A Radically Different Method of Moments

William A. Gardner
Department of Electrical and Computer Engineering, University of California Davis, Orcid number: 0000-0003-3840-7191

January 1, 2024

Abstract

The four primary moment-based probability density estimation and associated moment-based parameter estimation methods are briefly described as background for the introduction of a new method. This new method is radically different in approach yet provides a solution that requires essentially the same information as the existing methods: (1) Model moments with known dependence on unknown parameters and (2) associated sample moments. However, the new method, unlike the classical method of moments and its generalized counterparts, requires only the solution of simultaneous linear equations. A theoretical comparison between the new and old methods is made, and reference is made to the Author's earlier work on analytical comparisons with Bayesian parameter estimation and decision for time series data. The next step required for finding this new method's place in practice among the present four primary methods of moments is an extensive comparison of the performance of all these methods applied to a diverse variety of multivariate data sets. This next step is beyond the scope of this theoretical paper, which demonstrates for the first time that the method from engineering work on inference for time series data in the 1970s is a genuine method of moments for multivariate data though this was not originally recognized.

Keywords: Parameter estimation, Multivariate Model Fitting, Methods of Moments, Bayesian Inference

1 Introduction

1.1 Methods of Moments for Probability Approximation and Estimation, and Parameter Estimation

There is a plethora of methods for approximating probability density functions (PDFs), using known moments of the unknown PDF, and estimating PDFs from observed/measured data, some of which are based on first approximating the PDFs. Here is a brief description of the primary (most salient) methods:

Method 1, Characteristic Function Method: The first k moments can be used to determine the first k derivatives of the characteristic function at zero: $\mathrm{E}[X^k] = (-i)^k[\mathrm{d}^k\varphi_X/\mathrm{d}X^k](0)$. So, the first n terms of the characteristic function's Taylor series expansion around zero is given in terms of the first n moments. By inverse Fourier transforming this approximating Taylor expansion, we obtain an approximating PDF. Knowing the truncated Taylor series approximant is less accurate the further away from the origin we look, it is advisable to remove the more inaccurate part of this approximant by windowing it to an interval about the origin. When this is done, the impact on the PDF approximant is a convolution with the Fourier transform of the window function. This smoothing operation limits the degree of resolution of this PDF approximant. This method can be generalized to joint moments and joint PDFs by using joint characteristic functions. Once the PDF approximant in terms of moments is obtained, we have a probabilistic model that can be fitted to the data by replacing the model moments with sample moments to obtain a PDF estimate.

Method 2, Provost's Methodology: The least squares fit of an nth-order polynomial to a PDF with compact support is given by a weighted sum of Legendre Polynomials of orders 1 through n in terms of the first n moments of the PDF. Then by replacing these moments in this PDF model with sample moments, we obtain a PDF estimate from the available data. This is just one of a class of PDF approximation methods based on orthogonal polynomials, including Laguerre, Jacobi, and Hermite, some of which apply to approximating PDFs with semi-infinite support. These methods are all encompassed by a unified PDF approximation/estimation methodology, by which the exact density function whose first n moments are known can be approximated by means of the product of an assumed base density function, whose parameters are determined by matching moments, and a polynomial of degree n, whose coefficients are obtained by making use of the method of moments as well, Provost (2005).

Method 3, Pearson Method of Moments (MoM): When a PDF associated with data is unknown or intractable, but it is practical to obtain expressions for its moments up to but not including the values of a finite set of parameters, these parameters can be estimated using the MoM. This does not go as far as estimating the PDF itself, as in Methods 1 and 2, but there are many applications for which it is sufficient to estimate such parameters. The classical MoM first used by Pearson for parameter estimation near the turn of the Nineteenth Century is conceptually simple but does not offer a recipe for solving the (possibly unsolvable) set of simultaneous nonlinear equations that arise when the parameter-dependent model moments are equated to corresponding sample moments, Pearson (1936).

Method 4, Generalized Method of Moments (GMM): The GMM performs a weighted least squares fit, w.r.t. unknown-parameter values, of model moments to sample moments, and includes the MoM and several variations on the MoM as special cases, Lindsay (2014).

Methods 3 and 4 are parameter estimation methods, and Methods 1 and 2 can be made parameter estimation methods by performing the PDF estimation for each value of an unknown parameter (single or multiple), and then using the resultant estimated likelihood function or ratio of PDFs in a standard maximum-likelihood or Bayesian approach. We can call these parameter estimation methods collectively, Bayesian-Like (or Likelihood-Like) Methods of Moments. These well-known primary methods are complemented by the following essentially unknown method.

Method 5, Bayesian-Like MoM (BL-MoM): There is an essentially-unknown alternative to the above primary methods that uses optimally weighted sample moments to estimate the ratio of a likelihood function (parameter-conditional PDF) to the marginal (unconditional) PDF (see the ratio in the RHS of the equation below), which is tantamount to estimating the probability of each parameter value θ, given the observed data x (see the LHS of the equation below). This follows from the standard relationship

$$p(\theta|x) = \frac{p(x|\theta)}{p(x)}\, p(\theta)$$

where p is used to represent either a PDF or a probability mass function (PMF), specified by the symbol used for the argument. It is generally the case in this article that the p's in the ratio are PDFs and the other two p's are both PDFs or both PMFs. In the above relationship, we have

$$p(x) = \int p(x|\theta)\, p(\theta)\, \mathrm{d}\theta$$

where $p(\theta)$ is a PDF, and we replace the integral with a discrete sum when $p(\theta)$ is a PMF. For example, in a statistical decision application, discrete values of θ can correspond to a discrete set of competing hypotheses. Once this estimated *a posteriori* probability density (or mass) function in the LHS of the above relation is obtained, it can be substituted into any Bayes risk performance functional which can then be solved for a Bayesian-like parameter estimate or decision. Because the *a posteriori probability* $p(\theta|x)$ equals the aforementioned ratio of PDFs times the *a priori* probability $p(\theta)$, this *a posteriori* probability estimate is generally known only to within a factor, which is the *a priori* probability. When this factor is unknown, useful estimates can be obtained by treating this a priori probability as uniform over some user-specified admissible region of parameter space. This essentially converts a maximum *a posteriori* inference problem into a maximum likelihood problem.

This method appeared in disguised form, unknowingly hiding its equivalence to a method of moments, in engineering journals about half a century ago and has apparently not been recognized in the statistics literature over the long ensuing period, making it an essentially unknown MoM in the statistics community. As a result, there has apparently been no progress made toward comparing this likelihood or a posteriori PDF/PMF estimation method with the above-mentioned plethora of known methods for parameter estimation. This comparison is a sensible thing to do, given that the solution provided by the BL-MOM requires the same information as that required by the Methods 1, 3, 4 and less for Method 2, namely, model moments depending on unknown parameters, and sample moments. The primary differences are that the BL-MoM solution: 1) specifies optimally weighted sample moments in place of traditional sample moments, 2) requires only solution of simultaneous linear equations, and 3) is not restricted to use of polynomials; it can use any specified functions in place of polynomials, and it requires knowledge of only 1st and 2nd order moments of these functions of the original observations.

In this article, a theoretical comparison is made between the Pearson MoM and this Bayesian-Like MoM (BL-MoM). Because comparisons of such methods based on applications to data depend heavily on the particular data, any type of useful data-based or application-based comparison of methods 1- 5 is a substantial undertaking: Conclusions should not be drawn without comparison of results for a variety of types of data—something the Author has not attempted and something that will likely require the combined efforts of multiple investigators over an extended period of time. However, a few analytical comparisons between the BL-MoM method and the unadulterated Bayesian method of parameter estimation were made when the BL-MoM method was first introduced by the Author, and these are reported in the engineering publications Gardner (1973, 1976a, 1976b, 1981) and the unpublished engineering PhD dissertation Poulsen (1977).These publications report on the time-series version of the general BL-MoM concept with a single sample path of a time series, not on the equivalent multivariate BL-MoM concept with multiple multi-variate data samples, as explained in this article.

The purpose of this article is to give this new method visibility, without which the needed comparative studies will never be made. The motivation for making the needed comparative studies is provided by the list of theoretical benefits of the BL-MoM compared with the classical MoM (Method 3), which appear to carry over to the GMM (Method 4).

By way of explanation for the expansive style of this article, rather than simply presenting a terse mathematical statement of the assumptions made (known model moments and corresponding (weighted) sample moments, with weights to be solved for) and the mathematical method proposed (minimize mean square error of a specified error measure to obtain a set of simultaneous linear equations to solve), a choice has been made to provide expansive discussion of the fundamental differences in concept and method, relative to all the known alternatives, such as Methods 1 – 4. This narrative aspect is merited by the break from the tradition of the primary methods reviewed above.

1.2 Introduction to the Pearson MoM and the Alternative BL-MoM

The traditional Method of Moments is said to have been introduced by Pearson (1936) around the turn of the 19th Century for parameter estimation and also by Chebyshev in 1887 (see Wikipedia (2022)) for proving the Central Limit Theorem. This method for parameter estimation, when applied to either multivariate or time-series data, consists of equating sample moments measured from the data with theoretical moments obtained from a probabilistic model of the data, and then solving for the unknown values of parameters in the theoretical moment expressions. The theoretical moments can be interpreted as unconditional moments depending on unknown parameters, or moments conditioned on unknown values of random parameters. The choice of interpretation has no impact on the method. However, the latter interpretation can be used to formulate a radically new approach to parameter estimation based on concepts from Bayesian Inference.

The classical MoM and the more sophisticated Methods 2 and 4 are a mainstay of parameter estimation for probabilistic models of data in econometrics,biostatistics, and other fields for which knowledge of the likelihood function is often unavailable or complexity of the known likelihood function prevents its use for maximum-likelihood estimation.

In the Method of Moments, one can use the mean and centralized moments or the mean and non-centralized moments, and one can use as many moments as there are unknowns, and there are other variations that have been devised. One such variation uses the fact that the theoretical moments for an M-th order autoregressive time series model satisfy a set of $M + 1$ linear equations in $M + 1$ unknowns involving only 2nd-order moments, the autoregressive model coefficients, and these equations can be solved for these unknown coefficients. This method is very common in data modeling and time-series prediction and associated studies of causality. More generally, however, the Method of Moments and its generalization, Methods 3 and 4, require the solution of nonlinear equations.

In the alternative approach, conditional moments are used and the objective is not to match theoretical moments to samples moments but rather to estimate the posterior PDF of the unknown parameters using the observed data, and then select the values of the conditioning parameters that maximize the estimated posterior PDF, thereby obtaining the maximally "probable" solution for the parameter values, where the quotation marks denote the fact that the PDF used is only an estimated PDF.

In this alternative method, one can use moments of a linear combination of any user-specified nonlinear functions of the observations; the equations to be solved are always linear, regardless of the particular functional dependence of the theoretical conditional moments on the parameters. But, when polynomial nonlinearities are used, the higher the order of the polynomials, the higher the order of the required moments of the original data. The posterior PDF estimator requires orders of moments up to twice the order of the polynomial.

The alternative method is optimal in the Bayesian sense that its estimate of the posterior PDF is a minimum mean-squared-error estimate subject to the selected constraint on the nonlinearities used.

This method requires calculating the sum of the moments, conditioned on the unknown parameter values, weighted by a prior PDF for those parameters. But one can always use a uniform PDF over a sufficiently large finite region of the domain if there is no knowledge of a prior PDF. This is tantamount to switching from a Bayesian approach to a Maximum-Likelihood (ML) approach, since the posterior PDF, as a function of the unknown parameters, is proportional to the Likelihood Function over the admissible region of the domain of the uniform prior PDF. However, the ML approach here is still only Maximum-"Likelihood" because the likelihood function used is only an estimated likelihood function.

This alternative method can use a single sample of a stationary (or cyclostationary) sequence of random variables, which favors applications to time-series analysis, or it can use multiple samples of one or more random variables.

To provide the reader with a look ahead at what this new method offers in comparison with the classical MoM, Table 1 summarizes the key characteristics of these two methods.

Expanding on Table 1, the Radically Different MoM produces more than just a single estimate for each parameter; it produces an estimate of the posterior PDF of the unknown parameters. This "PDF" can be used to calculate most "probable" estimates of the unknown parameters (their posterior "PDF" modes) which equal maximum-"Likelihood" estimates when the prior PDF is chosen to be uniform, or it can produce minimum-"mean"-squared-error estimates (their posterior "means") or minimum-"mean"-absolute-value of error estimates (their posterior "median") where, in all cases, the quotation marks denote the fact that the posterior PDF used in the estimates is the MMSE estimate of the true posterior PDF, subject to a user specified constraint on the structure (functional form) of that estimator's dependence on the observed data.

Prior to recognizing the applicability of this method to multivariate statistics, it was devised and used to design linear and quadratic communications receivers for digitally modulated signals and was found to have strong resemblances to statistically optimum receivers that are linear or quadratic under the simplifying assumption (for the optimum receiver) of additive Gaussian noise Gardner (1973)-Gardner (1976*a*).

The purpose of this article is to show that this method is promising for not only statisti-

Table 1: Advantages of the Radically Different MoM

#	Comparison Basis	Classical Method of Moments (MoM)	Radically Different MoM
1.	Functionals of Data Used	Uses sample moments	Uses linear combinations of any specified functionals of samples, including for example optimally weighted sample moments
2.	Model Used	Probabilistic conditional moments of data	Probabilistic conditional 1st and 2nd order moments of specified functionals of data
3.	Nature of Equations to be Solved	Generally nonlinear, except for Auto Regressive models	Always linear
4.	Breadth of Optimality Criteria	Produces a single solution with no optimality properties in general	Produces different solutions for different choices of optimality criteria
5.	Convergence to ML or Min-Risk Estimate	Has no general relationship to ML or Min-Risk estimates, except asymptotically as the amount of data grows without bound	Converges to Bayesian Min-Risk estimate (when prior PDF is known) or ML estimate of parameters as order of polynomial estimator increases for any fixed finite amount of data.
6.	Use of Prior Information	Does not use prior information	Uses prior information in an optimal manner when available
7.	Number of Samples Used for Each Random Variable	Typically, as many as possible	One (e.g., for long time series) or many
8.	Philosophy of Approach	Purely ad hoc	Disciplined application of Bayesian methodology
9.	Ability to Address Dynamic as Well as Static Models	Is not convenient for tracking rapid changes in parameters of interest	Is inherently amenable to tracking rapidly changing parameters

cal inference based on single samples of time series data but also for multivariate statistical inference based on multiple samples in general. The particular applications studied in the original papers Gardner (1973)-Gardner (1976*a*), which focused on data communications system design, demonstrated that the new method is analytically tractable and can indeed produce useful parameter estimates. But, the applicability to multivariate statistical inference using multiple samples has not been recognized or pursued. The theoretical advantages of the alternative method identified in Table 1 provides strong motivation for showing when the new method can be expected to be competitive with the primary methods $1-4$ for application to multivariate statistics in general

The Radically Different MoM is no less different from the much newer Generalized MoM introduced by L. P. Hanson in 1982 Wikipedia (2022) than it is from the Classical MoM from a century earlier, and the Generalized MoM is equivalent to several other methods introduced 20-to-30 years earlier Wikipedia (2022).

There is a limitation to the applicability of this new method. As stated in row 2b of Table 1, the user must be able to calculate the moments specified in row 2a of this table, as explicit functions of the unknown parameters. This typically requires a model in which the unknown parameters appear explicitly in the data model. For example, if the unknown parameter is the variance of one random variable for which multiple samples are available, the expected values of any nonlinear functionals to be used for estimation (e.g. the squaring function), conditioned on knowledge of the variance, are not defined; e.g., the fourth moment conditioned on knowledge of the variance is undefined, except in very unique cases like jointly Gaussian variables.

Because there is much distracting detail in the following presentation of the derivation of the new MoM, a streamlined summary of this derivation is provided in the Appendix.

Readers may prefer to read the Appendix first in order to know in advance where the derivation is heading as it proceeds through the following sections of this paper.

2 Classical method of moments

Assume we have R observations (samples) $\{x_{k,r} : \; k = 1, 2, \ldots, K, \; r = 1, 2, \ldots, R\}$ of K random variables $\{X_k\}$ and a model of the functional dependence of these random variables on Q unknown parameters $\boldsymbol{\theta} = \{\theta_q : q = 1, 2, \ldots, Q\}$ and L random variables $\boldsymbol{Z} = \{Z_l : l = 1, 2, \ldots, L\}$

$$X_k = f_k(\boldsymbol{\theta}; \boldsymbol{Z}) \tag{1}$$

We briefly consider three alternative assumptions and then down-select to one:

1. the joint PDF of $\{X_k\}$ is known, or

2. the joint PDF of $\{Z_l\}$ is known and this enables calculation of the joint PDF of $\{X_k\}$, or

3. a formulaic probabilistic model of $\boldsymbol{X}$ is available and enables the calculation of the joint moments of $\{X_k\}$

Cases 1) and 2) are quickly dispensed with here because resorting to ad hoc methods in these cases is generally not necessary unless issues of complexity arise. To be more specific, we consider the well-known relationship among prior (before data observation) and posterior (after data observation) probabilities

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \tag{2}$$

where $p(\boldsymbol{\theta})$ is the prior PDF of the parameters, $p(\boldsymbol{\theta}|\boldsymbol{x})$ is the posterior PDF, $p(\boldsymbol{x}|\boldsymbol{\theta})$ is the Likelihood Function and $p(\boldsymbol{x})$ is the unconditional data PDF, which can be decomposed into conditional PDFs (likelihood function values) as follows:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{3}$$

or, for discrete-valued parameters,

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

In the abbreviated notation used here, the particular PDF function is denoted by the symbol used for its arguments.

Given knowledge of the functions $p(\boldsymbol{\theta})$ and $p(\boldsymbol{x}|\boldsymbol{\theta})$, the other two functions in (2) can be calculated, and one can choose to use an ML estimate or any Minimum-Bayes'-Risk estimate of the parameter vector $\boldsymbol{\theta}$.

Consequently, resorting to the ad hoc MoM is generally not necessary for parameter estimation unless these functions are not known or are exceedingly difficult to calculate, particularly the Likelihood Function.

For case 3), the classical MoM for estimating the values of $\{\theta_q\}$ is to:

(1) equate $M \geq Q$ calculated joint probabilistic moments of $\{X_k\}$ to the M corresponding sample moments of $\{x_{k,r} : k = 1, 2, \ldots, K; r = 1, 2, \ldots, R\}$. For example, some subset M of the $(K^2 + 1)/2$ unique moments from the set of K^2 2nd-order moments can be used:

$$\mathbb{E}\{X_j X_k|\boldsymbol{\theta}\} = \frac{1}{R}\sum_{r=1}^{R} x_{j,r}x_{k,r} \qquad \text{for } j, k = 1, 2, \ldots, K \tag{4}$$

Then,

(2) try to solve this set of M simultaneous equations.

3 Radically different method of moments

In preparation for introducing the alternative MoM, we briefly expand the above discussion of Cases 1) and 2). It follows from (2) that any difference between the ML estimate and the MAP estimate is completely determined by the prior PDF. In the event that the prior PDF is uniform over the region where the likelihood function reaches its maximum value, then the ML and MAP estimates are equal. In situations where knowledge of a non-uniform prior PDF is not available, it is common to assume it is uniform over a sufficiently large finite region A in the prior-PDF domain, Q-dimensional Euclidean space:

$$p(\theta) = \begin{cases} \frac{1}{|A|}, & \boldsymbol{\theta} \in A \\ 0, & \boldsymbol{\theta} \notin A \end{cases}$$

where $|A|$ denotes the volume of A. In this case, (2) reduces to

$$p(\theta|\boldsymbol{x}) = \begin{cases} \dfrac{p(\boldsymbol{x}|\boldsymbol{\theta})}{|A|p(\boldsymbol{\theta})}, & \boldsymbol{\theta} \in A \\ 0, & \boldsymbol{\theta} \notin A \end{cases}$$

and (3) reduces to

$$p(\boldsymbol{x}) = \frac{1}{|A|} \int_A p(\boldsymbol{x}|\boldsymbol{\theta}) \mathrm{d}\theta$$

It follows that the maxima of the likelihood function and the posterior PDF coincide, and these two alternative methods become one and the same:

$$\underset{\boldsymbol{\theta} \in A}{\operatorname{argmax}} \{p(\boldsymbol{\theta}|\boldsymbol{x})\} = \underset{\boldsymbol{\theta} \in A}{\operatorname{argmax}} \{p(\boldsymbol{x}|\boldsymbol{\theta})\}$$

Unfortunately, whether or not the prior PDF is known, if either the likelihood function or the posterior probability is unknown, neither the ML nor Min-Risk methods can be used, and the MoM or one of the more sophisticated Methods 2 and 4 are likely to be resorted to. However, the interpretation of $\boldsymbol{\theta}$ as a vector of random variables instead of unknown constants enables a radically different alternative to the MoM to be derived. In the original formulation of this method for time series analysis, the name *Structurally Constrained Bayesian Methodology* (SCBM) Gardner (1976*b*) was introduced. This descriptive name is also appropriate in the more general setting of multivariate statistics based on multiple samples; however, it has the disadvantage of suggesting that a full probabilistic model for the data is available, as it must be in classical Bayesian statistics. For this reason, the alternative name *Structurally Constrained Bayesian Method of Moments* is suggested here. The same acronym can be used. The substantive advantages of the SCBM over the classical MoM are described in Table 1 (entries 1, 3-6, 8, 9 in Table 1).

In the SCBM, knowledge of the posterior PDF required by the MAP and other Bayesian methods is replaced with the requirement of knowledge of moments of $\boldsymbol{X}$ conditioned on $\boldsymbol{\theta}$, as in the classical MoM. Such moments can often, in practice, be calculated from the model (1), even when the posterior PDF and the likelihood function cannot be calculated.

The SCBM specifies (1) a constraint that the estimator be confined to some linear space derived from the observations $\boldsymbol{X}$, and (2) a performance criterion for optimizing an estimate of the posterior PDF.

The following discussion explains the extension of the original work on the SCBM to multivariate statistics for which the classical MoM was devised. This discussion is not a necessary part of the Radically Different MoM when applied to the classical MoM data model. But it does lead to an understanding of why the SCBM is able to address dynamic as well as static MOM problems and produce tracking parameter estimates.

> *When more than one sample of the vector of observed random variables is available, say R as in* (1)*), each sample can be interpreted as originating from a distinct K-dimensional vector $\boldsymbol{X}_r$, all R of which are identically distributed and are concatenated to form the composite RK-dimensional column-vector of observations*

$\boldsymbol{X} = [\boldsymbol{X}_1^T \boldsymbol{X}_2^T \ldots \boldsymbol{X}_R^T]$. *In this case, each* $\boldsymbol{X}_r$ *can contain as few as* $K = 1$ *random variable,* X_r. *This enables the SCBM to accommodate scalar-valued time-series of observations — for which the* r*-th observed random variable* X_r *is the* r*-th time sample of a scalar-valued stochastic process — as well as the classical MoM setup involving multiple (R) statistical samples* $\{\boldsymbol{x}_r\}_{r=1}^{r=R}$ *of a single* K*-dimensional random vector* $\boldsymbol{X}$ *of observations, each sample vector* $\boldsymbol{\theta}$ *depending on the same unknown parameter vector in which time may play no role. The SCBM was originally proposed for time-series analysis for communications systems, i.e., for statistical signal processing. But, with this simple device of reindexing and re-interpreting, it becomes apparent that the methodology applies as well to the classic multivariate statistics problem for which the MoM was created.*

The stochastic process model with a single scalar-valued sample path of length KR *that is equivalent to the* K*-variate model with* R *sample vectors has a special temporal structure because the sequence of* RK *times samples has a block structure in which the joint PDF of any subset of time samples depends periodically on the time shift parameter with a period of* K. *That is, the process is Cyclostationary Gardner (2022), Gardner et al. (2006). And this is precisely the type of stochastic process for which the SCBM was originally created Gardner (1973)-Gardner (1976*a*). However, in the models originally addressed with the SCBM,–a sequence of* R *random* Q*-vector parameters* $\{\boldsymbol{\Theta}_r\}$*– is stationary whereas, for the classical MoM model, reinterpreted as a scalar-valued time series model, the* Q*-vector is fixed from one period to another* $\{\boldsymbol{\theta}_r = \boldsymbol{\theta}\}$ *and is not treated as a realization of a random vector. Therefore, instead of estimating a time-sequence of* Q*-vectors, as in the typical communications system application, there is only one* Q*-vector for all blocks of* K *time samples.*

*However, this reveals that the SCBM allows for the MoM data model to be generalized to allow for some evolution of parameter values as more samples are collected. This would be done by allowing the parameter vector to become dependent on the data-block index r so that the otherwise Static SCBM-based MoM becomes Dynamic and tracks evolving parameters. The case in which such changes occur slowly is accommodated by including high correlation in the stationary sequence of random vectors. For the other extreme of maximally rapid changes in the parameter vector, the parameter sequence can be modeled as independent and identically distributed. For this Dynamic MoM problem, the original time-series formulation Gardner (1973) and Gardner (1976*b*) is preferred to the classical MoM problem formulation.*

3.1 Structural Constraint

For each value of $\boldsymbol{\theta}$, the posterior probability estimator $\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X})$ must be some linear functional of some set of specified nonlinear functions $\{g_j(\boldsymbol{X})\}$ of the random variables $\boldsymbol{X}$ modeling the observations. A sufficiently general linear functional for many applications has the form

$$\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X}) = \sum_j H_j(\boldsymbol{\theta}) \cdot [g_j(\boldsymbol{X})] \tag{5}$$

The set of values of $\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X})$ for each specified value of $\boldsymbol{\theta}$ generated by all component linear functionals $\{H_j(\boldsymbol{\theta})\}$ of the specified set of nonlinear functions $\{g_j(\cdot)\}$ of $\boldsymbol{X}$ is a linear vector space Λ of random variables. For example, one can choose

$$\begin{aligned} g_1(\boldsymbol{X}) &= \boldsymbol{X} \\ g_2(\boldsymbol{X}) &= \boldsymbol{X}\boldsymbol{X}^T \end{aligned} \tag{6}$$

in which case (5) reduces to

$$\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X}) = \sum_k h_k(\boldsymbol{\theta}) X_k + \sum_{k,l} h_{k,l}(\boldsymbol{\theta}) X_k X_l \tag{7}$$

which is a multivariate polynomial of order 2. Of course, this linear-plus-quadratic form is easily generalized to higher order polynomials. In (7), $\{h_k(\boldsymbol{\theta})\}$ is a representation (kernel) of the functional $H_1(\boldsymbol{\theta})$ and $\{h_{k,l}(\boldsymbol{\theta})\}$ is a representation of the functional $H_2(\boldsymbol{\theta})$. In general, the functions $\{g_j(\boldsymbol{X})\}$ are each tensors of various dimensions, as illustrated in (6), and the functionals $\{H_j(\boldsymbol{\theta})\}$ each map these tensors into scalars.

For applications in which the observed data is a continuous-time stochastic process, $\{X(t) : t \in \{a, b\} \subset (-\infty, +\infty)\}$, (7) becomes

$$\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X}) = \int_a^b h(t; \boldsymbol{\theta})X(t)\mathrm{d}t + \int_a^b \int_a^b h(t, u; \boldsymbol{\theta})X(t)X(u)\mathrm{d}t\mathrm{d}u \qquad (8)$$

which is a 2^{nd}-order Volterra-like counterpart of a 2^{nd}-order polynomial. In this case, $\{g_j(\boldsymbol{X})\}$ and $\{H_j(\boldsymbol{\theta})\}$ are continuous counterparts of the discrete tensors and functionals described above.

This choice to constrain the estimator to be in a specified linear vector space facilitates the analytical optimization of the estimator.

> In the above example, (6), the solution uses non-centralized moments. A recommended alternative is to replace $\boldsymbol{X}$ with $\overline{\boldsymbol{X}} = \boldsymbol{X} - \mathbb{E}\{\boldsymbol{X}\}$. Also, the term $g_0(\overline{X}) = 1$ for which $H_0(\boldsymbol{\theta})\cdot g_0(\overline{\boldsymbol{X}}) = h(\boldsymbol{\theta})$, a constant scalar, can be included in (6). This adds to the RHS of (7) and (8) the constant term $h_0(\boldsymbol{\theta})$. These modifications are illustrated in Gardner (1976*a*), and they are made in some of the examples below. In addition, the quantity $p(\boldsymbol{\theta}|\overline{X})$ can be replaced with $p(\boldsymbol{\theta}|\overline{X}) - \mathbb{E}\{p(\boldsymbol{\theta}|\overline{X})\} = p(\boldsymbol{\theta}|\overline{X}) - p(\boldsymbol{\theta})$, which is done in the examples below. Observe that $p(\boldsymbol{\theta}|\overline{\boldsymbol{X}}) = p(\boldsymbol{\theta}|\boldsymbol{X})$.

This completes the explanation of the types of structural constraints imposed by the SCBM method. We now move on to a description of the optimality criterion.

3.2 Optimality Criterion

The performance criterion for optimizing the structurally constrained estimator of the posterior PDF arises from selecting squared error as a cost function. Then the Bayes Risk to be minimized, which is the expected value of the cost, is the Mean Squared Error (MSE):

$$\text{MSE} = \mathbb{E}\{[\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X}) - p(\boldsymbol{\theta}|\boldsymbol{X})]^2\} \tag{9}$$

This may seem strange at first glance because probabilities are not random variables. However, when a random variable $\boldsymbol{X}$ is substituted in place of a sample observation $\boldsymbol{x}$, inside the function $p(\boldsymbol{\theta}|\cdot)$, the function value becomes a random variable. For the parameter estimation problems of interest here, we have a set of multiple random variables indexed by the parameter vector $\boldsymbol{\theta}$.

The optimization problem before us is to find the estimator $\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X})$ for each value of $\boldsymbol{\theta}$ that minimizes the above MSE subject to the constraint that the random variable $\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X})$ is contained in the specified linear vector space Λ of all admissible estimators, which we denote by $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{X})$. The solution to this optimization problem is well-known to be the orthogonal projection of the vector $p(\boldsymbol{\theta}|\boldsymbol{X})$, generally outside of Λ, onto the hyperplane Λ contained in the linear space of *all* admissible functionals of the observables.

A technical detail here is that, in order to apply the classical orthogonal projection theorem, the linear space must be an inner-product space, and this in turn requires that the vectors in the space all have finite norms; in this application, this means the probability model of the nonlinearly transformed observed random variables $\{g_j(\boldsymbol{X})\}$ must have finite-mean-squared values. This puts constraints on both the nonlinearities used, $\{g_j(\cdot)\}$, and the probabilistic model of the original observations $\{p(\boldsymbol{X}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in A\}$. These constraints are to be expected: one cannot use mean-squared error if random variables of interest do not have finite mean-squared values. Nevertheless, even if $\boldsymbol{X}$ does not have finite-mean-squared values, the nonlinearities $\{g_j(\cdot)\}$ can be chosen such that $\{g_j(\boldsymbol{X})\}$ do have finite mean-squared values.

3.3 Solution for Optimum Posterior PDF Estimate

The necessary and sufficient condition that characterizes the orthogonal projection solution described above is the following Orthogonality Condition:

$$\mathbb{E}\{[\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X}) - p(\boldsymbol{\theta}|\boldsymbol{X})]\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{X})\} = 0 \qquad \forall \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{X}) \in \Lambda \tag{10}$$

By using the estimator characterization (5), this condition can be re-expressed as

$$\mathbb{E}\left\{\left[\sum_j H_j(\boldsymbol{\theta}) \cdot [g_j(\boldsymbol{X})] - p(\boldsymbol{\theta}|\boldsymbol{X})\right] g_k(\boldsymbol{X})\right\} = 0 \ \forall\{k\} \tag{11}$$

which is equivalent to

$$\begin{aligned} &\textstyle\sum_j H_j(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[g_j(\boldsymbol{X})]g_k(\boldsymbol{X})\right\} \\ &= \mathbb{E}\left\{p(\boldsymbol{\theta}|\boldsymbol{X})g_k(\boldsymbol{X})\right\} \ \ \forall\{k\} \end{aligned} \tag{12}$$

where (because $\mathbb{E}\{\cdot\}$ is linear) $H_j(\boldsymbol{\theta})$ operates on the quantity in square brackets *after* the expectation is executed, as made more explicit below in (19).

As a final step in simplifying these equations, we use the *magic relationship*:

$$\begin{aligned} \mathbb{E}\left\{p(\boldsymbol{\theta}|\boldsymbol{X})g_k(\boldsymbol{X})\right\} &= \mathbb{E}\left\{\frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})g_k(\boldsymbol{X})}{p(\boldsymbol{X})}\right\} \\ &= \int \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})g_k(\boldsymbol{x})}{p(\boldsymbol{x})}p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int p(\boldsymbol{x}|\boldsymbol{\theta})g_k(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \ p(\boldsymbol{\theta}) \\ &= \mathbb{E}\left\{g_k(\boldsymbol{X}|\boldsymbol{\theta})\right\} \ p(\boldsymbol{\theta}) \end{aligned} \tag{13}$$

in which the unknown posterior PDF vanishes and the assumed-known prior PDF (possibly a uniform PDF when it is not known) appears. Substituting (13) into (12) produces

$$\sum_j H_j(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[g_j(\boldsymbol{X})]g_k(\boldsymbol{X})\right\} = \mathbb{E}\left\{g_k(\boldsymbol{X})|\boldsymbol{\theta}\right\} \ p(\boldsymbol{\theta}) \ \forall\{k\} \tag{14}$$

The unconditional moments in the left member of this set of linear equations can be re-expressed in terms of conditional moments as follows:

$$\begin{gathered}\sum_j H_j(\boldsymbol{\theta}) \cdot \int \mathbb{E}\left\{[g_j(\boldsymbol{X})] g_k(\boldsymbol{X})|\widetilde{\boldsymbol{\theta}}\right\} p(\widetilde{\boldsymbol{\theta}}) \mathrm{d}\widetilde{\boldsymbol{\theta}} \\ = \mathbb{E}\left\{g_k(\boldsymbol{X})|\boldsymbol{\theta}\right\}\; p(\boldsymbol{\theta})\; \forall\{k\}\end{gathered} \tag{15}$$

This is a set of linear equations in the unknown linear functionals $\{H_j(\boldsymbol{\theta})\}$. Thus, regardless of the nonlinear functions (tensors) selected in the structural constraint, the equations to be solved are always linear. In addition, when $\{g_j(\boldsymbol{X})\}$ are comprised of homogeneous polynomials, as in the examples above, the linear equations are fully specified by moments of $\boldsymbol{X}$ conditioned on the parameters $\boldsymbol{\theta}$.

> *This latter observation reveals that the SCBM is a method of moments in the special case for which polynomial nonlinearities $\{g_j(\cdot)\}$ are selected, and the radical difference between the details of the SCBM and those of the classical MoM and more general Methods 1, 2, 4 explains why this method is called a radically different MoM.*
>
> *Another interesting observation that can be made from* (15) *is the fact that by using the SCBM, the otherwise required knowledge of the likelihood function—the data PDF conditioned on the parameter values—is replaced with the required knowledge of the 1^{st} and 2^{nd} order moments of prescribed nonlinear functions of the data which, for up-to-nth-order polynomial functions of the data, are 1^{st} through $2n^{th}$ order moments of the data. So, the required knowledge of likelihood functions—the data PDFs conditioned on parameter values—is replaced with the required knowledge of a finite set of data moments conditioned on parameter values. This is, after all, the essence of methods of moments.*

As mentioned in a previous section, when the prior PDF is known, this is additional information the SCBM uses, which the classical MoM does not use. And, in addition, when the prior PDF is not known, it can be assumed to be uniform over a user specified region of parameter space which the user can specify according to any relevant prior information.

To illustrate the design equation whose solution fully specifies the posterior PDF estimate for each set of parameter values $\boldsymbol{\theta}$ of interest, we consider here the example (6), modified by inclusion of the $k = 0$ term and replacement of $\boldsymbol{X}$ by $\overline{\boldsymbol{X}}$ as discussed in Section 3.1. Using (12), modified by replacement of $p(\boldsymbol{\theta}|\boldsymbol{X})$ with $p(\boldsymbol{\theta}|\boldsymbol{X}) - p(\boldsymbol{\theta})$, we obtain

$$\begin{gathered} H_0(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[g_0(\overline{\boldsymbol{X}})]g_k(\overline{\boldsymbol{X}})\right\} + H_1(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[g_1(\overline{\boldsymbol{X}})]g_k(\overline{\boldsymbol{X}})\right\} \\ +H_2(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[g_2(\overline{\boldsymbol{X}})]g_k(\overline{\boldsymbol{X}})\right\} \\ = \left(\mathbb{E}\{g_k(\overline{\boldsymbol{X}})|\boldsymbol{\theta}\} - \mathbb{E}\{g_k(\overline{\boldsymbol{X}})\}\right) p(\boldsymbol{\theta}) \text{ for } k = 0, 1, 2 \end{gathered} \tag{16}$$

which, using modified (6), is equivalent to

$$\begin{gathered} H_0(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[1]g_k(\overline{\boldsymbol{X}})\right\} + H_1(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[\overline{\boldsymbol{X}}]g_k(\overline{\boldsymbol{X}})\right\} \\ +H_2(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T]g_k(\overline{\boldsymbol{X}})\right\} \\ = \left(\mathbb{E}\{g_k(\overline{\boldsymbol{X}})|\boldsymbol{\theta}\} - \mathbb{E}\{g_k(\overline{\boldsymbol{X}})\}\right) p(\boldsymbol{\theta}) \text{ for } k = 0, 1, 2 \end{gathered} \tag{17}$$

which can be more explicitly expressed as

$$\begin{gathered} H_0(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[1]\right\} + H_1(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[\overline{\boldsymbol{X}}]\right\} \\ +H_2(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T]\right\} = \left(\mathbb{E}\{1|\boldsymbol{\theta}\} - \mathbb{E}\{1\}\right) p(\boldsymbol{\theta}) \\[2ex] H_0(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[1]\overline{\boldsymbol{X}}\right\} + H_1(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[\overline{\boldsymbol{X}}]\overline{\boldsymbol{X}}\right\} \\ +H_2(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T]\overline{\boldsymbol{X}}\right\} = \left(\mathbb{E}\{\overline{\boldsymbol{X}}|\boldsymbol{\theta}\} - \mathbb{E}\{\overline{\boldsymbol{X}}\}\right) p(\boldsymbol{\theta}) \\[2ex] H_0(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[1]\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T\right\} + H_1(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[\overline{\boldsymbol{X}}]\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T\right\} \\ +H_2(\boldsymbol{\theta}) \cdot \mathbb{E}\left\{[\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T]\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T\right\} \\ = \left(\mathbb{E}\{\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T|\boldsymbol{\theta}\} - \mathbb{E}\{\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T\}\right) p(\boldsymbol{\theta}) \end{gathered} \tag{18}$$

Using (7), we can now re-express the above set of linear equations more explicitly as follows:

$$
\begin{aligned}
& h(\boldsymbol{\theta}) + \textstyle\sum_{k,l} h_{k,l}(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_l\} = 0 \\
& \textstyle\sum_k h_k(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_j\} + \sum_{k,l} h_{k,l}(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_l\overline{X}_j\} \\
& \qquad = \mathbb{E}\{\overline{X}_j|\boldsymbol{\theta}\}p(\boldsymbol{\theta}) \quad \forall j \\
& h(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_j\overline{X}_i\} + \textstyle\sum_k h_k(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_j\overline{X}_i\} \\
& \qquad + \textstyle\sum_{k,l} h_{k,l}(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_l\overline{X}_j\overline{X}_i\} \\
& \qquad = \left(\mathbb{E}\{\overline{X}_j\overline{X}_i|\boldsymbol{\theta}\} - \mathbb{E}\{\overline{X}_j\overline{X}_i\}\right) p(\boldsymbol{\theta}) \quad \forall i, j
\end{aligned}
\tag{19}
$$

The unconditional moments in (19) can be characterized in terms of conditional moments using (3) as follows for example:

$$
\mathbb{E}\{X_kX_lX_jX_i\} = \int \mathbb{E}\{X_kX_lX_jX_i|\boldsymbol{\theta}\}p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{20}
$$

If no prior information is available for specifying a prior PDF, a uniform PDF can be used to obtain

$$
\int \mathbb{E}\{X_kX_lX_jX_i|\boldsymbol{\theta}\}p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \frac{1}{|A|}\int_A \mathbb{E}\{X_kX_lX_jX_i|\boldsymbol{\theta}\}\mathrm{d}\boldsymbol{\theta} \tag{21}
$$

The solutions to (19) are used in the estimator formula (7), modified by replacement of $\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X})$ with $\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X}) - p(\boldsymbol{\theta})$.

Example 1: Linear Estimator In the case of a constant-plus-linearly-constrained estimator of the posterior PDF, the design equation (18) reduces to

$$
\begin{aligned}
& h(\boldsymbol{\theta}) + \textstyle\sum_k h_k(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\} = 0 \\
& h(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_j\} + \textstyle\sum_k h_k(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_j\} = \mathbb{E}\{\overline{X}_j|\boldsymbol{\theta}\}p(\boldsymbol{\theta}) \quad \forall j
\end{aligned}
\tag{22}
$$

which has the explicit solution

$$
\begin{aligned}
& h(\boldsymbol{\theta}) = 0 \\
& \boldsymbol{h}(\boldsymbol{\theta}) = \left[\mathbb{E}\{\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T\}\right]^{-1}\mathbb{E}\{\overline{\boldsymbol{X}}|\boldsymbol{\theta}\}p(\boldsymbol{\theta})
\end{aligned}
\tag{23}
$$

Consequently, the posterior PDF estimator, given by the modified version of (7), reduces to

$$\widehat{p}(\boldsymbol{\theta}|\overline{\boldsymbol{X}}) = p(\boldsymbol{\theta}) + \boldsymbol{h}^T(\boldsymbol{\theta})\overline{\boldsymbol{X}} \tag{24}$$

Substituting (23) into (24) yields

$$\widehat{p}(\boldsymbol{\theta}|\boldsymbol{X}) = p(\boldsymbol{\theta}) \left(1 + \mathbb{E}\{\overline{\boldsymbol{X}}^T|\boldsymbol{\theta}\} \left[\mathbb{E}\{\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T\}\right]^{-1} \overline{\boldsymbol{X}}\right) \tag{25}$$

Denoting the square root of the inverse of the covariance matrix in (25) by $\boldsymbol{W}$, and denoting the centered decorrelated vector of observations by $\boldsymbol{Y} = \boldsymbol{W}\overline{\boldsymbol{X}}$, we can re-express (25) for a particular sample of data $\boldsymbol{x}$ as

$$\widehat{p}(\boldsymbol{\theta}|\boldsymbol{x}) = p(\boldsymbol{\theta}) \left(1 + \mathbb{E}\{\boldsymbol{Y}^T|\boldsymbol{\theta}\}\boldsymbol{y}\right) \tag{26}$$

> *In words, the constant plus linear estimator probabilistically centers the data and probabilistically decorrelates it and then empirically correlates it with its probabilistic mean conditioned on the parameter vector, multiplies this by the prior PDF and adds this to the prior PDF.*

For a pseudo-MAP estimator of $\boldsymbol{\theta}$, (25) yields

$$\widehat{\boldsymbol{\theta}}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{p(\boldsymbol{\theta}) \left(1 + \mathbb{E}\{\overline{\boldsymbol{X}}^T|\boldsymbol{\theta}\} \left[\mathbb{E}\{\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T\}\right]^{-1} \overline{\boldsymbol{x}}\right)\right\} \tag{27}$$

which can be re-expressed using (26) as

$$\widehat{\boldsymbol{\theta}}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{p(\boldsymbol{\theta}) \left(1 + \mathbb{E}\{\boldsymbol{Y}^T|\boldsymbol{\theta}\}\boldsymbol{y}\right)\right\} \tag{28}$$

Similarly, the pseudo-MMSE estimator, which is the pseudo-posterior mean, is given by

$$\widehat{\boldsymbol{\theta}}_{\text{MMSE}} = \int p(\boldsymbol{\theta}) \left(1 + \mathbb{E}\{\boldsymbol{Y}^T|\boldsymbol{\theta}\}\boldsymbol{y}\right) \boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta} \tag{29}$$

where $\mathrm{d}\boldsymbol{\theta} = \mathrm{d}\theta_1\mathrm{d}\theta_2 \ldots \mathrm{d}\theta_Q$.

Example 2: Linear Plus Quadratic Estimator In the case of a constant-plus-linearly-plus-quadratically constrained estimator of the posterior PDF, for the special case in which the odd-order unconditional moments of the observed data are zero, the design equations (19) reduce to the following equations:

$$
\begin{gathered}
h(\boldsymbol{\theta}) + \textstyle\sum_{k,l} h_{k,l}(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_l\} = 0 \\
\textstyle\sum_k h_k(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_j\} = \mathbb{E}\{\overline{X}_j|\boldsymbol{\theta}\}p(\boldsymbol{\theta}) \ \ \forall j \\
h(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_j\overline{X}_i\} + \textstyle\sum_{k,l} h_{k,l}(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_l\overline{X}_j\overline{X}_i\} \\
= \left(\mathbb{E}\{\overline{X}_j\overline{X}_i|\boldsymbol{\theta}\}-\mathbb{E}\{\overline{X}_j\overline{X}_i\}\right) p(\boldsymbol{\theta}) \ \forall i,j
\end{gathered}
\tag{30}
$$

which can be solved to obtain

$$
\begin{gathered}
h(\boldsymbol{\theta}) = -\textstyle\sum_{k,l} h_{k,l}(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_l\} \\
h_k(\boldsymbol{\theta}) = \textstyle\sum_j \left[\mathbb{E}\{\overline{X}_k\overline{X}_j\}\right]^{-1}\mathbb{E}\{\overline{X}_j|\boldsymbol{\theta}\}p(\boldsymbol{\theta}) \ \ \forall k \\
h_{k,l}(\boldsymbol{\theta}){=}\textstyle\sum_{j,i}\left[\mathbb{E}\{\overline{X}_k\overline{X}_l\overline{X}_j\overline{X}_i\}\right]^{-1} \\
\cdot\left[\left(\mathbb{E}\{\overline{X}_j\overline{X}_i|\boldsymbol{\theta}\}{-}\mathbb{E}\{\overline{X}_j\overline{X}_i\}\right) p(\boldsymbol{\theta}){-}h(\boldsymbol{\theta})\mathbb{E}\{\overline{X}_k\overline{X}_l\}\right] \forall k,l
\end{gathered}
\tag{31}
$$

Finally, the first and third equations in (31) can be combined to obtain the desired 3 explicit solutions for the unknown scalar, vector, and matrix defining the estimator. As can be seen, the solution for $h_{k,j}(\boldsymbol{\theta})$ requires the inversion of a rank-4 tensor. A standard approach to doing this is to represent the tensor in terms of matrices and use existing software to invert the matrices, and then convert those back to the desired inverse tensor. See, for example, the article Bu et al. (2014), and references therein, and Kisil et al. (2022).

Example 3: Higher-Order Polynomial Estimators Observe from (19) that the representations of the linear functionals $\{H_j(\boldsymbol{\theta})\}$ for homogeneous polynomial nonlinearities $\{g_j(\boldsymbol{X})\}$ are rank-1 tensors (vectors) in one linear design equation for 1st order polynomials, then rank-1 and rank-2 tensors (vectors and matrices) in two simultaneous linear design equations for 2nd order polynomials, then rank-1, rank-2, and (by extrapolating) rank-3 tensors in three linear design equations for 3rd order polynomials, etc; and the conditional moments of the modeled data defining these linear equations are rank-1 and rank-2 tensors for 1st-order polynomials, then rank-1 through rank-4 tensors for 2nd order polynomials, and then rank-1 through rank-6 tensors for 3rd order polynomials, etc.

> *This pattern enables one to simply write down the tensor design equations for any order polynomial estimator of the posterior PDF. All the analytical work has been done here, leaving for the user only the computational challenge of inverting tensors or otherwise solving explicit linear tensor equations.*

4 Summing up the Radically Different MoM

The SCBM can be summed up as follows:

- The Pseudo Min-Risk Estimate of a parameter vector is calculated from the structurally constrained Min-MSE estimate of the random posterior PDF in the same manner that the true Min-Risk parameter estimate would be computed from the true posterior PDF, were it available.

- The Min-MSE Posterior PDF Estimate is calculated from the structurally constrained formula

  $$\widehat{p}(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_j H_j[g_j(\boldsymbol{x})]$$

  in which the nonlinear functions (tensors) $\{g_j(\boldsymbol{x})\}$ are specified by the user (e.g., (7) or (8)).

- The linear functionals $\{H_j\}$ in this formula are the solutions to the set of simultaneous linear equations

  $$\sum_j H_j \cdot \int \mathbb{E}\left\{[g_j(\boldsymbol{X})]\, g_k(\boldsymbol{X})|\widetilde{\boldsymbol{\theta}}\right\} p(\widetilde{\boldsymbol{\theta}})\mathrm{d}\widetilde{\boldsymbol{\theta}} = \mathbb{E}\{g_k(\boldsymbol{X})|\boldsymbol{\theta}\}p(\boldsymbol{\theta}) \;\; \forall\{k\}$$

  (e.g., (19)). If the prior PDF is unknown, it is approximated with a uniform PDF over a user specified region of the parameter space (e.g., (21)).

- If the user specified nonlinear functions are multivariate polynomials, then all expected values in these linear equations are conditional moments obtained from a probabilistic model of the observations, justifying this as a *method of moments* (e.g., (19) - (21)).

- Moreover, for homogeneous polynomial nonlinearities, the linear design equations can be explicitly written down in terms of linear tensor equations, knowing nothing more than the specified order of the polynomial to be used. Similarly, the estimator formula can be explicitly written down as a polynomial in the observed data. The only work a user needs to do is solve the known simultaneous linear tensor equations and implement the polynomial posterior PDF estimator.

- As explained below in Section 7, the estimated posterior PDFs satisfy two of the three traditional axioms of probability

5 Options for SCBM Solutions for Parameter Estimates

Once we have the optimum estimate of the posterior PDF, we can proceed to choose a particular Bayesian Minimum-Risk performance criterion for estimating the parameters $\boldsymbol{\theta}$.

For example, we can choose the posterior mode (MAP) criterion described above which, for the assumption of uniform prior PDF, is equivalent to ML; or we can choose the posterior median, which derives from using the absolute value of the error in each element of the estimate of the vector $\boldsymbol{\theta}$ for the risk function. We also can use the posterior mean, which results from using the squared error of each element of the estimate of the vector $\boldsymbol{\theta}$. Some comparisons have been made between the pseudo posterior mode and pseudo posterior mean estimates in Gardner (1973), Gardner (1976*b*), and especially Gardner (1981). The results of these comparisons depend on the particular structural constraints chosen. Consequently, there may be low likelihood of obtaining any general comparative results on performance dependence on the selected type of risk. Nevertheless, the results in Gardner (1981) establish some conditions under which the estimated posterior mean is superior to the estimated posterior mode for the decision problem of classifying observed data into one of a finite number of specified classes. This is interesting since the mode seems like a more natural choice and actually is when the posterior probability is not just an estimate.

6 Application of SCBM to Decision Making

The Bayesian approach to minimum-risk decision making uses the same performance criterion as that it uses for parameter estimation. The primary difference is that the parameters for decision making are discrete-valued, and each discrete value corresponds to a particular hypothesis. The hypothesis that is decided to be the correct one minimizes the risk, given the particular observed data. Consequently, the SCBM described in this paper applies as well to decision making as it does to parameter estimation. This has been pursued in the early work reported in Gardner (1973), Gardner (1976*b*), Gardner (1981). The Author does not know of any formalism that has been formulated for a decision-making counterpart to the classical MoM formulated for parameter estimation. (However, one would expect that some work on this concept has been done.) Consequently, no complement to Table 1 that applies to decision making is included herein. Nevertheless, it seems likely that Table 1 may apply, as is, to both parameter estimation and decision making.

7 Properties of the SCBM Posterior PDF Estimator

It is shown in the original contribution Gardner (1976*b*) that the posterior PDF (and discrete probability mass function) estimates provided by the SCBM satisfy the traditional axioms of probability, regardless of the specific structural constraints chosen by the user, except for the positivity axiom. Another property of interest is revealed by the general solution (26) for a constant-plus-linear constraint, and this is that the posterior PDF estimate is explicitly specified in terms of the prior PDF and the conditional mean of the centered and decorrelated data. In all cases of essentially arbitrary nonlinearities in the structural constraints, the solution is fully specified in terms of the prior PDF and conditional first- and second-order moments of the nonlinearly transformed data. And for polynomial nonlinearities, these are equivalent to higher-order conditional moments of the model for the original random data, guaranteeing this is indeed a method of moments; however, in place of the sample moments of the data used in the Classical MoM, more general weighted averages of the data and products of the data with itself are used, and the weighting functions are optimized according to a Bayesian minimum-risk (minimum mean squared error) criterion.

8 Applications

To illustrate a nontraditional type of application of this alternative MoM, previously published work is referred to here. In Gardner (1973), Gardner (1976*b*) the problem of optimizing a digital communications system receiver is addressed. One of the models used for this is a continuous-time cyclostationary process defined for all time, and the unknown parameters in this process comprise an infinite sequence of discrete values from a finite alphabet of encoded symbols representing the information-bearing data being transmitted on a stream of pulses. Thus, this is an ongoing decision problem in which a decision as to which symbol was transmitted is made every symbol interval (after some delay required to process data following each symbol interval) . The data received for each symbol extends over multiple symbol intervals, creating what is called inter-symbol interference. As shown in Gardner (1973), Gardner (1976*b*), the solution for a constant-plus-linearly-constrained receiver has much in common with the min-risk receiver for additive Gaussian noise: It is comprised of a parallel bank of matched filters, each filter matched to one of the finite set of transmitted pulse shapes, followed by a symbol-rate time sampler and a multi-input/multi-output sampled-data filter which produces SCBM estimates of the posterior probabilities of the transmitted symbols. This portion of the receiver structure that follows the bank of matched filters is known as a Fractionally Spaced Equalizer, which attempts to remove the intersymbol interference; however, its function is seen here to be much more than a traditional channel equalizer. In fact, it is more akin to a discrete-time Wiener filter. These probability estimates can be used for making decisions on which of the symbols from the finite alphabet were transmitted or for estimating symbol values or estimating the entire transmitted signal.

Another application, addressed in Gardner (1976*a*), considers parameter estimation and decision making for marked and filtered Poisson processes, used to model optical communications signals transmitted over optical fibers. Results obtained for a linearly constrained receiver strongly paralleling those obtained in Gardner (1973), Gardner (1976*b*).

Yet another application to communications receiver design is addressed in Gardner (1976*b*), where a linear-plus-quadratically constrained receiver for noncoherent decision making for sinewave-carrier modulated signals is considered. Again, results obtained are similar to optimum receivers for signals in Gaussian noise.

9 Reflection

Some of the concepts used to formulate the SCBM parameter estimation method could be said to be twisted—they are quite unconventional. Seeking a new MoM within the Bayesian framework seems unmotivated and, at first glance, unlikely to succeed. Yet the Bayesian formulation is logical, and it leads to a tractable genuine MoM for two reasons:

1. The infrequently used concept that the posterior probability, with the conditioning quantity — which is normally a sample of a set of observed random variables — replaced with the observable random variables (not their samples), is itself a random variable and can be subjected to classical random variable estimation theory; though, it is uncommon to apply such theory to the problem of estimating an unknown deterministic function $u(\boldsymbol{X})$ of the observations, which is exactly what the posterior probability is. In fact, such a problem is generally unsolvable because it generally requires knowledge of the unknown function, even when the estimates are constrained to belong to a linear space derived from the observations, such as Λ herein. It appears, at first glance, by comparing (12) and (13), to be solvable under only one condition and this is that $u(\boldsymbol{x})$ is proportional to the ratio $p(\boldsymbol{x}|\boldsymbol{\theta})/p(\boldsymbol{x})$ of the likelihood function

to the unconditional PDF of the data, an example of which is the posterior PDF in which case the proportionality factor is the prior PDF $p(\boldsymbol{\theta})$. This condition is responsible for the disappearance of the unknown function $u(\boldsymbol{X}) = p(\boldsymbol{\theta}|\boldsymbol{X})$ in the RHS of the design equation (12) as per (13). However, a deeper look reveals that $u(\boldsymbol{X})$ and $p(\boldsymbol{\theta}|\boldsymbol{X})a$ for any scalar a can differ by any random variable that is orthogonal to $g_k(\boldsymbol{X})$ for all k. A good example is $u(\boldsymbol{X})$ equal to the event indicator function, $u(\boldsymbol{X}) = 1$ for all samples $\boldsymbol{X} = \boldsymbol{x}$ for which the event $\boldsymbol{\Theta} = \boldsymbol{\theta}$ occurs and $u(\boldsymbol{X}) = 0$ for all other $\boldsymbol{X} = \boldsymbol{x}$. It is easily shown that (12) reduces to (14) with this choice for $u(\boldsymbol{X})$. The reason for this is that $p(\boldsymbol{\theta}|\boldsymbol{X})$ is the orthogonal projection of this indicator function onto the space of all finite mean-square functions of $\boldsymbol{X}$ (see (Gardner 1989, pp. 427-428)). Therefore, the orthogonal projection of this indicator function onto the linear sub-space Λ is identical to the orthogonal projection of $p(\boldsymbol{\theta}|\boldsymbol{X})$ onto Λ.

2. The adoption of minimum-mean-squared error as an optimality criterion for estimating the function $u(\boldsymbol{X})$, together with the constraint on the estimator to a hyperplane in the space of all admissible functions $g(\boldsymbol{X})$ of the data. These two choices of formulation are responsible for the design equation (12) being a set of linear equations.

The observation above reveals that this alternative MoM could have been formulated in terms of estimating either any scaled version of the event indicator function or the ratio $p(\boldsymbol{x}|\boldsymbol{\theta})/p(\boldsymbol{x})$ instead of the posterior PDF $p(\boldsymbol{\theta}|\boldsymbol{x})$. In these cases, the prior PDF $p(\boldsymbol{\theta})$ disappears (with the appropriate scalar a) from the RHS of the general design equation (15), but not the LHS.

Because this methodology is so highly structured in terms of the algorithms required for implementation, namely linear equation solvers and multivariate polynomial functionals of the observations, it should be highly amenable to efficient algorithmic implementations in terms of either software computer applications or special purpose digital signal processing hardware.

As a final remark, it is mentioned that, unlike the Radically Different MoM, the Classical MoM and associated primary Methods 2 and 4 do not appear to be nearly as convenient a starting point for developing a tracking parameter estimator, regardless of how the memory of the sample moments calculator is adjusted, because every change in the sample moments requires the solution of a new set of generally nonlinear equations.

10 Conclusions

A method of parameter estimation using only specified moments of the observed data is described. It is radically different from the classical method of moments (MoM) introduced at the end of the 19th Century and shows promise for being competitive. The alternative method uses estimates of posterior PDF values of the unknown parameters – estimates that are constrained to be linear combinations of specified nonlinear transformations of the observed data. These estimates are the solutions to linear equations specified in terms of first- and second-order moments from a probabilistic model of the nonlinearly transformed data. For polynomial non-linearities up to the order n, these are equivalent to moments of the observed random variables up to the order $2n$, revealing that this general method includes an alternative MoM as a special case; however, in place of the sample moments of the data used along with the probabilistic moments in the Classical MoM, more general weighted averages of products of the data with itself are used, and the weighting functions are optimized according to a Bayesian minimum-risk criterion. The solution for the posterior PDF estimate is studied analytically. Results are encouraging.

APPENDIX: Outline of Derivation of New MoM

Background

- The Method of Moments (MoM) is a classical statistical technique for estimating the parameters of a probabilistic data model
- The MoM was introduced just prior to the turn of the 19th Century by K. Pearson and P. Chebyshev, independently
- It is designed for statistical inference where the available data consists of multiple samples of a set of random variables, with a partially specified probabilistic model
- **The partial model needed is a set of joint moments of various orders for the random variables, showing explicit dependence on unknown parameters**

The Classical Method of Moments

- The number of moments M needed is equal to the number of unknown parameters a in these moment models (formulas); e.g.,

$$\begin{aligned} M_{12} = \mathbb{E}\{X_1 X_2\} &= f(a_1, a_2, a_3) \\ M_2 = \mathbb{E}\{(X_2)^2\} &= g(a_2) \\ M_1 = \mathbb{E}\{(X_1)^2\} &= h(a_1) \end{aligned}$$

  for which f, g, h are known functions

- The statistics that are computed from the data consist of the sample moments corresponding to the theoretical moment models, e.g.,

$$m_{12} = \frac{1}{n}\sum_{j=1}^{n} x_1^j x_2^j$$
$$m_2 = \frac{1}{n}\sum_{j=1}^{n} (x_2^j)^2$$
$$m_1 = \frac{1}{n}\sum_{j=1}^{n} (x_1^j)^2$$

- **The inference procedure is to equate the computed sample moments to the theoretical moment formulas and attempt to solve these equations**

$$m_{12} = f(a_1, a_2, a_3)$$
$$m_2 = g(a_2)$$
$$m_1 = h(a_1)$$

- The tractability of this MoM depends on the particular nonlinear equations

An Alternative Approach

- I recently observed that **every multivariate statistical inference problem based on multiple samples can be *reformulated* as a problem of statistical inference for a single times series of data based on one sample path of the**

time series, consisting of concatenated time-series segments equal to a first sample of the ordered set of random variables, followed by a 2nd sample of the same random variables, and so on until all samples have been included, e.g.,

$$\{y_k\}_1^{16} = \{x_1^1, x_2^1, x_1^2, x_2^2, x_1^3, x_2^3, x_1^4, x_2^4, x_1^5, x_2^5, x_1^6, x_2^6, x_1^7, x_2^7, x_1^8, x_2^8\}$$

- The theoretical model for this time series is a single sample path of a cyclostationary stochastic process $\{Y_k\}$, with period equal to the number of random variables and with the time sequence of this set of random variables being i.i.d. from one period to the next: e.g., $\{x_1^1, x_2^1\}$ and $\{x_1^2, x_2^2\}$ are i.i.d.

- This is a special cyclostationary process because it contains the same unknown parameters in every period

- I generalized this model to allow the parameter values to change from one period to the next and modeled them as samples of a stationary sequence of random variables, which preserves the cyclostationarity

- Then I invoked an unusual methodology I had introduced in the early 1970s for this type of cyclostationary process model which I used for commonly encountered digital pulse-modulated signals used in communications transmission systems

- The unusual methodology uses Bayesian concepts to formulate the problem of estimating the parameter values (transmitted digits $\{a_i\}$) in terms of the sequence of posterior probabilities, which can be used to compute various minimum-risk parameter estimates, such as maximum-posterior-probability estimates and minimum-mean-squared-error estimates, e.g.,

$$\hat{a}_i = \max_{a_i} P(a_i | \{y_k\})$$

- Finally, I formulated an inference problem for estimating these posterior probabilities using structurally constrained minimum-MSE estimators: optimum linear combinations of any appropriate specified nonlinear transformation of the data samples

$$\hat{\hat{a}}_i = \max_{a_i} \hat{P}(a_i|\{y_k\})$$

- This particular formulation ensures the posterior probability estimates are always the solutions to sets of *simultaneous linear equations*

- By choosing polynomial nonlinearities, the equations are fully *specified by weighted sample moments* of the data; **this makes it a MoM**

- The weights are optimal in the sense of producing structurally constrained minimum-MSE estimates of the posterior probabilities

- In actuality, the reformulation process described above was performed in reverse order for the purpose of showing that **the original work on time series was equivalent to a radically new MoM**.

Summary

- A new Method of Moments has been introduced and it is radically different from the four primary methods.

- The **numerous advantages** of the new method are fully described in the Table 1 in Section 1

- The utility of the new method was studied back in the 1970s for estimating digital symbols in digital transmission systems developed by Bell Telephone Labs

- **But more diverse applications to various specific multivariate parameter estimation problems, and comparison with the classical MoM, has not yet been pursued**

What's Unusual About this Application of Bayes Minimum Risk Methodology?

- The quantities to be estimated, the posterior probabilities of parameters, are **deterministic functions of the observed data**.
- So, why do we need to estimate them?
- For the same reason we would choose to use the MoM: we do not know the complete probabilistic model for the data and, because of this, we cannot calculate these functions as in (2), (3)
- The particular way I set up the problem for estimating the unknown function

  $$P(a|\{y_k\})$$

  of the known data **requires knowledge of only moments of orders determined by the orders of the polynomial nonlinearities selected for the structural constraint**
- This was not foreseen, but rather was discovered during my open-ended investigation as a young naïve investigator in my first year as an assistant professor

References

Bu, C., Zhang, X., Zhou, J., Wang, W. & Wei, Y. (2014), 'The inverse, rank and product of tensors', *Linear Algebra and Its Applications* **446**, 269–280.

Gardner, W. A. (1973), 'The structure of least-mean-square linear estimators for synchronous M-ary signals (corresp.)', *IEEE Transactions on Information Theory* **19**(2), 240–243.

Gardner, W. A. (1976*a*), 'An equivalent linear model for marked and filtered doubly stochastic Poisson processes with application to MMSE linear estimation for synchronous m-ary optical data signals', *IEEE Transactions on Communications* **24**(8), 917–921.

Gardner, W. A. (1976*b*), 'Structurally constrained receivers for signal detection and estimation', *IEEE Transactions on Communications* **24**(6), 578–592.

Gardner, W. A. (1981), 'Design of nearest prototype signal classifiers (corresp.)', *IEEE Transactions on Information Theory* **27**(3), 368–372.

Gardner, W. A. (1989), *Introduction to random processes with applications to signals & systems*, McGraw-Hill.

Gardner, W. A. (2022), 'Cyclostationarity: educational website', *www.cyclostationarity.com* .

Gardner, W. A., Napolitano, A. & Paura, L. (2006), 'Cyclostationarity: Half a century of research', *Signal processing* **86**(4), 639–697.

Kisil, I., Calvi, G. G., Konstantinidis, K., Xu, Y. L. & Mandic, D. P. (2022), 'Accelerating tensor contraction products via tensor-train decomposition [tips & tricks]', *IEEE Signal Processing Magazine* **39**(5), 63–70.

Lindsay, B. G. (2014), 'Method of moments', *https://doi.org/10.1002/9781118445112.stat05908* .

Pearson, K. (1936), 'Method of moments and method of maximum likelihood', *Biometrika* **28**(1/2), 34–59.

Poulsen, R. (1977), Evaluation of the constrained Bayesian methodology for signal detection, Ph. D. Dissertation, Department of Electrical and Computer Engineering, University of California, Davis.

Provost, S. B. (2005), *https://www.researchgate.net/publication/242782657_Moment-Based_Density_Approximants* .

Wikipedia (2022), 'Generalized method of moments', *https://en.wikipedia.org/wiki/Generalized_method_of_moments* .