Joint distribution of correlated radar images

Norwegian Computing Center
APPLIED RESEARCH AND DEVELOPMENT

NOTE

SAMBA/30/03

Bård Storvik
Geir Storvik
Roger Fjørtoft

December 2003

Norwegian Computing Center
APPLIED RESEARCH AND DEVELOPMENT

NR Note

Title: Joint distribution of correlated radar images

Date: December 2003
Year: 2004
Note no.: SAMBA/30/03

Author: Bård Storvik <baard@nr.no>
Geir Storvik <storvik@nr.no>
Roger Fjørtoft <Roger.Fjortoft@cnes.fr>

Abstract: The literature on automatic analysis of remote sensing data has so far been dominated by methods that are applied to single images. With the ever-increasing number and diversity of earth observation satellites, it steadily becomes more important to be able to analyze compound data sets consisting of several images, possibly acquired by different sensors. The observed pixel values in optical images with several spectral bands are well modeled by a multivariate normal (Gaussian) distribution for each ground cover class, and the same model can be used for the joint distribution of a set of overlapping multispectral images. Likewise, multivariate complex circular Gaussian distributions can be used for single-look complex radar images. However, for detected radar images (amplitude or intensity) neither marginal nor joint distributions are normal. In this study we examine different ways of obtaining joint distributions for detected radar images, and we propose a transformation method that enables incorporation of inter-image covariance while keeping a good fit to the marginal distributions. The approach is studied for three different distribution families that are used to model radar image intensities: Gamma (classes with constant radar reflectivity), lognormal (an approximation) and K (textured classes with Gamma distributed radar reflectivity).

The approach basically consists of two steps. The marginal densities are assumed to come from parametric distributions. Based on this assumption, each marginal variable is transformed to a normal distributed variable. The joint distribution of the transformed variables is assumed multivariate normal with a certain covariance matrix. Transforming it back to the original scale will give a joint distribution with dependence, where the initial marginal distributions are not altered. The parameters of the new joint distribution can be estimated. Assuming marginal Gamma (or K) distributions and then using the proposed transformation method will give a flexible joint Gamma (or K) distribution incorporating inter-image dependence. If lognormal distributions are assumed marginally, the standard joint lognormal distribution appears when using the transformation method.

The joint distributions produced by the transformation method can e.g. be used in supervised classification of radar images. Results obtained on various data sets are presented.

Keywords: Radar images, transformation, multivariate distribution, inter-image correlation, normal scale, dependence.

Target group: All employees

Availability: Open

Project: EOTOOLS, WP2

Project no.: 830110

Research field: Remote Sensing

No. of pages: 34

Norwegian Computing Center
Gaustadalléen 23, P.O. Box 114 Blindern, NO-0314 Oslo, Norway
Telephone: +47 2285 2500, telefax: +47 2269 7660, http://www.nr.no

Contents

1 Introduction

Earth observation satellites acquire images of the earth's surface and atmosphere. Compared to optical sensors, radar has the advantage of being able to look through clouds. Moreover, as a radar provides its own illumination, it is independent of the sunlight and can record images also at night. Satellite borne radars are generally side-looking, and the synthetic aperture radar (SAR) principle is used to improve the resolution. The wavelength is typically of the order of decimeters.

The radar response of a surface is very much dependent on its structure, as well as its dielectric properties. A piece of surface that is perpendicular to the incoming radar beam will typically return a relatively strong response back to the satellite. This is for example the situation when reflectors are installed to calibrate the radar system. At the other extreme, a plane surface that is not perpendicular to the incoming beam (e.g. a lake without waves) will reflect little or nothing of the signal back to the satellite. Most natural surfaces are irregular at the scale of the wavelength, in which case the characteristic *speckle* phenomenon can be observed in the resulting image. Speckle is due to the constructive and destructive interferences of the responses of the elementary scatterers within a resolution cell, and it results in very strong fluctuations in the observed intensities for a given ground cover type. Speckle is often modeled as a strong multiplicative noise. The observed intensity is then the product of the characteristic radar reflectivity of the surface, and the speckle. Radar images rely on coherent illumination and are inherently complex (each pixel initially has an amplitude and a phase). However, in many cases only the amplitude or intensity are available for further analysis. For example, a technique called multi-looking is often used to reduce the speckle, in which case the phase is lost.

Let us now assume that we have a set of radar images that have been acquired over a given area, with approximately the same acquisition geometry. The images will generally appear somewhat different, e.g. because they were:

- not acquired simultaneously (multi-temporal)
- acquired by sensors with different wavelengths (multi-frequency)
- acquired with different polarization combinations (polarimetric)

For a given ground cover class (e.g. a certain kind of agricultural field or forest) the pixel values of the different images may then be correlated. This correlation can easily be taken into account for single-look complex (SLC) radar images, where a multivariate complex circular Gaussian distribution is well suited. However, for amplitude or intensity images (single- or multi-look) there is no straight-forward way of expressing the joint distribution.

Let us first briefly introduce the basic statistical properties of the pixel values of a radar image. The complex amplitude of a pixel in an SLC image is the sum of the contributions $a_k e^{-j\phi_k}$ of the N elementary scatterers within the resolution cell

$$Z \;=\; Ae^{-j\phi} = \sum_{k=1}^{N} a_k e^{-j\phi_k}$$

The real and imaginary parts of the complex amplitude $Z = Z_{\Re} + jZ_{\Im}$ can be written as

$$\begin{aligned} Z_{\Re} &= A\cos(\phi) \\ Z_{\Im} &= A\sin(\phi). \end{aligned}$$

To proceed further in finding the distribution of $Z_{\Re}$ and $Z_{\Im}$, some assumptions has to be made. Assuming that the speckle is *fully developed* (Goodman, 1984), i.e. basically that the observed surface is irregular at the scale of the wavelength and that the number of elementary scatterers N within each resolution cell is sufficiently big. From the irregularity assumption it follows that the amplitude a_k and phase ϕ_k are independent and identically distributed are reasonable assumptions. Furthermore, it also follows that the mean $E[\phi_k] = 0$ is a reasonable assumption. From the central limit theorem it follows that $Z_{\Re}$ and $Z_{\Im}$ are normal when N goes to infinity and $a_k\cos(\phi_k)$ and $a_k\sin(\phi_k)$ are identically distributed for all $k = 1, 2, \ldots$. Below the distribution of $Z_{\Re}$ and $Z_{\Im}$ are assumed normal. Define $E[a_k] = a$ and $E[a_k^2] = a^2$. It can then easily be shown that

$$\begin{aligned} E[Z_{\Re}] &= E[Z_{\Im}] = 0 \\ E[Z_{\Re}^2] &= E[Z_{\Im}^2] = \frac{N}{2}a^2, \end{aligned}$$

and that the real and imaginary parts are uncorrelated because

$$E[Z_{\Re}Z_{\Im}] = E[Z_{\Re}]E[Z_{\Im}] = 0.$$

As they are also independent we can write

$$f_{Z_{\Re},Z_{\Im}}(z_{\Re}, z_{\Im}) = \frac{1}{\pi R}e^{-\frac{z_{\Re}^2+z_{\Im}^2}{R}},$$

where $R = Na^2$ is called the *radar reflectivity of* the surface.

Transforming $Z_{\Re}$ and $Z_{\Im}$ to polar coordinates $A = \sqrt{Z_{\Re}^2 + Z_{\Im}^2}$ and $\phi = \tan^{-1}(Z_{\Re}/Z_{\Im})$, we get

$$\begin{aligned} f_{A,\phi}(x,y) &= f_{Z_{\Re},Z_{\Im}}(x\cos(y), x\sin(y))J \\ &= \begin{cases} \frac{x}{\pi R}e^{-\frac{x^2}{R}} & \text{for } x > 0 \text{ and } y \in [-\pi, \pi] \\ 0 & \text{elsewhere} \end{cases} \end{aligned}$$

where J is the Jacobi determinant. From this it follows that the scalar amplitude A is Rayleigh distributed

$$f_A(x) = \frac{2x}{R}e^{-\frac{x^2}{R}},\ x > 0$$

and that the phase ϕ is uniformly distributed

$$f_{\phi}(y) = \begin{cases} \frac{1}{2\pi} & \text{for } y \in [-\pi, \pi] \\ 0 & \text{elsewhere.} \end{cases}$$

The distribution of the intensity $I = A^2$ is exponential

$$f_I(x) = \frac{1}{R}e^{-\frac{x}{R}},\ x \geq 0.$$

Assume $I_1, \ldots, I_L$ are independent exponentials. The average value of L pixel intensities

$$\bar{I} = \frac{1}{L}\sum_{i=1}^{L} I_i \tag{1}$$

is Gamma distributed

$$f_{\bar{I}}(x) = \frac{1}{\Gamma(L)}\left(\frac{L}{R}\right)^{L} \exp(-\frac{Lx}{R})x^{L-1}, \; x \geq 0, \tag{2}$$

where $E[\bar{I}] = R$, $Var[\bar{I}] = R^2/L$, and x is a realization of $\bar{I}$. In practise the resolution cells are generally slightly overlapping, in which case there will be spatial correlation and the distribution of $\bar{I}$ is only approximately Gamma. However, in this study we will neglect spatial correlation and concentrate on the correlation between corresponding pixels in different images.

2 Multivariate Gamma case

In this section the aim is to explore the possibilities of extending the univariate Gamma distribution to a multivariate Gamma distribution. There are a number of various forms of multivariate Gamma distributions, see e.g. (Kotz et al., 2000). Each form will give different properties. Suppose $Y_0, Y_1, Y_2, \ldots, Y_k$ are independently Gamma distributed with with parameters $\theta_0, \theta_1, \ldots, \theta_K$. Let $X_k = Y_0 + Y_k$ for $k = 1, 2, \ldots, K$. In the general case, the multivariate distribution of $X_1, \ldots, X_K$ leads to very complicated expressions. However, $\theta_1 = \theta_2 = \cdots = \theta_k = 1$ (each have an exponential distribution) leads to an expression of the density. The expression involves integration. The restriction on the parameters is quite strong in such a way that the distribution is not flexible enough. The other multivariate Gamma distributions defined in Kotz and Johnson do not have easy expressions for the joint distribution. It is however essential to be able compute the joint density in order to classify the pixels. If the expression of the distribution is difficult, simulation will have to be used. One example is shown in the next subsection.

2.1 Wishart distribution

Suppose we have two bands

$$Z = \begin{pmatrix} Z_{\Re_1} \\ Z_{\Re_2} \\ Z_{\Im_1} \\ Z_{\Im_1} \end{pmatrix} \sim N_4\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho & 0 & 0 \\ \rho & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \rho \\ 0 & 0 & \rho & \sigma_2^2 \end{pmatrix}\right).$$

Marginally, the distribution of $I_i = A_i^2 = (Z_{\Re_i})^2 + (Z_{\Im_i})^2 \sim \exp(2\sigma_i)$ for $i = 1, 2$. Let

$$\Sigma_{12} = \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix}.$$

Define $\mathbf{Z}_{\Re} = (Z_{\Re_1}, Z_{\Re_2})$ and $\mathbf{Z}_{\Im} = (Z_{\Im_1}, Z_{\Im_2})$. Then $M_{\Re} = (\mathbf{Z}_{\Re})^T \mathbf{Z}_{\Re} \sim W(\Sigma_{12}, 1)$ and $M_{\Im} = (\mathbf{Z}_{\Im})^T \mathbf{Z}_{\Im} \sim W(\Sigma_{12}, 1)$ (W is for Wishart distribution). Furthermore, $M_{\Re} + M_{\Im} \sim W(\Sigma_{12}, 2)$. We can write $M_{\Re} + M_{\Im}$ as

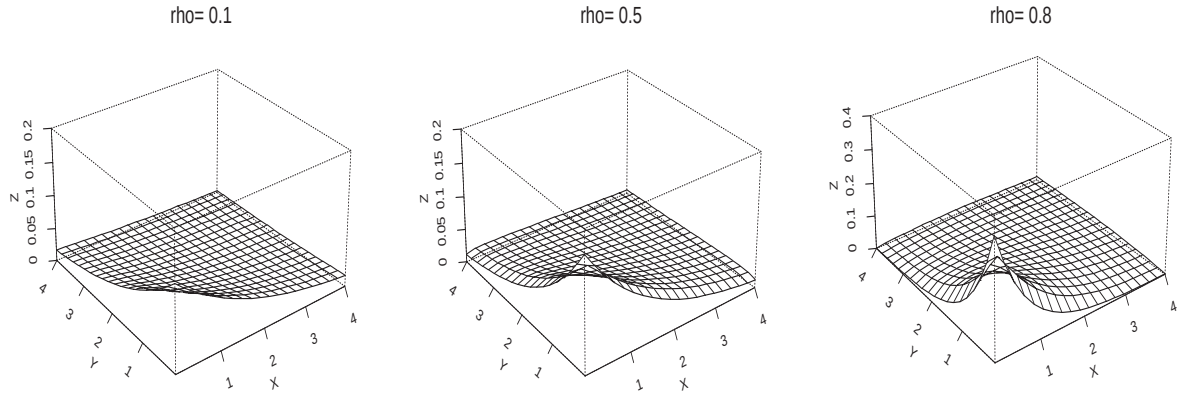
$$M = M_{\Re} + M_{\Im} = \begin{pmatrix} (Z_{\Re_1})^2 + (Z_{\Im_1})^2 & Z_{\Re_1} Z_{\Re_2} + Z_{\Im_1} Z_{\Im_1} \\ Z_{\Re_1} Z_{\Re_2} + Z_{\Im_1} Z_{\Im_1} & (Z_{\Re_2})^2 + (Z_{\Im_1})^2 \end{pmatrix} = \begin{pmatrix} I_1 & I_{12} \\ I_{12} & I_2 \end{pmatrix}.$$

The distribution of M is

$$f(M) = \frac{|M|^{-1/2} \exp\{-\frac{1}{2} \mathrm{tr} \Sigma_{12}^{-1} M\}}{4\pi^{1/2} |\Sigma_{12}| \prod_{i=1}^{2} \Gamma(\frac{1}{2}(3-i)}. \tag{3}$$

However, we want the distribution of the diagonal element of M, which is (I_1, I_2). This is a k-variate chi square distribution. From figure 1 it is possible to see that strong depend-

Figure 1: Density of Wishart distribution when $\rho = 0.1$, $\rho = 0.5$ and $\rho = 0.8$.



ence between real and imaginary values in the two bands will change the distribution of (I_1, I_2) dramatically compared to little dependence. A stronger dependence will make the distribution (I_1, I_2) more concentrated around 0.

2.2 Transformation

The transformation method that we introduce in this section is based on the following theorem:

Theorem 1 *Let X be a continuous cumulative distribution function F on $\mathcal{R}$, that is $F(x) = Pr\{X \leq x\}$. The inverse function F^{-1} is defined by:*

$$F^{-1}(u) = \inf\{x : F(x) = u,\, 0 < u < 1\}. \tag{4}$$

If U is a uniform $[0,1]$ random variable, then $F^{-1}(U)$ has distribution function F. Also, if X has distribution function F, then $F(X)$ is uniformly distributed on $[0,1]$.

Proof:
For all $x \in \mathcal{R}$, the first statement follows from

$$\begin{aligned} \Pr\{F^{-1}(U) \leq x\} &= \Pr\{\inf\{y : F(y) = U,\, 0 < u < 1\} \leq x\} \\ &= \Pr\{U \leq F(x)\} \\ &= F(x). \end{aligned}$$

For all $0 < u < 1$, the second statement follows from

$$\begin{aligned} \Pr\{F(X) \leq u\} &= \Pr\{X \leq F^{-1}(u)\} \\ &= F(F^{-1}(u)) \\ &= u. \end{aligned}$$

Let $Y = (Y_1, \ldots, Y_K)^t$ be multivariate normally distributed with mean vector 0 and correlation matrix Σ, that is $\mathrm{diag}(\Sigma) = 1$. Now, marginally $Y_k \sim N(0,1)$, that is Y_k is standard normally distributed. Suppose Φ is the cumulative distribution function of a standard normal distribution. It follows from the results above that the marginally transformed variables $\Phi(Y_k) = U_k$ is uniformly distributed on $[0,1]$. Let G_i be any cumulative distribution function, it then follows from the above results that $X_k = G_k^{-1}(\Phi(Y_k))$ has a cumulative distribution function G_k. Let g be the multivariate density function of $X = (X_1, \ldots, X_K)^t$ and g_k the marginal density function of X_k. Define

$$y_k(x) = \Phi^{-1}(G_k(x)). \tag{5}$$

and

$$\begin{aligned} y(x) &= (y_1(x_1), \ldots, y_K(x_K))^t \\ &= (\Phi^{-1}(G_1(x_1)), \ldots, \Phi^{-1}(G_K(x_K)))^t. \end{aligned} \tag{6}$$

It follows that the distribution of the transformed variables (Mardia et al., 1979):

$$\begin{aligned} g(x) &= |2\pi\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)\Sigma^{-1}y(x)\} \\ &\quad \times \begin{vmatrix} \frac{g_1(x_1)}{\phi(\Phi^{-1}(G_1(x_1)))} & & 0 \\ & \ddots & \\ 0 & & \frac{g_k(x_K)}{\phi(\Phi^{-1}(G_K(x_K)))} \end{vmatrix} \\ &= |2\pi\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)\Sigma^{-1}y(x)\} \times \prod_{k=1}^{K}\frac{g_k(x_k)}{\phi(y_k(x_k))} \\ &= \prod_{k=1}^{K} g_k(x_k) \times \frac{|2\pi\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)^t\Sigma^{-1}y(x)\}}{|2\pi I|^{-1/2}\exp\{-\frac{1}{2}y(x)^t y(x)\}} \\ &= \prod_{k=1}^{K} g_k(x_k) \times |\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)^t(\Sigma^{-1}-I)y(x)\}. \end{aligned}$$

The result is

$$g(x) = \prod_{k=1}^{K} g_k(x_k) \times |\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)^t(\Sigma^{-1}-I)y(x)\}. \quad (7)$$

Example 1. Marginal lognormal distributed variables
When data are lognormal, the distribution of the logarithm of the data is normal. The density function and the cumulative function G_k and g_k will then be

$$\begin{aligned} G_k(x) &= \Phi((\log(x)-\mu_k)/\sigma_k), \text{ for } x>0 \\ g_k(x) &= \frac{\phi((\log(x)-\mu_k)/\sigma_k)}{\sigma_k x}, \text{ for } x>0, \end{aligned}$$

where μ_k and σ_k are the parameters. It follows from (5) that the transformation in this case gives

$$\begin{aligned} y_k(x) &= \Phi^{-1}(G_k(x)) \\ &= \Phi^{-1}(\Phi((\log(x)-\mu_k)/\sigma_k))) \\ &= (\log(x)-\mu_k)/\sigma_k. \end{aligned}$$

The marginal density of

$$X_k = G_k^{-1}(\Phi(Y_k))$$

is lognormal. The result is

$$\begin{aligned} g(x) &= \prod_{k=1}^{K} g_k(x_k) \times |\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)^t(\Sigma^{-1}-I)y(x)\} \\ &= |2\pi\Sigma_1|^{-1/2}\exp\{-\frac{1}{2}y(x)^t y(x)\} \times |\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)^t(\Sigma^{-1}-I)y(x)\} \\ &= |2\pi\Sigma_1|^{-1/2}|\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)^t(\Sigma^{-1}-I+I)y(x)\} \\ &= |2\pi\Sigma_2|^{-1/2}\exp\{-\frac{1}{2}(\log(x)-\mu)^t\Sigma_2^{-1}(\log(x)-\mu)\}, \end{aligned}$$

where $\Sigma_1 = \text{diag}\{\sigma_1^2, \ldots, \sigma_k^2\}$ and $\Sigma_1^{1/2}\Sigma\Sigma_1^{1/2}$. $g(x)$ is the multivariate lognormal distribution. When data are normal, the transformation method yields a multivariate normal distribution.

Example 2. Marginal Gamma distributed variables
Suppose $X_1, X_2, \ldots, X_K$ all are marginally Gamma distributed. We want to build in dependence. Denote the density of X_k by

$$g_k(x) \;=\; g(x;\alpha_i,\beta_k) \;\;=\;\; \frac{\alpha_k^{\beta_k}}{\Gamma(\beta_k)} x^{\beta_k-1}\exp\{-\alpha_k x\}, \quad x>0. \tag{8}$$

Denote the cumulative distribution function of X_k by $G_k(x) = G(x;\alpha_k,\beta_k)$. Define

$$X_k \;\;=\;\; G_k^{-1}(\Phi(Y_k)),$$

where Y_k has cumulative distribution function Φ (standard normal). From (13) we get the joint distribution, which is

$$g(x) \;\;=\;\; \prod_{k=1}^{K} g_k(x_k) \times |\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)^t(\Sigma^{-1}-I)y(x)\}, \tag{9}$$

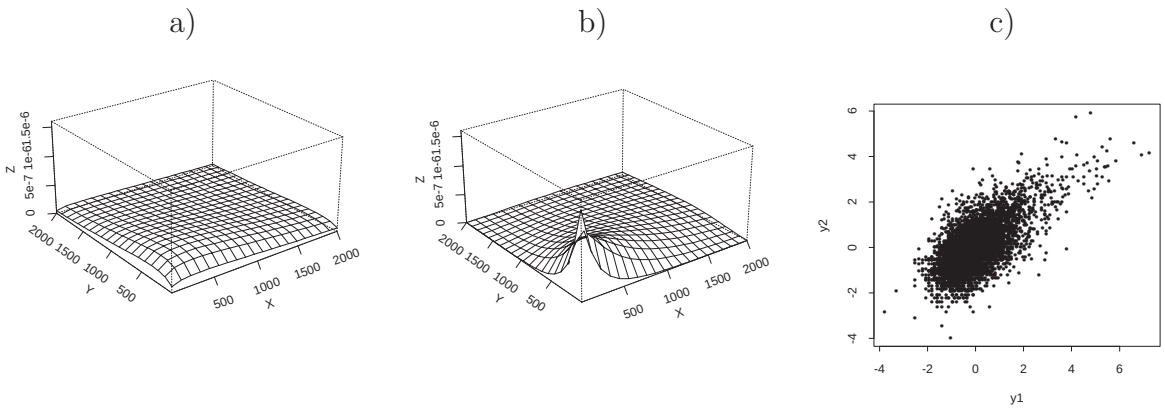
where $y(x)$ is defined in (6). The first term in (9) is a product of independent marginal gamma densities. The second term in (9), $Q(x) = |\Sigma|^{-1/2}\exp\{-\frac{1}{2}y(x)^t(\Sigma^{-1}-I)y(x)\}$, is a correction term which includes dependence on a normal scale. Note that the diagonal of Σ should be equal to 1, i.e., $\text{diag}(\Sigma) = 1$.

Example 3. Marginal K-distributed variables
Suppose $X_1, X_2, \ldots, X_K$ all are marginally K-distributed. We want to build in dependence. Denote the density of X_k by:

$$g_k(x) \;=\; g(x;\alpha_k,\beta_k,L_k) \;\;=\;\; \frac{\beta}{\sqrt{x}\Gamma(L_k)\Gamma(\alpha_k)}\left(\frac{\beta\sqrt{x}}{2}\right)^{\alpha+L_k-1}\mathcal{K}_{\alpha-L}(\beta\sqrt{x}), \quad x>0 \tag{10}$$

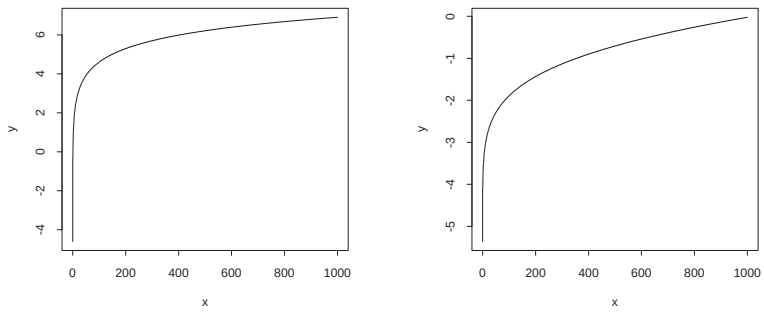
where $\mathcal{K}$ is here the modified Bessel function of the second kind. The same procedure as in example 1 and 2 can now be followed.

Figure 2: a) Gamma without dependence. b) Gamma with dependence. c) Scatter plot of transformed variables.



How will the transformation defined in (5) look like when we use lognormal and Gamma distribution?

Figure 3: Transformation to normal scale in the lognormal case (left panel) and Gamma case (right panel).



3 Meta-Gaussian distribution

We include this section which summarize the previous section. Note that a slightly new notation will be used. In order to combine different channels which are dependent, we have to model the dependence between the channels. From the introduction, the marginal distribution of radar images (pixel intensity values) can be modeled as gamma distributed. The approach now described is however more general and does not require gamma as marginal distribution. The idea is to transform each marginal value to Gaussian, measure the correlation on the Gaussian scale and transform back.

Let $\mathbf{X} = (X_1, ..., X_p)$ be a stochastic vector with marginal density g_j for the jth component X_j of $\mathbf{X}$. Let G_j be the cumulative distribution function corresponding to g_j and Φ the cumulative distribution function for the standard normal distribution. General probability theory shows that

$$Y_j = \Phi^{-1}(G_j(X_j)) \tag{11}$$

is a standard normally distributed variable. The meta-Gaussian approach is to model dependence between the different components of $\mathbf{X}$ through dependence between the components of $\mathbf{Y} = (Y_1, ..., Y_p)$. In particular, it is assumed that $\mathbf{Y}$ is a multivariate Gaussian distributed vector with expectation vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. In order to keep each Y_k standard normal, we require the diagonal elements of $\mathbf{\Sigma}$ to be equal to 1. Inverting (11), we obtain

$$X_j = G_j^{-1}(\Phi(Y_j)). \tag{12}$$

Further, by using standard results from probability theory on transformations, the multivariate density for $\mathbf{X}$ is

$$f(\mathbf{x};\boldsymbol{\gamma}) = |\mathbf{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}\mathbf{y}(\mathbf{x};\boldsymbol{\gamma})^T(\mathbf{\Sigma}^{-1} - \mathbf{I})\mathbf{y}(\mathbf{x};\boldsymbol{\gamma})\} \times \prod_{j=1}^{p} g_j(x_j;\boldsymbol{\gamma}_j), \tag{13}$$

where $\boldsymbol{\gamma}_j$ are the parameters describing the marginal distribution g_j, $\mathbf{y}(\mathbf{x};\boldsymbol{\gamma})=(y_1(x_1;\boldsymbol{\gamma}_1), \ldots, y_p(x_p;\boldsymbol{\gamma}_p))^T$ and $y_j(x_j;\boldsymbol{\gamma}_j) = G_j^{-1}(\Phi(x_j);\boldsymbol{\gamma}_j)$.

Note that for $\mathbf{\Sigma} = \mathbf{I}$, the distribution reduces to a product of independent marginals, making the interpretation of $\mathbf{\Sigma}$ similar to the correlation matrix for multivariate Gaussian distributions. Note further that no assumptions is made on g_j, except for the inverse of cumulative distribution G_j existing.

In practice, g_j will usually be chosen from a parametric family of distributions. If all g_j are Gaussian, the density (13) reduces to a Multivariate normal distribution. If all g_j are lognormal, we obtain the ordinary multivariate lognormal distribution. For g_j being Gamma distributions, we obtain a multivariate Gamma distribution. If some g_j are Gaussian and some are Gamma, a multivariate distribution combining Gaussian marginals with Gamma marginals is obtained. Such combinations is fruitful when combining optical with radar images.

In this paper we will concentrate on Gamma marginals and multivariate Gamma distributions.

4 Classification based on meta-Gaussian distributions

Using the framework in the previous section, we may for each class $k \in \{1, ..., K\}$ define a multivariate density $f_k(\mathbf{x})$ describing the distribution of a vector of observations $\mathbf{x}$ from class k. Define z_i to be the class of pixel i and $\mathbf{x}_i$ observations from pixel i. Neglecting contextual dependence, the Bayes rule for classification is

$$\hat{z}_i = \text{argmax}_k \pi_k f_k(\mathbf{x}). \tag{14}$$

Contextual classification methods can also be applied in the ordinary way. Assume e.g. a Potts model

$$p(\mathbf{z}) \propto e^{\sum_i \alpha_{z_i} + \beta \sum_{i \sim j} I(z_i = z_j)},$$

where $I(\cdot)$ is the indicator function and $i \sim j$ means that i and j are neighbors in a graph. Making the usual assumption on conditional independence of observations given classes, the posterior distribution for $\mathbf{z}$ is given by

$$p(\mathbf{z}|\mathbf{x}) \propto p(\ z) \prod_i f_{z_i}(\mathbf{x}_i). \tag{15}$$

Maximum a posteriori (MAP) estimates of $\mathbf{z}$ can be obtained by global maximization of (15). Such a maximization is recognized as a difficult problem and therefore approximative algorithms such as the Iterative Conditional Modes (ICM) Besag (1986) are usually applied. Note however that an efficient algorithm for obtaining global maxima has been presented in Storvik and Dahl (2000).

5 Estimation

In this section we consider parameter estimation. Assume a training set

$$T = \{\mathbf{x}_{k,i}, i = 1, ..., n_k, k = 1, ..., K\}$$

is available where $\mathbf{x}_{k,i}$ is a vector of observations from class k. All observation vectors are assumed independent, while dependence between components of the vectors is modeled through the meta-Gaussian distribution. We further assume the marginal distributions to be parametric with no common parameters between classes.

Under these assumptions, estimation can be performed separately on each class. Without loss of generality we therefore only consider one class. We further simplify the notation by suppressing the class index, i.e. our data is $\mathbf{x}_i, i = 1, ..., n$.

One obvious estimation method is Maximum Likelihood (ML). This is considered in section 5.1. However, finding the maximum likelihood estimates may be time consuming because we will have to optimize the whole likelihood wrt the parameters. Therefore, we will in section (5.2) introduce a simplified approach based on estimation functions (EF) which is much simpler and faster.

5.1 Maximum Likelihood

Assume each marginal density g_j contain a set of unknown parameters $\boldsymbol{\gamma}_j$ and write $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p)$. The full set of unknown parameters to be estimated is $(\boldsymbol{\gamma}, \boldsymbol{\Sigma})$.

Assume observations $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, where each $\mathbf{x}_i = (x_{i,1}, .., x_{1,p})$. Define

$$\mathbf{S}(\boldsymbol{\gamma}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} y(\mathbf{x}_i;\boldsymbol{\gamma})\mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma})^T.$$

From (13)

$$\begin{aligned}\log(f(\mathbf{x};\boldsymbol{\gamma},\boldsymbol{\Sigma})) = &-\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\mathbf{tr}(\boldsymbol{\Sigma}^{-1} - \mathbf{I})\mathbf{S}(\boldsymbol{\gamma}) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{p}\log(g_j(x_{i,j};\boldsymbol{\gamma}_j)) \end{aligned} \tag{16}$$

where

$$\mathbf{S}(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma})\mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma})^T. \tag{17}$$

In some cases (i.e for multivariate normal and multivariate lognormal distributions), analytical expressions for the ML estimates are available. In general, optimizing the log-likelihood (16) w.r.t. $(\boldsymbol{\gamma}, \boldsymbol{\Sigma})$ is a difficult task. This is mainly due to the restrictions on $\boldsymbol{\Sigma}$ (positive definite and 1 on the diagonal). Also, there might be some restrictions on γ depending on the marginal distributions.

The difficult part is the optimization wrt to $\boldsymbol{\Sigma}$, because of the constraints involved. These constraints can be handled by a transformation. Write first

$$\boldsymbol{\Sigma} = \mathbf{D}(\boldsymbol{\Sigma})^{-1/2}\boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{\Sigma})^{-1/2}, \tag{18}$$

where $\boldsymbol{\Sigma}$ is a positive definite matrix and $\mathbf{D}(\boldsymbol{\Sigma})$ is the diagonal matrix of $\boldsymbol{\Sigma}$. Then automatically the diagonal elements of $\boldsymbol{\Sigma}$ are all equal to 1. We can write $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ where L is a lower triangular matrix because $\boldsymbol{\Sigma}$ is positive definit. Defining $\mathbf{U} = \mathbf{L}^{-1}$, the log likelihood as a function of $\boldsymbol{\gamma}$ and $\mathbf{U}$ now becomes

$$\begin{aligned}\log(f(\mathbf{x};\boldsymbol{\gamma},\mathbf{U})) = &-\frac{n}{2}\log|\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{-1/2}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{-1/2}| \\ &-\frac{n}{2}\mathbf{tr}(\mathbf{D}(\mathbf{U}^T\mathbf{U})^{1/2}\mathbf{U}^T\mathbf{U}\mathbf{D}(\mathbf{U}^T\mathbf{U})^{1/2} - \mathbf{I})\mathbf{S}(\boldsymbol{\gamma}) \\ &+\sum_{i=1}^{n}\sum_{j=1}^{p}\log(g_j(x_{i,j};\boldsymbol{\gamma}_j)). \end{aligned} \tag{19}$$

The likelihood can be optimized using a standard numerical gradient based routine.

5.2 Estimation based on estimating functions

The ML estimates are cumbersome to obtain. In this section we will describe a simpler approach.

The maximum likelihood estimates can be considered as a solution to the equations

$$\frac{\partial \log(f(\mathbf{x};\boldsymbol{\gamma},\mathbf{U}))}{\partial \boldsymbol{\gamma}} = \mathbf{0}$$
$$\frac{\partial \log(f(\mathbf{x};\boldsymbol{\gamma},\mathbf{U}))}{\partial \mathbf{U}} = \mathbf{0}.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \mathbf{U})$. Now, define a estimating function as $h(\boldsymbol{\theta})$, where dependence on observations are suppressed from notation. The estimating equation is defined as

$$h(\boldsymbol{\theta}) = \mathbf{0}. \tag{20}$$

In referring to estimating equation we always mean the equation where the estimating function is set to zero. Furthermore, the estimating function estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is found by solving the estimating equation in (20). Now, there are a number of regularity conditions in order for the estimating function estimator to have the similar type of asymptotic features as the Maximum likelihood estimator (asymptotically consistent and normally distributed). Define the estimating function $h(\boldsymbol{\theta})$ as unbiased if

$$E_{\boldsymbol{\theta}}\{h(\boldsymbol{\theta})\} = \mathbf{0} \ \ \forall \boldsymbol{\theta}.$$

Furthermore, define the sensitivity, variability and Godambe information of $h(\boldsymbol{\theta})$ as

$$S_h(\theta) = E_{\boldsymbol{\theta}}\left\{\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}, \quad V_h(\theta) = E_{\boldsymbol{\theta}}\left\{h(\boldsymbol{\theta})h(\boldsymbol{\theta})^T\right\}$$
$$\text{and} \qquad J_h(\theta) = S_h^T(\boldsymbol{\theta})V_h^{-1}(\boldsymbol{\theta})S_h(\boldsymbol{\theta}),$$

respectively. The traditional regularity conditions on h is; second order moments exist, unbiasedness, the sensitivity and variability are non singular for all $\boldsymbol{\theta}$. The quality of the estimating function estimator is dependent on which estimating function that is chosen. Because the estimating function estimator has the Godambe information as the asymptotic variance, the Godambe information measures quality of the estimator.

If the estimating function is set equal to the score function, the maximum likelihood estimator appears when solving the estimating equation. Let $h(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), h_2(\boldsymbol{\theta}))^T$. Define $\mathbf{U}_0$ to be the values of $\mathbf{U}$ corresponding to independence between components, i.e. $\boldsymbol{\Sigma} = \mathbf{I}$. Let us define a particular set of estimating functions h_1 and h_2 as

$$h_1(\boldsymbol{\theta}) = \frac{\partial \log(f(\mathbf{x};\boldsymbol{\gamma},\mathbf{U}_0))}{\partial \boldsymbol{\gamma}} \tag{21}$$
$$h_2(\boldsymbol{\theta}) = \frac{\partial \log(f(\mathbf{x};\boldsymbol{\gamma},\mathbf{U}))}{\partial \mathbf{U}}. \tag{22}$$

The estimating function estimator is then found by the solving the estimating equation in (20). Compared to the ML-equations, the only change is that $\mathbf{U}$ is replaced by $\mathbf{U}_0$ in the first set of equations. Since $\mathbf{U}_0$ is fixed, (21) only involves $\boldsymbol{\gamma}$ and can be solved separately from (22), which is one of the main advantage of this approach.

Further, let the estimating function in (21) be used to solve a estimating function. This is equivalent to

$$\frac{\partial \sum_{i=1}^{n} \log(g_j(x_{i,j};\boldsymbol{\gamma}_j))}{\partial \boldsymbol{\gamma}_j} = \mathbf{0}, \quad j = 1, ..., p, \tag{23}$$

which corresponds to maximization of the marginal likelihoods for $\boldsymbol{\gamma}_j$.

Assume $\hat{\boldsymbol{\gamma}}$ is the solution of (23). The estimate for $\mathbf{U}$ can then be found from the estimating equation

$$\frac{\partial \log(f(\mathbf{x}; \hat{\boldsymbol{\gamma}}, \mathbf{U}))}{\partial \mathbf{U}} = \mathbf{0}.$$

This is still a complicated system to solve, but much simpler than the full Maximum likelihood approach. Experience with numerical procedures for solving this equation system shows that the computational gain is very large compared to full Maximum likelihood estimation. This is partly due to that solving the estimating equation of the estimating functions in (21) wrt $\boldsymbol{\gamma}$ and (22) wrt to $\mathbf{U}$, the information in the data can be described by a few sufficient statistics. This is in contrast to full ML estimation where the transformation $\mathbf{x} \rightarrow \mathbf{y}(\mathbf{x}; \boldsymbol{\gamma})$ has to be performed for each function evaluation.

In order to have consistent estimator, one important feature is to show that

$$E[\frac{\partial \sum_{i=1}^{n} \log(g_j(x_{i,j}; \boldsymbol{\gamma}_j))}{\partial \boldsymbol{\gamma}_j}] = \mathbf{0}, \quad j = 1, ..., p.$$

This equation corresponds to score functions for the marginal likelihoods, and therefore the statement is valid. Since the second set of equations is equivalent to the ML equations, also this set of estimation functions will have zero expectation. Using the general theory of estimation functions Liang and Zeger (1995) it can then be shown that the estimates obtained by solving the estimating equation of the estimating functions in (21) and (22) are asymptotically consistent and normally distributed. The asymptotic variances (Godambe information) for the estimates will differ compared to the ML estimates, but our experience is that the efficiency loss is small.

6 Test

The pixelwise Bayes classification rule (14) has been used to examine whether the use of meta-Gaussian distributions significantly improves the classification accuracy compared to marginal Gamma distributions that are assumed to be independent. It should be stressed that the focus is not on achieving the highest possible classification accuracy, but on revealing differences between the two approaches.

The data set considered here consists of a multi-temporal series of 6 ERS-1 images of Bourges, France. The images were acquired with monthly intervals during the summer season 1993, and 4-look amplitude images were generated from the original SLC images. The training set consists of vectors of amplitude observations from 21 523 pixels where the ground truth (class label) is known. The test data set contains 63 457 pixels. Table 1 contains the name, label value and number of pixels in training set and test set of each of the 15 classes.

The training set is used to estimate the parameters of the models and to construct the classification rule. The test set is used to find the probabilities of correct classification on the basis of the classification rule.

We compare two approaches. One consists in assuming that all components are independent with Gamma marginals. ML is used to estimate the parameters involved in this case. We will denote this method by independent maximum likelihood (IML). The other approach is the meta-Gaussian with Gamma marginals. For this model, both ML estimation and the use of estimation functions are considered. These methods are denoted by MML and MEF, respectively.

For both models, the marginal distributions (2) are described by parameters $\boldsymbol{\gamma}_j = (L_j, R_j)$. For the meta-Gaussian model, the dependence is described through the correlation matrices $\boldsymbol{\Sigma}$ on the Gaussian scale (one for each class).

The overall portion of correctly classified pixels in the test set for the three methods seen in Table 6 were 0.387 (IML), 0.400 (MML) and 0.397 (MEF), i.e., the differences are very small. Tables 7 (IML), 8 (MEF) and 9 (MML) shows the confusion matrices. The results for MML were similar to those of MEF. We would expect that high correlations within a class would give less confusion with other classes when taking the covariances into account (MEF) than when assuming independence (IML). This is mostly the case, but there are exceptions.

To further investigate the impact of the magnitude of the inter-image correlation, we performed classification into a reduced number of classes, corresponding to those having the strongest correlation between components, which were the ones with labels 6, 8, 17, 20, and 24. In this case, the overall portion of correctly classified pixels for the three classification rules seen in Table 6 were 0.411 (IML), 0.475 (MEF), and 0.458 (MML), i.e., a significant improvement is obtained by incorporating covariance through meta-Gaussian distributions. We also investigated the impact of the models on simulated data. The simulated data where obtained by simulating from a meta-Gaussian distribution (Gamma marginals) for each class. The class parameters where given from EF estimates from the real data, where the parameter estimates from the marginal distribution are kept untouched. Since higher covariance for each class is needed, we did the following. The cholescky decomposed covariance matrix where multiplied by a factor 5 on the off diagonal elements. This procedure will give higher correlations in the covariance matrix and the same relative change for each class. Note that marginal distribution of each class is Gamma with approximately the same parameter estimates as the real data. The only change is higher correlation between the channels. Note also that the same number of data as the real data are produced for each class. The overall portion of correctly classified pixels for the three classification rules were 0.420 (IML), 0.489 (MEF), and 0.483 (MML), i.e., a significant improvement is obtained by incorporating covariance through meta-Gaussian distributions.

To further investigate the impact of the magnitude of the inter-image correlation on the simulated data, we performed classification into a reduced number of classes, corresponding to those having the strongest correlation between components, which were the ones with labels 6, 8, 17, 20, and 24. In this case, the overall portion of correctly classified pixels for the three classification rules were 0.418 (IML), 0.513 (MEF), and 0.499 (MML), i.e., a significant improvement is obtained by incorporating covariance through meta-Gaussian distributions.

Note that if we compare the results from MEF and MLE there is very little difference.

Actually, MEF is doing a bit better in classification. This is probable due to a more robust procedure for giving estimates.

Table 1: Class labels, class name and number of pixels.

Class label	Class name	Number of pixels in training set	Number of pixels in test set
1	forest	2559	11985
3	orchard	48	66
4	hard wheat	2985	8195
5	soft wheat	2264	5782
6	maize	2876	10598
7	sunflower	2384	5479
8	barley	141	161
9	oilseed rape	2749	7012
10	peas	623	1573
11	clover	488	793
14	prairie	722	1899
17	bare soil	1162	2993
20	road	404	923
21	water	537	1990
24	urban area	1581	4008

Table 2: L estimates for each class.

Class	L_1	L_2	L_3	L_4	L_5	L_6
1	3.72	3.34	3.07	3.29	3.48	3.48
3	4.04	2.73	2.56	2.00	4.92	4.42
4	2.90	2.82	3.31	3.09	2.62	3.14
5	2.65	2.94	3.33	2.55	2.65	2.41
6	1.89	1.47	0.93	3.50	3.41	3.11
7	3.05	1.60	3.01	3.67	3.32	2.66
8	3.31	1.62	3.59	3.88	2.96	3.21
9	2.69	2.79	4.03	3.02	2.89	2.47
10	3.27	3.73	3.59	2.65	2.84	1.76
11	3.10	3.71	3.22	3.02	3.74	3.73
14	3.27	3.06	3.03	3.05	3.26	3.37
17	2.67	1.69	2.06	2.87	2.92	3.22
20	1.42	1.24	1.29	1.26	1.45	1.25
21	3.52	1.19	1.72	1.33	2.82	1.99
24	1.09	1.13	1.19	1.22	1.19	1.11

In Table 6 the probabilities of correct classification are shown for each of the 15 classes and for each of the 3 different classification rules.

Table 3: R estimates (scaled by a factor of 10000^{-1}).

Class	R_1	R_2	R_3	R_4	R_5	R_6
1	1.32	1.55	1.53	1.51	1.38	1.28
3	0.81	0.92	0.78	1.08	0.79	0.93
4	0.62	0.50	0.42	0.52	0.89	1.48
5	0.51	0.46	0.47	0.53	0.72	1.20
6	1.04	2.30	2.90	1.10	1.12	1.20
7	0.67	1.51	1.77	2.17	0.98	1.54
8	0.71	0.50	1.24	0.79	1.13	1.18
9	0.77	1.03	2.25	0.85	0.83	1.36
10	0.40	1.27	1.34	1.43	0.59	2.04
11	0.65	0.89	1.41	1.25	1.14	1.56
14	0.70	0.70	0.70	0.72	0.56	1.07
17	1.03	2.72	1.55	1.27	0.91	1.42
20	0.44	0.58	0.61	0.49	0.43	0.59
21	0.84	0.24	0.73	0.38	0.95	0.1
24	3.19	3.06	3.04	3.05	3.04	3.17

Table 4: Displaying probabilities of correct classification for each class and 4 methods.

Class	IML	EE	MLE
1	0.472	0.471	0.482
3	0.394	0.394	0.409
4	0.431	0.453	0.429
5	0.384	0.366	0.395
6	0.184	0.239	0.223
7	0.334	0.345	0.345
8	0.447	0.435	0.416
9	0.428	0.438	0.45
10	0.528	0.526	0.526
11	0.295	0.31	0.314
14	0.351	0.345	0.341
17	0.241	0.273	0.28
20	0.556	0.531	0.492
21	0.828	0.841	0.841
24	0.415	0.396	0.357
Total	0.387	0.400	0.397

Table 5: Displaying correlation of class 24.

1	0.245	0.123	0.187	0.392	0.519
0.245	1	0.085	0.130	0.300	0.546
0.123	0.085	1	0.592	0.549	0.559
0.187	0.130	0.592	1	0.679	0.691
0.392	0.300	0.549	0.679	1	0.896
0.519	0.546	0.559	0.691	0.896	1

Table 6: Displaying probabilities of correct classification for each class and 3 methods.

Class	IML	EF	ML
6	0.300	0.412	0.405
8	0.839	0.795	0.789
17	0.558	0.545	0.554
20	0.778	0.739	0.72
24	0.493	0.513	0.452
Total	0.411	0.475	0.458

Table 7: Displaying Confusion Matrix for IML Method. C is correct class while $\widehat{C}$ is estimated class of the IML method. Overall probability of correct classification is 0.387.

$C \setminus \widehat{C}$	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.472	0.128	0.007	0.003	0.021	0.058	0.034	0.067	0.021	0.075	0.030	0.063	0.002	0.002	0.018
3	0.076	0.394	0.015	0.061	0.061	0.030	0.045	0.030	0.106	0.045	0.106	0.000	0.030	0.000	0.000
4	0.006	0.050	0.431	0.265	0.002	0.001	0.048	0.002	0.017	0.019	0.109	0.002	0.042	0.005	0.002
5	0.007	0.048	0.249	0.384	0.003	0.001	0.048	0.004	0.017	0.023	0.106	0.002	0.095	0.010	0.003
6	0.130	0.088	0.018	0.005	0.184	0.053	0.052	0.102	0.048	0.100	0.033	0.111	0.004	0.001	0.072
7	0.080	0.069	0.011	0.005	0.028	0.334	0.043	0.057	0.120	0.158	0.019	0.048	0.002	0.001	0.023
8	0.062	0.087	0.056	0.037	0.000	0.012	0.447	0.050	0.012	0.075	0.118	0.012	0.019	0.012	0.000
9	0.070	0.055	0.009	0.009	0.042	0.036	0.099	0.428	0.080	0.083	0.051	0.023	0.007	0.003	0.004
10	0.010	0.051	0.009	0.010	0.006	0.116	0.025	0.074	0.528	0.076	0.053	0.022	0.013	0.000	0.005
11	0.076	0.098	0.050	0.019	0.015	0.086	0.097	0.103	0.086	0.295	0.058	0.009	0.003	0.001	0.004
14	0.007	0.096	0.091	0.197	0.001	0.006	0.064	0.028	0.046	0.023	0.351	0.005	0.077	0.008	0.000
17	0.158	0.098	0.014	0.007	0.058	0.065	0.038	0.079	0.069	0.073	0.054	0.241	0.007	0.002	0.038
20	0.017	0.039	0.015	0.135	0.003	0.010	0.026	0.011	0.016	0.014	0.131	0.008	0.556	0.013	0.005
21	0.002	0.004	0.005	0.020	0.000	0.001	0.005	0.003	0.001	0.001	0.008	0.000	0.121	0.828	0.003
24	0.278	0.059	0.008	0.008	0.015	0.044	0.024	0.023	0.012	0.039	0.030	0.034	0.005	0.004	0.415

Table 8: Displaying confusion matrix if the MEF method. C is correct class while $\widehat{C}$ is estimated class of MEF method. Overall probability of correct classification is 0.400.

$C \setminus \widehat{C}$	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.471	0.114	0.007	0.003	0.024	0.059	0.033	0.063	0.021	0.077	0.030	0.069	0.003	0.003	0.023
3	0.076	0.394	0.030	0.030	0.030	0.045	0.045	0.015	0.091	0.061	0.106	0.030	0.030	0.000	0.015
4	0.005	0.059	0.453	0.250	0.006	0.002	0.039	0.002	0.016	0.018	0.107	0.001	0.036	0.007	0.001
5	0.006	0.056	0.281	0.366	0.004	0.003	0.043	0.004	0.018	0.021	0.108	0.002	0.076	0.012	0.001
6	0.126	0.087	0.017	0.009	0.239	0.050	0.045	0.105	0.045	0.104	0.030	0.106	0.004	0.002	0.032
7	0.073	0.066	0.009	0.011	0.035	0.345	0.041	0.060	0.126	0.151	0.020	0.046	0.003	0.001	0.013
8	0.050	0.118	0.037	0.075	0.000	0.012	0.435	0.043	0.012	0.062	0.099	0.012	0.019	0.025	0.000
9	0.066	0.062	0.009	0.009	0.038	0.034	0.085	0.438	0.078	0.093	0.047	0.024	0.009	0.005	0.003
10	0.007	0.055	0.010	0.008	0.009	0.120	0.024	0.076	0.526	0.078	0.051	0.022	0.013	0.001	0.003
11	0.082	0.087	0.043	0.030	0.014	0.076	0.088	0.105	0.091	0.310	0.057	0.008	0.006	0.000	0.004
14	0.008	0.113	0.092	0.162	0.002	0.007	0.060	0.028	0.044	0.025	0.345	0.005	0.099	0.008	0.001
17	0.141	0.100	0.014	0.012	0.061	0.059	0.033	0.083	0.073	0.077	0.049	0.273	0.006	0.002	0.016
20	0.020	0.054	0.016	0.150	0.002	0.010	0.027	0.011	0.023	0.012	0.113	0.009	0.531	0.013	0.011
21	0.003	0.012	0.007	0.022	0.000	0.000	0.003	0.002	0.002	0.001	0.007	0.000	0.096	0.841	0.008
24	0.283	0.052	0.009	0.007	0.015	0.049	0.022	0.022	0.014	0.044	0.032	0.041	0.007	0.005	0.396

Table 9: Displaying confusion matrix. C is correct class while $\widehat{C}$ is estimated class of ML method. Overall probability of correct classification is 0.397.

$C \setminus \widehat{C}$	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.482	0.128	0.007	0.003	0.023	0.055	0.029	0.065	0.023	0.071	0.028	0.068	0.003	0.004	0.012
3	0.076	0.409	0.030	0.030	0.015	0.03	0.045	0.015	0.106	0.061	0.106	0.030	0.030	0.000	0.015
4	0.005	0.065	0.429	0.280	0.006	0.002	0.036	0.001	0.015	0.018	0.106	0.001	0.029	0.007	0.000
5	0.006	0.060	0.267	0.395	0.004	0.004	0.041	0.004	0.018	0.020	0.108	0.002	0.060	0.011	0.001
6	0.132	0.099	0.018	0.009	0.223	0.048	0.041	0.107	0.045	0.098	0.028	0.107	0.004	0.002	0.039
7	0.079	0.069	0.009	0.01	0.033	0.345	0.039	0.063	0.126	0.148	0.020	0.047	0.003	0.001	0.008
8	0.050	0.130	0.043	0.075	0.000	0.012	0.416	0.043	0.019	0.068	0.087	0.012	0.019	0.025	0.000
9	0.067	0.072	0.009	0.010	0.035	0.030	0.079	0.450	0.078	0.086	0.043	0.026	0.008	0.005	0.002
10	0.011	0.056	0.010	0.009	0.008	0.117	0.020	0.077	0.526	0.077	0.050	0.024	0.012	0.001	0.002
11	0.083	0.092	0.042	0.033	0.013	0.069	0.079	0.107	0.096	0.314	0.054	0.010	0.006	0.000	0.001
14	0.007	0.131	0.085	0.181	0.002	0.006	0.058	0.029	0.047	0.021	0.341	0.005	0.078	0.009	0.001
17	0.149	0.115	0.014	0.011	0.057	0.056	0.031	0.083	0.073	0.071	0.044	0.280	0.007	0.002	0.008
20	0.022	0.059	0.016	0.179	0.002	0.009	0.022	0.014	0.023	0.012	0.120	0.009	0.492	0.014	0.009
21	0.003	0.009	0.007	0.024	0.000	0.000	0.003	0.002	0.002	0.001	0.008	0.000	0.092	0.841	0.009
24	0.317	0.059	0.009	0.007	0.016	0.049	0.020	0.025	0.015	0.040	0.030	0.044	0.006	0.005	0.357

Table 10: Displaying confusion matrix. C is correct class while $\widehat{C}$ is estimated class of ICM method with Gamma. Overall probability of correct classification is 0.589.

$C \setminus \widehat{C}$	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.853	0.079	0.001	0.001	0.000	0.008	0.004	0.012	0.003	0.015	0.009	0.012	0.001	0.000	0.003
3	0.030	0.955	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.000	0.000	0.000	0.000
4	0.001	0.018	0.608	0.286	0.000	0.000	0.010	0.001	0.005	0.002	0.062	0.001	0.005	0.000	0.001
5	0.009	0.019	0.261	0.57	0.001	0.001	0.021	0.003	0.008	0.017	0.045	0.002	0.042	0.000	0.002
6	0.134	0.084	0.006	0.002	0.200	0.022	0.029	0.094	0.019	0.150	0.024	0.180	0.000	0.000	0.057
7	0.057	0.029	0.003	0.002	0.007	0.573	0.008	0.031	0.049	0.188	0.008	0.030	0.000	0.000	0.015
8	0.075	0.019	0.019	0.000	0.000	0.000	0.609	0.000	0.000	0.099	0.180	0.000	0.000	0.000	0.000
9	0.039	0.009	0.002	0.008	0.004	0.013	0.058	0.779	0.036	0.030	0.015	0.007	0.000	0.000	0.000
10	0.003	0.011	0.004	0.005	0.000	0.102	0.003	0.046	0.764	0.045	0.006	0.011	0.000	0.000	0.000
11	0.049	0.034	0.028	0.000	0.003	0.024	0.033	0.062	0.014	0.656	0.095	0.004	0.000	0.000	0.000
14	0.002	0.045	0.038	0.202	0.000	0.006	0.014	0.021	0.019	0.009	0.603	0.000	0.039	0.000	0.000
17	0.191	0.100	0.001	0.001	0.034	0.026	0.025	0.076	0.032	0.075	0.051	0.350	0.002	0.000	0.035
20	0.018	0.013	0.008	0.141	0.000	0.010	0.013	0.011	0.009	0.018	0.174	0.016	0.560	0.004	0.004
21	0.004	0.002	0.004	0.010	0.000	0.000	0.002	0.002	0.000	0.000	0.004	0.000	0.100	0.874	0.000
24	0.392	0.062	0.003	0.004	0.004	0.016	0.005	0.004	0.002	0.013	0.022	0.007	0.002	0.002	0.459

Table 11: Displaying confusion matrix. C is correct class while $\widehat{C}$ is estimated class of ICM method with Metagamma. Overall probability of correct classification is 0.620.

$C \setminus \widehat{C}$	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.871	0.051	0.000	0.001	0.003	0.009	0.003	0.014	0.003	0.017	0.008	0.015	0.001	0.000	0.003
3	0.061	0.742	0.000	0.000	0.000	0.030	0.015	0.000	0.030	0.030	0.091	0.000	0.000	0.000	0.000
4	0.001	0.027	0.636	0.251	0.000	0.001	0.008	0.001	0.006	0.004	0.054	0.000	0.010	0.000	0.001
5	0.008	0.027	0.304	0.503	0.001	0.001	0.015	0.003	0.012	0.014	0.055	0.001	0.054	0.001	0.001
6	0.119	0.090	0.008	0.002	0.277	0.024	0.027	0.096	0.014	0.161	0.015	0.161	0.000	0.000	0.006
7	0.052	0.029	0.002	0.001	0.009	0.606	0.010	0.031	0.064	0.154	0.007	0.024	0.001	0.000	0.009
8	0.075	0.068	0.012	0.000	0.000	0.000	0.578	0.019	0.000	0.093	0.155	0.000	0.000	0.000	0.000
9	0.028	0.019	0.003	0.007	0.003	0.010	0.048	0.794	0.036	0.033	0.010	0.007	0.001	0.000	0.001
10	0.001	0.013	0.004	0.005	0.001	0.087	0.007	0.047	0.772	0.037	0.015	0.011	0.000	0.000	0.000
11	0.025	0.053	0.032	0.003	0.000	0.020	0.039	0.049	0.025	0.68	0.066	0.003	0.000	0.000	0.006
14	0.003	0.051	0.038	0.123	0.000	0.005	0.017	0.019	0.031	0.015	0.622	0.000	0.075	0.000	0.000
17	0.177	0.112	0.003	0.001	0.033	0.021	0.023	0.087	0.035	0.064	0.049	0.388	0.001	0.000	0.007
20	0.015	0.046	0.013	0.072	0.002	0.011	0.024	0.008	0.005	0.014	0.119	0.009	0.648	0.007	0.009
21	0.002	0.002	0.009	0.004	0.000	0.000	0.002	0.001	0.001	0.000	0.001	0.000	0.091	0.888	0.000
24	0.302	0.034	0.004	0.002	0.005	0.010	0.006	0.003	0.002	0.013	0.018	0.007	0.006	0.001	0.587

Table 12: Displaying confusion matrix. C is correct class while $\widehat{C}$ is estimated class of no contextual method with Metagamma. Overall probability of correct classification is 0.404.

$C \setminus \widehat{C}$	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.462	0.116	0.007	0.003	0.024	0.059	0.034	0.064	0.021	0.077	0.030	0.069	0.003	0.004	0.027
3	0.091	0.424	0.015	0.045	0.000	0.045	0.045	0.015	0.091	0.061	0.106	0.045	0.000	0.000	0.015
4	0.005	0.059	0.452	0.236	0.006	0.002	0.045	0.002	0.016	0.018	0.105	0.001	0.045	0.007	0.001
5	0.006	0.056	0.276	0.337	0.005	0.003	0.045	0.004	0.019	0.020	0.107	0.001	0.106	0.012	0.002
6	0.126	0.088	0.017	0.007	0.253	0.051	0.047	0.103	0.046	0.104	0.030	0.111	0.004	0.002	0.013
7	0.073	0.066	0.009	0.008	0.033	0.351	0.044	0.060	0.124	0.151	0.023	0.044	0.000	0.001	0.013
8	0.050	0.118	0.043	0.062	0.000	0.012	0.46	0.031	0.012	0.068	0.099	0.012	0.012	0.019	0.000
9	0.065	0.061	0.009	0.008	0.037	0.034	0.090	0.441	0.078	0.092	0.045	0.027	0.008	0.005	0.003
10	0.008	0.055	0.010	0.007	0.010	0.123	0.024	0.077	0.53	0.078	0.054	0.022	0.004	0.000	0.000
11	0.074	0.091	0.044	0.025	0.018	0.073	0.086	0.103	0.093	0.318	0.058	0.006	0.005	0.000	0.005
14	0.008	0.116	0.095	0.147	0.004	0.006	0.063	0.027	0.050	0.023	0.335	0.004	0.113	0.009	0.001
17	0.140	0.098	0.013	0.010	0.058	0.060	0.036	0.081	0.072	0.079	0.050	0.276	0.005	0.002	0.019
20	0.020	0.052	0.013	0.096	0.003	0.010	0.028	0.012	0.022	0.012	0.098	0.009	0.603	0.014	0.009
21	0.003	0.009	0.008	0.013	0.001	0.001	0.005	0.003	0.002	0.001	0.009	0.000	0.089	0.854	0.005
24	0.241	0.054	0.009	0.004	0.016	0.043	0.023	0.023	0.014	0.042	0.029	0.033	0.011	0.005	0.453

Table 13: Displaying confusion matrix. C is correct class while $\widehat{C}$ is estimated class of no contextual IML method with Gamma. Overall probability of correct classification is 0.387.

$C \setminus \widehat{C}$	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.472	0.128	0.007	0.003	0.020	0.058	0.034	0.067	0.021	0.075	0.030	0.064	0.002	0.002	0.018
3	0.076	0.394	0.015	0.061	0.061	0.030	0.045	0.030	0.106	0.045	0.106	0.000	0.030	0.000	0.000
4	0.006	0.050	0.431	0.265	0.002	0.001	0.049	0.002	0.017	0.019	0.109	0.002	0.042	0.005	0.001
5	0.007	0.048	0.249	0.384	0.003	0.001	0.048	0.004	0.017	0.023	0.106	0.002	0.095	0.010	0.003
6	0.130	0.088	0.018	0.005	0.186	0.053	0.052	0.101	0.048	0.100	0.033	0.112	0.003	0.001	0.070
7	0.081	0.069	0.011	0.005	0.028	0.335	0.043	0.057	0.120	0.158	0.019	0.048	0.002	0.001	0.023
8	0.062	0.087	0.056	0.037	0.000	0.012	0.447	0.050	0.012	0.075	0.118	0.012	0.019	0.012	0.000
9	0.070	0.055	0.009	0.009	0.043	0.036	0.099	0.427	0.079	0.084	0.051	0.024	0.007	0.003	0.004
10	0.010	0.051	0.009	0.010	0.005	0.118	0.025	0.075	0.526	0.076	0.053	0.024	0.013	0.000	0.005
11	0.076	0.098	0.050	0.019	0.015	0.088	0.097	0.103	0.086	0.293	0.058	0.009	0.003	0.001	0.004
14	0.007	0.096	0.091	0.196	0.001	0.005	0.064	0.028	0.046	0.023	0.351	0.005	0.077	0.008	0.000
17	0.158	0.099	0.014	0.007	0.058	0.065	0.038	0.079	0.068	0.073	0.054	0.241	0.007	0.002	0.038
20	0.017	0.039	0.016	0.134	0.003	0.010	0.026	0.011	0.016	0.014	0.131	0.008	0.556	0.013	0.005
21	0.002	0.004	0.005	0.020	0.000	0.001	0.005	0.003	0.001	0.001	0.008	0.000	0.121	0.828	0.003
24	0.278	0.059	0.008	0.008	0.015	0.044	0.024	0.023	0.012	0.039	0.030	0.034	0.005	0.004	0.415

A Derivatives

Both the maximum likelihood and the estimating function estimator can be found by maximization of a target function. In the maximum likelihood case the target function is the likelihood or loglikelihood. For the estimating function estimator case, the target function depends on which estimating function that is decided on. Here also the likelihood is chosen, but some of the parameters are fixed. Maximization can be performed numerically by only target function evaluations or by also including evaluations of derivative of the target function. In this section we will find the derivatives of a target function which is the likelihood both for estimating function and maximum likelihood. We have by (19)

$$\log(f(\mathbf{x};\boldsymbol{\gamma},\mathbf{U})) = -\frac{n}{2}\log|\Sigma(U)| - \frac{n}{2}\mathbf{tr}A(U)\mathbf{S}(\boldsymbol{\gamma}) + \sum_{i=1}^{n}\sum_{j=1}^{p}\log(g_j(x_{i,j};\boldsymbol{\gamma}_j)), \tag{24}$$

where

$$\begin{aligned}\Sigma(U) =& \mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{-1/2}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{-1/2}\\ A(U) =& \mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2}\mathbf{U}^T\mathbf{U}\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2} - \mathbf{I}.\end{aligned}$$

Note that $\mathbf{D}$ in $\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})$ means diagonal matrix of $(\mathbf{U}^T\mathbf{U})^{-1}$. Furthermore, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1,\dots,\boldsymbol{\gamma}_p)$ and $\boldsymbol{\gamma}_j$ is a vector of parameters describing the marginal distribution of component j. The first term in (24) (leaving the constants behind) is

$$\begin{aligned}\log|\Sigma(U)| =& \log|\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{-1}| + \log|(\mathbf{U}^T\mathbf{U})^{-1}|\\ =& -\log|\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})| - \log|\mathbf{U}^T\mathbf{U}|\\ =& -\log|\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})| - 2\log|\mathbf{U}|\\ =& -\log|\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})|,\end{aligned} \tag{25}$$

where $\log|\mathbf{U}| = 0$ because $\mathbf{U}$ is a lower triangular matrix and has ones on the diagonal. Let $\mathbf{B} = \mathbf{B}(\mathbf{U}) = \mathbf{U}^T\mathbf{U}$. Then

$$\begin{aligned}\frac{\partial}{\partial\mathbf{U}}\log|\mathbf{D}(\mathbf{B}^{-1})| =& \frac{\partial\mathbf{B}}{\partial\mathbf{U}}\frac{\partial\mathbf{B}^{-1}}{\partial\mathbf{B}}\frac{\partial\mathbf{D}(\mathbf{B}^{-1})}{\partial\mathbf{B}^{-1}}\frac{\partial}{\partial\mathbf{D}(\mathbf{B}^{-1})}\log|\mathbf{D}(\mathbf{B}^{-1})|\\ =& \frac{\partial\mathbf{B}}{\partial\mathbf{U}}\frac{\partial\mathbf{B}^{-1}}{\partial\mathbf{B}}\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{-1}.\end{aligned} \tag{26}$$

The first term in (26) is

$$\frac{\partial\mathbf{B}}{\partial\mathbf{U}_{ij}} = \frac{\partial\mathbf{U}^T\mathbf{U}}{\partial\mathbf{U}_{ij}} = \mathbf{J}_{ji}\mathbf{U} + \mathbf{U}^T\mathbf{J}_{ij}, \tag{27}$$

where $\mathbf{J}_{ij}$ denotes a matrix with a 1 in the (i,j)th place and zeros elsewhere. The second term in (26) is

$$\frac{\partial\mathbf{B}^{-1}}{\partial\mathbf{B}_{ij}} = \left\{\begin{array}{ll} -\mathbf{B}^{-1}\mathbf{J}_{ii}\mathbf{B}^{-1}, & i=j\\ -\mathbf{B}^{-1}(\mathbf{J}_{ij}+\mathbf{J}_{ji})\mathbf{B}^{-1}, & i\neq j\end{array}\right\}. \tag{28}$$

Taking the derivatives of the second term in (24) (leaving the constants behind) gives

$$\begin{aligned}\frac{\partial}{\partial\mathbf{U}}\mathbf{tr}(\mathbf{A}(\mathbf{U})\mathbf{S}(\boldsymbol{\gamma})) =& \frac{\partial\mathbf{A}(\mathbf{U})}{\partial\mathbf{U}}\frac{\partial}{\partial\mathbf{A}(\mathbf{U})}\mathbf{tr}(\mathbf{A}(\mathbf{U})\mathbf{S}(\boldsymbol{\gamma}))\\ =& \frac{\partial\mathbf{A}(\mathbf{U})}{\partial\mathbf{U}}\left(2\mathbf{S}(\boldsymbol{\gamma}) - \mathbf{D}(\mathbf{S}(\boldsymbol{\gamma}))\right).\end{aligned} \tag{29}$$

The first term in (29) can be calculated in the following way

$$\begin{aligned}\frac{\partial \mathbf{A}(\mathbf{U})}{\partial \mathbf{U}} =& \frac{\partial \mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2}\mathbf{U}^T\mathbf{U}\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2}}{\partial \mathbf{U}} \\ =& \frac{\partial \mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2}}{\partial \mathbf{U}}\mathbf{U}^T\mathbf{U}\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2} \\ &+ \mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2}\frac{\partial \mathbf{U}^T\mathbf{U}}{\partial \mathbf{U}}\mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2} \\ &+ \mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2}\mathbf{U}^T\mathbf{U}\frac{\partial \mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2}}{\partial \mathbf{U}}.\end{aligned} \tag{30}$$

The first and third partial derivatives in (30) is

$$\begin{aligned}\frac{\partial \mathbf{D}((\mathbf{U}^T\mathbf{U})^{-1})^{1/2}}{\partial \mathbf{U}} =& \frac{\partial \mathbf{D}(\mathbf{B}^{-1})^{1/2}}{\partial \mathbf{U}} \\ =& \frac{\partial \mathbf{B}}{\partial \mathbf{U}}\frac{\partial \mathbf{B}^{-1}}{\partial \mathbf{B}}\frac{\partial \mathbf{D}(\mathbf{B}^{-1})^{1/2}}{\partial \mathbf{B}^{-1}} \\ =& \frac{\partial \mathbf{B}}{\partial \mathbf{U}}\frac{\partial \mathbf{B}^{-1}}{\partial \mathbf{B}}\frac{\partial \mathbf{D}(\mathbf{B}^{-1})}{\partial \mathbf{B}^{-1}}\frac{1}{2}\mathbf{D}(\mathbf{B}^{-1})^{-1/2},\end{aligned} \tag{31}$$

where $\frac{\partial \mathbf{D}(\mathbf{B}^{-1})}{\partial \mathbf{B}^{-1}}$ is the identity matrix. The first two terms in (31) are given in (27) and (28). The second partial derivatives in (30) is given in (27). This completes the derivatives of the log likelihood given in (24) wrt to $\mathbf{U}$. Both the maximum likelihood and estimation function estimator will have the same derivatives in this case. Furthermore, we need the derivatives of the log likelihood given in (24) wrt to $\boldsymbol{\gamma}$. Taking the derivatives of the second term in (24) wrt to $\boldsymbol{\gamma}$ (leaving the constants behind) gives

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\gamma}}\mathbf{tr}(\mathbf{A}(\mathbf{U})\mathbf{S}(\boldsymbol{\gamma})) =& \frac{\partial}{\partial \boldsymbol{\gamma}}\mathbf{tr}(\mathbf{S}(\boldsymbol{\gamma})\mathbf{A}(\mathbf{U})) \\ =& \frac{\partial \mathbf{S}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\frac{\partial}{\partial \mathbf{S}(\boldsymbol{\gamma})}\mathbf{tr}(\mathbf{S}(\boldsymbol{\gamma})\mathbf{A}(\mathbf{U})) \\ =& \frac{\partial \mathbf{S}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\left(2A(\mathbf{U}) - \mathbf{D}(\mathbf{A}(\mathbf{U}))\right).\end{aligned} \tag{32}$$

Note that the second term in (24) which has the partial derivatives given in (32) is only needed in the maximum likelihood case. This term is fixed wrt to $\boldsymbol{\gamma}$ in the estimating equation case. The partial derivative left to solve from (32) is

$$\begin{aligned}\frac{\partial \mathbf{S}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} =& \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\frac{\partial}{\partial \mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma})}\mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma})\mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma})^T \\ =& \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\left(\mathbf{Y}_{L,i}\mathbf{J}_{1,j} + \mathbf{J}_{j,1}\mathbf{Y}_{L,i}^T\right),\end{aligned} \tag{33}$$

where $\mathbf{Y}_{T,i} = (\mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma}), \ldots, \mathbf{y}(\mathbf{x}_i;\boldsymbol{\gamma}))^T$ is a $p \times p$ matrix and $\mathbf{Y}_{L,i}$ is lower triangular matrix of $\mathbf{Y}_{T,i}$. The partial derivative left to solve from (33) is

$$\begin{aligned}\frac{\partial y_j(x;\boldsymbol{\gamma}_j)}{\partial \boldsymbol{\gamma}_J} &= \frac{\partial}{\partial \boldsymbol{\gamma}_j}\Phi^{-1}(G_j(x;\boldsymbol{\gamma}_j)) \\ &= \frac{\partial G_j(x;\boldsymbol{\gamma}_j)/\partial \boldsymbol{\gamma}_j}{\phi(\Phi^{-1}(G_j(x;\boldsymbol{\gamma}_j)))}\end{aligned} \tag{34}$$

To be able to find the derivatives in (34), we need to know the specific marginal densities of the components. Therefore, we will give an example. Let all the marginals be *Gamma* distributed (see equation (8)). In the gamma case the derivatives of the marginal distributions has to be found. The marginal distributions are ($\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{0j}, \boldsymbol{\gamma}_{1j})$)

$$g_j(x; \boldsymbol{\gamma}_j) = \frac{\gamma_{0j}^{\gamma_{1j}}}{\Gamma(\gamma_{1j})} x^{\gamma_{1j}-1} \exp\{-\gamma_{0j} x\}, \quad x > 0.$$

The partial derivatives of the marginal distributions are

$$\begin{aligned} \frac{\partial}{\partial \gamma_{0j}} g_j(x; \gamma_{0j}, \gamma_{1j}) &= g_j(x; \gamma_{0j}, \gamma_{1j}) \left(\frac{\gamma_{1j}}{\gamma_{0j}} - x \right) \\ \frac{\partial}{\partial \gamma_{1j}} g_j(x; \gamma_{0j}, \gamma_{1j}) &= g_j(x; \gamma_{0j}, \gamma_{1j}) \left(\gamma_{0j} + \log(x) - \frac{\Gamma'(\gamma_{1j})}{\Gamma(\gamma_{1j})} \right). \end{aligned} \tag{35}$$

Using the results in (34) and (35) we get

$$\begin{aligned} \frac{\partial}{\partial \gamma_{0j}} y_j(x; \gamma_{0j}, \gamma_{1j}) &= \frac{\partial}{\partial \gamma_{0j}} \Phi^{-1}(G_j(x; \gamma_{0j}, \gamma_{1j}))) \\ &= \frac{\gamma_{1j}}{\gamma_{0j}} \frac{G_j(x; \gamma_{0j}, \gamma_{1j}) - G_j(x; \gamma_{0j}, \gamma_{1j}+1)}{\phi(\Phi^{-1}(G_j(x; \gamma_{0j}, \gamma_{1j})))}. \end{aligned}$$

The other partial derivative are not so easy to find and has to be solved numerically. Taking the derivatives of the third term in (24) wrt to $\boldsymbol{\gamma}$ involves derivatives of the marginal density functions. This are already found in (35).

B Further results

In this section we will explore radar images from both Feltwell and Bourges. This section was studied in the beginning of the project, mostly to learn about the subject and to decide on which path to continue for the rest of the project. Therefore, this section is put in the appendix so that it can be skipped to get the main message of the report. It is however kept in the report to show the development of the project.

B.1 Data

There are 3 (bands) radar images of Feltwell. In table 14, 6 classes are displayed with class name, class labels and number of pixels in each class.

Table 14: Displaying class label, class name and number of pixels. In parantese is the probability of class.

Class label	Class name	Number of pixels in training set	Number of pixels in test set
2		3595 (0.193)	3716 (0.198)
4		4642 (0.249)	4517 (0.241)
5		2512 (0.135)	2568 (0.137)
8		555 (0.030)	583 (0.031)
10		5697 (0.306)	5678 (0.303)
12		1643 (0.088)	1673 (0.089)

Table 15: Displaying probabilities of correct classification for each class and 5 methods.

Class	GammaMOM	GammaMLE	Lognormal0	BoxCox0	Lognormal1	BoxCox1	MixLognormal2	MixLognormal3
2	0.035	0.192	0.247	0.206	0.223	0.239	0.241	0.246
4	0.482	0.516	0.618	0.543	0.569	0.525	0.710	0.710
6	0.542	0.365	0.318	0.332	0.374	0.387	0.143	0.143
8	0.628	0.528	0.487	0.473	0.509	0.513	0.314	0.317
10	0.124	0.198	0.201	0.243	0.227	0.258	0.319	0.249
12	0.293	0.383	0.319	0.381	0.300	0.323	0.350	0.402
Total	0.281	0.323	0.346	0.339	0.344	0.350	0.376	0.361

The results from table (24) shows that there are not much to gain from using dependence between the different bands. The reason for this is the following. The means of the different classes are not very different. This means that the classification will have to use the covariance in order classify to the correct classes. Since the covariance between different bands are not very different for the different classes, there will not be much difference between assuming independence and dependence.

Table 16: Displaying estimated Gamma parameters $L_1, R_1, L_2, R_2, L_3, R_4$ by the method of moments for each class.

Class	L_1	R_1	L_2	R_2	L_3	R_3
2	0.6818	0.0004	1.1605	0.0003	0.8615	0.0006
4	1.8222	0.0006	2.1614	0.0004	1.9362	0.0007
6	2.1104	0.0009	2.2054	0.0004	1.9273	0.0010
8	0.9001	0.0009	1.1054	0.0005	0.7157	0.0008
10	0.5846	0.0004	0.7210	0.0004	0.9680	0.0009
12	0.8585	0.0007	0.6046	0.0003	0.6717	0.0007

Table 17: Displaying estimated Gamma parameters $L_1, R_1, L_2, R_2, L_3, R_4$ by the method of maximum likelihood for each class.

Class	L_1	R_1	L_2	R_2	L_3	R_3
2	1.5837	0.0009	1.5052	0.0004	1.9236	0.0014
4	2.1541	0.0007	2.1614	0.0004	2.1874	0.0008
6	2.1831	0.0009	2.0893	0.0004	2.1648	0.0012
8	1.3534	0.0013	1.6003	0.0007	1.3949	0.0016
10	1.3405	0.0010	1.1890	0.0007	1.5025	0.0014
12	1.3145	0.0011	0.9025	0.0005	1.2853	0.0014

We will now explore the Lognormal model further. The previous results are based on classification of a raw image. The next step is to filter pixel by pixel by taking into account the neighbors. In table 19, the image is filtered according to a different neighboring schemes. When no filtering is used, 1 neighbor is used. A filter with 3 neighbors will replace the target (the pixel in the middle) by a weighted mean of the three neighboring values.

The next step is to find a appropriate weight for the different neighbors. In the case of using k neighbor we denote the weight matrix as:

w_{11}	w_{12}	w_{13}	w_{14}	w_{15}
w_{21}	w_{22}	w_{23}	w_{14}	w_{25}
w_{31}	w_{32}	w_{33}	w_{34}	w_{35}
w_{41}	w_{42}	w_{43}	w_{44}	w_{45}
w_{51}	w_{52}	w_{53}	w_{54}	w_{55}

Depending on which neighboring strategy we use, some weights will be zero. In the case of eg using 3 neighbors there will only be positive weights to the following elements in the weight matrix given in (36) (w_{32}, w_{33}, w_{34}). There are two ways of filtering a lognormal model. Firstly, it is possible to take a weighted mean on log scale. Secondly, it is possible to take the log of the weighted mean. The first approach will be denoted by ML and the next will be denoted by LM. Specifying the weights will be done in the following way. First number after LM or ML will give the weight to the target pixel (w_{33}) and the last (or last two) number will specify the neighbor structure. LMn3 will give weights $(w_{23}, w_{33}, w_{34}) = (1/(2+n), n/(2+n), 1/(2+n))$. LMn5 will give weights $(w_{23}, w_{43}, w_{23}, w_{33}, w_{34}) = (1/(4+n), 1/(4+n), 1/(4+n), n/(4+n), 1/(4+n))$. LMn9 will give weights $(w_{22}, w_{23}, w_{24}, w_{42}, w_{43}, w_{44}, w_{23}, w_{33}, w_{34}) = (1/(8+n), 1/(8+n), 1/(8+n), 1/(8+n), 1/(8+n), 1/(8+n), 1/(8+n), n/(8+n), 1/(8+n))$.

LMn21 will give weights

0	$0.5/(n+15)$	$0.5/(n+15)$	$0.5/(n+15)$	0
$0.5/(n+15)$	$1/(n+15)$	$1/(n+15)$	$1/(n+15)$	$0.5/(n+15)$
$0.5/(n+15)$	$1/(n+15)$	$n/(n+15)$	$1/(n+15)$	$0.5/(n+15)$
$0.5/(n+15)$	$1/(n+15)$	$1/(n+15)$	$1/(n+15)$	$0.5/(n+15)$
0	$0.5/(n+15)$	$0.5/(n+15)$	$0.5/(n+15)$	0

Table 18: Displaying correlation matrix between the three bands for each of the 6 classes.

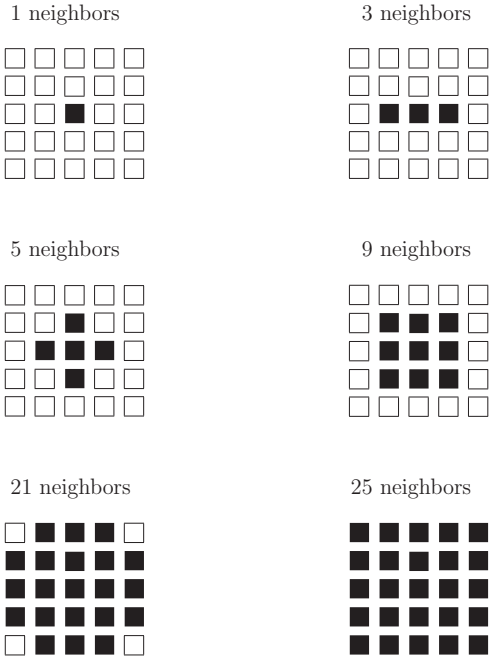
Band 1	Band 2	Band 3
Class 2		
1.000	0.302	0.519
0.302	1.000	0.128
0.519	0.128	1.000
Class 4		
1.000	0.142	0.574
0.142	1.000	0.122
0.574	0.122	1.000
Class 6		
1.000	0.186	0.379
0.186	1.000	0.146
0.379	0.146	1.000
Class 8		
1.000	0.403	0.601
0.403	1.000	0.432
0.601	0.432	1.000
Class 10		
1.000	0.185	0.713
0.185	1.000	0.240
0.713	0.240	1.000
Class 12		
1.000	0.478	0.650
0.478	1.000	0.358
0.650	0.358	1.000

LMn25 will give weights

$0.5/(n+17)$	$0.5/(n+17)$	$0.5/(n+17)$	$0.5/(n+17)$	$0.5/(n+17)$
$0.5/(n+17)$	$1/(n+17)$	$1/(n+17)$	$1/(n+17)$	$0.5/(n+17)$
$0.5/(n+17)$	$1/(n+17)$	$n/(n+17)$	$1/(n+17)$	$0.5/(n+17)$
$0.5/(n+17)$	$1/(n+17)$	$1/(n+17)$	$1/(n+17)$	$0.5/(n+17)$
$0.5/(n+17)$	$0.5/(n+17)$	$0.5/(n+17)$	$0.5/(n+17)$	$0.5/(n+17)$

In table (20) the results of the different methods are shown. The typical effect of replacing a mean of pixel values by the target of a pixel value is better classification when a wider range of neighbors are used. The difference between LM and ML is small.

Table 19: The pixel in the middle is filtered by a weighted mean of black neighbors.



Using the neighbor in filtering has a positive effect on better performances in classification. The total probability of correct classification is 0.34 for Lognormal model (with dependence) with one neighbor (no filtering). The probability goes up to 0.51 with filtering (25 neighbors). We will use the neighbors also in a slightly different setting. When classifying a pixel we have only used information about the pixel value for three bands. However, it is also possible to use the neighbor as part of the feature vector. Using 5 neighbor will give us a feature vector of length $3 \times 5 = 15$. From table (22) the best result is a probability of correct classification of 0.59.

Table 20: Displaying classification (with covariance) of the three bands for each of the 6 classes and for different neighboring methods.

Method	Class 2	Class 4	Class 6	Class 8	Class 10	Class 12	Total
LM13	0.361	0.644	0.461	0.626	0.310	0.293	0.430
ML13	0.368	0.626	0.453	0.633	0.291	0.342	0.424
LM23	0.354	0.644	0.452	0.636	0.310	0.298	0.428
ML23	0.356	0.625	0.447	0.617	0.282	0.337	0.418
LM15	0.414	0.680	0.505	0.657	0.314	0.321	0.459
ML15	0.411	0.652	0.483	0.654	0.300	0.366	0.448
LM25	0.409	0.675	0.502	0.664	0.315	0.328	0.457
ML25	0.399	0.656	0.480	0.668	0.291	0.369	0.445
LM35	0.390	0.668	0.489	0.655	0.305	0.327	0.447
ML35	0.387	0.654	0.469	0.666	0.286	0.359	0.437
LM19	0.449	0.721	0.566	0.686	0.294	0.378	0.484
ML19	0.440	0.701	0.542	0.717	0.294	0.433	0.480
LM39	0.423	0.708	0.546	0.693	0.284	0.377	0.470
ML39	0.419	0.685	0.522	0.712	0.281	0.415	0.464
LM59	0.413	0.688	0.524	0.677	0.278	0.368	0.457
ML59	0.398	0.667	0.484	0.698	0.282	0.389	0.447
LM121	0.486	0.758	0.639	0.709	0.260	0.469	0.507
ML121	0.475	0.742	0.589	0.774	0.295	0.507	0.511
LM221	0.473	0.756	0.622	0.717	0.256	0.470	0.502
ML221	0.473	0.738	0.587	0.762	0.286	0.500	0.506
LM321	0.464	0.745	0.605	0.724	0.252	0.458	0.493
ML321	0.461	0.724	0.578	0.755	0.277	0.493	0.495
LM125	0.481	0.763	0.651	0.726	0.249	0.470	0.507
ML125	0.476	0.746	0.601	0.765	0.293	0.524	0.514
LM225	0.477	0.759	0.646	0.731	0.251	0.474	0.505
ML225	0.469	0.742	0.598	0.771	0.290	0.518	0.510
LM325	0.467	0.749	0.628	0.720	0.249	0.469	0.497
ML325	0.461	0.730	0.593	0.755	0.283	0.502	0.501

Table 21: Displaying classification (assuming independence) of the three bands for each of the 6 classes.

Method	Class 2	Class 4	Class 6	Class 8	Class 10	Class 12	Total
LM13	0.373	0.707	0.392	0.580	0.295	0.311	0.434
ML13	0.390	0.689	0.369	0.573	0.289	0.330	0.429
LM23	0.376	0.702	0.386	0.575	0.300	0.313	0.433
ML23	0.383	0.690	0.364	0.563	0.281	0.332	0.425
LM15	0.425	0.743	0.396	0.580	0.308	0.301	0.455
ML15	0.436	0.727	0.366	0.562	0.310	0.333	0.452
LM25	0.417	0.739	0.390	0.581	0.310	0.301	0.452
ML25	0.432	0.729	0.373	0.595	0.307	0.326	0.452
LM35	0.407	0.733	0.392	0.581	0.304	0.304	0.448
ML35	0.416	0.719	0.37	0.578	0.304	0.325	0.445
LM19	0.447	0.770	0.425	0.606	0.325	0.296	0.475
ML19	0.473	0.752	0.373	0.588	0.339	0.338	0.477
LM39	0.442	0.763	0.418	0.588	0.320	0.299	0.470
ML39	0.465	0.749	0.376	0.592	0.330	0.337	0.472
LM59	0.422	0.750	0.413	0.581	0.313	0.306	0.461
ML59	0.436	0.734	0.374	0.588	0.317	0.342	0.459
LM121	0.476	0.783	0.462	0.605	0.280	0.321	0.477
ML121	0.505	0.754	0.378	0.597	0.340	0.364	0.486
LM221	0.477	0.781	0.455	0.603	0.279	0.326	0.476
ML221	0.500	0.756	0.375	0.603	0.339	0.360	0.485
LM321	0.478	0.780	0.451	0.612	0.281	0.328	0.476
ML321	0.492	0.758	0.387	0.621	0.334	0.356	0.484
LM125	0.478	0.778	0.467	0.621	0.249	0.349	0.471
ML125	0.502	0.750	0.390	0.587	0.331	0.372	0.484
LM225	0.478	0.781	0.466	0.626	0.251	0.352	0.472
ML225	0.503	0.753	0.392	0.594	0.333	0.365	0.486
LM325	0.475	0.780	0.465	0.621	0.254	0.345	0.471
ML325	0.496	0.754	0.392	0.610	0.330	0.365	0.484

Table 22: Displaying classification (assuming dependence) of the three bands times 5 neighbor for each of the 6 classes.

Method	Class 2	Class 4	Class 6	Class 8	Class 10	Class 12	Total
LM121	0.528	0.753	0.635	0.738	0.508	0.401	0.586
ML121	0.531	0.733	0.597	0.768	0.461	0.509	0.572
LM125	0.536	0.761	0.636	0.755	0.505	0.419	0.590
ML125	0.530	0.744	0.598	0.800	0.463	0.506	0.575
LM225	0.535	0.754	0.634	0.757	0.492	0.425	0.585
ML225	0.534	0.738	0.586	0.787	0.448	0.508	0.568

B.2 Bourges data set

The data set consists of a multitemporal series of 6 radar images of Bourges, France. The images are acquired in the summer season 1993. In table 1, 15 classes are displayed with class name, class labels and number of pixels in each class. In Table 16, the estimated values of the parameter L in the Gamma distribution for each class and each band are displayed.

Table 23: Displaying $L_1, L_2, \ldots, L_6$ for each class.

Class	L_1	L_2	L_3	L_4	L_5	L_6
1	3.71	3.33	3.04	3.28	3.47	3.47
3	3.94	2.64	2.92	2.36	5.16	4.63
4	2.86	2.80	3.28	2.66	2.20	2.85
5	2.60	2.52	3.05	2.23	2.46	1.79
6	0.82	0.90	0.35	3.22	3.46	2.62
7	2.56	1.34	2.79	3.65	3.22	2.42
8	3.29	1.24	3.51	3.94	2.74	3.17
9	2.66	2.77	4.04	2.98	2.85	2.43
10	3.24	3.73	3.57	2.61	2.81	1.36
11	2.54	3.64	3.02	2.64	3.78	3.52
14	3.28	3.04	3.00	2.68	3.11	3.17
17	2.10	1.21	1.53	2.55	2.75	3.03
20	1.05	1.04	1.22	0.94	1.33	0.98
21	3.49	0.66	1.92	1.26	2.59	0.90
24	0.24	0.26	0.35	0.42	0.36	0.24

In Table 18 the probabilities of correct classification are shown for each of the 15 classes and for each of the 5 different classification rules.

Table 24: Displaying probabilities of correct classification for each class and 4 methods.

Class	Gamma0 MOM	Gamma0 MLE	Lognormal0	Gamma1 MLE	Lognormal1	MixLognormal2	MixLognormal3
1	0.507	0.473	0.511	0.468	0.491	0.568	0.569
3	0.379	0.379	0.348	0.424	0.409	0.485	0.485
4	0.451	0.430	0.439	0.454	0.479	0.264	0.258
5	0.364	0.384	0.306	0.329	0.287	0.310	0.307
6	0.188	0.186	0.178	0.255	0.264	0.225	0.228
7	0.347	0.335	0.348	0.359	0.378	0.197	0.196
8	0.447	0.447	0.447	0.466	0.491	0.416	0.422
9	0.464	0.432	0.414	0.446	0.434	0.295	0.299
10	0.520	0.528	0.519	0.530	0.526	0.267	0.273
11	0.298	0.288	0.306	0.309	0.310	0.362	0.369
14	0.380	0.352	0.314	0.330	0.323	0.364	0.335
17	0.222	0.239	0.226	0.278	0.268	0.259	0.255
20	0.505	0.556	0.544	0.605	0.562	0.602	0.615
21	0.802	0.827	0.821	0.854	0.867	0.849	0.850
24	0.293	0.415	0.470	0.415	0.483	0.501	0.499
Total	0.391	0.388	0.388	0.403	0.412	0.360	0.359

In Table 25, the classification of probabilities for each class is displayed.

Table 25: Displaying classification within each class. Classification rule is Gamma0

Class	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.507	0.149	0.007	0.003	0.012	0.055	0.029	0.072	0.023	0.066	0.026	0.048	0.001	0.002	0.001
3	0.076	0.379	0.03	0.061	0.045	0.045	0.03	0.03	0.106	0.045	0.121	0	0.03	0	0
4	0.006	0.057	0.451	0.242	0.005	0.001	0.044	0.001	0.014	0.022	0.124	0.002	0.024	0.005	0
5	0.007	0.055	0.283	0.364	0.005	0.001	0.045	0.006	0.016	0.022	0.126	0.002	0.057	0.009	0.001
6	0.156	0.103	0.019	0.005	0.188	0.058	0.045	0.12	0.049	0.095	0.03	0.11	0.002	0.001	0.019
7	0.097	0.074	0.01	0.007	0.011	0.347	0.039	0.067	0.117	0.157	0.019	0.049	0.002	0.001	0.004
8	0.062	0.106	0.062	0.043	0.006	0.012	0.447	0.043	0.025	0.075	0.099	0.006	0.006	0.006	0
9	0.082	0.067	0.01	0.008	0.017	0.035	0.087	0.464	0.079	0.079	0.048	0.019	0.004	0.002	0
10	0.012	0.057	0.008	0.013	0.005	0.118	0.022	0.079	0.52	0.084	0.055	0.018	0.008	0	0.001
11	0.087	0.101	0.054	0.01	0.008	0.081	0.091	0.119	0.076	0.298	0.059	0.015	0.001	0.001	0
14	0.007	0.117	0.117	0.162	0.001	0.006	0.061	0.033	0.044	0.02	0.38	0.001	0.043	0.007	0
17	0.197	0.115	0.014	0.007	0.035	0.079	0.032	0.1	0.071	0.07	0.051	0.222	0.004	0.001	0.002
20	0.018	0.049	0.024	0.16	0.004	0.009	0.022	0.015	0.022	0.012	0.142	0.004	0.505	0.01	0.004
21	0.002	0.005	0.005	0.031	0	0.002	0.005	0.002	0.002	0.001	0.009	0	0.133	0.802	0.004
24	0.374	0.069	0.009	0.009	0.022	0.058	0.02	0.027	0.012	0.033	0.029	0.039	0.003	0.003	0.293

Table 26: Displaying classification within each class. Classification rule is Lognormal0

Class	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.511	0.091	0.006	0.002	0.026	0.059	0.026	0.062	0.019	0.066	0.022	0.07	0.002	0.003	0.037
3	0.136	0.348	0.045	0.045	0.015	0	0.045	0.03	0.121	0.076	0.091	0.03	0.015	0	0
4	0.008	0.07	0.439	0.205	0.001	0	0.066	0.003	0.014	0.032	0.106	0.003	0.048	0.005	0
5	0.008	0.063	0.272	0.306	0.001	0.002	0.068	0.006	0.016	0.035	0.108	0.003	0.099	0.01	0.003
6	0.163	0.065	0.016	0.004	0.178	0.061	0.038	0.1	0.049	0.104	0.02	0.099	0.005	0.002	0.096
7	0.106	0.053	0.01	0.009	0.028	0.348	0.041	0.05	0.108	0.15	0.013	0.047	0.005	0.001	0.031
8	0.068	0.099	0.075	0.031	0	0.012	0.447	0.043	0.019	0.093	0.068	0.012	0.025	0.006	0
9	0.103	0.043	0.007	0.007	0.041	0.045	0.085	0.414	0.082	0.088	0.033	0.029	0.012	0.005	0.007
10	0.016	0.044	0.01	0.007	0.005	0.146	0.021	0.07	0.519	0.074	0.04	0.024	0.02	0.001	0.004
11	0.116	0.073	0.038	0.01	0.014	0.111	0.081	0.095	0.088	0.306	0.044	0.014	0.005	0	0.005
14	0.013	0.115	0.106	0.14	0.002	0.005	0.077	0.041	0.062	0.046	0.314	0.008	0.064	0.008	0
17	0.185	0.08	0.01	0.006	0.06	0.077	0.031	0.083	0.068	0.074	0.041	0.226	0.007	0.002	0.051
20	0.02	0.049	0.014	0.127	0.004	0.01	0.03	0.015	0.028	0.02	0.117	0.008	0.544	0.009	0.007
21	0.002	0.005	0.003	0.023	0.001	0	0.005	0.002	0.001	0.001	0.009	0	0.13	0.821	0
24	0.268	0.047	0.008	0.005	0.011	0.04	0.021	0.022	0.01	0.037	0.025	0.025	0.006	0.004	0.47

Table 27: Displaying classification within each class. Classification rule is Lognormal1

Class	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.491	0.079	0.008	0.001	0.038	0.059	0.028	0.062	0.02	0.068	0.024	0.074	0.001	0.005	0.044
3	0.106	0.409	0.045	0.03	0	0.03	0.061	0.03	0.121	0.045	0.076	0.045	0	0	0
4	0.005	0.078	0.479	0.178	0.004	0.002	0.06	0.003	0.017	0.031	0.109	0.002	0.024	0.008	0
5	0.007	0.07	0.306	0.287	0.004	0.005	0.066	0.006	0.019	0.032	0.113	0.002	0.07	0.013	0
6	0.151	0.065	0.017	0.003	0.264	0.065	0.036	0.108	0.049	0.106	0.02	0.099	0.001	0.003	0.013
7	0.095	0.057	0.011	0.005	0.034	0.378	0.041	0.054	0.107	0.144	0.016	0.04	0	0.002	0.018
8	0.075	0.075	0.081	0.012	0.012	0.019	0.491	0.037	0.012	0.075	0.075	0.012	0.006	0.019	0
9	0.087	0.045	0.009	0.007	0.041	0.042	0.081	0.434	0.082	0.099	0.033	0.028	0.003	0.009	0.002
10	0.013	0.045	0.012	0.004	0.009	0.154	0.017	0.067	0.526	0.076	0.043	0.031	0.002	0.002	0
11	0.107	0.066	0.043	0.009	0.018	0.092	0.081	0.107	0.093	0.31	0.047	0.016	0.003	0	0.009
14	0.014	0.13	0.111	0.111	0.003	0.006	0.071	0.04	0.068	0.045	0.323	0.009	0.058	0.011	0.001
17	0.159	0.079	0.013	0.006	0.069	0.069	0.03	0.088	0.073	0.076	0.039	0.268	0.002	0.004	0.024
20	0.018	0.057	0.016	0.101	0.007	0.009	0.03	0.017	0.029	0.017	0.108	0.008	0.562	0.011	0.009
21	0.002	0.01	0.004	0.017	0.002	0	0.003	0.003	0.001	0.002	0.009	0	0.083	0.867	0
24	0.23	0.045	0.011	0.003	0.017	0.044	0.024	0.026	0.011	0.039	0.024	0.031	0.006	0.006	0.483

Table 28: Displaying classification within each class. Classification rule is Lognormal2

Class	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.568	0.1	0.006	0.003	0.031	0.018	0.019	0.032	0.01	0.07	0.031	0.047	0.003	0.001	0.065
3	0.121	0.485	0.03	0	0.015	0.03	0.045	0.045	0.03	0.03	0.121	0.045	0	0	0
4	0.009	0.128	0.264	0.276	0.007	0.001	0.04	0.002	0.003	0.031	0.19	0.001	0.041	0.007	0.001
5	0.012	0.122	0.163	0.31	0.006	0.002	0.05	0.002	0.006	0.034	0.186	0.002	0.091	0.014	0.001
6	0.245	0.095	0.01	0.008	0.225	0.041	0.028	0.071	0.015	0.109	0.031	0.099	0.007	0.001	0.017
7	0.176	0.124	0.009	0.008	0.037	0.197	0.04	0.037	0.037	0.224	0.025	0.059	0.002	0.001	0.026
8	0.106	0.106	0.043	0.05	0.006	0	0.416	0.031	0.012	0.106	0.112	0	0.012	0	0
9	0.182	0.094	0.009	0.007	0.034	0.02	0.065	0.295	0.03	0.136	0.056	0.046	0.018	0.002	0.004
10	0.05	0.084	0.016	0.008	0.018	0.146	0.008	0.047	0.267	0.186	0.082	0.065	0.019	0.004	0.001
11	0.193	0.126	0.023	0.019	0.011	0.029	0.043	0.055	0.028	0.362	0.084	0.009	0.008	0	0.01
14	0.023	0.191	0.056	0.09	0.006	0.005	0.046	0.015	0.015	0.046	0.364	0.011	0.119	0.012	0.001
17	0.257	0.11	0.009	0.004	0.049	0.027	0.024	0.049	0.023	0.085	0.055	0.259	0.008	0.002	0.041
20	0.03	0.078	0.023	0.075	0.003	0.004	0.015	0.01	0.004	0.017	0.099	0.01	0.602	0.022	0.008
21	0.003	0.013	0.015	0.029	0	0	0.008	0.001	0.007	0.001	0.007	0	0.07	0.849	0
24	0.255	0.061	0.006	0.003	0.014	0.024	0.015	0.01	0.006	0.034	0.031	0.024	0.012	0.003	0.501

Table 29: Displaying classification within each class. Classification rule is Lognormal3

Class	1	3	4	5	6	7	8	9	10	11	14	17	20	21	24
1	0.569	0.099	0.006	0.003	0.033	0.017	0.017	0.032	0.01	0.074	0.031	0.046	0.003	0.001	0.059
3	0.136	0.485	0.03	0.015	0.03	0.03	0.015	0.03	0.03	0.03	0.121	0.045	0	0	0
4	0.008	0.122	0.258	0.282	0.007	0.001	0.039	0.001	0.003	0.033	0.182	0.001	0.056	0.006	0.001
5	0.012	0.119	0.157	0.307	0.006	0.001	0.049	0.002	0.005	0.034	0.172	0.002	0.118	0.015	0.001
6	0.242	0.093	0.01	0.008	0.228	0.04	0.026	0.072	0.016	0.112	0.032	0.097	0.008	0.001	0.015
7	0.179	0.121	0.009	0.007	0.037	0.196	0.04	0.038	0.038	0.224	0.023	0.058	0.003	0.001	0.026
8	0.099	0.106	0.043	0.05	0.006	0	0.422	0.012	0.012	0.106	0.124	0	0.019	0	0
9	0.174	0.097	0.008	0.007	0.035	0.021	0.063	0.299	0.03	0.139	0.054	0.044	0.023	0.001	0.003
10	0.056	0.08	0.015	0.008	0.017	0.14	0.009	0.05	0.273	0.175	0.082	0.069	0.022	0.004	0.001
11	0.187	0.125	0.021	0.015	0.011	0.026	0.04	0.062	0.03	0.369	0.081	0.01	0.011	0	0.01
14	0.022	0.19	0.053	0.084	0.006	0.006	0.044	0.018	0.015	0.044	0.335	0.011	0.159	0.011	0.001
17	0.257	0.111	0.009	0.003	0.051	0.025	0.023	0.049	0.025	0.087	0.052	0.255	0.01	0.002	0.04
20	0.029	0.078	0.022	0.068	0.003	0.003	0.014	0.009	0.002	0.017	0.09	0.01	0.615	0.029	0.01
21	0.003	0.013	0.017	0.031	0	0	0.007	0.001	0.005	0.001	0.005	0	0.067	0.85	0
24	0.256	0.061	0.007	0.004	0.014	0.022	0.014	0.009	0.005	0.038	0.031	0.024	0.012	0.003	0.499

References

J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, B-48:259–302, 1986. 10

J. W. Goodman. Statistical properties of laser speckle patterns. In J. C. Dainty, editor, *Laser Speckle and Related Phenomena*. Springer-Verlag, New York, second edition, 1984. 2

S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continous Multivariate Distributions*, volume 1: Models an Applications. John Wiley Interscience Publication, New York, second edition, 2000. 3

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, New York, second edition, 1998.

K. Y. Liang and S. L. Zeger. Inference based on estimating functions in the presence of nuisance parameters (with comments). *Statistical Science*, 55:158–173, 1995. 13

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979. 6

G. Storvik and G. Dahl. Lagrangian based methods for finding MAP solutions for MRF models. *IEEE Trans. Image Processing*, 9(3):469–479, 2000. 10