NUWC-NPT Technical Report 11,807
1 June 2009

A Tutorial on EM-Based Density Estimation with Histogram Intensity Data

Phillip L. Ainsleigh
Sensors and Sonar Systems Department

Naval Undersea Warfare Center Division
Newport, Rhode Island

PREFACE

This report was partially funded by the Office of Naval Research (ONR-321US).

The technical reviewer for this report was Tod E. Luginbuhl (Code 1522).

Reviewed and Approved: 1 June 2009

David W. Grande
Head, Sensors and Sonar Systems Department

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (*DD-MM-YYYY*)	2. REPORT TYPE	3. DATES COVERED (*From – To*)
01-06-2009		

4. TITLE AND SUBTITLE

A Tutorial on EM-Based Density Estimation with Histogram Intensity Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Phillip L. Ainsleigh

5.d PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Undersea Warfare Center Division
1176 Howell Street
Newport, RI 02841-1708

8. PERFORMING ORGANIZATION REPORT NUMBER

TR 11,807

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
875 North Randolph Street
Arlington, VA 22203

10. SPONSORING/MONITOR'S ACRONYM

ONR

11. SPONSORING/MONITORING REPORT NUMBER

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report examines histogram estimation techniques in which intensity data are represented using a parameterized probability density function (PDF) model. It gives a high-level overview of histogram modeling, introducing the dominant issues and motivations, and then provides detailed mathematical developments of the histogram-based algorithms.

15. SUBJECT TERMS

Signal Processing; Histogram Estimation Methods; Histogram Modeling; Static-Mixture Histograms; Expectation-Maximization (EM) Algorithm

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Phillip L. Ainsleigh
(U)	(U)	(U)	(SAR)	69	19b. TELEPHONE NUMBER (*Include area code*) (401)-832-9201

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

A TUTORIAL ON EM-BASED DENSITY ESTIMATION WITH HISTOGRAM INTENSITY DATA

1. INTRODUCTION

Observed measurements from real-world physical systems are inherently stochastic because of noise in the propagation medium and measurement system, and because of random variabilities in the source-generating mechanism. Therefore, the essential problem in estimation is *not* to identify the "true value" of a variable of interest (such as spatial location or frequency), but to accurately characterize the probability density function (PDF) associated with that variable of interest. In most real-world situations, there are no such things as *numbers*; there are only *distributions*. This report is about the characterization of those distributions.

The notion that data collection is an exercise in distributions is consistent with the way observations of a physical variable are actually obtained. That is, observations from digital processing systems are usually obtained by partitioning the range of the physical variable into *bins*, and then measuring the energy that falls within each bin (or, equivalently, the *energy intensity* associated with each bin). Characteristics of the variable of interest are then inferred from the energy in these bins. The estimation error in this inference process is related to the amount of probability mass (i.e., area under the PDF) that is not encapsulated in the estimator. Thus, if the PDF governing the spread of energy in the variable of interest is concentrated enough that "nearly all" of the probability mass falls within a single bin, then viewing observed data as point measurements (i.e., numbers) is a reasonable approximation. In situations where the PDF has probability mass extending across several bins, however, such point approximations can lead to significant information loss and often to bias.

This report examines histogram estimation methods for representing intensity data using parameterized PDF models, which provide a mechanism to significantly reduce the information loss and bias that result from point approximations. While parametric density estimation has a long history, modern treatments of the problem usually trace back to the seminal paper by Dempster, Laird, and Rubin [1], who, among other things, applied the expectation-maximization (EM) algorithm for parameter estimation with histogram data. McLachlan and Jones [2] extended the algorithm in [1] by applying EM-based histogram estimation to static mixture densities. Luginbuhl [3], Luginbuhl and Willett [4], and Streit [5] took this a step further by applying histogram-

estimation methods for dynamic mixtures within the probabilistic multi-hypothesis tracking (PMHT) algorithm. The focus here is on the static-mixture histograms discussed by McLachlan and Jones [2], with the objective of providing enough detail to allow these models and algorithms to be applied as they stand (e.g., in signal classification applications) or extended for dynamic tracking contexts other than PMHT.

The organization of this report is intended to develop the concepts in increasing detail, ultimately linking all aspects of histogram-based algorithms back to fundamental statistical principles. The next section gives a high-level overview of histogram modeling, introducing the dominant issues and motivations. Sections 3 and 4 then provide detailed mathematical developments of the algorithms. In particular, histogram methods for non-mixture distributions are developed in section 3, and these are extended to mixture distributions in section 4. After a brief summary in section 5, a set of appendixes discuss supporting developments from optimization and distribution theory. In addition to reviewing material that is important to understanding the algorithms, these appendixes introduce much of the notation used in the body of the report. For example, the discussion of the EM algorithm in appendix A provides a notational and logical template that is used repeatedly when discussing histogram estimation algorithms in sections 3 and 4.

2. OVERVIEW OF HISTOGRAM MODELING

The objective is to define statistical models that characterize the behavior of variables of interest (such as bearing or frequency) for some class of sources, a problem that arises in a number of applications. For example, in maximum-likelihood classification (e.g., see [6]), PDF models are use to represent the behavior of the features under various class hypotheses. The classifier decision boundaries are then found at certain intersections of these class-conditional densities. As another example, probabilistic tracking algorithms (e.g., see [7]) estimate parameters in a dynamic PDF model at each time step. Regardless of the application, the goal when developing PDF models is usually to provide a "best fit" for recorded data. However, most sensors do not provide direct measurements of the variables of interest (e.g., recorded data are not tagged with location or frequency information). Information about the desired physical variable is usually derived by transforming the received data (e.g., beamforming of spatial array data or Fourier transform of time-domain data). An inherent characteristic of this transformation process is a partitioning of the variable domain into bins and the generation of output values that correspond to these bins. These transform outputs are usually further transformed to magnitude-squared, or *energy intensity*, data because the original transform outputs are complex quantities whose phase is very sensitive to noise. Therefore, PDF models are derived from data that take the form of energy as a function of transform bins (e.g., energy as a function of beam or frequency).

The traditional approach for processing this type of intensity data extracts "point observations" of the physical variable using a *peak estimator*, which selects one or more values of the variable for which the intensity data are locally maximum. While simple interpolation methods can significantly improve accuracy when signal energy extends over just a few bins, this approach is inappropriate when the spread in the energy extends over several transform bins. Histogram methods accommodate large intensity spreads by employing distribution models with nonzero second (and possibly higher) moment characteristics.

This section introduces the basic ideas and representation abilities of the histogram methods, as well as some of the issues that are addressed in the later sections. The first subsection introduces the histogram approximation. The second discusses issues that arise in the application of histogram methods to acoustic intensity data. The third motivates the use of mixture densities in a histogram context, and the last subsection discusses issues that arise when the range of available measurements does not fully cover the range of the PDF.

2.1 HISTOGRAM APPROXIMATION

The histogram algorithms discussed here fall in the general class of parametric statistical modeling techniques, where a PDF model $p(\mathbf{z};\boldsymbol{\Theta})$ is used to represent the energy distribution in the physical variable $\mathbf{z}$ of interest. The characteristics of the model are governed by the parametric structure of the model and by the values in the parameter set $\boldsymbol{\Theta}$, which must be estimated from observed data. The difficulty with this situation is that there are no direct observations of $\mathbf{z}$. The parameter vector $\boldsymbol{\Theta}$ must be estimated from the energy intensity data, denoted $\mathbf{S} = \{s_\ell : \ell = 1, \ldots, L\}$, where s_ℓ represents the energy in the ℓth bin and the collection of bins covers some range of interest in the values of $\mathbf{z}$.

A natural approach for overcoming this difficulty is to transform the PDF model $p(\mathbf{z};\boldsymbol{\Theta})$ into another valid PDF model $p(\mathbf{S};\boldsymbol{\Theta})$, allowing $\boldsymbol{\Theta}$ to be estimated from the intensity data. When a convenient mathematical form for $p(\mathbf{S};\boldsymbol{\Theta})$ is available, then various optimization methods can be applied directly (e.g., Newton or other derivative-based ascent method). Typically, however, the task of expressing $\mathbf{S}$ directly in terms of $\boldsymbol{\Theta}$ is highly nontrivial, and the resulting expression, if it can be obtained, is highly nonlinear. For this reason, histogram methods employ an approximate model in which $p(\mathbf{z};\boldsymbol{\Theta})$ is used in conjunction with a multinomial approximation to represent the within-bin and across-bin characteristics of the intensity data. Estimation algorithms are then developed using an iterative EM approach. A brief overview of the EM approach is provided in appendix A, which establishes a template for the algorithm descriptions in sections 3 and 4.

The use of the EM algorithm and multinomial approximation invokes a quantum image of intensity data, analogous to photons of light energy. The variable $\mathbf{z}$ of interest is considered to be a *feature* of the quantum particles, and the particles are assumed to be sorted into bins according to the value of $\mathbf{z}$ associated with each particle. The number of particles in each bin, denoted m_ℓ for the ℓth bin, is referred to synonymously as a *histogram intensity*, *bin intensity*, *histogram count*, *particle count*, or *bin count*. The symbol m_ℓ is used to emphasize the discrete nature of the histogram intensities, in contrast to real-valued energy intensities. The relationship of PDF model, point measurements, and histogram intensities is illustrated in figure 1.

For sensing modalities in which particle counts are actually observed (e.g., ionizing radiation), or when there exists a one-to-one physical correspondence between energy intensity $\mathbf{s}_\ell$ and particle count m_ℓ (e.g., electromagnetic energy), the histogram approach is generally valid. The only question in these cases concerns the appropriateness of the

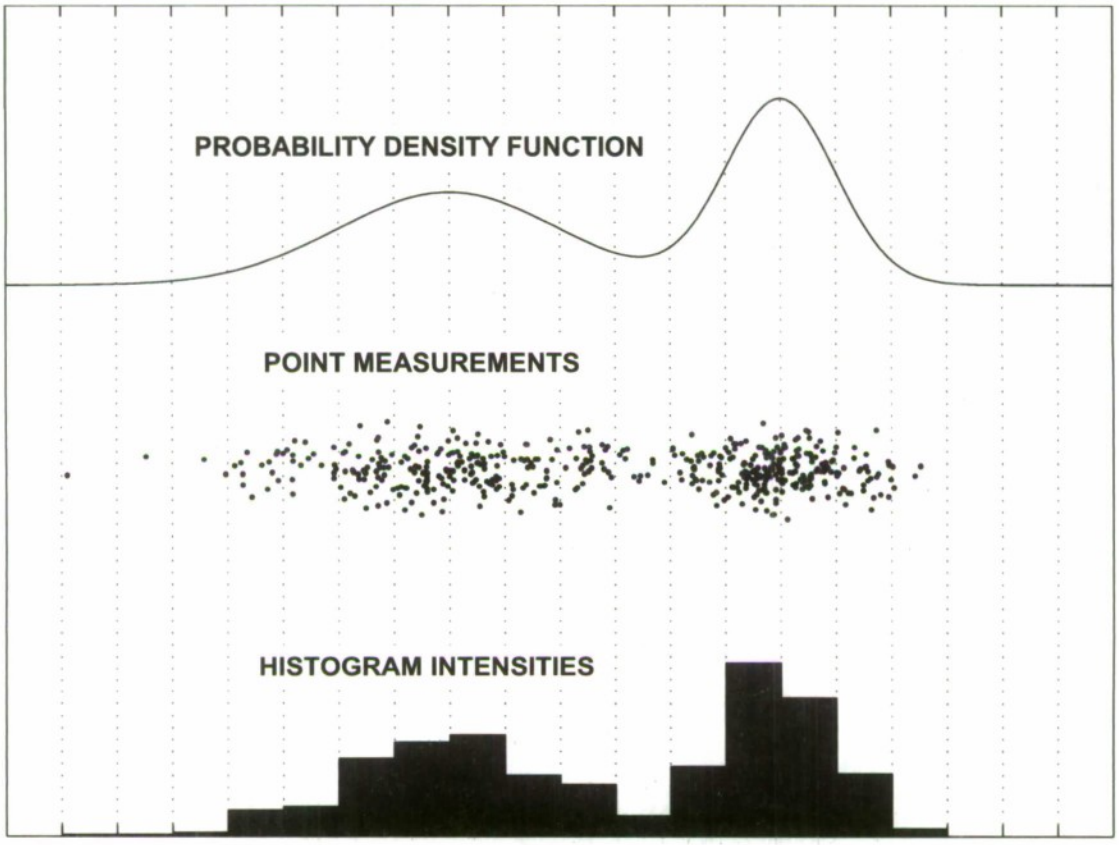

Figure 1: PDF, Measurements, and Histogram Intensities
The PDF (top) corresponds to a hypothetical scalar distribution. The points measurements (middle) are synthesized by sampling from the PDF, and these measurements are assigned to the various histogram bins with boundaries given by the vertical grid lines. The numbers of point measurements in each bin correspond to the histogram bin intensities (bottom). In practice, the point measurements are not observed (indeed, they may not even exist), such that histogram methods must estimate the PDF parameters from the intensity data. EM-based histogram methods postulate the existence of the point measurements in order to obtain an easier optimization problem. These point measurements are treated as missing data, however, and are marginalized from the objective function during each EM iteration. Note that the point measurements shown in this figure are one-dimensional in the horizontal axis; the vertical spread in the measurements is included merely to illustrate the clustering of measurements.

parametric form for $p(\mathbf{z};\boldsymbol{\Theta})$. When using the histogram model for acoustic intensity data, however, there is no physical mapping from $\mathbf{s}_\ell$ to m_ℓ. The histogram model is therefore inherently mismatched to the data in this case.

2.2 HISTOGRAM APPROXIMATION AND ACOUSTIC DATA

When applied to acoustic intensity data, the histogram model appears to exhibit some insurmountable flaws, namely, the presumption of particles and point measurements that don't really exist, as well as the real-valued nature of the energy intensities in a theory requiring integer count data. From the perspective of practical implementation, there are two factors that ameliorate these problems. First, the actual estimators for the model parameters in $\boldsymbol{\Theta}$ are obtained by integrating over the space of the point measurements, whereby the point measurements are marginalized out of the likelihood function. This marginalization takes place during the derivation of the estimators, such that the point measurement $\mathbf{z}$ never appears in any computational algorithm. Second, the parameter estimators end up being functions only of the relative intensities, which are the individual bin intensities divided by the overall intensity. The numerical algorithms do not care if these ratios are formed from integer-valued count data or real-valued energy data, since the ratio is, in general, real in both cases.

Now, just because the algorithm can be applied does not make it the right tool. It is therefore of interest to take a closer look at a common special case, specifically magnitude-squared discrete Fourier transform (DFT) data. The DFT forms the inner product of recorded time-domain data with a set of complex sinusoidal basis functions. Since, in practice, it is impossible to observe an infinite duration of the time-domain signal, it is impossible to localize the energy to a single frequency point. The DFT sample therefore represents the energy "in the vicinity of" a given frequency, such that the partitioning of the frequency domain into bins is a natural result of the transform process and the DFT samples correspond to the energy that is projected into each bin. This projection of energy into DFT bins has the flavor of a histogram by definition, although the histogram model takes this a step farther by assuming a quantization of the DFT bin energies to take integer values, as if there were some basic unit of acoustic energy (i.e., the acoustic equivalent of Planck's constant) and the quantized integers represent the number of these units that fall within each DFT bin. This quantization implies a set of "synthetic" particles for each bin, and the number of these synthetic particles is indeed a histogram count, even if it does not correspond to any known physical entity. The multinomial model at the heart of histogram methods then characterizes these synthetic count data.

The multinomial model starts out by treating histogram count data as statistically independent Poisson processes. The histogram count for each bin is then modeled using a conditional distribution, where the conditioning is on the total synthetic intensity in all bins (the quantized version of the total signal energy). Now, the total synthetic intensity is merely the sum of the individual intensities, and a sum of Poisson processes is itself a Poisson process whose expected value is the sum of the individual expected values [8]. Furthermore, since the conditioning process is equivalent to dividing probability mass functions, the multinomial model is effectively a series of ratios of Poisson mass functions. Note that, while the synthetic intensities for the various DFT bins are initially represented as independent processes, the bin intensities are *not* independent in the multinomial model because of the conditioning on the total intensity. That is, all bins affect all other bins through their contribution to the total intensity. Note also that modeling the quantized intensities with a multinomial distribution provides a statistical description that matches the first moment of the observed data but does not match any higher moments.

Given the subtleties of the multinomial approximation, its appropriateness must be evaluated on a case-by-case basis for different applications. That said, the histogram approach provides a way forward in cases where the alternatives are intractable. For example, an attempt to directly model the *energy* intensities would require a joint multivariate exponential distribution over all of the spectral or spatial bins (i.e., a joint distribution over easily thousands of variables, and more as resolution increases). Estimation in such high-dimensional spaces is impossible in most real-world scenarios.

In many applications, the quantization of the bin intensities is implicit because the algorithms depend only on the relative intensities. However, the quantization manifests itself in a very explicit way in Bayesian contexts where a prior distribution is imposed. The problem involves the relative weighting of the prior distribution and the measurement likelihoods when estimating the posterior parameter estimates. Specifically, the measurement likelihoods are weighted by the overall intensity of the bins (i.e., the total number of particles in all bins). The more "measurements" there are, the more the prior distribution is discounted. But when dealing with artificially quantized intensity data, the overall intensity depends on the unit energy associated with each particle, which is itself a function of the quantization. The net result is that the quantization unit becomes an explicit variable in the algorithm, making the relative weighting of the prior completely dependent on an algorithm design parameter. If a coarse quantization (i.e., a large particle energy unit) is assumed, then the synthetic particle counts will be

low and the prior distribution will dominate. If a very fine quantization (i.e., a small particle energy unit) is used, then the apparent number of particles is very high and the data overwhelm the prior. Indeed, in the limit as the quantization unit approaches zero, the synthetic particle counts become infinite and the estimator effectively ignores the prior distribution altogether, a problem that was observed by Streit in his work of histogram PMHT [5]. This same issue will arise with *any* dynamic mixture scenario in which the overall intensity varies with time, which includes just about all practical tracking applications.

As a final note, the dependence of the parameter estimators only on the relative intensities also means that the estimators are invariant to the true overall intensity. The degree of freedom introduced by this invariance allows the same model to represent large variabilities in different realizations of the distribution. To illustrate this, figures 2 and 3 show examples of synthetic histogram data with different values of signal-to-noise-ratio (SNR) and overall intensity. In figure 2, plots are shown with the same value of SNR, but with different values of the overall intensities. In contrast, figure 3 contains plots with the same overall intensity but with varying SNR. While there is significant variability in both cases, there is a subtle distinction. Low SNR generally means that there is *too much* of the *wrong kind* of data (i.e., noise), whereas low overall intensity indicates that there is *too little* of *every kind* of data. The histogram model accommodates both cases equally well.

2.3 HISTOGRAM METHODS AND MIXTURE MODELS

The Gaussian-signal-in-uniform-noise mixture model used to generate the above examples was chosen because of the basic role that such models play when using histogram techniques. Specifically, mixture models provide a convenient way to explicitly model noise, which is necessary because histogram models are very *data inclusive*. This is most easily seen by contrasting the histogram approach to an estimator that selects the single maximal peak. When such a peak estimator is applied to data with highly concentrated signal energy (e.g., narrowband signal spectra), a beneficial side effect is provided in the form of signal cleaning. That is, in cases where the *signal* intensity is largely confined to a single bin and that bin is correctly identified, most of the wideband noise is eliminated with the omitted bins. The histogram estimator, on the other hand, throws away nothing. It must therefore explicitly account for the noise to minimize noise-induced errors.

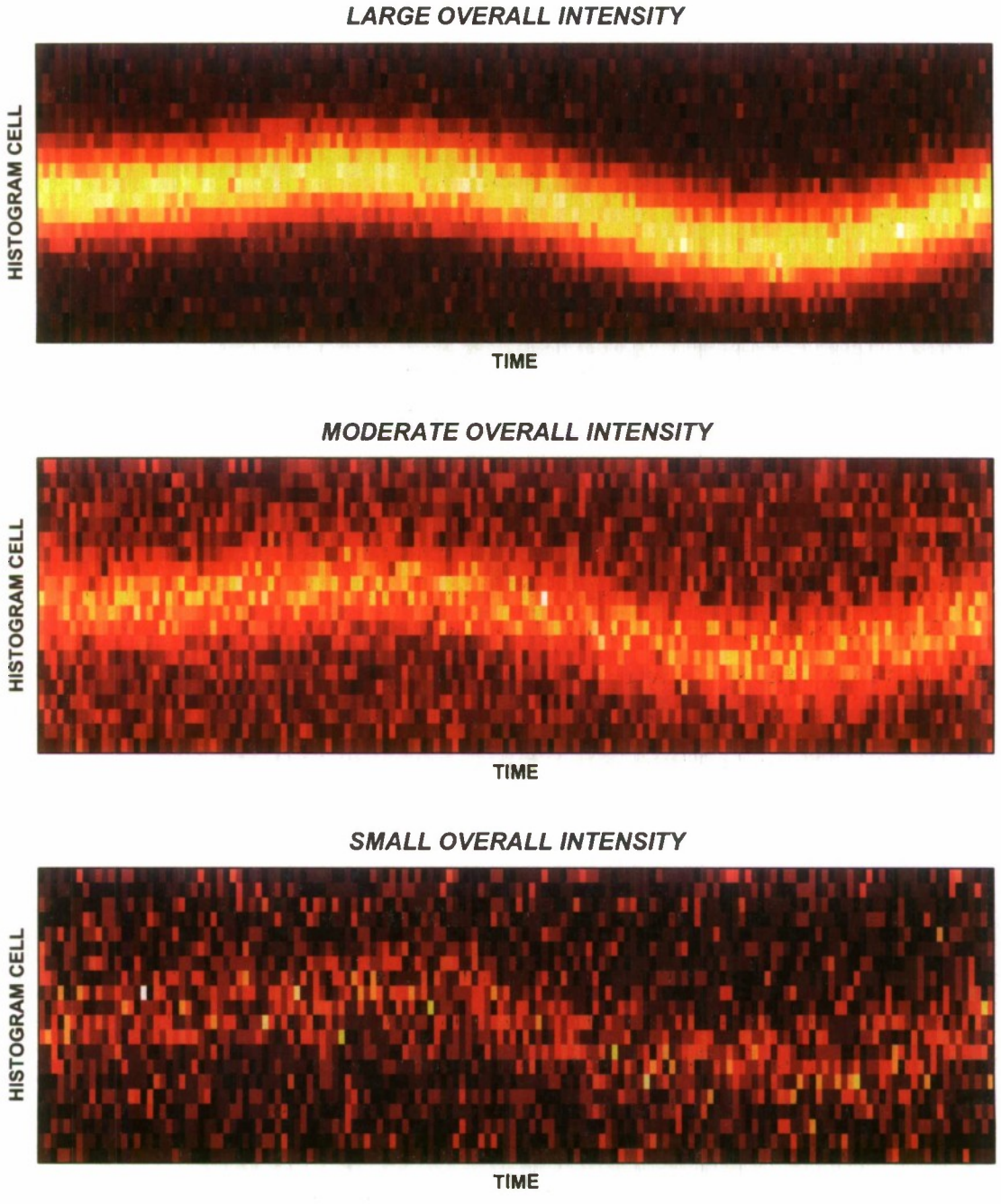

Figure 2: Histogram Data as a Function of Overall Intensity
Image plots of histogram intensity data are synthesized independently at each of 150 time points, with intensities at each time point obtained by sampling from a two-component mixture model containing a Gaussian signal in uniform noise. Samples are sorted into unit-width bins covering the range $[-10, 10]$. *The Gaussian component has constant variance* $\sigma^2 = 4$ *and a sinusoidally time-varying mean. The mixture components have constant probabilities* $\pi_s = 0.3$ *(for signal) and* $\pi_n = 0.7$ *(for noise). The ratio* π_s/π_n *is the (wide-band) SNR. Intensity gram plots are shown for overall intensities corresponding to 2500 point measurements per sample time (top), 250 measurements per sample time (middle), and 25 measurements per time (bottom).*

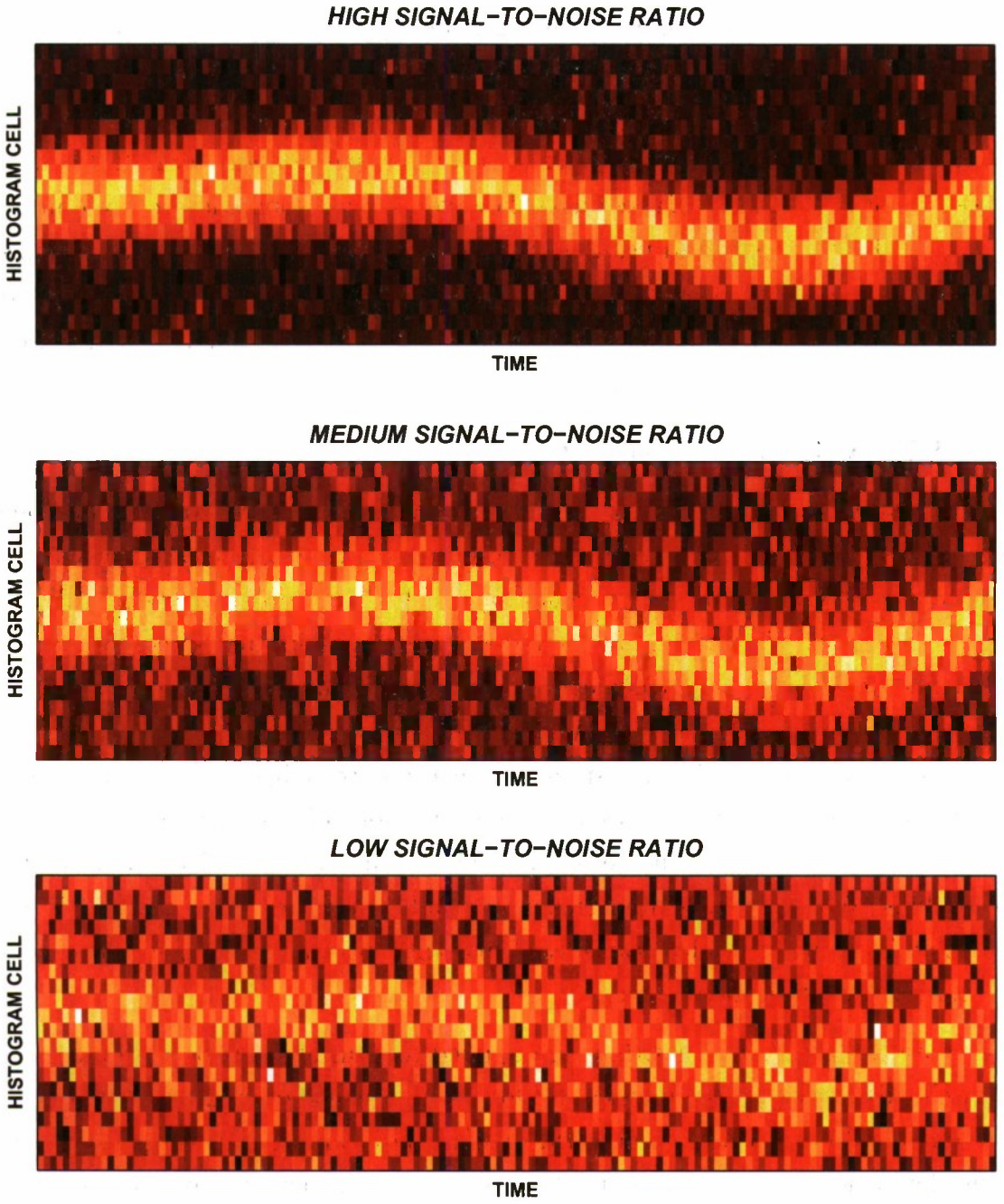

Figure 3: Histogram Data as a Function of SNR
Histogram data are synthesized using the same model as in figure 2. Here, the total number of measurements in all plots is $K = 250$ *(as in the middle plot figure 2). The signal-mode assignment probability varies, with values* $\pi_s = 0.5$ *(top),* $\pi_s = 0.3$ *(middle), to* $\pi_s = 0.1$ *(bottom), giving high, medium, and low signal-to-noise ratios, respectively.*

Finite mixture distributions, the topic of section 4, are ideally suited for noise modeling because they provide a natural mechanism for dividing the energy between a number of component distributions, one or more of which can be tailored to noise. The "basic" PDF model for histogram methods is therefore the Gaussian-signal-in-uniform-noise model. The use of a uniform noise distribution assumes that the data has been pre-whitened, for example, using a spectral or spatial normalizer. If it is more desirable to work with un-normalized data, however, a more sophisticated noise model is required and can be achieved using a mixture of uniform or Gaussian components. Multiple mixture modes can also be used to describe the signal energy itself, say for data containing energy from multiple sources of interest or signals whose energy in the variable of interest is inherently multi-modal.

2.4 TRUNCATED HISTOGRAM DATA

For a variety of reasons, intensity measurements may not be available for some histogram bins. This can happen unintentionally, for example, when processing spectral data from sensors with a limited frequency range. It can also occur intentionally, as when data are truncated or subsampled to reduce communication and/or computational requirements, or to isolate a phenomenon of interest. Figure 4 shows a hypothetical example of truncated histogram data. While this illustration focuses on histogram "edge effects" where the missing bin intensities are on the outer edges of the distribution, missing intensity values can also occur in the "interior" of the histogram, say, due to unintentional drop-outs in the sensor response or intentional subsampling of the histogram bins. When the histogram bins do not fully cover the range of the PDF, the histogram and its data are called *truncated*. This is in contrast to a *complete* histogram, whose bins do cover the entire range.

If complete-histogram algorithms are applied to truncated-histogram data, then a mismatch exists between the physical world (where a nonzero intensity is impossible in some bins) and the mathematical world (where nonzero intensities are possible, but none just happened to be observed in the data at hand). The size of the mismatch depends on the probability mass under the density function in the truncated regions. This probability mass may be very small, allowing the issue to be ignored in some applications. However, when modeling physical phenomena whose energy can approach the edges of the observable measurement space, or when sensor malfunction causes lost sensitivity in some interior region, then the probability mass in the truncated region can be significant. The EM algorithm is used to circumvent this problem by treating the unobserved intensities as missing data.

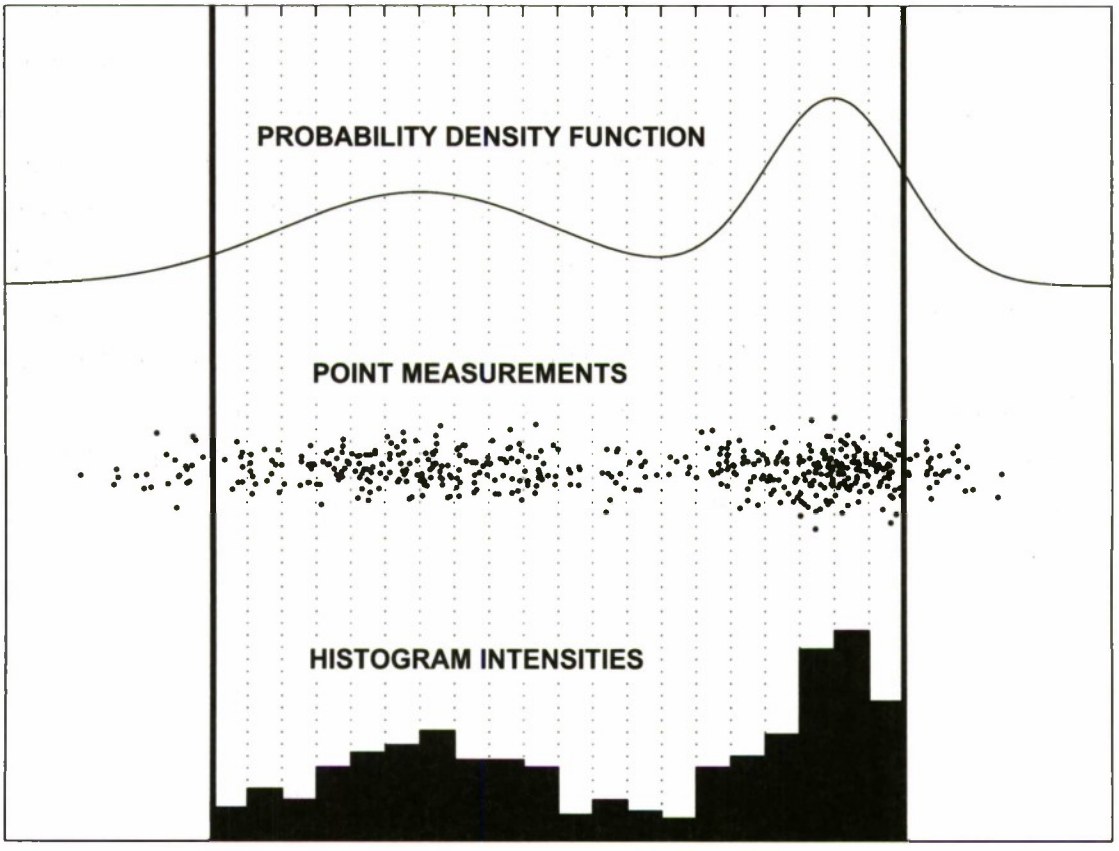

Figure 4: Truncated Histogram Data.
This figure shows data from the same scalar distribution as in figure 1, with the continuous PDF shown on top, the synthetic point measurements shown in the middle, and the histogram bin intensities shown on bottom. In this case, however, the histogram bins do not cover the entire region over which point measurements can occur. The energy associated with measurements falling outside of the observable region (i.e., outside of the bold vertical lines) is not included in any histogram bin and is therefore lost in the histogram model. The lost data are accounted for in the EM estimation algorithm by utilizing an augmented set of missing data. That is, the EM algorithm treats as missing data the intensity values in the truncated region, in addition to the point measurements that form the missing data in the case of a complete histogram.

The computational algorithm for the truncated histogram ends up being a simple modification of the algorithm for the corresponding complete histogram. The theoretical work needed to develop the algorithm, however, has a subtlety requiring careful analysis. In particular, it is impossible to know the total intensity (i.e., the sum of the bin intensities) when some of the bin intensities were not recorded, and the multinomial distribution is well defined only when the overall intensity is given. The uncertainty regarding the overall intensity requires the use of a negative binomial distribution and its multivariate extension, the negative multinomial distribution. This extension is discussed in section 3 and appendix C. Fortunately, the end result of that analysis is a simple extrapolation formula for obtaining the expected values of the missing intensities. The EM auxiliary function for the truncated histogram is then a linear function of the missing intensities, such that the maximization step in each iteration of the EM algorithm can be formulated in terms of the complete histogram. The algorithm then operates on an extended data set in which the observed intensities are augmented with *expected* intensities in the truncated regions.

3. ESTIMATION FROM HISTOGRAM DATA

The remainder of this report focuses on the mathematical developments related to histogram modeling and estimation. In all that follows, integer count data are assumed available for each bin. That is, the issues cited in the previous section related to converting from real-valued energy data to integer-valued histogram data are assumed to have been dealt with appropriately.

As mentioned in the previous section, one key distinction among histogram methods involves whether or not the histogram bins completely cover the range of possible measurement values *and* intensity data are available for all bins. When estimating the parameter set $\boldsymbol{\Theta}$, the values of $\mathbf{z}$ for which $p(\mathbf{z};\boldsymbol{\Theta})$ is well defined constitutes the *measurement space*, denoted $\mathcal{Z}$. If data are available for histogram bins that completely cover $\mathcal{Z}$, then the histogram and its data are *complete*. If there are regions of $\mathcal{Z}$ for which histogram data are lacking, then the histogram is *truncated*. This section first examines complete histograms and then extends those algorithms for truncated data. A general relationship between the two types of histograms is then discussed.

The notation used in this section and the logical steps in deriving algorithms closely parallel the discussion of the expectation-maximization (EM) method in appendix A. Note that EM theory defines *complete data* generically to indicate the concatenation of the observed and missing data, which is independent of whether the histogram bins fully cover the space of the physical variable. To distinguish these different types of complete data, data from a complete histogram will always be referred to as *complete-histogram data*. Any reference to "complete data" without the "histogram" modifier indicates the more generic notion from EM theory.

3.1 ESTIMATION FROM COMPLETE HISTOGRAM DATA

A complete histogram partitions the measurement space into a set of mutually exclusive and exhaustive regions, which are the histogram bins (or cells), denoted $\mathcal{Z}_\ell$ for $\ell = 1, \ldots, L$. The measurement space is thus decomposed as

$$\mathcal{Z} = \bigcup_{\ell=1}^{L} \mathcal{Z}_\ell \,, \tag{1}$$

where the non-overlapping nature of the $\mathcal{Z}_\ell$ is implicit. Consider now a collection of hypothetical point measurements $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$. The number of these measurements whose value falls in each bin comprise the histogram intensity data, denoted

$$\mathbf{M}_\mathrm{C} = \Big\{ m_\ell : \ell = 1, \ldots, L \Big\} , \tag{2}$$

where the subscript C indicates complete histogram data. The *overall intensity* is then defined as

$$K = \sum_{\ell=1}^{L} m_\ell \,. \tag{3}$$

Under a given set of parameter values, the unconditional probability that a measurement falls in the ℓth bin is

$$\Phi_\ell(\boldsymbol{\Theta}) = \int_{\mathcal{Z}_\ell} d\mathbf{z}\; p(\mathbf{z}\,;\boldsymbol{\Theta})\,, \tag{4}$$

which, for a complete histogram, satisfies

$$\sum_{\ell=1}^{L} \Phi_\ell(\boldsymbol{\Theta}) = 1\,. \tag{5}$$

Assuming that the measurements constitute K independent draws from L categories, the bin intensities are governed by the *multinomial distribution* discussed in appendix B; that is,

$$p(\mathbf{M}_\mathrm{C}\,;\boldsymbol{\Theta}) = c(\mathbf{M}_\mathrm{C}) \prod_{\ell=1}^{L} \Big\{\Phi_\ell(\boldsymbol{\Theta})\Big\}^{m_\ell}\,, \tag{6}$$

where $c(\mathbf{M}_\mathrm{C})$ is the multinomial coefficient defined in equation (127). From the standpoint of the EM-based algorithms developed in this section, equation (6) is the *observed data likelihood function* (ODLF) since it characterizes the observed histogram counts.

3.1.1 EM Algorithm with Missing Measurements

Definition of Missing Data and Auxiliary Function. The observed histogram intensities provide a relative measure of the number of unobserved point measurements that "fall into" each bin. The missing data for the EM algorithm are these unobserved measurements. Within a given bin (say, the ℓth), the measurements are denoted

$$\mathbf{Z}_\ell = \Big\{\, \mathbf{z}_{\ell k} \,:\, k = 1, \ldots, m_\ell \,\Big\}\,, \tag{7}$$

and the full set of missing data contains measurement sets for all bins as

$$\mathbf{Z} = \Big\{\, \mathbf{Z}_\ell \,:\, \ell = 1, \ldots, L \,\Big\}\,. \tag{8}$$

The complete data are the concatenated set containing the observed bin intensities and missing measurements, and its joint distribution $p(\mathbf{Z}, \mathbf{M}_\mathrm{C}\,;\boldsymbol{\Theta})$ is the *complete data*

likelihood function (CDLF). Given the CDLF and the posterior density $p(\mathbf{Z}|\,\mathbf{M}_C\,;\boldsymbol{\Theta}^*)$ for the missing data (with parameter estimates $\boldsymbol{\Theta}^*$ from the previous EM iteration), the auxiliary function is formed by taking the conditional expectation

$$Q_{\mathbf{Z}}(\boldsymbol{\Theta}\,;\boldsymbol{\Theta}^*,\mathbf{M}_C) \;=\; \int_{\mathcal{Z}} d\mathbf{Z}\; p(\mathbf{Z}|\,\mathbf{M}_C\,;\boldsymbol{\Theta}^*)\;\log\, p(\mathbf{Z},\mathbf{M}_C\,;\boldsymbol{\Theta})\,, \tag{9}$$

where $\int_{\mathcal{Z}} d\mathbf{Z}$ is the marginalization operator for the missing measurements. This marginalization is represented by the sequence of (possibly multivariate) integrations given by

$$\int_{\mathcal{Z}} d\mathbf{Z} \;=\; \left\{\int_{\mathcal{Z}_1} d\mathbf{z}_{11}\cdots\int_{\mathcal{Z}_1} d\mathbf{z}_{1m_1}\right\}\cdots\left\{\int_{\mathcal{Z}_L} d\mathbf{z}_{L1}\cdots\int_{\mathcal{Z}_L} d\mathbf{z}_{Lm_L}\right\}, \tag{10}$$

which can be expressed using the shorthand notation

$$\int_{\mathcal{Z}} d\mathbf{Z} \;\equiv\; \prod_{\ell=1}^{L}\prod_{k=1}^{m_\ell}\left\{\int_{\mathcal{Z}_\ell} d\mathbf{z}_{\ell k}\right\}. \tag{11}$$

Care must be exercised when using this shorthand notation because a set of nested integrals is clearly *not* equal to a product of single-variable integrals. However, due to the non-overlapping nature of the integration regions (i.e., the $\mathcal{Z}_k$), the cross terms in the nested integrals are zero, such that equation (10) is equivalent to equation (11) in this case.

CDLF and Posterior Distribution. With missing measurements, it is perhaps easiest to start with the posterior distribution of the missing data and then derive the CDLF from it. With no other knowledge, each measurement is governed by the distribution $p(\mathbf{z}_{\ell k}\,;\boldsymbol{\Theta})$. Because the measurement $\mathbf{z}_{\ell k}$ is known to reside in the ℓth bin, however, its distribution in this restricted region is

$$p(\mathbf{z}_{\ell k}|\,\mathcal{Z}_\ell\,;\boldsymbol{\Theta}) \;=\; \frac{p(\mathbf{z}_{\ell k}\,;\boldsymbol{\Theta})}{\Phi_\ell(\boldsymbol{\Theta})}\,. \tag{12}$$

The intensities provide the numbers of independent measurements in each bin. Given these, the likelihood of the group of measurements is just the product of the likelihoods of each measurement. The posterior of the measurements given the intensities is therefore

$$p(\mathbf{Z}|\,\mathbf{M}_C\,;\boldsymbol{\Theta}) \;=\; \prod_{\ell=1}^{L}\prod_{k=1}^{m_\ell} p(\mathbf{z}_{\ell k}|\,\mathcal{Z}_\ell\,;\boldsymbol{\Theta}) \;=\; \prod_{\ell=1}^{L}\prod_{k=1}^{m_\ell}\left\{\frac{p(\mathbf{z}_{\ell k}\,;\boldsymbol{\Theta})}{\Phi_\ell(\boldsymbol{\Theta})}\right\}. \tag{13}$$

The CDLF is derived from equations (6) and (13) by applying Bayes' rule and canceling terms, which gives

$$p(\mathbf{Z},\mathbf{M}_C\,;\boldsymbol{\Theta}) \;=\; p(\mathbf{M}_C\,;\boldsymbol{\Theta})\;p(\mathbf{Z}|\,\mathbf{M}_C\,;\boldsymbol{\Theta}) \;=\; c(\mathbf{M}_C)\prod_{\ell=1}^{L}\prod_{k=1}^{m_\ell} p(\mathbf{z}_{\ell k}\,;\boldsymbol{\Theta})\,. \tag{14}$$

The log-CDLF is then given by

$$\log p(\mathbf{Z}, \mathbf{M}_{\mathrm{C}}\,; \boldsymbol{\Theta}) \;=\; \log c(\mathbf{M}_{\mathrm{C}}) + \sum_{\ell=1}^{L} \sum_{k=1}^{m_\ell} \log p(\mathbf{z}_{\ell k}\,; \boldsymbol{\Theta})\,. \tag{15}$$

Dropping the term $\log c(\mathbf{M}_{\mathrm{C}})$, which does not depend on $\boldsymbol{\Theta}$, the auxiliary function becomes

$$Q_{\mathbf{Z}}(\boldsymbol{\Theta}\,; \boldsymbol{\Theta}^*, \mathbf{M}_{\mathrm{C}}) \;=\; \int_{\mathcal{Z}} d\mathbf{Z}\; p(\mathbf{Z}|\,\mathbf{M}_{\mathrm{C}}\,; \boldsymbol{\Theta}^*) \sum_{\ell=1}^{L} \sum_{k=1}^{m_\ell} \log p(\mathbf{z}_{\ell k}\,; \boldsymbol{\Theta})\,. \tag{16}$$

Conditional Expectation. Due to the independence of the measurements and the structure of the posterior distribution in equation (13), the conditional expectation operation can be expressed as

$$\int_{\mathcal{Z}} d\mathbf{Z}\; p(\mathbf{Z}|\,\mathbf{M}_{\mathrm{C}}\,; \boldsymbol{\Theta}^*) \;=\; \prod_{\ell=1}^{L} \prod_{k=1}^{m_\ell} \left\{ \frac{1}{\Phi_\ell(\boldsymbol{\Theta}^*)} \int_{\mathcal{Z}_\ell} d\mathbf{z}_{\ell k}\; p(\mathbf{z}_{\ell k}\,; \boldsymbol{\Theta}^*) \right\}, \tag{17}$$

where, noting the definition of $\Phi_\ell(\boldsymbol{\Theta}^*)$ in equation (4), each component in this expression satisfies

$$\frac{1}{\Phi_\ell(\boldsymbol{\Theta}^*)} \int_{\mathcal{Z}_\ell} d\mathbf{z}_{\ell k}\; p(\mathbf{z}_{\ell k}\,; \boldsymbol{\Theta}^*) \;=\; 1\,. \tag{18}$$

When the conditional expectation operation in equation (17) is applied to the log of the CDLF, it applies individually to each of the log terms that appear after the summations in equation (16). Furthermore, the components in the conditional expectation all integrate to one, except for those in which the indices of the integration variable match the indices of the log term being operated on. The auxiliary function therefore becomes

$$Q_{\mathbf{Z}}(\boldsymbol{\Theta}\,; \boldsymbol{\Theta}^*, \mathbf{M}_{\mathrm{C}}) \;=\; \sum_{\ell=1}^{L} \frac{1}{\Phi_\ell(\boldsymbol{\Theta}^*)} \sum_{k=1}^{m_\ell} \int_{\mathcal{Z}_\ell} d\mathbf{z}_{\ell k}\; p(\mathbf{z}_{\ell k}\,; \boldsymbol{\Theta}^*)\; \log\, p(\mathbf{z}_{\ell k}\,; \boldsymbol{\Theta})\,.$$

At this point, the indices ℓ and k on the measurements have no significance because the integration is carried out independently for each summand (i.e., the integration is "inside" the summation operations), and because there is no actual set of observed measurements into which one might have to index. The indices ℓ and k are therefore dropped from the measurements to obtain the expression

$$Q_{\mathbf{Z}}(\boldsymbol{\Theta}\,; \boldsymbol{\Theta}^*, \mathbf{M}_{\mathrm{C}}) \;=\; \sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)} \int_{\mathcal{Z}_\ell} d\mathbf{z}\; p(\mathbf{z}\,; \boldsymbol{\Theta}^*)\; \log\, p(\mathbf{z}\,; \boldsymbol{\Theta})\,. \tag{19}$$

At the level of generality considered thus far, further developments are impossible because maximization of the auxiliary function (the M-step) requires a particular form for $p(\mathbf{z}\,;\boldsymbol{\Theta})$. The M-step is carried out below, however, for the special case of a multivariate normal density.

M-Step for a Gaussian Model. As an example, consider estimating the mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{P}$ in the Gaussian distribution

$$p(\mathbf{z}\,;\boldsymbol{\Theta}) \;=\; \mathcal{N}(\mathbf{z};\boldsymbol{\mu},\mathbf{P}) \;=\; \left|\,2\pi\,\mathbf{P}\,\right|^{-1/2} \exp\left\{-\frac{1}{2}\,(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathbf{P}^{-1}\,(\mathbf{z}-\boldsymbol{\mu})\right\}. \tag{20}$$

In this case, the auxiliary function given in equation (19) becomes

$$Q_{\mathbf{Z}}(\boldsymbol{\Theta}\,;\boldsymbol{\Theta}^*,\mathbf{M}_{\mathrm{C}}) = \sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)} \int_{\mathcal{Z}_\ell} d\mathbf{z}\;\mathcal{N}(\mathbf{z};\boldsymbol{\mu}^*,\mathbf{P}^*)\;\log\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\mathbf{P})\,, \tag{21}$$

where $\boldsymbol{\mu}^*$ and $\mathbf{P}^*$ are the existing estimates from the previous EM iteration. The estimators for the mean and covariance are given in terms of a set of sufficient statistics for the posterior distribution. In addition to the bin probability

$$\Phi_\ell(\boldsymbol{\Theta}^*) \;=\; \int_{\mathcal{Z}_\ell} d\mathbf{z}\;\mathcal{N}(\mathbf{z};\boldsymbol{\mu}^*,\mathbf{P}^*)\,, \tag{22}$$

these sufficient statistics include the centroid and spread variables for each histogram cell, which are given respectively as

$$\boldsymbol{\omega}_\ell(\boldsymbol{\Theta}^*) \;=\; \int_{\mathcal{Z}_\ell} d\mathbf{z}\;\mathcal{N}(\mathbf{z};\boldsymbol{\mu}^*,\mathbf{P}^*)\;\mathbf{z}\,, \tag{23}$$

$$\boldsymbol{\Omega}_\ell(\boldsymbol{\Theta}^*) \;=\; \int_{\mathcal{Z}_\ell} d\mathbf{z}\;\mathcal{N}(\mathbf{z};\boldsymbol{\mu}^*,\mathbf{P}^*)\;\mathbf{z}\,\mathbf{z}^{\mathrm{T}}. \tag{24}$$

Computational algorithms for equations (22)–(24) are given for the case of scalar measurements in appendix E. The estimators for the mean vector and covariance matrix are derived in appendix D and are given by

$$\hat{\boldsymbol{\mu}} \;=\; \frac{1}{K}\sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)}\;\boldsymbol{\omega}_\ell(\boldsymbol{\Theta}^*)\,, \tag{25}$$

$$\hat{\mathbf{P}} \;=\; \frac{1}{K}\sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)}\;\tilde{\boldsymbol{\Omega}}_\ell(\boldsymbol{\Theta}^*,\hat{\boldsymbol{\mu}})\,, \tag{26}$$

where $\tilde{\boldsymbol{\Omega}}_\ell$ is the center-shifted second local moment, which is computed from the centroid and spread variables as

$$\tilde{\boldsymbol{\Omega}}_\ell(\boldsymbol{\Theta}^*,\hat{\boldsymbol{\mu}}) \;=\; \boldsymbol{\Omega}_\ell(\boldsymbol{\Theta}^*) - 2\,\hat{\boldsymbol{\mu}}\;\boldsymbol{\omega}_\ell(\boldsymbol{\Theta}^*) + \hat{\boldsymbol{\mu}}^2\;\Phi_\ell(\boldsymbol{\Theta}^*)\,. \tag{27}$$

As shown in appendix D, estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{P}}$ are located at a critical point of the auxiliary function, and, due to the concave structure of the auxiliary function, they also locate the global maximum.

3.2 ESTIMATION FROM TRUNCATED HISTOGRAM DATA

When histogram intensity data for various bins or regions are unavailable for some reason, then the estimator needs to acknowledge that there are unknown values in those bins or regions, instead of assuming that they are zero. The EM algorithm handily addresses this problem by letting the unobservable intensities be missing data.

To establish notation, let $\mathcal{Z}_{\mathrm{O}}$ denote the observable region of the measurement space, which is partitioned into L_{O} histogram cells as

$$\mathcal{Z}_{\mathrm{O}} = \bigcup_{\ell=1}^{L_{\mathrm{O}}} \mathcal{Z}_\ell \,. \tag{28}$$

Further, let $\mathcal{Z}_{\mathrm{T}}$ denote the truncated region, where particles are not counted (e.g., at the edge of an image or at either end of a frequency band). This region is partitioned into L_{T} cells as

$$\mathcal{Z}_{\mathrm{T}} = \bigcup_{\ell=L_{\mathrm{O}}+1}^{L} \mathcal{Z}_\ell \,, \tag{29}$$

where $L = L_{\mathrm{O}} + L_{\mathrm{T}}$ is the total number of cells that cover the concatenated space

$$\mathcal{Z} = \mathcal{Z}_{\mathrm{O}} \cup \mathcal{Z}_{\mathrm{T}} \,. \tag{30}$$

The unconditional probability of a particle in the observable region is

$$\Phi_{\mathrm{O}}(\boldsymbol{\Theta}) = \int_{\mathcal{Z}_{\mathrm{O}}} d\mathbf{z}\; p(\mathbf{z};\boldsymbol{\Theta}) = \sum_{\ell=1}^{L_{\mathrm{O}}} \Phi_\ell(\boldsymbol{\Theta}) \,, \tag{31}$$

and the particle probability for the truncated region is

$$\Phi_{\mathrm{T}}(\boldsymbol{\Theta}) = \int_{\mathcal{Z}_{\mathrm{T}}} d\mathbf{z}\; p(\mathbf{z};\boldsymbol{\Theta}) = \sum_{\ell=L_{\mathrm{O}}+1}^{L} \Phi_\ell(\boldsymbol{\Theta}) \,. \tag{32}$$

Since $\mathcal{Z}_{\mathrm{O}}$ and $\mathcal{Z}_{\mathrm{T}}$ satisfy equation (30), $\Phi_{\mathrm{O}}(\boldsymbol{\Theta})$ and $\Phi_{\mathrm{T}}(\boldsymbol{\Theta})$ satisfy

$$\Phi_{\mathrm{O}}(\boldsymbol{\Theta}) + \Phi_{\mathrm{T}}(\boldsymbol{\Theta}) = 1 \,. \tag{33}$$

The collection of observed intensities is denoted

$$\mathbf{M}_{\mathrm{O}} = \left\{ m_\ell : \ell = 1, \ldots, L_{\mathrm{O}} \right\}, \tag{34}$$

and the total number of observed particles is

$$K_{\mathrm{O}} = \sum_{\ell=1}^{L_{\mathrm{O}}} m_\ell \,. \tag{35}$$

Since the observed intensities represent K_{O} draws on L_{O} categories, $\mathbf{M}_{\mathrm{O}}$ is multinomially distributed. But since the unconditional probabilities of the observed histogram bins do not sum to one, the ODLF is given by

$$\begin{aligned} p(\mathbf{M}_{\mathrm{O}}\,;\boldsymbol{\Theta}) &= c(\mathbf{M}_{\mathrm{O}}) \prod_{\ell=1}^{L_{\mathrm{O}}} \left\{ \frac{\Phi_\ell(\boldsymbol{\Theta})}{\Phi_{\mathrm{O}}(\boldsymbol{\Theta})} \right\}^{m_\ell} \\ &= \frac{K_{\mathrm{O}}\,!}{\left\{ \prod_{\ell=1}^{L_{\mathrm{O}}} m_\ell\,! \right\}} \left\{ \Phi_{\mathrm{O}}(\boldsymbol{\Theta}) \right\}^{-K_{\mathrm{O}}} \prod_{\ell=1}^{L_{\mathrm{O}}} \left\{ \Phi_\ell(\boldsymbol{\Theta}) \right\}^{m_\ell} . \end{aligned} \tag{36}$$

The EM algorithm for optimizing this likelihood function is developed in two stages. First, the unobserved intensities are treated as the sole piece of missing data. This is then extended to the case of missing measurements and intensities.

3.2.1 EM Algorithm with Missing Hit Counts

Consider, for a moment, estimating the distribution parameters using an EM algorithm that treats the truncated intensities as missing data but does *not* include the measurements in the missing data. The resulting algorithm is of little practical interest since the M-step likely involves a difficult nonlinear problem. This situation is considered, however, to isolate the issues involved with truncated bin intensities. After getting an understanding of these issues in the present subsection, the next subsection takes on the case where both the truncated intensities and measurements are included in the missing data.

Definition of Missing Data and Auxiliary Function. For the present discussion, the missing data are the intensities in the truncated region, denoted

$$\mathbf{M}_{\mathrm{T}} = \left\{ m_\ell : \ell = L_{\mathrm{O}} + 1, \ldots, L \right\}, \tag{37}$$

and the auxiliary function is defined as

$$Q_{\mathbf{M}_{\mathrm{T}}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{M}_{\mathrm{O}}) = \sum_{\mathcal{M}_{\mathrm{T}}} p(\mathbf{M}_{\mathrm{T}} | \mathbf{M}_{\mathrm{O}}\,; \boldsymbol{\Theta}^*) \, \log p(\mathbf{M}_{\mathrm{T}}, \mathbf{M}_{\mathrm{O}}\,; \boldsymbol{\Theta}) \,. \tag{38}$$

The marginalization operator for the missing intensities is expressed using the shorthand notation

$$\sum_{\mathcal{M}_{\mathrm{T}}} \equiv \prod_{\ell=L_{\mathrm{O}}+1}^{L} \left\{ \sum_{m_\ell=0}^{\infty} \right\} \equiv \sum_{m_{L_{\mathrm{O}}+1}=0}^{\infty} \sum_{m_{L_{\mathrm{O}}+2}=0}^{\infty} \cdots \sum_{m_L=0}^{\infty} . \tag{39}$$

Similar to the notation introduced in equation (11), this shorthand notation is only accurate when the cross-terms in the nested summations are zero, in which case the nested sum does indeed reduce to a product of individual summations.

CDLF and Posterior Distribution. The posterior distribution of the missing intensities is derived in appendix C, and is given by

$$p(\mathbf{M}_{\mathrm{T}} | \mathbf{M}_{\mathrm{O}} ; \boldsymbol{\Theta}) = c^{-}(\mathbf{M}_{\mathrm{T}}, K_{\mathrm{O}}) \left\{ \Phi_{\mathrm{O}}(\boldsymbol{\Theta}) \right\}^{K_{\mathrm{O}}} \prod_{\ell=L_{\mathrm{O}}+1}^{L} \left\{ \Phi_{\ell}(\boldsymbol{\Theta}) \right\}^{m_{\ell}}, \tag{40}$$

where $c^{-}(\mathbf{M}_{\mathrm{T}}, K_{\mathrm{O}})$ is the *negative multinomial coefficient* defined in equation (140). The CDLF is given in terms of the ODLF and posterior distribution as

$$\begin{aligned} p(\mathbf{M}_{\mathrm{T}}, \mathbf{M}_{\mathrm{O}} ; \boldsymbol{\Theta}) &= p(\mathbf{M}_{\mathrm{O}} ; \boldsymbol{\Theta})\; p(\mathbf{M}_{\mathrm{T}} | \mathbf{M}_{\mathrm{O}} ; \boldsymbol{\Theta}) & (41) \\ &= c(\mathbf{M}_{\mathrm{O}})\; c^{-}(\mathbf{M}_{\mathrm{T}}, K_{\mathrm{O}}) \prod_{\ell=1}^{L} \left\{ \Phi_{\ell}(\boldsymbol{\Theta}) \right\}^{m_{\ell}}, & (42) \end{aligned}$$

whose logarithm is

$$\log p(\mathbf{M}_{\mathrm{T}}, \mathbf{M}_{\mathrm{O}} ; \boldsymbol{\Theta}) = \log c(\mathbf{M}_{\mathrm{O}}) + \log c^{-}(\mathbf{M}_{\mathrm{T}}, K_{\mathrm{O}}) + \sum_{\ell=1}^{L} m_{\ell}\; \log \Phi_{\ell}(\boldsymbol{\Theta}) . \tag{43}$$

Since the first two terms in equation (43) do not depend $\boldsymbol{\Theta}$, they can be dropped from the auxiliary function to obtain

$$Q_{\mathbf{M}_{\mathrm{T}}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{*}, \mathbf{M}_{\mathrm{O}}) = \sum_{\mathcal{M}_{\mathrm{T}}} p(\mathbf{M}_{\mathrm{T}} | \mathbf{M}_{\mathrm{O}} ; \boldsymbol{\Theta}^{*}) \sum_{\ell=1}^{L} m_{\ell}\; \log \Phi_{\ell}(\boldsymbol{\Theta}) . \tag{44}$$

Conditional Expectation. The expression in equation (44) is a linear function of the missing bin intensities, with respect to which the expectation is being taken. Because of this linearity, the expected value of the log-CDLF is obtained merely by substituting the expected values of the missing data, leading to the expression

$$Q_{\mathbf{M}_{\mathrm{T}}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{*}, \mathbf{M}_{\mathrm{O}}) = \sum_{\ell=1}^{L} \tilde{m}_{\ell}\; \log \Phi_{\ell}(\boldsymbol{\Theta}) , \tag{45}$$

where the expected intensities are defined by

$$\tilde{m}_{\ell} = \begin{cases} m_{\ell} & \text{if } \ell = 1, \ldots, L_{\mathrm{O}} \\ E\left\{ m_{\ell} | \mathbf{M}_{\mathrm{O}} ; \boldsymbol{\Theta}^{*} \right\} & \text{if } \ell = L_{\mathrm{O}} + 1, \ldots, L . \end{cases} \tag{46}$$

The expected intensities for the truncated cells are derived in appendix C and are given for $\ell = L_O + 1, \ldots, L$ as

$$E\Big\{m_\ell | \mathbf{M}_O ; \boldsymbol{\Theta}^*\Big\} = \left\{\frac{K_O}{\Phi_O(\boldsymbol{\Theta}^*)}\right\} \Phi_\ell(\boldsymbol{\Theta}^*) . \tag{47}$$

This expression extrapolates the intensities from the observed region into the truncated region by using the ratio $\{ K_O/\Phi_O(\boldsymbol{\Theta}^*)\}$ as a "probability-to-intensity" conversion factor and then applying this conversion factor to the unconditional probability in each truncated bin.

3.2.2 EM Algorithm with Missing Hit Counts and Measurements

The results of the previous subsection are now extended such that the missing data include both the truncated intensities and measurements.

Definition of Missing Data and Auxiliary Function. The missing data are defined by the sets

$$\begin{aligned} \mathbf{M}_T &= \Big\{ m_\ell : \ell = L_O + 1, \ldots, L \Big\} , \\ \mathbf{Z} &= \Big\{ \mathbf{z}_{\ell k} : \ell = 1, \ldots, L, \;\; k = 1, \ldots, m_\ell \Big\} . \end{aligned} \tag{48}$$

The auxiliary function is defined by

$$Q_{\mathbf{Z},\mathbf{M}_T}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{M}_O) = \sum_{\mathcal{M}_T} \int_{\mathcal{Z}} d\mathbf{Z} \; p(\mathbf{Z}, \mathbf{M}_T | \mathbf{M}_O; \boldsymbol{\Theta}^*) \; \log p(\mathbf{Z}, \mathbf{M}_T, \mathbf{M}_O ; \boldsymbol{\Theta}) , \tag{49}$$

where the marginalization operators for the missing data are again expressed using the shorthand notation

$$\begin{aligned} \sum_{\mathcal{M}_T} &\equiv \prod_{\ell=L_O+1}^{L} \left\{ \sum_{m_\ell=0}^{\infty} \right\} , \\ \int_{\mathcal{Z}} d\mathbf{Z} &\equiv \prod_{\ell=1}^{L} \prod_{k=1}^{m_\ell} \left\{ \int_{\mathcal{Z}_\ell} d\mathbf{z}_{\ell k} \right\} . \end{aligned}$$

CDLF and Posterior Distribution. The distribution $p(\mathbf{Z}, \mathbf{M}_T, \mathbf{M}_O ; \boldsymbol{\Theta})$ can be factored using Bayes' rule as

$$\begin{aligned} p(\mathbf{Z}, \mathbf{M}_T, \mathbf{M}_O ; \boldsymbol{\Theta}) &= p(\mathbf{M}_O ; \boldsymbol{\Theta}) \; p(\mathbf{M}_T | \mathbf{M}_O ; \boldsymbol{\Theta}) \; p(\mathbf{Z} | \mathbf{M}_T, \mathbf{M}_O; \boldsymbol{\Theta}) \\ &= p(\mathbf{M}_O ; \boldsymbol{\Theta}) \; p(\mathbf{M}_T | \mathbf{M}_O ; \boldsymbol{\Theta}) \; p(\mathbf{Z} | \mathbf{M}_C; \boldsymbol{\Theta}) , \end{aligned} \tag{50}$$

where $p(\mathbf{Z}|\,\mathbf{M}_{\mathrm{T}},\mathbf{M}_{\mathrm{O}};\boldsymbol{\Theta}) = p(\mathbf{Z}|\,\mathbf{M}_{\mathrm{C}};\boldsymbol{\Theta})$ because, once the missing intensities are given, the concatenated set $\{\mathbf{M}_{\mathrm{O}},\mathbf{M}_{\mathrm{T}}\}$ corresponds to the set of intensities for the complete histogram. Equation (50) serves as the starting point both for obtaining the posterior distribution of the missing data and for evaluating the CDLF to be substituted into equation (49). A suitable expression for this latter purpose is obtained by substituting equations (36), (40), and (13), and canceling terms, which gives

$$p(\mathbf{Z},\mathbf{M}_{\mathrm{T}},\mathbf{M}_{\mathrm{O}}\,;\boldsymbol{\Theta}) \;=\; c(\mathbf{M}_{\mathrm{O}})\;\, c^{-}(\mathbf{M}_{\mathrm{T}},K_{\mathrm{O}})\;\prod_{\ell=1}^{L}\prod_{k=1}^{m_\ell} p(\mathbf{z}_{\ell k}\,;\boldsymbol{\Theta})\,. \tag{51}$$

The log of the CDLF is therefore given by

$$\log p(\mathbf{Z},\mathbf{M}_{\mathrm{T}},\mathbf{M}_{\mathrm{O}}\,;\boldsymbol{\Theta}) \;=\; \log c(\mathbf{M}_{\mathrm{O}}) + \log c^{-}(\mathbf{M}_{\mathrm{T}},K_{\mathrm{O}}) + \sum_{\ell=1}^{L}\sum_{k=1}^{m_\ell} \log p(\mathbf{z}_{\ell k}\,;\boldsymbol{\Theta})\,, \tag{52}$$

such that, after ignoring constant terms $\log c(\mathbf{M}_{\mathrm{O}})$ and $\log c^{-}(\mathbf{M}_{\mathrm{T}},K_{\mathrm{O}})$, the auxiliary function becomes

$$Q_{\mathbf{Z},\mathbf{M}_{\mathrm{T}}}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{*},\mathbf{M}_{\mathrm{O}}) \;=\; \sum_{\mathcal{M}_{\mathrm{T}}}\int_{\mathcal{Z}} d\mathbf{Z}\; p(\mathbf{Z},\mathbf{M}_{\mathrm{T}}|\,\mathbf{M}_{\mathrm{O}};\boldsymbol{\Theta}^{*})\;\sum_{\ell=1}^{L}\sum_{k=1}^{m_\ell} \log p(\mathbf{z}_{\ell k}\,;\boldsymbol{\Theta})\,. \tag{53}$$

The missing-data posterior distribution is obtained by dividing equation (50) by $p(\mathbf{M}_{\mathrm{O}}\,;\boldsymbol{\Theta})$, which gives

$$p(\mathbf{Z},\mathbf{M}_{\mathrm{T}}|\,\mathbf{M}_{\mathrm{O}};\boldsymbol{\Theta}^{*}) \;=\; p(\mathbf{M}_{\mathrm{T}}|\,\mathbf{M}_{\mathrm{O}}\,;\boldsymbol{\Theta}^{*})\; p(\mathbf{Z}|\,\mathbf{M}_{\mathrm{C}};\boldsymbol{\Theta}^{*})\,. \tag{54}$$

Conditional Expectation. The expectation operation in equation (53) can be decomposed as

$$\sum_{\mathcal{M}_{\mathrm{T}}}\int_{\mathcal{Z}} d\mathbf{Z}\; p(\mathbf{Z},\mathbf{M}_{\mathrm{T}}|\,\mathbf{M}_{\mathrm{O}};\boldsymbol{\Theta}^{*}) \;=\; \sum_{\mathcal{M}_{\mathrm{T}}} p(\mathbf{M}_{\mathrm{T}}|\,\mathbf{M}_{\mathrm{O}};\boldsymbol{\Theta}^{*})\;\int_{\mathcal{Z}} d\mathbf{Z}\; p(\mathbf{Z}|\,\mathbf{M}_{\mathrm{C}};\boldsymbol{\Theta}^{*})\,, \tag{55}$$

allowing the auxiliary function to be expressed as

$$\begin{aligned} Q_{\mathbf{Z},\mathbf{M}_{\mathrm{T}}}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{*},\mathbf{M}_{\mathrm{O}}) \;=\; & \sum_{\mathcal{M}_{\mathrm{T}}} p(\mathbf{M}_{\mathrm{T}}|\,\mathbf{M}_{\mathrm{O}};\boldsymbol{\Theta}^{*}) \\ & \times \int_{\mathcal{Z}} d\mathbf{Z}\; p(\mathbf{Z}|\,\mathbf{M}_{\mathrm{C}};\boldsymbol{\Theta}^{*})\;\sum_{\ell=1}^{L}\sum_{k=1}^{m_\ell} \log p(\mathbf{z}_{\ell k}\,;\boldsymbol{\Theta})\,. \end{aligned} \tag{56}$$

Noting equation (16), the second line of this expression is just the auxiliary function for the complete histogram, such that

$$Q_{\mathbf{Z},\mathbf{M}_{\mathrm{T}}}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{*},\mathbf{M}_{\mathrm{O}}) \;=\; \sum_{\mathcal{M}_{\mathrm{T}}} p(\mathbf{M}_{\mathrm{T}}|\,\mathbf{M}_{\mathrm{O}};\boldsymbol{\Theta}^{*})\; Q_{\mathbf{Z}}(\boldsymbol{\Theta}\,;\boldsymbol{\Theta}^{*},\mathbf{M}_{\mathrm{C}})\,. \tag{57}$$

Substituting the final form for $Q_{\mathbf{Z}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{M}_{\mathrm{C}})$ given in equation (19) again leaves an expectation of a function that is linear in the missing intensities. Paralleling the case in which only the intensities are missing, the auxiliary function simplifies to

$$Q_{\mathbf{Z},\mathbf{M}_{\mathrm{T}}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{M}_{\mathrm{O}}) = \sum_{\ell=1}^{L} \frac{\tilde{m}_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)} \int_{\mathcal{Z}_\ell} d\mathbf{z}\; p(\mathbf{z}; \boldsymbol{\Theta}^*) \log p(\mathbf{z}; \boldsymbol{\Theta}), \tag{58}$$

where $\tilde{m}_\ell$ is defined in equation (46). Once again it is impossible to go any further without imposing a form on $p(\mathbf{x}; \boldsymbol{\Theta})$. For the special case of Gaussian measurements, maximization of equation (58) over the mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{P}$ yields estimates that are equal to those in equation (25) and equation (26), but with m_ℓ replaced with $\tilde{m}_\ell$ and K replaced with $\tilde{K} = \sum_{\ell=1}^{L} \tilde{m}_\ell$. The Gaussian parameter estimates after each iteration of the EM algorithm with truncated histogram data are thus

$$\hat{\boldsymbol{\mu}} = \frac{1}{\tilde{K}} \sum_{\ell=1}^{L} \frac{\tilde{m}_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)}\; \boldsymbol{\omega}_\ell(\boldsymbol{\Theta}^*), \tag{59}$$

$$\hat{\mathbf{P}} = \frac{1}{\tilde{K}} \sum_{\ell=1}^{L} \frac{\tilde{m}_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)}\; \tilde{\boldsymbol{\Omega}}_\ell(\boldsymbol{\Theta}^*, \hat{\boldsymbol{\mu}}), \tag{60}$$

where $\boldsymbol{\omega}_\ell(\boldsymbol{\Theta}^*)$ and $\tilde{\boldsymbol{\Omega}}_\ell(\boldsymbol{\Theta}^*, \hat{\boldsymbol{\mu}})$ are defined in equations (23) and (27), respectively.

3.3 RELATIONSHIP BETWEEN COMPLETE AND TRUNCATED HISTOGRAMS

In the analysis of truncated histograms, the auxiliary function with missing measurements and histogram intensities reduced to an expected version of the auxiliary function for the complete histogram. This occured because of the similarity of the CDLF for the two situations. It is useful for later developments with mixtures to flesh out this equivalence more generally. Consider the generic problem in which the missing data contain some set of variables $\boldsymbol{\Xi}$. For a complete histogram, the CDLF in most cases of practical interest can be expressed as

$$\begin{aligned} p(\boldsymbol{\Xi}, \mathbf{M}_{\mathrm{C}}\,; \boldsymbol{\Theta}) &= p(\mathbf{M}_{\mathrm{C}}\,; \boldsymbol{\Theta})\; p(\boldsymbol{\Xi} | \mathbf{M}_{\mathrm{C}}\,; \boldsymbol{\Theta}) \\ &= c(\mathbf{M}_{\mathrm{C}}) \prod_{\ell=1}^{L} \Big\{ \Phi_\ell(\boldsymbol{\Theta}) \Big\}^{m_\ell}\; p(\boldsymbol{\Xi} | \mathbf{M}_{\mathrm{C}}\,; \boldsymbol{\Theta}). \end{aligned} \tag{61}$$

The CDLF for the truncated histogram, on the other hand, is given by

$$\begin{aligned} p(\boldsymbol{\Xi}, \mathbf{M}_{\mathrm{T}}, \mathbf{M}_{\mathrm{O}}\,; \boldsymbol{\Theta}) &= p(\mathbf{M}_{\mathrm{T}}, \mathbf{M}_{\mathrm{O}}\,; \boldsymbol{\Theta})\; p(\boldsymbol{\Xi} | \mathbf{M}_{\mathrm{C}}\,; \boldsymbol{\Theta}) \\ &= c(\mathbf{M}_{\mathrm{O}})\; c^{-}(\mathbf{M}_{\mathrm{T}}, K_{\mathrm{O}}) \prod_{\ell=1}^{L} \Big\{ \Phi_\ell(\boldsymbol{\Theta}) \Big\}^{m_\ell}\; p(\boldsymbol{\Xi} | \mathbf{M}_{\mathrm{C}}\,; \boldsymbol{\Theta}). \end{aligned} \tag{62}$$

With the exception of the coefficients, these two expressions are identical. Substituting the definitions of the multinomial and negative multinomial coefficients in the CDLF for the truncated histogram gives

$$\begin{aligned} c(\mathbf{M}_{\mathrm{O}})\ c^{-}(\mathbf{M}_{\mathrm{T}}, K_{\mathrm{O}}) &= \frac{K_{\mathrm{O}}!}{\left\{\prod_{\ell=1}^{L_{\mathrm{O}}} m_\ell!\right\}} \frac{(K-1)!}{(K_{\mathrm{O}}-1)!\left\{\prod_{\ell=L_{\mathrm{O}}+1}^{L} m_\ell!\right\}} \\ &= \left(\frac{K_{\mathrm{O}}}{K}\right) \frac{K!}{\left\{\prod_{\ell=1}^{L} m_\ell!\right\}} \\ &= \left(\frac{K_{\mathrm{O}}}{K}\right) c(\mathbf{M}_{\mathrm{C}}), \end{aligned} \tag{63}$$

such that the two CDLFs are related by the expression

$$p(\Xi, \mathbf{M}_{\mathrm{T}}, \mathbf{M}_{\mathrm{O}}\,; \Theta) = \left(\frac{K_{\mathrm{O}}}{K}\right) p(\Xi, \mathbf{M}_{\mathrm{C}}\,; \Theta)\,. \tag{64}$$

This expression gives an indication of the likelihood "loss" that results from the unobservability of some of the intensities. This relationship between the CDLFs results in a similar relationship between the auxiliary functions for the two problems. The auxiliary function for the truncated histogram can be written as

$$\begin{aligned} Q_{\Xi,\mathbf{M}_{\mathrm{T}}}(\Theta; \Theta^*, \mathbf{M}_{\mathrm{O}}) &= \sum_{\mathcal{M}_{\mathrm{T}}} p(\mathbf{M}_{\mathrm{T}} | \mathbf{M}_{\mathrm{O}}\,; \Theta^*) \int d\Xi\ p(\Xi | \mathbf{M}_{\mathrm{C}}; \Theta^*)\ \log p(\Xi, \mathbf{M}_{\mathrm{T}}, \mathbf{M}_{\mathrm{O}}\,; \Theta) \\ &= \sum_{\mathcal{M}_{\mathrm{T}}} p(\mathbf{M}_{\mathrm{T}} | \mathbf{M}_{\mathrm{O}}\,; \Theta^*) \int d\Xi\ p(\Xi | \mathbf{M}_{\mathrm{C}}; \Theta^*) \\ &\qquad \times \left\{ \log\left(\frac{K_{\mathrm{O}}}{K}\right) + \log p(\Xi, \mathbf{M}_{\mathrm{C}}\,; \Theta) \right\}. \end{aligned}$$

The random variable K is a function of Θ^*, but not of Θ. Since K_{O} is observed, the term $\log(K_{\mathrm{O}}/K)$ is independent of Θ, and the auxiliary function effectively becomes

$$Q_{\Xi,\mathbf{M}_{\mathrm{T}}}(\Theta; \Theta^*, \mathbf{M}_{\mathrm{O}}) = \sum_{\mathcal{M}_{\mathrm{T}}} p(\mathbf{M}_{\mathrm{T}} | \mathbf{M}_{\mathrm{O}}\,; \Theta^*)\ Q_{\Xi}(\Theta; \Theta^*, \mathbf{M}_{\mathrm{C}})\,.$$

Furthermore, the parameter estimators for the complete histogram usually depend linearly on m_ℓ, as was the case in the previous subsection. The approach used above therefore generalizes to most cases of interest; that is, the truncated-histogram estimators are obtained by substituting the expected intensities into the corresponding complete-histogram estimators. One thus never need explicitly derive the estimator for the truncated histogram, so long as the complete-histogram estimator exhibits this linear dependence on the cell intensities. Since this is also the case with the mixture models

discussed in the next section, the results are formulated for complete histograms only. The qualifiers (i.e., subscripts I, C, and T) on histogram-related variables are therefore dropped in the next section, with the understanding that variable m_ℓ in a given expression is an observed intensity for cells in $\mathcal{Z}_\mathrm{O}$ and an expected intensity for cells in $\mathcal{Z}_\mathrm{T}$.

4. HISTOGRAM MIXTURE MODELS

This section discusses parameter estimation for finite mixture models, which introduce another form of missing data, namely the *mode assignments*. Subsection 1 outlines the signal-in-noise mixture model; subsection 2 reviews estimation from point measurements, which illustrates in isolation the issues surrounding mode assignment uncertainty. Subsection 3 then generalizes this to estimation from histogram intensity data, wherein the point measurements and modes assignments are missing data.

4.1 MODEL DEFINITION

Let the distribution of interest (i.e., the "signal distribution") be denoted $p_S(\mathbf{z};\boldsymbol{\theta}_S)$. At issue is the fact that, in noisy data, some of the measurements do not belong to $p_S(\mathbf{z};\boldsymbol{\theta}_S)$, but instead belong to some extraneous distribution $p_0(\mathbf{z};\boldsymbol{\theta}_0)$ (i.e., the "noise distribution"). If one attempts to estimate $\boldsymbol{\theta}_S$ in the signal distribution without accounting for the noise measurements, then the resulting parameter estimates are distorted. Noise measurements are accommodated using the two-mode "signal-or-noise" mixture distribution given by

$$p(\mathbf{z}\,;\pi_0,\boldsymbol{\theta}_0,\boldsymbol{\theta}_S) \;=\; \pi_0\; p_0(\mathbf{z};\boldsymbol{\theta}_0) + (1-\pi_0)\; p_S(\mathbf{z}\,;\boldsymbol{\theta}_S)\,, \tag{65}$$

where π_0 is the unconditional probability that a measurement belongs to the noise. In general, the model distributions $p_S(\mathbf{z};\boldsymbol{\theta}_S)$ and $p_0(\mathbf{z};\boldsymbol{\theta}_0)$ can have different parametric structures (e.g., Gaussian signal distribution and uniform noise distribution).

From this simple statement of the model, more sophisticated models are obtained by letting the individual signal or noise components themselves be mixtures, resulting in a mixture of mixtures. This is most valuable when $p_S(\mathbf{z};\boldsymbol{\theta}_S)$ and/or $p_0(\mathbf{z};\boldsymbol{\theta}_0)$ are either unknown or so complicated that they lead to intractable estimation problems. In the discussion below, a separate mixture is included for the signal distribution only; a mixture model for the noise component is constructed similarly. Thus, the signal distribution is expressed in terms of modal distributions as

$$p_S(\mathbf{z}\,;\boldsymbol{\theta}_S) \;=\; \sum_{j=1}^{J} \psi_j\; p_j(\mathbf{z};\boldsymbol{\theta}_j)\,, \tag{66}$$

where $p_j(\mathbf{z};\boldsymbol{\theta}_j)$ is the jth modal distribution and ψ_j is the unconditional probability that the jth mode is valid (i.e., that $\mathbf{z}$ is actually governed by the jth modal distribution). The collection of these probabilities satisfies the constraint $\sum_{j=1}^{J}\psi_j = 1$. Substituting equation (66) into equation (65), and defining $\pi_j = (1-\pi_0)\,\psi_j$ for $j = 1,\ldots,J$, results

in an overall mixture distribution, including signal and noise models, given by

$$p(\mathbf{z}\,;\Theta) \;=\; \sum_{j=0}^{J} \pi_j \; p_j(\mathbf{z};\boldsymbol{\theta}_j)\,. \tag{67}$$

This model is parameterized by the set $\Theta = \{\boldsymbol{\pi}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_S\}$, where $\boldsymbol{\pi} = \{\pi_0, \pi_1, \ldots, \pi_J\}$ is the collection of unconditional mode assignment probabilities, $\boldsymbol{\theta}_0$ contains the noise distribution parameters, and $\boldsymbol{\theta}_S = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J\}$ is the collection of parameters for all modes of the signal distribution. The right-hand side of equation (67) is a convex combination since $\sum_{j=0}^{J} \pi_j = 1$.

4.2 ESTIMATION FROM POINT MEASUREMENTS

The assignment uncertainty issue is the defining characteristic of finite mixture densities. This issue is now considered in isolation by considering the problem in which a set $\mathbf{Z}$ of independent point measurements are observed. While the resulting algorithm is of value in certain cases in which feature values of observed data samples are directly observed (e.g., see [9]), this is a rather artificial problem in the context of histogram modeling and is presented merely as a stepping stone to the more general problem description in the next subsection. In any case, when observations of the point measurements are assumed to be available, an optimal parameter estimation algorithm maximizes the ODLF given by

$$p(\mathbf{Z}\,;\Theta) \;=\; \prod_{k=1}^{K} \left\{ \sum_{j_k=1}^{J} \pi_{j_k} \; p_{j_k}(\mathbf{z}_k;\boldsymbol{\theta}_{j_k}) \right\}. \tag{68}$$

Direct optimization of this likelihood function with respect to Θ is generally difficult because of the interaction between mixture modes.

4.2.1 EM Algorithm with Missing Assignments

The EM algorithm for mixtures circumvents the mode interaction issue by treating as missing data the *mode assignments* (i.e., the j_k). The idea behind this choice is that, if the mode assignment were known for each measurement (i.e., if the component distribution actually governing each measurement were known), then the measurements could be grouped according to mode and the parameter estimation problem would decompose into J independent problems that are easier to solve. The EM algorithm does induce such a decomposition, albeit in each iteration of an iterative procedure.

Definition of Missing Data and Auxiliary Function. The set of missing data corresponding to the observed measurements in $\mathbf{Z}$ is denoted as

$$\mathbf{J} = \left\{ j_1, j_2, \ldots, j_K \right\}, \tag{69}$$

such that the auxiliary function is given by

$$Q_{\mathbf{J}}(\boldsymbol{\Theta};\boldsymbol{\Theta}^*,\mathbf{Z}) \;=\; \sum_{\mathcal{J}} \; p(\mathbf{J}\,|\,\mathbf{Z};\boldsymbol{\Theta}^*) \;\log\, p(\mathbf{Z},\mathbf{J};\boldsymbol{\Theta})\,. \tag{70}$$

Given that the measurements are independent, the joint density is a product of terms that allows the missing-data marginalization operation to be expressed using the shorthand notation

$$\sum_{\mathcal{J}} \;\equiv\; \prod_{k=1}^{K} \left\{ \sum_{j_k=1}^{J} \right\}. \tag{71}$$

CDLF and Posterior Distribution. The CDLF is obtained by noting the independence of the measurements and by applying Bayes' rule on a measurement-by-measurement basis, which gives

$$p(\mathbf{J},\mathbf{Z}\,;\boldsymbol{\Theta}) \;=\; \prod_{k=1}^{K} \; p(\,j_k;\boldsymbol{\Theta})\; p(\mathbf{z}_k\,|\,j_k;\boldsymbol{\Theta}) \;=\; \prod_{k=1}^{K} \; \pi_{j_k}\; p_{j_k}(\mathbf{z}_k;\boldsymbol{\theta}_{j_k})\,, \tag{72}$$

where $p(\,j_k;\boldsymbol{\Theta}) = \pi_{j_k}$ and $p(\mathbf{z}_k\,|\,j_k;\boldsymbol{\Theta}) = p_{j_k}(\mathbf{z}_k;\boldsymbol{\theta}_{j_k})$. The log-CDLF is therefore

$$\log p(\mathbf{J},\mathbf{Z}\,;\boldsymbol{\Theta}) \;=\; \sum_{k=1}^{K} \; \Big\{\, \log \pi_{j_k} + \log p_{j_k}(\mathbf{z}_k;\boldsymbol{\theta}_{j_k}) \,\Big\}\,. \tag{73}$$

Bayes' rule then gives the posterior distribution as to obtain

$$p(\mathbf{J}\,|\,\mathbf{Z}\,;\boldsymbol{\Theta}^*) \;=\; \frac{p(\mathbf{J},\mathbf{Z}\,;\boldsymbol{\Theta}^*)}{p(\mathbf{Z}\,;\boldsymbol{\Theta}^*)} \;=\; \prod_{k=1}^{K} \; \gamma_{kj_k}(\boldsymbol{\Theta}^*)\,, \tag{74}$$

where $\gamma_{kj_k}(\boldsymbol{\Theta}^*) = p(\,j_k\,|\,\mathbf{z}_k;\boldsymbol{\Theta}^*)$ is the single-measurement posterior assignment probability, which is defined (and computed) as

$$\gamma_{kj_k}(\boldsymbol{\Theta}^*) \;=\; \frac{\pi^*_{j_k}\; p_{j_k}(\mathbf{z}_k;\boldsymbol{\theta}^*_{j_k})}{\Big\{\sum_{i=1}^{J} \pi^*_i\; p_i(\mathbf{z}_k;\boldsymbol{\theta}^*_i)\Big\}}\,. \tag{75}$$

For all values of k, these posterior probabilities satisfy

$$\sum_{j_k=1}^{J} \; \gamma_{kj_k}(\boldsymbol{\Theta}^*) \;=\; 1\,. \tag{76}$$

Conditional Expectation. Given the structure of equation (74), the conditional expectation in equation (70) is a sequence of single-variable expectation operations, which is expressed as a product of operators as

$$\sum_{\mathcal{J}} \; p(\mathbf{J}\,|\,\mathbf{Z}\,;\boldsymbol{\Theta}^*) \;=\; \prod_{k=1}^{K} \left\{ \sum_{j_k=1}^{J} \; \gamma_{kj}(\boldsymbol{\Theta}^*) \right\}. \tag{77}$$

If evaluated as it stands, then all of the single-variable expectations in this expression sum to one (hence the overall expression is one). Similarly, when the conditional-expectation operator is applied to the log-CDLF in equation (73), all of the single-variable expectations sum to one, except those for which the marginalization index matches the index j_k appearing in the log-CDLF. The auxiliary function therefore reduces to

$$Q_{\mathbf{J}}(\Theta; \Theta^*, \mathbf{Z}) = \sum_{k=1}^{K} \sum_{j_k=1}^{J} \gamma_{kj_k}(\Theta^*) \left\{ \log \pi_{j_k} + \log p_{j_k}(\mathbf{z}_k; \boldsymbol{\theta}_{j_k}) \right\}. \tag{78}$$

At this point, the index k on the assignment variable adds no meaning since the summation covers all possible values of the assignment. Furthermore, since j_k is missing data, there is no actual "observed value" for j_k that is tied to a particular measurement $\mathbf{z}_k$. This index is therefore dropped to obtain

$$Q_{\mathbf{J}}(\Theta; \Theta^*, \mathbf{Z}) = \sum_{k=1}^{K} \sum_{j=1}^{J} \gamma_{kj}(\Theta^*) \left\{ \log \pi_j + \log p_j(\mathbf{z}_k; \boldsymbol{\theta}_j) \right\}. \tag{79}$$

Without the k-dependence, the summation over j can be moved in front of the summation over k, and the auxiliary function can be decomposed into a collection of component auxiliary functions, each of which depends on a separate subset of the model parameters. This decomposition is defined as

$$Q_{\mathbf{J}}(\Theta; \Theta^*, \mathbf{Z}) = \sum_{j=1}^{J} Q_{\mathbf{J}}(\pi_j; \Theta^*, \mathbf{Z}) + \sum_{j=1}^{J} Q_{\mathbf{J}}(\boldsymbol{\theta}_j; \Theta^*, \mathbf{Z}), \tag{80}$$

where

$$Q_{\mathbf{J}}(\pi_j; \Theta^*, \mathbf{Z}) = \sum_{k=1}^{K} \gamma_{kj}(\Theta^*) \log \pi_j, \tag{81}$$

$$Q_{\mathbf{J}}(\boldsymbol{\theta}_j; \Theta^*, \mathbf{Z}) = \sum_{k=1}^{K} \gamma_{kj}(\Theta^*) \log p_j(\mathbf{z}_k; \boldsymbol{\theta}_j). \tag{82}$$

Given the mutually exclusive nature of the parameters in these various components, the M-step of the EM algorithm decomposes into a set of independent optimization problems, one for each component. The assignment probabilities are obtained by maximizing equation (81), and the parameters in each mode density are obtained by maximizing equation (82).

Estimation of Assignment Probabilities. Equation (81) is maximized using Lagrange multiplier techniques to obtain the estimator

$$\hat{\pi}_j = \frac{\kappa_j(\Theta^*)}{K}, \tag{83}$$

where $\kappa_j(\Theta^*)$ is the *effective number of measurements* corresponding to mode j given by

$$\kappa_j(\Theta^*) = \sum_{k=1}^{K} \gamma_{kj}(\Theta^*) . \tag{84}$$

The effective numbers of measurements for all modes satisfy

$$\sum_{j=1}^{J} \kappa_j(\Theta^*) = K . \tag{85}$$

Gaussian Mode Estimation. Each component in equation (82) is used to individually estimate the parameters in one of the modal distributions. For Gaussian mode densities, $p_j(\mathbf{z};\boldsymbol{\theta}_j) = \mathcal{N}(\mathbf{z};\, \boldsymbol{\mu}_j, \mathbf{P}_j)$, the optimization problem decomposes into J independent (single-mode) Gaussian estimation problems, which are discussed in appendix D. Drawing on the results of that appendix, equation(82) is maximized by

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{\kappa_j(\Theta^*)} \sum_{k=1}^{K} \gamma_{kj}(\Theta^*)\ \mathbf{z}_k , \tag{86}$$

$$\hat{\mathbf{P}}_j = \frac{1}{\kappa_j(\Theta^*)} \sum_{k=1}^{K} \gamma_{kj}(\Theta^*)\ \left(\mathbf{z}_k - \hat{\boldsymbol{\mu}}_j\right)\left(\mathbf{z}_k - \hat{\boldsymbol{\mu}}_j\right)^{\mathrm{T}} . \tag{87}$$

These Gaussian parameter estimators take the above form regardless of the parametric structure of any other modes. That is, not all modes need be Gaussian.

4.3 ESTIMATION FROM HISTOGRAM DATA

To set the stage for histogram mixture estimation, the unconditional probability of any given histogram cell under the mixture model in equation (67) is given by

$$\Phi_\ell(\Theta) = \int_{\mathcal{Z}_\ell} p(\mathbf{z};\Theta) = \sum_{j=0}^{J} \pi_j \int_{\mathcal{Z}_\ell} d\mathbf{z}\ p_j(\mathbf{z};\boldsymbol{\theta}_j) . \tag{88}$$

This can be alternatively expressed by defining the "mode-bin probability"

$$\phi_{\ell j}(\boldsymbol{\theta}_j) = \int_{\mathcal{Z}_\ell} d\mathbf{z}\ p_j(\mathbf{z};\boldsymbol{\theta}_j) , \tag{89}$$

such that the overall bin probability is

$$\Phi_\ell(\Theta) = \sum_{j=0}^{J} \pi_j\ \phi_{\ell j}(\boldsymbol{\theta}_j) . \tag{90}$$

The "observed" data are the intensity vector

$$\mathbf{M} = \left\{ m_\ell : \ell = 1, \ldots, L \right\}, \tag{91}$$

and the total number of measurements is

$$K = \sum_{\ell=1}^{L} m_\ell . \tag{92}$$

As discussed in section 3.3, the intensity vector may contain place-holders for expected intensities in the truncated region of a truncated histogram.

4.3.1 EM Algorithm with Missing Assignments and Measurements

Definition of Missing Data and Auxiliary Function. For histogram estimation of mixture-distribution parameters, the missing data include the measurements and mode assignments, which are denoted

$$\mathbf{Z} = \left\{ \mathbf{z}_{\ell k} : \ell = 1, \ldots, L, \;\; k = 1, \ldots, m_\ell \right\},$$
$$\mathbf{J} = \left\{ j_{\ell k} : \ell = 1, \ldots, L, \;\; k = 1, \ldots, m_\ell \right\}.$$

The auxiliary function is then defined as

$$Q_{\mathbf{J},\mathbf{Z}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{M}) = \int_{\mathcal{Z}} d\mathbf{Z} \sum_{\mathcal{J}} p(\mathbf{J}, \mathbf{Z} | \mathbf{M}; \boldsymbol{\Theta}^*) \, \log p(\mathbf{J}, \mathbf{Z}, \mathbf{M}; \boldsymbol{\Theta}), \tag{93}$$

where the marginalization operators for $\mathbf{J}$ and $\mathbf{Z}$ were defined in equations (71) and (11), respectively, and are given for the current case as

$$\int_{\mathcal{Z}} d\mathbf{Z} \equiv \prod_{\ell=1}^{L} \prod_{k=1}^{m_\ell} \left\{ \int_{\mathcal{Z}_\ell} d\mathbf{z}_{\ell k} \right\},$$
$$\sum_{\mathcal{J}} \equiv \prod_{\ell=1}^{L} \prod_{k=1}^{m_\ell} \left\{ \sum_{j_{\ell k}=0}^{J} \right\}.$$

CDLF and Posterior Distribution. The CDLF is obtained using Bayes' rule as

$$\begin{aligned} p(\mathbf{J}, \mathbf{Z}, \mathbf{M}; \boldsymbol{\Theta}) &= p(\mathbf{M}; \boldsymbol{\Theta}) \, p(\mathbf{J}, \mathbf{Z} | \mathbf{M}; \boldsymbol{\Theta}) \\ &= c(\mathbf{M}) \prod_{\ell=1}^{L} \prod_{k=1}^{m_\ell} \pi_{j_{\ell k}} \, p_{j_{\ell k}}(\mathbf{z}_{\ell k}; \boldsymbol{\theta}_{j_{\ell k}}), \end{aligned} \tag{94}$$

whose logarithm is

$$\log p(\mathbf{J}, \mathbf{Z}, \mathbf{M}\,; \boldsymbol{\Theta}) \;=\; \log c(\mathbf{M}) + \sum_{\ell=1}^{L} \sum_{k=1}^{m_\ell} \Big\{ \log \pi_{j\ell k} + \log p_{j\ell k}(\mathbf{z}_{\ell k}\,; \boldsymbol{\theta}_{j\ell k}) \Big\}. \qquad (95)$$

The posterior distribution of the missing data is then obtained by dividing equation (94) by the definition of $p(\mathbf{M}\,; \boldsymbol{\Theta})$, given in equation (6), which yields

$$p(\mathbf{J}, \mathbf{Z} | \mathbf{M}; \boldsymbol{\Theta}^*) \;=\; \prod_{\ell=1}^{L} \prod_{k=1}^{m_\ell} \left\{ \frac{\pi^*_{j\ell k}\, p_{j\ell k}(\mathbf{z}_{\ell k}\,; \boldsymbol{\theta}^*_{j\ell k})}{\Phi_\ell(\boldsymbol{\Theta}^*)} \right\}. \qquad (96)$$

Conditional Expectation. Given equation (96), the conditional expectation operator is a product of operators given by

$$\int_{\mathcal{Z}} d\mathbf{Z} \sum_{\mathcal{J}} p(\mathbf{J}, \mathbf{Z} | \mathbf{M}; \boldsymbol{\Theta}^*) \;=\; \prod_{\ell=1}^{L} \prod_{k=1}^{m_\ell} \left\{ \frac{1}{\Phi_\ell(\boldsymbol{\Theta}^*)} \sum_{j\ell k=0}^{J} \pi^*_{j\ell k} \int_{\mathcal{Z}_\ell} d\mathbf{z}_{\ell k}\; p_{j\ell k}(\mathbf{z}_{\ell k}\,; \boldsymbol{\theta}^*_{j\ell k}) \right\}. \qquad (97)$$

Except when operating on some function containing $\mathbf{z}_{\ell k}$, each term in the large brackets is unity since

$$\Phi_\ell(\boldsymbol{\Theta}^*) \;=\; \sum_{j\ell k=0}^{J} \pi^*_{j\ell k} \int_{\mathcal{Z}_\ell} d\mathbf{z}_{\ell k}\; p_{j\ell k}(\mathbf{z}_{\ell k}\,; \boldsymbol{\theta}^*_{j\ell k}). \qquad (98)$$

Therefore, as in previous cases, when the conditional expectation operator is applied to the CDLF, all components of the expectation marginalize to one except those for which the ℓ and k in the expectation match the ℓ and k in the CDLF. In fact, it would be more correct notationally (but much uglier) to denote the indices in the expectation operation as ℓ' and k' to distinguish them from the ℓ and k that appear in the CDLF, and then introduce a Kronecker delta function for bookkeeping. Whether said in symbols or words, the net result is that all expectation terms marginalize to one except those for which $\ell' = \ell$ and $k' = k$. Noting all of these "unit marginalizations" and dropping the term $\log c(\mathbf{M})$, the auxiliary function reduces to

$$Q_{\mathbf{J},\mathbf{Z}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{M}) \;=\; \sum_{\ell=1}^{L} \sum_{k=1}^{m_\ell} \frac{1}{\Phi_\ell(\boldsymbol{\Theta}^*)} \sum_{j\ell k=0}^{J} \pi^*_{j\ell k} \int_{\mathcal{Z}_\ell} d\mathbf{z}_{\ell k}\; p_{j\ell k}(\mathbf{z}_{\ell k}\,; \boldsymbol{\theta}^*_{j\ell k})$$
$$\times \Big\{ \log \pi_{j\ell k} + \log p_{j\ell k}(\mathbf{z}_{\ell k}\,; \boldsymbol{\theta}_{j\ell k}) \Big\}.$$

At this point, the indices ℓ and k on the missing variables j and $\mathbf{z}$ impart no information since the summation over the assignment variable and the integration over the

measurement covers all possible values, independent of ℓ or k. They are thus dropped to obtain

$$Q_{\mathbf{J},\mathbf{Z}}(\boldsymbol{\Theta};\boldsymbol{\Theta}^*,\mathbf{M}) = \sum_{\ell=1}^{L}\sum_{k=1}^{m_\ell}\frac{1}{\Phi_\ell(\boldsymbol{\Theta}^*)}\sum_{j=0}^{J}\pi_j^*\int_{\mathcal{Z}_\ell} d\mathbf{z}\; p_j(\mathbf{z}\,;\boldsymbol{\theta}_j^*)\Big\{\log\pi_j + \log p_j(\mathbf{z}\,;\boldsymbol{\theta}_j)\Big\}.$$

Without the dependence on ℓ and k, the summation over j can be moved forward, and the sum over k becomes a constant multiplier m_ℓ, allowing the auxiliary function to be expressed as

$$Q_{\mathbf{J},\mathbf{Z}}(\boldsymbol{\Theta};\boldsymbol{\Theta}^*,\mathbf{M}) = \sum_{j=0}^{J} Q_{\mathbf{J},\mathbf{Z}}(\pi_j;\boldsymbol{\Theta}^*,\mathbf{M}) + \sum_{j=0}^{J} Q_{\mathbf{J},\mathbf{Z}}(\boldsymbol{\theta}_j;\boldsymbol{\Theta}^*,\mathbf{M}), \tag{99}$$

where the components in this expression are defined as

$$Q_{\mathbf{J},\mathbf{Z}}(\pi_j;\boldsymbol{\Theta}^*,\mathbf{M}) = \pi_j^*\sum_{\ell=1}^{L}\frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)}\int_{\mathcal{Z}_\ell} d\mathbf{z}\; p_j(\mathbf{z}\,;\boldsymbol{\theta}_j^*)\;\log\pi_j\,, \tag{100}$$

$$Q_{\mathbf{J},\mathbf{Z}}(\boldsymbol{\theta}_j;\boldsymbol{\Theta}^*,\mathbf{M}) = \pi_j^*\sum_{\ell=1}^{L}\frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)}\int_{\mathcal{Z}_\ell} d\mathbf{z}\; p_j(\mathbf{z}\,;\boldsymbol{\theta}_j^*)\;\log p_j(\mathbf{z};\boldsymbol{\theta}_j)\,. \tag{101}$$

With the exception of the factor π_j^*, the form of this last component is identical to the auxiliary function $Q_{\mathbf{Z}}(\boldsymbol{\theta}_j;\boldsymbol{\Theta}^*,\mathbf{M})$ for the non-mixture case. During the M-step, when equation (99) is differentiated with respect to $\boldsymbol{\theta}_j$ to find a critical point, the derivative is π_j^* times the derivative of $Q_{\mathbf{Z}}(\boldsymbol{\theta}_j;\boldsymbol{\Theta}^*,\mathbf{M})$. When equated to zero, the π_j^* term drops out and optimization of $Q_{\mathbf{J},\mathbf{Z}}(\pi_j;\boldsymbol{\Theta}^*,\mathbf{M})$ is achieved by independently optimizing $Q_{\mathbf{Z}}(\boldsymbol{\theta}_j;\boldsymbol{\Theta}^*,\mathbf{M})$ for each $\boldsymbol{\theta}_j$, making all of the earlier (non-mixture) results applicable. That said, the discussion below for Gaussian densities includes the factor π_j^* in order for the effective number of measurements to fall out as an intermediate variable. In effect, this just means that the expressions for $\hat{\boldsymbol{\mu}}_j$ and $\hat{\mathbf{P}}_j$ include a term $\pi_j^*/\pi_j^* = 1$.

Estimation of Assignment Probabilities. Optimization of the assignment probabilities is performed using equation (100) directly. A convenient form for that expression is obtained by noting the definition of $\Phi_\ell(\boldsymbol{\Theta})$, and defining

$$\gamma_{\ell j}(\boldsymbol{\Theta}^*) = \frac{\pi_j^*\int_{\mathcal{Z}_\ell} d\mathbf{z}\; p_j(\mathbf{z}\,;\boldsymbol{\theta}_j^*)}{\left\{\sum_{j=0}^{J}\pi_j^*\int_{\mathcal{Z}_\ell} d\mathbf{z}\; p_j(\mathbf{z}\,;\boldsymbol{\theta}_j^*)\right\}}. \tag{102}$$

With these so defined, the expression for $Q_{\mathbf{J},\mathbf{Z}}(\pi_j;\boldsymbol{\Theta}^*,\mathbf{M})$ then becomes

$$Q_{\mathbf{J},\mathbf{Z}}(\pi_j;\boldsymbol{\Theta}^*,\mathbf{M}) = \left\{\sum_{\ell=1}^{L} m_\ell\;\gamma_{\ell j}(\boldsymbol{\Theta}^*)\right\}\log\pi_j\,. \tag{103}$$

Aside from the weight m_ℓ, this is identical in form to the case with observed measurements. The EM update for the assignment probability is thus given by

$$\hat{\pi}_j \;=\; \frac{1}{K} \sum_{\ell=1}^{L} m_\ell \; \gamma_{\ell j}(\boldsymbol{\Theta}^*) \,. \tag{104}$$

As was the case for observed measurements, this expression is independent of the form of the modal distributions.

Gaussian Mode Estimation. When the jth mode is Gaussian, the corresponding component of equation (101) becomes

$$Q_{\mathbf{J},\mathbf{z}}(\boldsymbol{\theta}_j; \boldsymbol{\theta}_j^*, \mathbf{M}) \;=\; \pi_j^* \sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)} \int_{\mathcal{Z}_\ell} d\mathbf{z} \; \mathcal{N}\big(\mathbf{z}; \boldsymbol{\mu}_j^*, \mathbf{P}_j^*\big) \; \log \mathcal{N}\big(\mathbf{z}; \boldsymbol{\mu}_j, \mathbf{P}_j\big) \;. \tag{105}$$

Substituting the Gaussian density in the log-CDLF term of the auxiliary function gives

$$\begin{aligned} Q_{\mathbf{z}}(\boldsymbol{\theta}_j \,; \boldsymbol{\Theta}^*, \mathbf{M}) \;=\; & \frac{\pi_j^*}{2} \log \Big| \mathbf{P}_j^{-1} \Big| \sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)} \int_{\mathcal{Z}_\ell} d\mathbf{z} \; \mathcal{N}\big(\mathbf{z}; \boldsymbol{\mu}_j^*, \mathbf{P}_j^*\big) \\ & - \frac{\pi_j^*}{2} \sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)} \int_{\mathcal{Z}_\ell} d\mathbf{z} \; \mathcal{N}\big(\mathbf{z}; \boldsymbol{\mu}_j^*, \mathbf{P}_j^*\big) \; \big(\mathbf{z} - \boldsymbol{\mu}_j\big)^{\mathrm{T}} \mathbf{P}_j^{-1} \big(\mathbf{z} - \boldsymbol{\mu}_j\big) \,. \end{aligned} \tag{106}$$

To parallel the non-mixture case, it is convenient to define the weighted "mode-specific" local moments

$$\boldsymbol{\omega}_{\ell j}(\boldsymbol{\theta}_j^*) \;=\; \int_{\mathcal{Z}_\ell} d\mathbf{z} \; \mathcal{N}\big(\mathbf{z}; \boldsymbol{\mu}_j^*, \mathbf{P}_j^*\big) \; \mathbf{z} \,, \tag{107}$$

$$\boldsymbol{\Omega}_{\ell j}(\boldsymbol{\theta}_j^*) \;=\; \int_{\mathcal{Z}_\ell} d\mathbf{z} \; \mathcal{N}\big(\mathbf{z}; \boldsymbol{\mu}_j^*, \mathbf{P}_j^*\big) \; \mathbf{z}\,\mathbf{z}^{\mathrm{T}} \,. \tag{108}$$

Also, the effective number of measurements corresponding to the jth mode is

$$\kappa_j(\boldsymbol{\Theta}^*) \;=\; \pi_j^* \sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)} \; \phi_{\ell j}(\boldsymbol{\theta}_j^*) \,. \tag{109}$$

While the meaning of this variable is the same as in section 4, the definition is altered relative to equation (84) to accommodate the missing measurements.

Again drawing on the results of appendix D, the estimators for the mean and covariance of the jth (Gaussian) component are

$$\hat{\boldsymbol{\mu}}_j \;=\; \frac{1}{\kappa_j(\boldsymbol{\Theta}^*)} \sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)} \; \pi_j^* \, \boldsymbol{\omega}_{\ell j}(\boldsymbol{\theta}_j^*) \,, \tag{110}$$

$$\hat{\mathbf{P}}_j \;=\; \frac{1}{\kappa_j(\boldsymbol{\Theta}^*)} \sum_{\ell=1}^{L} \frac{m_\ell}{\Phi_\ell(\boldsymbol{\Theta}^*)} \; \pi_j^* \, \tilde{\boldsymbol{\Omega}}_{\ell j}(\boldsymbol{\theta}_j^*, \hat{\boldsymbol{\mu}}_j) \,, \tag{111}$$

where $\tilde{\Omega}_{\ell j}(\theta_j^*, \hat{\mu}_j) = \int_{\mathcal{Z}_\ell} d\mathbf{z}\, \mathcal{N}(\mathbf{z}; \mu_j^*, \mathbf{P}_j^*)\ (\mathbf{z} - \hat{\mu}_j)(\mathbf{z} - \hat{\mu}_j)^{\mathrm{T}}$ is the mode-specific center-shifted second local moment, computed as

$$\tilde{\Omega}_{\ell j}(\theta_j^*, \hat{\mu}_j) \;=\; \Omega_{\ell j}(\theta_j^*) - 2\,\hat{\mu}_j\ \omega_{\ell j}(\theta_j^*) + \hat{\mu}_j^2\ \phi_{\ell j}(\theta_j^*)\,. \tag{112}$$

Estimating a Scalar Gaussian Signal in Uniform Noise. If the measurement space is scalar, the signal distribution contains a single Gaussian component, and the noise is considered to be uniform over the observable region of measurement space, then the model distribution is the two-component mixture given by

$$p(\mathbf{z}; \pi_s, \mu, \sigma) \;=\; \pi_s \mathcal{N}(z; \mu, \sigma^2) + (1 - \pi_s)\left(\frac{1}{w}\right)\,, \tag{113}$$

where $w = \max(\mathcal{Z}_{\mathrm{O}}) - \min(\mathcal{Z}_{\mathrm{O}})$ is the width of the observable region and of the uniform noise distribution. Note that the above expression has been parameterized in terms of the mixing weight π_s for the signal component, as opposed to parameterizing in terms of $\pi_0 = 1 - \pi_s$ as was done earlier in the section to accommodate mixture signal distributions.

One nice property of the model in equation (113) is that the noise component contains no unknown parameters. The unconditional bin probabilities for the noise component can therefore be computed once at the outset of the estimation algorithm; they need not be updated in each EM iteration since nothing changes. Figures 5 and 6 show EM iteration results for the Gaussian-signal-in-uniform-noise model under two scenarios. In generating both figures, the model was used to synthesize a number of measurements, these measurements were binned into a set of histogram cells, and the histogram intensities were used to estimate the model parameters. For figure 5, a very large number of synthetic measurements was generated and binned such that the histogram has a large overall intensity and the individual intensities are quite accurate. For figure 6, a much smaller number of synthetic measurements was generated such that the histogram has a small overall intensity and the individual intensities are noisy. As one might expect, the final EM estimates computed from the noisy histogram data are degraded from those computed from the accurate histogram intensities. They are still quite close to the true values, however. Furthermore, as figure 6 suggests, the estimate of the mean (i.e., the location parameter) is far superior to what would be obtained if the location of the largest histogram count were chosen (i.e., peak picking).

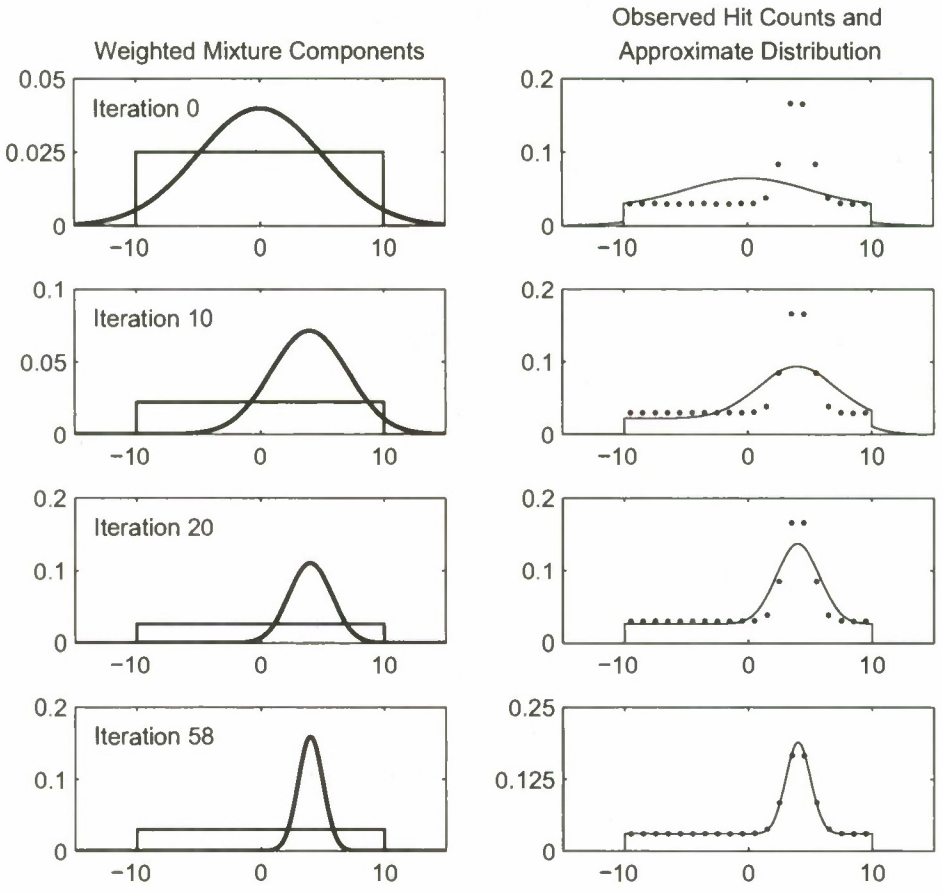

Figure 5: Gaussian-Uniform Mixture Estimation, High-Intensity Data *Histogram-based parameter estimation for a two-component mixture distribution with Gaussian and uniform components. Histogram data were obtain by binning* $K = 10^6$ *measurements that were generated from the ideal mixture distribution with parameters* $\mu = 4$, $\sigma = 1$, *and* $\pi_s = 0.4$. *The large number of measurements gives "clean" histogram data, and the estimated parameters are extremely accurate at* $\mu = 3.99937$, $\sigma = 1.00260$, *and* $\pi_s = 0.39925$. *The first and last rows correspond to the initial and final parameter estimates, respectively. The second and third rows correspond to intermediate iterations of the EM algorithm.*

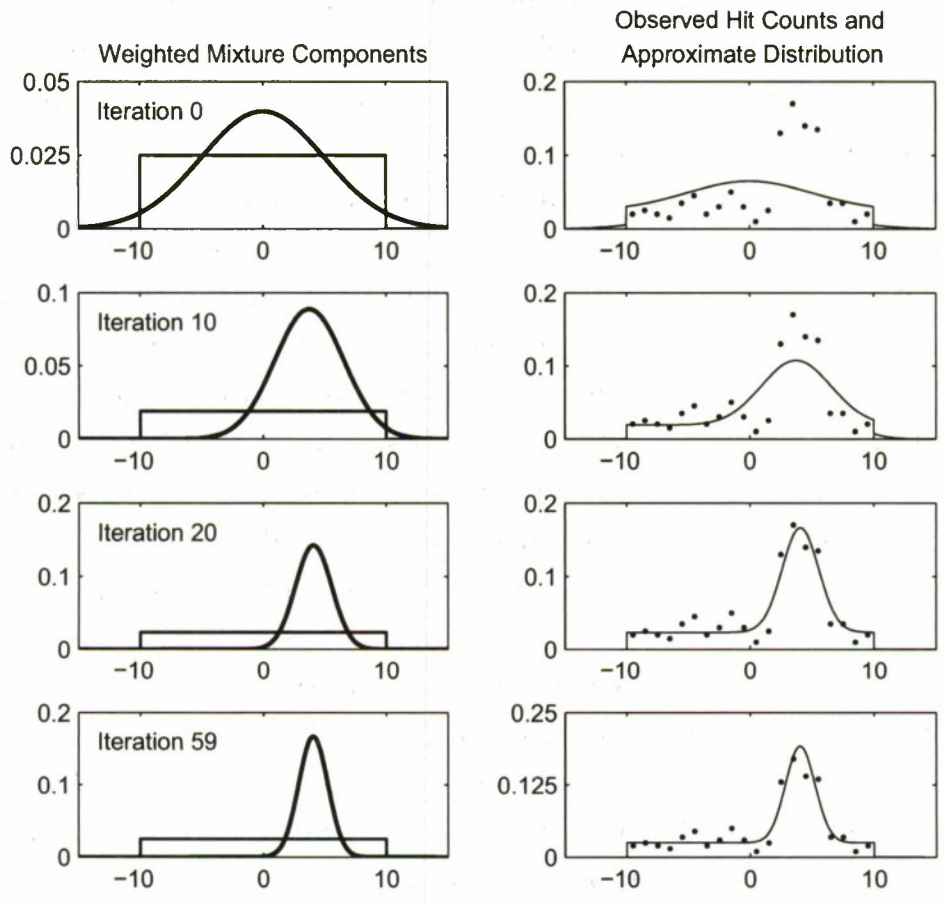

Figure 6: Gaussian-Uniform Mixture Estimation, Low-Intensity Data *Histogram estimation of parameters in the two-component mixture distribution described in the caption of Figure 5. In this case, however, parameters are estimated from "noisy" histogram data with* $K = 200$. *The final parameter estimates are* $\mu = 4.03696$, $\sigma = 1.18893$, *and* $\pi_s = 0.49568$.

5. SUMMARY

This report has examined histogram estimation techniques for inferring the statistical properties of physical variables from observed intensity data. The motivation for using these estimation algorithms is to reduce bias and other artifacts that can occur when traditional peak-picking or simple interpolation algorithms are applied to intensity data, particularly when the data exhibit significant spreading in the variable of interest. The report opened with a high-level consideration of the theory, highlighting some of the primary characteristics, issues, capabilities, and limitations of the theory and algorithms. Great emphasis was placed on the fact that histogram techniques are based on discrete stochastic theory (for integer-valued intensity data) and the issues that arise when the theory and techniques are applied to real-valued acoustic energy data. As it turns out, this does not present a significant difficulty except in Bayesian contexts where the maximum likelihood (ML) estimate must be weighted with a prior distribution. The relative weighting of the prior and ML distributions remains a significant open issue when using histogram methods with real-valued energy data.

After the introduction and consideration of issues involved with application of the methods, the histogram estimation algorithms were developed from first statistical principles by drawing on significant background material provided in a set of appendixes. The algorithms were "built up" by considering the simplest cases first and then adding complexity. The in-depth tutorial examination given here provides the detailed theoretical developments that were omitted from earlier works such as McLachlan and Jones [2]. This report also limited the discussion to static distributions, which are subtle and interesting enough to warrant consideration in and of themselves. Limiting the discussion to the static case also avoids the significant unresolved issues and notational baggage that is incurred when histogram techniques are considered in the context of dynamic tracking algorithms (e.g., see [3] and [5]). The overall goal of this report was to arm the reader with adequate background and insights to apply and extend these powerful and versatile methods for a variety of applications.

APPENDIX A
EXPECTATION-MAXIMIZATION ALGORITHMS

The goal of this appendix is to introduce the basic ideas, terminology, and notation for the expectation-maximization (EM) approach to algorithm design, which is the common thread through all of the algorithms discussed in this report. The EM approach provides a general template for generating iterative maximum-likelihood algorithms that estimate the parameter vector $\boldsymbol{\Theta}$ in a probability density function (PDF) $p(\mathbf{x};\boldsymbol{\Theta})$, given observations of the variable random variable $\mathbf{x}$. The EM approach is most useful when it is difficult to directly maximize $p(\mathbf{x};\boldsymbol{\Theta})$ with respect to $\boldsymbol{\Theta}$, but there exists an auxiliary variable $\boldsymbol{\xi}$, such that maximizing the joint PDF $p(\mathbf{x},\boldsymbol{\xi};\boldsymbol{\Theta})$ is straightforward. The approach replaces a difficult nonlinear optimization problem with an iterative sequence of easier problems. The EM method is extensively used in modern statistical analysis because it greatly simplifies algorithm development for many problems, it has guaranteed convergence under some fairly general conditions, and many statistical models have obvious choices for missing data. The general properties of the EM algorithm are discussed in [1], and numerous applications and extensions are discussed in the text by McLachlan and Krishnan [10].

Observed, Missing, and Complete Data. EM algorithms all share the characteristic of having observed, missing, and complete data. The *observed data* consist of samples of data that are at hand, representing either physical measurements from a sensor or features that are computed from physical measurements. The collection of such samples is denoted $\mathbf{X} = \{\mathbf{x}_n : n = 1,\ldots,N\}$. The density function $p(\mathbf{X};\boldsymbol{\Theta})$ is referred to as is the *observed-data likelihood function* (ODLF).

The *missing data* contain the information that, were it known in addition to the observed data, causes the difficult estimation problem to reduce to a simpler one. For a given set of observed data $\mathbf{X}$, the corresponding collection of missing data is denoted $\boldsymbol{\Xi} = \{\boldsymbol{\xi}_m : m = 1,\ldots,M\}$. The concatenated set containing both the observed and missing data is referred to as the *complete data*, and the joint distribution $p(\mathbf{X},\boldsymbol{\Xi};\boldsymbol{\Theta})$ is referred to as the *complete-data likelihood function* (CDLF). As mentioned above, the fundamental premise of EM is that $p(\mathbf{X},\boldsymbol{\Xi};\boldsymbol{\Theta})$ is much easier to optimize with respect to $\boldsymbol{\Theta}$ than is $p(\mathbf{X};\boldsymbol{\Theta})$.

Iterative Structure and Auxiliary Function. The EM algorithm is iterative. It requires an initial estimate for $\boldsymbol{\Theta}$, which is obtained using an application-dependent suboptimal algorithm (the preferred method), or by random selection within some reasonable range of values. Denoting the initial estimate by $\boldsymbol{\Theta}^0$, the EM algorithm generates

the sequence of estimates defined by

$$\boldsymbol{\Theta}^{i+1} = \arg\max_{\boldsymbol{\Theta}} Q_{\boldsymbol{\Xi}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^i, \mathbf{X}), \tag{114}$$

where $Q_{\boldsymbol{\Xi}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^i, \mathbf{X})$ is the so-called *auxiliary function*, which will be formally defined in a moment. While the literature typically denotes the auxiliary function simply as $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*)$, this report discusses a number of different auxiliary functions with different sets of observed and missing data. To more easily distinguish these various cases, the notation used here includes a subscript to denote the missing data and an additional argument to denote the observed data; thus the subscript $\boldsymbol{\Xi}$ and argument $\mathbf{X}$ in $Q_{\boldsymbol{\Xi}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{X})$.

When iterating equation (114) to optimize the ODLF, the iterations are terminated either at a pre-selected number of iterations or when some set of convergence criteria is satisfied. When convergence tests are used, they are typically based on the relative change in the observed-data likelihood and/or values of the estimates, similar to convergence tests for standard nonlinear optimization techniques like the Newton and gradient ascent algorithms.

Equation (114) includes the iteration number as an index. When using the EM approach, there are often a large number of other indices that track a number of other characteristics, making index variables a rare notational commodity. It is therefore convenient to define the parameter estimates as $\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{i+1}$ and $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}^i$, such that the "typical" EM iteration is

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} Q_{\boldsymbol{\Xi}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{X}). \tag{115}$$

The auxiliary function $Q_{\boldsymbol{\Xi}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{X})$ is defined as the expectation of the log of the CDLF, conditioned on the posterior distribution of the missing data, which is stated mathematically as

$$\begin{aligned} Q_{\boldsymbol{\Xi}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*, \mathbf{X}) &= E_{\boldsymbol{\Xi}|\mathbf{X};\boldsymbol{\Theta}^*}\Big\{ \log p(\mathbf{X}, \boldsymbol{\Xi}; \boldsymbol{\Theta}) \Big\} \\ &= \int d\boldsymbol{\Xi}\; p(\boldsymbol{\Xi}|\mathbf{X}, \boldsymbol{\Theta}^*) \log p(\mathbf{X}, \boldsymbol{\Xi}; \boldsymbol{\Theta}). \end{aligned} \tag{116}$$

Here the integral notation $\int d\boldsymbol{\Xi}$ is used to indicate marginalization over the missing data $\boldsymbol{\Xi}$; in general, this is a sequence of continuous integrations, discrete summations, or both. Since missing data with both continuous and discrete variables are common, sequences that mix integrations and summations are also common.

The basic idea behind EM is well illustrated from the viewpoint of *iterative minorization* (IM), which is an even more general class of algorithms to which the EM

method belongs. Figure A-1 shows two iterations of a generic IM algorithm. As discussed in the figure caption, each iteration maximizes an auxiliary function that "minorizes" the function being maximized, which for the EM algorithm is the ODLF. That the EM auxiliary function minorizes the ODLF follows from the nature of missing data, in the sense that including a stochastic nuisance parameter in a model always drives down the likelihood relative to the nuisance-free model because of the added uncertainty associated with the nuisance parameter.

Expectation and Maximization Steps. Each EM iteration consists of an expectation step (E-step) and a maximization step (M-step). The E-step involves evaluating equation (116), and the M-step involves computing estimates that maximize equation (116) with respect to $\boldsymbol{\Theta}$. The process for deriving the auxiliary function in the E-step is fairly universal across applications and is largely an exercise in conditional probability. The steps are to (1) define the observed data and the ODLF, (2) define the missing data and the CDLF, (3) determine the posterior distribution of the missing data, and (4) carry out the conditional expectation to obtain the auxiliary function. In many applications, the missing data are chosen such that the conditional distribution $p(\mathbf{X}|\boldsymbol{\Xi};\boldsymbol{\Theta})$ can be written in closed form, and there exists a known unconditional (prior) distribution $p(\boldsymbol{\Xi};\boldsymbol{\Theta})$. In such cases, the CDLF is obtained as

$$p(\mathbf{X},\boldsymbol{\Xi};\boldsymbol{\Theta}) = p(\mathbf{X}|\boldsymbol{\Xi};\boldsymbol{\Theta})\, p(\boldsymbol{\Xi};\boldsymbol{\Theta})\,. \tag{117}$$

Given the ODLF and CDLF, the posterior for the missing data is given by

$$p(\boldsymbol{\Xi}|\mathbf{X};\boldsymbol{\Theta}) = \frac{p(\mathbf{X},\boldsymbol{\Xi};\boldsymbol{\Theta})}{p(\mathbf{X};\boldsymbol{\Theta})}\,. \tag{118}$$

In other applications, it may be more convenient to first derive the posterior distribution $p(\boldsymbol{\Xi}|\mathbf{X};\boldsymbol{\Theta})$ and then obtain the CDLF as

$$p(\mathbf{X},\boldsymbol{\Xi};\boldsymbol{\Theta}) = p(\boldsymbol{\Xi}|\mathbf{X};\boldsymbol{\Theta})\, p(\mathbf{X};\boldsymbol{\Theta})\,. \tag{119}$$

The preferred order for evaluating these distributions depends on the structure of the missing data and the distributions involved, but some form of these steps will be encountered whenever EM is applied in a new situation. The final stage in deriving the auxiliary function involves carrying out the conditional expectation, which is usually simplified by noting independence or conditional independence among the variables. More will be said about this process in the context of particular problems.

While the details of the M-step cannot be specified without formulating a particular parameterization of the CDLF, these details typically follow from standard con-

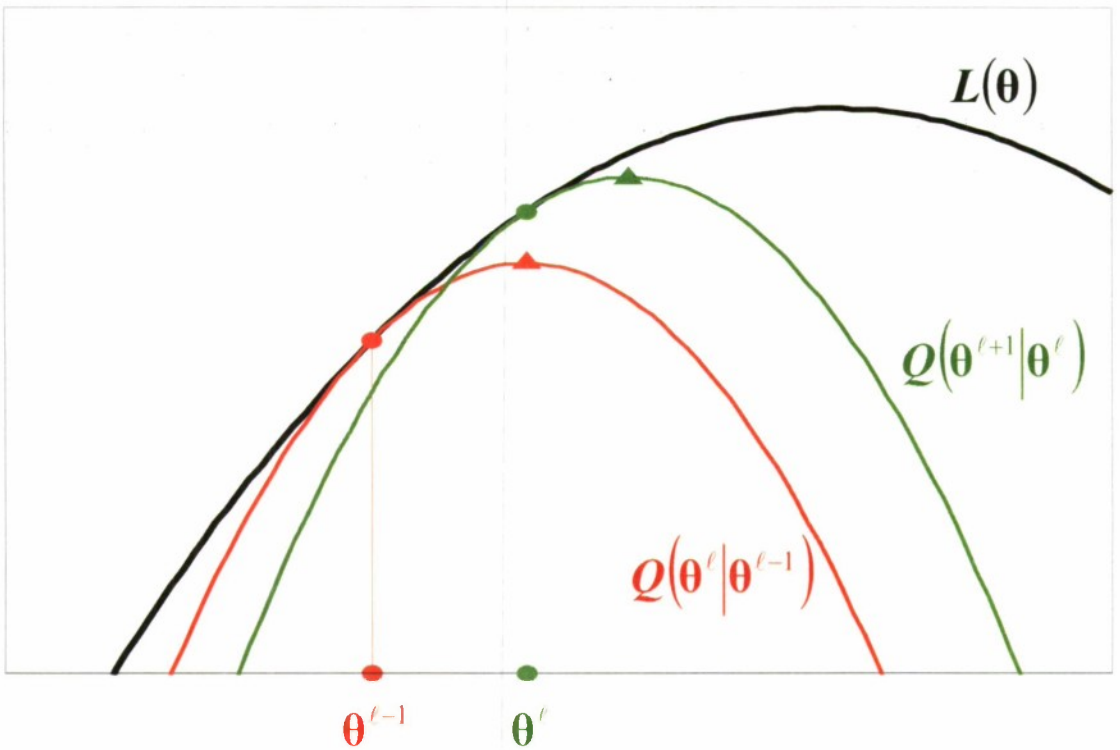

Figure A-1: Iterative Minorization (IM)

Two IM iterations are shown; the first depicted in red and the second in green. The function $L(\theta)$ to be maximized is shown in black. The first iteration maximizes an auxiliary function $Q(\theta^{\ell}|\theta^{\ell-1})$, which "minorizes" the likelihood function at $\theta^{\ell-1}$; that is, $Q(\theta^{\ell}|\theta^{\ell-1})$ is strictly less than $L(\theta)$ at every point θ except $\theta^{\ell-1}$, where the two functions share the same function value and first derivative (i.e., the two functions are tangent at $\theta^{\ell-1}$). Under these minorization constraints, maximization of $Q(\theta^{\ell}|\theta^{\ell-1})$ is guaranteed to increase $L(\theta)$ unless $\theta^{\ell-1}$ is already a stationary point of $L(\theta)$, in which case no changes take place. The second iteration repeats the process with the tangent constraint imposed at the point θ^{ℓ} produced by the first iteration.

strained or unconstrained optimization methods. For example, a stationary point is obtained by equating to zero the derivative of $Q_{\Xi}(\Theta; \Theta^*, \mathbf{X})$ with respect to each element of Θ. The stationary point is a maximum if the second derivative (Hessian) matrix is negative definite (i.e., all of its eigenvalues are strictly negative) at the stationary point. Fortunately, the auxiliary function is *concave* for many problem formulations involving Gaussian or Gaussian-mixture model densities, which means that a stationary point for such an objective function is the unique maximum. Appendix D demonstrates that the auxiliary function for Gaussian histogram estimators is, indeed, concave.

APPENDIX B
COMBINATORIAL PROBABILITY DISTRIBUTIONS

Histogram-based methods are inherently combinatorial, since they consist of looking at all combinations of ways that events can be grouped into a collection of histogram cells. This appendix reviews the binomial, negative binomial, and multinomial distributions to establish notation and to give an easy single reference for these basic distributions.

Binomial Distribution. Perhaps the simplest experimental situation involves a series of Bernoulli trials, whose outcomes are limited to one of two categories, "success" or "failure". A common question with regard to Bernoulli trials concerns the number of successes that might be observed over the course of n independent trials. This situation is described by the binomial distribution. Let the probability of success in any single trial be denoted by ϕ, such that the probability of failure is $(1-\phi)$. To have exactly m successes in n independent trials, there must also be $(n-m)$ failures. The probability of the m successes is ϕ^m; the probability of the $(n-m)$ failures is $(1-\phi)^{(n-m)}$; and there are $b(m,n)$ ways in which this situation can occur in n trials, where $b(m,n)$ is the binomial coefficient:

$$b(m,n) \;=\; \frac{n\,!}{m\,!\,(n-m)\,!}\,. \tag{120}$$

The binomial distribution measures the probability of the event "m successes out of n trials." The distribution is therefore defined as

$$p(m;\phi,n) \;=\; b(m,n)\;\phi^m\;(1-\phi)^{(n-m)}\,, \tag{121}$$

which satisfies the marginalization constraint

$$\sum_{m=0}^{n} p(m;\phi,n) \;=\; 1\,. \tag{122}$$

Negative Binomial Distribution. In the binomial distribution, the number of trials is a parameter, and number of successes is a random variable. Suppose, however, that the question is turned around to ask "How many trials are needed to achieve a certain number of successes?" Here, the number of successes is the parameter, and the number of trials is the random variable. This situation is described by the negative binomial distribution

$$p(n\,;\phi,m) \;=\; b^{-}(n,m)\;\phi^m\;(1-\phi)^{(n-m)}\,, \tag{123}$$

where $b^-(n,m)$ is the *negative binomial coefficient*, defined as

$$b^-(m,n) \;=\; b(n-m-1,n-1) \;=\; \frac{(n-1)!}{(m-1)!\,(n-m)!}\,. \tag{124}$$

The negative binomial distribution satisfies the marginalization constraint

$$\sum_{n=m}^{\infty} p(n;\phi,m) \;=\; 1\,. \tag{125}$$

Multinomial Distribution. Now consider an independent series of trials in which each trial has L categories of possible outcomes, rather than simply success or failure. With no other information, the probability of an outcome occuring in the ℓth category is denoted ϕ_ℓ, and the collection of probabilities for all categories is $\boldsymbol{\phi} = \{\phi_\ell : \ell = 1,\ldots,L\}$. When actually performing a series of trials, an observed outcome in one of the categories is referred to as a "hit" in that category. The number of hits observed in each category over the series of trials is denoted m_ℓ for $\ell = 1,\ldots,L$. These "hit counts" are collected in the vector $\mathbf{M} = \{m_\ell : \ell = 1,\ldots,L\}$, whose statistical properties are governed by the multinomial distribution

$$p(\mathbf{M};\boldsymbol{\phi}) \;=\; c(\mathbf{M}) \prod_{\ell=1}^{L} \phi_\ell^{m_\ell}\,, \tag{126}$$

where $c(\mathbf{M})$ is the *multinomial coefficient*, defined as

$$c(\mathbf{M}) \;=\; \frac{\left(\sum_{\ell=1}^{L} m_\ell\right)!}{\left\{\prod_{\ell=1}^{L} m_\ell!\right\}}\,. \tag{127}$$

Unlike the binomial distribution, the number of trials is not an explicit parameter because the multinomial distribution implicitly assumes that the sum of the hit counts for all categories equals the number of trials (i.e., that all outcomes are counted). If this is not the case in a given application, then the multinomial distribution is not the correct model.

Equation (126) assumes that any outcome not fitting into one of the defined categories has zero probability of occurrence. Alternatively stated, it assumes that, with probability one, all possible outcomes fall within one of the categories, such that

$$\sum_{\ell=1}^{L} \phi_\ell \;=\; 1\,. \tag{128}$$

If, on the other hand, events occur with nonzero probability that do not fit any of the categories, then the statistical model must be modified. In particular, consider the case

of "pre-screened" data, where any trial whose outcome does not fit a given category is ignored entirely; not only is the hit not recorded, but the counter that records the total number of trials is not incremented. The sum of hit counts equals the number of recorded trials, so the general form of the multinomial distribution is still valid. The probabilities for the categories, however, must be normalized to account for the fact that the universe of outcomes has been projected down to the union of the known categories. Defining the probability that a hit occurs in any of the categories as

$$\phi = \sum_{\ell=1}^{L} \phi_\ell \tag{129}$$

allows the multinomial distribution for this case to be defined as

$$p(\mathbf{M}\,;\boldsymbol{\phi}) \;=\; c(\mathbf{M}) \prod_{\ell=1}^{L} \left(\frac{\phi_\ell}{\phi}\right)^{m_\ell} \;=\; c(\mathbf{M})\; \phi^{-m} \prod_{\ell=1}^{L} \phi_\ell^{m_\ell}\,, \tag{130}$$

where m is the total number of trials that pass the screening process; that is,

$$m \;=\; \sum_{\ell=1}^{L} m_\ell\,. \tag{131}$$

Negative Multinomial Distribution. When applying the multinomial distribution to pre-screened data, the distribution is useful for making inferences only about things that occur within the observable categories. It is sometimes desirable to extrapolate inference into categories that cannot be observed. This situation can be thought of in terms of Bernoulli trials by grouping the observable categories into one single *super-category*. The universe of possible outcomes can then be partitioned into two classes, with a measurement either falling within the observed super-category (a "success") or falling outside of this super-category (a "failure"). Treating the sum of observed hit counts as the number of successes in a sequence of Bernoulli trials, the unknown total number of trials required to have generated these successes is a random variable governed by the negative binomial distribution in equation (123), with ϕ defined by equation (129) and m given by equation (131). This analysis can be taken a step further by subdividing the space of unobservable outcomes and making inferences about these unobservable categories. This last situation is described by the negative multinomial distribution, which is discussed further in appendix C.

APPENDIX C
STATISTICS OF MISSING HIT COUNTS

This appendix derives the posterior distribution $p(\mathbf{M}_\mathrm{T}\,|\,\mathbf{M}_\mathrm{O};\boldsymbol{\theta})$ of the missing intensities given the observed histogram counts, and then uses this posterior to obtain the expected intensities in the unobserved histogram cells.

Posterior Distribution. Development of the posterior density involves making a number of structural observations concerning the component distributions involved. This discussion is facilitated by defining the "number of truncated measurements" as the discrete random variable

$$K_\mathrm{T} = \sum_{\ell=L_\mathrm{O}+1}^{L} m_\ell \,. \tag{132}$$

The total number of (observed and unobserved) measurements is then the discrete random variable defined by

$$K_\mathrm{C} = K_\mathrm{O} + K_\mathrm{T} \,. \tag{133}$$

Observing that the regions $\mathcal{Z}_\mathrm{O}$ and $\mathcal{Z}_\mathrm{T}$ are disjoint in $\mathcal{Z}$, the posterior distribution satisfies

$$p(\mathbf{M}_\mathrm{T}\,|\,\mathbf{M}_\mathrm{O};\boldsymbol{\theta}) = p(\mathbf{M}_\mathrm{T}\,|\,K_\mathrm{O}\,;\boldsymbol{\theta}) \,. \tag{134}$$

That is, from the standpoint of $\mathcal{Z}_\mathrm{T}$, it does not matter how the measurements in $\mathcal{Z}_\mathrm{O}$ are distributed within $\mathcal{Z}_\mathrm{O}$; only the total number of hits in $\mathcal{Z}_\mathrm{O}$ is important since that provides evidence for inferring the total number of truncated measurements. The second structural observation concerns the appearance of these hit-count totals in the distribution functions. For example, note that in the presence of $\mathbf{M}_\mathrm{T}$, the variable K_T carries no additional statistical information, such that

$$p(\mathbf{M}_\mathrm{T}\,|\,K_\mathrm{O}\,;\boldsymbol{\theta}) = p(\mathbf{M}_\mathrm{T}, K_\mathrm{T}\,|\,K_\mathrm{O}\,;\boldsymbol{\theta}) \;\delta\left(K_\mathrm{T} - \sum_{\ell=L_\mathrm{O}+1}^{L} m_\ell\right) ,$$

where $\delta(\cdot)$ is the Kronecker delta function, whose value is one for an argument of zero and zero otherwise. The delta function is introduced to ensure that the distribution has nonzero probability *only* when equation (132) is satisfied. Having said this, this delta function will be dropped to ease the notational burden, with the understanding that equation (132) is indeed satisfied. A similar observation can be made concerning K_C

and K_T. That is, given the observed hit total K_O, the variables K_C and K_T carry the same information, such that

$$p(K_T | K_O ; \boldsymbol{\theta}) = p(K_C | K_O ; \boldsymbol{\theta}) \, \delta(K_C - K_O - K_T) . \tag{135}$$

As before, the Kronecker delta is dropped, with the understanding that equation (133) is satisfied. Given these observations, the posterior distribution for the missing data is given by

$$\begin{aligned} p(\mathbf{M}_T | K_O ; \boldsymbol{\theta}) &= p(\mathbf{M}_T, K_T | K_O ; \boldsymbol{\theta}) \\ &= p(\mathbf{M}_T | K_T, K_O ; \boldsymbol{\theta}) \, p(K_T | K_O ; \boldsymbol{\theta}) \\ &= p(\mathbf{M}_T | K_T ; \boldsymbol{\theta}) \, p(K_C | K_O ; \boldsymbol{\theta}) , \end{aligned} \tag{136}$$

where the first term in the last expression follows because once K_T is given, K_O does not bring any additional information concerning $\mathbf{M}_T$. The variable $\mathbf{M}_T$ merely splits the K_T measurements among the L_T bins in $\mathcal{Z}_T$, independent of what happens in $\mathcal{Z}_O$. The conditional distribution $p(\mathbf{M}_T | K_T ; \boldsymbol{\theta})$ thus corresponds to K_T independent draws from L_T categories, and is given by the multinomial distribution

$$p(\mathbf{M}_T | K_T ; \boldsymbol{\theta}) = c(\mathbf{M}_T) \left\{ \phi_T(\boldsymbol{\theta}) \right\}^{-K_T} \prod_{\ell=L_O+1}^{L} \left\{ \phi_\ell(\boldsymbol{\theta}) \right\}^{m_\ell} . \tag{137}$$

The distribution $p(K_C | K_O ; \boldsymbol{\theta})$ pertains to the total number of Bernoulli trials K_C that must be executed in order to achieve K_O successes, where a "success" is a hit anywhere in $\mathcal{Z}_O$. The total number of hits is therefore governed by the negative binomial distribution in equation (123). Noting the relationships between K_O, K_T, and K_C, and those between $\phi_O(\boldsymbol{\theta})$ and $\phi_T(\boldsymbol{\theta})$, the negative binomial distribution for this situation is expressed as

$$p(K_C | K_O ; \boldsymbol{\theta}) = b^-(K_C, K_O) \left\{ \phi_O(\boldsymbol{\theta}) \right\}^{K_O} \left\{ \phi_T(\boldsymbol{\theta}) \right\}^{K_T} , \tag{138}$$

where $b^-(K_C, K_O)$ is the negative binomial coefficient defined in equation (124). Substituting equation (137) and equation (138) into equation (136) then gives the posterior distribution of the missing intensities as a *negative multinomial distribution*, which is defined for the present situation as

$$\boxed{p(\mathbf{M}_T | K_O ; \boldsymbol{\theta}) = c^-(\mathbf{M}_T, K_O) \left\{ \phi_O(\boldsymbol{\theta}) \right\}^{K_O} \prod_{\ell=L_O+1}^{L} \left\{ \phi_\ell(\boldsymbol{\theta}) \right\}^{m_\ell} ,} \tag{139}$$

where $c^-(\mathbf{M}_\mathrm{T}, K_\mathrm{O})$ is the *negative multinomial coefficient*, defined as

$$c^-(\mathbf{M}_\mathrm{T}, K_\mathrm{O}) \;=\; c(\mathbf{M}_\mathrm{T})\; b^-(K_\mathrm{C}, K_\mathrm{O}) \;=\; \frac{(K_\mathrm{C}-1)!}{(K_\mathrm{O}-1)!\,\left\{\prod_{\ell=L_\mathrm{O}+1}^{L} m_\ell!\right\}}\,. \tag{140}$$

Expected Values. The expected intensities are obtained by computing the first moment of the posterior distribution. Dropping the explicit dependence on the parameter vector $\boldsymbol{\theta}$ and expressing the negative multinomial coefficient directly in terms of the missing intensities instead of total hit counts allows the posterior to be written as

$$p(\mathbf{M}_\mathrm{T}\,|\,K_\mathrm{O}) \;=\; \frac{\left(K_\mathrm{O}-1+\sum_{\ell=L_\mathrm{O}+1}^{L} m_\ell\right)!}{(K_\mathrm{O}-1)!\,\left\{\prod_{\ell=L_\mathrm{O}+1}^{L} m_\ell!\right\}}\;\phi_\mathrm{O}^{K_\mathrm{O}}\prod_{\ell=L_\mathrm{O}+1}^{L}\phi_\ell^{m_\ell}\,. \tag{141}$$

The expected value of the (typical) missing hit count, m_ζ is defined for $L_\mathrm{O} < \zeta \le L$ as

$$E\Big\{\, m_\zeta|\,\mathbf{M}_\mathrm{O}\,\Big\} \;=\; \sum_{\mathcal{M}_\mathrm{T}} m_\zeta\; p(\mathbf{M}_\mathrm{T}|\,K_\mathrm{O})\,, \tag{142}$$

where the marginalization operator was defined in equation (39), which is repeated here for convenience as

$$\sum_{\mathcal{M}_\mathrm{T}} \;\equiv\; \prod_{\ell=L_\mathrm{O}+1}^{L}\left\{\sum_{m_\ell=0}^{\infty}\right\}. \tag{143}$$

The key piece of information for taking the expectation is the normalization property of the negative multinomial distribution, that is,

$$\sum_{\mathcal{M}_\mathrm{T}} p(\mathbf{M}_\mathrm{T}|\,K_\mathrm{O}) \;=\; 1\,. \tag{144}$$

Forming and simplifying the product $m_\zeta\, p(\mathbf{M}_\mathrm{T}|\,K_\mathrm{O})$ is greatly facilitated by defining the variable

$$\bar{m}_\ell \;=\; \begin{cases} m_\ell & \text{if } \ell \neq \zeta \\ m_\ell - 1 & \text{if } \ell = \zeta\,. \end{cases} \tag{145}$$

With this variable so defined, the product $m_\zeta\, p(\mathbf{M}_\mathrm{T}|\,K_\mathrm{O})$ can be written as

$$m_\zeta\, p(\mathbf{M}_\mathrm{T}|\,K_\mathrm{O}) \;=\; \frac{\left(K_\mathrm{O}-1+\sum_{\ell=L_\mathrm{O}+1}^{L} m_\ell\right)!}{(K_\mathrm{O}-1)!\,\left\{\prod_{\ell=L_\mathrm{O}+1}^{L} \bar{m}_\ell!\right\}}\;\phi_\mathrm{O}^{K_\mathrm{O}}\prod_{\ell=L_\mathrm{O}+1}^{L}\phi_\ell^{m_\ell}\,. \tag{146}$$

The desire is to obtain an expression that is equal to within a scale factor of the negative multinomial distribution with the variable $\bar{m}_\ell$ instead of m_ℓ. If such an expression

can be obtained, then the negative multinomial portion marginalizes to unity and the desired expectation is the scale factor. To get the numerator of the negative multinomial coefficient in terms of $\bar{m}_\ell$ while retaining the desired structure, it is required to compensate by introducing the variable

$$\bar{K}_{\mathrm{O}} = K_{\mathrm{O}} + 1 \tag{147}$$

such that the numerator is given by

$$\left(K_{\mathrm{O}} - 1 + \sum_{\ell=L_{\mathrm{O}}+1}^{L} m_\ell \right) ! = \left(\bar{K}_{\mathrm{O}} - 1 + \sum_{\ell=L_{\mathrm{O}}+1}^{L} \bar{m}_\ell \right) ! . \tag{148}$$

Now, the marginalization property of the negative multinomial distribution is valid regardless of the actual value of K_{O}, so long as the structure is retained and the same value appears everywhere. The new observed count variable is thus substituted and compensated everywhere it appears in equation (146) to obtain

$$m_\zeta \, p(\mathbf{M}_{\mathrm{T}} | \, K_{\mathrm{O}}) = \frac{\left(\bar{K}_{\mathrm{O}} - 1 + \sum_{\ell=L_{\mathrm{O}}+1}^{L} \bar{m}_\ell \right) !}{\left\{ \frac{(\bar{K}_{\mathrm{O}}-1)!}{K_{\mathrm{O}}} \right\} \left\{ \prod_{\ell=L_{\mathrm{O}}+1}^{L} \bar{m}_\ell ! \right\}} \, \frac{\phi_{\mathrm{O}}^{\bar{K}_{\mathrm{O}}}}{\phi_{\mathrm{O}}} \prod_{\ell=L_{\mathrm{O}}+1}^{L} \phi_\ell^{m_\ell} . \tag{149}$$

Substituting and compensating the variable $\bar{m}_\ell$ in the final product term and rearranging then yields

$$m_\zeta \, p(\mathbf{M}_{\mathrm{T}} | \, K_{\mathrm{O}}) = \frac{K_{\mathrm{O}}}{\phi_{\mathrm{O}}} \phi_\zeta \left\{ \frac{\left(\bar{K}_{\mathrm{O}} - 1 + \sum_{\ell=L_{\mathrm{O}}+1}^{L} \bar{m}_\ell \right) !}{(\bar{K}_{\mathrm{O}} - 1) ! \left\{ \prod_{\ell=L_{\mathrm{O}}+1}^{L} \bar{m}_\ell ! \right\}} \, \phi_{\mathrm{O}}^{\bar{K}_{\mathrm{O}}} \prod_{\ell=L_{\mathrm{O}}+1}^{L} \phi_\ell^{\bar{m}_\ell} \right\} . \tag{150}$$

Since the term in brackets is exactly the negative multinomial distribution in the variables $\bar{m}_\ell$ and $\bar{K}_{\mathrm{O}}$, and the scale factor is independent of any of the truncated m_ℓ, the bracketed term marginalizes under the expectation, leaving the desired expression

$$\boxed{E\Big\{ m_\zeta | \mathbf{M}_{\mathrm{O}} \Big\} = \left\{ \frac{K_{\mathrm{O}}}{\phi_{\mathrm{O}}} \right\} \phi_\zeta .} \tag{151}$$

APPENDIX D
ESTIMATING GAUSSIAN PARAMETERS

This appendix derives estimates and proves the maximality of those estimates for parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ in the objective function

$$Q(\boldsymbol{\theta}) = \sum_{\ell=1}^{L} \rho_\ell \int_{\mathcal{Z}_\ell} d\mathbf{z}\, \mathcal{N}(\mathbf{z};\, \boldsymbol{\zeta}, \mathbf{S})\ \log \mathcal{N}(\mathbf{z};\, \boldsymbol{\mu}, \boldsymbol{\Gamma})\,, \tag{152}$$

where ρ_ℓ is a non-negative weighting function, $\mathcal{N}(\mathbf{z};\boldsymbol{\zeta},\mathbf{S})$ is the *observed measurement density*

$$\mathcal{N}(\mathbf{z};\boldsymbol{\zeta},\mathbf{S}) = \left|2\pi\,\mathbf{S}\right|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\zeta})^{\mathrm{T}}\,\mathbf{S}^{-1}\,(\mathbf{z}-\boldsymbol{\zeta})\right\}, \tag{153}$$

and $\log\mathcal{N}(\mathbf{z};\,\boldsymbol{\mu},\boldsymbol{\Gamma})$ is the model likelihood function (i.e., the log of the model density function). Omitting the constant $\log(2\pi)$ term, which does not affect the parameter estimation problem, the model likelihood is

$$\log\mathcal{N}(\mathbf{z};\,\boldsymbol{\mu},\boldsymbol{\Gamma}) = \frac{1}{2}\log\left|\boldsymbol{\Gamma}\right| - \frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\,\boldsymbol{\Gamma}\,(\mathbf{z}-\boldsymbol{\mu})\,, \tag{154}$$

which is parameterized here in terms of the inverse covariance matrix (or *information matrix*) $\boldsymbol{\Gamma} = \mathbf{P}^{-1}$ to ease the math. The parameter estimates that maximize the objective function are identified below. The integrations are carried out first to obtain an expression that is cast in terms of a set of *effective measurements*. The resulting expression is then maximized by finding the stationary point of the function and by proving that the function is concave, such that the stationary point is the unique maximum.

Effective Measurements. Given the form of the model likelihood function, it is convenient to similarly decompose the objective function into determinant and quadratic-form components, giving

$$Q(\boldsymbol{\theta}) = \eta_1(\boldsymbol{\theta}) - \eta_2(\boldsymbol{\theta})\,, \tag{155}$$

where

$$\eta_1(\boldsymbol{\theta}) = \frac{1}{2}\sum_{\ell=1}^{L} \rho_\ell \int_{\mathcal{Z}_\ell} d\mathbf{z}\, \mathcal{N}(\mathbf{z};\,\boldsymbol{\zeta},\mathbf{S})\ \log\left|\boldsymbol{\Gamma}\right|, \tag{156}$$

$$\eta_2(\boldsymbol{\theta}) = \frac{1}{2}\sum_{\ell=1}^{L} \rho_\ell \int_{\mathcal{Z}_\ell} d\mathbf{z}\, \mathcal{N}(\mathbf{z};\,\boldsymbol{\zeta},\mathbf{S})\ (\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\,\boldsymbol{\Gamma}\,(\mathbf{z}-\boldsymbol{\mu})\,. \tag{157}$$

The determinant component is rewritten by defining the bin probability value

$$\boxed{\phi_\ell = \int_{\mathcal{Z}_\ell} d\mathbf{z}\, \mathcal{N}(\mathbf{z};\, \boldsymbol{\zeta}, \mathbf{S})} \tag{158}$$

and the effective number of measurements

$$\boxed{\kappa = \sum_{\ell=1}^{L} \rho_\ell\, \phi_\ell} \tag{159}$$

to obtain

$$\eta_1(\boldsymbol{\theta}) = \frac{\kappa}{2} \log \left| \boldsymbol{\Gamma} \right| . \tag{160}$$

Equation (157) is rewritten by expressing the quadratic form as the trace of an outer product, which gives

$$\begin{aligned}
\eta_2(\boldsymbol{\theta}) &= \frac{1}{2} \sum_{\ell=1}^{L} \rho_\ell \int_{\mathcal{Z}_\ell} d\mathbf{z}\, \mathcal{N}(\mathbf{z};\, \boldsymbol{\zeta}, \mathbf{S})\ \mathrm{tr}\left\{ \boldsymbol{\Gamma} \left(\mathbf{z}\,\mathbf{z}^{\mathrm{T}} - 2\,\mathbf{z}\,\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\,\boldsymbol{\mu}^{\mathrm{T}} \right) \right\} \\
&= \frac{1}{2} \sum_{\ell=1}^{L} \rho_\ell\ \mathrm{tr} \left\{ \boldsymbol{\Gamma} \int_{\mathcal{Z}_\ell} d\mathbf{z} \mathcal{N}(\mathbf{z};\, \boldsymbol{\zeta}, \mathbf{S}) \left(\mathbf{z}\,\mathbf{z}^{\mathrm{T}} - 2\,\mathbf{z}\,\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\,\boldsymbol{\mu}^{\mathrm{T}} \right) \right\} \\
&= \frac{1}{2} \sum_{\ell=1}^{L} \rho_\ell\, \phi_\ell\ \mathrm{tr} \left\{ \boldsymbol{\Gamma} \left(\boldsymbol{\Omega}_\ell - 2\,\boldsymbol{\omega}_\ell\,\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\,\boldsymbol{\mu}^{\mathrm{T}} \right) \right\} \qquad (161) \\
&= \frac{1}{2} \sum_{\ell=1}^{L} \rho_\ell\, \phi_\ell\ \mathrm{tr} \left\{ \boldsymbol{\Gamma} \left[\tilde{\boldsymbol{\Omega}}_\ell + (\boldsymbol{\omega}_\ell - \boldsymbol{\mu}) (\boldsymbol{\omega}_\ell - \boldsymbol{\mu})^{\mathrm{T}} \right] \right\} , \qquad (162)
\end{aligned}$$

where

$$\boxed{\boldsymbol{\omega}_\ell = \frac{1}{\phi_\ell} \int_{\mathcal{Z}_\ell} d\mathbf{z}\, \mathcal{N}(\mathbf{z};\, \boldsymbol{\zeta}, \mathbf{S})\ \mathbf{z} ,} \tag{163}$$

$$\boxed{\boldsymbol{\Omega}_\ell = \frac{1}{\phi_\ell} \int_{\mathcal{Z}_\ell} d\mathbf{z}\, \mathcal{N}(\mathbf{z};\, \boldsymbol{\zeta}, \mathbf{S})\ \mathbf{z}\,\mathbf{z}^{\mathrm{T}} ,} \tag{164}$$

$$\boxed{\tilde{\boldsymbol{\Omega}}_\ell = \boldsymbol{\Omega}_\ell - \boldsymbol{\omega}_\ell\,\boldsymbol{\omega}_\ell^{\mathrm{T}} .} \tag{165}$$

Vector $\boldsymbol{\omega}_\ell$ is the normalized first moment (centroid) and matrix $\boldsymbol{\Omega}_\ell$ is the normalized second moment of $\mathcal{N}(\mathbf{z};\, \boldsymbol{\zeta}, \mathbf{S})$ when restricted to the region $\mathcal{Z}_\ell$. Vector $\boldsymbol{\omega}_\ell$ is also referred

to as an "effective measurement" for reasons that will be made clear in a moment. Matrix $\tilde{\mathbf{\Omega}}_\ell$ is the second central moment (covariance matrix) of $\mathcal{N}(\mathbf{z};\, \boldsymbol{\zeta}, \mathbf{S})$ in $\mathcal{Z}_\ell$.

An interesting form of the objective function is obtained by defining the weighted sum of covariance matrices

$$\tilde{\mathbf{\Omega}} = \frac{1}{2} \sum_{\ell=1}^{L} \rho_\ell \, \phi_\ell \, \tilde{\mathbf{\Omega}}_\ell \tag{166}$$

such that, when $\eta_2(\boldsymbol{\theta})$ is re-combined with the determinant term $\eta_1(\boldsymbol{\theta})$, the overall objective function becomes

$$Q(\boldsymbol{\theta}) = \operatorname{tr}\left\{\mathbf{\Gamma}\,\tilde{\mathbf{\Omega}}\right\} + \sum_{\ell=1}^{L} \rho_\ell \, \phi_\ell \, \log \mathcal{N}(\boldsymbol{\omega}_\ell; \boldsymbol{\mu}, \mathbf{\Gamma}) \; . \tag{167}$$

Aside from the trace term, which reflects the cumulative measurement uncertainty within all bins, the objective function is equivalent to one in which the effective measurements $\boldsymbol{\omega}_\ell$ are observed point measurements. Indeed, from the standpoint of estimating the mean, this problem is identical to one in which the $\boldsymbol{\omega}_\ell$ are observed point measurements since the trace term is not a function of $\boldsymbol{\mu}$.

Location of the Stationary Point. The stationary point of the objective function is found by equating to zero its partial derivatives with respect to the various parameters. Noting equation (161) and the identity

$$\frac{\partial}{\partial \mathbf{x}} \operatorname{tr}\left\{\mathbf{A}\,\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\right\} = 2\,\mathbf{A}\,\mathbf{x}, \tag{168}$$

the derivative of the objective function with respect to the mean is given by

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} = \frac{\partial \eta_2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} = \mathbf{\Gamma} \sum_{\ell=1}^{L} \rho_\ell \, \phi_\ell \, (\boldsymbol{\omega}_\ell - \boldsymbol{\mu}) . \tag{169}$$

Equating to zero gives the stationary value for the mean as

$$\boxed{\hat{\boldsymbol{\mu}} = \frac{1}{\kappa} \sum_{\ell=1}^{L} \rho_\ell \, \phi_\ell \, \boldsymbol{\omega}_\ell .} \tag{170}$$

The information matrix is estimated by differentiating $Q(\boldsymbol{\theta})$ with respect to $\mathbf{\Gamma}$, equating the result to zero, and substituting in place of $\boldsymbol{\mu}$ the stationary value $\hat{\boldsymbol{\mu}}$. Expressing the objective function as the sum of equations (160) and (162), and drawing

upon the derivative identities

$$\frac{\partial}{\partial \mathbf{A}} \log \left| \mathbf{A} \right| = \mathbf{A}^{-1} \tag{171}$$

and (172)

$$\frac{\partial}{\partial \mathbf{A}} \operatorname{tr} \{\mathbf{A}\,\mathbf{B}\} = \mathbf{B} \tag{173}$$

gives the desired derivative as

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\Gamma}} = \frac{\kappa}{2}\,\boldsymbol{\Gamma}^{-1} - \frac{1}{2}\sum_{\ell=1}^{L} \rho_\ell\,\phi_\ell\,\left\{\tilde{\boldsymbol{\Omega}}_\ell + (\boldsymbol{\omega}_\ell - \boldsymbol{\mu})\,(\boldsymbol{\omega}_\ell - \boldsymbol{\mu})^{\mathrm{T}}\right\}. \tag{174}$$

Equating to zero and substituting $\hat{\boldsymbol{\mu}}$ then gives the stationary value

$$\boxed{\hat{\mathbf{P}} = \hat{\boldsymbol{\Gamma}}^{-1} = \frac{1}{\kappa}\sum_{\ell=1}^{L} \rho_\ell\,\phi_\ell\,\left\{\tilde{\boldsymbol{\Omega}}_\ell + (\boldsymbol{\omega}_\ell - \hat{\boldsymbol{\mu}})\,(\boldsymbol{\omega}_\ell - \hat{\boldsymbol{\mu}})^{\mathrm{T}}\right\}.} \tag{175}$$

Maximality of the Stationary Point. A stationary point is a local maximum of the objective function if the Hessian matrix of the objective function is negative definite at the stationary point. Construction of the Hessian matrix, however, requires taking second derivatives of the function with respect to all possible pairwise combinations of the elements of the mean vector and information matrix, which can be a formidable task. If the objective function is *radially concave*, however, then it has the nice characteristic that there is only one stationary point and it is the unique *global* maximum. The remainder of this section parallels a proof by Liporace [11] of a test for radial concavity, which does not require forming all of the second partial derivatives.[1] Now, if the function fails the concavity test, a critical point may still be a local maximum, so the test is not conclusive in that regard. If, on the other hand, the function passes the test, then no further analysis is needed.

The basic idea behind the test for radial concavity, which is illustrated in figure D-1, is to formulate a family of one-dimensional trajectories along the function that cover every possible direction from the stationary point, and then show that these one-dimensional trajectories are all individually concave. If the objective function is continuous at the stationary point, and the "hill curves downward" in every direction

[1] Indeed, the test outlined below would be a test of absolute concavity were it not for the pathological family of functions in which non-concave behavior can occur along contour lines that are perpendicular to radial lines emanating from the critical point, which is the motivation for the term *radially concave*. Even in this pathological case, however, the stationary point is still the unique maximum.

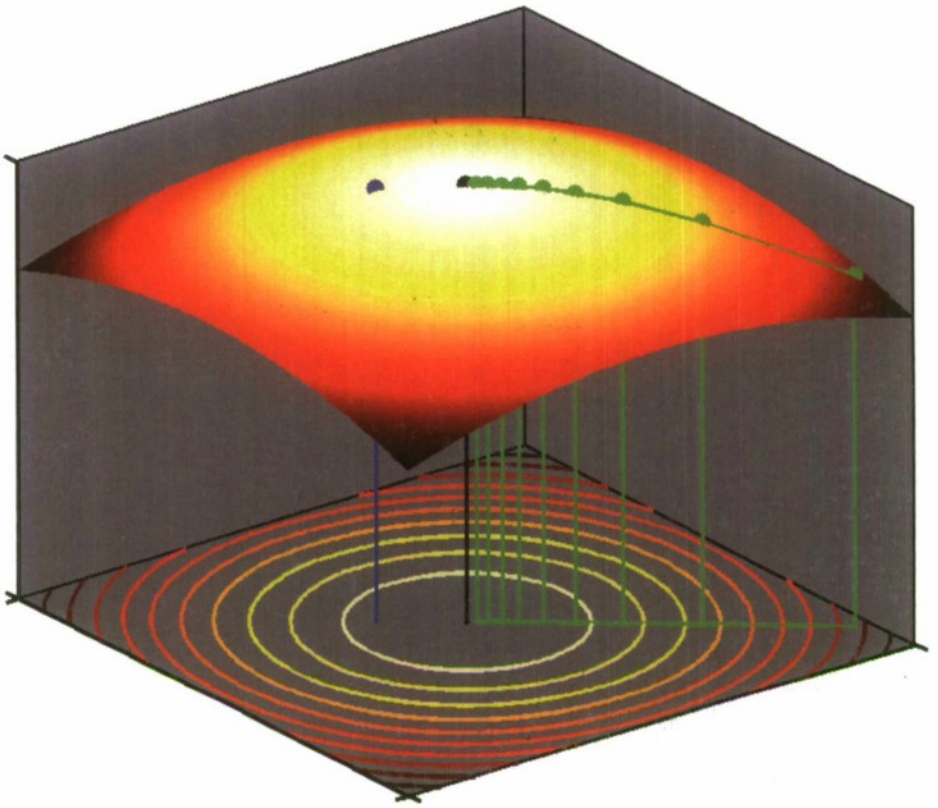

Figure D-1: A Test for Radial Concavity.
The surface plot and equal-value contour lines depict a hypothetical concave function $f(\mathrm{x})$ *of the two-dimensional variable* x. *The "stem lines" indicate the locations of the stationary point* $\hat{\mathrm{x}}$ *(black), an arbitrarily chosen reference point* x_r *(blue), and the points* $\vec{\mathrm{x}}$ *(green) that satisfy the constraint* $\hat{\mathrm{x}} = \lambda \mathrm{x}_r + (1-\lambda)\vec{\mathrm{x}}$ *for* $\lambda = 0.1, 0.2, \ldots, 0.9$. *Over the continuum* $0 < \lambda < 1$, *the locus of* $\vec{\mathrm{x}}$ *forms a one-dimensional trajectory (a subsegment of which is indicated by lower green line) that emanates outward from* $\hat{\mathrm{x}}$ *along the line passing through* $\hat{\mathrm{x}}$ *and* x_r, *but in the direction opposite to* x_r. *The upper ends of the stem lines correspond to the function values* $f(\hat{\mathrm{x}})$, $f(\mathrm{x}_r)$, *and* $f(\vec{\mathrm{x}})$. *Given particular values of* $\hat{\mathrm{x}}$ *and* x_r, $\vec{\mathrm{x}}$ *is completely determined by* λ, *such that* $f(\vec{\mathrm{x}})$ *can be expressed as* $\vec{f}(\lambda)$ *(upper green line), which follows the curvature of* $f(\mathrm{x})$ *along the one-dimensional trajectory. By varying* x_r, *this trajectory can be made to extend from* $\hat{\mathrm{x}}$ *in every possible direction. Thus, if* $\vec{f}(\lambda)$ *is necessarily concave with respect to* λ, *regardless of the value of* x_r, *then the overall function is radially concave and the stationary point is necessarily the unique maximum.*

away from the stationary point, then the stationary point must be the unique maximum. This approach reduces the maximality test to evaluating a single second derivative with respect to a scalar variable.

To set up the concavity test, each set of parameter values in $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Gamma}\}$ is viewed as a "point" in the domain of the objective function (i.e., parameter space). From any such point, a family of radial trajectories is obtained by defining a "reference point" $\boldsymbol{\theta}_r = \{\boldsymbol{\mu}_r, \boldsymbol{\Gamma}_r\}$ and a set of "locus points" $\vec{\boldsymbol{\theta}} = \left\{\vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}\right\}$ that satisfy

$$\boldsymbol{\mu} = \lambda \boldsymbol{\mu}_r + (1 - \lambda)\, \vec{\boldsymbol{\mu}}\,, \tag{176}$$

$$\boldsymbol{\Gamma} = \lambda \boldsymbol{\Gamma}_r + (1 - \lambda)\, \vec{\boldsymbol{\Gamma}}\,, \tag{177}$$

where λ is a scalar variable satisfying $0 < \lambda < 1$. Substitution of these expressions into the objective function generates a *modified objective function* over a restricted parameter space. For a given $\boldsymbol{\theta}_r$, the values of $\vec{\boldsymbol{\theta}}$ fall along a radial line emanating from (but not including) $\boldsymbol{\theta}$, in the direction opposite $\boldsymbol{\theta}_r$. The direction of $\vec{\boldsymbol{\theta}}$ from $\boldsymbol{\theta}$ is thus a function of the reference point location. The distance of $\vec{\boldsymbol{\theta}}$ from $\boldsymbol{\theta}$ varies as a function of λ, with a scale factor equal to the distance from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_r$. The reference point $\boldsymbol{\theta}_r$ is treated as an implicit parameter that is arbitrarily chosen but is a fixed constant once it is chosen. The modified function is then defined as

$$\vec{Q}(\lambda, \boldsymbol{\theta}) = Q\left(\lambda \boldsymbol{\theta}_r + (1 - \lambda)\, \vec{\boldsymbol{\theta}}\right), \tag{178}$$

where the short-hand notation $\boldsymbol{\theta} = \lambda \boldsymbol{\theta}_r + (1 - \lambda)\, \vec{\boldsymbol{\theta}}$ is used to indicate that equations (176) and (177) are satisfied. To show that the objective function is concave, the modified objective function is twice differentiated with respect to λ. The resulting second derivative is evaluated at the stationary point $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, where it is shown to be manifestly negative. The information term $\eta_1(\boldsymbol{\theta})$ and error term $\eta_2(\boldsymbol{\theta})$ are again treated individually. Both terms contribute negative values to the overall second derivative.

The first component is expressed in terms of λ by substituting equation (177) into equation (160), which gives

$$\vec{\eta}_1(\lambda, \boldsymbol{\theta}) = \frac{\kappa}{2} \log \left| \lambda \boldsymbol{\Gamma}_r + (1 - \lambda)\, \vec{\boldsymbol{\Gamma}} \right|. \tag{179}$$

When evaluating this determinant, it is helpful to observe that the constraint on the information matrices in equation (177) is satisfied for all λ, which implies that $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_r$, and $\vec{\boldsymbol{\Gamma}}$ are all diagonalized by the same set of eigenvectors. Therefore, there exists an orthogonal matrix $\mathbf{U}$ for which

$$\mathbf{U}\, \boldsymbol{\Gamma}\, \mathbf{U}^{\mathrm{T}} = \mathbf{D} = \mathbf{U} \left\{ \lambda \boldsymbol{\Gamma}_r + (1 - \lambda)\, \vec{\boldsymbol{\Gamma}} \right\} \mathbf{U}^{\mathrm{T}} = \lambda \mathbf{D}_r + (1 - \lambda)\, \vec{\mathbf{D}}\,, \tag{180}$$

where $\mathbf{D}=\mathrm{diag}(d_1,\ldots,d_M)$, $\mathbf{D}_r=\mathrm{diag}(d_{r,1},\ldots,d_{r,M})$, and $\vec{\mathbf{D}}=\mathrm{diag}(\vec{d}_1,\ldots,\vec{d}_M)$ contain the eigenvalues of $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_r$, and $\vec{\boldsymbol{\Gamma}}$, respectively. The first component of the modified objective function is thus given by

$$\vec{\eta}_1(\lambda,\boldsymbol{\theta}) \;=\; \frac{\kappa}{2}\log\left|\,\lambda\,\mathbf{D}_r+(1-\lambda)\,\vec{\mathbf{D}}\,\right| \;=\; \frac{\kappa}{2}\sum_{i=1}^{M}\log\left\{\lambda\,d_{ri}+(1-\lambda)\,\vec{d}_i\right\}\,, \qquad (181)$$

which is differentiated twice with respect to λ to obtain

$$\frac{\partial^2}{\partial\lambda^2}\,\vec{\eta}_1(\lambda,\boldsymbol{\theta}) \;=\; -\kappa\sum_{i=1}^{M}\frac{\left(d_{ri}-\vec{d}_i\right)^2}{\left\{\lambda\,d_{ri}+(1-\lambda)\,\vec{d}_i\right\}^2} \;=\; -\kappa\sum_{i=1}^{M}\frac{1}{\vec{d}_i^{\,2}}\left(d_{ri}-\vec{d}_i\right)^2\,. \qquad (182)$$

The numerator in the summand of equation (182) is strictly positive since d_{ri} and $\vec{d}_i$ cannot be equal. In general, the denominator is only non-negative (i.e., $\boldsymbol{\Gamma}$ will have zero-valued eigenvalues *somewhere*), but $\vec{d}_i^{\,2}$ is strictly positive at the stationary point, as long as the number of independent "measurements" $\boldsymbol{\omega}_\ell$ is larger than the dimension of the measurement space. Finally, since κ is also a positive number, the overall expression for the second derivative is therefore negative, regardless of the value of $\boldsymbol{\Gamma}_r$.

Turning now to the quadratic-form component, the manipulations involved with evaluating $\eta_2(\boldsymbol{\theta})$ are eased by defining the intermediate variables

$$\boldsymbol{\delta} \;=\; \vec{\boldsymbol{\mu}}-\boldsymbol{\mu}_r\,, \qquad (183)$$

$$\boldsymbol{\Delta} \;=\; \vec{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}_r\,, \qquad (184)$$

$$\mathbf{y}_\ell \;=\; \boldsymbol{\omega}_\ell-\vec{\boldsymbol{\mu}}\,, \qquad (185)$$

such that

$$\boldsymbol{\Gamma} \;=\; \lambda\,\boldsymbol{\Gamma}_r+(1-\lambda)\vec{\boldsymbol{\Gamma}} \;=\; \vec{\boldsymbol{\Gamma}}-\lambda\,\boldsymbol{\Delta}\,, \qquad (186)$$

$$\boldsymbol{\omega}_\ell-\boldsymbol{\mu} \;=\; \boldsymbol{\omega}_\ell-\lambda\,\boldsymbol{\mu}_r-(1-\lambda)\vec{\boldsymbol{\mu}} \;=\; \mathbf{y}_\ell+\lambda\,\boldsymbol{\delta}\,. \qquad (187)$$

The modified function is evaluated by substituting equations (176) and (177) into equation (162) and collecting terms in λ to obtain

$$\vec{\eta}_2(\lambda,\boldsymbol{\theta}) \;=\; \frac{1}{2}\sum_{\ell=1}^{L}\rho_\ell\;\phi_\ell\;\mathrm{tr}\Big\{\,\vec{\boldsymbol{\Gamma}}\,\tilde{\boldsymbol{\Omega}}_n+\vec{\boldsymbol{\Gamma}}\,\mathbf{y}_n\,\mathbf{y}_n^{\mathrm{T}}+\lambda\left(2\vec{\boldsymbol{\Gamma}}\,\mathbf{y}_n\,\boldsymbol{\delta}-\boldsymbol{\Delta}\,\tilde{\boldsymbol{\Omega}}_n-\boldsymbol{\Delta}\,\mathbf{y}_n\,\mathbf{y}_n^{\mathrm{T}}\right)$$

$$+\;\lambda^2\left(\vec{\boldsymbol{\Gamma}}\,\boldsymbol{\delta}\,\boldsymbol{\delta}^{\mathrm{T}}-2\boldsymbol{\Delta}\,\mathbf{y}_n\boldsymbol{\delta}^{\mathrm{T}}\right)-\lambda^3\left(\boldsymbol{\Delta}\,\boldsymbol{\delta}\,\boldsymbol{\delta}^{\mathrm{T}}\right)\Big\}\,. \qquad (188)$$

Differentiating twice with respect to λ and performing some algebra then gives

$$\frac{\partial^2}{\partial \lambda^2}\,\vec{\eta}_2(\lambda,\boldsymbol{\theta}) \;=\; \kappa\,\boldsymbol{\delta}^{\mathrm{T}}\,\boldsymbol{\Gamma}\,\boldsymbol{\delta} - 2\,\lambda\,\boldsymbol{\delta}^{\mathrm{T}}\,\boldsymbol{\Delta}\left\{\sum_{\ell=1}^{L}\rho_\ell\;\phi_\ell\;\left(\boldsymbol{\omega}_n - \boldsymbol{\mu}\right)\right\}. \tag{189}$$

At the stationary point, the mean vector must satisfy $\sum_{\ell=1}^{L}\rho_\ell\,\phi_\ell\;\left(\boldsymbol{\omega}_\ell - \hat{\boldsymbol{\mu}}\right) = 0$, such that the term in braces in equation (189) is zero. Combining these results with equation (182) gives

$$\frac{\partial^2}{\partial \lambda^2}\,\vec{Q}(\lambda,\hat{\boldsymbol{\theta}}) \;=\; -\kappa\sum_{i=1}^{M}\frac{1}{\hat{d}_i^2}\left(d_{ri} - \vec{d}_i\right)^2 - \kappa\,\boldsymbol{\delta}^{\mathrm{T}}\,\hat{\boldsymbol{\Gamma}}\,\boldsymbol{\delta} \;<\; 0\,, \tag{190}$$

where $\boldsymbol{\delta}^{\mathrm{T}}\,\hat{\boldsymbol{\Gamma}}\,\boldsymbol{\delta}$ is necessarily positive because $\hat{\boldsymbol{\Gamma}}$ is positive definite. This expression is independent of λ and its sign is independent of the value of $\boldsymbol{\theta}_r$. The original objective function therefore must be concave, and the stationary point is the unique maximum.

APPENDIX E
GAUSSIAN MOMENTS IN AN INTERVAL

When implementing histogram-based estimation methods, it is necessary to evaluate the bin probabilities and the first two local moments in each bin, which are developed in this appendix for the scalar Gaussian density

$$p(x;\mu,\sigma^2) \;=\; \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\,(x-\mu)^2\right\}. \tag{191}$$

The integral expressions required for these probabilities are not very exciting or interesting, but they are quite useful. In what follows, the bin-probability and local-moment computations are outlined for a "typical" bin interval $I = [a, b)$.

Bin Probability. The bin probability is given by the integral

$$\phi(I,\mu,\sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b \exp\left\{-\frac{1}{2\sigma^2}\,(x-\mu)^2\right\} dx\,, \tag{192}$$

which is evaluated by performing the change of variables

$$y \;=\; \frac{1}{\sqrt{2\sigma^2}}\,(x-\mu) \;\Longrightarrow\; x \;=\; \sqrt{2\sigma^2}\,y+\mu \;\Longrightarrow\; dx \;=\; \sqrt{2\sigma^2}\,dy \tag{193}$$

to obtain

$$\phi(I,\mu,\sigma^2) \;=\; \frac{1}{\sqrt{\pi}} \int_\alpha^\beta e^{-y^2}\,dy\,, \tag{194}$$

where the modified integration boundaries are

$$\boxed{\alpha \;=\; \frac{1}{\sqrt{2\sigma^2}}\,(a-\mu)\,,} \tag{195}$$

$$\boxed{\beta \;=\; \frac{1}{\sqrt{2\sigma^2}}\,(b-\mu)\,.} \tag{196}$$

Given a computational routine for the error function

$$\mathrm{erf}(u) \;=\; \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2}\,dx\,, \tag{197}$$

the bin probability is computed as

$$\boxed{\phi(I,\mu,\sigma^2) \;=\; \frac{1}{2}\Big\{\,\mathrm{erf}(\beta) - \mathrm{erf}(\alpha)\Big\}.} \tag{198}$$

First Moment. The first moment in the interval is

$$\omega(I,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b x \, \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx\,. \tag{199}$$

Adding and subtracting $\mu \int_a^b e^{-(x-\mu)^2/(2\sigma^2)} dx$ yields

$$\omega(I,\mu,\sigma^2) = \tilde{\omega}(I,\mu,\sigma^2) + \mu\,\phi(I,\mu,\sigma^2)\,, \tag{200}$$

where $\tilde{\omega}(I,\mu,\sigma^2)$ is the first central local moment

$$\tilde{\omega}(I,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b (x-\mu) \, \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx\,. \tag{201}$$

This central moment is evaluated by performing the change of variables

$$y = \frac{1}{2\sigma^2}(x-\mu)^2 \Longrightarrow (x-\mu) = \pm\sqrt{2\sigma^2}\, y^{1/2} \Longrightarrow dx = \pm\frac{\sigma}{\sqrt{2}}\, y^{-1/2}\, dy \tag{202}$$

to obtain

$$\tilde{\omega}(I,\mu,\sigma^2) = \frac{\sigma}{\sqrt{2\pi}} \int_{\alpha^2}^{\beta^2} e^{-y}\, dy = \frac{\sigma}{\sqrt{2\pi}} \left(e^{-\alpha^2} - e^{-\beta^2}\right). \tag{203}$$

The first moment is therefore

$$\boxed{\omega(I,\mu,\sigma^2) = \frac{\sigma}{\sqrt{2\pi}} \left(e^{-\alpha^2} - e^{-\beta^2}\right) + \mu\,\phi(I,\mu,\sigma^2)\,.} \tag{204}$$

Second Moment. The second moment in the interval is

$$\Omega(I,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b x^2 \, \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx\,. \tag{205}$$

Completing the square in the moment variable yields

$$\Omega(I,\mu,\sigma^2) = \tilde{\Omega}(I,\mu,\sigma^2) + 2\mu\,\omega(I,\mu,\sigma^2) - \mu^2\,\phi(I,\mu,\sigma^2)\,, \tag{206}$$

where $\tilde{\Omega}(I,\mu,\sigma^2)$ is the second central local moment

$$\tilde{\Omega}(I,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b (x-\mu)^2 \, \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx\,. \tag{207}$$

This expression is simplified using the change of variables defined in equation (193) to obtain

$$\tilde{\Omega}(I,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_\alpha^\beta \left(2\sigma^2\, y^2\right) e^{-y^2} \left(\sqrt{2\sigma^2}\, dy\right) = \frac{2\sigma^2}{\sqrt{\pi}} \int_\alpha^\beta y^2\, e^{-y^2}\, dy\,, \tag{208}$$

where α and β are defined in equations (195) and (196), respectively. Using the integration-by-parts formula $\int u\,dv = u\,v - \int v\,du$ with $dv = e^{-y^2}dy$ and $u = y^2$ (such that $v = -e^{-y^2}/(2y)$ and $du = 2y\,dy$), the second central moment is

$$\begin{aligned}\tilde{\Omega}(I,\mu,\sigma^2) &= \frac{2\sigma^2}{\sqrt{\pi}}\left\{-\frac{1}{2}\,y\,e^{-y^2}\Big|_{\alpha}^{\beta} + \int_{\alpha}^{\beta} e^{-y^2}\,dy\right\} \\ &= \frac{\sigma^2}{\sqrt{\pi}}\left(\alpha\,e^{-\alpha^2} - \beta\,e^{-\beta^2}\right) + 2\sigma^2\,\phi(I,\mu,\sigma^2)\,. \end{aligned} \tag{209}$$

The overall second moment is therefore given by

$$\Omega(I,\mu,\sigma^2) = \frac{\sigma^2}{\sqrt{\pi}}\left(\alpha\,e^{-\alpha^2} - \beta\,e^{-\beta^2}\right) + 2\sigma^2\,\phi(I,\mu,\sigma^2) + 2\mu\,\omega(I,\mu,\sigma^2) - \mu^2\,\phi(I,\mu,\sigma^2)\,.$$

Substituting the definition of $\omega(I,\mu,\sigma^2)$ given in equation (204), collecting like terms, and defining the function

$$\boxed{c(x) = \sqrt{\frac{2}{\pi}}\,\sigma\,\mu + \sqrt{\frac{1}{\pi}}\,\sigma^2\,x} \tag{210}$$

gives the second moment as

$$\boxed{\Omega(I,\mu,\sigma^2) = c(\alpha)\,e^{-\alpha^2} - c(\beta)\,e^{-\beta^2} + \left(2\sigma^2 + \mu^2\right)\,\phi(I,\mu,\sigma^2)\,.} \tag{211}$$

Underflow Issues. The expressions given above for the bin probabilities and moments become unstable when interval I is far removed from the mean μ. In these regions, underflow issues become a dominating factor. This difficulty is easily avoided, however, by restricting the range of integration to those bins that lie within, say, ± 8 standard deviations of the mean, effectively treating the remaining bins as a "set of measure zero."

REFERENCES

1. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood Estimation from Incomplete Data," *J. Royal Stat. Soc. (B)*, vol. 39, no. 1, 1977, pp. 1–38.
2. G. J. McLachlan and P. N. Jones, "Fitting Mixture Models to Grouped and Truncated Data via the EM Algorithm," *Biometrics*, vol. 44, 1988, pp. 571–578.
3. T. E. Luginbuhl, *Estimation OF General Discrete-Time FM Processes*, Ph.D. thesis, University of Connecticut, 1999.
4. T. E. Luginbuhl and P. Willet, "Estimating the Parameters of General Frequency Modulated Signals," *IEEE Trans. Signal Process.*, vol. 52, no. 1, January 2004, pp. 117–131.
5. R. L. Streit, "Tracking on Intensity-Modulated Data Streams," NUWC-NPT Technical Report 11,221, Naval Undersea Warfare Center, Newport, RI, 2000.
6. K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, New York, 1990.
7. Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic Press, New York, 1988.
8. J. F. C. Kingman, *Poisson Processes*, Oxford University Press, New York, 1993.
9. R. A. Redner and H. F. Walker, "Mixture Densities, Maximum Likelihood, and the EM Algorithm," *SIAM Review*, vol. 26, no. 2, 1984, pp. 195–239.
10. G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, Inc., New York, 1997.
11. L. A. Liporace, "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," *IEEE Trans. Informat. Theory*, vol. 28, no. 5, 1982, pp. 729–734.

INITIAL DISTRIBUTION LIST

Addressee	No. of Copies
Office of Naval Research (Attn: John Tague)	1
NSWC Panama City (Attn: James Cobb)	1
Defense Technical Information Center	2
Center for Naval Analyses	1
University of Florida (Attn: K.C. Slatton)	1
University of Connecticut (Attn: Peter Willett)	1
Metron Inc. (Attn: Roy Streit)	1