



HAL
open science

Probabilistic Modelling of Printed Dots at the Microscopic Scale

Quoc Thong Nguyen, Yves Delignon, François Septier, Anh Thu Phan Ho

► **To cite this version:**

Quoc Thong Nguyen, Yves Delignon, François Septier, Anh Thu Phan Ho. Probabilistic Modelling of Printed Dots at the Microscopic Scale. *Signal Processing: Image Communication*, 2018, 62, pp.129-138. 10.1016/j.image.2018.01.003 . hal-01686159v2

HAL Id: hal-01686159

<https://hal.science/hal-01686159v2>

Submitted on 24 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Probabilistic Modelling of Printed Dots at the Microscopic Scale

Quoc Thong Nguyen^a, Yves Delignon^b, François Septier^b, and Anh Thu Phan Ho.^c

^a*Université de Bretagne Sud, Laboratoire de Mathématiques de Bretagne Atlantique, UMR CNRS 6205,
Campus de Tohannic, Vannes, France*

^b*IMT Lille Douai, Université Lille, CNRS UMR 9189 - CRIStAL, F-59000 Lille, France*

^c*Université de La Rochelle, Laboratoire L3i, La Rochelle, France*

Abstract

Microscopic analysis of paper printing shows regularly spaced dots whose random shape depends on the printing technology, the configuration of the printer as well as the paper properties. The modelling and identification of paper and ink interactions are required for qualifying the printing quality, for controlling the printing process and for application in authentication as well. This paper proposes an approach to identify the authentic printer source using micro-tags consisting of microscopic printed dots embedded in the documents. These random shape features are modelled and extracted as a signature for a particular printer. In the paper, we propose a probabilistic model consisting of vector parameters using a spatial interaction binary model with inhomogeneous Markov chain. These parameters determine the location and describe the diverse micro random structures of microscopic printed dots. A Markov chain Monte Carlo (MCMC) algorithm is thus developed to approximate the Minimum Mean Squared Error estimator. The performance is assessed through numerical simulations. The real printed dots from the common printing technologies (conventional offset, waterless offset, inkjet, laser) are used to assess the effectiveness of the model.

Keywords: Probabilistic model; Bernoulli process; Metropolis Hastings within Gibbs; Microscopic printing; Markov chain.

*Corresponding author

Email address: quoc-thong.nguyen@univ-ubs.fr (Quoc Thong Nguyen)

1. Introduction

Preventing and discouraging unauthorised printed materials is the expected achievement of the authentication methods. Papers dealing with this subject can be categorised in two strategies. The first one consists in embedding an extrinsic signature such as secure tags in the prints [1, 2]. While the second one, which is the motivation of this paper, consists in characterising the intrinsic features of printing and scanning process [3]. As examples of intrinsic features, the signature consisting on the banding artefact coming from fluctuation of the optical photoconductor angular velocity in case of laser printers [4], dimple effect which is specific to inkjet printers or also texture features [5], or the analysis of quality signature of the unique print to differentiate one printer technology/supplier from another [6]. Authentication can be performed based on paper statistics, for example, the paper statistics is extracted by capturing microstructure images from different carton packages under different light conditions and with different cameras. These microstructure images are considered as digital fingerprinting on which the authentication frameworks are derived [7]. In [8], the authors proposed the intrinsic feature of paper based on a texture speckle pattern, a random bright/dark region formation at the microscopic level when light falls on to the paper. The extracted signature of printer is a unique pattern of misplaced toner powder on each paper [9]. For electrophotographic printers, the intrinsic signals extracted from banding artefact are used to identify the device [10]. Printer source identification is an important forensic tool. The techniques using characters for identification of source scanner and printer have been proposed [11]. Source identification using printed character in different languages has been recently studied [12, 13, 14, 15, 16].

Besides the role in print authentication, depending on the purposes of the problems, there are a number of applications concerning the printing models. A specific model is considered in the scale of the experiments/applications. For instance, in the human vision scale, halftoning technique is a method that takes advantage of the optical illusion of human vision to display continuous tone grayscale image with only black or/and white dots [17, 18, 19]. The scale of the model is the resolution of the printer. An image of print not only contains the properties of a printing process but also includes

image degradation; Baird proposed a physics-based model in which the parameters defining imaging defects are described [20]. Under the microscopic scale printing, a probabilistic model for the average coverage of the toner particles ink of electrophotographic printing process is introduced by Norris *et al.* [21]. However, the model
35 has a limited validation, and the estimation procedure has not been considered. For authentication purpose, we propose a probabilistic modelling of microscopic printed dots.

The shapes of microscopic printed dots can be considered as the intrinsic feature of the printing process. At the microscopic scale, a dot is a random pattern whose shape
40 depends on the technology, the setting of the printer, the ink quality and/or the paper properties. From a statistical point of view, the digital acquisition of these random dots can be modelled as a spatial interaction binary model based on an exponential power kernel that depends on location parameter and shape parameters. The accurate and consistent model can help to create an authentic signature to improve the authentication
45 procedure. The model proposed in our previous works [22, 23] consists in a spatial distribution model that simulates the randomness of the shape of printed dots. The random shape is manipulated by the parameters in this probabilistic model. In [24], the numerical estimation procedure using MCMC algorithm is developed. However, in order to cover the diverse random shape of printed dots, the surrounding pixels or the
50 conditional probabilities has to be taken into account.

In this paper, we propose a spatial interaction binary model that describes more precisely the random shape of the microscopic dots by taking into account the neighbourhood of the pixels. In particular, the Markov property between pixels in the image is considered. This property helps to describe more accurately the intrinsic features of
55 the print, which can thus improve the discrimination power. An MCMC algorithm is developed for estimating the model parameters. Finally, we analyse the performance of the proposed unsupervised identification algorithm. The rest of this paper is organised as follows. Section 2 is devoted to visually describe the randomness of shape of microscopic dots. The formulation of spatial stochastic model is discussed in Section 3. The
60 Metropolis-Hastings within Gibbs algorithm for the estimation is developed in Section 4. In Section 5, the Posterior Cramér - Rao bound used to evaluate the performance of

the algorithm is derived. Section 6 is devoted to the performance of the new algorithm which are analysed through simulations, and the benefit of such a model and estimation algorithm are illustrated in the case of real microscopic scale printed dots. The basic authentication of printer identification using the model is also described in Section 7. Concluding remarks are given in Section 8.

2. Printed dots at the microscopic scale

As aforementioned, we consider the shape of microscopic printed dots as an intrinsic feature to perform authentication on printer identification. This section is devoted to describing the random shape of microscopic printed dots. We then demonstrate intuitively that the shapes are different between different printers, which can benefit printers discrimination. Let us start by the observation of printed dots at the microscopic scale. A single printed dot is created by printing an image containing one black pixel. In our study, the microscopic printed dots are captured using an optical Zeiss Microscope with an AxioCam camera. The camera has a maximum resolution of 2464 x 2056 (5Mp) with a sensor size of 8.5 mm x 7.1 mm and a pixel size of 3.45 μm x 3.45 μm . Since the focus of the study is on the physical shape of coverage area of the ink, the collected images are only in black and white.

The images on the left column in Fig. 1 illustrate the dots printed by offset, laser, inkjet printers respectively. Random shapes of these dots are observed under microscope scale. The physical size of the dots is very small. For example, size of a dot from the offset printer (1200 dpi) is theoretically 21.16 μm . The printing resolution is an important element in this study. If the resolution is too high, the outcome will be similar to the dust noise from the printer. For instance in our study, the dots printed from a laser printer under resolution 1200 dpi are not distinguishable from the noise. The resolution can neither be too low, since the shape of the dots are no more random except for the edge. Therefore, the resolution has to be chosen manually for a specific printer.

Visually, from the left column of Fig. 1, we can see that the shapes of printed dots from different printers are dissimilar. Fig. 2 illustrates the shape of dots printed by

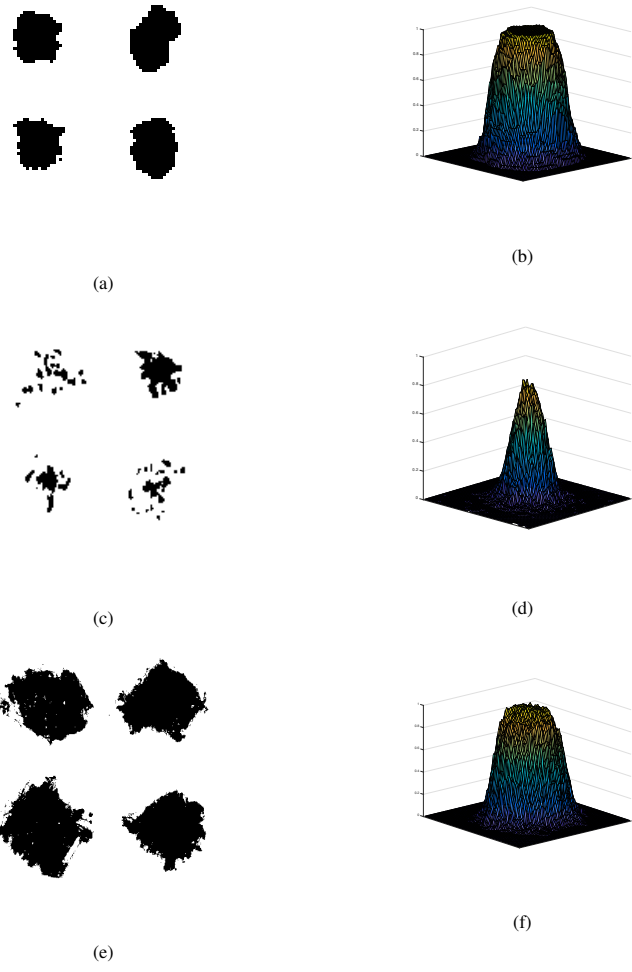


Figure 1: Binary image of 4 single dots printed by an offset printer (a), the empirical probabilities of appearance of black pixels in the given position (profile) of 100 samples of one printed dot with the resolution 1200 dpi (b). The dots from a laser printer with the resolution 600 dpi (c), and the profile (d). The dots from an inkjet printer at 720 dpi (e), and the profile (f).

two different laser printers at the same resolution 600 dpi. On the right column in Fig. 1, each image illustrates the frequency of appearance of black pixels at each position (profile) from 100 captured samples. As observed, the probabilities of emergence of the black pixels are lower when the pixels are further from the centres. The shape of the profile contains the characteristics of the printed dots, which means the shape

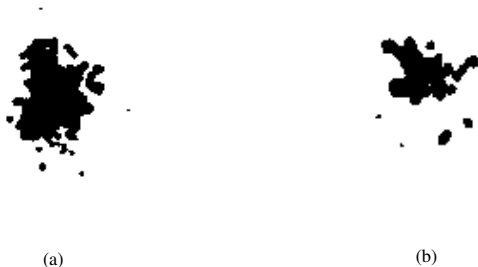


Figure 2: The dots printed by two laser printers, Dell laser printer (a) and HP laser printer (b). The conditions are identical, resolution is 600 dpi.

of the dots contains the identity of the source printer in some sense. The model in [22, 23] introduced the parameters controlling the width of the droplets, the density of the black pixels and the average number of black pixels. However, the interaction with the surrounding pixels was not taken into account. In particular, when a pixel is black, its neighbour is more likely to be blackened. In addition, when the pixels are further from the centre of the dots, this influence of the black pixel(s) on their vicinity is also smaller. Including these properties could help the model to have a more precise simulation, thus describing in more details the identity of source printer. These properties are detailed in the next Section.

3. Spatial interaction binary model with Markov chain

Let S be a finite set corresponding to N pixels of a binary image \mathbf{U} that displays K dots. We consider the random process $\mathbf{U} = (U_s)_{s \in S}$, where each U_s is a Bernoulli random variable which takes its value from $\{0, 1\}$, respectively black state and white state. Let V_s be a neighbourhood of the pixel s , the geometric shape of V_s is independent of $s \in S$. \mathbf{U} is modelled by a Markov random field

$$P(U_s = u_s \mid (U_t = u_t)_{t \neq s}) = P(U_s = u_s \mid (U_t = u_t)_{t \in V_s}), \quad (1)$$

i.e. the probability that the pixel s takes the value u_s conditional on all other pixels of the image equals to the probability of u_s conditional on the values of the pixels in the neighbourhood V_s . This probability also has to depend on the distance of the pixel s to

the centre μ , which means $P(U_s = 0 | (U_t = u_t)_{t \neq s})$ decreases when the pixel s goes
 110 further from the centre, i.e. the Euclidean distance $\|s - \mu\|$ increases.

A two-dimensional image can be transformed into a one-dimensional chain by Hilbert curve [25] (see Fig. 3). Because the targeted interactions (statistical dependencies) between pixels within the microscopic dot are well preserved in the spatial structure by Hilbert curve, the Markov chain with Hilbert path is used in many image processing applications [26, 27, 28]. Moreover, Pseudo-Hilbert Scan algorithm can be applied to arbitrarily-sized image [29]. In convenient index, let $\mathbf{U} = (U_1, U_2, \dots, U_N)$ be a vector of random variables ordered according to the Hilbert curve transformation of the image. In case the neighbourhood V_s is restricted to one pixel, the image \mathbf{U} becomes a Markov chain

$$P(\mathbf{U} = \mathbf{u}) = P(U_1 = u_1) \prod_{i=2}^N P(U_i = u_i | U_{i-1} = u_{i-1}). \quad (2)$$

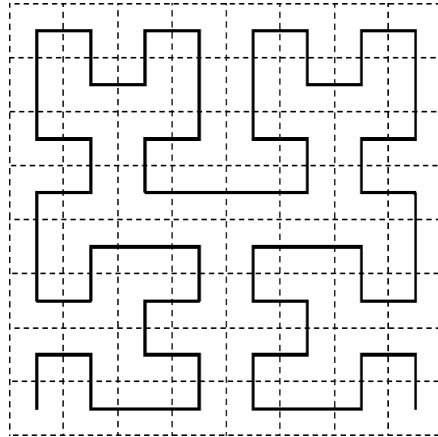


Figure 3: Scanning of 8×8 image

The Markov chain (2) encompasses the parameters of the initial probability

$$P(U_1 = 0), \quad (3)$$

and transition probabilities

$$P(U_i = 0 \mid U_{i-1} = u), \quad (4)$$

for $1 < i \leq N$. The parametric probabilities for (3), (4) are

$$P(U_1 = 0) = 1 - \prod_{k=1}^K (1 - p_k(s_1)) \quad (5)$$

$$P(U_i = 0 \mid U_{i-1} = u) = 1 - \prod_{k=1}^K (1 - p_k(s_i)) \lambda^{(1-u)}. \quad (6)$$

The function $p_k(\cdot)$, named kernel, measures the ability to blacken the surrounding pixels of the k -th dot. The kernel $p_k(s)$ has to decrease with respect to the distance $\|s - \mu_k\|$, so that, as requisite property, the initial and transition probabilities also decrease with respect to the distance. The support of $p_k(\cdot)$ is a subset of $(0,1]$.

The positive parameter λ characterises the interaction between pixels when the neighbour pixel is black, $U_{i-1} = 0$. With $\lambda = 1$, the probabilities (6) only depend on the relative position to the centres, $P(U_i \mid U_{i-1}) = P(U_i)$, which is the approach proposed in [22, 23]. In case $\lambda < 1$, the smaller λ leads to a higher chance of being black of current pixel U_i , since $P(U_i = 0 \mid U_{i-1} = 0)$ is greater with smaller λ (see Eq. (6)). When the neighbour pixel is white, $U_{i-1} = 1$, the probabilities depend on the distances of the current pixel to the dots centres. With these properties, the dots are more compact, and the density of black pixels at the areas closer to the centres is higher. The new proposed parameter λ helps to describe a wider variety of shape of dots.

The transition probabilities depend on $p_k(s)$ which relates to the position of the pixels, this implies that \mathbf{U} is a non-stationary Markov chain. The expected number of black pixels of the image, denoted n_b , is calculated as

$$n_b = \sum_{i=1}^N \mathbb{E}(\delta(U_i)) = \sum_{i=1}^N P(U_i = 0), \quad (7)$$

with $\delta(x) = 1$ if $x = 0$, $\delta(x) = 0$ elsewhere. The probability $P(U_1 = 0)$ is computed

by (5), while the other terms are calculated by (6) with the recurrence relation and the law of total probability

$$P(U_i = 0) = P(U_i = 0 | U_{i-1} = 0)P(U_{i-1} = 0) + P(U_i = 0 | U_{i-1} = 1)P(U_{i-1} = 1), \quad (8)$$

with $2 \leq i \leq N$. Let us note that with smaller λ , $P(U_i = 0 | U_{i-1} = 0)$ is greater, that leads to the value of $P(U_i = 0)$ also becoming greater. This implies that smaller λ statistically increases the number of black pixels. In the following, the kernel p_k is chosen as a Gaussian power kernel

$$p_k(s) = \eta \exp \left\{ - \left(\frac{\|s - \mu_k\|^2}{\sigma^2} \right)^\beta \right\}. \quad (9)$$

130 Fig. 4 illustrates the simulation of one dot with σ^2, β equal to 70 and 5 respectively and various values of λ and η . It can be seen that the dots with small λ are more continuous, in other words, the black pixels make their neighbourhood more likely to be black. The parameter η controls the amplitude of the kernel and the density of black pixels increases with η . When η is large, the density of black pixels near the centre is
 135 almost independent of λ .

4. Estimation

A Bayesian framework is employed to estimate the model parameters. Let us recall our problem, the parameters of interest in (5), (6) are $\theta_{\mathbf{U}} = [\{\mu_k\}, \sigma^2, \beta, \eta, \lambda]$ which belong to a $2K + 4$ dimensional space. The Minimum Mean Squared Error (MMSE) estimator $\hat{\theta}$ of $\theta_{\mathbf{U}}$ given the observed data \mathbf{u} is

$$\hat{\theta}(\mathbf{u}) = \int \theta p(\theta | \mathbf{u}) d\theta, \quad (10)$$

where $p(\theta | \mathbf{u})$ is the posterior distribution. To approximate the integral (10), an MCMC algorithm is applied [30]. Since the vector of parameters is of high-dimension,







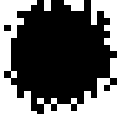


η	λ		
	1	0.6	0.2
0.3			
0.7			
1			

Figure 4: The simulation of the model (5), (6) with Gaussian power kernel with $\sigma^2 = 70$, $\beta = 5$ and different values of λ , η .

it is more efficient to use Gibbs sampling procedure which consists in sampling successively each parameter from their conditional posterior distribution. Unfortunately, since no closed form expression is available for these distributions, we propose to use a Metropolis-Hastings within Gibbs algorithm [30]. The chosen strategy is to run a random walk as proposal distribution:

$$\theta^* = \theta^{(t)} + \epsilon_t, \quad (11)$$

where ϵ_t is a zero mean normal random vector with a diagonal covariance matrix. We generate $\theta_2 = \log(\sigma^2)$, $\theta_3 = \log(\beta)$ instead of σ^2, β to ensure their positivity; a transformation is also applied for η and λ to generate θ_4, θ_5 respectively, i.e.

140 $\eta = \frac{1}{1+e^{\theta_4}}, \lambda = \frac{1}{1+e^{\theta_5}}$. For convenient notation, we rewrite the vector of interested

parameters as $\theta = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5]$. The Metropolis-Hastings within Gibbs using random walk is summarised in Algo. 1.

Algorithm 1 Sampling algorithm with random walk

- 1: **Initialise** $\theta^{(0)} = [\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}, \theta_5^{(0)}]$
- 2: **for** $i = 1$ to N_iter **do**
- 3: **for** $b = 1$ to 5 **do**
- 4: Generate $\theta_b^* \sim \mathcal{N}(\theta_b^{(i-1)}, \sigma_b^2)$
- 5: Compute the acceptance ratio

$$\alpha(\theta_b^{(i-1)}, \theta_b^*) = \min \left\{ \frac{p(\theta_b^* | \theta_{-b}^{(i-1)}, \mathbf{u})}{p(\theta_b^{(i-1)} | \theta_{-b}^{(i-1)}, \mathbf{u})}, 1 \right\}$$

- 6: Decide

$$\theta_b^{(i)} = \begin{cases} \theta_b^* & \text{with probability } \alpha(\theta_b^{(i-1)}, \theta_b^*) \\ \theta_b^{(i-1)} & \text{otherwise} \end{cases}$$

- 7: **end for**
 - 8: **end for**
Where $\theta_{-b}^{(i-1)} = [\theta_1^{(i-1)}, \dots, \theta_{b-1}^{(i-1)}, \theta_{b+1}^{(i-1)}, \dots, \theta_5^{(i-1)}]$.
-

The initialisation of θ proceeds as following. The centres are initialised by the careful seeding [31]. The other parameters are respectively initialised as $\sigma^2 = 0.6 \times \frac{N}{\pi K}$, $\beta = 2$. Since η and λ are in $(0, 1]$, intuitively, we may initialise both of them equal to 0.5. From Bayes' law, we have:

$$p(\theta_b | \theta_{-b}, \mathbf{u}) = \frac{p(\mathbf{u} | \theta_b, \theta_{-b}) p(\theta_b, \theta_{-b})}{p(\theta_{-b}, \mathbf{u})}, \quad (12)$$

where $p(\mathbf{u} | \theta_b, \theta_{-b})$ is the likelihood and $p(\theta_b, \theta_{-b})$ is the prior. Since all parameters are independent, the acceptance ratio in Alg. 1 is rewritten as

$$\alpha(\theta_b^{(i-1)}, \theta_b^*) = \min \left\{ \frac{p(\mathbf{u} | \theta_b^*, \theta_{-b}) p_b(\theta_b^*)}{p(\mathbf{u} | \theta_b^{(i-1)}, \theta_{-b}) p_b(\theta_b^{(i-1)})}, 1 \right\}, \quad (13)$$

with p_b the prior of parameter θ_b . The prior distributions are chosen with fixed pa-

rameters by experience. Each centre $\{\mu_k\}$ has multivariate normal distribution, σ^2 is
 145 log-normal distributed, β follows gamma law, while both η and λ are beta-distributed.
 Let us note that when the priors are closed to uniform, the acceptance ratio mainly depends on the likelihood ratio. The choice of the variances of the priors depends on the confidence of the initialisation and the experience.

5. Posterior Cramér - Rao bound

Let \mathbf{u} be the observation vector and θ be the r -dimensional vector of random parameters (with $r = 2k + 4$). The joint probability density of the pair (\mathbf{U}, θ) is $p(\mathbf{U}, \theta)$. Let $\hat{\theta}(\mathbf{u})$ denote an estimate of θ . The Posterior Cramér-Rao Bound (PCRB) states that [32, 33]:

$$E[(\hat{\theta}(\mathbf{U}) - \theta)(\hat{\theta}(\mathbf{U}) - \theta)^T] \geq I^{-1}, \quad (14)$$

150 where I is $r \times r$ Fisher information matrix (FIM)

$$\begin{aligned} I &= E [\nabla_{\theta} \log p(\mathbf{U}, \theta) \nabla_{\theta}^T \log p(\mathbf{U}, \theta)] \\ &= -E [\Delta_{\theta}^{\theta} \log p(\mathbf{U}, \theta)], \end{aligned} \quad (15)$$

where $\Delta_{\theta}^{\theta} := \nabla_{\theta} \nabla_{\theta}^T$ is the second derivative operator and ∇_{θ} is the gradient operator with respect to θ .

Using the fact that $p(\mathbf{U}, \theta) = p(\mathbf{U} | \theta)p(\theta)$, the expression of the FIM in (15) can be expressed as:

$$\begin{aligned} I &= -E [\Delta_{\theta}^{\theta} \log p(\mathbf{U} | \theta)] - E [\Delta_{\theta}^{\theta} \log p(\theta)] \\ &= I_d + I_p, \end{aligned} \quad (16)$$

155 where I_p stands for the *a priori* information matrix and I_d is the "standard" FIM ($I_d(\theta)$ is used in the derivation of the CRB) averaged over the *a priori* distribution of θ :

$$I_d = \int_{\Theta} I_d(\theta) p(\theta) d\theta. \quad (17)$$

In many special cases, when only the i -th component of θ , say θ_i , is of interest, the PCRB is expressed as

$$E[(\hat{\theta}_i(\mathbf{U}) - \theta_i)^2] \geq [I^{-1}]_{ii} \quad (18)$$

6. Experimental results and performances

In this section, we aim at assessing both the performance of the estimators and the adequacy of the model. The Metropolis-Hasting within Gibbs algorithm is analysed, the performance of the algorithm is evaluated from the images generated by the parametric model. The experiment on the real captured dots is also conducted.

6.1. Synthetic Data

The simulated images are generated with a size of 64×64 . As mentioned previously, the priors are fixed, the expected values of the prior are set by the initial values, and the variance are chosen large in order to consider a highly uncertain prior knowledge. The estimation is implemented with 20000 MCMC iterations and a burn-in period of 5000, which means that 15000 last samples are used to approximate the parameters over 20000 generated samples.

6.1.1. Analysis of the estimation with respect to β

In this analysis, only β is varying. The simulated images of single dot are generated. The true centres μ of the dot is $(32, 32)$. With $\sigma^2 = 50, \eta = 0.8, \lambda = 0.6$, along with three values of $\beta, \beta \in \{0.8, 2, 5\}$, Tab. 1 shows the mean squared errors of the estimators. Larger value of β leads to smaller errors for $\hat{\sigma}^2$ and $\hat{\eta}$. In contrast, the errors of $\hat{\beta}$ and $\hat{\lambda}$ are larger with large β . With a large value of β , the simulated dots are different mainly near the edges. It means that the information of λ is less significant, which makes the estimation of λ less consistent. While with smaller β , the contribution of λ to the connected black pixels is more visible (see Fig. 5). On the other hand, the simulated dots look similar for large values of β (see Fig. 6), which also makes the estimation of β less consistent.

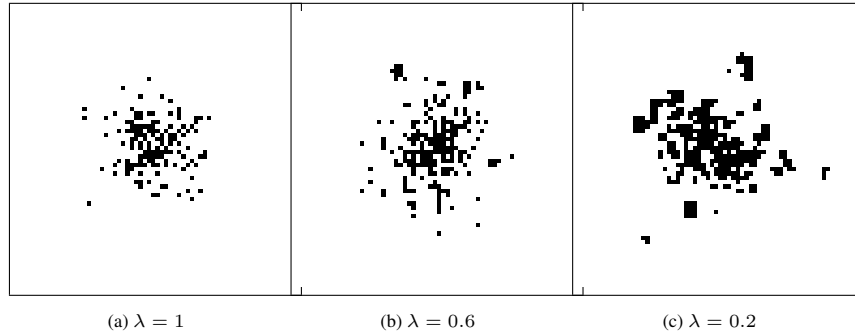


Figure 5: The role of λ with small $\beta = 0.8$.

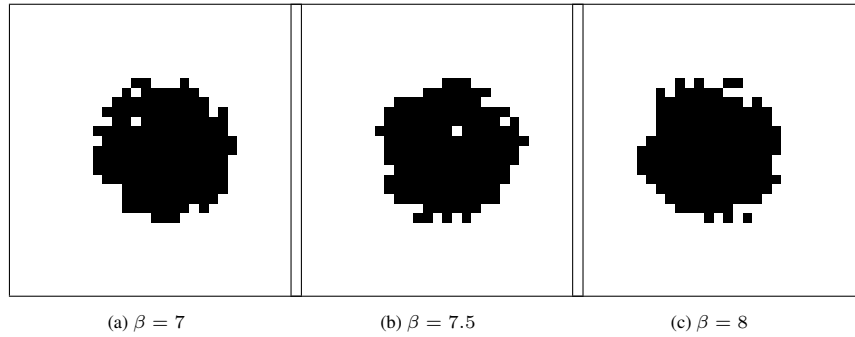


Figure 6: Realization with three different values of β .

180 6.1.2. Analysis of the estimation with respect to η

The MSE of the estimators are given in Tab. 2. The simulations of single dot were conducted with $\beta = 2, \sigma^2 = 50, \lambda = 0.6$, along with three values of $\eta, \eta \in \{0.5, 0.8, 0.97\}$. The estimators of σ^2 have smaller MSE with large η , which means the black pixels give a significant information to the estimation procedure. The MSE of $\hat{\lambda}$ increases due to the fact that when η is large, the role of λ is visually presented around the edge of the dots. This phenomenon can be observed in Fig. 4.

	σ^2	β	η	λ
0.8	218.51	0.029	0.022	0.0025
2	45.12	0.23	0.011	0.0068
5	13.20	2.011	0.0037	0.0025

Table 1: The mean squared errors of the estimators with various values of $\beta, \sigma^2 = 50, \eta = 0.8, \lambda = 0.6$.

	σ^2	β	η	λ
0.5	100.9	0.2041	0.0065	0.0058
0.8	45.12	0.23	0.011	0.0068
0.97	49.85	0.1845	0.032	0.01

Table 2: The mean squared errors of the estimators with various values of $\eta, \sigma^2 = 50, \beta = 2, \lambda = 0.6$.

6.1.3. Analysis of the estimation with respect to λ

The simulations of one dot were conducted with $\beta = 2, \sigma^2 = 50, \eta = 0.8$, along with three values of $\lambda, \lambda \in \{0.2, 0.6, 0.9\}$. Tab. 3 shows the MSE of the estimators. The estimators of σ^2, β, η have decreasing MSE when λ increases. Since the number of black pixels decreases when the value of λ increases, the information for the estimation of σ^2, β, η exposes clearer on the realisations. On the other hand, the estimation of λ is better for small values, because the connection between black pixels strongly emerges.

	σ^2	β	η	λ
0.2	66.06	0.247	0.026	0.0017
0.8	45.12	0.23	0.011	0.0068
0.9	36.57	0.199	0.0113	0.0181

Table 3: The mean squared errors of the estimators with various values of $\lambda, \beta = 2, \sigma^2 = 50, \eta = 0.8$.

6.1.4. Posterior Cramér-Rao bound for single dot case

The PCRb is used to assess the precision of the estimators. Since the posterior information matrix I in (16) cannot be calculated analytically, it is approximated by Monte Carlo simulation. The random values of vector of parameters θ generated N times by the priors. For each generated vector of parameter, M images are simulated to approximate the standard Fisher information matrix $I_d(\theta)$ in (18).

	μ_x	μ_y	σ^2	β	η	λ
PCRb	0.111	0.109	9.621	0.002	0.002	0.003
MSE	5.180	5.605	33.146	2.843	0.044	0.054

Table 4: The approximated PCRb and the Mean Squared Error (MSE) of each parameter in case of single dot.

In Tab. 4, we demonstrate the approximated values of the Posterior Cramér-Rao

Bound in case of single dot with $N = 100$ generated random values of vector θ . The values of priors are as follows: centre (μ_x, μ_y) has multivariate normal distribution with mean $(32, 32)$ and covariance matrix $diag(5, 5)$; σ^2 is log-normal distributed with mean 50 and variance 100; β follows gamma law with mean 2 and variance 1.5; η and λ are beta-distributed $B(3, 1.5)$. For each value of θ , $M = 40$ images are generated to approximate matrix $I_d(\theta)$. Specifically, the size of each image is 64×64 . With each generated image, the parameters are estimated using Algo. 1. In total, 40×100 images are used to approximate the Mean Squared Error. As we observe in this example, σ^2 is estimated least accurate.

6.1.5. The performance with many dots

The impact of the number of dots, K , is analysed. Instead of one dot on each image as the analyses above, each image contains multiple simulated dots. The experiment is set up with $\sigma^2 = 50$, $\beta = 2$, $\eta = 0.8$, $\lambda = 0.6$, and all generated dots are completely separated from each other. As demonstrated in Tab. 5, the estimators are clearly more consistent for the images with more dots. The accuracy and consistency of $\hat{\beta}$, $\hat{\eta}$, $\hat{\lambda}$ is crucial in authentication. Therefore, analysing many dots is a good mean to have a more consistent estimation of the parameters. Since more dots give more information about the shape, which are determined by the parameters of the model, the estimators are more consistent.

	1 dot	2 dots	4 dots	7 dots
σ^2	6.03	3.08	2.42	1.92
β	0.25	0.19	0.18	0.15
η	0.08	0.04	0.04	0.03
λ	0.08	0.06	0.05	0.03

Table 5: The standard deviation of the estimators with different number of cluster-dots K , $\sigma^2 = 50$, $\beta = 2$, $\eta = 0.8$, $\lambda = 0.6$.

6.2. Estimation results from printed dots

The dots from four printers of four popular technologies are taken: conventional offset, waterless offset, laser and inkjet printings. The resolutions of each printing process for each sample are respectively 1200 dpi with offset printers, 600 dpi with

225 laser printer, and 720 dpi with inkjet one. For each printer, 10 images of size 64x64
of one dot are collected for the experiment. Fig. 7 represents the images of the real
dots and the realisation generated from the estimated parameters obtained from the
developed MCMC method. Metropolis-Hasting within Gibbs algorithm is conducted
with 30000 MCMC iterations with 10000 as burn-in period. The realisation from the
230 Spatial binary model [23] is illustrated to compare the improvement as well.

	Conventional Offset		Waterless offset	
	Avg.	Std.	Avg.	Std.
β	3.71	1.4	6.55	1.74
η	0.96	0.05	0.99	0.0004
λ	0.67	0.15	0.52	0.13

	Laser		Inkjet	
	Avg.	Std.	Avg.	Std.
β	1.21	0.92	2.28	0.51
η	0.59	0.32	0.99	0.01
λ	0.30	0.06	0.20	0.04

Table 6: The estimators of the real dots by MCMC.

Tab. 6 gives the numerical values of the estimators by MCMC estimation method
as well as the standard deviation values. Fig. 7 visually compares the printed dots,
the realisations from the proposed Markov spatial model and from the spatial binary
model that does not take into account the neighbour dependence developed in [22,
235 23]. The results are very promising when the interaction is taken into account. The
realisation from Markov spatial model is more realistic to the real dot than the one
from spatial binary model. With the offset and inkjet images, the fragment near the
border is resembled better, which makes them more similar to the real dots. The shape
of the dot from inkjet printer is more complicated to simulate due to the concave parts.
240 With the print from laser printer, we can see the substantial improvement using the
Markov spatial model w.r.t. the spatial binary model. The simulated dot is much more
realistic and continuous.

In Fig. 8 and Tab. 7, we give other estimation of four printed dots with the resolu-
tion 600 dpi from HP-600 M620 laser printer. We can again have a visual comparison
245 to show the improved simulation using the Markov spatial model. The capability of

the proposed model to capture more details can clearly increase the power of discrimination between printers. Moreover, since the characteristics of printed dots are represented by the parameters, the proposed model may significantly reduce training data than using machine learning generative models to characterise the printed dots.

	Estimators	Error with 99% of confidence
σ^2	66.82	± 0.019
β	1.36	$\pm 3.6 \times 10^{-4}$
η	0.76	$\pm 2.6 \times 10^{-4}$
λ	0.35	$\pm 1.2 \times 10^{-4}$

Table 7: The estimators of σ^2 , β , η , and λ from four printed dots.

250 7. Application in printer source identification

The capability of simulating the random shape of printed dots of the model and the performance of the estimation are demonstrated in previous Section. In this part, we use the typical random shape which is characterised by the parameters as a "signature" to identify the printer source. The vector of three parameters (β, η, λ) can be used as
 255 a characteristic of a printer with a specific setting. These three parameters are dimensionless, which means that the size of image of dots does not affect the value of these parameters. With the model, we are able to set up an authentic tag from an exact printer by setting specific values for the parameters of the printer (ink toner, paper...). Under a specific setting, a particular printer statistically creates the same shape of microscopic
 260 dots. In this experiment, we applied the maximum likelihood classification method to discriminate the printed dots from different sources [34, 35]. Fig. 9 summarises the procedure of the proposed printer identification method.

The estimations of the shape parameters are displayed in Fig. 10 that shows respectively β vs. λ , β vs. η and λ vs. η . Each printer is 3D Gaussian cluster in the space
 265 of parameters. Their location and shape are respectively given by their means (Tab. 6) and their covariance matrices (Tab. 8). For each printer (conventional offset, waterless offset, laser and inkjet), we take 10 other images for the discrimination experiment. In total, 20 images for each printer with 10 images are used to approximate the parameters of 3D Gaussian cluster. In this specific experiment, all the dots are printed on

270 uncoated paper, and other conditions are identical during the experiments. In this case, the normal distributions of the shape parameters enable to use the maximum likelihood classification method. All the samples are correctly classified (100%) with respect to their printer sources. The confusion matrix is presented in Tab. 9

$$\left| \begin{array}{ccc} 1.910 & 0.011 & 0.095 \\ 0.011 & 6e-4 & 0.001 \\ 0.095 & 0.001 & 0.033 \end{array} \right| \quad \left| \begin{array}{ccc} 3.034 & 6e-4 & 0.193 \\ 6e-4 & e-7 & 4e-5 \\ 0.193 & 4e-5 & 0.018 \end{array} \right|$$

(a) Conventional offset

(b) Waterless offset

$$\left| \begin{array}{ccc} 1.199 & -0.190 & 0.030 \\ -0.190 & 0.096 & 0.006 \\ 0.030 & 0.006 & 0.007 \end{array} \right| \quad \left| \begin{array}{ccc} 0.253 & 0.001 & 0.015 \\ 0.001 & 1.2e-5 & e-4 \\ 0.015 & e-4 & 0.002 \end{array} \right|$$

(c) Laser

(d) Inkjet

Table 8: The covariance matrices calculated from the estimators of four printers.

Actual/Predicted	Con. offset	Wat. offset	Laser	Inkjet
Con. offset	20	0	0	0
Wat. offset	0	20	0	0
Laser	0	0	20	0
Inkjet	0	0	0	20

Table 9: Confusion matrix for ML classification of four different types of printer: conventional offset, waterless offset, laser and inkjet.

Other experiment is also conducted on two laser printers, a HP laser printer and a
 275 Dell laser printer. The resolution of 600 dpi is set up for both printers, and the samples are printed on uncoated paper. The relation between estimated parameters of each printer is given in Fig. 11 which clearly separates two printers. Their location and shape of the 3D Gaussian clusters of each printer are respectively given by the means in Tab. 10 and the covariance matrices in Tab. 11. Similar to the previous experiment,
 280 20 images are captured from each printer with 10 images are used to approximate the parameters of 3D Gaussian cluster. The maximum likelihood classification method also gives a completely accurate classified result (Tab. 12). These results point out the relevance of the model for printed dots at the microscopic scale.

As demonstrated in Fig. 8, Tab. 7 and the analysis in the simulation experiment

285 section, the estimation of parameters is more consistent with many dots. Therefore, the analysis of many dots will help to improve the accuracy of discrimination an exact printer source of a document from the other printers. In practice, a micro-tag consisting in printed dots can be embedded in an authentic document.

	β	η	λ
HP	0.668	0.225	0.129
Dell	2.218	0.947	0.116

Table 10: The mean values of the estimators of HP laser printer and Dell laser printer.

0.0147	0.001	-0.0003	0.243	0.054	-0.003
0.001	0.009	0.0009	0.054	0.028	0.001
-0.0003	0.0009	0.0007	-0.003	0.001	0.0007
(a) HP			(b) Dell		

Table 11: The covariance matrices calculated from the estimators of HP and Dell laser printers.

Actual/Predicted	HP laser	Dell laser
HP laser	20	0
Dell laser	0	20

Table 12: Confusion matrix for ML classification of two laser printers.

8. Concluding remarks

290 In this paper, we have proposed a novel statistical model that considers the spatial interaction of printed dots at microscopic scale. This model allows to describe more accurately the formation of the shape of the binary dot. The spreading of the ink on the surface was described as the emergence of the black pixels and the interaction between pixels that depends on the distance between the centres of the dots. All these features
295 are modelled by non-stationary Markov chain.

An estimation method for the parameters based on the Metropolis-Hasting within Gibbs algorithm is proposed. The performance of this approach is assessed, not only with simulated data but also with real dots coming from four different printing technologies. Thanks to the proposed model and inference algorithm, the realisations are

300 very close to the real images. The fragmentation on the border of the dot is better
treated, and the density of the black pixels is more realistically simulated with the
newly introduced interaction. Authentication of printers from micro-prints has shown
promising results owing to the accuracy of both the Markov spatial model and the
estimation algorithm. Future research directions are consisting in extending this model
305 to grayscale images and colour images.

Acknowledgments

Q-T. Nguyen wants to dedicate this work to Prof. Y. Delignon, 2nd author of the
paper who passed away before this work can be published.

References

- 310 [1] J. Picard, C. Vielhauer, N. Thorwirth, Towards fraud-proof id documents using
multiple data hiding technologies and biometrics, in: *Electronic Imaging 2004*,
International Society for Optics and Photonics, 2004, pp. 416–427. doi:10.
1117/12.525446.
- [2] R. Villán, S. Voloshynovskiy, O. Koval, T. Pun, Multilevel 2-d bar codes: to-
ward high-capacity storage modules for multimedia security and management,
315 *Information Forensics and Security, IEEE Transactions on* 1 (4) (2006) 405–420.
doi:10.1109/TIFS.2006.885022.
- [3] P.-J. Chiang, N. Khanna, A. K. Mikkilineni, M. V. O. Segovia, S. Suh, J. P. Alle-
bach, G. T.-C. Chiu, E. J. Delp, Printer and scanner forensics, *Signal Processing*
320 *Magazine, IEEE* 26 (2) (2009) 72–83. doi:10.1109/MSP.2008.931082.
- [4] G. N. Ali, A. K. Mikkilineni, J. P. Allebach, E. J. Delp, P.-J. Chiang, G. T.
Chiu, Intrinsic and extrinsic signatures for information hiding and secure printing
with electrophotographic devices, in: *NIP & Digital Fabrication Conference*, Vol.
2003, Society for Imaging Science and Technology, 2003, pp. 511–515.

- 325 [5] S.-J. Ryu, H.-Y. Lee, D.-H. Im, J.-H. Choi, H.-K. Lee, Electrophotographic printer identification by halftone texture analysis, in: *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, IEEE, 2010, pp. 1846–1849. doi:10.1109/ICASSP.2010.5495377.
- [6] J. Oliver, J. Chen, Use of signature analysis to discriminate digital printing technologies, in: *NIP & Digital Fabrication Conference, Vol. 2002*, Society for Imaging Science and Technology, 2002, pp. 218–222.
330
- [7] S. Voloshynovskiy, M. Diephuis, F. Beekhof, O. Koval, B. Keel, Towards reproducible results in authentication based on physical non-cloneable functions: The forensic authentication microstructure optical set (famos), in: *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*, IEEE, 2012, pp. 43–48. doi:10.1109/WIFS.2012.6412623.
335
- [8] A. Sharma, L. Subramanian, E. A. Brewer, Paperspeckle: microscopic fingerprinting of paper, in: *Proceedings of the 18th ACM conference on Computer and communications security*, ACM, 2011, pp. 99–110. doi:10.1145/2046707.2046721.
340
- [9] B. Zhu, J. Wu, M. S. Kankanhalli, Print signatures for document authentication, in: *Proceedings of the 10th ACM conference on Computer and communications security*, ACM, 2003, pp. 145–154. doi:10.1145/948109.948131.
- [10] A. K. Mikkilineni, G. N. Ali, P.-J. Chiang, G. T. Chiu, J. P. Allebach, E. J. Delp, Signature-embedding in printed documents for security and forensic applications, in: *Electronic Imaging 2004*, International Society for Optics and Photonics, 2004, pp. 455–466. doi:10.1117/12.531944.
345
- [11] N. Khanna, A. K. Mikkilineni, G. T. C. Chiu, J. P. Allebach, E. J. Delp, Survey of Scanner and Printer Forensics at Purdue University BT - Computational Forensics: Second International Workshop, IWCF 2008, Washington, DC, USA, August 7-8, 2008. *Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 22–34. doi:10.1007/978-3-540-85303-9_3.
350

- [12] M.-J. Tsai, J.-S. Yin, I. Yuadi, J. Liu, Digital forensics of printed source identification for chinese characters, *Multimedia tools and applications* 73 (3) (2014) 2129–2155. doi:10.1007/s11042-013-1642-2.
- 355
- [13] M.-J. Tsai, C.-L. Hsu, J.-S. Yin, I. Yuadi, Japanese character based printed source identification, in: *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on, IEEE, 2015*, pp. 2800–2803. doi:10.1109/ISCAS.2015.7169268.
- [14] M.-J. Tsai, C.-L. Hsu, J.-S. Yin, I. Yuadi, Digital forensics for printed character source identification, in: *Multimedia and Expo (ICME), 2016 IEEE International Conference on, IEEE, 2016*, pp. 1–6. doi:10.1109/ICME.2016.7552892.
- 360
- [15] M.-J. Tsai, I. Yuadi, Source identification for printed arabic characters, in: *PROCEEDINGS OF THE 9TH IEEE INTERNATIONAL CONFERENCE ON UBI-MEDIA COMPUTING” UMEDIA-2016”*, 2016, pp. 49–53.
- 365
- [16] M.-J. Tsai, I. Yuadi, Digital forensics of microscopic images for printed source identification, *Multimedia Tools and Applications* (2017) 1–30doi:10.1007/s11042-017-4771-1.
- [17] R. A. Ulichney, Review of halftoning techniques, in: *Electronic Imaging, International Society for Optics and Photonics, 1999*, pp. 378–391. doi:10.1117/12.373419.
- 370
- [18] L. Velho, J. d. M. Gomes, Digital halftoning with space filling curves, in: *ACM SIGGRAPH Computer Graphics, Vol. 25, ACM, 1991*, pp. 81–90. doi:10.1145/127719.122727.
- 375
- [19] T. N. Pappas, D. L. Neuhoff, Model-based halftoning, in: *Electronic Imaging’91, San Jose, CA, International Society for Optics and Photonics, 1991*, pp. 244–255. doi:10.1117/12.44360.

- [20] H. S. Baird, Document image defect models, in: *Structured Document Image Analysis*, Springer, 1992, pp. 546–556. doi:10.1007/978-3-642-77281-8_26.
- [21] M. Norris, E. H. B. Smith, Printer modeling for document imaging, in: *Proceedings of the 2004 International Conference on Imaging Science, Systems, and Technology (CISST'04)*, 2004.
- [22] Q.-T. Nguyen, Y. Delignon, L. Chagas, F. Septier, Printer technology authentication from micrometric scan of a single printed dot, in: *Proc.SPIE*, Vol. 9028, 2014, pp. 9028 – 9028 – 7. doi:10.1117/12.2039989.
URL <https://doi.org/10.1117/12.2039989>
- [23] T. Q. Nguyen, Y. Delignon, L. Chagas, F. Septier, Printer identification from micro-metric scale printing, in: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE, 2014, pp. 6236–6239. doi:10.1109/ICASSP.2014.6854803.
- [24] Q. T. Nguyen, Y. Delignon, L. Chagas, F. Septier, Modélisation de points imprimés à l'échelle micro-métrique, in: *XXVème Colloque GRETSI*, 2015.
- [25] D. Hilbert, Ueber die stetige abbildung einer line auf ein flächenstück, *Mathematische Annalen* 38 (3) (1891) 459–460. doi:10.1007/978-3-662-25726-5_1.
- [26] S.-i. Kamata, M. Niimi, E. Kawaguchi, Interactive analysis of multi-spectral images using a hilbert curve, in: *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, Vol. 1, IEEE, 1994, pp. 93–97. doi:10.1109/ICPR.1994.576234.
- [27] R. Fjortoft, Y. Delignon, W. Pieczynski, M. Sigelle, F. Tupin, Unsupervised classification of radar images using hidden markov chains and hidden markov random fields, *Geoscience and Remote Sensing, IEEE Transactions on* 41 (3) (2003) 675–686. doi:10.1109/TGRS.2003.809940.

- [28] N. Dridi, Y. Delignon, W. Sawaya, C. Garnier, Em-based joint symbol and blur estimation for 2d barcode, in: Image and Signal Processing and Analysis (ISPA), 2011 7th International Symposium on, IEEE, 2011, pp. 32–36.
- 410 [29] J. Zhang, S.-i. Kamata, Y. Ueshige, A pseudo-hilbert scan algorithm for arbitrarily-sized rectangle region, in: Advances in Machine Vision, Image Processing, and Pattern Analysis, Springer, 2006, pp. 290–299. doi:10.1007/11821045_31.
- [30] C. Robert, G. Casella, Monte Carlo Statistical Methods, Springer Science & Business Media, 2013. doi:10.1007/978-1-4757-4145-2.
- 415 [31] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, in: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.
- [32] P. Tichavský, C. H. Muravchik, A. Nehorai, Posterior cramer-rao bounds for discrete-time nonlinear filtering, Signal Processing, IEEE Transactions on 46 (5) 420 (1998) 1386–1396. doi:10.1109/78.668800.
- [33] H. L. Van Trees, Detection, estimation, and modulation theory, John Wiley & Sons, 2004.
- [34] G. M. Foody, N. Campbell, N. Trodd, T. Wood, Derivation and applications of probabilistic measures of class membership from the maximum-likelihood classification, Photogrammetric engineering and remote sensing 58 (9) (1992) 1335– 425 1341.
- [35] A. Asmala, Analysis of maximum likelihood classification on multispectral data, Applied Mathematical Sciences 6 (129-132) (2012) 6425–6436.

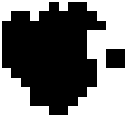

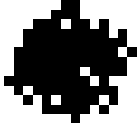


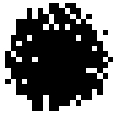






	Real	Markov spatial model	Spatial binary model
Conventional offset			
Waterless Offset			
Laser			
Inkjet			

Figure 7: The estimation from four printing processes and the realisations from the models.

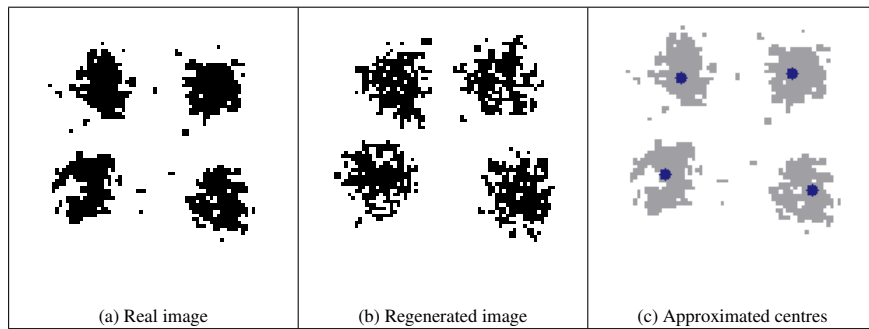


Figure 8: The real dots vs. the regenerated image, the approximated centres are demonstrated.

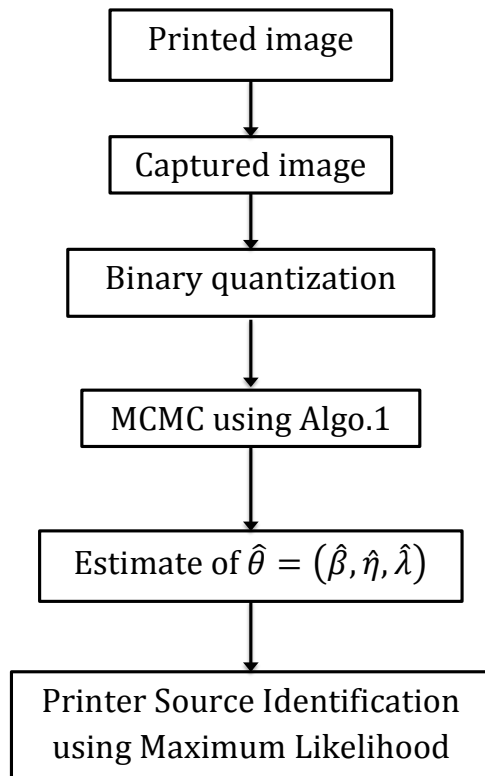


Figure 9: Procedure of printer identification scheme.

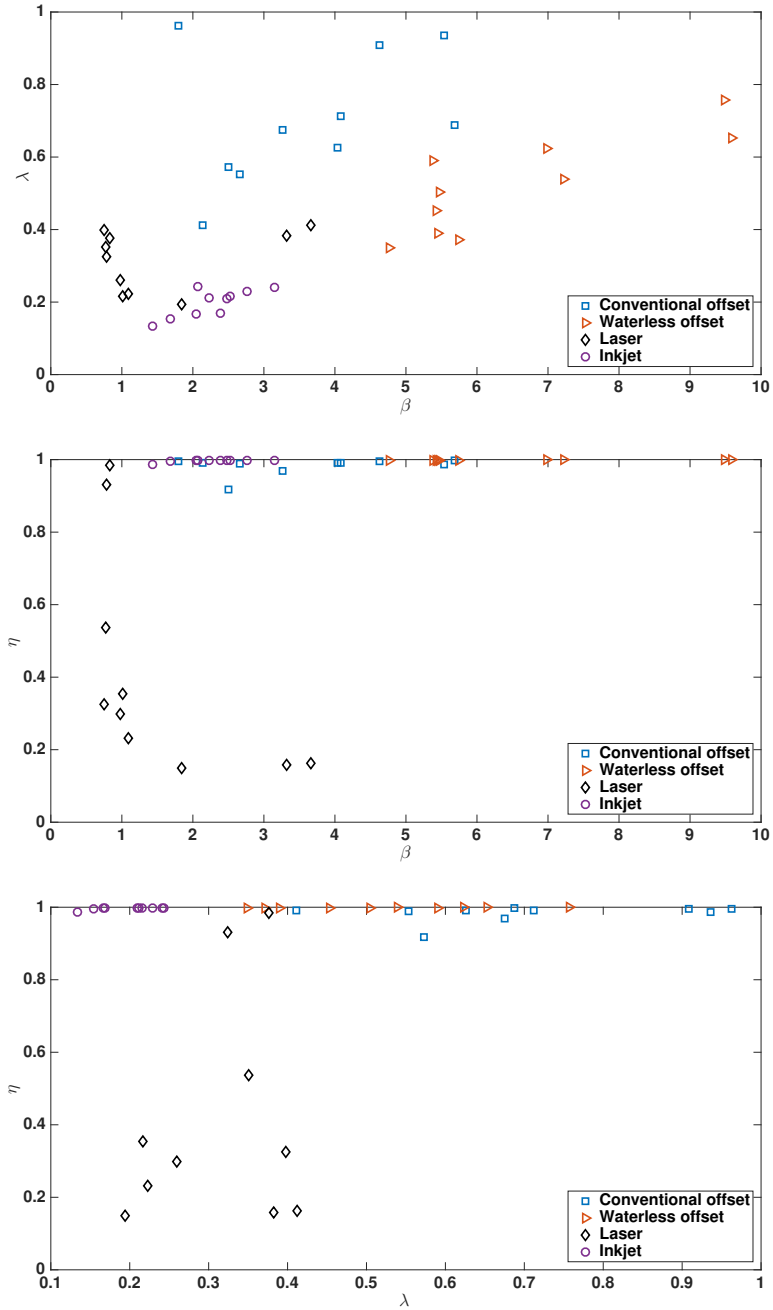


Figure 10: Relation between the estimated parameters of the samples from four printers with four printing technologies.

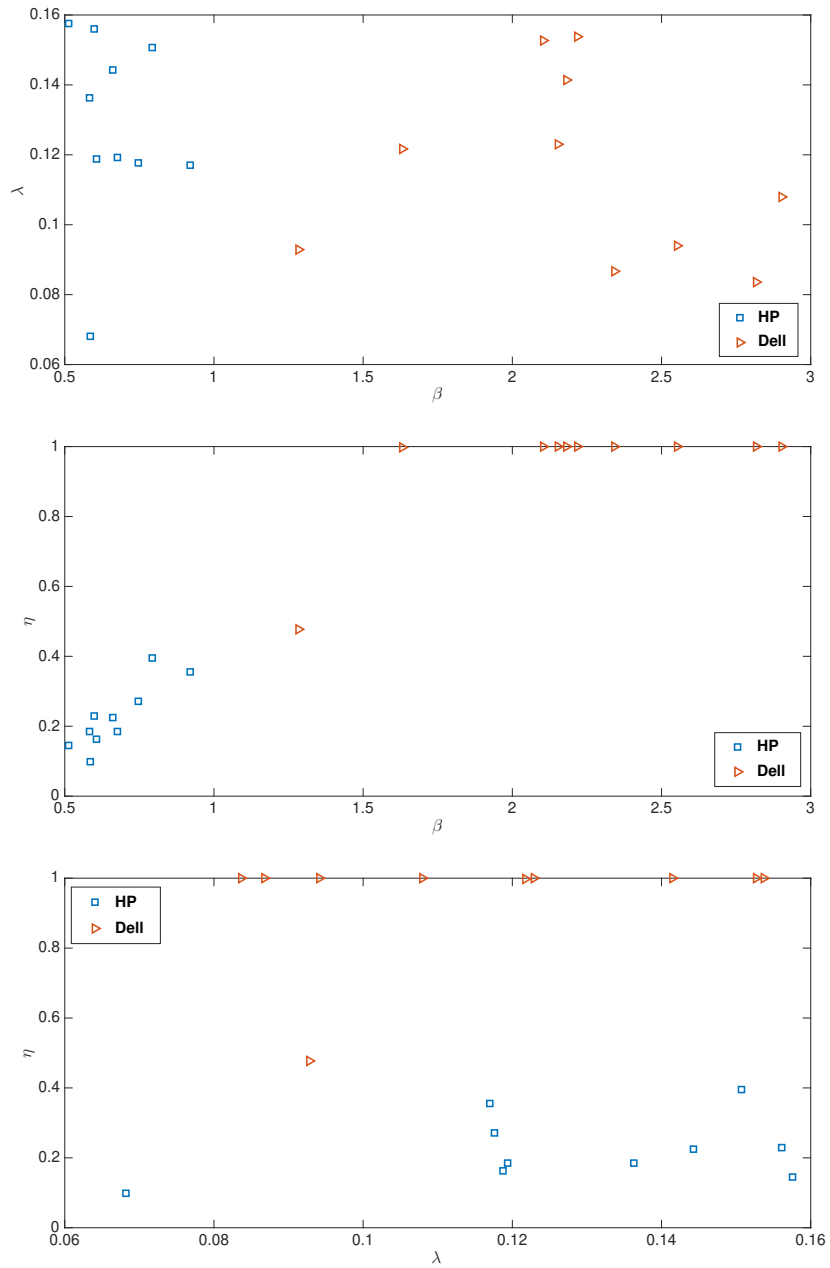


Figure 11: Relation between the estimated parameters of the samples from a HP laser printer and a Dell laser printer.