) model, motivated by a distribution function to fit mortality tables is a precursor to other models such as the logistic model and its ramifications ([Johnson and Kotz 1970](#)). With this genesis, logistic regression models pertain to stochastic growth curves in more general formulations. Likewise, Poisson regression models pertain to such GCM ([Sen et al. 2010](#)).

This volume has a primary emphasis on GCM in conventional agricultural setups with emphasis on the *elephant foot yam*. In modern interdisciplinary research, typically, high-dimensional data models are encountered where some times the sample size may be relatively smaller, thus giving rise to the so-called *high dimension low sample size* (HDLSS) models. This is particularly the case with bioinformatics and toxico-genomics studies. Even in many socioeconomic investigations, HDLSS models are encountered in very nonstandard setups. The scope for traditional MANOCOVA tools in HDLSS has been critically appraised in the recent past ([Sen 2006, 2008](#)). In this context, the dimension reduction can be effectively done with appropriate GCM in beyond parametrics setups ([Sen et al. 2007](#)). Nevertheless, conventional statistical tools are of very limited utility in such HDLSS-related GCM setups. This study focuses on some high-dimensional models arising in some socioeconomic research problems where GCM may have a natural appeal. The next section is devoted to the preliminary notion on the evolution of GCM from simple parametric to beyond parametric setups, encompassing HDLSS models as well. Some of these beyond parametrics perspectives are elaborated in Sect. 2.3. The main results on GCM approach on some general socioeconomic models are disseminated in Sect. 2.4. The concluding section is devoted to some general observations and remarks.

## 2.2 Preliminary Notion

Typically, GCM relates to some multi-sample or blocked design models where (in a general setup), there are  $n$  observations, each observation has  $p$  characteristics and observed at  $q$  time points. For example, in an environmental health hazard study for identifying environmental dioxin pollution (Chen et al. 2012), *fingerprint analysis* comparing the polychlorinated *dibenzo-p-dioxin* and *dibenzofuran* (PCDD/F) congener profile patterns of collected samples with those of potential dioxin emission source(s), has been advocated as an important tool. There are  $p$  ( $= 17$ ) PCDD/F congeners comprising a fingerprint and data collected in a longitudinal setup. This typically relate to a MANOVA model, albeit the sample sizes are small, and moreover, the underlying distributions are distinctly not multivariate normal; multivariate gamma distributions appear to be more reasonable in this setup for which the dispersion matrix depends on the mean levels and shape parameters, and hence, the homogeneity of the dispersion matrices may not hold. For a stochastic  $p$ -vector  $\mathbf{X}$  following a multivariate normal distribution with mean vector  $\boldsymbol{\mu}$  and dispersion matrix  $\boldsymbol{\Sigma}$ , in a conventional setup, it is tacitly assumed that the dispersion matrix does not depend on the mean vector. This basic assumption may not generally hold in GCM where heteroscedasticity, possible collinearity and nonlinear relationship of dispersion matrix and mean vectors may mar the simplicity of the standard GCM analysis schemes. In the above cited PCDD/F model, we have nonnegative component variables which brings the relevance of *compositional data models*. If the  $p$  coordinate variables of  $\mathbf{X}$  are independent gamma variables with shape parameters  $\alpha_1, \dots, \alpha_p$  respectively (all positive), and a scale parameter  $\nu (> 0)$ , and if we define the proportion vector as  $\mathbf{Y} = (\mathbf{X}'\mathbf{1})^{-1}\mathbf{X}$ , then  $\mathbf{Y}$  has the Dirichlet distribution whose mean vector and the (singular) covariance matrix depend on the scale as well as shape parameters. This particular feature not only renders a singular covariance matrix but also invalidates the routine adaption of the so-called *principal component model* (PCM) or *canonical correlation analysis*. Bearing in mind such examples, we first consider a simple GCM and motivate more general ones arising thereof.

Let  $t_{i1}, \dots, t_{iq}$  be the time points for the  $i$ th subject and let

$$\mathbf{Y}_i = ((Y_{ijk}))_{j=1, \dots, p; k=1, \dots, q}, \quad i = 1, \dots, n, \quad (2.1)$$

where  $Y_{ijk} = Y_{ij}(t_{ik}), k = 1, \dots, q$ . In a balanced design,  $t_{ik} = t_k, \forall i = 1, \dots, n; k = 1, \dots, q$ . In a conventional parametric setup, it is typically assumed that  $\mathbf{Y}_i$  has a matrix-valued multi-normal distribution with unknown mean (matrix)  $\boldsymbol{\Theta}_i$  and unknown dispersion matrix  $\boldsymbol{\Gamma}$  (of order  $pq \times pq$ ), for  $i = 1, \dots, n$ . It is thus tacitly assumed that the dispersion matrix  $\boldsymbol{\Gamma}$  is common for all observations; this condition as noted earlier is violated for multivariate gamma and other non-normal distributions. In a conventional GCM setup,

$$\boldsymbol{\Theta}_i = \nu \mathbf{A}_i + \mathbf{B} \mathbf{C}_i, \quad i = 1, \dots, n, \quad (2.2)$$

where  $\mathbf{v}$  is a  $p \times k$  matrix of unknown (intercept) parameters,  $\mathbf{A}_i$  are matrices of known design variables (constants),  $\mathbf{C}_i$  is a  $r \times q$  matrix of known constants, and  $\mathbf{B} = ((\beta_{jl}))$  is a  $p \times r$  matrix of unknown parameters. In the  $k$  sample model, the  $j$ th column of  $\mathbf{A}_i$  is equal to  $\mathbf{1}$  and other columns are  $\mathbf{0}$ , according the  $i$ th observation belongs to the  $j$ th sample or not. In more complex designs, the choice of the known  $\mathbf{A}_i$  and  $k$  depends on the design matrix. Typically,  $r < q$  (to facilitate dimension reduction in a GCM setup). In the balanced design case, the  $\mathbf{C}_i$  are all equal. By (2.1) and (2.2), we have

$$\mathbf{Y}_i = \mathbf{v}\mathbf{A}_i + \mathbf{B}\mathbf{C}_i + \mathbf{E}_i, i = 1, \dots, n, \quad (2.3)$$

where the  $\mathbf{E}_i$  are independent and identically distributed random matrices with null mean and dispersion matrix  $\mathbf{\Gamma}$ . One may use the vec notation to convert these  $\mathbf{Y}_i$  into  $pq$ -vectors and then apply the usual MANOVA tools to draw statistical conclusions on  $\mathbf{B}$ . Typically,  $r$  is much smaller than  $q$ , and hence, the GCM approach works out well in having a more powerful statistical analysis when the postulated model in (2.2) holds. We refer to Rao (1965) and Gnanadesikan et al. (1971) for a systematic account of these developments.

In most of the fields of application, be it in biometry or clinical trials, system biology or bioinformatics, or the vast area of modern interdisciplinary research, usual MANOVA model assumptions are mostly untenable. In multivariate normal models, the covariance matrix is functionally independent of the mean vector, but this is not generally true for other multivariate distributions. We may refer to the fingerprint analysis problem where not only multi-normality assumption may be dubious but also homogeneity of the dispersion matrices is untenable. In the univariate setup, the Box and Cox (1964) transformation has been widely used to achieve approximate linearity of the model and improve normality approximation of such transformed variables. However, such nonlinear transformations while improving the normality approximation may adversely affect the underlying additivity structures as well as the homoscedasticity assumption. In some simple univariate models, Bartlett variance stabilizing transformations work out well. But such transformations are of not much help in stabilizing the dispersion matrices. For example, in multivariate gamma distributions, the dispersion matrix may functionally depend on the mean vector and hence the homoscedasticity condition may not hold. Because of these impasses, in multivariate GCM, in beyond parametrics approaches, some alternative analysis schemes are advocated; these are to be considered in the next section.

### 2.3 Beyond Parametrics Formulations

Whereas in parametric GCM, conventionally, it is assumed that the error distributions are (multi-)normal, in beyond parametrics, not only this multi-normality assumption is deemphasized but also other robustness issues are appropriately appraised. In this perspective, first consider the conventional parametric models.

In the balanced design case, all the  $\mathbf{C}_i$  are the same, and without loss of generality it may be assumed that they are of rank  $r(\leq q)$ . We make a similar assumption for the general unbalanced case as well. Let us then consider a set of  $q \times q$  matrices  $\mathbf{L}_i$  and partition it as

$$\mathbf{L}_i = (\mathbf{L}_{i1}, \mathbf{L}_{i2}), \quad i = 1, \dots, n, \quad (2.4)$$

where

$$\mathbf{L}_{i1} = \mathbf{C}_i'(\mathbf{C}_i\mathbf{C}_i')^{-1} \quad (2.5)$$

is of order  $q \times r$  and  $\mathbf{L}_{i2}$  is of order  $q \times (q - r)$  for  $i = 1, \dots, n$ . Let then

$$\mathbf{Z}_i = \mathbf{Y}_i\mathbf{L}_i = (\mathbf{Z}_{i1}, \mathbf{Z}_{i2}), \quad (2.6)$$

for  $i = 1, \dots, n$ . As in [Potthoff and Roy \(1964\)](#) and [Rao \(1965\)](#), we note that the  $\mathbf{B}\mathbf{C}_i\mathbf{L}_{i1} = \mathbf{B}$  while we choose the  $\mathbf{L}_{i2}$  in such a way that  $\mathbf{C}_i\mathbf{L}_{i2} = \mathbf{0}$  are null matrices of order  $r \times (q - r)$ . By (2.3) and (2.6), we have on writing  $\mathbf{E}_i^* = \mathbf{E}_i\mathbf{L}_i = (\mathbf{E}_{i1}^*, \mathbf{E}_{i2}^*)$  and  $\mathbf{A}_i^* = \mathbf{A}_i\mathbf{L}_i$ ,

$$\mathbf{Z}_i = \mathbf{v}\mathbf{A}_i^* + \mathbf{B}(\mathbf{I}_r, \mathbf{0}) + (\mathbf{E}_{i1}^*, \mathbf{E}_{i2}^*), \quad (2.7)$$

for  $i = 1, \dots, n$ . This perfectly fits in to a MANOCOVA model which under the multi-normality condition has been thoroughly studied in the literature ([Rao 1965](#), and others).

In a simple nonparametric approach ([Puri and Sen 1971](#)), it is assumed that the  $\mathbf{E}_i^*$  have jointly a  $pq$  variate continuous distribution, for all  $i = 1, \dots, n$ . Then linear rank statistics are constructed for each of the  $pq$  coordinates of the  $\mathbf{Z}_i$ ,  $1 \leq i \leq n$  of which  $pr$  statistics relate to the case where  $\mathbf{B}$  is present in addition to the partitioned part of  $\mathbf{v}\mathbf{A}_i^*$ , while the remaining  $p(q - r)$  linear rank statistics relate to the part where  $\mathbf{B}$  does not appear but the complementary part of  $\mathbf{v}\mathbf{A}_i^*$  appears. If the null hypothesis of relates to  $H_0 : \mathbf{B} = \mathbf{0}$ , i.e., no regression on the time points, then we can proceed in two ways. From the first part, we use the  $R$ -estimators of  $\mathbf{B}$  as in [Jurečková and Sen \(1996\)](#) and use a Wald-type test statistic. Alternatively, assuming  $\mathbf{B} = \mathbf{0}$ , we estimate  $\mathbf{v}$  from the entire set of linear rank statistics. In the second place, we align the  $\mathbf{Z}_i$  by using these  $R$ -estimators of  $\mathbf{v}$ , and on the aligned linear rank statistics for the  $p \times r$  sub-matrix, we construct an aligned rank MANOCOVA test statistic as in [Puri and Sen \(1971, 1985\)](#) wherein the  $p(q - r)$  linear rank statistics are treated as covariate statistics. Such tests are based on the [Chatterjee and Sen \(1964\)](#) rank permutation principle and are conditionally distribution-free under hypotheses of invariance. For large sample sizes, under these hypotheses of invariance, they have approximately chi-square distribution with appropriate degrees of freedom. Being based on the marginal ranks of the  $\mathbf{Z}_i$ ,  $i = 1, \dots, n$ , these tests are robust against plausible model departures.

A second approach is based on *rank* (or *R*-) *estimators* of  $\mathbf{v}$ ,  $\mathbf{B}$  from individual observations. Note that by (2.7), for each  $i (= 1, \dots, n)$ , we can obtain linear estimates of  $\mathbf{v}$  and  $\mathbf{B}$  from  $\mathbf{Z}_i$ . Given these  $n$  independent estimators of  $\mathbf{B}$ , it may be possible to use the weighted least squares methodology to obtain a combined sample estimator of  $\mathbf{B}$  and also, to estimate its dispersion matrix (of order  $pr \times pr$ ). Tests for suitable hypotheses on  $\mathbf{B}$  can be based on these estimates using the classical Wald statistics. Such tests are, however, generally not robust due to the poor robustness properties of the estimated dispersion matrix. On the other hand, based on these  $\mathbf{B}_i, i = 1, \dots, n$  (all of the order  $p \times r$ ), the general theory of *R*-estimators developed in detail in Jurečková and Sen (1996) can be incorporated to obtain robust estimators of  $\mathbf{B}$  and also to test for suitable hypotheses on  $\mathbf{B}$ .

We illustrate this methodology with a simple situation where the  $n$  observations can be regarded as the composite of  $k (\geq 2)$  samples of sizes  $n_1, \dots, n_k$ , respectively, so that  $n = \sum_{s=1}^k n_s$ . For an observation from the  $s$ th sample, referred to (2.2), the  $\mathbf{A}_i$  are equal to some  $\mathbf{A}_s$ , for  $s = 1, \dots, k$ . In this setup,  $t = p + k$  where the additional  $k$  relates to the individual population effects (vectors). As such, we may proceed to test for the null hypothesis that the  $k$  columns of  $\mathbf{v}$  are the same, treating  $\mathbf{B}$  as a nuisance parameter (matrix). Thus, using *R*-estimators of  $\mathbf{B}$ , we may use aligned rank test based on the  $k \times q$  linear rank statistics. We refer to Puri and Sen (1985) and omit the details. Alternatively, if the null hypothesis relates to  $\mathbf{B} = \mathbf{0}$  (i.e., no regression over time), treating  $\mathbf{v}$  as nuisance, then one can use aligned rank statistics. The dimension reduction (from  $pq$  to  $pr$  when  $r \ll q$ ) generally leads to increased statistical precision.

A further source of concern is the very basic assumption of linear models in GCM. It is not uncommon in toxicology and *physiologically based pharmacokinetics* (PBPK) models to have distinct nonlinear GCM where even if normality assumption can be approximately justified, the homogeneity of the error variances may not be tenable. Further, in PBPK and certain systems models, often (stochastic) differential equations (SPDE) are incorporated to explain better the underlying kinetics. In such a case, typically, the response pattern is nonlinear and multidimensional (viz., Mandal et al. 2012). Though such nonlinear systems are often approximated by linear ones (under the usual delta method), the reliability and validity of such linearization may be open to questions. In PBPK modeling, the composite response is the synergic and chain body resistance and metabolic changes through a number of organs along with their impact on the blood circulation system. As such, a multicomponent model is usually advocated, though in most of the mathematical modeling, for drawing statistical conclusions a simplistic approach is considered. A suitable growth model connecting the impact of these organs in relation to the body reaction to external stimulus will certainly be a better solution. The GCM approach has therefore a natural appeal in this context.

In HDLSS models, though it may be tempting to use *projection pursuit* for dimension reduction, its scope may be limited to distributions admitting linear structure so that the classical *principal component model* (PCM)-based statistical methodology is appropriate. In many contexts, this linear manifold is not tenable and hence this approach may not be adaptable either. In some simple models, this



approach was elaborated by Sen (1973). There is scope for expansion in more general GCM (viz., Preisser et al. 2011). They treated the *Gingivitis* problem resulting from absention from brushing or flossing of tooth. It had 7 time points consisting of the *induction* and *resolution phases* and 31 biomarkers on 22 subjects. Although the area under the individual subject and biomarker data were initially considered, the number of subjects (22) being smaller than the number of data points (186), conventional MANOVA tools were not adaptable. Moreover, the assumption of multinormality was difficult to justify. First, in the usual way in GCM, a dimension reduction was suggested, resulting in 4 response variables for each biomarker. Second, signed-rank tests were used in the univariate as well as multivariate setups (viz., Sen and Puri 1967), resulting in more robust procedures. Further, the Chen–Stein theorem (Chen 1975) was adapted to control Type I error and related measures. This produced a better inferential procedure.

In the environmental pollution problem (Chen et al. 2012), the GCM appeal was overwhelming. However, in that *fingerprint* analysis comparing the *poly-chlorinated dibenzo-p-dioxin* and *dibenzofuran* (PCDD/F) congener profile patterns of collected samples with those of potential dioxin emission sources, there were 17 PCDD/F congeners comprising a fingerprint which did not look like to have multi-normal distribution. They differ in their emission rate and exposure pattern, and hence, it was decided to have the proportion of these 17 compounds relative to their sum. This has of course given rise to a response vector on a 16-simplex (in a *compositional data model*) for which the variance-covariance matrix is intricately dependent on the mean vector and a multi-normal distribution is far from being tenable. Based on plausible assumptions, multivariate gamma-type distributions were thought to be more appropriate. That led to the so-called *Dirichlet* type distribution. A discouraging feature is that the dispersion matrix for this multivariate random vector depends not only on the mean vector but also on the shape parameters of the underlying gamma distributions. As such, conventional growth curve model-based analysis was not pursued. It turned out that the usual procedure based on multivariate ranks (Puri and Sen 1971) have much more robustness perspectives, thus performing better than the multi-normality-based likelihood ratio type tests. Thus, beyond parametrics seems to have a better appeal.

A third illustration relates to rank analysis of covariance (R-ANOCOVA) in some nonstandard data models (Sen et al. 2013). There, the R-ANOCOVA has been extended to a more general class of linear or nonlinear models (including measurement errors or misspecified models). This would make GCM for such more nonstandard cases manageable under beyond parametric schemes.

## 2.4 GCM in Health Care Studies

The development and management of a health care plan is a global problem, albeit drastically different from one country to another, or even within a country, from one region to another. A health care plan may either pertain to a general (overall)



population or certain subclass, termed a target population, demarcated by various socioeconomic or demographic features. Some of these features are qualitative or categorical while some others are quantitative. A health care plan is designed to assess the need for welfare or financial support for the target population for needy people when inflicted with certain type of disease or disorder. Of course, to run that, it needs sustainable funding through health insurance, government support, and other resources. Therefore, it is needed to have a complete inventory of diseases and disorders which are to be covered under the health care plan. It also needs to assess the available resources to cover the cost of providing health care for the target population. Such resources not only include the financial aspects but also the availability of ambulatory care personnel and facility, medical and paramedical personnel, general awareness of the population for some of the pertinent health hazards, lifestyle of the population concerned, and a thousand and one other associated factors, some of which may not even be properly ascribable. In this respect, the *quality of life* (QoL) and general attitude towards life have an important bearing too. There is naturally a temporal factor that relates to the adequacy or deterioration of a health care plan over time as is commonly perceived in many countries (Sen 2012). The growth of population susceptible to various diseases and disorders, by sector, spatial and temporal factors, the temporal change in the enrolment and compliance to a health care plan, growth of various burdens of disease (including virus mutations which may alter the nature of some of these diseases), (mal-)nutrition, poverty and affluence and other factors have significant bearing on such health care plans.

In most of the countries in the Western Europe, the social welfare system provides a significant support to available health care plans, although such schemes are difficult to implement in developing countries, especially, the over-populated ones, including the Indian subcontinent and China. The burden of population and the vast inequality of wealth and living standards create impasses for a unified health plan that could suit equally well the people from all walks of life. In capitalistic countries, USA is no exception, a health care insurance plan is not affordable across the various sectors of the population, and no wonder, still a big number of people are deprived of equitable health care insurance and facilities. The prevalence of certain diseases or disorders can impact a health care plan drastically. For example, diabetes is a major concern in India, China, and many other countries where consumption of carbohydrates is significantly higher. In this respect, the familial or genetic effects are very much noticeable. Breast cancer is more likely for daughters of mothers who has had such affliction. The fast changing lifestyle of a major sector of population be it in the West or in the third world countries is having an impactful aftermath on many cardio-vascular diseases. Hypertension is another big concern. On top of that, HIV (AIDS) has become a global threat, and all over the world, is having a huge toll in terms of mortality and morbidity. Arthritis and gout affect a significant part of any population, especially at golden ages. For cholera, quite prevalent in the coastal areas of the Indian subcontinent, it has been observed that there has been a mutation in the microbes which can now fight back many of the drugs (salines) which were quite effective a few years ago. Arsenic contamination of ground water is a major

health concern in a vast coastal area in the eastern part of India as well as the entire southern part of Bangladesh. Most of the working class people have their daily need of drinking, cooking, washing clothes and dishes too, and even bathing, intake a perceptible amount of arsenates which may not only have carcinogenic impact on their skin, hands and feet but also have impactful effect on their ingestion system. Combined with that improper disposition of human waste adds more misery to this contamination. Dementia, Parkinson's disease, and Alzheimer may be occurring at a higher incidence rate. Smoking and lung cancer may be good relation although they have not been linked causally. Environmental smoking effect is a significant health hazard, more so in metropolitan areas where automobile exhausts contribute liberally to this pollution. In any composite health care plan, the galaxy of diseases and disorders need to accounted for, although the prevalence pattern and relative cost for cure could be quite dissimilar.

The models discussed in Sect. 2.3 can be adapted for such health care plans. However, a much more complex and interacting modeling is necessary. First and foremost, let  $\mathbf{D}(s, t)$  stand for the galaxy of diseases or disorders, at time  $t$ ,  $t \in T$ ,  $s \in \mathcal{S}$ , which are to be covered under the plan. Here  $T$  stands for the time domain and  $\mathcal{S}$  stands for the domain of other spatial as well as explanatory variables. Secondly, some of the diseases or disorders are chronic and have long-range impact, while some others are relatively short duration with a (stochastically) much smaller in-disease period. Therefore, it may be better to include statistical information on the *time under treatment or service* of various diseases and disorders. In this respect, the age at onset, duration of the service and the level (ambulatory, in-house assistance or hospitalization) distributions are needed to be charted. The prevalence of various diseases or disorders may vary considerably across the demographic and economic strata of an overall population. The coverage of health plans may also depend on such socioeconomic strata. Thus, we will have a multi-dimensional stochastic vector, say,  $\mathbf{W}(s, t)$ ,  $t \in T$ ,  $s \in \mathcal{S}$  wherein all the other information are to be included as covariates. There may be a growth of prevalence of the diseases or disorders (in some cases the opposite way), and the information on available (para-)medical or clinical help and the associated cost analysis all are needed for an in-depth assessment. The (age-specific) *life expectancy* as further categorized by sex, ethnicity, and other demographic features can be viewed as a very useful piece of information in this respect. This needs development of a suitable index of *health status* of individuals covered under the health care system that can be incorporated in the formulation of a general *exposure risk* measure  $\mathbf{R}(s, t)$ ,  $t \in T$ ,  $s \in \mathcal{S}$  whose distribution over the target population constitutes an essential component of a stochastic modeling of the overall picture. Some other factors like most of the high-cost surgeries need to be attuned to a possible health plan in such a way that a complete coverage may push up the cost factor so much that on cost ground such plans may not be affordable for a greater part of the society. Therefore, sustainability and afford-ability issues are to be weighed in objectively so as to make a plan adaptable. That also needs statistical modeling.

No plans can be sustained without a complete provision of funding through health insurance, cost sharing by the patients and government or other funding. An accounting of the relative support, their potential change over time and their matching the cost of providing the health care service is therefore desirable before any undertaking can be planned. As such, we have a complex of variable, some being response variables while others as covariates or explanatory variables, and statistical appraisal of this picture is a prerequisite. This is needed to model a composite *cost-factor* analysis based on a stochastic time-dependent  $\mathbf{C}(s, t), s \in \mathcal{S}, t \in T$  which are to be attuned to the other stochastic matrices described before.

Statistically speaking, we need to have the collection of stochastic systems:

$$(\mathbf{D}(s, t), \mathbf{W}(s, t), \mathbf{R}(s, t), \mathbf{C}(s, t)), t \in T, s \in \mathcal{S}, \quad (2.8)$$

which are to be incorporated in to a growth model for a composite model. It is also necessary to account for  $U(s, t), t \in T, s \in \mathcal{S}$ , the cost for providing health care contrasted with the resources to match that factor. This is intricately related to fixation of the health insurance premiums, projection of clinical and medical personnel cost and revenue sharing from other sources. Even in USA and other developed countries in the West, the escalating health care cost is a nightmare for concerned administrations; the problem is undoubtedly much more complex in the Indian sub-continent and China. This is highly a nonlinear system, and routine use of standard MANOCOVA or GCM may be grossly inappropriate. It may be appealing to incorporate some SPDE (as in the PBPK models). However, given the usual assumptions of white noises following suitable Gaussian laws in such SPDE, it could be difficult to formulate computationally manageable methodological justifications of SPDE sans those Gaussian components.

An essential feature of these stochastic processes is that they are not stationary even in a very broad sense. Time dependence of not only the basic marginal functions but also their association structures may generally cause tremendous roadblocks to implement standard GCM models even in a component-wise formulation. Generally, these stochastic processes have some tendency to acquire some aggregative effects, resulting in usually nonlinear trends. Thus detrending is an essential task. In the presence of nonlinear trends, usual parametric models may not only be inadequate but also too irrelevant. Beyond parametrics approaches based on *wave-length* methodology and *nonparametric smoothing* are therefore advocated. That may invariably need relatively much larger sample size and could run into cost constraints. It seems that taking into account the basic extraneous factors a multidimensional, nonstationary, and non-Gaussian process with appropriate systematic factors (most relevant to the GCM) can only be done in a more nonparametric setup with adherence to local (sub- or semi-)martingale features may lead to more meaningful resolutions. The basic issue may be can there be sufficient statistical validation and interpretation of data collection and monitoring to induce the impact of GCM in this largely exploratory field?

## 2.5 Concluding Remarks

It is indeed a challenge, especially in the developing countries, to collect reliable data sets pertaining to the detailed statistical perspective as listed in the preceding section. In most of the cases, there may be data sets pertaining to marginal morbidity and mortality rates due to various (competing) causes such as the major diseases or disorders but not that much of their synergic effects, and on top of that, very little information on the health care facilities, insurance coverage, actual illness and disease-free state sojourns, cost of services and individual health insurance premium, etc. In health care and health services, especially for the senior people, composite impact of more than one disease or disorder needs to be investigated. This information can only be obtained through intensive sample surveys. The sampling frame, cost of sample survey, adequacy of sample size information, possible adjustment for non-responses, and the need for follow-up sampling, all are to be formulated in a sound statistical manner. Collecting the relevant information from census or official publications is likely to be grossly incomplete. In USA and some other countries, the Bureau of Census, regularly conducts sample surveys to update the census figures and collect some additional information. Still, they are not enough to chart out the whole complex of growth models presented in Sect. 2.4. In the Indian subcontinent, possibly the State Statistical Bureaus and the Central Statistical Organization can undertake a network of sampling scheme but would probably require statistical expertise to do it in depth and in a valid way to match the need of the general objectives of health care plans and health study protocols. There has been a sustained development of statistical thinking in public health (Sen and Rao 2000) but their adaption in health care system is one step further that requires immediate attention. My feeling is that this is a more complex problem beyond the reach of these organizations present state of activities. On top of that some other public health enterprises in India may not have the expertise and resources to undertake such schemes. It is my hope that given such exploratory studies, the implementation of actual health care plans will be facilitated. A much more detailed statistical study is indeed needed and intended in the near future.

## References

- Box, G.E.P. (1950). Problem in the analysis of growth and wear curves. *Biometrics*, 6, 362–389.
- Box, G.E.P., & Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series*, 26, 211–252.
- Chatterjee, S.K., & Sen, P.K. (1964). Nonparametric tests for the bivariate two-sample problem. *Calcutta Statistical Association Bulletin*, 22, 13–50.
- Chen, C.C., Sen, P.K., & Wu, K.-Y. (2012). Permutation tests for homogeneity of fingerprint patterns of dioxin congener profiles. *Environmetrics*, in press.
- Chen, L.H.Y. (1975). Poisson approximation for dependent trials. *Annals of Probability*, 3, 534–545.

- Cole, J.W.L., & Grizzle, J.E. (1966) Applications of multivariate analysis of variance to repeated measurements experiments. *Biometrics*, 22, 810–828.
- Geisser, S. (1970). Bayesian analysis of growth curves. *Sankhya, Ser. A*, 32, 53–64.
- Geisser, G. (1981). Growth curve analysis. In P.R. Krishnaiah (Ed.) *Handbook of Statistics, Vol. 1: Analysis of Variance* (pp. 89–115). Amsterdam: North Holland.
- Ghosh, M., Grizzle, J.E., & Sen, P.K. (1973). Nonparametric methods in longitudinal studies. *Journal of the American Statistical Association*, 68, 29–36.
- Gnanadesikan, R., Srivastava, J.N., Roy, S.N., Foulkes, E.B., & Lee, E.T. (1971). *Analysis and design of certain quantitative multiresponse experiments*. New York: Pergamon Press.
- Gompertz, B. (1825). *Philosophical Transactions of the Royal Society, Ser. A*, 115, 513–580.
- Grizzle, J.E., & Allen, D.M. (1969). Analysis of growth and dose-response curves. *Biometrics*, 25, 357–381.
- Johnson, N.L., & Kotz, S. (1970). *Distributions in statistics: continuous univariate distributions*. New York: Wiley.
- Jurečková, J., & sen, P.K. (1996). *Robust statistical procedures: asymptotics and interrelations*. New York: Wiley.
- Khatri, C.G. (1966). A note on a MANOVA model applied to problems in growth curves. *Annals of the Institute of Statistical Mathematics*, 18, 75–86.
- Mandal, S., Sen, P.K., & Peddada, S.D. (2012). Statistical inference for dynamic systems governed by differential equations with application to toxicology. (under preparation).
- Potthoff, R.F., & Roy, S.N. (1964). Generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51, 313–326.
- Preisner, J., Sen, P.K., & Offenbacher, S. (2011). Multiple hypothesis testing for experimental gingivitis based on Wilcoxon signed rank statistics. *Statistics in Biopharmaceutical Research*, 3, 372–384.
- Puri, M.L., & Sen, P.K. (1971). *Nonparametric methods in multivariate analysis*. New York: Wiley.
- Puri, M.L., & Sen, P.K. (1985). *Nonparametric methods in general linear models*. New York: Wiley.
- Rao, C.R. (1958). Comparison of growth curves. *Biometrics*, 14, 1–16.
- Rao, C.R. (1959). Some problems involving linear hypotheses in multivariate analysis. *Biometrika*, 46, 49–58.
- Rao, C.R. (1965). Theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52, 447–458.
- Sen, P.K. (1973). Some aspects of nonparametric procedures in multivariate statistical analysis. In D.G. Kabe, & R.P. Gupta (Eds.) *Multivariate statistical analysis* (pp. 231–240). Amsterdam: North Holland.
- Sen, P.K. (1985). Nonparametric procedures for some miscellaneous problems. In P.R. Krishnaiah, & P.K. Sen (Eds.) *Handbook of statistics, Vol. 4: Nonparametric methods* (pp. 699–739). Amsterdam: North Holland.
- Sen, P.K. (2006). Robust statistical procedures for high-dimensional data models with applications to genomics. *Austrian Journal of Statistics*, 35, 197–214.
- Sen, P.K. (2008). Kendall's tau in high-dimensional genomic parsimony. In *Institute of mathematical statistics, collection series, vol. 3* (pp. 251–266).
- Sen, P.K. (2012). Development and management of national health plans: Health economics and Statistical perspectives. In Y.P. Chaubey (Ed.) *Some topics on current issues in mathematical and statistical methods*. Singapore: World Scientific Press.
- Sen, P.K., Singer, J.M., & Pedroso de Lima, A.C. (2010). *Finite Sample to Asymptotic Methods in Statistics*. Cambridge University Press, New York.
- Sen, P.K., Jurečková, J., & Picek, J. (2013). Rank tests for corrupted linear models. *Journal of the Indian Statistical Association*, 51(1), in press.
- Sen, P.K., & Puri, M.L. (1967). On the theory of rank order tests for location in the multivariate one sample problem. *Annals of Mathematical Statistics*, 38, 1216–1228.
- Sen, P.K., & Rao, C.R. (Eds.) (2000). *Handbook of statistics, vol. 18: Bioenvironmental and public health sciences*. Amsterdam: North Holland.

- Sen, P.K., Tsai, M.-T., & Jou, Y.S. (2007). High-dimension low sample size perspectives in constrained statistical inference: The SARSCoV-2 genome in illustration. *Journal of the American Statistical Association*, 102, 685–694.
- Timm, N.H. (1981). Multivariate analysis of variance of repeated measurements. In P.R. Krishnaiah (Ed.) *Handbook of statistics, vol. 1: Analysis of variance* (pp. 41–87). Amsterdam: North-Holland.
- Zerbe, G.O., & Walker, S.H. (1977). A randomization test for comparison of groups of growth curves with different polynomial design matrices. *Biometrics*, 33, 653–659.

Advances in Growth Curve Models

Topics from the Indian Statistical Institute

Dasgupta, R. (Ed.)

2013, XIII, 270 p. 147 illus., 73 illus. in color., Hardcover

ISBN: 978-1-4614-6861-5