Gaussian Mixtures and their Applications to Signal Processing

K.N. Plataniotis and D. Hatzinakos
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Ontario, M5S 3G4, Canada

March 10, 2000

Chapter 1

Gaussian Mixtures and their Applications to Signal Processing

Abstract

There is a number of engineering applications in which a function should be estimated from data. Mixtures of distributions, especially Gaussian mixtures, have been used extensively as models in such problems where data can be viewed as arising from two or more populations mixed in varying proportions [1]-[3]. The objective of the chapter is to highlight the use of mixture models as a way to provide efficient and accurate solutions to problems of important engineering significance. Using the Gaussian mixture formulation, problems are treated from a global viewpoint that readily yields and unifies previous, seemingly unrelated results. The proposed work reviews the existing methodologies, examines current trends, provides connections with other methodologies and practices, and discusses application areas.

Contents

Nomeclature

Φ	family of distributions
$F(x\|\hat{\theta})$	conditional distribution
$\hat{\theta}$	unknown parameter
$\hat{\theta}$	estimated value
$N(x;\mu_i,\Sigma_i)$	$d-$ dimensional Gaussian
μ	mean value
Σ	covariance
w_i	mixing coefficient
ML	maximum likelihood
EM	expectation-maximization
$log\Lambda$	log likelihood
Z	missing data (EM algorithm)
LRT: likelihood ratio test	
N_g	number of mixture components
RBF	radial basis functions
PNN	probabilistic neural network
$x(k)$	state vector
$Z^k = z(1), z(2), ..., z(k), ...$	observation record
$\hat{x}(k\|k) = E(x(k)\|Z^k)$	mean-squared-error filtered estimate
EKF	Extended Kalman Filter
$J_h(x\bar{(k)})$	Jacobian matrix
AGSF	Adaptive Gaussian Sum Filter
CMKF	verted measurement Kalman filter
ϵ-mixture	ϵ-contaminated Gaussian mixture model
$i(k)$	intersymbol inference
$n(k)$	thermal noise
NMSE	normalized mean square error

1.1 Introduction

Central to unsupervised learning in adaptive signal processing, stochastic estimation, and pattern recognition is the determination of the underlying probability density function of the quantity of interest based on available measurement data [4]. If no a-priori knowledge of the functional form of the requested density is available non-parametric techniques should be used. Therefore, over the years a number of techniques ranging from data histograms and kernel estimators, to neural network and fuzzy system based approximators have been proposed [4], [7], [8]. On the other hand if some impartial a-priori knowledge regarding the data characteristics is available, the requested probability function is assumed to be of a known functional form but with a set of unknown parameters. The parametrized function provides a partial description where the full knowledge of the underlying phenomenon is achieved through the specific values of the parameters.

Let x be an d-dimensional vector with a probability distribution $F(x)$ and probability density $f(x)$. In most engineering problems, a density such as the multidimensional Gaussian is assumed. More often families of parametric distributions are used [14]. In this case the family is considered to be a linear combination of given distributions. This family is often called parametric since their members can be characterized by a finite number of parameters [9]-[12], [15], [18], [23]-[26].

The family of distributions considered in this chapter can be defined as:

$$\Phi = [F(x|\theta); \theta \in \Theta] \tag{1.1}$$

Suppose that a sequence of random identically distributed observation's $x_1, x_2, ..., x_n$ are drawn form $F(x|\hat{\theta})$ with $\hat{\theta}$ unknown to the observer. An estimate of the unknown parameter $\hat{\theta}$ which can be obtained as a function of the observations can be used to completely characterize the mixture [10], [18], [16].

Let us assume that associated with each one of the random samples $x_1, x_2...$ is a probability distribution with the possibility some of the samples being from $F(x|\theta^1)$, some from $F(x|\theta^2)$, etc, with θ^1, θ^2 are different realizations of the unknown parameter θ. In other words any sample x could be from any of the member distributions in the parametric family Φ [5]. Defining a mixing distribution $G(\theta)$ which describes the probability that point θ characterizes the mixture, the sample x can be considered as having a distribution

$$H(x) = \int F(x|\theta) dG(\theta) \tag{1.2}$$

which is called a mixture. In most engineering applications a finite number of points $\theta^1, \theta^2, ..., \theta^g$ are assumed. Then the mixing distribution is expressed as:

$$G(\theta) = \sum_{i=1}^{Ng} P(\theta^i)\delta(\theta - \theta^i) \tag{1.3}$$

Substituting (3) into the mixture expression of (2) the finite mixture

$$H(x) = \sum_{i=1}^{Ng} F(x|\theta^i)P(\theta^i) \tag{1.4}$$

can be obtained. The parameter points used to discritize the mixture can be known a-priori with the only unknown elements in the above expression the mixing parameters $P(\theta^i)$. In such a scenario the distributions used in the mixture (basis functions) are determined a-priori. Thus, only the mixture coefficients are fit to the observations, usually through the minimization of an error

criterion. Alternatively, the basis functions themselves (through their parameters) are adapted to the data in addition to the mixing coefficients. In such a case, the optimization of the mixture parameters becomes a difficult non-linear problem and the type of the basis function selected as well as the type of the optimization strategy used becomes very important. Because of their simplicity and their efficient representation in terms of the first two moments Gaussian densities are most often used as the basis functions [5].

The discussion in this chapter is intended to provide a perspective on the Gaussian mixture approach to developing solutions and methodologies for signal processing problems. We will discuss in detail a number of engineering areas of application of finite Gaussian mixtures. In engineering applications the finite mixture representation can be used to: (i) directly represent the underlying physical phenomenon, e.g. tracking in a multi-target environment, medical diagnosis, etc., and (ii) indirectly model underlying phenomena that do not necessarily have a direct physical interpretation, e.g. outlier modeling in communication channels. The problem of tracking a target using polar coordinate measurements in used here to demonstrate the applicability of the Gaussian mixture model to model an actual physical phenomenon. The process of tracking a target involves the reception and processing of received signals. The Gaussian mixture model is used to approximate the densities involved in the derivation of the optimal Bayesian estimator needed to provide reliable and cost effective estimates of the state of the system. In addition, we also discuss in detail the problem of narrowband interference suppression as an example of indirect application of the Gaussian mixture model. Spread-spectrum communication systems often use estimation techniques to reject narrowband interference. The basic assumption is that the direct sequence spread-spectrum signal along with the background noise can be viewed as non-Gaussian measurement noise. The Gaussian mixture framework is then used to model the non-Gaussian measurement channels. Similar treatment of signals can easily be extended to any application subject to onnlinear effects or non-Gaussian measurements, e.g. biomedical systems. For example, Gaussian mixtures have been used to model random noise, magnetic field inhomogeneites and biological variations of the tissue in magenetic resonance imaging (MRI) as well as computerized tomography (CT) [27]-[30].

After a brief review of the mathematical aspects of Gaussian mixtures, three methodologies for estimating mixture parameters are discussed. Particular emphasis is placed on the expectation/maximization (EM) algorithm and its applicability to the problem of adaptive mixture parameter determination. Computational issues are also analyzed with emphasis on the computer generation of mixture variables. Then the framework is applied to two problems and numerical results are presented. The results included in the chapter are meant to be illustrative rather than exhaustive. Finally, to demonstrate the versatility and the powerful nature of the framework connections with other research areas are drawn with particular emphasis on the connection between Gaussian mixtures and the radial-basis functions (RBF) networks.

1.2 Mathematical Aspects of Gaussian Mixtures

1.2.1 The approximation Theorem

In an adaptive signal processing, unsupervised learning environment the usefulness of the Gaussian mixture model depends on two factors. First, whether or not the approximation is sufficiently powerful to represent a broad class of density functions, most notably those that are encountered in engineering applications. Secondly, if such an approximation can be obtained in a reasonable manner through a parameter estimation scheme which allows the user to compute the optimal values of the mixture parameters from a finite set of data samples [6], [51]-[53], [13].

Regarding the first question a Gaussian mixture can be constructed to approximate arbitrary well any given density. This can be proven by utilizing the Wiener's theorem of approximation, or by considering delta functions of a positive type. This methodology, first presented in [8], [51], is reviewed in this chapter. The resulting class of density functions is rich enough to approximate all density functions of engineering interest [8], [51].

We start reviewing the methodology by briefly discussing the characteristics and properties of Delta functions. *Delta families of positive type* are families of functions which converge to a delta, (impulse) function as a parameter characterizing the family converges to a limit value. Specifically, let δ_λ be a family of functions on the interval $(-\infty, \infty)$ which are integrable over every interval. This is called *delta family of positive type* if the following conditions are satisfied.

1. $\int_{-a}^{a} \delta_\lambda(x)\, dx \to \lambda$ as $\lambda \to \lambda_0$ for some a.
2. For every constant $\gamma > 0$ δ_λ tends to zero uniformly for $\gamma \leq |x| \leq \infty$ as $\lambda \to \lambda_0$.
3. $\delta_\lambda(x) \geq 0$ for all x and λ.

If such a function required to satisfy the condition that

$$\int_{-\infty}^{\infty} \delta_\lambda(x)\, dx = 1 \tag{1.5}$$

then it defines a probability density function for all λ. It can seen by inspection that the Gaussian density tends to the delta function as the variance tends to zero, and therefore can be used as a basis functions for approximation purposes [51], [73].

Using the delta families, the following result can be used for the approximation of an arbitrary density function p.
The sequence $p_\lambda(x)$ which is formed by the convolution of δ_λ and p

$$p_\lambda(x) = \int_{-\infty}^{\infty} \delta_\lambda(x-u) p(u)\, du \tag{1.6}$$

converges uniformly to $p(x)$ on every interior subinterval of $(-\infty, \infty)$.

When the density p has a finite number of discontinuities the above holds true except at the points of discontinuity. Since the Gaussian density can be used as a delta family of positive type, the approximation p_λ can be written as follows:

$$p_\lambda(x) = \int_{-\infty}^{\infty} N_\lambda(x-u) p(u)\, du \tag{1.7}$$

which forms the basis for the Gaussian sum approximation. The term $\delta_\lambda(x-u)p(u)$ is integrable on $(-\infty, \infty)$ and it is at least piecewise continuous. Thus, $p_\lambda(x)$ itself can itself be approximated on any finite interval by a *Riemann* sum. In particular if a bounded interval (a, b) is considered the function is given as:

$$p_{\lambda,n}(x) = \frac{1}{k} \sum_{i=1}^{n} N_\lambda(x - x_i)[\xi_i - \xi_{i-1}] \tag{1.8}$$

where the interval (a, b) is divided into n subintervals by selecting points such that:

$$a = \xi_0 < \xi_1 < \xi_2 < ... < \xi_n = b \tag{1.9}$$

In each such subinterval, a point x_i is chosen such as:

$$p(x_i)[\xi_i - \xi_{i-1}] = \int_{-\xi_{i-1}}^{\xi_i} p(x)\,dx \tag{1.10}$$

which is possible by the mean value theorem. The normalization constant k ensures that the density $p_{\lambda,n}$ is a density function.

Consequently, an approximation of p_λ over some bounded interval (a, b) can be written as:

$$p_{\lambda,n}(x) = \sum_{i=1}^{n} w_i N_{\sigma_i}(x - x_i) \tag{1.11}$$

where $\sum_{i=1}^{n} w_i = 1$ and $w_i \geq 0$ for all i.

The relation between the last two equations is obvious by inspection. However, in the last equation the variance σ_i can vary from one term to another. This has be done to obtain greater flexibility for a approximations using Gaussian mixtures with finite number of terms. As the number of terms in the mixture increase, it is necessary to require that σ_i tend to become equal and vanish.

Under this framework, an unknown $d-$ dimensional distribution (density function) can be expressed as a linear combination of Gaussian terms. The form of the approximation is a s follows:

$$p(x) = \sum_{i=1}^{Ng} \omega_i N(x; \mu_i, \Sigma_i) \tag{1.12}$$

where $N(.)$ represents a $d-$ dimensional Gaussian density defined as:

$$N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{0.5}|\Sigma|^{0.5}} exp(-0.5(x-\mu)^\tau \Sigma^{-1}(x-\mu)) \tag{1.13}$$

where μ, Σ are the mean and covariance of the Gaussian basis fu nctions and w_i in (5) is the i^{th} mixing coefficient (weight) with the assumption that $\omega_i \geq 0$, $\forall i = 1, 2, ..Ng$ and $\sum_{i=1}^{Ng} \omega_i = 1$.

1.2.2 The identifiability problem

The problem most often encountered in the context of finite mixtures is that of the identifiability, meaning the uniqueness of representation in the mixture [20]- [22], [31]. If the Gaussian mixture of (12) is identifiable then the equation:

$$\sum_{i=1}^{M} w_i F(x|\theta^i) = \sum_{j=1}^{M'} w'_j F(x|\theta^j) \tag{1.14}$$

implies that:

1. $M = M'$
2. for each $i, 1 \leq i \leq M$, there exists uniquely j, $1 \leq j \leq M'$ such that $w_i = w'_j$ and $F(x|\theta^i) = F(x|\theta^j)$,

There exist extensive literature on the problem of mixture identifiability. A necessary and sufficient condition that the class Φ of all finite mixtures be identifiable is Φ be a linearly independent set

over the field of real numbers (Teicher, Yakowitz and Spragins) [20], [31], [5]. Using the above conditions the identifiability of several common distribution functions have been investigated. Among the class of all finite mixtures, that of Gamma distributions, the one-dimensional Cauchy distribution, the one-dimensional Gaussian and finally the multi-dimensional Gaussian family is identifiable.

The following theorem, discusses the identifiability problem [21], [31]:

Theorem:

A necessary and sufficient condition that the class of all finite mixtures of the family $\aleph$ be identifiable is that F be a linearly independent set over the field of real numbers.

Proof

Necessity:

Let $\sum_{i=1}^{M} \alpha_i F_i = 0 \ \ \forall x$, α_i real numbers be a linear relation in $\aleph$. Assume that the α_i's are subscripted so that $\alpha_i < 0$ if $i < N$. We then have

$$\sum_{i=1}^{N} \alpha_i F_i + \sum_{i=N+1}^{M} \alpha_i F_i = 0 \longrightarrow \sum_{i=1}^{N} |\alpha_i| F_i = \sum_{i=N+1}^{M} |\alpha_i| F_i$$

Since the F_i are distribution functions **d.f** or **c.d.f**, $F_i(\infty) = 1$, thus

$$\sum_{i=1}^{N} |\alpha_i| = \sum_{i=N+1}^{M} |\alpha_i| = b > 0$$

Therefore, if we define $w_i = \frac{|\alpha_i|}{b}$ we have that

$$\sum_{i=1}^{N} w_i^1 F_i = \sum_{i=N+1}^{M} w_i F_i$$

Since by definition $w_i > 0$ and $\sum_{i=1}^{N} w_i^1 = \sum_{i=N+1}^{M} w_i = 1$ the coefficients satisfy the requirements for mixing parameters.

The relation $\sum_{i=1}^{N} w_i^1 F_i = \sum_{i=N+1}^{M} w_i F_i$ a sserts that there exist two distinct representation of a finite mixture so that *leph* cannot be identifiable.

Since the proof of necessity requires that $\aleph$ is identifiable, we are led to a contradiction which followed from assuming that the members of the family are linearly dependent. Consequently, the conclusion follows that the member of the family form a linearly independent set over the field of real numbers.

Sufficiency:

If a given mixture is a linearly independent set then it can be considered as a basis which span the family $\aleph$. If there were two distinct representations of the same mixture, this contradict the unique representation property of a basis. This does not mean that there exists only one representation of the mixture but rather , that given a basis for span the family consisting of $(F_i)_{i=1}^{M}$ the relation $\sum_{i=1}^{N} w_i^1 F_i = \sum_{i=N+1}^{M} w_i F_i$ implies always that $w_i^1 = w_i$. The unique representation property of a basis, allows the conclusion that if F is a linearly independent set is sufficient for identifiability.

The problem of identifiability it is of significant practical importance in all practical applications of mixtures. Without resolving the problem of the unique characterization of the mixture model, a reliable estimation procedure to determine its parameters cannot be defined. There are many classes of mixture models in which we are unable to define a unique representation. A simple example of such non-identifiable mixture is the uniform distribution which it can be expressed as a mixture of two other uniform distributions, e.g. $U(;0.5,0.5) = 0.5U(x;0.25,0.25) + 0.5U(x;0.75,0.25)$. However, by utilizing the theorems summarized above it has been proven that the class of all finite mixtures of Gaussian (normal) distributions is identifiable [1], [5].

1.3 Methodologies for mixture parameter estimation

The problem of determining the parameters of the mixture to best approximate a given density function can be solved in more than one ways. There exist considerable literature on mixture parameter estimation with a variety of different approaches ranging from the moments method [46], to the moment generation function [3], graphical methods [47], Bayesian methods [9] and the different variations of the maximum likelihood method [1], [4], [10], [32]. In this chapter we will concentrate on the maximum likelihood approach.

There are two different methodologies in estimating the parameters of the Gaussian mixture by using the maximum likelihood principle. The first approach is the iterative one, in which the parameter values are refined by processing the data iteratively. Alternatively, one can use a recursive approach, refining the mixture parameter values with each new available data value. A recursive procedure requires that the latest value of a parameter within the mixture model depends only on the previous value of the estimate and the current data sample. Generally speaking, an iterative procedure will produce better results than a recursive one. On the other hand, the recursive parameter estimator is usually much faster than the iterative one. In the case of the Gaussian mixture approximation, we are interested to estimate, from the data, the mixing coefficients (weights) and, if needed, the first two moments of the Gaussian basis functions.

The method of choice for the estimation of the Gaussian mixture parameters is currently the Expectation/Maximization (EM) algorithm [32], [33]. This is an iterative procedure, which starts with an initial estimate of the mixture's parameters. Based on that initial guess the method constructs a sequence of estimates by first evaluating the expectation of the log likelihood of the current estimate, and then proceeds by determining the new parameter value which maximizes this expectation. Although the EM methodology is most often used we continue our analysis by reviewing first the classical maximum likelihood approach to the problem of mixture parameter estimation. In this approach estimates of the mixture parameters are obtained by maximizing the marginal likelihood function of (n) independent observations drawn form the mixture. A detail description of the method follows in the next section.

1.3.1 The maximum likelihood approach

Let us assume that a set of unlabeled data samples $(x_1, x_2, ..., x_n)$ are drawn from a Gaussian mixture density

$$p(x) = \sum_{i=1}^{Ng} p(\omega_i)p(x|\omega_i) = \sum_{i=1}^{Ng} \omega_i N(x, \theta_i) \tag{1.15}$$

with $\sum_{i=1}^{Ng} \omega_i = 1$, $\omega_i \geq 0$ for all i and θ the unknown parameter vector which summarizes the uncertainty on the me an value and the variance (covariance) of the Gaussian basis function. By applying the Bayes rule the following relation hold:

$$p(\omega_i|x) = \frac{p(\omega_i)p(x|\omega_i)}{p(x)} = \frac{\omega_i N(x, \theta_i)}{\sum_{j=1}^{N} \omega_j N(x, \theta_j)} \tag{1.16}$$

We are seeking parameters θ and ω which minimize the log-likelihood of the available samples:

$$log\Lambda = \sum_{k=1}^{n} logp(x_k) \tag{1.17}$$

Using Lagrange multipliers (17) can be rewritten as follows:

$$log\hat{\Lambda} = \sum_{k=1}^{n} logp(x_k) - \lambda(\sum_{i=1}^{n} \omega_i - 1) \tag{1.18}$$

Taking the partial derivative with respect to ω_i, and setting it equal to 0 we have the following expression:

$$\hat{\omega}_i = \frac{1}{\lambda}\sum_{k=1}^{n} p(\omega_i|x_k) \tag{1.19}$$

To obtain estimates of the generic basis parameter θ the partial derivative with respect to θ is set equal to 0:

$$\sum_{k=1}^{n} p(\omega_i|x_k)\frac{\partial}{\partial\theta_i}N(x_k,\theta_i) = 0 \tag{1.20}$$

For the case of a multi-dimensional Gaussian density, the parameter vector θ is comprised by the mean value and the covariance matrix. Taking the partial deriva tives of the logarithm with respect to their elements together we have the following relations:

$$\hat{\mu}_i = \frac{\sum_{k=1}^{n} p(\omega_i|x_k)x_k}{\sum_{k=1}^{n} p(\omega_i|x_k)} \tag{1.21}$$

$$\hat{\Sigma}_i = \frac{\sum_{k=1}^{n} p(\omega_i|x_k)(x_k - \hat{\mu}_i)^\tau(x_k - \hat{\mu}_i)}{\sum_{k=1}^{n} p(\omega_i|x_k)} \tag{1.22}$$

The system of equations (16), (21), (22), can be solved using iterative methods. However, when such an approach is used singular solutions may occur since a component density centered on a single design sample may have a likelihood that approaches infinity as the variance (covariance) of the component approaches zero. The simplest way to avoid this problem is to utilize a new set of design data samples for each iteration of the solution, making in this way impossible for a single data sample to dominate the whole component density.

Although simple in concept, this method does not work well in practice. Therefore, alternative solutions have been developed to alleviate the problem. Among them is the stochastic gradient descent solution discussed in [23] and reviewed in the next section.

1.3.2 The stochastic gradient descent approach

Let us start for the generic parametric-update formula devised through the utilization of the maximum likelihood solution. For both the mean and the variance (covariance) the update equation has the following form:

$$\theta_n = \frac{\sum_{k=1}^{n} p(\omega|x_k)\theta(x_k)}{\sum_{k=1}^{n} p(\omega|x_k)} \tag{1.23}$$

After some simple algebraic manipulation, a recursive expression for the θ_{n+1} as a function of θ_n can be obtained as:

$$\theta_{n+1} = \theta_n + \gamma_{n+1}(\theta_{n+1} - \theta_n) \tag{1.24}$$

with

$$\gamma_{n+1} = \frac{p(\omega|x_{n+1})}{\sum_{k=1}^{n+1} p(\omega|x_{n+1})} \tag{1.25}$$

The last equation can also be formulated in a recursive format. However, the denominator for the calculation of the correction term is not bounded for growing data sets (n) and thus such an estimation procedure would require infinite memory. Therefore, if we assume only a finite sample set with samples drawn unbiased from the unknown distribution and with the fixed set size (n) large, then approximately the correction factor can be calculated as follows:

$$\gamma_{n+1} \approx \frac{p(\omega|x_{n+1})}{(n+1)p(\omega)} \tag{1.26}$$

By utilizing equations (24), (26) explicit time update equations for the parameters of the Gaussian mixture can be written. Although it maybe impossible to obtain convergence from only one iteration if the design set is too small, acceptable estimates can be obtained if the data samples are drawn with replacement until a stable solution is obtained.

1.3.3 The Expectation/Maximization (EM) approach

As before we assume that a set of unlabeled data samples $(x_1, x_2, ..., x_n)$ are drawn from a Gaussian mixture density

$$p(x) = \sum_{i=1}^{Ng} p(\omega_i)p(x|\omega_i) = \sum_{i=1}^{Ng} \omega_i N(x, \theta_i) \tag{1.27}$$

with $\sum_{i=1}^{Ng} \omega_i = 1$, $\omega_i \geq 0$ for all i and θ_i is the unknown parameter vector consisting of the elements of the mean value μ_i and the distinct elements of the covariance (variance) Σ_i of the Gaussian basis function $N(x; \theta_i)$. The EM algorithm utilizes the concept of missing data which in our case is the knowledge of which Gaussian function each data sample is coming from. Let us assume that the variable Z_j provides the density membership for the j^{th} sample available. In other words, if $Z_{ij} = 1$ then x_j has a density $N(x, \theta_i)$. The values of the Z_{ij} are unknown and are treated by EM as missing information on to be estimated along with the parameters θ and ω of the mixture model. The likelihood of the model parameters θ, w given the joint distribution of the data set and the missing values Z can be defined as:

$$logL(\theta, w|(x_1, x_2, ..., x_n), Z) = \sum_{i=1}^{n} \sum_{j=1}^{Ng} Z_{ij} log(p(x_i|\theta_j)\omega_j) \tag{1.28}$$

The EM algorithm iteratively maximizes the expected *log likelihood* over the conditional distribution of the missing data, Z given (i) the observed data $x_1, x_2, ..., x_n$ and (ii) the current estimates of the mixture model parameters θ and ω. This is achieved by repeatedly applying the E-step and the M-step of the algorithm. The E-step of EM finds the expected value of the log likelihood over the values of the missing data Z given the observed data and the current parameters $\theta = \theta^0$ and $\omega = \omega^0$.

It can be shown that the following equation holds true:

$$Z_{ij}^0 = \frac{p(x_i|\theta_j^0)\omega_j^0}{\sum_{t=1}^{g} pp(x_i|\ theta a_t^0)\omega_t^0} \tag{1.29}$$

with $i = 1, 2, ..., n$ and $t = 1, 2, ..., N$. The M-step of the EM algorithm maximizes the log-likelihood over θ and ω in order to find the next estimates for them, the so-called θ^1 and ω^1.

The maximization over ω leads to a solution

$$\omega_{ji}^{1} = \sum_{i=1}^{n} \frac{Z_{ij}^{n}}{n} \tag{1.30}$$

We can then maximize over the parameters θ by maximizing the terms of the log-likelihood separately over each θ_j with $j = 1, 2, ..., g$. Therefore, evaluation of this step means calculations of the:

$$\theta_j^1 = \max_{\theta_j} \sum_{i=1}^{n} Z_{ij}^0 log(p(x_i|\theta_j) \tag{1.31}$$

For the case of Gaussian mixtures the solution to the M-step of the algorithm exists in closed form. Thus, at the $(k+1)^{th}$ iteration the current estimates for the mixture coefficients, the elemental means and the covariance matrices are given as:

$$\omega_j(k+1) = \sum_{i=1}^{n} \frac{\hat{\tau}_j(k+1)}{n} \tag{1.32}$$

$$\hat{\tau}_i(k+1) = \frac{\omega_j(k)}{N}(x_j; \theta_i(k)) \sum_{i=1}^{g} \omega_j(kN(x_j; \theta_i(k)) \tag{1.33}$$

$$\mu_i(k+1) = \frac{\sum_{j=1}^{n} \hat{\tau}_j(k+1) x_j}{n\omega_i(k+1)} \tag{1.34}$$

$$\Sigma_i(k+1) = \frac{\sum_{j=1}^{n} \hat{\tau}_j(k+1)(x_j - \mu_i(k+1))(x_j - \mu_i(k+1))^{\tau}}{n\omega_i(k+1)} \tag{1.35}$$

The EM algorithm increases the likelihood function of the data at each iteration, and under suitable regularity conditions converges to a stationary parameter vector [32]. The convergence properties of the algorithm have been discussed extensively in the literature. The EM algorithm produces a monotonic increasing sequence of likelihoods, thus if the algorithm converges it will reach a stationary point in the likelihood function, which however can be different from the global maximum. However, like any other optimization algorithm the EM algorithm depends on the provided initial values to determine the solution. Given a specific test of initial conditions it may converge to the optimal solution, while for another set of initial parameters it may find only a suboptimal one. The final set of values as well as the number of iterations needed for the convergence of the EM algorithm is thus greatly affected from the initial parameter values. Therefore, the initial placement of the Gaussian components are of paramount importance for the convergence of the algorithm.

In the problem of function approximation or distribution modeling in which the EM algorithm is used to guide the function approximation a good starting point for the elemental Gaussian terms may be near the means of the actual underlying component Gaussian terms. To this end many different techniques have been devised over the years. Among them clustering techniques, such as the different variants of the K-means algorithm are the most popular [4]. In this approach the components of the underlying distribution which generates the data are considered as data clusters and pattern recognition techniques are used to identify them. When the K-means algorithm is used to identify initial values for the EM algorithm, the number of Gaussian functions in the mixture (clusters) has to be specified in advance. Having the number of clusters predefined, an iterative procedure is invoked to move the cluster centers in order to minimize the mean square error between cluster centers and available data points. The procedure can be described as follows:

1. Randomly select N_g data points as the initial starting locations of the elemental Gaussian terms (clusters).
2. Assign a novel data point x_j to cluster center μ_i if $|x_j - \mu_i| \leq |x_j - \mu_l|$ for all $l = 1, ..., N_g$, $l \neq j$.
3. Calculate the new mean value of the data points associated with the center μ_i.
4. Repeat steps (1), (2) until the centers are stationary.

Although the above algorithm is simple and works well in many practical applications it has several drawbacks. The procedure itself depends on the initial conditions and it can converge to different solutions depending on which initial data points were selected as initial cluster centers. Thus, if it is used as initial starting point for the EM algorithm then the varying final configuration of the cluster centers produced by the K-means algorithm may lead to variations in the final Gaussian mixture generated by the EM algorithm [38].

Alternatively, scale space techniques can be utilized to determine the Gaussian term parameters from the available data samples. Such techniques initially motivated by the use of Gaussian filters for edge detection can be provide constructing descriptions of signals and functions by decomposing the data histogram into sums of Gaussian distributions [39], [40]. The scale-space description of a given data set indicated the zero-crossing points of the second derivatives of the data at varying resolutions [41]. When scale space techniques are used to determine the parameters of a Gaussian mixture we are particularly interested in the location of zero-crossings in the second derivative and the sign of the third derivative at the zero crossing. By determining the second derivatives of the data waveform and locating the zero-crossing points the number of Gaussian terms present in the approximating Gaussian mixture can be identified. The sign of the waveform's second derivative can be used to determine where the function is convex or concave [40].

In general to determine a (N_g) components normal mixture, $(3N_g - 1)$ parameters must be estimated. The direct calculation of these parameters as a function of the location of the zero-crossing points form a system of $(3N_g - 1)$ simultaneous nonlinear equations. To overcome the computational complexity of a direct estimation, a two stage procedure was proposed in [40]. In this approach a rough estimate of the parameter values are obtained based on the zero-crossing locations. With this initial set as a starting point the EM algorithm is utilized to provide the final set of the Gaussian mixture parameters. The procedure can be summarized as follows:

- At any scale, sign changes will alternate left to right. Odd (even) numbered zero-crossings will thus correspond to lower (upper) turning points.
- Given the locations of upper and lower turning points the point halfway between the turning point pair is used to provide the initial estimate of the mean μ_i.
- Half the distance between turning point pairs is used as an estimate of the standard deviation (covariance)
- Given these initial estimates of the parameters which determine the mixture, the EM algorithm is used to calculate the optimal set of parameters.

In summary, clustering techniques, such as the $K-means$ algorithm or scale space filters can be used to provide initial values for the EM algorithm. Changes in the initial conditions will result in varying final Gaussian mixtures and although there is no guarantee that the final mixture chosen is optimal those which are based on initial sets selected from these algorithms are usually better.

Finally, to improve the properties of the EM algorithm, a stochastic version of the algorithm, the so-called Stochastic EM (SEM) algorithm has been proposed in the literature [42]. Stochastic perturbation and sampling methodologies are used in the context of SEM to reduce the dependence on the initial values and to speed up convergence. If the initial parameters are sufficiently close to the actual values, the convergence is exponential for Gaussian mixtures. Although the dependence on the initial values is largely reduced in the SEM algorithm, SEM seems not appropriate for small sample records [43].

1.3.4 The EM algorithm for adaptive mixtures

The problem of determining the number (N_g) of components in the Gaussian mixture when the mixing coefficients, means and variances (covariances) of the elemental Gaussian terms are also unknown parameters to be determined from the data is a difficult but important one. Most of the studies undertaken in the past concern the problem of testing the hypothesis of $(N_g = g_1)$ versus the alternative $(N_g = g_2)$ with the two numbers $1 \leq g_1 \leq g_2$. If the classical likelihood approach is utilized to determine the rest of the parameters in the mixture, the maximum likelihood ratio test (LRT) can be used to determine the actual number of the components in the mixture. The LRT test rejects the hypothesis $H^n_{g_1}$ and decides for $H^n_{g_2}$ whether the likelihood ratio $\lambda = \frac{\Lambda_{g_1}}{\Lambda_{g_2}} \leq 1$ is to small or, equivalently, the log likelihood statistic is too large [43].

Recently, adaptive version of the EM algorithm have also appear in the literature in an attempt to circumvent the problem of determining the number of components in the mixture. The so-called adaptive mixture is essentially a recursively calculated Gaussian mixture with the ability to create new terms or drop existing terms as dictated by the data. In the case of multivariate Gaussian basis functions examined here, a recursive formulation of the EM algorithm can be used to evaluate the number of basis function as well as their parameters at every time instant. The parameter update equations are summarized below:

$$\hat{\tau}_i(k+1) = \frac{\omega_j(k) N(x_{k+1}; \theta_i(k))}{\sum_{i=1}^{N} \omega_j(k) N(x_{k+1}; \theta_i(k))} \tag{1.36}$$

$$\omega_j(k+1) = \omega_j(k) + \frac{1}{n}(\hat{\tau}_i(k+1) - \omega_j(k)) \tag{1.37}$$

$$\mu_i(k+1) = \mu_i(k) + \frac{\hat{\tau}_i(k+1)}{n\omega_j(k)}(x_{k+1} - \mu_i(k)) \tag{1.38}$$

$$\Sigma_i(k+1) = \Sigma_i(k) + \frac{\hat{\tau}_i(k+1)}{n\omega_j(k)}((x_{k+1} - \mu_i(k+1))(x_{k+1} - \mu_i(k+1))^{\tau}) - \Sigma_i(k)) \tag{1.39}$$

with the time index k defined over the interval $k = 1, 2, ..., n$. Given a new data point at a certain time instant k the algorithm either updated the parameters of the existing basis on the mixture by utilizing the equations above or a new term is added to the mixture. The addition of a new term should be based on the utilization of an appropriate measure as to the likelihood that the current measurement has been drawn form the existing model. One such measure

proposed is the Mahalanobis distance between the observation and each of the existing basis in the Gaussian mixture. For the Gaussian basis mixtures considered here the squared Mahalanobis distance between a data point x_j and a Gaussian basis function with mean value μ_i and covariance Σ_i is given as $d_M^2 = (x_j - \mu_i)^\tau {\Sigma_i}^{-1}(x_j - mu_i)$. Thus, if the distance between a new point and each basis function in the current Gaussian mixture exceed a predefined threshold then a new term is created with its mean value given by the location of the point and a covariance which is based on the covariances of the surrounding terms and their mixing coefficients [34]. After the insertion of the new term, the mixing (weighting) coefficients of the Gaussian basis functions are re-normalized appropriately.

1.4 Computer generation of mixture variables

It is of paramount importance in many practical applications to generate random variables which can be described in terms of mixtures. The availability of such techniques will not only help the practitioner to understand the applications of mixtures to a variety of engineering problems but it can also provide insights useful for modifying or extending mixture methodologies.

Let us assume that the mixture model $f(.) = \sum_{j=1}^{Ng} \omega_j f_j(.)$ is available. It can be seen that the mixture is defined in terms of three distinguishable steps:

1. The number of elements present in the mixtures Ng (typically a finite number is selected).
2. The mixture weights ω_j, $j = 1, 2, ..., Ng$ which regulate the contribution of each element in the final outcome.
3. The elements (elemental density functions), $f_j(.)$, $j = 1, 2, ..., Ng$ of the mixture.

To generate a random variable X from a given mixture the following steps should be performed:

1. Generate an element identifier $J = P(J = j) = \omega_j$. In most applications the number N_g of mixture elements is chosen to be 2, in which case the identifier can be simply generated as a result of a comparison of a uniform $(0, 1)$ variable with ω_j. In the case of $g > 2$ the identifier may be generated by one of several discrete variable generating techniques.
2. Generate realizations X_j, $f_j(.)$ for $j = 1, 2, ..., g$.
3. Using steps (1),(2) calculate X, $f(.)$.

By the application of this method the resulting random variable X has the desired distribution $f(.)$ since by construction follows the distribution:

$$\sum_{j=1}^{Ng} f_j(.)P(J = j) = \sum_{j=1}^{Ng} f_j(.)\omega_j = f(.) \tag{1.40}$$

The above described methodology can be utilized to generate random variables from a given mixture model and is used in the simulation studies reporting in this survey.

In this section an application example is used to demonstrate the applicability of the above generation method. The problem selected is that of 'glint noise generation'. In radar target applications the observation noise is highly non-Gaussian. It is well documented in the literature that the so-called 'glint noise' possess the characteristics of a long-tailed distribution [63]-[65]. Conventional minimum mean square estimators, can be seriously degraded if non-Gaussian noise is present. Therefore, it is of paramount importance the accurate modeling of the non-Gaussian noise phenomenon prior to the development of any efficient tracking algorithm. Many different models have been used for the non-Gaussian glint noise present in target tracking applications. Among them a mixture approach, originally proposed by Hewer, Martin and Zeh [64], which argue that the radar glint noise can be modeled as a mixture of background Gaussian noise with outliers. Their results were based on the analysis of the *QQ-plots* of glint noise records [65]. Examination of such records reveals that the glint *QQ-plot* is fairly linear around the origin, an indication that the distribution is Gaussian-like around its mean. However, in the tail region, the plot deviates from linearity and indicates a non-Gaussian, long-tailed character. The data in the tail region is essentially associated with the glint spikes and are considered to be outliers. These outliers have a considerable influence on conventional target tracking filters, such as the Kalman filter which are quite non-robust. The effect of the glint spikes is even greater on the sample variance (covariance) used in the derivation of the filter's gain. It is not difficult to be seen that variances (covariances) which are quadratic functions of the data are more sensitive to outliers than the sample means. Therefore, the glint spikes can be modeled as a Gaussian noise with large variance (covariance) resulting in an overall glint noise model which can be considered as a Gaussian mixture with the two components used to model the background (thermal) Gaussian noise and the glint spikes respectively. The weighting coefficients in the mixture (percentage of contamination) can be used to model the non-Gaussian nature of the glint spikes. Therefore, the glint noise model model can be generated as the mixture of two Gaussian distributions, each with zero mean and with fixed variance (covariance).

In most studies the variances (covariances) are proportional to each other. Assuming that the Gaussian terms are denoted as $N_1(0, \sigma_1)$ and $N_2(0, \sigma_2)$ the mixture distribution has the following form:

$$f(k, \sigma_1, \sigma_2)) = (1-k)N_1(0, \sigma_1) + kN_2(0, \sigma_2) \tag{1.41}$$

with $0 < k < 1$. A random variable X of this distribution can be generated by first selecting un iformly a sample U from the interval $[0, 1]$. If $U > k$ then X is generated by independent sample from $N_1(0, \sigma_1)$. Otherwise, the requested variable X is a sample from $N_2(0, \sigma_2)$. In a first experiment it is assumed that the regulatory coefficient is the unknown parameter in the mixture. The weighting coefficient assumes the values of $k = 0.1$, $k = 0.2$, and $k = 0.3$ respectively. The variances of the two components are given by $\sigma_1 = 1.0$ and $\sigma_2 = 100.0$. The resulting noise profiles can be seen in Fig. 1. In a second experiment we assume that the weighting coefficient in the mixture is known and the only parameter is the variance of the second component in the mixture. We assume that the variance of the first component is fixed, $\sigma_1 = 1.0$. The variance σ_2 of the second component assumes the values of 10.0, 100.0 and 1000.0. By varying the parameters of the second term in the mixture a different noise profile can be obtained. It is evident from Fig. 2 that by increasing the contribution or the variance of the second component in the mixture, the resulting profile deviates more from Gaussian becoming increasingly non-Gaussian.

1.5 Mixture Applications

In this section we will describe, in detail, three areas of application of Gaussian mixture models where the model is used either to represent the underlying physical phenomenon or to assist in the development of an efficient and cost effective algorithmic solution. The application areas considered are those of target tracking in polar coordinates and stochastic estimation for nonlinear systems, non-Gaussian (impulsive) noise modeling and inter-symbol interference rejection, and neural networks for function approximation. These applications were selected mainly due to its importance to the signal processing community. It should be emphasized at this point that Gaussian mixtures have been applied to a number of different areas. Such areas include among other electrophoresis, medical diagnosis and prognosis, econometric applications such as switching econometric models, astronomy, geophysical applications, and applications in agriculture. The interest reader should refer to the extensive summary of references to Gaussian mixture applications provided in [1] for mixture applications.

Apart from that Gaussian mixture models are essential tools in other literatures, such as neural networks, where Radial Basis Functions (RBF) networks and Probabilistic Neural Networks (PNN) are based on Gaussian mixture models, fuzzy systems where fuzzy basis functions are often constructed to imitate the Gaussian mixture model, and image processing/computer vision where Gaussian mixtures can used to model image intensities and to assist in the estimation of the optical flow [66]-[73], [23], [25], [74], [75].

In the next few paragraphs we consider three case studies in order to illustrate the effectiveness of the Gaussian mixture approach in solving difficult signal processing problems. The problem of target tracking in polar coordinates is considered in the next section.

1.5.1 Applications to non-linear filtering

Estimation (filtering) theory has received considerable attention in the past four decades, primarily due to its practical significance in solving engineering and scientific problems. As a result of the combined research efforts of many scientists in the field, numerous estimation algorithms have been developed. These can be classified into two major categories. Namely, linear and nonlinear filtering algorithms corresponding to linear (or linearized) physical dynamic models with Gaussian noise statistics and to nonlinear or non-Gaussian physical models [48], [49]. The most challenging problem arising in stochastic estimation and control is the development of an efficient estimation (filtering) algorithm which can provide estimates of the state of a dynamical system when non-linear dynamic models coupled with non-Gaussian statistics are assumed. We seek therefore, the optimal, in the minimum mean square sense, estimator of the state vector $x(k)$ of a dynamic system which can be described by the following set of equations:

$$x(k+1) = f(x(k), v(k), k) \tag{1.42}$$

$$z(k+1) = h(x(k), w(k), k) \tag{1.43}$$

where $f(.)$ is the nonlinear function which describes the state evolution over time, and $v(k)$ is the state process noise which can be of a non-Gaussian nature. In most cases the state noise is modeled as additive white Gaussian noise with covariance $Q(k)$. The only information available about this system is a sequence of measurements $z(1), z(2), ..., z(k), ...$ obtained at discrete time intervals. The measurement equation (43) describes the observation model which transforms the plant state vector into the measurement space. Most often the observation matrix $h(.)$ it is assumed to be nonlinear with additive measurement noise $w(k)$. The additive measurement noise

is considered to be white Gaussian with noise covariance $R(k)$ and uncorrelated to the state noise process. The initial state vector $x(0)$, which is in general unknown, is modeled as a random variable, Gaussian distributed with mean value $\hat{x}(0)$ and covariance $P(0)$. It is considered uncorrelated to the noise processes $\forall k > 0$.

Given the set of measurements $Z^k = [z(1), z(2), ..., z(k-1), z(k)]$, we desire the *mean-squared-error* optimal filtered estimate $\hat{x}(k|k)$ of $x(k)$:

$$\hat{x}(k|k) = E(x(k)|Z^k) \tag{1.44}$$

of the system state.

For the case of linear dynamics and additive Gaussian noise the problem was first solved by Kalman through his well known filter [49]. The so-called Kalman filter is the optimal recursive estimator for the this case. However, if the dynamics of the system are non-linear and/or the noise processes in (42)-(43) are non-Gaussian, the degradation in the performance of the Kalman filter will be rather dramatic [50].

The requested state estimate in (44) can be obtained recursively through the application of the Bayes theorem as follows:

$$\hat{x}(k|k) = E(x(k)|Z^k) = \int_{-\infty}^{\infty} x(k) f(x(k)|Z^k)\, dx \tag{1.45}$$

$$f(x(k), z(k)|Z^{k-1}) = f(x(k)|z(k), Z^{k-1}) f(z(k)|Z^{k-1}) = f(z(k)|x(k), Z^{k-1}) f(x(k)|Z^{k-1}) \tag{1.46}$$

$$f(x(k)|z(k), Z^{k-1}) = \frac{f(z(k)|x(k), Z^{k-1}) f(x(k)|Z^{k-1})}{f(z(k)|Z^{k-1})} = \frac{f(z(k)|x(k), Z^{k-1}) f(x(k)|Z^{k-1})}{\int f(z(k)|x(k), Z^{k-1}) f(x(k)|Z^{k-1})\, dx(k)} \tag{1.47}$$

Based on the assumptions of the model, the density function $f(x(k)|z(k))$ can be considered as Gaussian with mean value $h(x(k))$ and covariance $R(k)$

$$f(x(k)|z(k)) = \frac{1}{(2\pi)^m} |R(k)|^{-0,5} exp(-0.5||z(k) - h(x(k)||^2_{R^{-1}(k)}) \tag{1.48}$$

In a similar manner the density $f(x(k)|x(x-1))$ can be considered Gaussian with mean value $f(x(k-1))$ and covariance $Q(k-1)$. Given the fact that the initial conditions are assumed Gaussian and thus:

$$f(x(0)|z(0)) = \frac{f(x(0)) f(z(0)|x(0))}{f(z(0))} \tag{1.49}$$

a set of equations which can be used to recursively evaluate the state estimate is now available [48]-[55].

The above estimation problem is solvable only when the density $f(x(k)|z(k))$ can be evaluated for all k. However, this is possible only for a linear state space model and if the a-priori noise and state distributions are Gaussian in nature. In this case, the relations describing the conditional mean and covariance are the well known Kalman Filter equations [54]. To overcome the difficulties associated with the determination of the integrals in (46)-(47) suboptimal estimation procedures have been developed over the years [51]-[55]. The most commonly used involves the assumption that the a-priori distributions are Gaussian and that the nonlinear system can be linearized relative to the latest available state estimate resulting to a Kalman-like filter, the so-called 'Extended' Kalman Filter (EKF). Although EKF performs well in many practical applications, there are numerous situations in which unsatisfactory results have been reported. Thus, a number of different methodologies have been appeared on the literature. Among them, the Gaussian sum filter which

utilizes the approximation theorem reported in section II to approximate (46)-(47). This estimation procedure utilizes a Gaussian mixture to approximate the a-posteriori density $f(x(k)|z(k), Z^{k-1})$ in conjunction with the linearization procedure used in EKF. This so-called Gaussian sum approach assumes that at a certain time instant k the one step ahead predicted density $f(x(k)|Z^{k-1})$ can be written in the form of a Gaussian mixture [51], [52], [54].

Then given the next available measurement and the nonlinear model, the filtering density $f(x(k)|z(k), Z^{k-1})$ is calculated as:

$$f(x(k)|z(k), Z^{k-1}) = c(k)\sum_{i=1}^{Ng} \omega_i N((x(k) - a_i), B_i) f((z(k) - h(x(k))) \tag{1.50}$$

Parallelizing the EKF operation, the Gaussian sum filter linearizes $h(x(k))$ relative to a_i so that $f((z(k) - h(x(k)))$ can be approximated by a Gaussian-like function in the region around each a_i. Once the a-posteriori density $f(x(k)|z(k), Z^{k-1})$ is in the form of a Gaussian mixture, the prediction step of the nonlinear estimator can be performed in the same manner by linearizing $f(x(k+1)|x(k))$ about each term in the Gaussian mixture defined to approximate $f(x(k)|z(k), Z^{k-1})$.

In this review a nonlinear filter based on Gaussian mixture models is utilized to provide efficient, computationally attractive solution to the radar target tracking problem. In tracking applications target motions is usually best modeled in a simple fashion using Cartesian coordinates. However, the target position measurements are provided in polar coordinates (range and azimuth) with respect to the sensor location. Due to the geometry of the problem and the nonlinear relationship between the two coordinate systems, tracking in Cartesian coordinates using polar measurements can be seen as a nonlinear estimation problem, which is described in terms of the following non-linear state space model:

$$x(k+1) = F(k+1, k)x(k) + G(k+1, k)v(k) \tag{1.51}$$

where $x(k)$ is the vector of Cartesian coordinates target states, $F(.)$ is the state transition matrix, $G(.)$ is the noise gain matrix, and $v(k)$ is the system noise process which is modeled as a zero-mean white Gaussian random process with covariance matrix $Q(k)$.

The polar coordinate measurement of the target position is related to the Cartesian coordinate target state as follows:

$$z(k) = h(x(k)) + w(k) \tag{1.52}$$

where $z(k)$ is the vector of polar coordinates measurement, $h(\cdot)$ is the Cartesian-to-polar coordinate transformation, and $w(k)$ is the observation noise process which is assumed to be zero-mean white Gaussian noise process with covariance matrix $R(k)$. Thus, target tracking becomes the problem of estimating the target states $x(k)$ from the noisy polar measurements $z(k)$, $k = 1, 2, \ldots$.

A Gaussian mixture model can be used to approximate the densities involved in the derivation of the optimal Bayesian estimator of (46)-(47) when is applied to the tracking problem.

To evaluate the state prediction density $p(x(k)|Z^{k-1})$ efficiently we will assume the conditional density $p(x(k-1)|Z^{k-1})$ to be Gaussian with mean $\hat{x}(k-1|k-1)$ and covariance matrix $P(k-1|k-1)$. Based on this assumption the state prediction density is a Gaussian density with

$$\hat{x}(k|k-1) = F\hat{x}(k-1|k-1) \tag{1.53}$$

$$P(k|k-1) = FP(k-1|k-1)F^T + GQ(k)G^T \tag{1.54}$$

Given the state space model of the the problem, the function $p(z(k)|x(k))$ can be defined by the measurement equation and the known statistics of the measurement noise $w(k)$

$$\begin{aligned} p(z(k)|x(k)) &= \int p(z(k)|x(k), w(k))p(w(k)|x(k))dw(k) \\ &= \int \delta(x(k) - h(x(k)) - w(k))p_w(w(k))dw(k) \\ &= p_w(x(k) - h(x(k))) \end{aligned} \tag{1.55}$$

Thus, the function $p(z(k)|x(k))$ can be obtained by applying the transformation $w(k) = z(k) - h(x(0))$ to the density function $p_w(w(k))$. Utilizing this observation, we select some initial parameters $\tilde{\alpha}_{k,i}$, $\tilde{m}_{k,i}$ and $\tilde{B}_{k,i}$ from the known statistics of the noise $w(k)$, and transform these parameters from the $w(k)$-space to the $f(k)$-space based on the transformation $w(k) = z(k) - h(x(k))$ $z(k) - h(x(k))$ and finally collect them as a Gaussian mixture approximation for the function $p(z(k)|x(k))$ (see Fig. 3).

The Gaussian mixture procedure used to approximate the non-linear prediction density $p(x(k)|Z^{k-1})$ is summarized as follows:

1. For initialization, select the parameters $\tilde{\alpha}_{k,i}$, $\tilde{m}_{k,i}$ and $\tilde{B}_{k,i}$ for a prescribed value of N such that the following sum-of-squared error is minimized.

$$\sum_{j=1}^{K} \left| p_w(w_{k,jj}) - \sum_{i=1}^{N} \tilde{\alpha}_{k,i} \mathcal{N}(w_{k,j} - \tilde{m}_{k,i}, \tilde{B}_{k,i}) \right|^2 < \epsilon \tag{1.56}$$

   where $w_{k,j} : j = 1, \ldots, K$ is the set of uniformly spaced points distributed through the region containing non-negligible probability and ϵ is the prescribed accuracy.

2. For each new measurement $z(k)$, update the new parameters $\alpha_{k,i}$, $m_{k,i}$ and $B_{k,i}$ such that

$$p(z(k)|x(k)) \approx \sum_{i=1}^{N} \alpha_{k,i} N(m_{k,i} - D(x(k)), B_{k,i}) \tag{1.57}$$

   where

$$\begin{aligned} m_{k,i} &= h^{-1}(\bar{m}_{k,i}) && (1.58) \\ \bar{m}_{k,i} &= z(k) - \tilde{m}_{k,i} && (1.59) \\ B_{k,i} &= \left[J_h(m_{k,i})^T \tilde{B}_{k,i}^{-1} J_h(m_{k,i}) \right]^{-1} && (1.60) \\ \beta_{k,i} &= |J_h(m_{k,i})| && (1.61) \\ \alpha_{k,i} &= \beta_{k,i} \tilde{\alpha}_{k,i} && (1.62) \end{aligned}$$

   Here, we assume the function is invertible; however, if the inverse does not exist, then we must choose $m_{k,i}$ to be the most likely solution given $m_{k,i} = h(\tilde{m}_{k,i})$. Moreover, $J_h(x(k))$, $He_h(m_{k,i})$ are the Jacobian and the Hessian of the function $h(x(k))$ respectively, evaluated as:

$$
\begin{aligned}
J_{F_i}(m_{k,i}) &= \frac{\partial F_i(x_n)}{\partial x_n}\Big|_{x_n=m_{n,i}} \\
&= \frac{1}{2} J_h(m_{n,i})^T \tilde{B}_{n,i}^{-1}(m_{n,i} - h(m_{n,i})) \qquad (1.63) \\
He_{F_i}(m_{n,i}) &= \frac{\partial^2 F_i(x_n)}{\partial x_n \partial x_n^T}\Big|_{x_n=m_{n,i}} \\
&= -\left[He_h(m_{n,i})^T \tilde{B}_{n,i}^{-1}(m_{n,i} - h(m_{n,i})) + J_h(m_{n,i})^T \tilde{B}_{n,i}^{-1} J_h(m_{n,i})\right] \qquad (1.64)
\end{aligned}
$$

Given the form of the approximation the algorithmic description of the non-linear Adaptive Gaussian Sum Filter (AGSF) for one processing cycle is as follows (see Fig. 4):

1. Assume that at time k the mean $\hat{x}(k-1|k-1)$ and the associated covariance matrix $P(k-1|k-1)$ of the conditional density $p(x(k-1)|Z^{k-1})$ are available.

   The predictive mean $\hat{x}(k|k-1)$ and the corresponding covariance matrix $P(k|k-1)$ of the predictive density $p(x(k)|Z^{k-1})$ are determined through (53)-(54) using the state equation of the model.

2. The density $p(x(k)|Z^k)$ is approximated systematically by a weighted sum of Gaussian terms.

3. The Gaussian terms in the mixture are passed to a bank of N Kalman filters which evaluate the parameters for the Gaussian mixture approximation for the density $p(x(k)|Z^k)$.

4. The Gaussian mixture approximation for the density $p(x(k)|Z^k)$ is collapsed into one equivalent Gaussian term with mean $\hat{x}(k|k)$ and covariance $P(k|k)$

A two-dimensional long range target tracking application is simulated to demonstrate the performance of the adaptive Gaussian sum filter on target state estimation. The target trajectory is modeled by the second-order kinematic model of (51) with a process noise of standard variation 0.01 m/s^2 in each coordinate. The measurements are modeled according to equation (52). The standard deviations for range errors is assumed to be 50 m and two standard deviations of bearing error are used $\sigma_\theta = 2.5^o$ and 5.73^o. The parameters of the model are defined as follows:

$$
\begin{aligned}
\mathbf{x}_{n+1} &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}_n + \begin{bmatrix} 1/2 & 0 \\ 1 & 0 \\ 0 & 1/2 \\ 0 & 1 \end{bmatrix} \mathbf{w}_n \qquad (1.65) \\
\mathbf{z}_n &= \begin{bmatrix} \sqrt{x_n^2 + y_n^2} \\ \tan^{-1} y_n/x_n \end{bmatrix} + \mathbf{v}_n \qquad (1.66) \\
\mathbf{Q} &= \begin{bmatrix} 0.0001 & 0 \\ 0 & 0.0001 \end{bmatrix} \\
\mathbf{R} &= (1) \begin{bmatrix} 2500 & 0 \\ 0 & 0.037 \end{bmatrix}, \qquad (2) \begin{bmatrix} 2500 & 0 \\ 0 & 0.01 \end{bmatrix}
\end{aligned}
$$

The adaptive Gaussian sum filter (AGSF) is compared with the Extended Kalman filter (EKF), and the converted measurement Kalman filter (CMKF) in this experiment. All these filters are initialized with the same initial filtered estimate $\hat{\mathbf{x}}_{0|0}$ and the same initial error covariance $\mathbf{P}_{0|0}$

based on the first two measurements. The initial number of Gaussian terms in the preprocessing stage is 30. After preprocessing the number of the Gaussian terms used in the implementation of the AGSF is 9. The results presented here are based on 1500 measurements averaged over 1000 independent Monte Carlo realizations of the experiment with the sampling interval of one second and with two different measurement noise levels. In order to generate the measurement record the initial state $\mathbf{x}_0$ is assumed Gaussian with an average range of 50 km and an average velocity of 20 m/s. For each Monte Carlo realization of the experiment the initial value is chosen randomly from the assumed Gaussian distribution.

The position errors and the velocity errors for the three filters are shown in Figs. 5 and 6 respectively for $\sigma_\theta = 2.5^o$. The error is defined as the root mean square of the difference between the actual value and the estimated value. The Gaussian sum approach converges faster and yields estimates of smaller error than the EKF and the CMKF does. For $\sigma_\theta = 2.5^o$ the CMKF converges faster than the EKF initially but it ceases to converge after the first 400 measurements. The EKF on the other hand is very steady and consistent. As σ_θ increases to 5.72^o $(0.1rad)$ the EKF starts to diverge due to the fact that the EKF is extremely sensitive to the initial filter conditions. When the cross-error gets too large, the wrong set of initial conditions can lead to divergence. The CMKF however it seems to be more robust to inconsistent initial conditions. The AGSF due to its parallel nature and the fact that the Bayes rule operates as a correcting/adjusting mechanism is also in position to compensate for inconsistent initial conditions.

1.5.2 Non-Gaussian noise modeling

The Gaussian mixture density approximation has been extensively used to accomplish practical models for non-Gaussian noise sources in a variety of applications. The appearance of the noise and its effect is related to its characteristics. Noise signals can be either periodic in nature or random. Usually noise signals introduced during signal transmission are random in nature resulting in abrupt local changes in the transmitting sequence. These noise signals cannot be adequately described in terms of the commonly used Gaussian noise model. Rather, they can be characterized as impulsive sequences (interferences) which occur in the form of short time duration, high energy spikes attaining large amplitudes with probability higher than the probability predicted by a Gaussian density model.

These are various sources that can generate such non-Gaussian noise signals. Among others, man made phenomena, such as car ignition systems, industrial machines in the vicinity of the signal receiver, switching transients in power lines and various unprotected electric switches. In addition, natural causes, such as lightning in the atmosphere and ice cracking in the antarctic region also generate non-Gaussian, long-tailed type of noise.

Several models have been used to date to model non-Gaussian noise environments. Some of these models have been developed directly from the underlying physical phenomenon. On the other hand, empirically devised noise models have been used over the years to approximate many non-Gaussian noise distributions. Based on the density approximation theorem presented above, any non-Gaussian noise distribution can be expressed as, or approximated sufficiently well, by a finite sum of known Gaussian pdfs. The Gaussian sum model has been used in the development of approximate empirical distributions which relate to many physical non-Gaussian phenomena.

The most commonly used empirical model is the ϵ-mixture or ϵ-contaminated Gaussian mixture model in which the noise pdf has the form of:

$$f(x) = (1-\epsilon)f_b(x) + \epsilon f_o(x) \tag{1.67}$$

where $\epsilon \in [0,1]$ is the mixture weighting coefficient. The mixing parameter ϵ regulates the contribution of the non-Gaussian component and usually it varies between 0.01 to 0.25.

The $f_b(x)$ pdf is usually taken to be a Gaussian pdf representing background noise. Among the choices for the contaminating pdf are various 'heavy-tailed' distributions, such as the Laplacian, or the double exponential. However, most often f_o is taken to be Gaussian with variance σ_o^2 taken to be many times the variance of f_o, σ_b^2. The ratio $k = \frac{\sigma_o^2}{\sigma_b^2}$ has generally been taken to be between 1 and 10,000. Although the parameters of the mixture model are not directly related to the underlying physical phenomenon, the model is widely used in a variety of applications, primarily due to its analytic simplicity. The flexibility of the model allows for the approximation of many different naturally occurring noise distribution shapes. This approach has been used to model non-Gaussian measurement channels in narrowband interference suppression, a problem of considerable engineering interest [60].

Spread-spectrum communication systems often use estimation techniques to reject narrowband interference. Recently, the interference rejection problem has been formulated as a non-linear estimation problem using a state-space representation [58]. Following the state-space approach, the narrowband interference is modeled as the state trajectory and the combination of the direct-sequence spread spectrum signal with the background noise is treated as non-Gaussian measurement noise.

The basic idea is to spread the bandwidths of transmitting signals so that they are much greater than the information rate. The problem of interest is the suppression of a narrowband interferer in a direct-sequence spread-spectrum (DS/SS) system operating as an N^{th} order autoregressive process of the form:

$$i_k = \sum_{n=1}^{N} \Phi_n i_{k-n} + e_k \tag{1.68}$$

where e_k is a zero mean white Gaussian noise process and $\Phi_1, \Phi_2, ..., ., \Phi_{N-1}, \Phi_N$ are the autoregressive parameters known to the receiver.

The discrete time model arises when the received continuous time signal is passed through an integrate-and-dump filter operating at the chip rate [59].

The Direct Sequence Spread Spectrum (DS/SS) modulation waveform is written as:

$$m(t) = \sum_{k=0}^{N_c-1} c_k q(t - k\tau_c) \tag{1.69}$$

where N_c is the pseudo-noise chip sequence used to spread the transmitted signal and $, q()$ is a rectangular pulse of duration τ_c. he transmitted signal can be then expressed as:

$$s(t) = \sum_{k} b_k m(t - kT_b) \tag{1.70}$$

where $b(k)$ is the binary information sequence and $T_b = N_c \tau_c$ is the bit duration. Based on that, the received signal is defined as:

$$z(t) = as(t - \tau) + n(t) + i(t) \tag{1.71}$$

where a is an attenuation factor, τ is a delay offset, $n(t)$ is wideband Gaussian noise and $i(t)$ is narrow-band interference. Assuming that $n(t)$ is band-limited and hence white after sampling, with $\tau = 0$ and $a = 1$ for simplicity, if the received signal is chip-matched and sampled at the chip

rate of the pseudo-noise sequence, the discrete time sequence resulting from the continuous model above can be re-written as follows:

$$z(k) = s(k) + n(k) + i(k) \tag{1.72}$$

The system noise contains an inference component $i(k)$ and a thermal noise component $n(k)$. We assume binary signaling and a processing gain of K chips/bit so that during each bit interval, a pseudo-random code sequences of length K is transmitted. The code sequences can be denoted as:

$$S^K = [s_1(1), s_1(2), ..., s_1(K)] \tag{1.73}$$

with $s_1 \in (+1, -1)$.

Based on this, a state space representation for the received signal and the interference can be constructed as follows:

$$\begin{aligned} x(k) &= \Phi x(k-1) + v(k) \\ z(k) &= Hx(k) + w(k) \end{aligned} \tag{1.74}$$

with $x(k) = [i_k, i_{k-1}, ..., i_{k-N+1}]^\tau$, $v(k) = [e_k, 0, ..., 0]^\tau$, $H = [1, 0, ..., 0]$, and

$$\Phi = \begin{vmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_N \\ 1. & 0. & \cdots & 0. \\ \cdots & \cdots & \cdots & \cdots \\ 0. & 0. & \cdots & 1. \end{vmatrix}$$

.

The additive observation noise $w(k)$ in the state space model is defined as:

$$v(k) = n(k) + s(k)$$

Since the first component of the system state $x(k)$ is the interference $i(k)$, an estimate of the state contains an estimate of $i(k)$ which can be subtracted from the received signal in order to increase the system's performance. The additive observation (measurement) noise $v(k)$ is the sum of two independent variables, one is Gaussian distributed and the other takes on values -1 or -1 with equal probability. Therefore its density is the weighted sum of two Gaussian densities (Gaussian sum) [59], [60]:

$$f(w(k)) = (1-\epsilon)N(\mu, \sigma_n^2) + \epsilon N(-\mu, \lambda\sigma_n^2) \tag{1.75}$$

with $\epsilon = 0.5$ and $\mu = 1$.

In summary, the narrowband interference is modeled as the state trajectory and the combination of the DS/SS signal and additive Gaussian noise is treated as non-Gaussian measurement noise. Nonlinear statistical estimators can be used then to estimate the narrowband interference and to subtract it from the received signal. Due to the nature of the non-Gaussian measurement noise a non-linear filter should be used to provide the estimates. The non-linear filter takes advantage of the Gaussian-mixture representation of the measurement noise to provide on-line estimates of the inter-symbol interference. By collapsing the Gaussian mixture at every step through the utilization of the Bayes theorem, a Kalman-like recursive filter with constant complexity can be devised.

For the state space model of (72) if the measurement noise is expressed in terms of the Gaussian mixture of (73) an estimate $\hat{x}(k|k)$ of the system state $x(k)$ at time instant k, can be computed recursively by an Adaptive Gaussian Sum Filter (AGSF) as follows:

$$\hat{x}(k|k) = \hat{x}(k|k-1) + K(k)(z(k) - \hat{z}(k|k-1) \tag{1.76}$$

$$P(k|k) = (I - K(k)H(k))P(k|k-1) \tag{1.77}$$

$$\hat{x}(k|k-1) = \Phi(k, k-1)\hat{x}(k-1|k-1) \tag{1.78}$$

$$P(k|k-1) = \Phi(k, k-1)P(k-1|k-1)\Phi(k, k-1)^{\tau} + Q(k-1) \tag{1.79}$$

with initial conditions $\hat{x}(0|0) = \hat{x}(0)$ and $P(0|0) = P(0)$.

$$K(k) = P(k|k-1)H^{\tau}(k|k-1)P_z^{-1}(k|k-1) \tag{1.80}$$

$$\hat{z}(k|k-1) = \sum_{i=1}^{Ng} \omega_i(k)\hat{z}_i(k|k-1) \tag{1.81}$$

$$\hat{z}_i(k|k-1) = H(k)\hat{x}(k|k-1) + \mu_i \tag{1.82}$$

$$P_{zi}(k|k-1) = H(k)P(k|k-1)H^{\tau}(k) + R_i \tag{1.83}$$

In case of (73), $Ng = 2$, with $\mu_i = \mu$ and $R_1 = \sigma_n^2$, $R_2 = \lambda\sigma_n^2$.

The corresponding innovation covariance and the a-posteriori weights used in the Bayesian decision module are defined as:

$$P_z(k|k-1) = \sum_{i=1}^{Ng} (P_{zi}(k|k-1) + (\hat{z}(k|k-1) - \hat{z}_i(k|k-1))(\hat{z}(k|k-1) - \hat{z}_i(k|k-1))^{\tau})\omega_i(k) \tag{1.84}$$

$$\omega_i(k) = \frac{((2\pi)^{-m}|P_{zi}|^{-1}exp(-0.5(||z(k) - \hat{z}_i(k|k-1)||^2_{P_{zi}^{-1}(k|k-1)})))a_i}{c(k)} \tag{1.85}$$

where, $|.|$ denotes the determinant of the matrix, and $||.||$ the inner product. The parameter a_i are the initial weighting coefficients used in Gaussian mixture which describes the additive measurement noise. In case of (73) $a_1 = (1 - \epsilon)$ and $a_2 = \epsilon$.

Finally, the normalization factor $c(k)$ is calculated recursively as follows:

$$c(k) = \sum_{i=1}^{Ng} ((2\pi)^{-m}|P_{zi}|^{-1}exp(-0.5(||z(k) - \hat{z}_i(k|k-1)||^2_{P_{zi}^{-1}(k|k-1)})))a_i \tag{1.86}$$

Simulation results are included here to demonstrate the effectiveness of such an approach. In this study the interferer is found by channeling white noise through a second-order infinite-duration impulse response (IIR) with two poles at 0.99:

$$i_k = 1.98i_{k-1} - 0.9801i_{k-2} + e_k \tag{1.87}$$

where e_k is zero mean white Gaussian noise with variance 0.01. The regulatory coefficient ϵ used in the Gaussian mixture of (73) is set to be $\epsilon = 0.2$ and the ratio λ is taken to be $\lambda = 10$

or $\lambda = 10,000$ with $\sigma_n = 1.0$. The non-Gaussian measurement noise profile, for a single run, is depicted in Fig. 7 ($\lambda = 10$) and in Fig. 10 for $\lambda = 10,000$.

The normalized mean square error (*NMSE*) is utilized for filter comparison purposes in all experiments. The data were averaged through Monte Carlo techniques. Given the form of the state vector, the first component of $x(k)$ is used in the evaluation analysis. The NMSE is therefore defined as:

$$NMSE = \frac{1}{MCRs}\left(\sum_{k=1}^{MCRs} \frac{\left(x^k{}_{1r} - \hat{x}^k_{1j}\right)^2}{x^{k2}_{1r}}\right)$$

where MCRs is the number of Monte Carlo runs, x_{1r} the actual value and $\hat{x}_{1j}$ is the outcome of the j-filter under consideration.

In this experiment, 100 independent runs (Monte Carlo runs), each 1000 samples in length were considered. Due to its high complexity and the unavailability of suitable nonlinear transformation for the 'score function' the Masreliez filter was not included in these simulation studies.

Two different plot types are reported in the paper. First, state estimation plots for single Monte Carlo runs are included to facilitate the performance of the different estimation schemesi (Figs. 8, 11). In addition, the normalized mean square error plots for all the simulation studies are also reported (Figs. 9, 12). From the plots included in the chapter we can clearly see the improvement accomplished by the utilization of the new filter versus the Kalman filter and the Masreliez filter. The effects have appeared more pronounced at more dense non-Gaussian (impulsive) environments. This trend was also verified during the error analysis utilizing the Monte Carlo error plots (Figs. 9, 12).

1.5.3 Radial Basis Function Networks

Although Gaussian mixtures have been used for many years in adaptive signal processing, stochastic estimation, statistical pattern recognition, Bayesian analysis, and decision theory only recently have been considered by the neural networks community as a valuable tool for the development of a rich class of neural nets, the so-called Radial Basis functions (RBF) networks [72]. RBF networks can be used to provide an effective and computationally efficient solution to the interpolation problem. In other words given a sequence of (n) available data points $X = (x_1, x_2, ..., x_n)$ (which can be vectors) and the corresponding (n) measurement values $Y(y_1, y_2, ..., y_n)$ the objective is to define a function F satisfying the interpolation condition $F(x_i) = y_i$, $i = 1, 2, ..., n$. The RBF neural approach consists of choosing F from a linear space of dimension (n) which depends on the data points x_i [73]. The basis of this linear space is chosen to be the set of radial functions. Radial functions are a special class of functions in which their response decreases or increases monotonically with distance from a central point. The central point, the distance scale as well as the shape of the radial function are parameters of the RBF neural model. Although many radial functions have defined and used in the literature, the typical one is the Gaussian which, in the case of a scalar input, is defined as:

$$f(x; c, r) = exp\left(-\frac{(x-c)^2}{r^2}\right) \tag{1.88}$$

with parameters the center c and the radius r. A single layer network consisting of such Gaussian basis functions is usually called Radial Basis Function (RBF) net in the neural network literature. Optimization techniques can be used to adjust the parameters of the basis functions in order to achieve better results. Assuming that the number of basis (Gaussian) functions is fixed the

interpolation problem is formulated as follows:

$$F(x) = \sum_{i=1}^{Ng} \omega_i f(x; c_i, r) \tag{1.89}$$

Although the number Ng of elemental Gaussian terms in the mixture expression can be defined a-priori, it can also be considered as a parameter. In such a case the smallest possible number of Gaussian bases is targeted.

In this setting the problem is the equivalent of solving of set of $(3N_g)$ nonlinear equations using (n) data points. Thus, the problem is to determine the Gaussian centers and radius along with the mixture parameters from the sample data set.

One of the most convenient ways to implement this is to start with an initial set of parameters and then iteratively modified them until a local minimum is reached in the error function between the available data set and the approximating Gaussian mixture.

However, defining a smooth curve from available date is an ill-posed problem in the sense that the information in the data may not be sufficient to uniquely reconstruct the function mapping in regions where data samples are not available. Moreover, if the available data set is subject to measurement errors or stochastic variations, additional steps, such as introduction of penalty terms in the error function are needed in order to guarantee good results. In a general $d - dimensional$ space the Gaussian radial basis can be as $f(x) = exp(-0.5||x - \mu_i||_{\Sigma_i^{-1}})$ where μ_i and Σ_i represent the mean vector and the covariance matrix of the i^{th} radial basis function.

The quadratic term in the Gaussian basis function form can be written as an expanded form

$$||x - \mu_i||_{\Sigma_i^{-1}} = \sum_{k=1}^{d} \sum_{j=1}^{d} \lambda_{ikj}(x_j - \mu_{ij})(x_k - \mu_{ik}) \tag{1.90}$$

with μ_{ij} the j^{th} element of the mean vector μ_i and λ_{kj} the (j,k) element of the shape matrix Σ_i^{-1}. The elements of the shape function can be evaluated in terms of the marginal standard deviations σ_{ij}, σ_{ik} and the correlation coefficient. Assuming that the shape matrix is positive diagonal, a much simpler expression can be obtained. In such a case, the output of the i^{th} Gaussian basis function can be defined as:

$$o_i = exp(-0.5 \sum_{k=1}^{d} \frac{(x_k - \mu_{ik})^2}{\sigma_{ik}}) \tag{1.91}$$

with $1 \leq i \leq Ng$.

The output of the i^{th} Gaussian basis function forms a hyper-ellipsoid in the $d-$ dimensional space with the mean and the variance the parameters which determine the geometric shape and the position of that hyper-ellipsoid. Therefore, the Radial basis network consists of an array of Gaussian functions determined by some parameter vectors [68].

$$F(x) = \sum_{i=1}^{Ng} \omega_i exp(-0.5 \sum_{k=1}^{d} \frac{(x_k - \mu_{ik})^2}{\sigma_{ik}}) \tag{1.92}$$

Radial basis function networks have extensively be used to approximate non-linear functions [78]. In most cases single hidden layer structures with Gaussian units are used due to their simplicity and fast training. To demonstrate the function approximation capabilities of the RBF network, a simple scalar example is considered. The RBF network consists of 5 Gaussian units equally

weighted. Figure 13 depicts the initial placement of the five Gaussian terms, as well as the overall function to be approximated. It can be seen from the plot that the basis functions are equally distributed on the interval $[50 - 200]$. Figure 14 depicts the final location of the Gaussian basis functions. The unequal weights and the shifted placement of the basis functions provides an efficient and cost effective approximation to the original function.

The deterministic function approximation approach is probably not the best way to characterize an RBF network when the relationship between the input and output parameters is a statistical rather a deterministic one. In was suggested in [79] that in this case it is better to consider the input and output pair $x, F(x)$ as realizations of random vectors which are statistically dependent. In such a case, if a complete statistical description of the data is available, the output value can be estimated given only the input values. However, since complete statistical description is seldom available in most cases the optimal statistical estimator cannot be realized. One way to overcome the problem is to assume a certain parametric model and use the data to construct a model which fits the data reasonably well [80]. A number of different neural network based on parametric modeling of data have been proposed in the literature. Among them the so-called probabilistic neural networks (PNN) [25], [73] and the Gaussian-mixture (GM) model of [81], [80]. The GM model is a parametric probabilistic model based on the the Gaussian mixture model discussed through out this chapter. In the context of GM it is assumed that the available input/output pairs result from a mixture of N_g populations of Gaussian random vectors, each one with a probability of occurrence of ω_i, $i = 1, ..., N_g$. Given that assumption a Gaussian-mixture basis function network (GMBFN) [80] can be used to provide estimate of the output variable given a set of input values and the set of N_g Gaussian bases. The GMBFN parallelizes the Gaussian mixture models used in the development of non-linear statistical estimators. Parameter estimation techniques, such as the EM algorithm discussed in this survey can be used to estimate the parameters of the GMBFN model during training. The GMBFN network can be viewed as the link between the Gaussian mixture models used in statistical signal processing and the RBF networks used for function approximation. This type of networks has been shown to have good approximation capabilities in nonlinear mappings and has been proven to provide efficient solutions in application problems, such as channel equalization and image restoration.

1.6 Concluding Remarks

In this article we reviewed some of the issues related to the Gaussian mixture approach and its applications to signal processing. Due to the nature of the Gaussian mixture model special attention was given to nonlinear, non-Gaussian signal processing applications. Novel signal processing techniques were developed to provide effective, simple, and computationally attractive solutions in important application problems, such as target tracking in polar coordinates and interference rejection in impulsive channels. Emphasis was also given on theoretical results, such as the approximation theorem and the EM algorithm for mixture parameter estimation. Although these issues are not related to any particular practical application, they can provide the practitioner with the necessary tools needed to support a successful application of Gaussian mixtures.

The authors' intention was to illustrate the applicability of the Gaussian mixture methodology in signal processing applications and to highlight the similarities between Gaussian mixture models used in statistical signal processing and neural network methodologies, such as RBF used in function approximation and optimization. Since mixture model analysis yields a large number of theorems, methods, applications and test procedures, there is much pertinent theoretical work as well as research on Gaussian mixture applications which has been omitted for reasons of space and time.

Apart from the practical problems discussed here there are a large class of problems that appear to be amenable to solution by Gaussian mixtures. Among them emerging areas of significant importance, such as data mining, estimation of video flow and modeling of (computer) communication channels. It is the authors' belief that Gaussian mixture models provide effective tools foe these emerging signal processing applications and thus surveys on Gaussian mixture analysis and applications can contribute to further advances in these emerging research areas.

Bibliography

[1] D.M. Tittirington, A.F.M. Smith, U.F. Makov, **Statistical Analysis of Finite Distributions**,N.Y. Willey, 1985.

[2] G.J. McLachlam, K.E. Basford, **Mixture Models: Inference and Applications to Clustering**, Marcel dekker, 1988.

[3] B.S. Everitt, D.J. Hand, **Finite Mixture Distributions**, Chapman and Hall, London, 1981.

[4] R.O. Duda, P.E. Hart, **Pattern Recognition and Scene Analysis**, Wiley, N.Y., 1973.

[5] E.A. Patrick, **Fundamentals of Pattern Recognition**, Prentice-Hall, 1972.

[6] R.S. Bucy, P.D. Joseph, **Filtering for Stochastic Process with Application to Guidance**, Interscience, New York, 1968.

[7] J. Koreyaar, **Mathematical Methods, Vol. 1**, pp. 330-333, Academic Press, 1968.

[8] W. Feller, **An Introduction to Probability and Its Applications, Vol. II**, p. 249, John Wiley, 1966.

[9] M. Aitkin, D.B. Rubin, 'Estimation and hypothesis testing in finite mixture models', J. R. Stat. Soc. B, vol. 47, pp. 67-75, 1985.

[10] K.E. Basford, G.J. McLachlan, 'Likelihood estimation with normal mixture models', Appl. Statistics, vol. 34, pp. 282-289, 1985.

[11] J. Behboodian, 'On a mixture of normal distributions', Biometrika, vol. 57, pp. 215-217, 1970.

[12] J.G. Fryer, C.A. Robertson, 'A comparison of some methods for estimating mixed normal distributions', Biometrika, vol. 59, pp. 639-648, 1972.

[13] A.S. Zeevi, R. Meir, 'Density estimation through convex combination of densities: Appr oximation and estimation bounds', Neural Networks, vol. 10, no. 1, pp. 99-109, 1997.

[14] S.S. Gupta, W.T. Huang, 'On mixtures of distributions: A survey and some new results on ranking and selection', Sankhya B, vol. 43, pp. 245-290, 1981.

[15] V. Husselblad, 'Estimation of finite mixtures of distributions from exponential family', Journal of American Statistical Association, vol. 64, pp. 1459-1471, 1969.

[16] R.J. Hathaway, 'Another interpretation of the EM algorithm for mixture distributions', Statistics & Probability Letters, vol. 4, pp. 53-56.

[17] R.A. Maronna, 'Robust M-estimators of multivariate location and scatter', Ann. Statist. vol. 4, pp. 51-67, 1976.

[18] R.A. Render, H.F. Walker, 'Mixture densities, maximum likelihood and the EM algorithm', SIAM Review, pp. 195-293, 1984.

[19] T. Hasties, R. Tibshitani, 'Discriminant analysis by Gaussian mixtures', to appear J. R. Stat. soc. B, 1996.

[20] S.J. Yakowitz, J.D. Sprangins, 'On the identifiability of finite mixtures', Ann. Math. Stat., vol. 39, pp. 209-214, 1968.

[21] S.J. Yakowitz, 'Unsupervised learning and the identification of finite mixtures', IEEE Trans. on Information Theory, vol. IT-16, pp. 258-263, 1970.

[22] S.J. Yakowitz, 'A consistent estimator for the identification of finite mixtures', Ann. Math. Stat., vol. 40, oo. 1728-1735, 1968.

[23] H.G.C. Traven, 'A neural network approach to statistical pattern classification by semiparametric estimation of probability density function', IEEE Trans. on Neural Networks, vol. 2, no. 3, pp. 366-377, 1991.

[24] H. Amindavar, J.A. Ritchey, 'Pade approximations of probability functions', IEEE Trans. on Aerospace and Electronics Systems, vol. AES-30, pp. 416-424, 1994.

[25] D.F. Specht, 'Probabilistic neural networks', Neural Networks, vol. 3, pp. 109-118, 1990.

[26] T.Y. Young, G. Copaluppi, 'Stochastic estimation of a mixture of normal density functions using an information criterion', IEEE Trans. on Information Theory, vol. IT-16, pp. 258-263, 1970.

[27] J.C. Rajapakse, J.N. Gieldd, J.L. Rapaport, 'Statistical approach to segmentation of single-channel cerebral MR images', IEEE Trans. on Medical Imaging, vol. 16, no. 2, pp. 176-186, 1997.

[28] P. Schroeter, J.M. Vesin, T. Langenberger, R. Meuli, 'Robust parameter estimation of intensity distributions for brain magnetic resonance images, IEEE Trans. on Medical Imaging, vol. 27, no. 2, pp. 172-186, 1998.

[29] J.C. Rajapakse, F. Kruggel, 'Segmentation of MR images with intensity inhomogeneities', Image and Vision Computing, vol. 16, no. 3, pp. 165-180, 1998.

[30] S.G. Sanjay, T.J. Hebert, 'Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm', IEEE Trans. on Image Processing, vol. 7, no. 7, pp. 1024-1028, 1998.

[31] H. Teicher, 'Identifiability of finite mixtures', Ann. Stat., vol. 34, pp. 1265-1269, 1963.

[32] A.P. Dempster, N.M. Laird, D.B. Rubin, 'Maximum likelihood from incomplete data via the EM algorithm', J.R. Stat. Soc. B, vol. 39, pp. 1-38, 1977.

[33] T.K. Moon, 'The Expectation-Maximization algorithm', IEEE Signal Processing Magazine, pp. 47-60, 1997.

[34] C.E. Priebe, 'Adaptive Mixtures', Journal of the American Statistical Association, vol. 89, pp. 796-806, 1994.

[35] J. Diebolt, C. Robert, 'Estimation of finite mixture distributions through Bayesian sampling', J. Roy. Statist. Soc. B, vol. 56, pp. 363-375, 1994.

[36] M. Escobar, M. West, 'Bayesian density estimation and inference using mixtures', i Journal of the American Statistical Association, vol. 90, pp. 577-588, 1995.

[37] D.M. Titterington, 'Some recent research in the analysis of mixture distribution', Statistics, vol. 21, pp. 619-640, 1990.

[38] P. McKenzie, M. Alder, 'Initializing the EM algorithm for use in Gaussian mixture modelling', in Pattern Recognition in Practice IV, E.S. Gelsema and L.N. Kanal Editors, pp. 91-105, 1994.

[39] A. Witkin, 'Scale space filtering', in Proceedings of the International Joint Conference on artificial Intelligence, IJCAI-83, pp. 1019-1022, 1983.

[40] M.J. Carlotto, 'Histogram analysis using a scale space approach', IEEE Trans. on Pattern Recognition and Machine Intelligence, vol. PAMI-9, no.1, pp. 121-129, 1987.

[41] A. Goshtasby, W.D. O'Neill, 'Curve fitting by a sum of Gaussians', Graphical Models and Image Processing, vol. 56, no.4, pp. 281-288, 1994.

[42] G. Celeux, J. Diebolt, 'The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem'. Computational Statistics Quarterly, vol. 2, pp. 35-52, 1986.

[43] H.H. Bock, 'Probabbility models and hypotheses testing in partitioning cluster analysis', in *Clustering and Classification*, P. Arabie, L.J. Hubert, G. DeSoete (Eds.), pp. 377-453, Word Scientific Publishers, 1996.

[44] J.L. Solka, W.L. Poston, E.J. Wegman, B.C. Wallet, 'A new iterative adaptive mixture type estimator', to appear in the Proceedings of the 28^{th} Symposium on the Interface, 1996.

[45] C.E. Priebe, D.M. Marchette, 'Adaptive mixtures: Recursive nonparametric pattern recognition', Pattern Recognition, vol. 24, pp. 1197-1209, 1991.

[46] D.B. Cooper, P.W. Cooper, 'Nonsupervised adaptive signal detection and pattern recognition', Information and Control, vol. 7, pp. 416-444, 1964.

[47] B.S. Everitt, **Graphical techniques for multivariate data**, Heinemann, London, 1978.

[48] R.S. Bucy, 'Linear and non-linear filtering', Proc. IEEE, vol. 58, pp. 854-864, 1970.

[49] D.G. Lainiotis, 'Partitioning: A unifying framework for adaptive systems I: Estimation,' Proceedings of IEEE, vol. 64, pp. 1126-1143, 1976.

[50] R.S. Bucy, K.D. Senne, 'Digital synthesis of non-linear filters', Automatica, vol. 7, pp. 287-298, 1971.

[51] H.W. Sorenson, D.L. Alspach, 'Recursive Bayesian estimation using Gaussian sums', Automatica, vol. 7, pp. 465-479, 1971.

[52] H.W. Sorenson, A.R. Stubberud, 'Nonlinear filtering by approximation of a-posteriori density', Int. J. Control, vol. 18, pp. 33-51, 1968.

[53] T. Numera, A.R. Stubberud, 'Gaussian sum approximation for non-linear fixed point prediction', Int. J. Control, vol. 38, pp. 1047-1053, 1983.

[54] D.L. Alspach, 'Gaussian sum approximations in nonlinear filtering and control', Information Sciences, vol. 7, pp. 271-290, 1974.

[55] T.S. Rao, M. Yar, 'Linear and non-linear filters for linear, but non-Gaussian processes', Int. J. Control, vol. 39, pp. 235-246, 1983.

[56] D. Lerro, Y. Bar-Shalom, 'Tracking with debiased consistent converted measurements versus EKF', IEEE Trans. on Aerospace and Electronic Systems, vol. 29, no. 3, pp. 1015-1022, 1993.

[57] Wing Ip Tam, K.N. Plataniotis, D. Hatzinakos, 'An adaptive Gaussian sum algorithm for target tracking', Signal Processing, vol. 77, no. 1, pp. 85-104, August 1999.

[58] R. Vijayan, H.V. Poor, 'Nonlinear techniques for interference suppression in spread-sp ectrum systems', IEEE Trans. on Communication, vol. COM-38, pp. 1060-1065, 1990.

[59] K.S. Vastola, 'Threshold detection in narrowband non-Gaussian noise,' IEEE Trans. on Communication, vol. COM-32, pp. 134-139, 1984.

[60] L.M. Garth, H.V. Poor, 'Narrowband interference suppression in impulsive environment', IEEE Trans. on Aerospace and Electronics Systems, vol. AES-28, pp. 15-33, 1992.

[61] C.J. Masreliez, 'Approximate non-Gaussian filtering with linear state and observation relations', IEEE Trans. on Automatic Control, vol. AC-20, pp. 107-110, 1975.

[62] Wen R. wu, A. Kundu, 'Recursive filtering with non-Gaussian noises', IEEE Trans. on Signal processing, vol. 44, no. 4, pp. 1454-1468, 1996.

[63] W.R. Wu, P.P. Cheng, 'A nonlinear IMM algorithm for maneuvering target tracking', IEEE Trans. on Aerospace and Electronics Systems, vol. AES-30, pp. 875-885, 1994.

[64] G.A. Hewer, R.D. Martin, J. Zeh, 'Robust preprocessing for Kalman filtering of glint noise', IEEE Trans. on Aerospace and Electronic Systems, vol. AES-23, pp. 120-128, 1987.

[65] Z.M. Durovic, B.D. Kovacevic, 'QQ-plot approach to robust Kalman filtering', Int. J. of Control, vol. 61, no. 4, pp. 837-857, 1994.

[66] A.R. Webb, 'Functional approximation by feed-forward networks: A least squares approach to generalization', IEEE Trans. on Neural Networks, vol. 5, pp. 363-371, 1994.

[67] J. Mooddy, C.J. Darken, 'Fast learning in networks of locally-tuned processing units', Neural Computation, vol. 1, pp. 281-294, 1989.

[68] L. Jin, M.M. Gupta, P.N. Nikiforuk, 'Neural networks and fuzzy basis functions for functional approximation', in **Fuzzy logic and Intelligent Systems**, H. Li, M.M. Gupta (Eds.), Kluwer Academic Publishers, 1996.

[69] T. Poggio, F. Girosi, 'Networks for approximation and learning', Proc. of IEEE, vol. 78, pp. 1481-1497, 1990.

[70] D.a. Cohn, Z. Ghrahramani, M.I. Jordan, 'Active learning with statistical models', Journal of Artificial Intelligence Research, vol. 4, pp. 129-145, 1996.

[71] M.I. Jordan, C.M. Bishop, 'Neural Networks', Mach. Int. of Technology, A.I. Memo No. 1562, 1996.

[72] B. Mulgrew, 'Applying Radial Basis Functions', IEEE Signal Processing Magazine, pp. 50-65, 1996.

[73] H.M. Kim, J.M. Mendel, 'Fuzzy basis functions: Comparisons with other basis functions', IEEE Trans. on Fuzzy Systems, vol. 3, pp. 158-168, 1995.

[74] A. Jepson, M. Black, 'Mixture models for image representation', Technical Report ARK96-PUB-54, Department of Computer Science, University of Toronto, March 1996.

[75] P. Kontkane, P. Myllymaki, H. Tirri, 'Predictive data mining with finite mixtures', Proceedings of the 2^{nd} International Conference on Knowledge iDiscovery and data Mining, pp. 176-182, 1996.

[76] I. Caballero, C.J. Pantaleon-Prieto, A. Artes-Rodriguez, 'Sparce deconvolution using adaptive mixed-Gaussian models', Signal processing, vol. 54, pp. 161-172, 1996.

[77] Y. Zhao, X. Zhuang, S.J. Ting, 'Gaussian mixture density modeling of non-Gaussian source for autoregressive process', IEEE Trans. on Signal processing, vol. 43, no. 4, pp. 894-903, 1995.

[78] J. Park, I.W. Sandberg, 'Universal approximation using radial-basis function networks', Neural Computations, vol. 3, pp. 246-257, 1991.

[79] I. Cha, S.A. Kassam, 'Gaussian-mixture basis function networks for nonlinear signal processing', Proceedings 1995 Workshop on Nonlinear Signal Processing, vol. 1, pp. 44-47, 1995.

[80] I. Cha, S.A. Kassam, 'RBNF restoration of nonlinear degraded images', IEEE Trans. on Image Processing, vol. 5, no. 6, pp. 964-975, 1996.

[81] R.A. Render, R.J. Hathaway, J.C. Bezdeck, 'Estimating the parameters of mixture models with modal estimators', Commun. Stat. Part A: Theory and Methods, vol. 16, pp. 2639-2660, 1987.

1.7 List of Figures

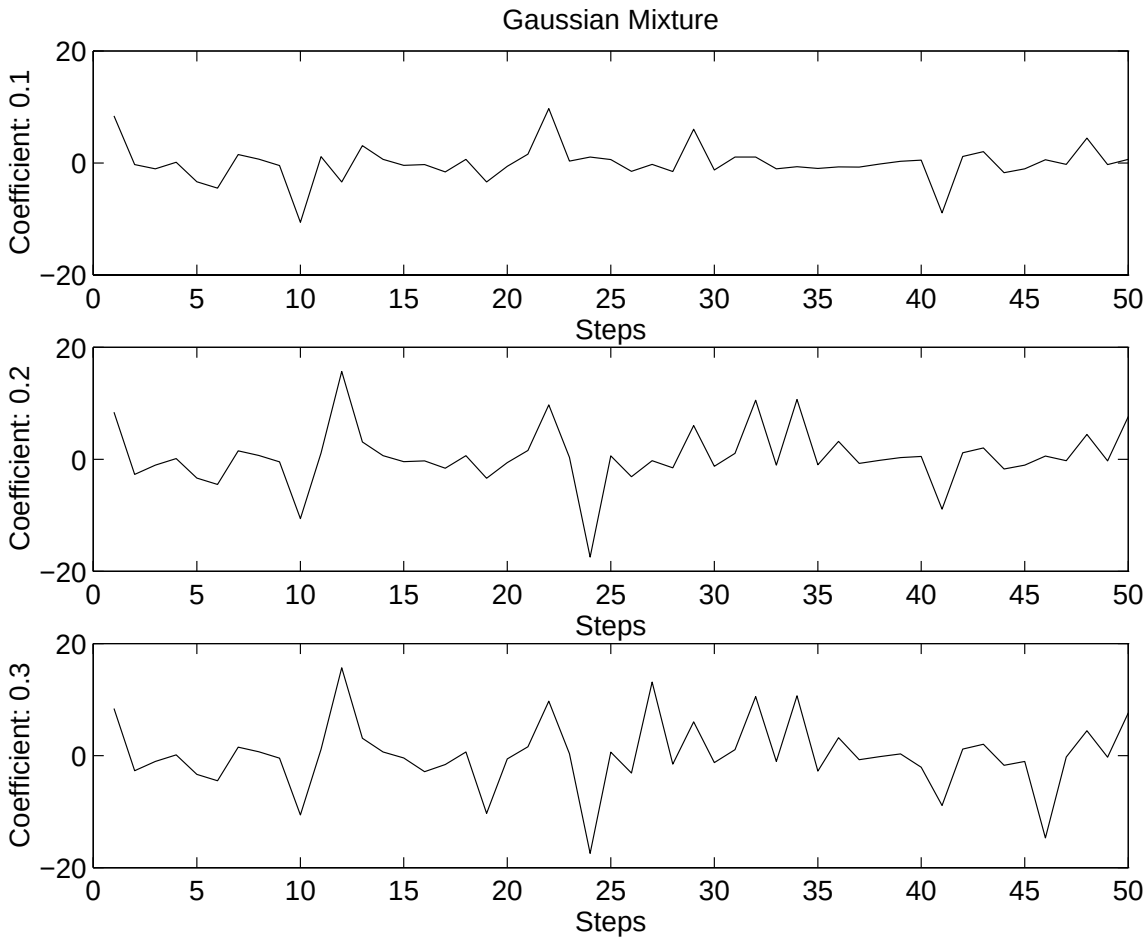

Figure 1.1: Gaussian mixture generation: The Effect of the Weighting Coefficient

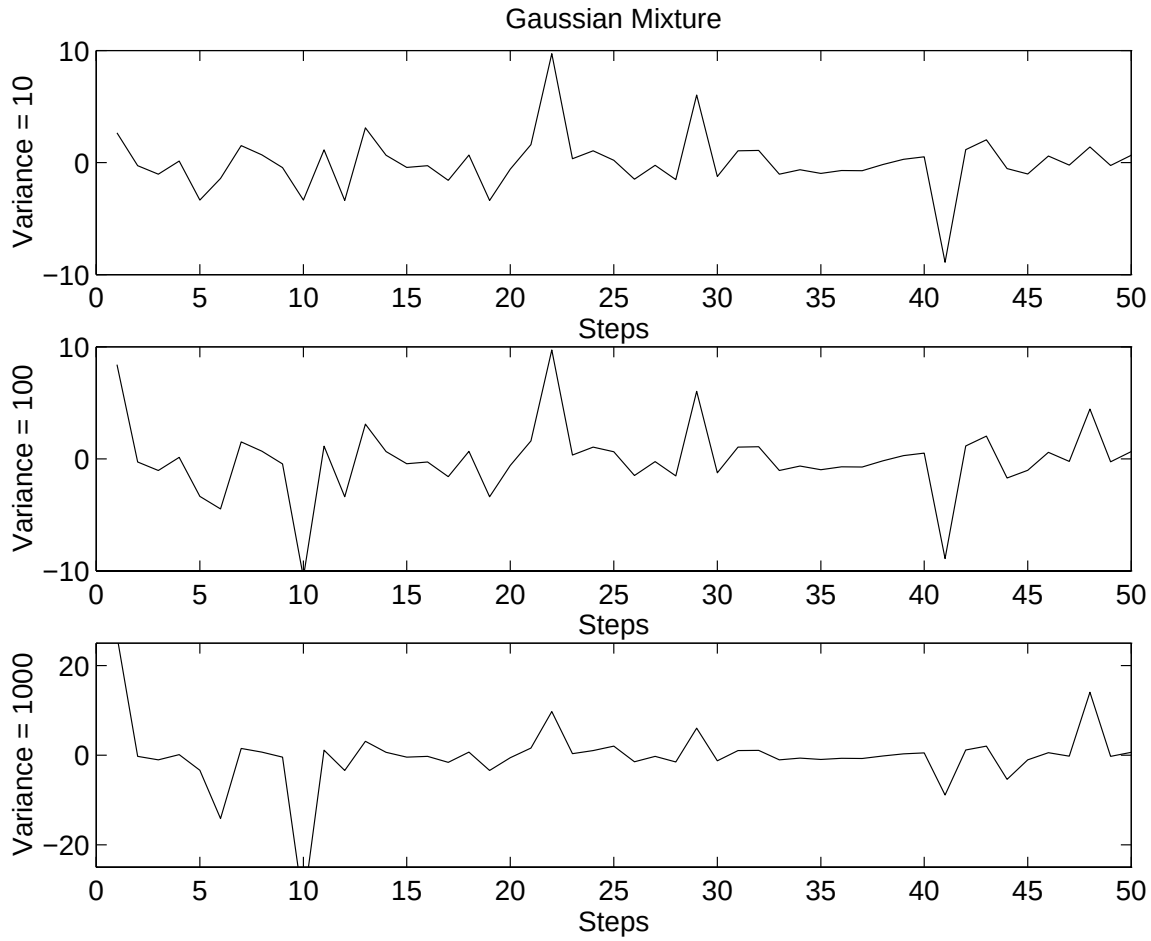

Figure 1.2: Gaussian Mixture Generation: The Effect of the Variance

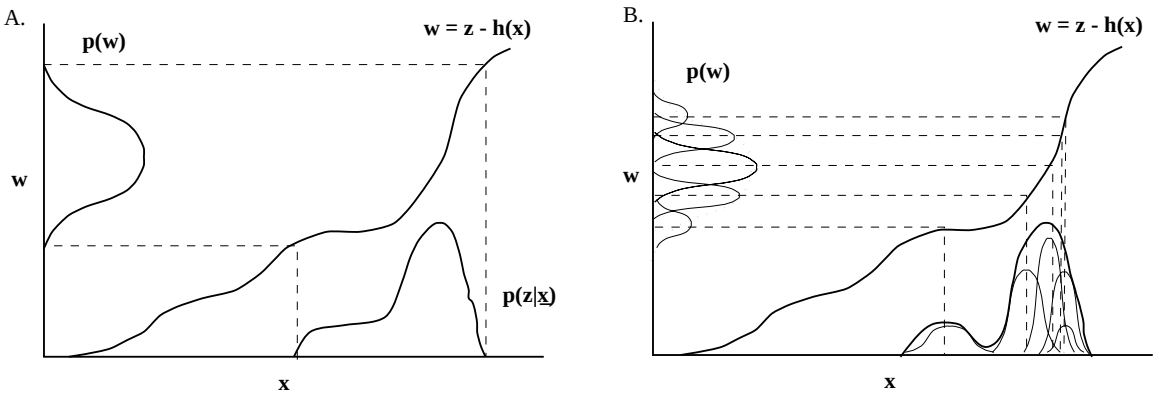

Figure 1.3: Gaussian Mixture Approximation of $p(z(k)|x(k))$

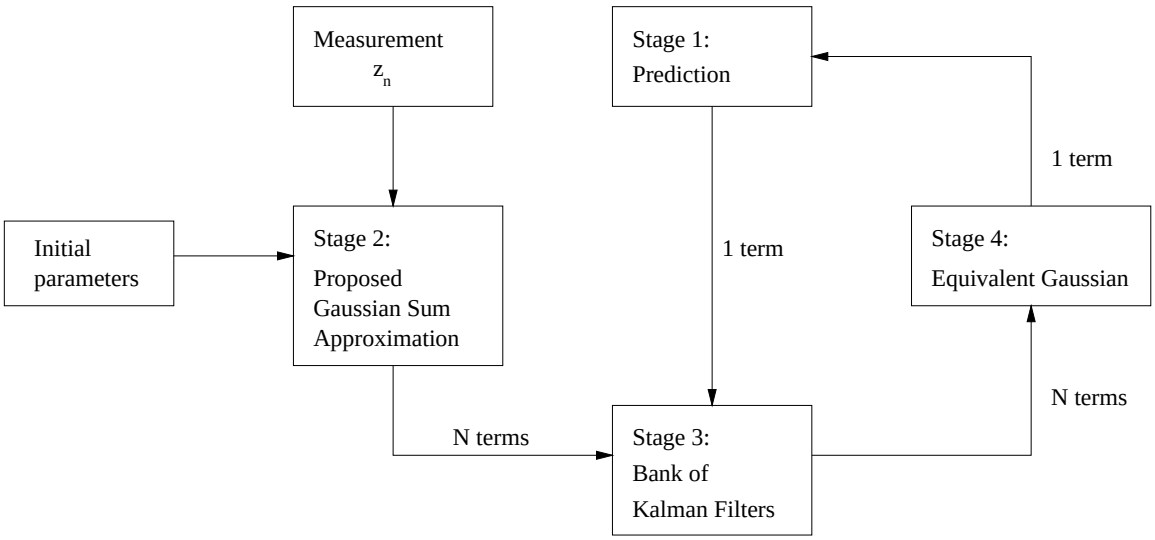

Figure 1.4: The Adaptive Gaussian Sum Filter (AGSF)

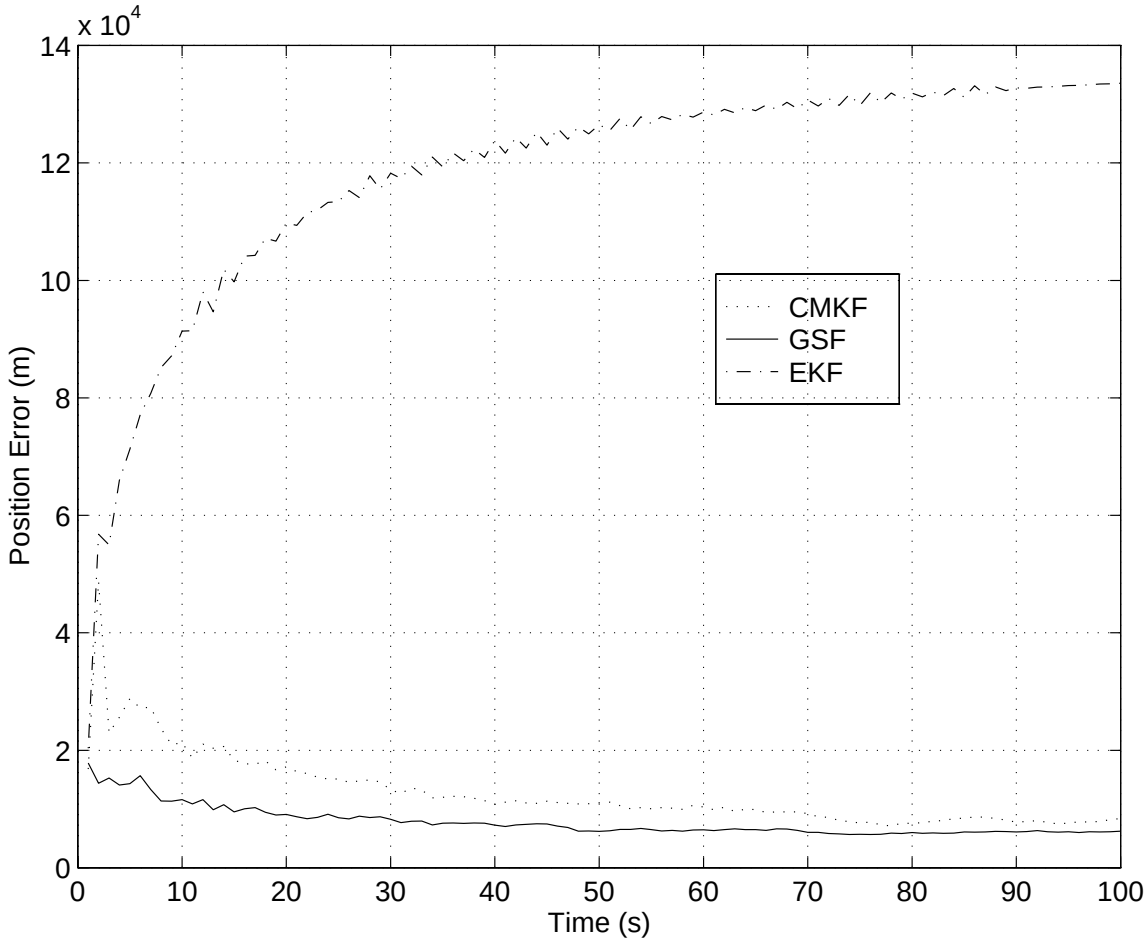

Figure 1.5: Target Tracking: Comparison of Position Errors

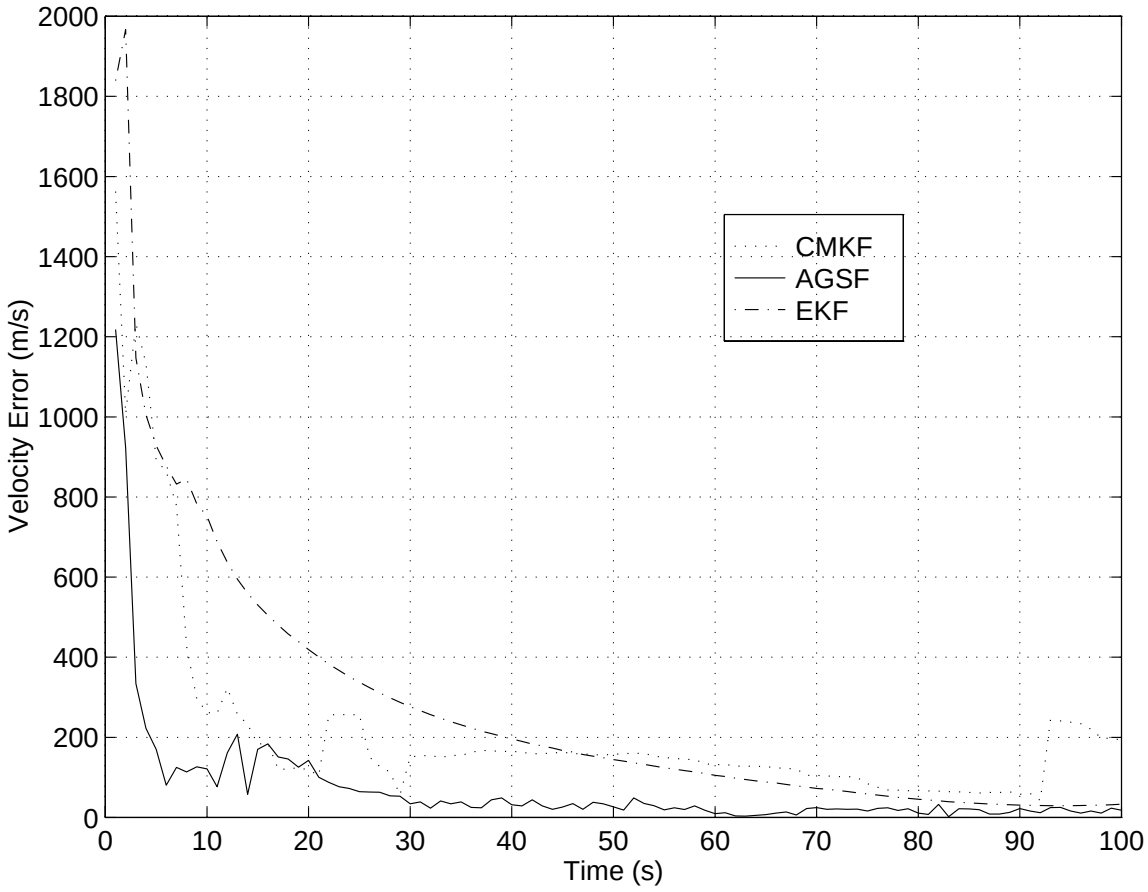

Figure 1.6: Target Tracking: Comparison of Velocity Errors

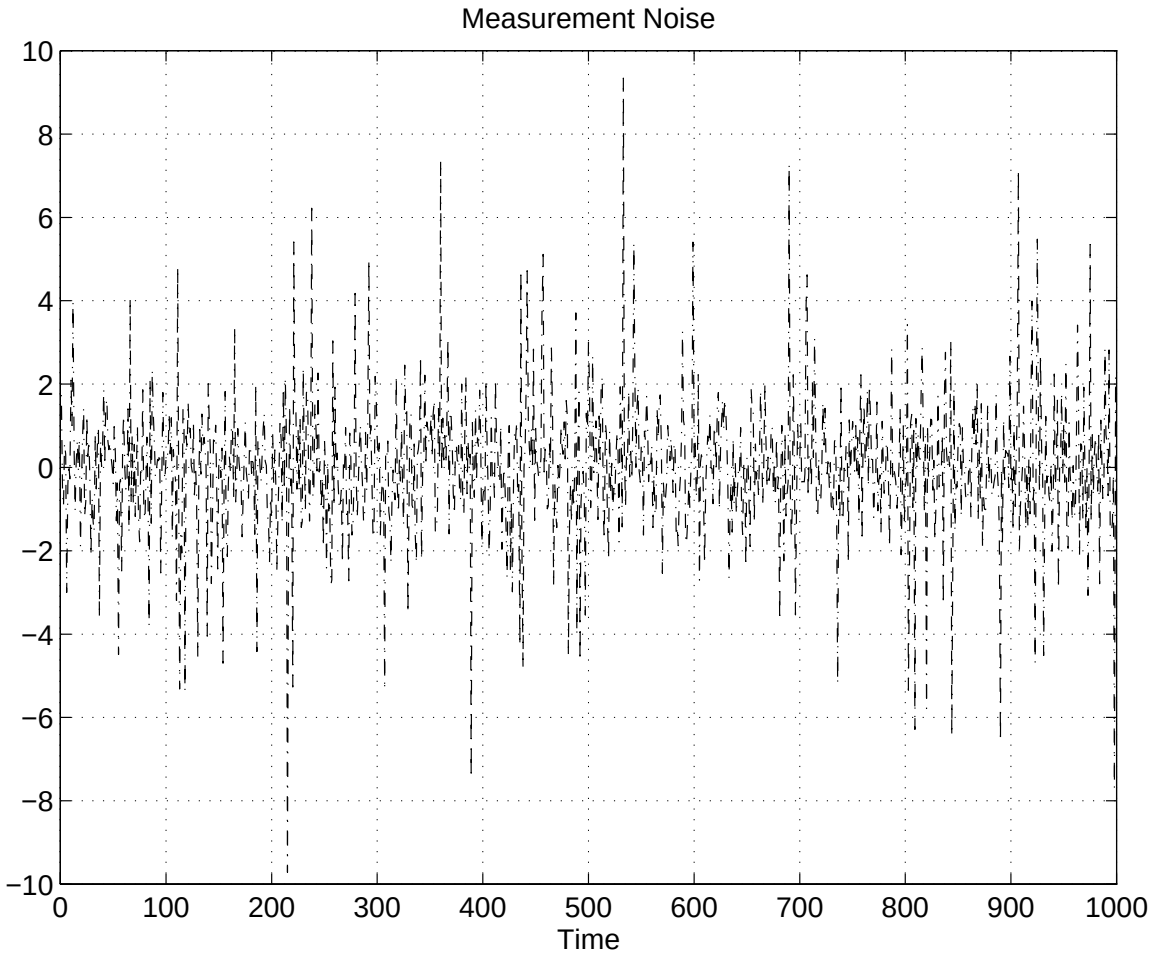

Figure 1.7: Intersymbol Interference-I: Measurement Noise Profile

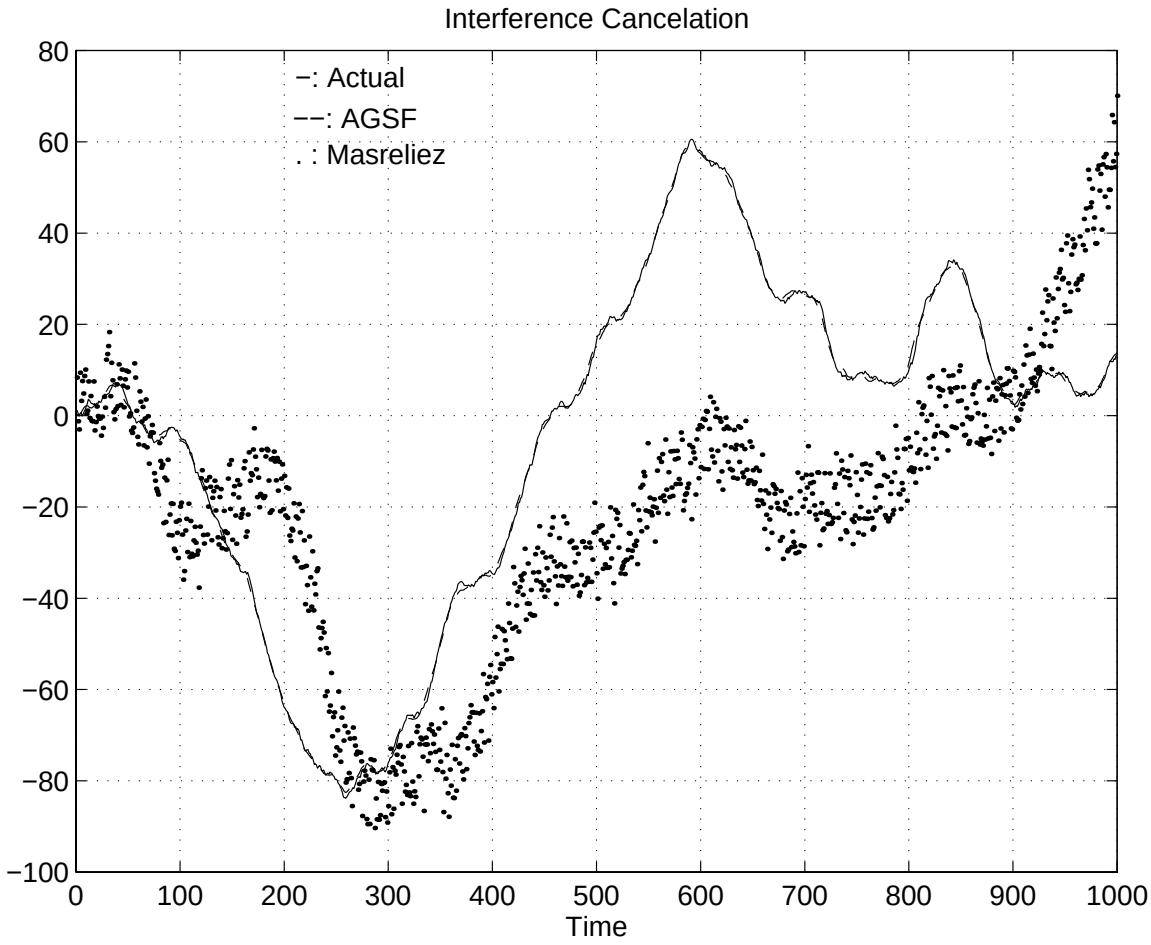

Figure 1.8: Intersymbol Interference-I: Performance Comparison

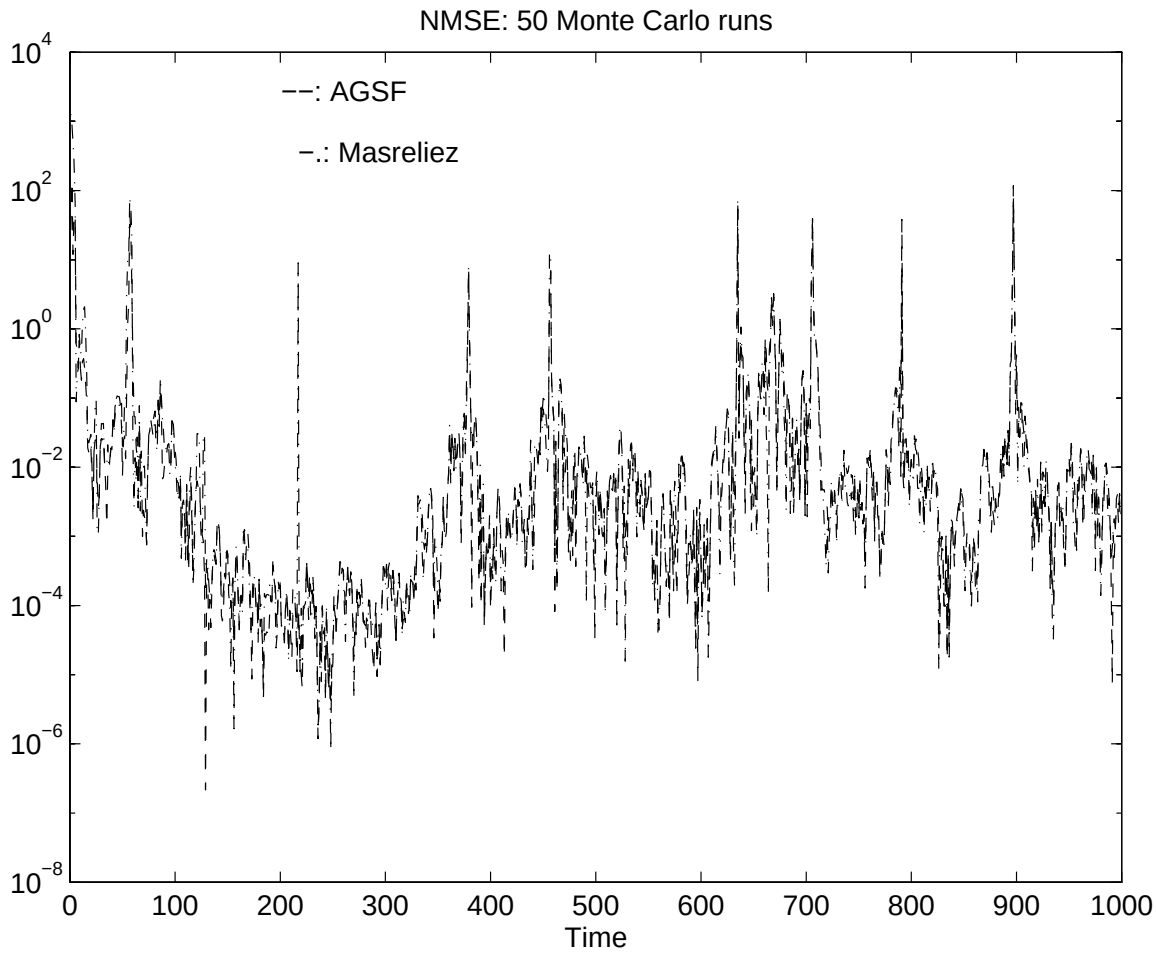

Figure 1.9: Intersymbol Interference-I: Monte Carlo Evaluation

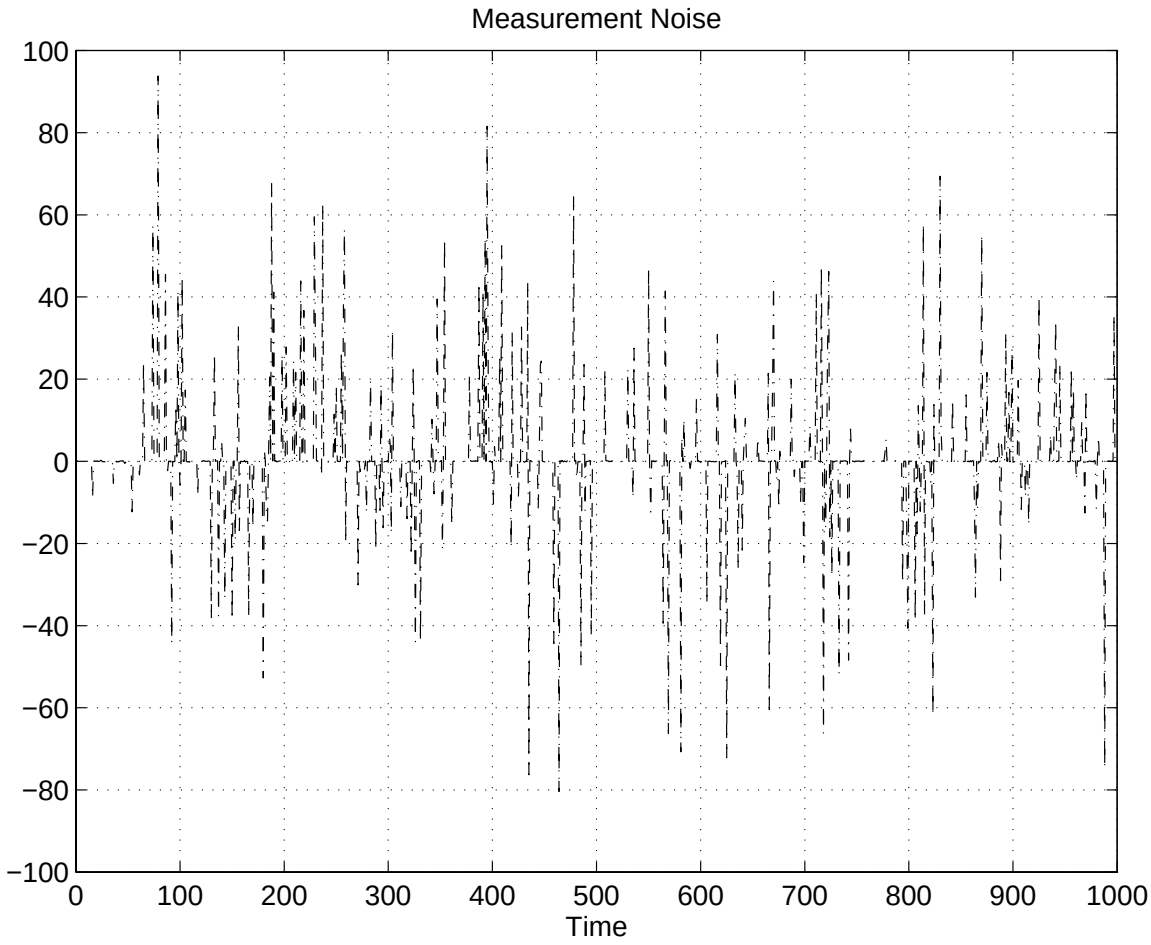

Figure 1.10: Intersymbol Interference-II: Measurement Noise Profile

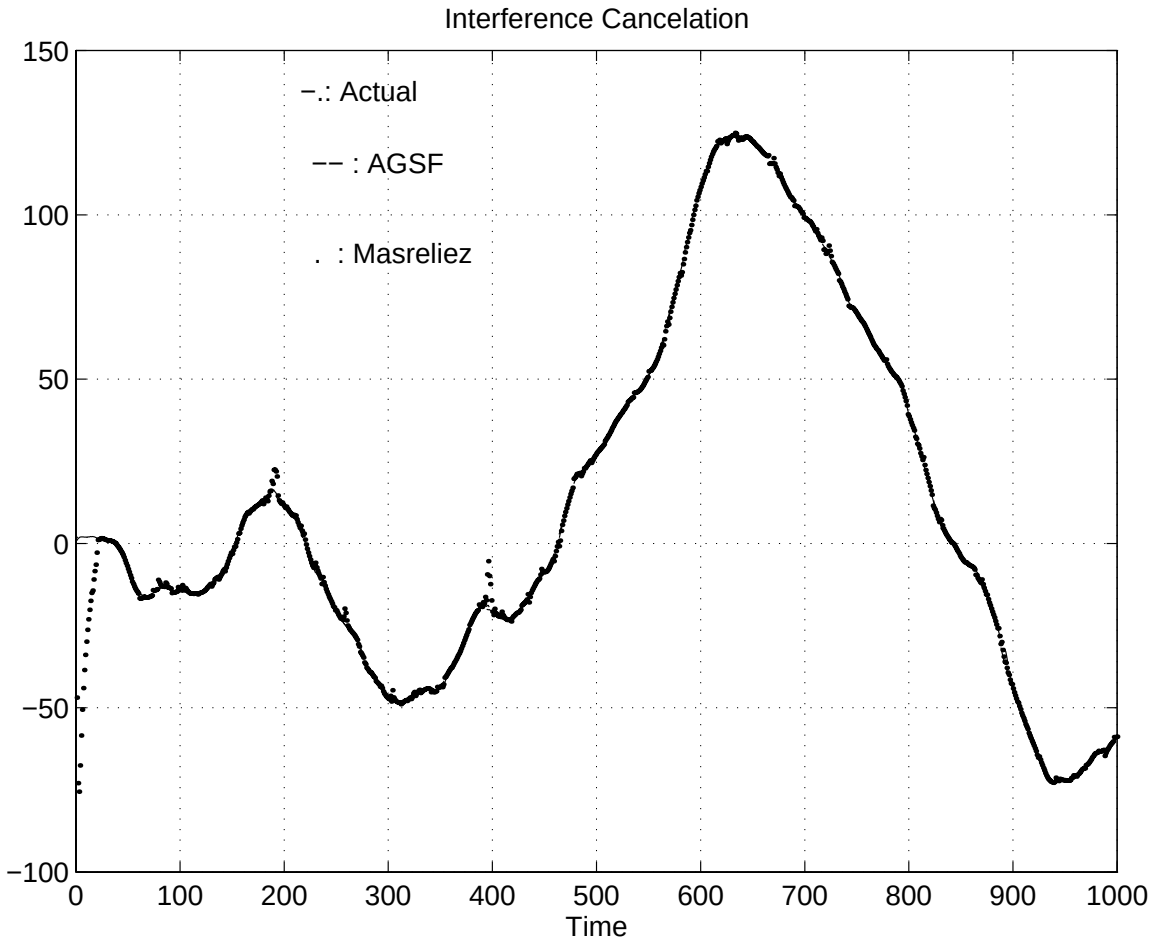

Figure 1.11: Intersymbol Interference-II: Performance Comparison

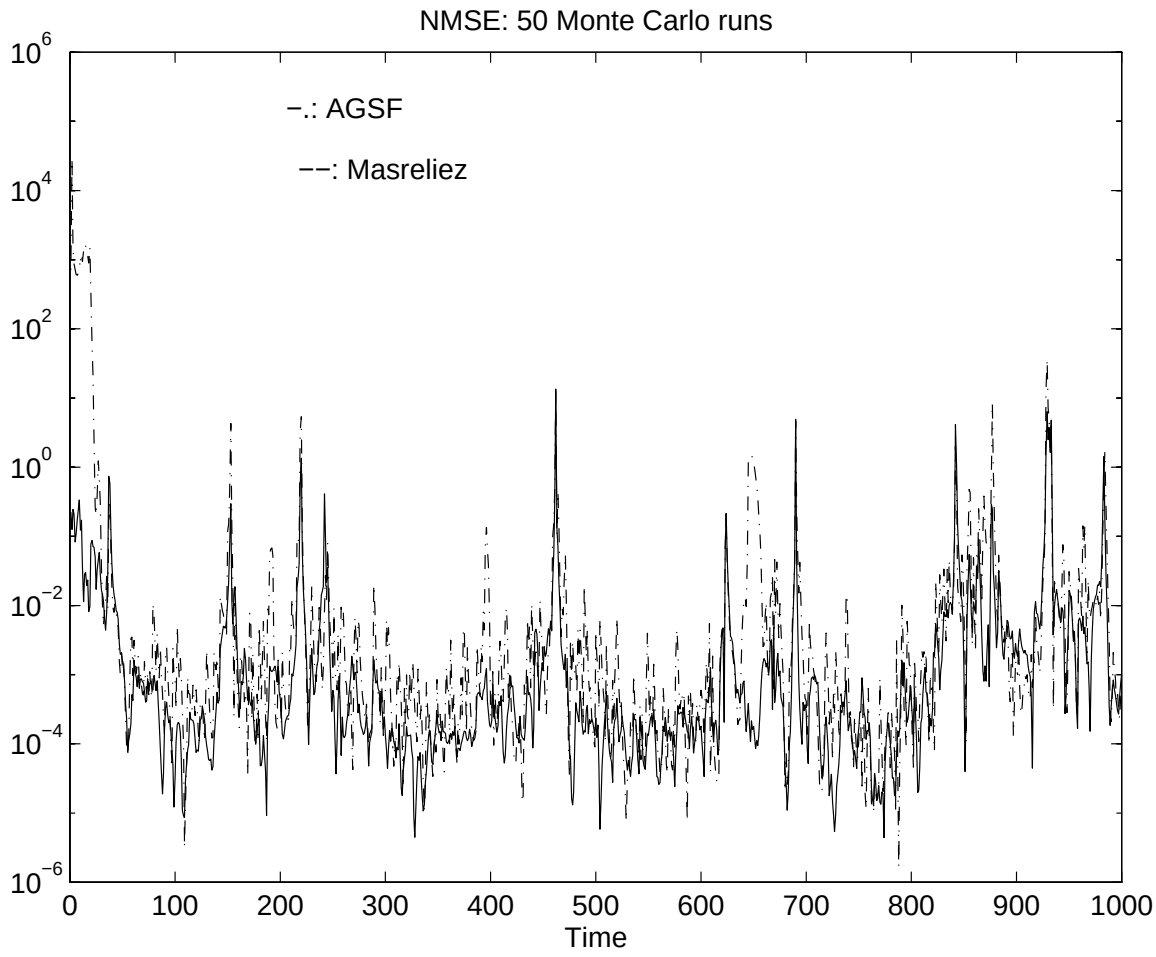

Figure 1.12: Intersymbol Interference-II: Monte Carlo Evaluation

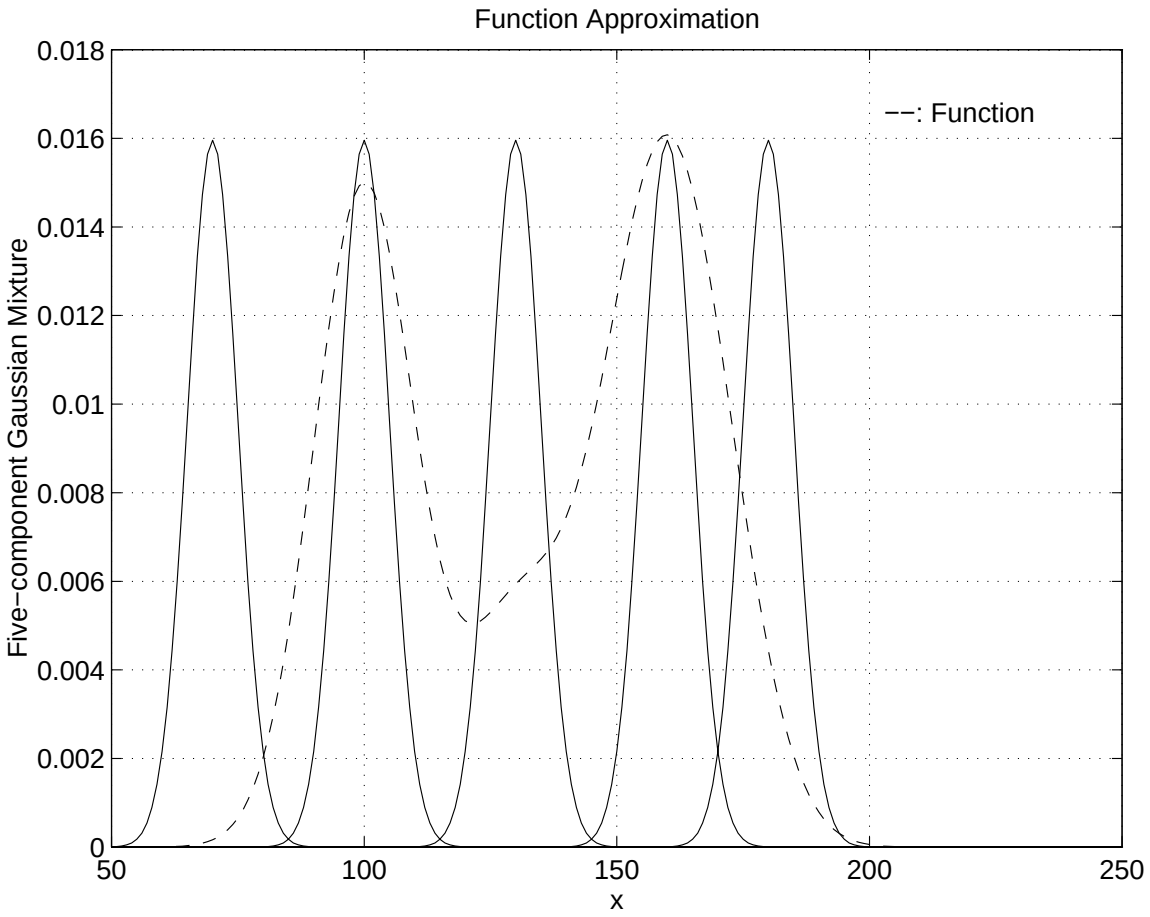

Figure 1.13: Function Approximation via RBF Nets: Initial Placement of the Gaussian Terms

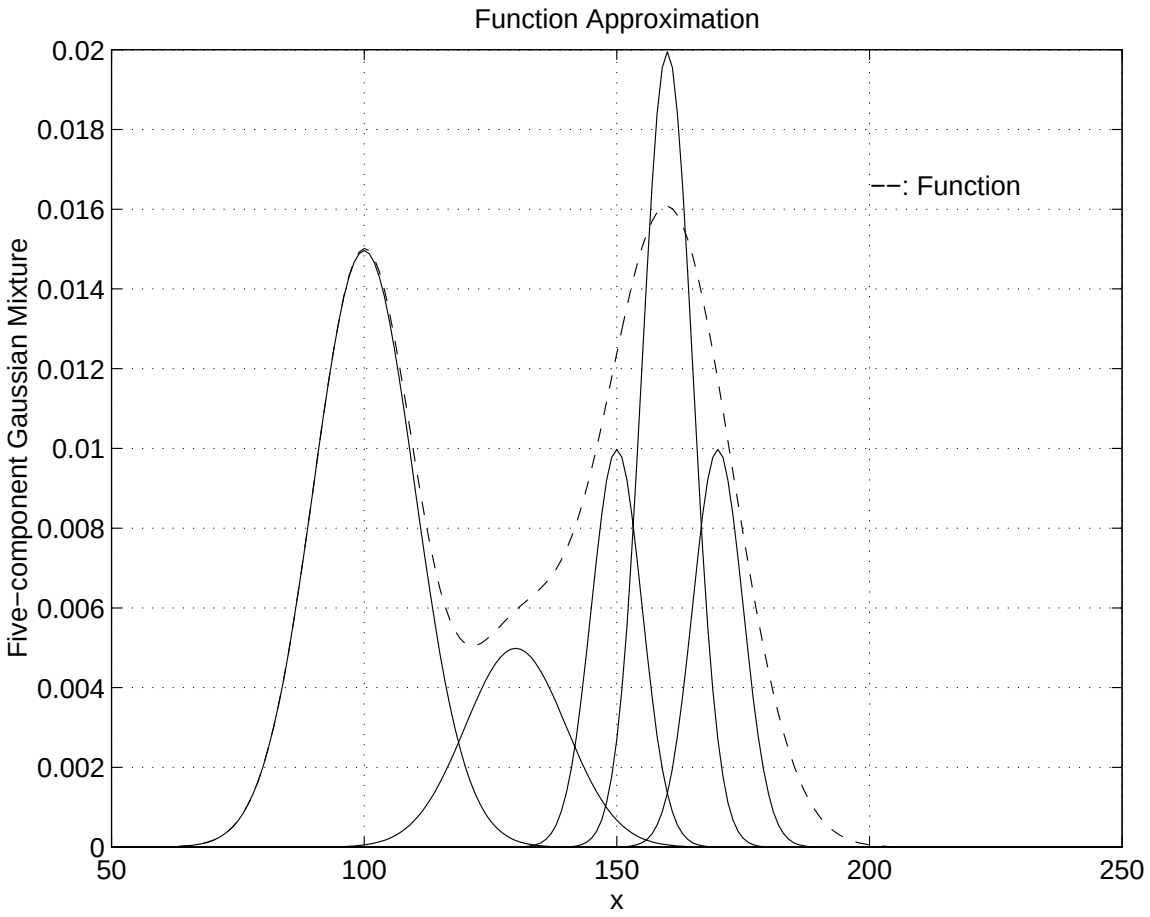

Figure 1.14: Function Approximation via RBF Nets: Final Placement of the Gaussian Terms