Moments and Moment Matrices for Invariant Classification of Noise Contaminated Spatial Pattern

Alfred O. Hero, J. O'Neill, W. J. Williams, H. Hadinejad-Mahram

COMMUNICATIONS & SIGNAL PROCESSING LABORATORY
Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109-2122

Technical Report No. 307, Feb. 1997

Abstract

We introduce methods for detection and classification of noise contaminated patterns which are based on performing subspace decomposition on a matrix of higher order spatial moments. The subspace decomposition permits signal-alone moments to be recovered from signal-plus-noise observations arising from an unknown mixture distribution. The methods use normalized power moments and normalized factorial moments as pattern descriptors. While the set of L-th order factorial moments are in one-to-one correspondence with the set of L-th order power moments, digital computation of factorial moments is more numerically stable permitting a larger number of moments to be utilized. We illustrate these methods for wordspotting in automatic processing of documents, and for digital modulation classification in communications.

I. Introduction

Moment methods of spatial pattern classification have been applied to a great variety of areas including: character recognition [1], edge detection [2], galaxy cluster analysis [3], [4], and word spotting [5].

Common justifications for spatial moments are: 1) they provide a non-parametric pattern description; 2) combinations of moments have been identified with important invariances such as rotation, scale, and translation [6], [7]; 3) sample moments can usually be treated as jointly Gaussian random variables simplifying construction of statistical tests of significance. With the introduction of non-negative definite higher moment matrices described in this paper, we provide another justification: moments can be used to effectively separate signal pattern from noise background via noise subspace processing.

Noise subspace processing has been successfully applied to achieve noise reduction for many signal processing problems. This type of processing involves performing eigencompositions on non-negative definite covariance matrices formed from second order moment (autocorrelation) lag products. A longstanding problem has been to generalize this technique to accomodate higher order moments. In [8] a generalization was obtained by applying a generalized singular value decomposition to a tensor of moments of fixed order indexed by multivariate lag parameters. Here we introduce a noise subspace technique which uses a generalization in a different direction: eigendecomposition of a non-negative definite symmetric matrix of moments of different orders at zero lag.

The crux of the technique is to arrange the higher order spatial moments into a non-negative definite matrix, which we call the moment matrix. Noise robustification of the moment estimates is accomplished by applying the Cholesky factor of the noise-alone moment matrix as a prewhitener, performing an eigendecomposition on the whitened moment matrix, and identifying and eliminating the noise-dependent subspace. Renormalization of the remaining eigenvalues and dewhitening leads to recovery of the noiseless moment matrix and the noiseless spatial moments. The recovered moments can be more reliably used in algorithms for constructing moment invariants, Neyman-Pearson and constant false alarm rate (CFAR) detectors, Zernicke moments, etc.

The main disadvantage of standard power moment descriptors is that under finite register length arithmetic, they become computationally unstable as the moment orders get even moderately large. For this reason it is advantageous to work with factorial moments. While power moments and factorial moments are mathematically equivalent, in the sense that they are related through a 1-1 transformation, digital computation of factorial moments is more accurate. This means that for a fixed register length computer a larger number of factorial moments can be reliably computed resulting in more accurate pattern matching. Alternatively, for a fixed number of moments factorial moments can be reliably computed using shorter register lengths leading to potential savings in hardware cost and power dissipation.

Factorial moments come in two varieties, ascending and descending, and have mainly been applied to integer-valued scalar random variables [9]. Descending factorial moments appear to have been used in more applications than ascending factorial moments. One dimensional descending factorials (also known as negative factorials [10]) have been used in statistics in diverse areas such as identification of parameters in discrete mixture distributions [11] and moment expansions for stochastic processes [12]. One dimensional descending factorial moments have also been used in many applications in physics including: neutron-neutron coincidence counting [13], discriminants of non-classical quantum states in coupled lasers [14], galaxy distribution theory [4], discriminants

[1]This work was supported in part by NSA contract MDA904-95-C-2157 and AFOSR contract F49620-96-0028.

of fractal behavior in high energy physics [15], and statistics of nuclear interactions [16]. In this paper we introduce both ascending and descending spatial factorial moments in the more general context of multiple dimensions and non-integer random variables.

The main focus of this paper is moment matrices for univariate and bivariate mixture densities. Extensions are indicated in the appendix for trivariate mixture densities. Univariate densities arise in application areas involving multiple observations of a single random variable, e.g. the interarrival distribution of a Poisson process in a queuing network or echo arrivals in seismic exploration. Bivariate densities arise when the observation histogram is indexed by two parameters such as: one dimensional continuous random processes, e.g. speech signals (time and amplitude); or two dimensional "0,1" processes, e.g. binary document images (row and column locations of the active pixels) or in phase and quadrature signal constellations in digital communications (locations of time samples in the complex plane). Trivariate densities are required when the histogram is triply indexed such as occurs for gray scale images (row, column and intensity). Examples are provided to illustrate how moment matrices can be applied to these problems.

II. Moment Representations of Multivariate Distributions

Let $f = f(\underline{x})$ be the density function of a random point $\underline{X}$ in n dimensional space $\mathbb{R}^n$. It is well known that the density function $f(\underline{x})$ can be characterized by its set of higher order moments of mixed orders. For clarity, in the sequel we will specialize to the case of *spatial random variables* evolving in the plane ($\underline{x} = (x, y)$). The univariate moment matrices associated with a random variable X are simply the upper $(L+1) \times (L+1)$ blocks of the bivariate moment matrices discussed below (see Proposition 2 in Appendix B). The case of higher dimensional variables, e.g. 3D variables $\underline{x} = (x, y, z)$, is more notationally burdensome. The trivariate case is treated in the appendix (Appendix C).

A. Bivariate Power Moments

The bivariate power moment (PM) $\mu_{X,Y}(k_1, k_2)$ of (integer) order k_1, k_2 is defined as as the expectation

$$\mu_{X,Y}(k_1, k_2) = E[X^{k_1} Y^{k_2}]. \tag{1}$$

Since the bivariate monomials form a basis for the space of all three-dimensional square integrable functions, the set of L^2 power moments $\{\mu_{X,Y}(k_1, k_2)\}_{k_1,k_2=1}^{L}$ completely characterizes the density in the limit as $L \to \infty$.

B. Bivariate Factorial Moments

For a bivariate probability density function $f(x, y)$ the bivariate *fractional factorial moment* (FFM) $\delta_{X,Y}$ of (non-negative integer) orders k_1, k_2 and fraction parameters s_1, s_2 is defined as the expectation

$$\delta_{X,Y}(k_1, k_2; s_1, s_2) = E[(X)_{s_1}^{k_1} \ (Y)_{s_2}^{k_2}]. \tag{2}$$

where

$$(x)_s^k = x(x+s)\cdots(x+s(k-1)) \tag{3}$$

is the Pochhammer symbol [17] denoting a fractional factorial with fraction s and order k. By convention $(x)_s^0 = 1$. For $s > 0$, $(x)_s^k$ is called an ascending factorial while for $s < 0$ it is called a descending factorial [9].

The descending factorials are related to the ascending factorials by the simple relation

$$(x)_{-s}^k = (-1)^k \ (-x)_s^k, \quad s > 0. \tag{4}$$

When x is a positive integer and $s = 1$ or $s = -1$ the FMM reduces to the standard (non-fractional) ascending and descending factorial moments: $(x)_1^k = \frac{(x+k-1)!}{(x-1)!}$ and $(x)_{-1}^k = \frac{x!}{(x-k)!}$, respectively.

C. Finite Register Length Effects

Lemma 1 in Appendix A establishes that the set of L^2 PM's $\{\mu_{X,Y}(k_1, k_2)\}_{k_1 k_2=1}^{L}$ and the set of L^2 FFM's $\{\delta_{X,Y}(k_1, k_2; s_1, s_2)\}_{k-1,k_2=1}^{L}$ are equivalent in the sense that they are related through a 1-1 transformation. However, as we shall discuss here, the FFM's are less sensitive than PM's to errors due to finite register length computation.

III. Non-negative Definite Moment Matrices

Here we introduce non-negative definite matrices of higher order moments which generalize the notion of a covariance matrix.

A. Bivariate Power Moment Matrix

For a planar random vector (X, Y) the $(2L+1) \times (2L+1)$ bivariate power moment matrix $\mathbf{P}_{X,Y}^{2L+1}$ is defined as the expectation over X, Y of the dyadic outer product of the $2L+1$ element vector $\underline{U} = [1, X, X^2, \ldots, X^L, Y, Y^2, \ldots, Y^L]^T$, or more explicitly

$$\mathbf{P}_{X,Y}^{2L+1} = \tag{5}$$
$$E \begin{bmatrix} 1 & X & \cdots & X^L & Y & \cdots & Y^L \\ X & X^2 & \cdots & X^{L+1} & XY & \cdots & XY^L \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X^L & X^{L+1} & \cdots & X^{2L} & X^L Y & \cdots & X^L Y^L \\ & & & & Y^2 & \cdots & Y^{L+1} \\ & (sym) & & & \vdots & \ddots & \vdots \\ & & & & Y^{L+1} & \cdots & Y^{2L} \end{bmatrix}.$$

B. Bivariate Factorial Moment Matrix

A factorial moment matrix can be defined analogously to the power moment matrix by taking the expectation of the dyadic product $\underline{V}\underline{V}^T$ where $\underline{V} = [1, (X)_{s_1}^1, (X)_{s_1}^2, \ldots, (X)_{s_1}^L, (Y)_{s_1}^1, (Y)_{s_1}^2, \ldots, (Y)_{s_1}^L]^T$. The elements of the moment matrix which are subject to the maximum roundoff error are the $(L+1) \times (L+1)$ and $(2L+1) \times (2L+1)$ elements $[(X)_{s_1}^L]^2$ and $[(Y)_{s_2}^L]^2$, respectively. Below we give a better behaved factorial moment matrix whose roundoff error is dominated by $(X)_s^{2L}$, $s > 0$. When X and Y are prescaled to the interval $0 < X < 1$, $0 < Y < 1$, which we will do in the next section, the latter has much smaller roundoff error than $[(X)_s^L]^2$.

For a non-negative planar random vector (X, Y) and $s_1, s_2 \geq 0$ the $(2L+1) \times (2L+1)$ bivariate ascending factorial moment matrix is defined as the following expectation over X, Y

$$\mathbf{F}_{X,Y}^{2L+1}(s_1,s_2) = \tag{6}$$

$$E\left[\begin{array}{cccc|cccc} 1 & (X)_{s_1}^1 & \cdots & (X)_{s_1}^L & (Y)_{s_2}^1 & \cdots & (Y)_{s_2}^L \\ (X)_{s_1}^1 & (X)_{s_1}^2 & \cdots & (X)_{s_1}^{L+1} & (X)_{s_1}^1(Y)_{s_2}^1 & \cdots & (X)_{s_1}^1(Y)_{s_2}^{L+1} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\ (X)_{s_1}^L & (X)_{s_1}^{L+1} & \cdots & (X)_{s_1}^{2L} & (X)_{s_1}^L(Y)_{s_2}^1 & \cdots & (X)_{s_1}^L(Y)_{s_2}^L \\ & & & & (Y)_{s_2}^2 & \cdots & (Y)_{s_2}^{L+1} \\ & & (sym) & & \vdots & \ddots & \vdots \\ & & & & (Y)_{s_2}^{L+1} & \cdots & (Y)_{s_2}^{2L} \end{array}\right].$$

IV. Noise Mitigation via Eigendecomposition

Let $f_s(x,y)$ and $f_n(x,y)$ be two probability densities and consider the case where the observations X, Y are generated by the additive mixture model

$$f(x,y) = \beta f_s(x,y) + (1-\beta) f_n(x,y), \tag{7}$$

where $\beta \in [0,1]$. The mixture model arises in many applications [18], [19], [20]. Typically the density $f_n(x,y)$ is a known "nominal" or "noise" density, $f_s(x,y)$ is an unknown "signal" density of interest, and β plays the role of a proportional signal-to-noise ratio. Two such examples will be treated in section VI.

Assume that the L^2 power moments $\{\mu_n(i,j)\}_{i,j=1}^L$ (or equivalent factorial moments) of the "noise density" are known. Here we will present a method to exactly recover both the mixture parameter β and the L^2 power moments $\{\mu_s(i,j)\}_{i,j=1}^L$ of the "signal density" from the L^2 power moments $\{\mu_{X,Y}(i,j)\}_{i,j=1}^L$ of $f(x,y)$. This method relies on the symmetric non-negative definiteness of the PM matrix (5) and FMM matrix (6), which is established in Appendix B, the known Cholesky decomposition of the noise alone PM or FMM matrices, and the fact that, by construction, the (1,1) elements of the PM and FMM matrices are equal to 1.

We start from the definition (7). For any function $g(X,Y)$ we have the additive signal-plus-noise decomposition of the expectation

$$\begin{aligned} E[g(X,Y)] &= \beta E[g(X,Y)|\text{signal alone}] \\ &\quad +(1-\beta)E[g(X,Y)|\text{noise alone}] \end{aligned} \tag{8}$$

where $E[g(X,Y)|\text{signal alone}] = \int\int g(x,y) f_s(x,y)dxdy$ and $E[g(X,Y)|\text{noise alone}] = \int\int g(x,y) f_n(x,y)dxdy$.

Let $\mathbf{M}$, $\mathbf{M}_s$, and $\mathbf{M}_n$ denote the signal-plus-noise, signal-alone, and noise-alone moment matrices ($\mathbf{M}$, $\mathbf{M}_s$, and $\mathbf{M}_n$ denote either power moment or factorial moment matrices). From (8) we have the key result

$$\mathbf{M} = \beta \mathbf{M}_s + (1-\beta)\mathbf{M}_n$$

Now let $\mathbf{C}$ be the upper triangular Cholesky factor of $\mathbf{M}_n$, i.e. $\mathbf{M}_n = \mathbf{C}^T\mathbf{C}$, and define $\tilde{\mathbf{M}} = \mathbf{C}^{-T}\mathbf{M}\mathbf{C}^{-1}$. Then we have

$$\tilde{\mathbf{M}} = \beta\tilde{\mathbf{M}}_s + (1-\beta)\mathbf{I} \tag{9}$$

where $\mathbf{I}$ is the $(2L+1)\times(2L+1)$ identity matrix, and $\tilde{\mathbf{M}}_s = \mathbf{C}^{-T}\mathbf{M}_s\mathbf{C}^{-1}$ is the *whitened moment matrix.*

Except for the presence of the scalar mixture parameter β, (9) is in the form of a standard additive decomposition of a "measurement covariance" matrix $\tilde{\mathbf{M}}$ into "signal covariance" $\tilde{\mathbf{M}}_s$ plus "white noise covariance" $\mathbf{I}$. While the latter is full rank with $2L+1$ identical eigenvalues $\{1,\ldots,1\}$, the (ordered) eigenvalues of the signal matrix $\tilde{\mathbf{M}}_s$ are non-negative and typically fall off rapidly to zero. As an approximation we assume that the eigenvalues $\{\gamma_i^s\}_{i=1}^q$ of $\tilde{\mathbf{M}}_s$ are equal to zero for $i > q$, some q in the range $1 \leq q < 2L+1$.

Now from (9) the eigenvectors $\{\underline{\xi}_i\}_{i=1}^{2L+1}$ of $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{M}}_s$ are identical and hence the eigendecomposition of $\tilde{\mathbf{M}}$ is of the form

$$\begin{aligned} \tilde{\mathbf{M}} &= \sum_{i=1}^{2L+1} \gamma_i \underline{\xi}_i\underline{\xi}_i^T \\ &= \sum_{i=1}^{q} [\beta\gamma_i^s + (1-\beta)]\, \underline{\xi}_i\underline{\xi}_i^T + (1-\beta)\sum_{i=q+1}^{2L+1} \underline{\xi}_i\underline{\xi}_i^T \end{aligned}$$

Thus only the q largest eigenvalues $\gamma_i = \beta\gamma_i^s + (1-\beta)$ of $\tilde{\mathbf{M}}$ are related to the signal moments and the rest are pure noise eigenvalues $\{(1-\beta)\ldots,(1-\beta)\}$. In particular, at least if we know β *a priori*, then $\tilde{\mathbf{M}}_s$ can be exactly recovered from the eigendecomposition of $\tilde{\mathbf{M}}$ via

$$\tilde{\mathbf{M}}_s = \frac{1}{\beta}\sum_{i=1}^{q}[\gamma_i - (1-\beta)]\, \underline{\xi}_i\underline{\xi}_i^T. \tag{10}$$

However, it turns out that prior knowledge of β is not required. Indeed, since the $(1,1)$ element of $\mathbf{M}_s$ is equal to 1, and $\underline{e}_1^T\mathbf{C}\underline{\xi}_i = \underline{e}_1^T\underline{\xi}_i = \xi_{i1}$, we can determine β from the linear equation $\beta = \sum_{i=1}^q [\gamma_i - (1-\beta)]\,|\xi_{i1}|^2$ giving:

$$\beta = \frac{\sum_{i=1}^q \left[\gamma_i - |\xi_{i1}|^2\right]}{1-\sum_{i=1}^q |\xi_{i1}|^2}.$$

V. Scaling for Numerical Stability

It will be necessary to center and scale the variates X, Y to improve numerical stability of the computations. The least numerically stable step in the signal moment recovery method described above is computation of the Cholesky inverse $\mathbf{C}^{-1}$. Thus scaling should improve the conditioning of $\mathbf{M}$ ($\mathbf{P}_{X,Y}^{2L+1}$ and $\mathbf{F}_{X,Y}^{2L+1}$) while minimizing roundoff errors in the computation of its elements.

A. Scaling the Power Moment Matrix

When X and Y are contained in the interval $[x_{min}, x_{max}]$ and $[y_{min}, y_{max}]$, respectively, we can improve conditioning of $\mathbf{C}$ by centering and scaling X and Y into the interval $[-1,1]$. This forces $C_{ij} = 0$ for all odd off-diagonal entries ($i+j$ odd integer) and all other entries to be bounded in magnitude by 1. By making $\mathbf{C}$ more sparse its upper triangular Cholesky decomposition inherits sparseness and becomes better conditioned, see Fig. 1. The centering and scaling is done by making the transformations $\tilde{X} = \frac{X-(x_{max}+x_{min})/2}{\sigma_x}$ and $\tilde{Y} = \frac{Y-(y_{max}+y_{min})/2}{\sigma_y}$ where $\sigma_x = (x_{max}-x_{min})/2$ and $\sigma_y = (y_{max}-y_{min})/2$. After normalization the $(2L+1)\times(2L+1)$ power moment matrix $\mathbf{P}_{X,Y}^{2L+1}$ becomes

$$\tilde{\mathbf{P}}_{X,Y}^{2L+1} = \mathbf{P}_{\tilde{X},\tilde{Y}}^{2L+1}. \tag{11}$$

Note that $\tilde{\mathbf{P}}$ remains non-negative definite.

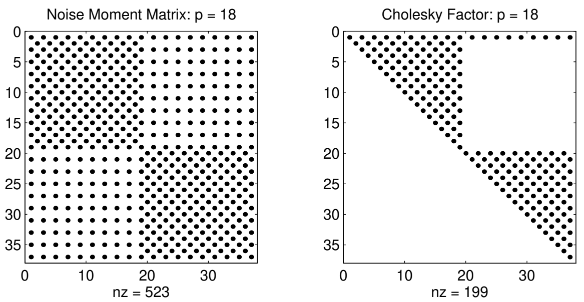

Fig. 1. The noise alone factorial moment matrix and its Cholesky factor have sparse structure which can be exploited to reduce roundoff error and explore structure of moment invariants.

B. Scaling the Factorial Moment Matrix

Since $\mathbf{F}_{X,Y}^{2L+1}$ is not defined for negative valued X, Y construction of an appropriately scaled factorial moment matrix requires more care than for the case of the power moment matrix. First of all we need center and scale the integer variates X and Y differently. Define the variates $\tilde{X}$ and $\tilde{Y}$ analogously to above except that σ_x and σ_y are replaced by

$$\sigma_x = \exp\left\{\frac{1}{2L}\sum_{k_1=1}^{2L}[(x_{max}-x_{min})/2+(k_1-1)s_1]\right\}$$

and similarly for σ_y. This scaling ensures that the maximum magnitude entry of $\mathbf{F}_{X,Y}$ will be bounded by 1. Next, for any real number $T = T^+ - T^-$, we define T^+ as the positive part of T and T^- as the absolute value of the negative part of T. Define the $(2L+1)\times(2L+1)$ normalized spatial factorial moment matrix

$$\tilde{\mathbf{F}}_{X,Y}^{2L+1} = \mathbf{F}_{\tilde{X}^+,\tilde{Y}^+}^{2L+1}(1/\sigma_f) + \mathbf{D}\,\mathbf{F}_{\tilde{X}^-,\tilde{Y}^-}^{2L+1}(1/\sigma_f)\,\mathbf{D}^T \quad (12)$$

where $\mathbf{D} = \text{diag}_{k=0,\ldots,2L}((-1)^k)$ is a diagonal matrix with alternating $+1$ and -1 along the diagonal. The non-negative definiteness of $\mathbf{F}_{X,Y}^{2L+1}$ is obviously preserved under the transformation (12) since $\tilde{\mathbf{F}}_{X,Y}^{2L+1}$ is the sum of two non-negative definite matrices.

The step sizes s_1 and s_2 need to be selected so as to minimize the spread between the maximum and minimum eigenvalues of $\tilde{\mathbf{F}}_{X,Y}^{2L+1}$ thereby minimizing the condition number of the Cholesky factor$\mathbf{C}$. Note that as s_1 and s_2 approach zero $\tilde{\mathbf{F}}_{X,Y}^{2L+1}$ becomes equivalent to $\mathbf{P}_{X,Y}^{2L+1}$ and the condition of $\mathbf{C}$ becomes identical for the PM and FMM approaches. We have observed from experiments that the condition number of $\mathbf{C}$ improves as s_1, s_2 increases over $[0, 10]$. However, it was also observed that the roundoff error in computing the elements of $\tilde{\mathbf{F}}_{X,Y}^{2L+1}$ also increases for large values of s_1 and s_2. The choices $s_1 = 2/(x_{max}-x_{min})$ and $s_2 = 2/(y_{max}-y_{min})$ appear to achieve a good compromise for a wide variety of values of L.

VI. Applications

For illustration, we apply the power moment and factorial moment matrix analysis to two applications areas: wordspotting in noise degraded documents and modulation classification in digital communications.

A. Wordspotting Example

Given a group of documents, it is frequently desirable to know whether any of the documents contain a certain word or set of words [21], [1]. It would be useful if this process could be automated and work reliably on documents with different fonts, font sizes, and noise contamination, e.g. in faxed documents. here we illustrate the moment matrix methodology for a simple wordspotting example.

A.1 Experiments

We obtained an electronic copy (ascii) of Thomas Payne's Common Sense, segmented the ascii text, and generated postscript versions of each word in Helvetica 48 point font. homogeneous salt and pepper noise were added modulo-2 to the bitmaps of each word. Note that modulo-2 addition produces noise which is not strictly additive or linear. Raw moments of various mixed orders were computed empirically and sample power and sample factorial moments matrices were constructed using Matlab 4.0. Note that the number of pixels, or window size, for each word depends on the number of letters in the word, the presence of capitalization, punctuation, etc. To standardize the computation the bitmap coordinates for each word were scaled to a square of length 1 on a side. The Cholesky factor $\mathbf{C}$ of the spatially homogeneous noise moment matrix was applied to prewhiten the empirical word moment matrix. An eigendecomposition was performed on the prewhitened empirical moment matrix, with the signal subspace dimension determined by a threshold rule, and the original noiseless moment matrix was recovered by eliminating the noise subspace and renormalization, as discussed in the previous section of this paper.

A representative example of the noiseless and noise degraded bitmaps moment matrices is shown in Figs. 2-4 for SNR $(\beta) = 0.5$. It was found that $L = 23$ could be used in the factorial moment matrix without running into run time errors due to numerical roundoff. This is to be compared to the upper limit of $L = 20$ encountered for the power moment matrix computations. Note that this 15% increase in L translates into an over 30% increase in the number of distinct mixed raw moments that can be used for discrimination ($23^2 - 1 = 528$ as compared to $20^2 - 1 = 399$). The noiseless whitened power moment matrices appear to better distinguish between the two words as compared to the noiseless factorial matrices. However the noisy factorial moment matrices are observed to be much less sensitive to the noise contamination. Our results suggest that the whitened moment matrices may themselves be an excellent feature space for robust discrimination between words.

We defined a simple discriminant based on computing the mean square distances between noisy moment matrix to the corresponding noiseless moment matrices for "will" or "War." The discrimination was implemented using a minimum distance decision rule. Three classes of discriminants were compared: (D1) mean square distances between the raw empirical moment matrix $\mathbf{M}$ and the noiseless moment matrices $\mathbf{M}_s$ for the two words; (D2) mean square distances between the recovered prewhitened signal moment matrices $\tilde{\mathbf{M}}_s$ and noiseless prewhitened moment matrices $\mathbf{C}^{-T}\mathbf{M}_s\mathbf{C}^{-1}$ for the two words; (D3) the mean square distances between the recovered signal moment matrix $\hat{\mathbf{M}}_s = \mathbf{C}^T\tilde{\mathbf{M}}_s\mathbf{C}$ and the noiseless moment matrices $\mathbf{M}_s$ for the two words. In each case a mask was used to screen out elements of the moment matrices which were not substantially different from "will" to "War."

The probability of decision error for each of discriminators rules is shown in Figs. 5 and 6 as a function of SNR. For unwhitened moment matrix discriminators D1 and D3 we compare probability of error using only lower order moments extracted from the noisy and recovered 37×37 moment matrices, respectively. The number used range from 3 moments $(\mu_{X,Y}(1,0), \mu_{X,Y}(1,1), \mu_{X,Y}(0,1))$, denoted by $L=1$ in the figures, to 99 moments $(\mu_{X,Y}(i,j), i,j=0,\ldots,9, i,j$ not simultaneously equal to zero), denoted by $L=9$ in the figures. The performance of D1 is uniformly worse than that of D3 for all L values. Note that the use of more moments in the raw moment discriminant D1 actually degrades discrimination performance. This is consistent with the well known variance increase in estimation of higher order statistics as the order increases [22]. Interestingly, the opposite trend is observed in the whitened moment discriminant D3 where variance reduction has been acheived in the higher order moments via the subspace eigendecomposition of the raw 37×37 moment matrix. Finally, as expected, note that D2 attains very low probability of error by using minimum distance discrimination directly in the whitened moment matrix domain. We suspect that the reason that D3 is incapable of matching the excellent performance of D2 is due to poor condition number of the Cholesky factor $\mathbf{C}$ used to recover the cleaned raw moment matrix from the cleaned whitened moment matrix.

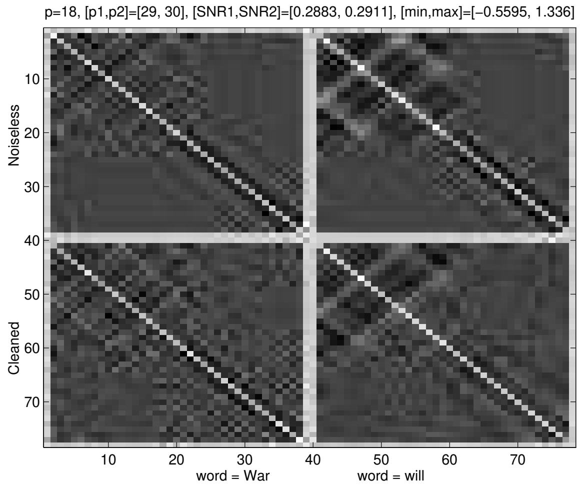

Fig. 3. Top: Noiseless whitened power moment matrices for words *will* and *War* for $L=18$. Bottom: denoised empirical moment matrix from noisy vesrions of words shown in Fig 2. Note that distinctive features of noiseless whitened power moments are recovered after denoising.

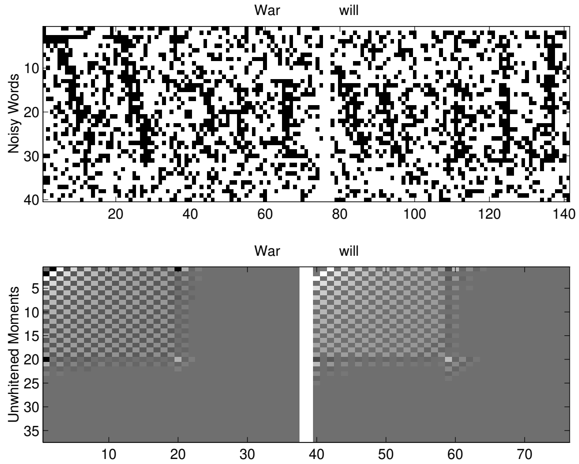

Fig. 2. Top: the words *will* and *War* in helvetica font and severe additive noise. Bottom: the unwhitened power moment matrices empirically estimated from the data strongly resemble noise alone moment matrices and no distinctive features are evident between words.

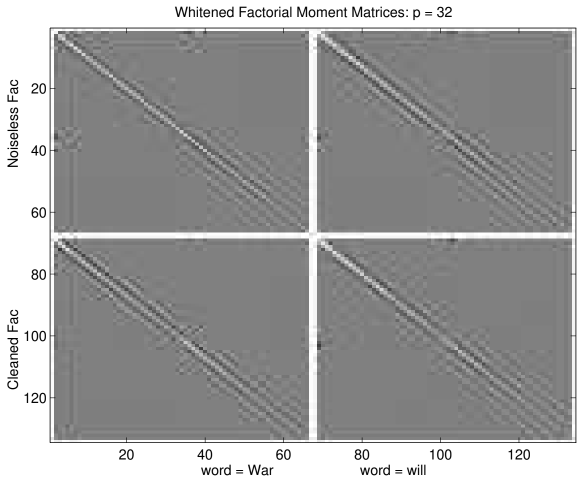

Fig. 4. Same as Fig. 3 except that $L=32$ factorial moment matrix is shown.

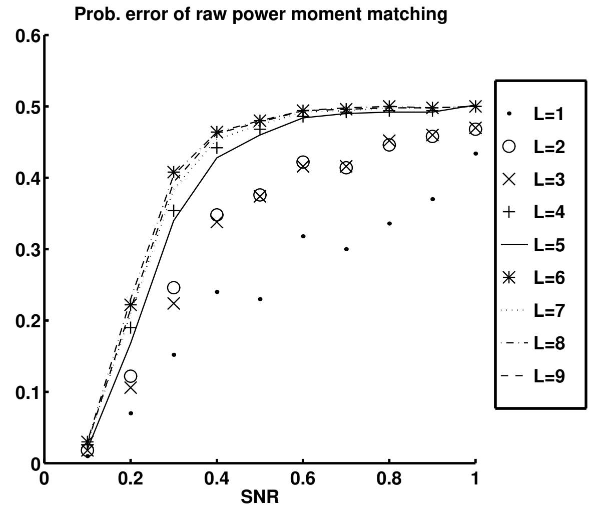

Fig. 5. Probability of error curves for D1 = minimum distance decision rule based on raw moments. Here we are matching different numbers of empirical moments to corresponding noiseless moments obtained from Fig. 2 without additive noise. SNR is the relative number of bit flip errors as compared to active pixels in noiseless bitmap.

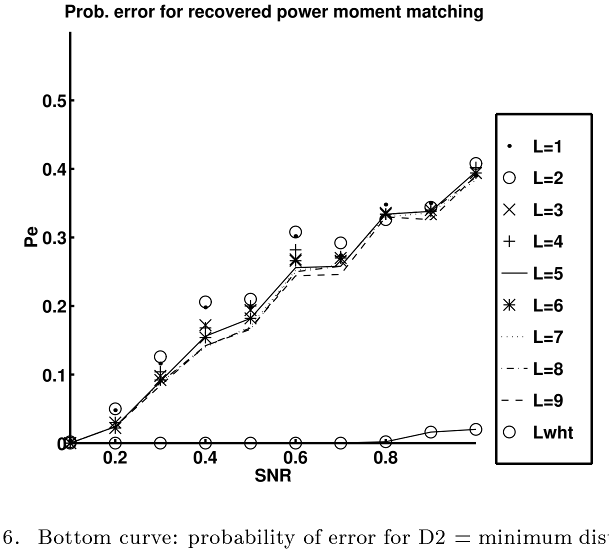

Fig. 6. Bottom curve: probability of error for D2 = minimum distance decision rule based on moment matching in denoised prewhitened moment matrix domain. Here we are matching the empirical whitened moment matrices to noiseless whitened moment matrices shown in Fig. 3. Upper curves: same for D3 = minimum distance decision rule based on moment matching in recovered moment matrix domain.

B. Modulation Classification Example

A typical RF receiver uses an in-phase and quadrature (IQ) detector. The transmitter produces a carrier modulated signal $s(t)$ at carrier frequency f_o Hz. The transmitted signal passes through a noisy bandpass channel before arriving at the receiver as the noisy bandpass signal $w(t) = s(t) + n(t)$. The in phase (w_I) and quadrature (w_Q) baseband components are detected separately by two orthogonal frequency mixers, each centered at frequency $w_o = 2\pi f_o$ radians, each followed by an integrator. These components are each sampled at integer multiples of the symbol period to produce a sequence of complex measurements $y_I(t_i)+jy_Q(t_i)$, $i = 1, \ldots, N$. In the absence of noise and intersymbol interference the N points in the complex IQ plane specify a complex signal constellation which depends on the type of modulation used.

Automated detection and classification of modulation type is an important problem arising in non-cooperative communications environments, surveillance, and multi-user networks. A wide variety of techniques have been proposed for constant and non-constant envelope signals such as M-ary PSK, FSK and QAM including: zero crossing analysis [23], likelihood ratio approximation [24], [25], wavelet expansions [26], higher order correlations [27], and moment methods [28]. Most classification methods, including those cited above, are applied to the in-phase and quadrature (IQ) components of the received bandpass communications signal which can be plotted in the complex plane.

Here we apply the moment matrix classifier methodology to classification of modulation types for M-ary PSK, FSK and QAM in the IQ plane in the presence of noise, unknown phase angle, and unknown magnitude. More detailed coverage of this method can be found in [29]. Our application of moment matrices can be viewed as a generalization of the approach of [28], which was based solely on the eighth order phase moment. In particular, the use of moment matrices allows us to use linear combinations of a large number of different orders (more than 100) of joint phase and magnitude moments in addition to providing a denoising procedure for extracting signal-alone joint moments from noise contaminated measurements.

The IQ plane is composed of $2^B \times 2^B$ pixels, where B is number of bits of the A/D converter which samples and quantizes the output of the baseband IQ signal. Therefore, unlike the relatively low spatial resolution wordspotting example, for a 16 bit A/D the number of pixels is huge. To reduce computation, it will be more convenient to down sample the IQ plane to fewer pixels. As each downsampled pixel can contain multiple samples, the downsampled IQ plane produces a Gray scale image for which the three dimensional representation (X, Y, Z) (2 spatial coordinates and one grey scale coordinate) described in the previous Section and Appendix C will be adopted.

B.1 Polar Power Moments

In addition, communications signals in the IQ plane have special structure that can be exploited: the IQ signal samples are always distributed symmetrically about the origin $(0, 0)$. In this subsection we represent the IQ samples in polar coordinates relative to the known origin in order to express rotation in cartesian coordinates as angular-translation in polar coordinates. By so doing it is a simple matter to enforce rotation invariance via retaining magnitude-only information of the higher order moments, inducing discrimination invariance with respect to unknown phase.

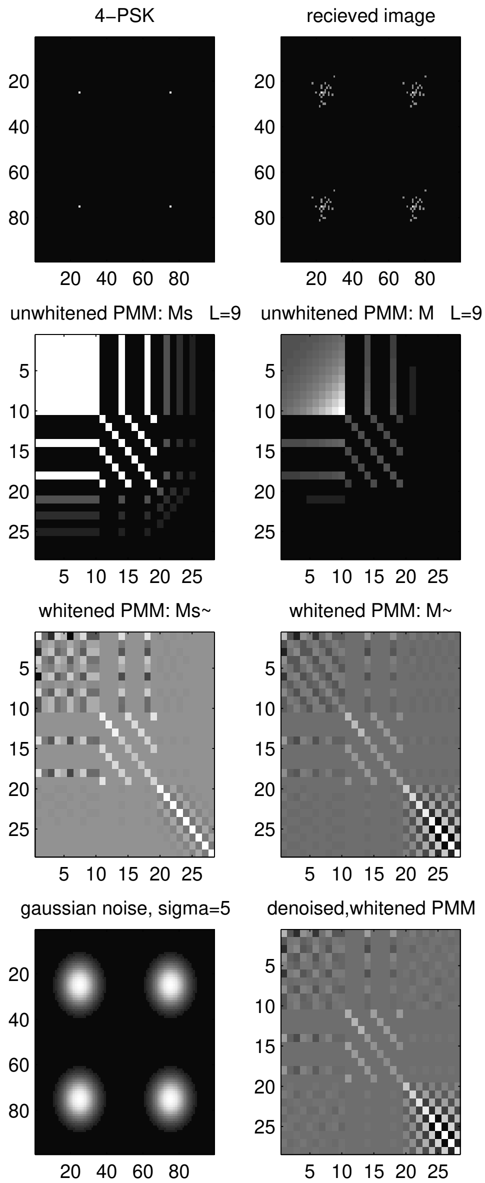

Fig. 7. First row: signal alone (left) and noise contaminated IQ images for 4-PSK, second row: unwhitened polar PMMs, third row: whitened polar PMMs, fourth row: spatial distributuion of the noise (left) and the denoised, whitened polar PMM.

Define complex moments of the phase-magnitude polar representation of the IQ image

$$m^{p,q,s}_{r,\theta,z} = E\left[r^p \, e^{j\theta q} z^s\right]$$

where r, θ denotes magnitude and phase of a pixel location, and z denotes the grey level of the image. In order to make these moments rotation invariant, the magnitude squared of the complex moments $\left|m^{p,q}_{r,\theta}\right|^2$ can be used in the moment matrix. Specifically, define $\mathbf{M}$ as the $(3L+1)\times(3L+1)$ non-negative definite Hermitian matrix of complex moments $\mathbf{M} = ((m^{p,q,s}_{r,\theta,z}))_{p,q,s}$ where $p \in \{0,\ldots,L\}$ and $q,s \in \{1,\ldots,L\}$.

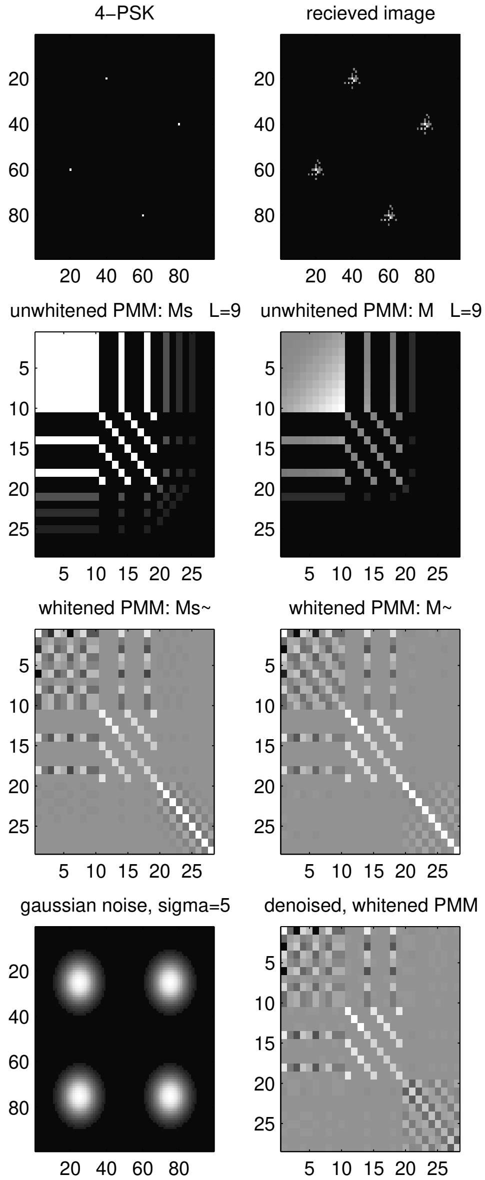

Fig. 8. The order of the panels is the same as in figure 7. Note that there is now a mismatch between the model noise and the actual noise vaiance of the received signal. Note further the rotation invariance of polar PMM.

Then the rotation invariant moment matrix is defined by the direct product matrix $\mathbf{M}^2 = \mathbf{M}^H \circ \mathbf{M}$, where $\circ$ denotes element by element multiplication. Note that $\mathbf{M}^2$ is symmetric and non-negative definite since the direct product preserves non-negative definiteness.

B.2 Experiments

The moment matrix techniques described in the previous section were implemented in the complex IQ plane. Here the number of symbols transmitted is $N = 25$. Note that the channel noise produces a spatial blurring of the signal constellation in the IQ plane, i.e. the noise in the IQ plane is not spatially homogeneous and is signal dependent. For the experiments below we fixed an assumed signal modulation (4-PSK at unit power) and noise variance ($\sigma^2 = 25$) and generated the Cholesky factors of the associated moment matrix. This Cholesky factor was then used to "whiten" the empir-

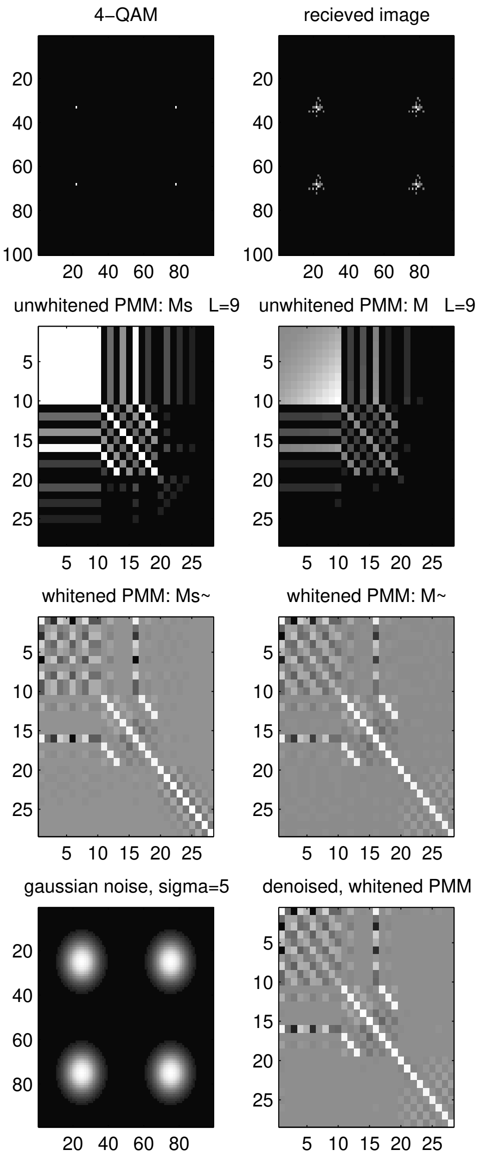

Fig. 9. First row: signal alone (left) and noise contaminated IQ images for 4-QAM, second row: unwhitened polar PMMs, third row: whitened polar PMMs, fourth row: spatial distributuion of the noise (left) and the denoised, whitened PMM. Here the mismatched Cholesky factor is used, i.e. the Cholesky factor corresponding to 4-PSK, to whiten the PMM for 4-QAM.

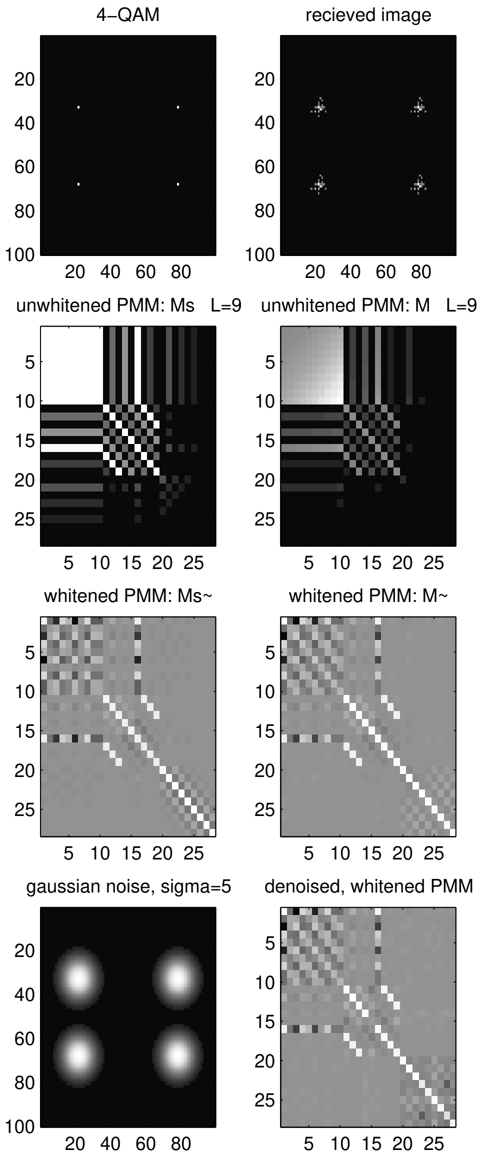

Fig. 10. The order of the panels is the same as in figure 9. Here, the 4-QAM cholesky factor is used. There is hardly any difference between the moment matrices of figure 9 and the moment matrices in this figure.

ically calculated moment matrices for each of the measured signal types. Note that in each of the experiments described below this whitening is mismatched to the actual noise distribution in the IQ plane. Nonetheless, we will show that an improvement in modulation discrimination results even with a moderate amount of mismatch.

Figures 7 and 8 illustrate the IQ images and moment matrices for the cases that $s(t)$ is a 4-PSK signal with noise power of $\sigma^2 = 25$, while Figs. 9 and 10 are for 4-QAM with $\sigma^2 = 25$). While the 4-QAM constellation shown is not desirable due to its poor minimum distance properties, such a constellation can arise when the amplification factors of the I and Q stages of the transmitter become uncalibrated due to parameter drift. This constellation was chosen for these experiments due to its close similarity to 4-PSK, making discrimination particularly difficult. The first column of each of the figures 7 and 8 are, from top to bottom: the signal alone represented in the IQ plane, the unwhitened power moment matrix of the signal, the whitened power moment matrix of the signal, and the assumed spatial distribution of the noisy measurements used to calculate the Cholesky factor of the moment matrix. The second column of these figures are, from top to bottom: the noise contaminated signal in the IQ plane, the unwhitened moment matrix estimated from the $N = 25$ noisy measurements, the estimated whitened moment matrix, and the resultant cleaned power moment matrix after noise subspace processing.

By comparing the signal alone unwhitened moment matrix to that of the noise contaminated signal it is evident that the raw moments are quite sensitive to additive noise (compare left and right panels in the second rows of Figs. 7-10). On the other hand, by comparing the left and right panels in each of the third rows of the figures, it is evident that the prewhitened moment matrices are much less sensitive to noise. This can be explained by the fact that, as the variables

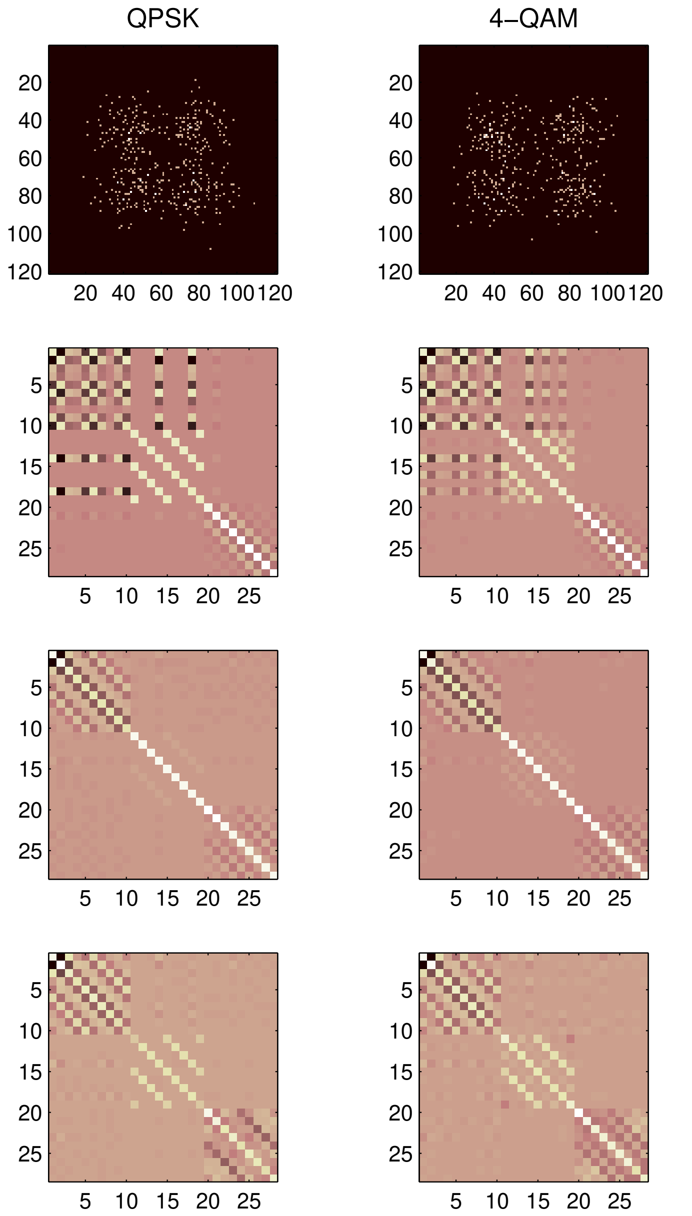

Fig. 11. From top to bottom: Noisy IQ-images, whitened noiseless $\mathbf{M}^2$s, whitened noisy $\mathbf{M}^2$s, denoised $\mathbf{M}^2$s.

$\underline{x} = (r, \theta, z)$ are prenormalized to the interval $[-1, 1]$, the higher order moments pairs $E[r^p e^{j\theta q}]$, $E[e^{j\theta q} z^s]$ and $E[r^p z^s]$ are exponentially decreasing to zero as p, q, s become large. The prewhitening of the moment matrix via Cholesky decomposition rescales all of these moments to produce entries of comparable magnitudes and thus all high and low order moments are put on equal footing. This matrix rescaling can be interpreted as a generalization of variance normalization, such as those used to produce correlation coefficients and spectral coherence functions, which have been widely used to study dependencies in two or more random variables of greatly different average magnitudes.

It is noteworthy that even with noise mismatch the whitened and cleaned moment matrices provide a stable representation of any specified noise contaminated modulation (note similarity of 3rd rows of Figs. 7 and 8 for 4-PSK with two different received noise powers) while they provide a good degree of discrimination between different modulation formats (note dissimilarity of 3rd rows of Figs. 7 and 9). Note also that, by construction, the moment representations are invariant to rotation and scale of the signal constellation due to unknown carrier phase angle or signal amplitude. Figure 10 shows the moment matrices for the case of 4-QAM where the exact constellation noise distribution, shown in the lower left hand panel, is used to construct the matrix in the Cholesky decomposition.

In Fig. 11 the additive noise level has been increased to $\sigma_n^2 = 81$ which, as can be seen from the similarities between the polar moment matrices, stretches the discrimination capability of the rotation invariant moment method to its limit. In this high noise regime, the imposition of rotation invariance has effectively robbed the moments of their ability to accurately discriminate between the QPSK and QAM signals. Thus for operation in high noise environments the computationally simple magnitude squared of complex moments is not effective. In the next section we introduce a discriminant based on the outputs of a bank of non-invariant moment matrix computers, each matched to a particular IQ rotation. This will be called the moment matrix filter-bank classifier. As will be seen this produces a more sensitive, but also more computationally intensive, discriminant for low SNR situations.

B.3 Probability of Classification Error

Fig. 12 shows the probability of correct classification for the moment matrix filter-bank classifier as compared to two popular methods of modulation classification for discriminating between BPSK and QPSK signals. The curves are indexed by a function of the symbol-carrier-to-noise ratio (SCNR) defined as

$$\gamma_s = \frac{A^2 T_s}{2\sigma_n^2}.$$

where A^2 is amplitude of the received signal envelope, T_s is the symbol period, and σ_n^2 is the variance of the additive Gaussian noise in the signal passband. Shown in Fig. 12 are correct classification probability curves for the eighth order phase moment discriminant of [28], labeled PMC in the figure, the log-likelihood function classifier of [24], labeled q_2, and the power moment matrix discriminant, labeled PMM, implemented with $L = 2$, i.e. a 7×7 moment matrix. Both coherent and incoherent receivers were simulated, the curves for the coherent system (known IQ rotation) are subscripted cs while the curves for the incoherent system (unknown IQ rotation) are subsripted ns. The curve for PMC corresponds to the coherent case - as proposed in [28] the PMC method is not implementable for non-coherent systems. The performance curves for q_2^{cs}, q_2^{ns} and PMC are directly extracted from the simulations shown in [24, Fig. 4]. The PMM method was implemented by computing the Mahalanobis distance between each moment matrix filter-bank output and the associated noiseless moment matrix template. The filter bank was comprised of 12 filters performing empirical and theoretical moment matrix computations as described in Sections III and IV. Each of these moment matrices were whitened using the Cholesky decomposition of a noise matrix corresponding to one of 12 rotations of the IQ plane by increments of approx 15° (for more details see [29]. In all cases, the number of samples N in the IQ plane is equal to 100 and the number of experiments is equal to 1000. Observe from the probability of correct classification curves in Fig. 12 that the PMM method performs almost as well as the LLFC method and significantly outperforms the PMC method for the case of coherent reception. For non-coherent reception the PMM method attains a probability of correct classification that is only slightly poorer than in the coherent case and is significantly better than the LLFC.

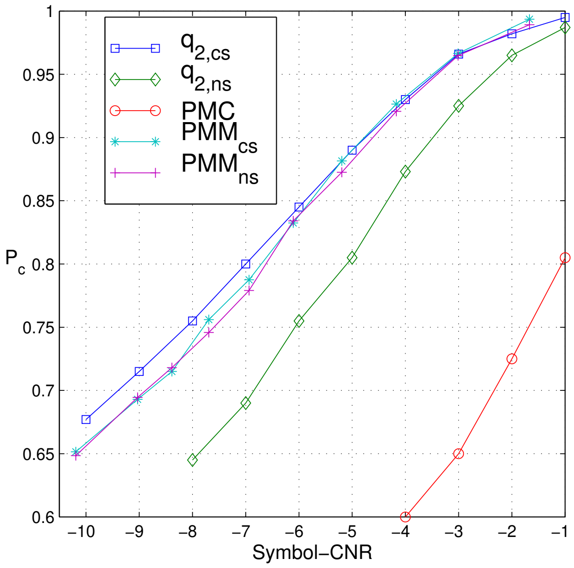

Fig. 12. Performance Comparison

VII. Conclusion

In this report a new method for performing feature classification in multi-dimensional spatial feature space has been described. The method is based on selecting principal components from a Cholesky-whitened symmetric non-negative definite matrix of higher order moments in feature space. Both power moments and factorial moments were introduced, the latter having more stable numerical properties for higher orders. Two very different applications were investigated to illustrate the application of these methods. In the wordspotting application the use of principal components was seen to counterbalance the increased variance inherent to the high dimensional moment matrix discriminant. By performing discrimination directly on the principal components of the whitened moment matrix substantial improvement in probability of error was observed relative to the standard unwhitened raw moment discriminants. In the digital communications example, application of these moment matrix methods to the IQ complex plane produced a low complexity modulation discriminant which outperforms the well known maximum likelihood approach of Polydoros for incoherent receivers.

Appendix A

Lemma 1: The set of L^2 power moments $\{\mu_{X,Y}(k_1,k_2)\}_{k_1,k_2=1}^L$ defined in (1) and the set of L^2 factorial moments $\{\delta_{X,Y}(k_1,k_2;s_1,s_2)\}_{k_1,k_2=1}^L$ defined in (2) are equivalent.

Proof

The proof consists of deriving a 1-1 transformation between each of the $(L+1)\times(L+1)$ matrices of moments $\mathrm{M}_{X,Y}^L = ((\mu_{X,Y}(k,l)))_{kl=0}^L$, $\mathrm{D}_{X,Y}^L = ((\delta_{X,Y}(k,l,s_1,s_2)))_{kl=0}^L$. It is easily shown that

$$(x)_s^k = \sum_{l=0}^k d_k^{(l)} x^l,$$

where the $d_k^{(l)}$ are coefficents, related to the *Stirling numbers* [17], which obey the recursions

$$d_{k+1}^{(l)} = d_k^{(l-1)} + skd_k^{(l)}, \quad l=1,\ldots,k+1, \quad k=1,2,\ldots$$

$$d_1^{(k)} = \begin{cases} 1, & k=1 \\ 0, & o.w. \end{cases}$$

Note that when $s>0$ the coefficients $d_k^{(l)}$ are non-negative.

In particular, we have the relations

$$\begin{bmatrix} (x)_s^L \\ \vdots \\ (x)_s^0 \end{bmatrix} = \begin{bmatrix} 1 & d_L^{(L-1)} & \cdots & \cdots & d_L^{(0)} \\ 0 & 1 & d_{L-1}^{(L-2)} & \cdots & d_{L-1}^{(0)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & d_1^{(0)} \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} x^L \\ \vdots \\ x^0 \end{bmatrix} \quad (13)$$

Let the $(L+1)\times(L+1)$ upper triangular matrix in (13) be denoted $\mathbf{U}$. This matrix is obviously invertible since it has ones along its diagonal. Now consider the expectation of the outer product of the vectors $[(X)_s^L,\ldots,(X)_s^0]$ and $[(Y)_s^L,\ldots,(Y)_s^0]$ which is equal to the matrix of factorial moments $\mathrm{D}_{X,Y}^L$. In view of relation (13) we have the identity

$$\mathrm{D}_{X,Y}^L = \mathbf{U}\mathrm{M}_{X,Y}^L\mathbf{U}^T$$

relating the $(L+1)\times(L+1)$ matrices $\mathrm{D}_{X,Y}^L$ and $\mathrm{M}_{X,Y}^L$. Since $\mathbf{U}$ is invertible we have the inverse mapping $\mathrm{M}_{X,Y}^L = \mathbf{U}^{-1}\mathrm{D}_{X,Y}^L\mathbf{U}^{-T}$ which establishes that there exists a 1-1 transformation between the power moments and the factorial moments. □

Appendix B

The power moment matrix $\mathbf{P}_{X,Y}^{2L+1} = E[\underline{U}\,\underline{U}^T]$ is the expectation of a rank one outer product where $\underline{U} = [1,X,\ldots,X^L,Y,\ldots,Y^L]^T$. Thus $\mathbf{P}_{X,Y}^{2L+1}$ is obviously a non-negative definite matrix. Likewise for the outer product factorial moment matrix $\mathbf{P}_{(X),(Y)}^{2L+1} = E[\underline{V}(s_1)\underline{V}^T(s_2)]$, where $\underline{V} = [1,(X)_{s_1}^1,\ldots,(X)_{s_1}^L,(Y)_{s_2}^1,\ldots,(Y)_{s_2}^L]^T$.

We next turn to the ascending factorial moment matrix $\mathbf{F}_{X,Y}^{2L+1}(s_1,s_2)$. Note that while the power moment matrix is constructed as the expectation of a rank one outer product, i.e. it is a covariance-type matrix, the factorial moment matrix is not expressible in this way. Thus the proof of non-negative definiteness is more involved than before.

Note that for positive real m and $|s|=1$ the factorial moment $(m)_s^k = m(m+s)\ldots(m+s(k-1))$ can be expressed in the form

$$(m)_s^k = \begin{cases} \frac{\Gamma(m+k)}{\Gamma(m)}, & s=1 \\ \frac{\Gamma(m-k)}{\Gamma(m)}, & s=-1 \end{cases}, \quad (14)$$

where

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \tag{15}$$

is Euler's (complete) Gamma function ($\Gamma(n+1) = n!$ for n a non-negative integer).

Lemma 2: Let $m > 0$ and fix an integer $L \geq 1$. The following $(L+1)\times(L+1)$ Hankel matrix is positive definite

$$\mathbf{G}(m) \stackrel{\text{def}}{=} \begin{bmatrix} \Gamma(m) & \cdots & \Gamma(m+L) \\ \vdots & \Gamma(m+i+j) & \vdots \\ \Gamma(m+L) & \cdots & \cdots \Gamma(m+2L) \end{bmatrix}$$

Proof:

Let $\underline{z} = [z_0, \ldots, z_L]^T$ be an arbitrary real vector. Consider the following function

$$f(z) = \int_0^\infty \left(q(t) \sum_{i=0}^L z_i t^i \right)^2 dt, \tag{16}$$

where

$$q(t) = \sqrt{t^{m-1} e^{-t}}.$$

Since $q(t) > 0$ and $g_i(t) = t^i$, $i = 0, \ldots, L$, is a linearly independent set of functions: $f(z) > 0$ for $\underline{z}$ not identically zero. The right hand side of (16) expands to the form

$$\sum_{i,j=0}^L z_i z_j \int_0^\infty t^{m+i+j-1} e^{-t} dt = \underline{z}^T \mathbf{G}(m) \underline{z},$$

where the Euler integral representation of the Gamma function (15) has been identified. Combining the positivity of (16) with the above we have $\underline{z}^T \mathbf{G}(m) \underline{z} > 0$ which establishes the lemma. □

For $0 \leq x < \infty$ and $s \geq 0$ define the $(2L+1)\times(2L+1)$ Hankel matrix

$$\mathbf{A}(x,s) = \begin{bmatrix} 1 & (x)_s^1 & \cdots & \cdots & (x)_s^L \\ (x)_s^1 & (x)_s^2 & \cdots & \cdots & (x)_s^{L+1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ (x)_s^L & (x)_s^{L+1} & \cdots & \cdots & (x)_s^{2L} \end{bmatrix} \tag{17}$$

$$= \begin{bmatrix} 1 & \underline{a}^T(x,s) \\ \underline{a}(x,s) & \mathbf{H}(x,s) \end{bmatrix} > 0 \tag{18}$$

where $\mathbf{H}(x,s) \stackrel{\text{def}}{=} \left((x)_s^{i+j}\right)_{i,j=1,\ldots,L}$ is an $L\times L$ Hankel matrix and $\underline{a}(x,s) = [(x)_s^1, \ldots, (x)_s^L]^T$ is an L-element vector.

Lemma 3: For $x, s \geq 0$ the matrices $\mathbf{A}(x,s)$, $\mathbf{H}(x,s)$ and $\mathbf{H}(x,s) - \underline{a}(x,s)\underline{a}^T(x,s)$ are all non-negative definite. They are positive definite for $x > 0$.

Proof: The case $s = 0$ is quickly disposed of since for this case $\mathbf{A}(x,s)$ reduces to the outer product $\underline{U}\underline{U}^T$ where $\underline{U} = [1, x, \ldots, x^L]^T$. When $x = 0$, $\mathbf{A}(x,s)$ is non-negative definite since it is a matrix of zeros except for a 1 at the (1,1) element. We henceforth assume $x, s > 0$. Let $\mathbf{S} \stackrel{\text{def}}{=} \text{diag}_{k=0,\ldots,L}(s^{k/2})$ and let $m \stackrel{\text{def}}{=} x/s$. Since $(x)_s^k = x(x+s)\ldots(x+(k-1)s) = s^k(m)(m+1)\ldots(m+k-1)$

$$\begin{aligned} \mathbf{A}(x,s) &= \mathbf{S}^T \mathbf{A}(m,1) \mathbf{S} \\ &= \mathbf{S}^T \mathbf{G}(m) \mathbf{S} \, \frac{1}{\Gamma(m)} \end{aligned}$$

where $\mathbf{G}(m)$ is the positive definite matrix defined in Lemma 2. Thus $\mathbf{A}(x,s)$ is positive definite. The positive definiteness of $\mathbf{H}(x,s)$ and $\mathbf{H}(x,s) - \underline{a}(x,s)\underline{a}^T(x,s)$ follows immediately from [30, Thm. 7.7.6] and the partition representation (18) of $\mathbf{A}(x,s)$. □

The $(L+1)\times(L+1)$ univariate factorial matrix $\mathbf{F}_{X,X}^{L+1}(s)$ is the statistical expectation over X of $\mathbf{A}(X,s)$. Note that it follows from Lemma 2 and the assumption $E[X^{2L}] < \infty$ that all elements of $\mathbf{F}_{X,X}^{L+1}(s)$ exist and are finite. Hence we have

Proposition 1: Fix an integer $L \geq 1$ and assume $X \geq 0$, $0 < E[X^{2L}] < \infty$, and $s \geq 0$. Then the univariate factorial moment matrix $\mathbf{F}_{X,X}^{L+1}(s)$ is finite and non-negative definite.

Next we use Lemma 3 to establish the same result for the bivariate factorial matrix $\mathbf{F}_{X,Y}^{2L+1}(s_1,s_2)$ which is the statistical expectation over X and Y of the $(2L+1)\times(2L+1)$ matrix

$$\mathbf{M}_{X,Y} = \begin{bmatrix} 1 & \underline{a}^T(X,s_1) & \underline{a}^T(Y,s_2) \\ \underline{a}(X,s_1) & \mathbf{H}(X,s_1) & \underline{a}(X,s_1)\underline{a}^T(Y,s_2) \\ \underline{a}(Y,s_2) & \underline{a}(Y,s_2)\underline{a}^T(X,s_1) & \mathbf{H}(Y,s_2) \end{bmatrix}. \tag{19}$$

Proposition 2: Fix an integer $L \geq 1$ and assume $X, Y \geq 0$, $0 < E[X^{2L}], E[Y^{2L}] < \infty$, and $s_1, s_2 \geq 0$. Then the bivariate factorial moment matrix $\mathbf{F}_{X,Y}^{2L+1}(s_1,s_2)$ is finite and non-negative definite.

Proof:

When either $X = 0$ or $Y = 0$ (w.p.1), $\mathbf{M}_{X,Y}$ is equivalent to a block diagonal matrix equivalent to $\mathbf{A}$, defined in (17), plus zero padding and the proposition follows by Lemma 3. Thus we assume that $X, Y > 0$ (w.p.1) in what follows. We first show the positive definiteness of the $2L\times 2L$ lower block of $\mathbf{M}_{X,Y}$

$$\mathbf{Q} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{H}(X,s_1) & \underline{a}(X,s_1)\underline{a}^T(Y,s_2) \\ \underline{a}(Y,s_2)\underline{a}^T(X,s_1) & \mathbf{H}(Y,s_2) \end{bmatrix}.$$

To simplify notation denote $\mathbf{H}(X,s_1)$ by $\mathbf{H}_x$, $\underline{a}(X,s_1) = \underline{a}_x$, etc. Note that by Lemma 3 $\mathbf{H}_y$ is positive definite. By [30, Thm. 7.7.6] positive definiteness of $\mathbf{Q}$ will follow from positive definiteness of $\mathbf{H}_x - \underline{a}_x\underline{a}_y^T\mathbf{H}_y^{-1}\underline{a}_y\underline{a}_x^T$. But again by Lemma 3 $\mathbf{A}_y$ is positive definite and hence $1 - \underline{a}_y^T\mathbf{H}_y^{-1}\underline{a}_y > 0$. Hence $\mathbf{H}_x - \underline{a}_x\underline{a}_y^T\mathbf{H}_y^{-1}\underline{a}_y\underline{a}_x^T > \mathbf{H}_x - \underline{a}_x\underline{a}_x^T > 0$.

Now

$$\mathbf{Q} - \begin{bmatrix} \underline{a}_x \\ \underline{a}_y \end{bmatrix} [\underline{a}_x^T, \underline{a}_y^T] = \begin{bmatrix} \mathbf{H}_x - \underline{a}_x\underline{a}_x^T & \mathrm{O} \\ \mathrm{O} & \mathbf{H}_y - \underline{a}_y\underline{a}_y^T \end{bmatrix},$$

is positive definite by Lemma 3. Hence $\mathbf{M}_{X,Y}$, defined in (19), is positive definite for $X, Y > 0$ and non-negative definite for $X, Y = 0$. Since $\mathbf{F}_{X,Y}^{2L+1}(s_1,s_2) = E[\mathbf{M}_{X,Y}]$ the Lemma is established. □

Appendix C

For a three dimensional random variable (X, Y, Z) the trivariate power moment matrix is a $(3L(L+1)+1) \times (3L(L+1)+1)$ matrix defined as the statistical expectation of the outer product

$$\mathbf{P}_{X,Y,Z}^{3L(L+1)+1} = E[\underline{W}\underline{W}^T],$$

where $\underline{W} = [1, \underline{X}, \underline{Y}, \underline{Z}, \text{vec}\underline{XY}^T, \text{vec}\underline{YZ}^T, \text{vec}\underline{XZ}^T]^T$. where, $\underline{X} = [X, \ldots, X^L]$, $\underline{Y} = [Y, \ldots, Y^L]$, $\underline{Z} = [Z, \ldots, Z^L]$, and, for a $m \times n$ matrix $\mathbf{B}$, vec$(\mathbf{B})$ denotes the mn element row vector formed by concatenation of all m rows of $\mathbf{B}$.

The $(3L(L+1)+1) \times (3L(L+1)+1)$ trivariate factorial moment matrix $\mathbf{F}_{X,Y,Z}^{3L(L+1)+1}(s_1, s_2, s_3, s_4, s_5, s_6)$ is defined as the expectation over X, Y, Z of the following generalized version of the matrix $\mathbf{M}_{X,Y}$ in (19)

$$\mathbf{M}_{X,Y,Z} = \begin{bmatrix} 1 & \underline{a}_x^T & \underline{a}_y^T & \underline{a}_z^T & \underline{a}_{xy}^T & \underline{a}_{yz}^T & \underline{a}_{xz}^T \\ \underline{a}_x & \mathbf{H}_x & \underline{a}_x\underline{a}_y^T & \underline{a}_x\underline{a}_z^T & \underline{a}_x\underline{a}_{xy}^T & \underline{a}_x\underline{a}_{yz}^T & \underline{a}_x\underline{a}_{xz}^T \\ \underline{a}_y & \underline{a}_y\underline{a}_x^T & \mathbf{H}_y & \underline{a}_y\underline{a}_z^T & \underline{a}_y\underline{a}_{xy}^T & \underline{a}_y\underline{a}_{yz}^T & \underline{a}_y\underline{a}_{xz}^T \\ \underline{a}_z & \underline{a}_z\underline{a}_x^T & \underline{a}_z\underline{a}_y^T & \mathbf{H}_z & \underline{a}_z\underline{a}_{xy}^T & \underline{a}_z\underline{a}_{yz}^T & \underline{a}_z\underline{a}_{xz}^T \\ \underline{a}_{xy} & \underline{a}_{xy}\underline{a}_x^T & \underline{a}_{xy}\underline{a}_y^T & \underline{a}_{xy}\underline{a}_z^T & \mathbf{H}_{xy} & \underline{a}_{xy}\underline{a}_{yz}^T & \underline{a}_{xy}\underline{a}_{xz}^T \\ \underline{a}_{yz} & \underline{a}_{yz}\underline{a}_x^T & \underline{a}_{yz}\underline{a}_y^T & \underline{a}_{yz}\underline{a}_z^T & \underline{a}_{yz}\underline{a}_{xy}^T & \mathbf{H}_{yz} & \underline{a}_{yz}\underline{a}_{xz}^T \\ \underline{a}_{xz} & \underline{a}_{xz}\underline{a}_x^T & \underline{a}_{xz}\underline{a}_y^T & \underline{a}_{xz}\underline{a}_z^T & \underline{a}_{xz}\underline{a}_{xy}^T & \underline{a}_{xz}\underline{a}_{yz}^T & \mathbf{H}_{xz} \end{bmatrix},$$

where $\underline{a}_x = \underline{a}(X, s_1)$, $\underline{a}_y = \underline{a}(Y, s_2)$, $\underline{a}_z = \underline{a}(Z, s_3)$, $\underline{a}_{xy} = \underline{a}(XY, s_4)$, $\underline{a}_{yz} = \underline{a}(YZ, s_5)$ and $\underline{a}_{xz} = \underline{a}(XZ, s_6)$, and similarly for the 6 $\mathbf{H}$ matrices. The proof of non-negative definiteness of the trivariate factorial moment matrix is a straightforward extension of the bivariate case.

Note that all moments of the form $E[X^{k_1}Y^{k_2}Z^{k_3}]$ for $0 \le k_1, k_2, k_3 \le L$ can be extracted from either of these trivariate moment matrices. Hence the moment matrices completely characterize the trivariate density as $L \to \infty$. In many applications the pairwise higher order moments of the form $E[X^{k_1}Y^{k_2}], E[Y^{k_2}Z^{k_3}]$ and $E[X^{k_1}Z^{k_3}]$ may be sufficient to perform pattern discrimination. In this case only the upper $(3L+1) \times (3L+1)$ block submatrices of $\mathbf{P}_{X,Y,Z}$ or $\mathbf{F}_{X,Y,Z}$ need be computed and used.

References

[1] S. Kahan, T. Pavlidis, and H. Baird, "On the recognition of printed characters of any font and size," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 9, no. 2, pp. 274–288, Mar, 1987.

[2] S. Ghosal and R. Mehrortra, "Orthogonal moment operators for subpixel edge detection," *Pattern Recognition*, vol. 26, no. 2, pp. 295–306, 1993.

[3] I. Szapudi and A. Szalay, "Higher order statistics of the galaxy distribution using generating functions," *Astrophysical Journal*, no. 1, pp. 43–56, 1993.

[4] P. Carruthers, "Galaxy correlations, counts, and moments," *Astrophysical Journal*, no. 1, pp. 24–29, 1991.

[5] J. O'Neill, A. O. Hero, and W. J. Williams, "Word spotting via spatial point processes," in *IEEE Int. Conf. on Image Processing*, volume 2, pp. 215–219, Laussane, Switzerland, Sept. 1996.

[6] M. Hu, "Pattern recognition by moment invariants," *Proc. of the Institute of Radio Engineers (IRE)*, p. 1428, 1961.

[7] Y. Li, "Reforming the theory of invariant moments for pattern recognition," *Pattern Recognition*, no. 7, pp. 723–730, 1992.

[8] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Independent component analysis based on higher-order-only statistics," in *Proc. 1996 8th IEEE Sig. Proc. Workshop on Statist. Sig. and Array Processing (SSAP'96)*, pp. 356–359, Corfu, Greece, 1996.

[9] N. L. Johnson, S. Kotz, and A. W. Kemp, *Univariate discrete distributions (2nd Ed)*, Wiley, New York, 1992.

[10] Y. Lepage, "Negative factorial moments of positive random variables," *Indust. Math.*, no. 2, pp. 95–100, 1978.

[11] G. Cammilleri, "Evaluation des parametres des fonctions composées resultant du mélange de deux courbes élémentaires, avec la méthode des moments factoriels," *Publ. Inst. Statist. Univ. Paris*, no. 2, pp. 17–35, 1982.

[12] B. Blaszczyszyn, "Factorial moment expansion for stochastic systems," *Stochastic Process. Appl.*, pp. 321–335, 1995.

[13] M.-S. Lu and T. Teichmann, "Statistical modelling of coincidence counting," *Nucl. Instrum. and Meth.*, no. 2, pp. 544–550, 1993.

[14] A. Dodson and R. Vyas, "Homodyne photon statistics of the subthreshold degenerate parametric oscillator," *Phys. Rev. A*, pp. 3396–3412, 1993.

[15] I. Dremin, "Fractality and the anomalous dimension of quantum chromodynamics," *IFIP Trans. A*, pp. 101–107, 1994.

[16] EMU01 Collaboration, "Systematic investigation of scaled factorial cumulant moments for nucleus-nucleus interactions," *Physical Rev. D*, no. 9, pp. 3726–3732, 1993.

[17] M. Abromowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover, New York, N.Y., 1965.

[18] R. Redner and H. Walker, "Mixture densities, maximum likelihood, and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, April 1984.

[19] C. E. Priere and D. J. Marchette, "Adaptive mixture density estimation," *Pattern Recognition*, vol. 26, no. 5, pp. 771–785, May 1993.

[20] R. Samadani, "Finite mixtures algorithm for finding proportions in SAR images," *IEEE Trans. on Image Processing*, vol. 4, no. 8, pp. 1182–1186, aug 1995.

[21] F. Chen, L. Wilcox, and D. Bloomberg, "Word spotting in scanned images using hidden Markov modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, volume V, pp. 1–4, Minneapolis, MN, 1993.

[22] D. R. Brillinger, *Time Series: Data Analysis and Theory*, Springer-Verlag, New York, 1981.

[23] S.-L. Hsue and S. Soliman, "Automatic modulation recognition of digitally modulated signals," in *Proceedings of MILCOM*, pp. 645–647, Boston, MA, 1989.

[24] A. Polydoros and K. Kim, "On the detection and classification of quadrature digital modulations in broad-band noise," *IEEE Trans. on Communications*, vol. 38, pp. 1199–1211, Aug. 1990.

[25] C. Long, K. Chugg, and A. Polydoros, "Further results in likelihood classification of qam signals," in *Proceedings of MILCOM*, pp. 57–61, Long Branch, NJ, 1994.

[26] Y.-C. Lin and C.-C. J. Kuo, "Practical psk modulation classifier using wavelets," in *Proceedings of SPIE*, pp. 492–503, Bellingham, WA, 1995.

[27] B. F. Beidas and C. L. Weber, "Higher-order correlation-based approach to modulation classification of digitally frequency-modulated signals," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 89–101, Jan. 1995.

[28] S. S. Soliman and S.-Z. Hsue, "Signal classification using statistical moments," *IEEE Trans. on Communications*, vol. 40, pp. 908–916, May. 1992.

[29] H. Hadinejad-Mahram, "A novel pattern recognition based digital modulation classification apprach," Technical Report Studienarbeit submitted in partial fulfilment of requirements for Diplom Ingenieur, Dept. Electrical Engineering, University of Technology, Darmstadt, DL, mar 1998.

[30] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge, 1985.