Second year review

Ondřej Mandula

12th July 2011

Contents

1 Introduction **1**

1.1 Localisation microscopy 1

1.2 PALM, STORM 1

1.3 Quantum dots 1

1.4 Localisation microscopy using Quantum Dots 2

1.5 Non-negative matrix factorisation 2

1.6 Comparison NMF to the Richardson Lucy deconvolution 2

2 Model comparison problem **4**

2.1 Gamma Poisson (GaP) model 4

2.2 Variational treatment of the GaP model 5

2.3 Analysis of the residuals 7

2.4 Results 7

2.5 Conclusions 12

3 Theoretical limits of the LM **13**

3.1 Cramer Rao bound 13

3.2 Fundamental resolution measure (FREM) 13

3.3 An alternative derivation of the FREM 13

3.4 FREM for blinking sources 15

3.5 Simulations 16

3.6 Conclusions 18

4 Out of focus PSF **19**

4.1 Results 19

4.2 Conclusions 21

References **22**

A NMF as a minimisation of Kullback–Leibler divergence **25**

B Variational approximation for GaP model **25**

C Resolution limit for the blinking QDs **31**

Outline The second year review report summarises work on my PhD project I have done since September 2010. Section 1 gives a short introduction to localisation microscopy methods and is put into context of the recent research. Section 2 refers to the problem of the model selection for localisation microscopy data with overlapping sources. Theoretical limits of the localisation microscopy are discussed in Section 3. Section 4 demonstrates ability of the NMF algorithm to separate sources with different PSFs.

1 Introduction

1.1 Localisation microscopy

Localisation microscopy (LM) provides a conceptually simple way to super-resolution microscopy as a single source can be localised with an uncertainty much smaller then a classical Rayeigh's resolution criterion (approximately $\lambda_{em}/2$, where λ_{em} is an emission wavelength of the source [Rayleigh 1896; Born et al. 1975]). The Rayleigh criterion neglects the stochastic nature of the photon-detection process and does not take the total photon count (intensity of the source) into account. As shown in [Thompson 2002; Ram et al. 2006b] the variance of a single source localisation is inversely proportional to the number of photons we can collect from this source. For sufficiently intense sources the localisation precision can significantly surpass the Rayleigh resolution limit [Gordon et al. 2004; Qu et al. 2004; Lidke et al. 2005; Ober et al. 2004].

1.2 PALM, STORM

In the real biological samples the sources are usually dense and highly overlapping making the localisation of the individual sources very difficult. PALM (Photo Activation Localisation Microscopy) [Hess et al. 2006] and STORM (STochastic Optical Resolution Microscopy) [Bates et al. 2007] separate the individual emitters by activating only a small subset of them at a time. If the activated subset is small enough the individual fluorophores are well spatially dispersed and the individual sources can be identified and localised. This requires a control of the sources activation/excitation. It can be achieved either by using photo-activable (PA) fluorophores (PALM, STORM) or by reversive photo - bleaching [Baddeley et al. 2009]. Repetition of this activation - localisation cycles can provide super-resolution images of biological samples [Shroff et al. 2008; Huang et al. 2008].

1.3 Quantum dots

Quantum dots (QD) are an order of magnitude brighter and several orders of magnitude more photo-bleaching resistant fluorophores compared to the organic dyes used in LM [Jaiswal and Simon 2004; Michalet et al. 2005]. QDs also provide a wide absorption and a narrow excitation spectrum which is very convenient for multicolour imaging. All these properties make QD very attractive for biological research.

Under a continuous excitation QDs exhibit a blinking behaviour (fluorescence intermittency). They switch between "ON" episodes ($1/\tau_{\mathrm{ON}}^{m}$) of a rapid absorption-fluorescence cycling and "OFF" episodes ($1/\tau_{\mathrm{OFF}}^{m}$) where no light is emitted despite the continuous excitation. Both ON-time and OFF-time probability densities follow an inverse power law $P(\tau_{\mathrm{ON/OFF}}) \propto 1/\tau_{\mathrm{ON/OFF}}^{m}$ [Kuno et al. 2001; Stefani et al. 2009]. However, the blinking process is not yet fully understood .

Despite all the advantages the QD can potentially provide, they are not suitable for the standard LM methods (PALM/STORM). The QD blinking behaviour is difficult to control and the overlapping sources cannot be separated and localised individually.

1.4 Localisation microscopy using Quantum Dots

In 2005 there has been published a method exploiting the fluorescence intermittency ('blinking') of QDs under continuous excitation [Lidke et al. 2005]. A time series of the blinking QDs was recorded and analysed using Independent Component Analysis (ICA); FastICA algorithm [Hyvärinen and Oja 2000] has been used. Localisation of two quantum dots separated down to 23 nm (corresponding to $\lambda_{\mathrm{em}}/30$) has been reported [Lidke et al. 2005]. Further exploration of the technique for more than two sources and for different configuration of the experiment can be found in [Lidke and Heintzmann 2007].

A Bayesian approach to the blinking QD data has been presented [Harrington et al. 2008]. A model consisting of PSFs with known shape (all sources are assumed to have PSF of the same form) and individual intensities (normally distributed) was fitted to the data. Localisation of several QDs within the diffracted limited volume has been shown.

A method using QDs for measurement of sub-resolution distances has been published in [Lagerholm et al. 2006]. However, discrete ON-OFF blinking is required (only one source being ON and others OFF) as opposed to [Lidke et al. 2005; Harrington et al. 2008] where only fluctuation of the individual sources is needed.

1.5 Non-negative matrix factorisation

Non-negative matrix factorisation (NMF) seems to be a very natural model for describing the blinking QD data. The expectation value of noisy spatio-temporal $N \times T$ data matrix $\boldsymbol{D}$ (N - total number of pixels, T - number of time slices) is assumed to be decomposed into the $N \times K$ spatial component matrix $\boldsymbol{W}$ (images of the individual sources) and the $K \times T$ temporal component matrix $\boldsymbol{H}$ (intensities of the sources).

$$\mathbb{E}\left[\boldsymbol{D}\right] = \mathbb{E}\left[d_{xt}\right] = \left(\boldsymbol{W}\boldsymbol{H}\right)_{xt} = \sum_{k=1}^{K} w_{xk} h_{kt}, \tag{1.1}$$

with non-negativity constraints d_{xt}, w_{xk} and $h_{kt} \geq 0$.

The NMF algorithm shown in [Lee and Seung 2001] minimises the Kullback–Leibler (KL) divergence between the data and the model. It can be shown that this is equivalent to the maximisation log-likelihood function of the model Eq (1.1) under the assumption of the Poisson noise (Appendix A). The details of the the algorithm are described in the first year review.

1.6 Comparison NMF to the Richardson Lucy deconvolution

An observed 'blurred' (diffraction limited) image $\boldsymbol{i}$ ($N \times 1$ vector) can expressed as a (discretised) convolution

$$i_x = \sum_{j=1}^{N} o_j w_{x-j},$$

where $\boldsymbol{o}$ ($N \times 1$) is the original (unblurred) object which represents locations and intensities of fluorescent sources. $\boldsymbol{w}$ ($N \times 1$) is an image of point spread function (PSF) centred in the middle of the image Richardson [Richardson 1972] and Lucy [Lucy 1974] published an iterative deconvolution technique for astronomical images with known PSF. They used Bayes theorem as a 'hint' for an iterative update of $\boldsymbol{o}$. This update is usually referred to as Richardson-Lucy (RL) deconvolution algorithm and is identical to the Lee-Seung NMF update with generalised KL-divergence objective function [Lee and Seung 2001]. However, $\boldsymbol{o}$ and $\boldsymbol{w}$ represent different objects then the matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ in the NMF model. In the RL deconvolution one PSF $\boldsymbol{w}$ is shared by all sources. In the NMF each source has its own PSF (columns of $\boldsymbol{W}$).

Holmes [Holmes 1992] derived the RL updates based on maximum likelihood estimation of the model with Poisson noise using the expectation-maximisation algorithm. He also proposed an update for $\boldsymbol{w}$ so that the method can be used as a blind deconvolution algorithm (PSF is not known). They are sometimes referred to as a 'blind RL algorithm'. The updates for $\boldsymbol{o}$ and $\boldsymbol{w}$ are technically identical to the Lee and Seung NMF updates (KL divergence as an objective function). Modified updates imposing radial symmetry constraints on PSF were also proposed.

There exist several modified updates derived using EM algorithm which impose some constraints on $\boldsymbol{o}$ or $\boldsymbol{w}$. [Joshi and Miller 1993] gives updates where Good's roughness measure ($\int \frac{|\nabla f(x)|^2}{f(x)} dx$) on the original image $\boldsymbol{o}$ is used as a regularisation term. This biases the solution towards the 'smooth' images and avoids speckle artefacts in the reconstructions that are sometimes experienced in deconvolution methods.

[Fish et al. 1995] use RL 'blind' algorithm (updates on both $\boldsymbol{o}$ and $\boldsymbol{w}$) but after some number of iterations they fit some approximation of the PSF to the estimated $\boldsymbol{w}$ and uses this fit as a new $\boldsymbol{w}$. They claim that in noisy images this 'semi-blind' deconvolution can perform better than the one with known PSF.

The comparison of the regularised RL versions and some other deconvolution techniques has been shown in [van Kempen et al. 1997; Verveer et al. 1999]. RL usually performs well for noisy images.

2 Model comparison problem

NMF requires a prior knowledge about the number of components to be separated (K in Eq.(1.1) - rank of the factorisation). For noise-free data it is possible to estimate the number of sources by analysing principal components (PC), for example. However, in the noisy case the estimation of K is difficult.

The standard NMF algorithm maximises the likelihood of the model Eq.(1.1) under the assumption of the Poisson noise corrupted data (Appendix A). The likelihood function increases with higher K as the more flexible model fits the data better. The Bayesian Information Criterion (BIC) is a rough approximation of the Bayesian treatment penalising the complexity of the model. By evaluating the model for different values of K we can compare the BIC score. However, as shown in the first year review the BIC might be too crude approximation for the correct estimation of the data dimensionality.

A generative model underlying NMF is presented in section 2.1. A variational approximation for the Bayesian treatment of the problem (section 2.2) can provide estimation of K in the QD data.

A different approach to the model selection is shown in section 2.3. Analysis of the correlations in residuals (data-model) is used to estimate the number of components K.

These two different approaches have been applied to the simulated data with different densities of the sources. The results are shown in section 2.4.

2.1 Gamma Poisson (GaP) model

Gamma-Poisson (GaP) model [Canny 2004] has been proposed as a probabilistic model for documents. It represents a generative probabilistic model for NMF Eq.(1.1). The entries h_{kt} (intensities of the sources) are treated as hidden variables generated from a Gamma distribution

$$p(h_{kt}|\alpha_k, \beta_k) = \frac{h_{kt}^{\alpha-1}\beta^{\alpha}\exp(-\beta h_{kt})}{\Gamma(\alpha)},$$

and the data d_{xt} modelled as a Poisson variable with mean $\sum_k w_{xk}h_{kt}$

$$p(d_{xt}|w_{xk}, h_{kt}) = \frac{(\sum_k w_{xk}h_{kt})^{d_{xt}}\exp(-\sum_k w_{xk}h_{kt})}{d_{xt}!}, \tag{2.1}$$

where w_{xk}, α_k and β_k are the parameters of the model.

The likelihood function of this model is given by

$$p(D, H|W, K, \theta) = \prod_{t=1}^{T}\prod_{k=1}^{K} p(h_{kt}|\alpha_k, \beta_k)\prod_{x=1}^{N} p(d_{xt}|w_{xk}, h_{kt}), \tag{2.2}$$

and the log-likelihood

$$\log p(D, H|W, K, \theta) = \sum_{t=1}^{T}\left\{\sum_{k=1}^{K}\log p(h_{kt}|\alpha_k, \beta_k) + \sum_{x=1}^{N}\log p(d_{xt}|w_{xk}, h_{kt})\right\}.$$

The coupling between W and H in Eq.(2.1) prevents from integrating out the hidden variables H in Eq.(2.2) [Blei et al. 2003]. This is problematic as in the expectation maximisation (EM) algorithm [Bishop 2006] it is necessary to evaluate the term $\mathbb{E}_{h_{kt}}[\log\sum_k w_{xk}h_{kt}]$. (In [Canny 2004] a crude approximation is used $\mathbb{E}_{h_{kt}}[\log\sum_k w_{xk}h_{kt}] \approx \log\mathbb{E}_{h_{kt}}[\sum_k w_{xk}h_{kt}]$.)

2.2 Variational treatment of the GaP model

Variational treatment of the problem is proposed in [Buntine and Jakulin 2006]. The detailed derivation is provided in Appendix B.

A new latent $N \times K$ matrix V (entries v_{xk}) is introduced such that

$$\begin{array}{c} \sum_{x=1}^{N} v_{xk}^{(t)} = c_k^{(t)} \\ \sum_{k=1}^{K} v_{xk}^{(t)} = d_{xt} \end{array},$$

where the discrete latent vector c_k gives the intensity of the kth component in the (t)th time slice and d_{xt} are the entries of the data matrix D. The distribution of the underlying GaP model now becomes

$$\begin{aligned} h_{kt} &\sim \text{Gamma}(h_{kt}; \alpha_k, \beta_k) \\ c_k^{(t)} &\sim \text{Po}(c_k^{(t)}; h_{kt}) \\ v_{x,k}^{(t)} &\sim \text{Multinom}(v_{xk}^{(t)}; w_{xk}, c_k^{(t)}) \end{aligned}$$

and the likelihood of the GaP model with the latent matrix V is then

$$\begin{aligned} p(V, H|\alpha, \beta, W, K) &= \prod_{t=1}^{T} \prod_{k=1}^{K} p(h_{kt}|\alpha_k, \beta_k) \prod_{x=1}^{N} p(v_{1k}^{(t)}, v_{2k}^{(t)} ... v_{N,k}^{(t)}|h_{kt}, w_{xk}) \\ &= \prod_{kt} \text{Gamma}(h_{kt}; \alpha_k, \beta_k) \prod_{x} \text{Po}(c_k^{(t)}; h_{kt}) \times \text{Multinom}(v_{xk}^{(t)}; w_{xk}, c_k^{(t)}) \end{aligned}$$

explicitly

$$p(V, H|\alpha, \beta, W, K) = \prod_{kt} \frac{\beta_k^{\alpha_k} h_{kt}^{c_k^{(t)}+\alpha_k-1} \exp(-(\beta_k+1)h_{kt})}{\Gamma(\alpha_k)} \prod_{x} \frac{w_{xk}^{v_{xk}^{(t)}}}{v_{xk}^{(t)}!}. \tag{2.3}$$

D is derived from V ($d_{xt} = \sum_k v_{xk}^{(t)}$) so it is not explicitly represented.

A factored posterior approximation is made for the latent variable to find expectations as part of an optimisation step.

$$p(V, H|\alpha, \beta, W, K) \approx q(V, H) = q_V(V) q_H(H), \tag{2.4}$$

where the optimal solution is given by [Bishop 2006].

$$\begin{aligned} \log q_H^*(H) &= \mathbb{E}_{V \sim q_V} \left[\log p(V, H, D|W, \alpha, \beta)\right] + \text{const} \\ \log q_V^*(V) &= \mathbb{E}_{H \sim q_H} \left[\log p(V, H, D|W, \alpha, \beta)\right] + \text{const} \end{aligned}$$

The likelihood of the data is bounded by

$$p(D|W, \alpha, \beta, K) \geq \mathcal{L}(q, W, \alpha, \beta), \tag{2.5}$$

where

$$\mathcal{L}(q, W, \alpha, \beta) = \mathbb{E}_{V,H \sim q(V,H)} \left[\log p(H, V, D|W, \alpha, \beta, K)\right] + C \tag{2.6}$$

is a variational lower bound [Bishop 2006] on $\log p(D|W, \alpha, \beta, K)$ and the constant C contains the entropy terms of q_H and q_V which are constant wrt W.

The factorised form Eq.(2.4) and the likelihood Eq.(2.3) suggest

$$\begin{aligned} q_H(H) &= \prod_k \mathrm{Gamma}(h_{kt}; a_k^{(t)}, b_k) \\ q_V(V) &= \prod_{xk} \mathrm{Mutlinom}(v_{xk}^{(t)}; n_{xk}^{(t)}, d_x) \end{aligned}$$

and the update rules for the parameters n_{xk}, a_k and b_k can be derived (for details see Appendix B)

$$\begin{aligned} n_{xk}^{(t)} &= \frac{1}{z_x} W_{xk} \exp(\psi_0(a_k^{(t)}) - \log b_k) \\ a_k^{(t)} &= \sum_{x=1}^{N} n_{xk}^{(t)} d_{xt} + \alpha_k \\ b_k &= 1 + \beta_k \end{aligned} \tag{2.7}$$

where z_x is the normalisation constant $z_x = \sum_k n_{xk}$ and ψ_0 is digamma function (logarithmic derivation of the gamma function).

Maximising the lower bound Eq.(2.6) with respect to w_{xk} gives

$$w_{xk} = \frac{\sum_t n_{xk}^{(t)} d_{xt}}{\lambda_k}, \tag{2.8}$$

where $\lambda_k = \sum_x w_{xk}$ is the normalisation constant.

The variational lower bound on the log-likelihood Eq.(2.6) then becomes

$$\mathcal{L} = \sum_t \left\{ \sum_k \mathbb{E}_H \left[\log h_{kt} \right] (\alpha_k - a_k^{(t)}) + \sum_k \log \frac{\Gamma(a_k^{(t)}) \beta_k^{\alpha_k}}{\Gamma(\alpha_k) b_k^{a_k^{(t)}}} + \sum_x d_{xt} \log z_x - \log \prod_x d_{xt}! \right\}. \tag{2.9}$$

The algorithm is summarised in Algorithm 1 and has been implemented in MATLAB.

Algorithm 1 Variational approximation updates.

Repeat until convergence:

1. For each time slice t of the stack: update $n_{xk}^{(t)}$ and $a_k^{(t)}$ according to Eq.(2.7). (Variational E step)
2. Update W according to the Eq.(2.8). (Variational M step.)
3. Compute the variational lower bound Eq.(2.9) and check for convergence.

GaP model treats the intensities h_{kt} as latent variables. $\mathcal{L}$ Eq.(2.6) is the lower bound for the data log-likelihood $p(D|K)$ Eq.(2.5) with latent variables h_{kt} integrated out and parameters α, β and W omitted for clarity. We can consider several models with different number of components K. If we assign equal prior probabilities $p(K)$ to each of the model, then the maximisation of the lower bound $\mathcal{L}$ is equivalent to maximising $p(K|D) \propto p(D|K)p(K)$ because $p(D|K) \geq \mathcal{L}$ and $p(K) =$ constant. We can then pick the model with K giving the highest probability $p(K|D)$. Thus the variational lower bound computed for different values of K can be used for the K selection [Bishop 2006]. However, we assumed here that the difference between $\mathcal{L}$ and $p(D|K)$ remains approximately constant over some range of Ks. In reality, the $\mathcal{L}$ is the lower bound of $p(D|K)$ and the value of K which gives maximum $\mathcal{L}$ does not necessarily be the one which maximises $p(D|K)$ and in turn $p(K|D)$.

This bayesiand treatment should be contrasted with the maximum-likelihood approach (such as NMF) where h_{kt} (and w_{xk}) are treated as parameters of the model and the likelihood of the model is maximised with respect to these parameters. The more complex models with higher K will always yield lower residual error and thus will be assign higher likelihood. The ML methods favour severely overfited models.

2.3 Analysis of the residuals

A different approach to estimating the number of components K is to analyse the $N \times T$ residual matrix $\boldsymbol{S}$ (entries s_{xt}). After optimising the parameters of the model Eq.(1.1) for different values of K, we compute a normalised residual matrix

$$s_{xt} = \frac{d_{xt} - \sum_{k=1}^{K} w_{xk} h_{kt}}{\sqrt{\sum_{k=1}^{K} w_{xk} h_{kt}}}.$$

The normalisation $\left(\sum_{k=1}^{K} w_{xk} h_{kt}\right)^{-1/2}$ is applied in order to standardise the residuals of Poisson distributed data ($\mathrm{var}(x) = \mathrm{mean}(x)$). We can then compute the $N \times N$ correlation matrix

$$\boldsymbol{C}_S = \boldsymbol{S}\boldsymbol{S}^T,$$

and the $N \times N$ matrix of the correlation coefficients $\boldsymbol{R}_S$ with entries

$$r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii} c_{jj}}}. \tag{2.10}$$

Underestimation of the number of sources (K too small) will lead to correlations between some pixels as the model will try to explain multiple sources by one object. By increasing K the correlations should are expected to decrease until we reach a sufficient number of sources to explain the data and the residuals become uncorrelated.

2.4 Results

In the following section simulated data of 10 sources randomly spatially distributed are considered. Evaluation of the model with a classic NMF algorithm and the variational approximation is presented. K was varied in the range $K = \{4, 5, 6 ... 17\}$.

Simulations

Data were simulated according to the model Eq.(1.1) with parameter set to the values shown in Table 2.1:

1. $K_{true} = 10$ spatial coordinates $[x_k,\ y_k]$ confined to a circular region with radii δ, 1.5δ and 2δ were generated. δ was equal to the Rayleigh resolution limit ($\delta = 0.61\frac{\lambda_{em}}{NA}$). x and y coordinates were drawn from a uniform distribution subject to spatial constraint to the circular area. This represent three datasets with different densities of the sources.

2. An in focus PSF was centred at each coordinate $[x_k,\ y_k]$ (W in Eq.(1.1)).

3. Intensities of the individual sources over time h_{kt} were generated from a uniform distribution over an interval $[0,\ ...\ I_{max}]$ where I_{max} is the maximum intensity of the sources.

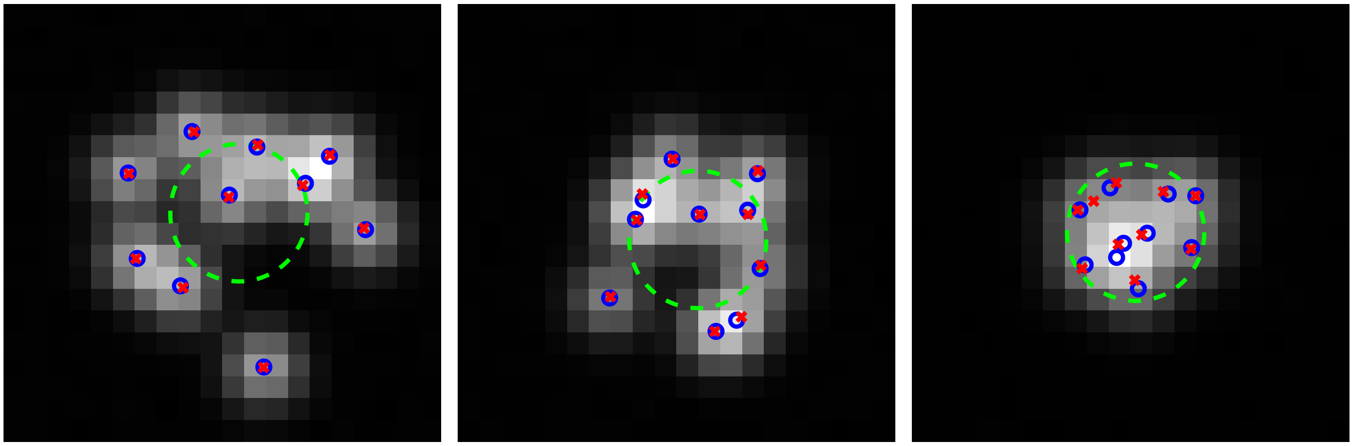
Parameter	Value	Description
T	10^3	Number of time slices in the sequence
K_{true}	10	Number of sources in the simulated data
b	10^2 photons	Uniform background added to each time slice
I_{max}	$1.5 \cdot 10^3$ photons	Maximum intensity of a single source in one time slice
λ_{em}	655 nm	Emission wavelength
NA	1.2	Numerical aperture of the objective
pixel-size	106 nm	Size of a pixel in the sample plane
δ	333 nm (3.1 pixels)	Radius of the region containing the sources ($\delta = 0.61\frac{\lambda_{em}}{NA}$)

Table 2.1: Parameters of the simulation

4. A homogeneous background of b was added to each frame of the time sequence.

5. Intensity in each pixel was corrupted with Poisson noise.

Each dataset with different sources densities was simulated seven-times with different geometrical configurations of the sources. An example of one simulation for each density is shown in Fig. 2.1. A circle with a radius δ is shown in green. The mean intensity image ($\frac{1}{T}\sum_t d_{xt}$ - time sequence of QDs averaged over time) is shown in as grey scale image.



(a) 2δ **(b)** 1.5δ **(c)** δ

Figure 2.1: Example of the simulated data (10 sources) with different densities.The sources are contained in a circular area with a radius 2δ (left), 1.5δ (middle) and δ (right). Mean intensity over time is shown as a grey scale image. The true positions of the sources are shown as blue circles, estimated positions as red crosses. The green circle indicates an area with a radius δ (Rayleigh resolution limit).

Principal component analysis (PCA) for the simulated data is shown in Fig. 2.2a-c. The first 18 PC eigenvalues are plotted. For the the dataset 2δ (Fig. 2.2a) we observe a distinct 'kink' and for $k > 10$ the PC values are negligible. For this dataset we can easily estimate K. However when the density of the sources becomes higher the 'kink' is much less pronounced (Fig. 2.2b-c) and the estimation of K from the PCA analysis becomes impossible.

Evaluation

The classic NMF updates [Lee and Seung 2001] and the variational updates [Buntine and Jakulin 2006] has been used. Each individual data set was evaluated three times with different random initialisations (see below) and the result with the highest likelihood

$$\log p(D|WH, K) = \sum_{x,t} \left\{ d_{xt} \log \sum_{k=1}^{K} w_{xk}h_{kt} - \sum_{k=1}^{K} w_{xk}h_{kt} - \log d_{xt}! \right\} \tag{2.11}$$

has been chosen.

Results were compared to the situation when we set $W = W_{true}$ and $H = H_{true}$, where W_{true} and H_{true} are the matrices used for data simulation. Comparison to the results obtained when W and H were initialised with W_{true} and H_{true} are also shown.

NMF updates

We used the standard NMF updates

$$\begin{aligned} w_{xk} &= \frac{w_{xk}}{\sum_{t=1}^{T} h_{kt}} \left[(D./WH)H^\top \right]_{xk} \\ h_{kt} &= \frac{h_{kt}}{\sum_{x=1}^{N} w_{xk}} \left[W^\top (D./WH) \right]_{kt} \end{aligned} \tag{2.12}$$

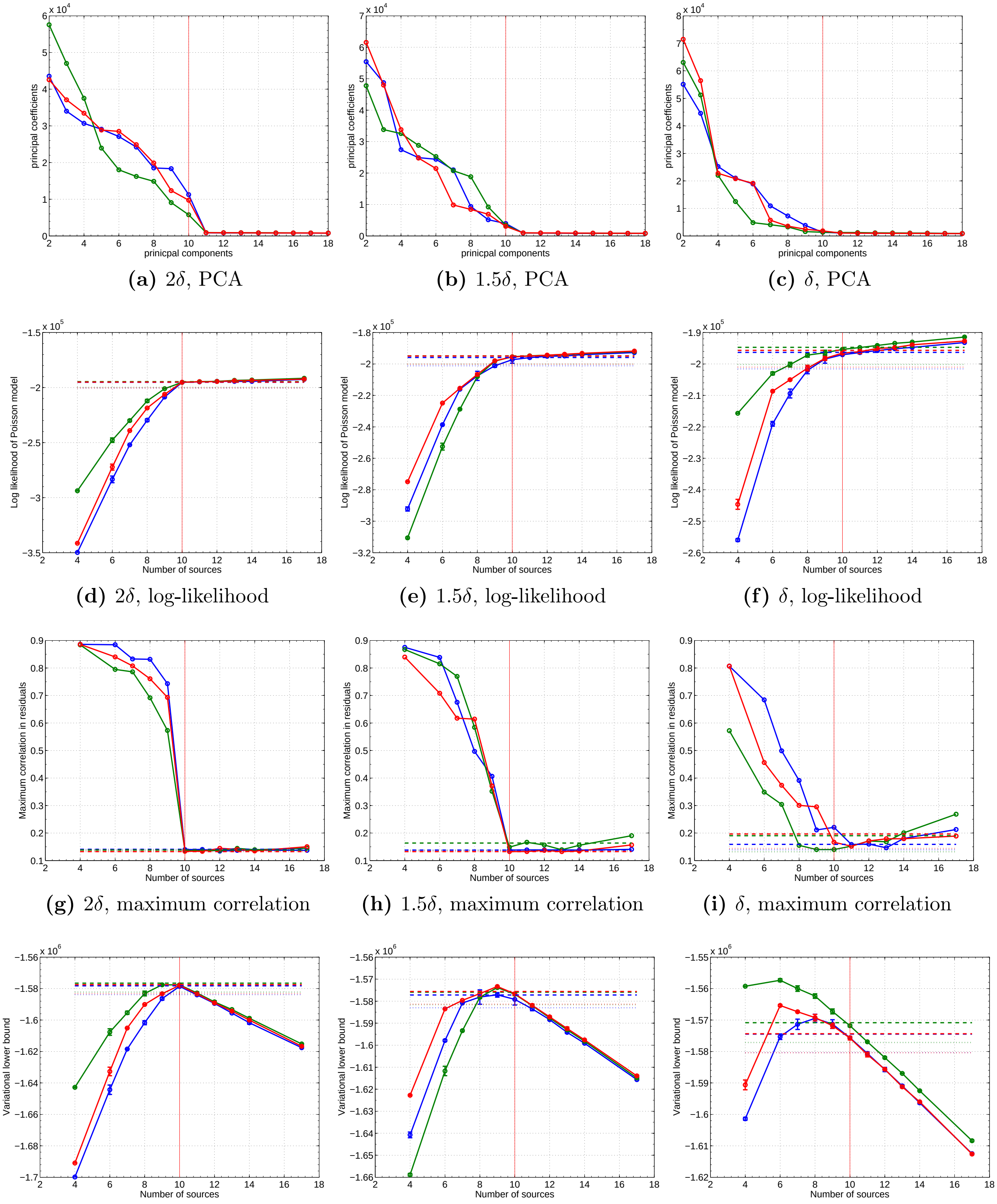
to estimate the non-negative matrices W (images of the individual sources) and H (their intensities) of the model Eq.(1.1). The './' operation refers to an element-wise division. The matrix elements w_{xk} and h_{kt} were initialised at random from uniform distribution. The columns of the matrix W were normalised such that $\sum_x w_{xk} = 1$ and the columns of the matrix H were normalised such that $\sum_{x,k} w_{xk}h_{kt} = \sum_k h_{kt} = \sum_x d_{xt}$.

One component was added as a homogeneous, static background: $w_{x(K+1)} = \frac{1}{N}$, $h_{(K+1)t} = b$, where the background value b was estimated from the dark regions of the data averaged over time $\frac{1}{T} \sum_t d_{xt}$. This component was not updated over the iterations.

We run each NMF evaluation with five partial restarts - after each run convergence we restarted the values of h_{kt} while keeping w_{xk}. This heuristic procedure helps to reach better local minimum (discussed in the first year report.)

Different values of $K = \{4, 5, \ldots 18\}$ were used and the log-likelihood Eq.(2.11) has been reported (Fig. 2.2d-f). We observe an increase of the likelihood function with increasing K. The estimated position of the separated sources (W) with $K = K_{true}$ is shown as blue circles in Fig. 2.1.

We analysed the correlation-coefficient matrix $\boldsymbol{R}_S$ of the residuals Eq (2.10) and plotted the maximum entries in $\boldsymbol{R}_S$ ($\max(r_{ij})$) in Fig. 2.2g-i (minimum values of $\max(r_{ij})$ over three evaluations with different random initialisations is shown). There is a significant drop in the correlation values for $K = 10$ for the first two datasets. For the densest data Fig. 2.1c the correlations drop is less clear and becomes difficult to estimate (Fig. 2.2i). The maximum correlations in the data when W_{true} and H_{true} (the matrices used for data generation) were used are plotted as dotted lines. The results when W and H were initialised with W_{true} and H_{true} are shown as dashed lines. We can observe some increase of the residual correlations after NMF updates with 'true' initialisation. These correlations are higher then in the results evaluated with random initialisation. This might suggest that, for example, the highly overlapping sources can be more efficiently explained by one source and the remaining components used for some minor correction in the model.



(a) 2δ, PCA

(b) 1.5δ, PCA

(c) δ, PCA

(d) 2δ, log-likelihood

(e) 1.5δ, log-likelihood

(f) δ, log-likelihood

(g) 2δ, maximum correlation

(h) 1.5δ, maximum correlation

(i) δ, maximum correlation

(j) 2δ, lower bound

(k) 1.5δ, lower bound

(l) δ, lower bound

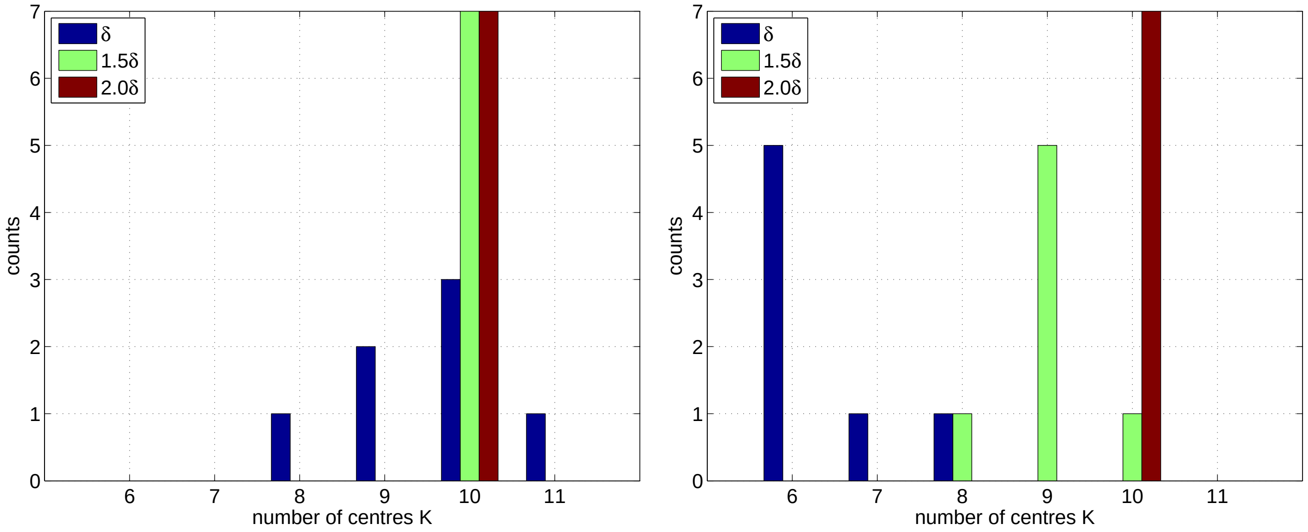
Figure 2.2: K estimation for sources contained in a circular area with radius 2δ (left column), 1.5δ (middle column) and δ (right column) as shown in Fig. 2.1. The true value $K_{true} = 10$ is indicated with a red vertical line. Values for the true W_{true} and H_{true} ($K = K_{true}$) are plotted as horizontal dotted lines. Values for $W$and H evaluated when initialised with W_{true} and H_{true} are plotted as horizontal dashed lines. Three examples for datasets with different geometrical configurations of the sources are shown in each plot.

Variational updates

Algorithm 1 was used for evaluation of data for different K as above. We used an uniform initialisation of $a_k^{(t)} = \frac{1}{K}\left(\sum_k \alpha_k + \sum_x d_{xt}\right)$ from Eq.(2.7). $n_{jk}^{(t)}$ was initialised as a uniform random and normalised $\sum_k n_{jk} = 1$. The parameters of the prior Gamma distribution were set to value $\alpha_k = 1$ and $\beta_k = 2/I_{max}$. This creates a rather flat exponential distribution over positive values of h_{kt} with mean $\alpha/\beta = I_{max}/2$ with I_{max} from the Table (2.1). Each dataset was evaluated three times with different random initialisation of $n_{jk}^{(t)}$.

Similar to the NMF updates, one component as a static homogeneous background (a_{K+1}, $w_{(K+1),x}$) has been added and not updated.

In Fig. 2.2j-l the variational lower bound $\mathcal{L}(K)$ computed for the GaP with different values of K (section 2.2) is shown. For the easy (sparse) data set Fig. 2.1a $\mathcal{L}(K)$ peaks for $K = K_{true}$ (Fig. 2.2j), however, for the more dense data $\mathcal{L}(K)$ reaches the maximum for $K < K_{true}$. For data evaluated when initialised with true values W_{true} and $\mathbb{E}[H] = H_{true}$ ($K = K_{true}$) the lower bound is only slightly higher then for data with random initialisation (dashed line in Fig. 2.2j-l). For δ and 1.5δ datasets this value do not exceed the values for model $K < K_{true}$. This suggests that there is not enough evidence in the data for $K = K_{true}$ as the data can be sufficiently explained with less complicated model $K < K_{true}$.



(a) Maximum correlations in residuals

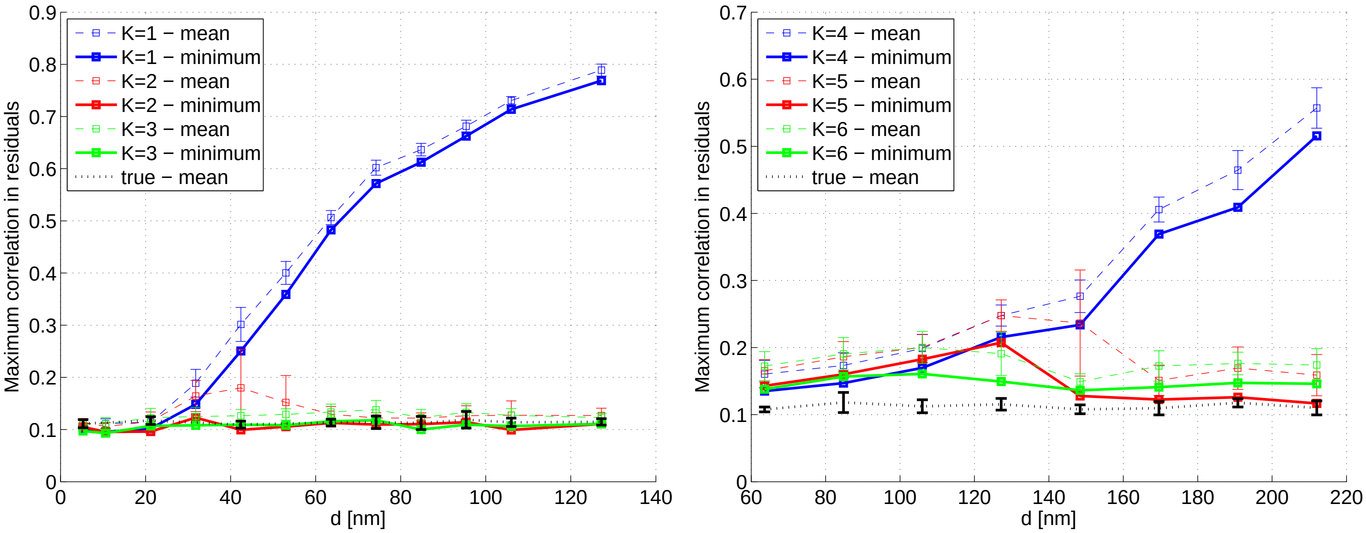
(b) Variational lower bound

Figure 2.3: Histograms of the K estimations for data with three different sources density.

We estimated K from the maximum of the variational lower bound Fig. 2.2j-k and from the estimated 'kink' in the maximum correlations in residuals 2.2g-i for the seven datasets with different geometrical configurations of the sources. The histograms for datasets with three different sources density is shown in Fig. 2.3.

Resolution limit of the residual correlations

From the comparison in Fig. 2.3 it seems that the analysis of the residual correlations gives more precise K estimation. In order to investigate what is the 'resolution' of this K estimation we simulated datasets consisting of $K_{true} = 2$ and $K_{true} = 5$. The sources were distributed uniformly on a circle with diameter d. Each data set was simulated 10 times as a different draws from a Poisson distribution (with identical parameters). Each dataset was evaluated with NMF for $K = \{K_{true} - 1, K_{true}, K_{true} + 1\}$. The maximum correlations coefficients in residuals $\max_{ij}(r_{ij})$ has been recorded. The mean (and standard deviation) and the minimum values over 10 evaluation



(a) $K_{true} = 2$ **(b)** $K_{true} = 5$

Figure 2.4: Maximum residual correlations. The mean (and standard deviation) and the minimum values of 10 evaluations of the datasets are shown. The maximum residual correlations of the true model (the one used for the generation of the data) are shown as black dotted line.

of the datasets for simulations with different d are shown in Fig. 2.4. The maximum residual correlations of the true model (the one used for the generation of the data) are plotted as a baseline for comparison.

The correlations in residuals are much higher when $K < K_{true}$. Once $K = K_{true}$ the correlations drop significantly to a baseline level. Further increase $K > K_{true}$ does not results in lower correlations. This can be observed only if the distance d exceeds a certain 'resolution' limit d_l. From the Fig. 2.4 we can estimate $d_l \approx 40\,\text{nm}$ $(\lambda_{em}/16)$ for two sources and $d_l \approx 150\,\text{nm}$ for five sources.

2.5 Conclusions

The variational lower bound provides correct estimates only for relatively easy data (Fig. 2.1a). K is systematically underestimated for the data with higher densities of the sources (Fig. 2.3b). This is intuitive as several highly overlapping sources corrupted with noise can be approximated with one larger source. Comparison to the results evaluated after initialisation with the 'true' values (matrices used for data simulation) suggests that the higher values for $K < K_{true}$ is due to the lack of evidence in the data rather then convergence to a 'wrong' solution (Fig. 2.2k-l). The value of the lower bound for the 'true' initialisation do not exceed the values for the model with $K < K_{true}$. When the data do not provide enough evidence, the bayesian approach tend to prefer the less complicated model (underestimation of K).

The more accurate estimation of K were achieved by analysis of the correlations in the residuals (Fig. 2.3a), however, the 'kink' can be difficult to estimate for the dense data (Fig.2.2i) and a visual inspection of the curves is ofter required. Some automated procedure for the 'kink' detection is needed.

It is worth emphasising that NMF nor GaP models do not make any assumptions about the time structure of the (latent) variables h_{kt}. This information might provide stronger evidence for the correct K estimation.

3 Theoretical limits of the LM

Results from the variational approximation naturally bring a question about actual limits of the LM method. How close the sources can be in order to be possible to resolve them with the LM method? What are the limiting factors? Does the fluorescence intermittency (blinking) allow for higher resolution? We tried to address these question by examining the Cramér–Rao lower bound for the variance of the estimator on distance between two sources.

3.1 Cramer Rao bound

If $\mathcal{L}(\theta) = \log p(x|\theta)$ is a log-likelihood function for data X, then a covariance matrix $\boldsymbol{Q}$ of an unbiased estimator of $\hat{\theta}$ is bounded by [Rao 1945; Cover and Thomas 1991]

$$\boldsymbol{Q} \geq \boldsymbol{I}^{-1}(\theta), \tag{3.1}$$

where $\boldsymbol{I(\theta)}$ is the Fisher information matrix

$$I_{ij}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j}\right] = \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial\theta_i}\frac{\partial \mathcal{L}}{\partial\theta_j}\right]. \tag{3.2}$$

The inequality Eq.(3.1) is in the sense that $\boldsymbol{Q} - \boldsymbol{I}^{-1}(\theta)$ is a non-negative definite matrix.

3.2 Fundamental resolution measure (FREM)

For PALM/STORM (section 1.2) the spatial resolution limit is determined by the localisation precision for an individual source. The Cramér–Rao bound lower for the position estimation of a single source detected by a CCD camera is derived in [Ram et al. 2006b,a]. The variance is shown to be proportional to $1/\Lambda$ where Λ is the intensity of the source.

A fundamental resolution measure (FREM) for two sources separated by a distance d is shown as an alternative to the Rayleigh resolution criterion considering the photon statistics on the detector (CCD camera). The Fisher information is derived as

$$I(d) = \frac{1}{4}\sum_{n=1}^{N} \frac{\left(\Lambda_1 \int_{C_n} \partial_x q(x-\frac{d}{2})dx - \Lambda_2 \int_{C_n} \partial_x q(x+\frac{d}{2})dx\right)^2}{\Lambda_1 \int_{C_n} q(x-\frac{d}{2})dx + \Lambda_2 \int_{C_n} q(x+\frac{d}{2})dx}, \tag{3.3}$$

where Λ_i is the intensity of the ith source, $q(x)$ is a response function of a source ($\int q(x)dx = 1$) and C_n is an area of the nth pixel. The variance of the estimator on d is then

$$\mathrm{var}(d) = I^{-1}(d).$$

A short summary is shown in Appendix C. There are certain problems with this formula, though. The limit $d \to 0$ gives the Fisher information $I(d) \to 0$ ($\mathrm{var}(d) \to \infty$) for situation when $\Lambda_1 = \Lambda_2$. However, the variance remains finite ($I(d) \neq 0$) when $\Lambda_1 \neq \Lambda_2$ (see discussion in Appendix C). The formula also gives $I(d) \neq 0$ even for the situation when one of the sources is not present ($\Lambda_i = 0$). This stems from the fact that the response function of the sources are assumed to be located at $\pm\frac{d}{2}$ which implicitly assumes the knowledge of the origin location. Only one source is then needed to determined the distance $\frac{d}{2}$.

3.3 An alternative derivation of the FREM

We derived an alternative FREM formula for two sources situation (details in Appendix C). We assume two sources located at positions c_1 and c_2. The response functions (we assume an identical

PSF $q(x)$ for both sources) are $f_1 = q(x - c_1)$ and $f_2 = q(x - c_2)$. The distance between the two sources is $d = c_1 - c_2$ and the variance of d is given by

$$\mathrm{var}(d) = Q_{11} + Q_{22} - 2Q_{12},$$

where $\boldsymbol{Q}$ is the covariance matrix $\boldsymbol{Q} = \boldsymbol{I}^{-1}(\theta)$ and $\boldsymbol{I}(\theta)$ is a (symmetric) Fisher information matrix Eq.(3.2).

For a Poisson distributed pixel counts (this assumes that data are corrupted with Poisson noise only)

$$p(n_p|\theta_p) = \mathrm{Po}(n_p; \lambda_p),$$

where λ_p is the mean value of the pth pixel:

$$\lambda_p = \int_{C_p} \lambda(x)dx = \int_{C_p} \Lambda_1 f_1(x) + \Lambda_2 f_2(x)dx. \tag{3.4}$$

The log-likelihood function for N pixels

$$\log p(n|\theta) = \sum_{p=1}^{N} \mathrm{Po}(n_p; \lambda_p), \tag{3.5}$$

and the Fisher information matrix becomes (Appendix C)

$$I_{ij}(\theta) = \sum_{p=1}^{N} \frac{1}{\lambda_p} \frac{\partial \lambda_p}{\partial \theta_i} \frac{\partial \lambda_p}{\partial \theta_j}.$$

The covariance matrix $\boldsymbol{Q}$ is then

$$\boldsymbol{Q} = \boldsymbol{I}^{-1}(\theta) = \frac{1}{I_{11}I_{12} - I_{12}^2} \begin{pmatrix} I_{22} & -I_{12} \\ -I_{12} & I_{11} \end{pmatrix},$$

and the variance of $d = c_1 - c_2$

$$\mathrm{var}(d) = (1, -1)^T \cdot \boldsymbol{Q} \cdot (1, -1) = \frac{I_{11} + I_{22} + 2I_{12}}{I_{11}I_{12} - I_{12}^2}. \tag{3.6}$$

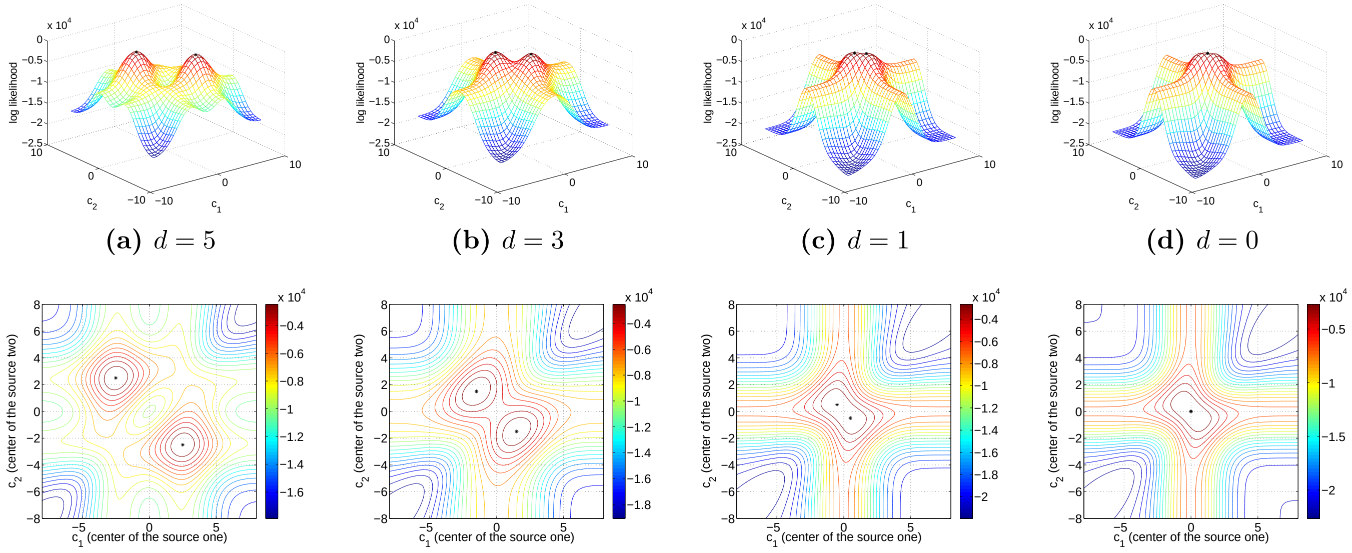
For a symmetrical PSF $q(x-d) = q(x+d)$ we show (Appendix C) that the entries of the Fisher information matrix

$$\begin{aligned} I_{ii} &= \sum_{n=1}^{N} \frac{(\Lambda_i q_n')^2}{\Lambda_i q_n + \Lambda_j q_n(d)} \\ I_{ij} &= \sum_{n=1}^{N} \frac{\Lambda_i \Lambda_j q_n'(0) q_n'(d)}{\Lambda_i q_n + \Lambda_j q_n(d)} \text{ for } i \neq j, \end{aligned} \tag{3.7}$$

where $q_n(d) = \int_{C_n} q(x - d)dx$ ($q_n(0) = q_n$) and $q_n'(d) = \int_{C_n} \frac{\partial q(x-d)}{\partial x} dx$ ($q_n'(0) = q_n'$).

As shown in Appendix C variance defined in Eq.(3.6) have very reasonable behaviour in the limits: the limit $d \to 0$ gives $\mathrm{var}(d) \to \infty$ for any value Λ_i, Λ_j. The variance is also infinite if one of the sources is zero $\Lambda_i = 0$ as we do not make any assumption about the symmetry with respect to the origin ($d = c_1 - c_2$).

For sources which are well separated $d \to \infty$ the off-diagonal elements of the Fisher information matrix vanish ($I_{ij} = 0$ for $i \neq j$) and the variance becomes $\mathrm{var}(d) = \mathrm{var}(c_1) + \mathrm{var}(c_2)$ (sum of the variances for localisation of individual sources).



(a) $d = 5$ **(b)** $d = 3$ **(c)** $d = 1$ **(d)** $d = 0$

(e) $d = 5$ **(f)** $d = 3$ **(g)** $d = 1$ **(h)** $d = 0$

Figure 3.1: Expected log-likelihood Eq.(3.8) as a function of $\mathbf{c} = (c_1, c_2)$ for different separation d between the two sources (located at $\mathbf{c}^{true}$). The true sources position is marked with black asterisk.

In order to understand the behaviour of the Fisher Information matrix we visualised in Fig. 3.1 an expected log-likelihood Eq. (3.5) as a function of the parameter $\mathbf{c} = (c_1, c_2)$.

$$\mathbb{E}_{\mathrm{Po}(n;\lambda^{true})}\left[\sum_{p=1}^{N} \log \mathrm{Po}\left(n_p; \lambda_p(\mathbf{c})\right)\right] = \sum_{p=1}^{N}\left(\lambda_p^{true} \log \lambda_p(\mathbf{c}) - \lambda_p(\mathbf{c})\right) + A, \tag{3.8}$$

where λ^{true} is the true intensity profile sources and A is independent on $\mathbf{c}$. The Fisher Information matrix Eq. (3.2) for $\theta = (c_1, c_2)$ is a Hessian (curvature) of this surface at (c_1^{true}, c_2^{true}). The parameters from Tab. 2.1 have been used.

We can observe that for small, non zero d there is a saddle point for $c_1 = c_2$ (Fig. 3.1b). This settle is less and less pronounces as the separation d gets smaller (Fig. 3.1c) and eventually degenerates into a flat 'crest' for $d = 0$ (Fig. 3.1d). For $d = 0$ ($c_1^{true} = c_2^{true}$) the Hessian becomes singular as the 'crest' is flat in the proximity of $\mathbf{c}^{true}$ along the diagonal direction $c_1 = c_2$ and the variance $\mathrm{var}(d)$ diverges.

3.4 FREM for blinking sources

One of the fundamental questions is whether the fluorescence intermittency (QD blinking) allows for higher resolution compared to the situation when the sources remains static over time. To address this question we assume a simple model of Poisson distributed data with mean values λ_n shown in Eq.3.4. We assume the intensity vector of the two sources $\mathbf{\Lambda} = (\Lambda_1, \Lambda_2)$ to be a random variable distributed over four distinctive states

$$\left\{\mathbf{\Lambda}^1 = (\Lambda_1, 0),\ \mathbf{\Lambda}^2 = (0, \Lambda_2),\ \mathbf{\Lambda}^3 = (\Lambda_1, \Lambda_2),\ \mathbf{\Lambda}^4 = (0, 0)\right\}.$$

This simulates a simple blinking model of two QDs. If the state of $\mathbf{\Lambda}$ were known we would write the likelihood function as

$$l(\theta) = p(n, \mathbf{\Lambda}|\theta) = \prod_{p=1}^{N} p(n_p|\theta, \mathbf{\Lambda})p(\mathbf{\Lambda}),$$

and the expected Fisher information matrix would become (Appendix C)

$$I(\theta) = \sum_{i=1}^{4} p(\mathbf{\Lambda}^i) \sum_{p=1}^{N} \frac{1}{\lambda_p(\theta, \mathbf{\Lambda}^i)} \left(\frac{\partial \lambda_p(\theta, \mathbf{\Lambda}^i)}{\partial \theta} \right)^2 , \tag{3.9}$$

which is the expectation value (with respect to the states $\mathbf{\Lambda}$) of the Fisher information matrix Eq.(3.7).

However, we assume that the variable $\mathbf{\Lambda}$ is fully described by the probability $p(\mathbf{\Lambda})$ over the states. The exact state in each situation is unknown. Therefore we have to integrate over $\mathbf{\Lambda}$ and the likelihood function is then

$$l(\theta) = \prod_{p=1}^{N} p(n_p|\theta) = \prod_{p=1}^{N} \sum_{i=1}^{4} p(n_p, \mathbf{\Lambda}^i|\theta) = \prod_{p=1}^{N} \sum_{i=1}^{4} p(n_p|\theta, \mathbf{\Lambda}^i) p(\mathbf{\Lambda}^i). \tag{3.10}$$

This complicates the evaluation of the Fisher information matrix Eq.(3.2) because of the summation within the logarithm in the log-likelihood

$$\mathcal{L}(\theta) = \log l(\theta) = \sum_{p} \log \sum_{i=1}^{4} p(n_p|\theta, \mathbf{\Lambda}^i) p(\mathbf{\Lambda}^i).$$

In Appendix C we show that the Fisher information matrix for $p(\mathbf{\Lambda}^i) = \frac{1}{4}$ for all i is given by

$$I_{rs}(\theta) = \sum_{p=1}^{N} \mathbb{E}_p \left[\frac{\left(\sum_{i=1}^{4} \frac{\partial \mathrm{Po}(\lambda_p^i)}{\partial c_r} \right) \left(\sum_{l=1}^{4} \frac{\partial \mathrm{Po}(\lambda_p^l)}{\partial c_s} \right)}{\left(\sum_{j=1}^{4} \mathrm{Po}(\lambda_p^j) \right)^2} \right] , \tag{3.11}$$

where $\lambda_p^i = \lambda_p(\mathbf{\Lambda}^i)$ is the mean intensity in the pth pixel when $\mathbf{\Lambda}$ is in the state $\mathbf{\Lambda}^i$. $\mathbb{E}_p\,[.]$ represents the expectation value with respect to $p(n_p|\theta)$ in Eq.(3.10). Expressing the derivatives and the expectation value gives

$$I_{rs}(\theta) = \frac{1}{2} \sum_{p=1}^{N} \left(\frac{\partial \lambda_p^r}{\partial c_r} \right) \left(\frac{\partial \lambda_p^s}{\partial c_s} \right) \sum_{n_p \geq 0} \left[\frac{\left(\sum_{i=\{r,3\}} \mathrm{Po}(n_p; \lambda_p^i) \frac{(n_p - \lambda_p^i)}{\lambda_p^i} \right) \left(\sum_{i=\{s,3\}} \mathrm{Po}(n_p; \lambda_p^i) \frac{(n_p - \lambda_p^i)}{\lambda_p^i} \right)}{\sum_{j=1}^{4} \mathrm{Po}(n_p; \lambda_p^j)} \right] .$$

In Appendix C we show that the limit $d \to 0$ gives $\mathrm{var}(d) \to \infty$ and the limit $d \to \infty$ gives $\mathrm{var}(d) = \frac{1}{I_{11}} + \frac{1}{I_{22}}$. It is also shown that for the the limit $d \to \infty$ the variance in the static case $\mathrm{var}^{\mathrm{static}}(d)$ is a lower bound for the blinking case: $\mathrm{var}(d) > \mathrm{var}^{\mathrm{static}}(d)$. This assumes that the total number of detected photons is equal for static and blinking case.

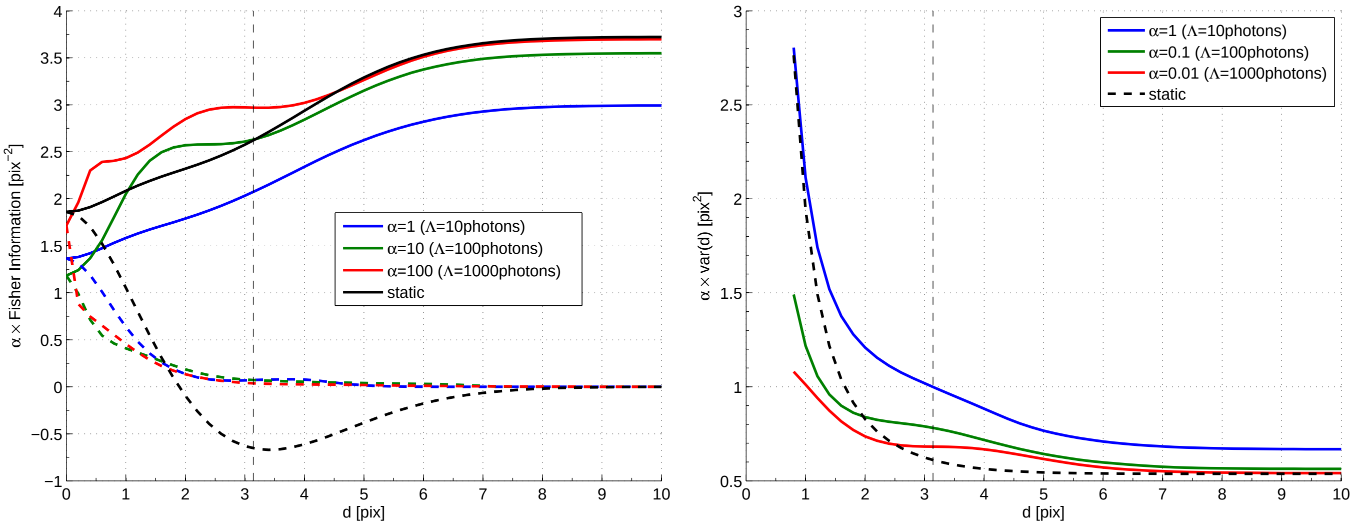
3.5 Simulations

We made a numerical computation of the $\mathrm{var}(d)$ for two sources with equal intensity $\Lambda_1 = \Lambda_2 = \Lambda$. The parameters of the simulated sources are shown in Tab. 2.1. We kept the total intensity (total photon count) equal for both static and the blinking case.

In the static case Eq.(3.6) the variance scales linearly with $1/\Lambda$

$$\mathrm{var}^{\mathrm{static}}(d) \propto \frac{1}{\Lambda}$$

as each entry of the Fisher information matrix $I_{pq} \propto \Lambda$ (Eq.(3.7)). The shape of the curve remains unchanged for different Λ.



(a) Fisher Information matrix **(b)** var(d)

Figure 3.2: Left: Fisher information matrix entries. Solid lines show diagonal entries, broken lines show the off-diagonal elements. Right: variance of the distance estimation (the values must be scaled by a factor α for each line as shown in the legend). Coloured lines show the blinking situation Eq.(3.11) for different intensities Λ of the sources. The black dashed line shows the results for the static case Eq.(3.7). As the curve for the static case does not change the shape for different intensities it is valid for all three different intensities when scaled by the corresponding factor α. This allows for a direct pairwise comparison between each coloured (blinking sources) and the dashed curve (static sources). The vertical broken line indicates the Rayleigh resolution limit.

In the blinking case Eq.(3.11) the dependency on Λ is complicated as the expectation in Eq.(3.11) cannot be simplified and var(d) depends on Λ through the parameter $\lambda_p^i(\Lambda)$ of the Poisson distribution in Eq.(3.11). This gives rise to a highly non-linear relationship between var(d) and Λ.

The comparison of the blinking and the static case for three different values of $\Lambda = 10$, 10^2 and 10^3 photons is shown in Fig. 3.2. The intensity of the static sources was set to $\Lambda/2$ to keep the total number of detected photons constant for the static and blinking case (the blinking sources are on average 'ON' only half of the time and so the average intensity is $\Lambda/2$).

All the curves for the blinking situation for different Λ are plotted as coloured lines in one graph Fig. 3.2. Each coloured curve corresponds to $\Lambda = 10$, 10^2 and 10^3 photons and the values of the var(d) (y-axes) must be scaled by a factor $\alpha = 1$, 10^{-1} and 10^{-2}, respectively. The curve for var$^{\text{static}}(d)$ is plotted as a dashed line. As the curve does not change the shape for different Λ, one curve represents all the three situations when correctly scaled by a factor α. This allows for a pairwise comparison between the static and the blinking case for different Λ. For example, the graph is valid for the green var(d) and the dashed curve var$^{\text{static}}(d)$, when the blue and red curves are ignored and the y-axes is scaled by $\alpha = 10^{-1}$.

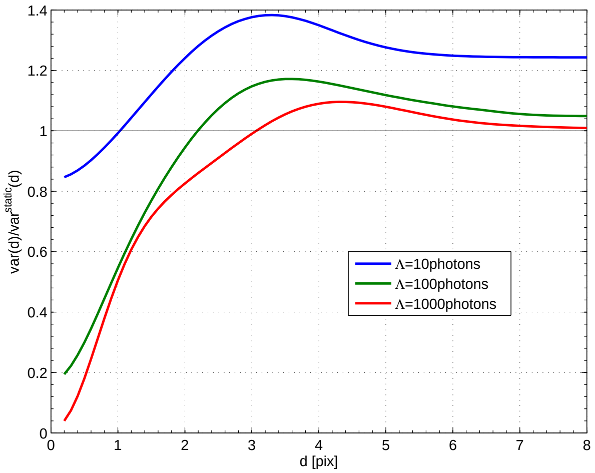


Figure 3.3: Ratio var(d)/var$^{\text{static}}(d)$. The blinking situation gives smaller variance where var(d)/var$^{\text{static}}(d) < 1$.

For small separation of the sources we observe that the blinking situation gives smaller values of variance (coloured curve is under the dashed curve) and thus blinking should allow for localisation with higher precision. The intersection between the coloured and the dashed curve depends on the intensity of the sources Λ. The higher the intensity Λ is the smaller the var(d) results from the blinking situation. Also the region over which the blinking is preferable is larger as we observe that the intersection point between the dashed and the coloured curve shifts towards larger d when increasing Λ. For large separations the static case gives smaller variance and is the lower bound on the blinking case as shown in Appendix C.

The ratio of the coloured (blinking) and the dashed (static) lines for variance (Fig. 3.2b) gives direct performance comparison between the blinking and the static case (Fig. 3.3). In the region where the values are smaller then one, the blinking of the sources gives better localisation precision. For small separation ($d < 50$ nm - 1 pixel corresponds to 106 nm) and for bright sources ($\Lambda > 1000$ detected photons/source) the blinking results in more then 4 times better variance and so more then $\sqrt{4} = 2$ better localisation precision (standard deviation).

3.6 Conclusions

The alternative derivation of the FREM provides a correction to the formula published in [Ram et al. 2006b]. Results presented in the Fig. 3.2 suggest that the intensity blinking of the sources can significantly increase the localisation precision compared to the static situation. This effect is stronger for small separation $d < 50$nm and bright sources ($\Lambda > 1000$ photons). For well separated sources ($d \to \infty$ limit) the static situation provides lower bound for the localisation precision. The intersection points of the curves for static and blinking model depends on the intensity of the sources (Fig. 3.2b).

The results here, however, are computed for the model with no background. More investigation will be needed to explore the effect of the background intensity.

4 Out of focus PSF

The point spread function (PSF) is a rather complicated 3D object. A scan along the axial direction through a PSF of an optical microscope (Fig. 4.1) shows the characteristic ringings for out-of-focus PSF. In a real biological sample PSF from different focal planes can overlap. As the NMF does not make any assumption about the PSFs of the individual sources (parameter matrix W), it is possible to separate overlapping QDs located in different axial positions. This can be possible used for the 3D localisation of the sources. Moreover the PSF of the individual sources can differ due to the aberrations caused by imperfections of the microscope or by inhomogenities of the refractive index in the sample. However, the PSF of each source must remain constant during the data acquisition.

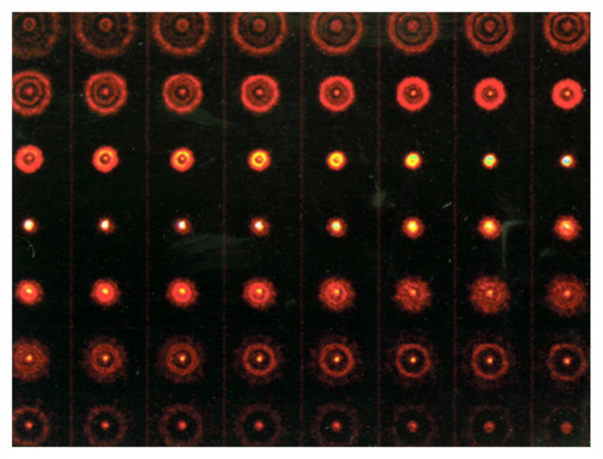


Figure 4.1: Point spread function for different focal planes. Source: www.invitrogen.com

4.1 Results

We recorded a sample of the QDs deposited on a cover slip. Slight variations in the focal plane occurs between the individual QDs which results in different PSFs for each source.

We acquired 10^3 images with 100 ms exposure time (the total acquisition time is about 2 mins). Several images from the time stack are shown in Fig. 4.2. The PCA coefficients of the data are shown in Fig. 4.3a.

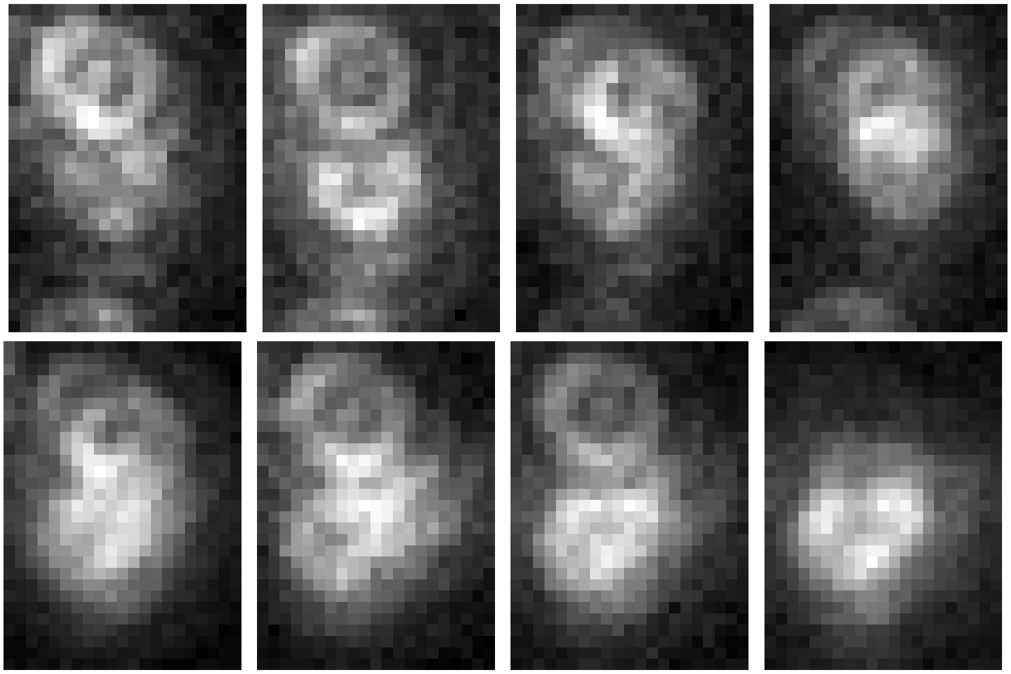
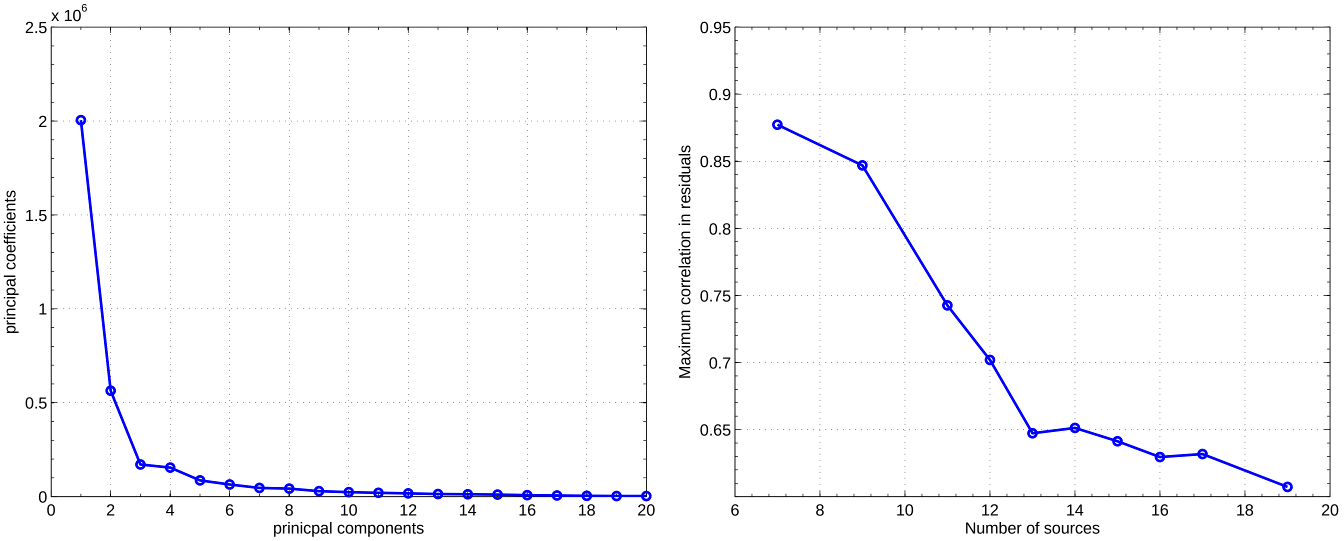


Figure 4.2: Eight images (out of 10^3) of the blinking QD time series.

We used the standard NMF algorithm [Lee and Seung 2001] with different number of sources $K = \{7, 8 \ldots 19\}$ and computed the maximum correlation coefficient of the residuals Eq.(2.10).

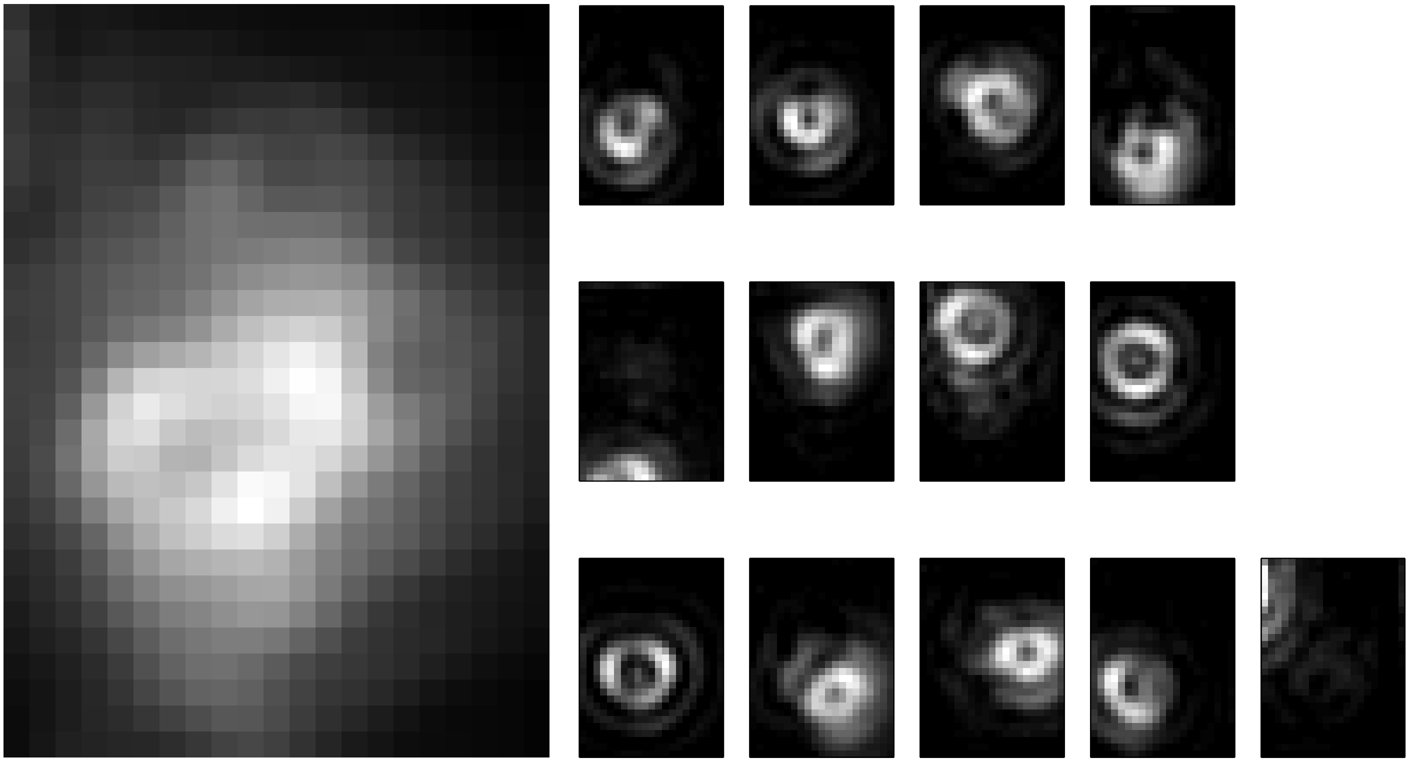
The results are shown in Fig. 4.3b. Unlike the PCA (Fig. 4.3a) we can observe a 'kink' for $K = 13$. Increasing K does not lead to a further decrease of the correlations in residuals. Separated individual PSF (W) for $K = 13$ are shown in Fig. 4.4b. The separated individual PSFs correspond to QDs at different axial positions (see Fig. 4.1).



(a) PCA

(b) Correlations

Figure 4.3: Estimation of K



(a) Mean intensity image

(b) Separated sources

Figure 4.4: Left: mean intensity image correspond to a standard wide-field image. Right: separated individual sources for $K = 13$.

4.2 Conclusions

We demonstrated the capability of NMF to separate individual highly overlapping sources with different PSFs (Fig. 4.4b). The separated sources (Fig. 4.4b) correspond to the real out-of-focus PSFs (Fig. 4.1). K (number of sources in data) has been estimated from the analysis of the residual correlations (Fig. 4.3b).

As the NMF does not make any assumption about the shape of the sources it should be capable of separating the PSF distorted by various aberrations as well. In [Huang et al. 2008], for example, the optical astigmatism is introduced in STORM data in order the PSF changes its elliptical shape for different axial positions. This allows for 3D localisation of the sources. NMF might separate the overlapping aberrated sources and allow this method to use QDs instead of the photo-activable organic fluorophores.

References

Baddeley, D., Jayasinghe, I. D., Cremer, C., Cannell, M. B., and Soeller, C. (2009). Light-induced dark states of organic fluochromes enable 30 nm resolution imaging in standard media. *Biophysical journal*, 96(2):L22–4.

Bates, M., Huang, B., Dempsey, G. T., and Zhuang, X. (2007). Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science (New York, N.Y.)*, 317(5845):1749–53.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Born, M., Wolf, E., and Bhatia, A. B. (1975). *Principles of optics*, volume 10. Pergamon Pr.

Buntine, W. and Jakulin, A. (2006). Discrete component analysis. In Saunders, C., Grobelnik, M., Gunn, S., and Shawe-Taylor, J., editors, *Subspace, Latent Structure and Feature Selection*, pages 1–33. Springer.

Canny, J. (2004). GaP: a factor model for discrete data. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 122–129. ACM.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, volume 1 of *Wiley Series in Telecommunications*. Wiley.

Fish, D. A., Brinicombe, A. M., Pike, E. R., and Walker, J. G. (1995). Blind deconvolution by means of the Richardson–Lucy algorithm. *J. Opt. Soc. Am. A*, 12(1):58–65.

Gordon, M., Ha, T., and Selvin, P. (2004). Single-molecule high-resolution imaging with photobleaching. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6462.

Harrington, P., Anderson, J., Rieger, B., Lidke, D., and Lidke, K. A. (2008). Poster: A Bayesian Approach to Fluorescence Intermittency Based Localization Microscopy. *Supplement of Biophysical Journal*, 96:20–20.

Hess, S. T., Girirajan, T. P. K., and Mason, M. D. (2006). Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical journal*, 91(11):4258–72.

Holmes, T. J. (1992). Blind deconvolution of quantum-limited incoherent imagery: maximum-likelihood approach: errata. *J. Opt. Soc. Am. A*, 9(11):2097.

Huang, B., Wang, W., Bates, M., and Zhuang, X. (2008). Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science (New York, N.Y.)*, 319(5864):810–3.

Hyvärinen, a. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13(4-5):411–30.

Jaiswal, J. K. and Simon, S. M. (2004). Potentials and pitfalls of fluorescent quantum dots for biological imaging. *Trends in cell biology*, 14(9):497–504.

Joshi, S. and Miller, M. I. (1993). Maximum α posteriori estimation with Good's roughness for three-dimensional optical-sectioning microscopy. *J. Opt. Soc. Am. A*, 10(5):1078–1085.

Kuno, M., Fromm, D. P., Hamann, H. F., Gallagher, A., and Nesbitt, D. J. (2001). "On"/"off" fluorescence intermittency of single semiconductor quantum dots. *The Journal of Chemical Physics*, 115(2):1028.

Lagerholm, B. C., Averett, L., Weinreb, G. E., Jacobson, K., and Thompson, N. L. (2006). Analysis method for measuring submicroscopic distances with blinking quantum dots. *Biophysical journal*, 91(8):3050–60.

Lee, D. and Seung, H. (2001). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

Lidke, K. and Heintzmann, R. (2007). Localization fluorescence microscopy using quantum dot blinking. In *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, pages 936–939. IEEE.

Lidke, K. a., Rieger, B., Jovin, T. M., and Heintzmann, R. (2005). Superresolution by localization of quantum dots using blinking statistics. *Optics Express*, 13(18):7052.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79:745.

Michalet, X., Pinaud, F. F., Bentolila, L. a., Tsay, J. M., Doose, S., Li, J. J., Sundaresan, G., Wu, a. M., Gambhir, S. S., and Weiss, S. (2005). Quantum dots for live cells, in vivo imaging, and diagnostics. *Science (New York, N.Y.)*, 307(5709):538–44.

Ober, R. J., Ram, S., and Ward, E. S. (2004). Localization accuracy in single-molecule microscopy. *Biophysical journal*, 86(2):1185–200.

Qu, X., Wu, D., Mets, L., and Scherer, N. (2004). Nanometer-localized multiple single-molecule fluorescence microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31):11298.

Ram, S., Sally Ward, E., and Ober, R. J. (2006a). A Stochastic Analysis of Performance Limits for Optical Microscopes. *Multidimensional Systems and Signal Processing*, 17(1):27–57.

Ram, S., Ward, E. S., and Ober, R. J. (2006b). Beyond Rayleigh's criterion: a resolution measure with application to single-molecule microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12):4457–62.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull Calcutta Math Soc*, 37:81–91.

Rayleigh, L. (1896). On the theory of optical images with special reference to the optical microscope. *Phil Mag*, 5(42):167–195.

Richardson, W. H. (1972). Bayesian-Based Iterative Method of Image Restoration. *J. Opt. Soc. Am.*, 62(1):55–59.

Shroff, H., Galbraith, C., Galbraith, J., and Betzig, E. (2008). Live-cell photoactivated localization microscopy of nanoscale adhesion dynamics. *Nature Methods*, 5(5):417–423.

Stefani, F. D., Hoogenboom, J. P., and Barkai, E. (2009). Beyond quantum jumps: Blinking nanoscale light emitters. *Physics Today*, 62(2):34.

Thompson, R. (2002). Precise Nanometer Localization Analysis for Individual Fluorescent Probes. *Biophysical Journal*, 82(5):2775–2783.

van Kempen, G. M. P., van Vliet, L. J., Verveer, P. J., and van der Voort, H. T. M. (1997). A quantitative comparison of image restoration methods for confocal microscopy. *Journal of Microscopy*, 185(3):354–365.

Verveer, P. J., Gemkow, M. J., and Jovin, T. M. (1999). A comparison of image restoration approaches applied to three-dimensional confocal and wide-field fluorescence microscopy. *Journal of microscopy*, 193(1):50–61.

Zhang, B., Zerubia, J., and Olivo-Marin, J. (2007). Gaussian approximations of fluorescence microscope point-spread function models. *Applied Optics*, 46(10):1819–1829.

A NMF as a minimisation of Kullback–Leibler divergence

A generalised Kullback-Liebler (KL) divergence between two (un normalised) discrete variables P_i and Q_i is defined

$$KL(P \parallel Q) = \sum_i \left(Q_i \log \frac{Q_i}{P_i} - Q_i + P_i \right)$$

NMF algorithm [Lee and Seung 2001] minimised the KL divergence between the data matrix D and the factorised model WH (Eq.(1.1))

$$\min_{W,H\geq 0} KL(D \| WH) = \min_{W,H\geq 0} -\sum_{xt} \left(d_{xt} \log \sum_{k=1}^{K} w_{xk} h_{kt} - \sum_{k=1}^{K} w_{xk} h_{kt} \right) + C \tag{A.1}$$

where C is a constant independent on W and H.

Log-likelihood function of the model Eq.(1.1) under an assumption of Poisson noise

$$\log p(D|W,H) = \sum_{x,t} \log \left(\frac{d_{xt}^{\sum_k w_{xk} h_{kt}} e^{-\sum_k w_{xk} h_{kt}}}{d_{xt}!} \right) = \sum_{xt} \left(d_{xt} \log \sum_{k=1}^{K} w_{xk} h_{kt} - \sum_{k=1}^{K} w_{xk} h_{kt} \right) + D$$

where D is a constant independent on W and H. Comparison with Eq.(A.1) shows that minimisation of the KL divergence between data and the model is equivalent to the maximisation of the log-likelihood function of the model with assumption of the Poisson noise.

B Variational approximation for GaP model

This is a derivation of the variational approximation of the GaP model [Buntine and Jakulin 2006]. From the main text there is a different notation: data $d \rightarrow w$, latent variables (intensities) $h \rightarrow l$, parameters of the model (PSFs of the individual sources) $w \rightarrow \theta$.

Gamma-Poisson (GP) model [Canny 2004]:

$$\mathbb{E}_{w \sim p(w|l,\theta)} [w_j] = \sum_{k=1}^{K} \theta_{jk} l_k$$

- w_j word count of jth word in a document

$$w_j \sim \text{Po}(w_j; (\theta \mathbf{l})_j) = \frac{(\theta \mathbf{l})_j^{w_j} \exp(-(\theta \mathbf{l})_j)}{w_h!}$$

- l_k component scores (vector $\mathbf{l}$) that indicate amount of the component in the document

$$l_k \sim \text{Gamma}(l_k; \alpha_k, \beta_k) = \frac{l_k^{\alpha_k - 1} \beta_k^{\alpha_k} \exp(-\beta_k l_k)}{\Gamma(\alpha_k)}$$

- θ component loading matrix of size $J \times K$. θ_{jk} controls partition of the kth component in the jth word

The log-likelihood of this model:

$$\log p(\mathbf{w}, l | \theta, \text{GP}, \text{K}) = \sum_{k=1}^{K} \left\{ \alpha_k \log(\beta_k) + (\alpha_k - 1) \log l_k - \beta_k l_k - \log \Gamma(\alpha_k) + \sum_{j=1}^{J} \left[w_j \log(\theta \mathbf{l})_j - (\theta \mathbf{l})_j - \log w_j! \right] \right\} \tag{B.1}$$

$$= \sum_{k=1}^{K} \text{log likelihood of } l_k + \sum_{j=1}^{J} \text{log likelihood of } w_j \text{ given } \mathbf{l}$$

Components assignment for words.

Introducing a discrete latent vector $\mathbf{c}$ whose total count is $\sum_j w_j$. The count c_k gives the count of words in the document appearing in the kth component. It is derived from a latent matrix $\mathbf{V}$ of size $J \times K$ (entries v_{jk}).

$$\sum_{j=1}^{J} v_{jk} = c_k$$
$$\sum_{k=1}^{K} v_{jk} = w_j$$

The distribution underlying the GP model now becomes

$$\begin{aligned} l_k &\sim \text{Gamma}(l_k; \alpha_k, \beta_k) \\ c_k &\sim \text{Po}(c_k; l_k) \\ v_{j,k} &\sim \text{Multinom}(v_{jk}; \theta_{jk}, c_k) = c_k! \prod_j \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!} \end{aligned}$$

Proof:

We have $p(c_k|l_k) = \text{Po}(c_k; l_k)$ and $p(v_{jk}|c_k) = \text{Binom}(v_{jk}; \theta_{jk}, c_k) = \binom{c_k}{v_{jk}} \theta_{jk}^{v_{jk}} (1-\theta_{jk})^{c_k - v_{jk}}$ (probability of having v_{jk} counts in c_k counts). Then:

$$\begin{aligned} p(v_{jk}|l_k) &= \sum_{c_k} p(v_{jk}|c_k) p(c_k|l_k) \\ &= \sum_{c_k = v_{jk}}^{\infty} \frac{c_k!}{v_{jk}!(c_k - v_{jk})!} \theta_{jk}^{v_{jk}} (1-\theta_{jk})^{c_k - v_{jk}} \times \frac{l_k^{c_k} \exp(-l_k)}{c_k!} \\ &= \frac{\exp(-l_k)\theta_{jk}^{v_{jk}}}{v_{jk}!} \sum_{c_k = v_{jk}}^{\infty} \frac{l_k^{c_k}(1-\theta_{jk})^{c_k - v_{jk}}}{(c_k - v_{jk})!} && |\alpha_{jk} = c_k - v_{jk} \\ &= \frac{\exp(-l_k)(\theta_{jk} l_k)^{v_{jk}}}{v_{jk}!} \sum_{\alpha_{jk}=0}^{\infty} \frac{(l_k - \theta_{jk} l_k)^{\alpha_{jk}}}{(\alpha_{jk})!} \\ &= \frac{\exp(-l_k)(\theta_{jk} l_k)^{v_{jk}}}{v_{jk}!} \exp(l_k - \theta_{jk} l_k) \\ &= \frac{(\theta_{jk} l_k)^{v_{jk}} \exp(-\theta_{jk} l_k)}{v_{jk}!} \end{aligned}$$

and so $p(v_{jk}|l_k) \sim \text{Po}(v_{jk}; \theta_{jk} l_k)$.

Now sum of two independent Poisson distributed variables $Z = X_1 + X_2$ ($X_i \sim \text{Po}(x; \lambda_i)$)is Poisson distributed:

$$\begin{aligned} p(Z) &= \sum_{x_1=0}^{z} p(X_1) p(Z - X_1) \\ &= \sum_{x_1=0}^{z} \frac{\lambda_1^{x_1} e^{-\lambda_1}}{x_1!} \frac{\lambda_2^{z-x_1} e^{-\lambda_2}}{(z-x_1)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{z!} \sum_{x_1=0}^{z} \frac{z!}{x_1!(z-x_1)!} \lambda_1^{x_1} \lambda_2^{z-x_1} \\ &= \frac{(\lambda_1+\lambda_2)^z e^{-(\lambda_1+\lambda_2)}}{z!} \end{aligned}$$

for more by induction.

So $w_j = \sum_{k=1}^{K} v_{jk}$ is Poisson distributed:

$$w_j \sim \mathrm{Po}(w_j; \sum_{k=1}^{K} \theta_{jk} l_k)$$

The joint distribution for v_{jk} (each is Poisson):

$$\begin{aligned}
p(v_{1,k}, v_{2,k}...v_{J,k}|l_k, \theta_{jk}) &= \prod_{j=1}^{J} \frac{(\theta_{jk} l_k)^{v_{jk}} \exp(-\theta_{jk} l_k)}{v_{jk}!} \\
&= e^{-l_k \sum_j \theta_{jk}} l_k^{\sum_j v_{jk}} \prod_j \frac{\theta^{v_{jk}}}{v_{jk}!} \qquad\qquad |\sum_j \theta_{jk} = 1, \sum_j v_{jk} = c_k \\
&= \frac{l_k^{c_k} e^{-l_k}}{c_k!} c_k! \prod_j \frac{\theta^{v_{jk}}}{v_{jk}!} \\
&= \mathrm{Po}(c_k; l_k) \times \mathrm{Multinom}(v_{jk}; \theta_{jk}, c_k)
\end{aligned}$$

The likelihood of GaP model with latent matrix V is then

$$\begin{aligned}
p(V, l|\alpha, \beta, \theta, K) &= \prod_k p(l_k|\alpha_k, \beta_k) \prod_{jk} p(v_{1k}, v_{2k}...v_{J,k}|l_k, \theta_{jk}) \\
&= \prod_k \mathrm{Gamma}(l_k; \alpha_k, \beta_k) \prod_{jk} \mathrm{Po}(c_k; l_k) \times \mathrm{Multinom}(v_{jk}; \theta_{jk}, c_k)
\end{aligned}$$

explicitly:

$$p(V, l|\alpha, \beta, \theta, K) = \prod_k \frac{\beta_k^{\alpha_k} l_k^{c_k + \alpha_k - 1} \exp(-(\beta_k + 1) l_k)}{\Gamma(\alpha_k)} \prod_{jk} \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!} \tag{B.2}$$

and

$$\log p(V, l|\alpha, \beta, \theta, K) = \sum_k \Bigg\{ (c_k + \alpha_k - 1) \log l_k - (\beta_k + 1) l_k + \alpha_k \log \beta_k - \log \Gamma(\alpha_k) + \sum_j [v_{jk} \log \theta_{jk} - \log v$$

w_j is derived from V so it is not represented.

It is possible to integrate out l (not sure about discrete values...?):

$$\begin{aligned}
p(V|\alpha, \beta, \theta, K) &= \int_0^\infty p(V, l|\alpha, \beta, \theta, K) dl \\
&= \prod_{jk} \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!} \prod_k \frac{\beta_k}{\Gamma(\alpha_k)} \int_0^\infty \left[l_k^{c_k + \alpha_k - 1} \exp(-(\beta_k + 1) l_k) \right] dl_k
\end{aligned}$$

and

$$\begin{aligned}
\int_0^\infty \left[l_k^{c_k + \alpha_k - 1} \exp(-(\beta_k + 1) l_k) \right] dl_k &= \int_0^\infty l_k^{z-1} \exp(-(\beta_k + 1) l_k) dl_k \quad |c_k + \alpha_k = z \\
&= \frac{1}{(\beta_k + 1)^z} \int_0^\infty t^{z-1} \exp(-t) dt | (\beta_k + 1) l_k = t \\
&= \frac{1}{(\beta_k + 1)^z} \Gamma(z)
\end{aligned}$$

so

$$p(V|\alpha,\beta,\theta,K) = \prod_k \frac{\beta_k}{(\beta_k+1)^{c_k+\alpha_k}} \frac{\Gamma(c_k+\alpha_k)}{\Gamma(\alpha_k)} \prod_{jk} \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!}$$

EM algorithm

The term $l_k^{(c_k+\alpha_k-1)} = l_k^{(\sum_j v_{jk}+\alpha_k-1)}$ in Eq.(B.2) links together l_k and V and prevents simple evaluation of $\mathcal{Q}(\theta,\theta^{\text{old}}) = \mathbb{E}_{p(V,l|\theta^{\text{old}})}\left[\log p(V,l|\theta,...)\right]$ in the EM algorithm because of the term $\mathbb{E}_{p(V,l|\theta^{\text{old}})}\left[v_{jk}\right]$. It comes from the Poisson term $\text{Po}(c_k;l_k)$ in $p(V,l|\alpha,\beta,\theta,K)$.

In the likelihood Eq.(B.1) is problematic the term $w_k \log \sum_k \theta_{jk} l_k$. (In [Canny 2004] is the term $\mathbb{E}_l\left[\log \sum_k \theta_{jk} l_k\right]$ approximated by $\log \mathbb{E}_l\left[\sum_k \theta_{jk} l_k\right]$ which might be quite crude.)

Variational Approximation

Factorised approximate posterior distribution for latent variables:

$$p(l,V|w,\alpha,\beta,\theta,K) \approx q(l,V) = q_l(l)q_V(V)$$

Optimal solution [Bishop 2006] (p.466 Eq. (10.9))

$$\log q_l^*(l) = \mathbb{E}_{V\sim q_V}\left[\log p(V,l,w|\theta,\alpha,\beta)\right] + \text{const} \tag{B.4}$$

$$\log q_V^*(V) = \mathbb{E}_{l\sim q_l}\left[\log p(V,l,w|\theta,\alpha,\beta)\right] + \text{const} \tag{B.5}$$

The lower bound is given by [Bishop 2006] (p.465 Eq. (10.3))

$$\mathcal{L}(q,\theta) = \sum_z q(Z) \log \frac{p(X,Z|\theta)}{q(Z)} = \sum_z q(Z) \log p(X,Z|\theta) + H(q_l) + H(q_V)$$

where

$$H(q_l) = -\mathbb{E}_{l\sim q_l}\left[\log q_l\right]$$
$$H(q_V) = -\mathbb{E}_{V\sim q_V}\left[\log q_V\right]$$

are the entropy terms.

$$\log p(w|\theta,\alpha,\beta,K) \geq \mathbb{E}_{l,V\sim q(l,V)}\left[\log p(l,V,w|\theta,\alpha,\beta,K)\right] + C \tag{B.6}$$

The functional form of the complete likelihood suggests

$$q_l(l) = \prod_k \text{Gamma}(l_l;\alpha_k,\beta_k) = \prod_k \frac{l_k^{a_k-1} b_k^{a_k} \exp(-b_k l_k)}{\Gamma(a_k)} \tag{B.7}$$

$$q_V(V) = \prod_{jk} \text{Mutlinom}(v_{jk};n_{jk},w_j) = \prod_{jk} \frac{w_j!}{v_{jk}!} n_{jk}^{v_{jk}} \tag{B.8}$$

with $\sum_k n_{jk} = 1$.

Then from Eq.(B.4), (B.7) and (B.3) keeping terms dependent on l

$$(a_k-1)\log l_k - b_k l_k + \text{const} = (\sum_j \mathbb{E}_V\left[v_{jk}\right] + \alpha_k - 1)\log l_k - (\beta_k+1)l_k + \text{const}$$

where $c_k = \sum_j v_{jk}$. Form Eq.(B.5), (B.8) and (B.3) keeping terms dependent on V

$$v_{jk}\log n_{jk} - \log v_{jk}! + \text{const} = v_{jk}\mathbb{E}_l\left[\log l_k\right] + v_{jk}\log\theta_{jk} - \log v_{jk}! + \text{const}$$

so the rewrite rules for the parameters:

$$\begin{aligned} n_{jk} &= \frac{1}{z_j}\theta_{jk}\exp(\mathbb{E}_l\left[\log l_k\right]) \\ a_k &= \sum_j n_{jk}w_j + \alpha_k \\ b_k &= 1 + \beta_k \end{aligned} \tag{B.9}$$

where z_j is the normalisation constant ($\sum_k n_{jk} = 1$) so $z_j = \sum_k \theta_{jk}\exp(\mathbb{E}_l\left[\log l_k\right])$ and $\sum_j \mathbb{E}_V\left[v_{jk}\right] = \sum_j n_{jk}w_j$ (Eq.(B.8)). $\mathbb{E}_{l\sim q_l}\left[\log l_k\right] = \psi_0(a_k) - \log b_k$ where ψ_0 is digamma function (logarithmic derivation of the gamma function) and so

$$n_{jk} = \frac{1}{z_j}\theta_{jk}\exp(\psi_0(a_k) - \log b_k)$$

Now recompute model parameter θ by maximising lower bound Eq.(B.6) (with constraints $\sum_j \theta_{jk} = 1$). Keeping only term dependent on θ_{jk}:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{j,k} \mathbb{E}_{q_V(V)}\left[v_{jk}\right]\log\theta_{jk} + \text{const} \\ &= \sum_{j,k} n_{jk}w_j\log\theta_{jk} + \text{const} \end{aligned}$$

(from Eq.(B.8) $\mathbb{E}_{q_V(V)}\left[v_{jk}\right] = w_j n_{jk}$)

$$0 = \frac{\partial}{\partial\theta_{mn}}\left[\sum_{j,k} n_{jk}w_j\log\theta_{jk} + \lambda_k(1 - \sum_p \theta_{pk})\right]$$

we get

$$\theta_{mn} = \frac{n_{mn}w_m}{\lambda_n}$$

and from normalisation constraints $\lambda_n = \sum_m n_{mn}w_m$.

If we take likelihood function over all documents ($i = 1 : L$) each $w_j \to w_{j(i)}$ and $n_{jk} \to n_{jk(i)}$ then we get

$$\theta_{mn} = \frac{\sum_i n_{mn(i)}w_{m(i)}}{\lambda_n} \tag{B.10}$$

Buntine [Buntine and Jakulin 2006] even introduce prior on $\theta_{jk} \sim \mathrm{Dirichlet}(\theta_{\mathrm{jk}};\gamma,\mathrm{J}) = \mathrm{C}(\gamma_{\mathrm{j}})\prod_{\mathrm{j=1}}^{\mathrm{J}}\theta_{\mathrm{jk}}^{\gamma_{\mathrm{j}}-1}$. This is incorporated into the complete log-likelihood function $p(V,l,w,\theta|\alpha,\beta,K)$ so that lower bound $\mathbb{E}_{l,V\sim q(l,V)}\left[\log p(l,V,w,\theta|\alpha,\beta,K)\right]$ and terms dependent on θ:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i,j,k} \mathbb{E}_{q_V(V)}\left[v_{jk(i)}\right]\log\theta_{jk} + (\gamma_j - 1)\log\theta_{jk} + \text{const} \\ &= (\sum_{i,j,k} n_{jk(i)}w_{j(i)} + \gamma_j - 1)\log\theta_{jk} + \text{const} \end{aligned}$$

and by maximising with normalisation constraints:

$$\theta_{mn} \propto \sum_i n_{mn(i)}w_{m(i)} + \gamma_j \tag{B.11}$$

The lower bound Eq.(B.6)

$$\mathcal{L}(\theta) = \mathbb{E}_{l,V\sim q(l,V)} \left[\sum_k (c_k + \alpha_k - 1) \log l_k - (\beta_k + 1) l_k + \log \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} + \sum_j \left[v_{jk} \log \theta_{jk} - \log v_{jk}! \right] \right] + C$$

$$= \sum_k \mathbb{E}_l \left[\log l_k \right] (\sum_j \mathbb{E}_V \left[v_{jk} \right] + \alpha_k - 1) - (\beta_k + 1) l_k + \log \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)}$$

$$+ \sum_j \left[\mathbb{E}_V \left[v_{jk} \right] (\log n_{jk} + \log z_j - \mathbb{E}_l \left[\log l_k \right] - \mathbb{E}_V \left[\log v_{jk}! \right]) \right] + C$$

$$= \sum_k \mathbb{E}_l \left[\log l_k \right] (\alpha_k - 1) - (\beta_k + 1) l_k + \log \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} + \sum_j \left[\mathbb{E}_V \left[v_{jk} \right] (\log n_{jk} + \log z_j - \log v_{jk}!) \right] + C$$

where Eq.(B.9) for θ was used and $c_k = \sum_j v_{jk}$ and where $C = H(q_l) + H(q_V)$ from Eq.(B.6):

$$H(q_l) = -\sum_k \left\{ (a_k - 1) \mathbb{E}_l \left[\log l_k \right] - b_k \mathbb{E}_l \left[l_k \right] - \log \frac{b_k^{a_k}}{\Gamma(a_k)} \right\}$$

$$H(q_V) = -\sum_{jk} \left\{ -\mathbb{E}_V \left[\log v_{jk}! \right] + \mathbb{E}_V \left[v_{jk} \right] \log n_{jk} + \log w_j! \right\}$$

Including these terms we get

$$\mathcal{L} = \sum_k \mathbb{E}_l \left[\log l_k \right] (\alpha_k - a_k) + \sum_j w_j \log z_j + \sum_k \log \frac{\Gamma(a_k) \beta_k^{\alpha_k}}{\Gamma(\alpha_k) b_k^{a_k}} - \log \prod_j w_j! \tag{B.12}$$

where Eq.(B.9) for b_k and $\sum_k n_{jk} = 1$ was used.

After initialisation the algorithm then repeats until convergence:

1. For each document: update n_{jk} and a_k according to Eq.(B.9) (variational E step).
2. Update θ according to Eq.(B.10) or (B.11) (variational M step).
3. Compute lower bound on log-probability Eq.(B.12) and check for convergence.

C Resolution limit for the blinking QDs

Poisson random variable

This is derivation of the fisher information for Poisson distributed variable X with mean λ.

$$X \sim \text{Po}(n, \lambda) = p(n|\theta) = \frac{\lambda^n e^{-\lambda}}{n!}$$

Likelihood of the Poisson distributed variable with detection n_k in K pixels:

$$l(\theta) = \prod_{k=1}^{K} l_k = \prod_{k=1}^{K} \frac{\lambda_k^{n_k} e^{-\lambda_k}}{n_k!} \tag{C.1}$$

where $l_k(\theta) = p(n_k|\theta)$ to emphasise the dependency on the parameter θ.

Log-likelihood:

$$\mathcal{L} = \sum_k (n_k \log \lambda_k - \lambda_k - \log n_k!)$$

Fisher Information for a Poisson variable

Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}}{\partial \theta^2}\right] = \mathbb{E}\left[\left(\frac{\partial \mathcal{L}}{\partial \theta}\right)^2\right] = \mathbb{E}\left[\left(\sum_k \frac{\partial \log(l_k)}{\partial \theta}\right)^2\right] = \mathbb{E}\left[\left(\sum_k \frac{1}{l_k}\frac{\partial l_k}{\partial \theta}\right)^2\right] \tag{C.2}$$

$$\begin{aligned} I(\theta) &= \mathbb{E}\left[\left(\sum_k \frac{1}{l_k}\frac{\partial l_k}{\partial \theta}\right)\left(\sum_m \frac{1}{l_m}\frac{\partial l_m}{\partial \theta}\right)\right] \\ &= \mathbb{E}\left[\sum_k \frac{1}{l_k^2}\left(\frac{\partial l_k}{\partial \theta}\right)^2\right] + \mathbb{E}\left[\sum_k \sum_{m \neq k} \frac{1}{l_k}\frac{\partial l_k}{\partial \theta}\frac{1}{l_m}\frac{\partial l_m}{\partial \theta}\right] \end{aligned}$$

as n_k are iid then the second term can be expressed as

$$\mathbb{E}\left[\sum_k \sum_{m \neq k} \frac{1}{l_k}\frac{\partial l_k}{\partial \theta}\frac{1}{l_m}\frac{\partial l_m}{\partial \theta}\right] = \sum_k \sum_{m \neq k} \mathbb{E}_k\left[\frac{1}{l_k}\frac{\partial l_k}{\partial \theta}\right]\mathbb{E}_m\left[\frac{1}{l_m}\frac{\partial l_m}{\partial \theta}\right]$$

where

$$\mathbb{E}_k\left[f(n_k)\right] = \sum_{n_k \geq 0} p(n_k|\theta) f(n_k)$$

But

$$\mathbb{E}_k\left[\frac{1}{l_k}\frac{\partial l_k}{\partial \theta}\right] = \sum_{n_k} l_k \frac{1}{l_k}\frac{\partial l_k}{\partial \theta} = \sum_{n_k} \frac{\partial l_k}{\partial \theta} = \frac{\partial \sum_{n_k} l_k}{\partial \theta} = 0$$

as $\sum_{n_k} l_k = \sum_{n_k} p(n_k|\theta) = 1$. The Fisher Information can then be expressed

$$\begin{aligned} I(\theta) &= \mathbb{E}\left[\sum_k \frac{1}{l_k^2}\left(\frac{\partial l_k}{\partial \theta}\right)^2\right] \\ &= \sum_{k=1}^{k} \sum_{n_k \geq 0} l_k \frac{1}{l_k^2}\left(\frac{\partial l_k}{\partial \theta}\right)^2 \\ &= \sum_{k=1}^{k} \sum_{n_k \geq 0} \frac{1}{l_k}\left(\frac{\partial l_k}{\partial \theta}\right)^2 \end{aligned}$$

Derivatives of likelihood Eq.(C.1):

$$\frac{\partial l_k}{\partial \theta} = \frac{l_k(n_k - \lambda_k)}{\lambda_k}\frac{\partial \lambda_k}{\partial \theta}$$

And we get:

$$\begin{aligned} I(\theta) &= \sum_{k=1}^{k}\sum_{n_k \geq 0} \frac{l_k(n_k-\lambda_k)^2}{\lambda_k^2}\left(\frac{\partial \lambda_k}{\partial \theta}\right)^2 \\ &= \sum_{k=1}^{k} \frac{1}{\lambda_k^2}\left(\frac{\partial \lambda_k}{\partial \theta}\right)^2 \mathbb{E}_k\left[(n_k-\lambda_k)^2\right] \end{aligned}$$

for Poisson $\mathrm{var}(n) = \mathrm{mean}(n) = \lambda$ gives

$$\mathbb{E}_k\left[(n_k-\lambda_k)^2\right] = \mathrm{var}(n_k) = \lambda_k$$

and

$$I(\theta) = \sum_{k=1}^{K}\frac{1}{\lambda_k}\left(\frac{\partial \lambda_k}{\partial \theta}\right)^2 \tag{C.3}$$

This is the pixelised version (detection of the photons in K detectors - CCD camera and $\lambda_k = \int_{C_k}\lambda(x)dx$ where C_k is an area of the pixels of the detector).

Non pixelised version [Ram et al. 2006b]s

$$I(\theta) = \int \frac{1}{\lambda(x)}\left(\frac{\partial \lambda(x)}{\partial \theta}\right)^2 dx$$

Two sources separated by a distance d

These are comment on Fisher Information estimation as described in [Ram et al. 2006b].

For two sources separated by a distance d we have a mean value of the intensity:

$$\lambda = \Lambda_1 f_1 + \Lambda_2 f_2$$

where f_i and Λ_i is the response function and intensity, respectively, of the source i. For translationally invariant PSF and in-focus sources: $f_1 = q(x - \frac{d}{2})$ and $f_2 = q(x + \frac{d}{2})$

$$\lambda(d) = \Lambda_1 q(x - \frac{d}{2}) + \Lambda_2 q(x + \frac{d}{2})$$

where q is the PSF of the sources. For pixelised version (integral over pixel area C_k)

$$\lambda_k(d) = \Lambda_1 \int_{C_k} q(x - \frac{d}{2})dx + \Lambda_2 \int_{C_k} q(x + \frac{d}{2})dx$$

so we get (as described inRam et al. [2006b])

$$I(d) = \frac{1}{4}\sum_{k=1}^{K}\frac{\left(\Lambda_1 \int_{C_k}\partial_x q(x-\frac{d}{2})dx - \Lambda_2 \int_{C_k}\partial_x q(x+\frac{d}{2})dx\right)^2}{\Lambda_1 \int_{C_k} q(x-\frac{d}{2})dx + \Lambda_2 \int_{C_k} q(x+\frac{d}{2})dx} \tag{C.4}$$

Limit $d = 0$ If $\Lambda_1 = \Lambda_2$ then $I(d=0) = 0$ which means $\mathrm{var}(d=0) \to \infty$. (This does not hold for $\Lambda_1 \neq \Lambda_2$).

Limit $d \to \infty$ When sources are far apart then the mixing term in nominator in (C.4) $\Lambda_1\Lambda_2\partial_x q(x-\frac{d}{2})\partial q(d+\frac{d}{2})=0$ as the $\partial_x q(x-\frac{d}{2})$ $(q(x-\frac{d}{2}))$ and $\partial_x q(x+\frac{d}{2})$ $(q(x+\frac{d}{2})$) do not have any overlap. The (C.4) then decomposes into two individual terms (sum of Fisher Information for localisation of individual sources.)

$$
\begin{aligned}
I(d) &= \frac{1}{4}\sum_{k=1}^{K}\left[\frac{\left(\Lambda_1\int_{C_k}\partial_x q(x-\frac{d}{2})dx\right)^2}{\Lambda_1\int_{C_k}q(x-\frac{d}{2})dx}+\frac{\left(\Lambda_2\int_{C_k}\partial_x q(x+\frac{d}{2})dx\right)^2}{\Lambda_2\int_{C_k}q(x+\frac{d}{2})dx}\right] \\
&= \frac{1}{4}\sum_{k=1}^{K}\frac{\left(\int_{C_k}\partial_x q(x)dx\right)^2}{\int_{C_k}q(x)dx}\left[\Lambda_1+\Lambda_2\right]
\end{aligned}
$$

Limit $\Lambda_i = 0$ If $\Lambda_1 = 0$ or $\Lambda_2 = 0$ $I(d) \neq 0$. So the variance is finite even if one of the sources is not present.

An alternative way to derive Fisher information for two sources separated by d:

This is a suggestion how to fix the problems with limits for Fisher Information derived above. This gives infinite variance when one of the sources is no present. Also fix weird behaviour of the $I(d)$ for $d=0$.

For two sources $f_1 = q(x-c_1)$ and $f_2 = q(x-c_2)$ we have $\lambda = \Lambda_1 f_1 + \Lambda_2 f_2$. The distance between the two sources is $d = c_1 - c_2$. This is a linear combination $\boldsymbol{a}^T\cdot\boldsymbol{c}$ of the variable $\boldsymbol{c}=(c_1,c_2)$ where $\boldsymbol{a}=(1,-1)$. The variance of d is given by

$$
\operatorname{var}(d) = \operatorname{var}(\boldsymbol{a}^T\cdot\boldsymbol{c}) = \boldsymbol{a}^T\cdot\boldsymbol{Q}\cdot\boldsymbol{a} = Q_{11}+Q_{22}-2Q_{12}
$$

where $\boldsymbol{Q}$ is a covariance matrix $\boldsymbol{Q}=\boldsymbol{I}^{-1}(\theta)$ and $\boldsymbol{I}(\theta)$ is the Fisher information matrix (symmetric $I_{12}=I_{21}$)

$$
\boldsymbol{I}(\theta) = \begin{pmatrix} I_{11} & I_{12} \\ I_{12} & I_{22} \end{pmatrix}
$$

given by generalisation of Eq.(C.3)

$$
I_{ij}(\theta) = \sum_{k=1}^{K}\frac{1}{\lambda_k}\frac{\partial\lambda_k}{\partial\theta_i}\frac{\partial\lambda_k}{\partial\theta_j}
$$

The covariance matrix $\boldsymbol{Q}$ is then

$$
\boldsymbol{Q} = \boldsymbol{I}^{-1}(\theta) = \frac{1}{I_{11}I_{12}-I_{12}^2}\begin{pmatrix} I_{22} & -I_{12} \\ -I_{12} & I_{11} \end{pmatrix}
$$

and the variance of $d = c_1 - c_2$

$$
\operatorname{var}(d) = (1,-1)^T\cdot\boldsymbol{Q}\cdot(1,-1) = \frac{I_{11}+I_{22}+2I_{12}}{I_{11}I_{12}-I_{12}^2} \tag{C.5}
$$

The individual terms of the Fisher Information matrix

$$
I_{11} = \sum_{k=1}^{K}\frac{1}{\lambda_k}\left(\frac{\partial\lambda_k}{\partial c_1}\right)^2 = \sum_{k=1}^{K}\frac{\left(\Lambda_1 q_k'(c_1)\right)^2}{\Lambda_1 q_k(c_1)+\Lambda_2 q_k(c_2)}
$$

where

$$q_k(c) = \int_{C_k} q(x-c)dx \quad , \ (q_k(0) = q_k)$$
$$q'_k(c) = \int_{C_k} \frac{\partial q(x-c)}{\partial x}dx, \ (q'_k(0) = q'_k)$$

If this keeps translational invariance (non-pixelised version does as $\int_{\mathbb{R}} g(x+c)dx = \int_{\mathbb{R}} g(x)dx$) then

$$I_{11} = \sum_{k=1}^{K} \frac{(\Lambda_1 q'_k)^2}{\Lambda_1 q_k + \Lambda_2 q_k(-d)}$$

where $d = c_1 - c_2$ and

$$I_{22} = \sum_{k=1}^{K} \frac{(\Lambda_2 q'_k)^2}{\Lambda_2 q_k + \Lambda_1 q_k(d)}$$

For symmetrical PSF $q(x-d) = q(x+d)$ we have

$$I_{ii} = \sum_{k=1}^{K} \frac{(\Lambda_i q'_k)^2}{\Lambda_i q_k + \Lambda_j q_k(d)} \tag{C.6}$$

And the cross term $(i \neq j)$

$$I_{ij} = \sum_{k=1}^{K} \frac{\Lambda_i \Lambda_j q'_k q'_k(d)}{\Lambda_i q_k + \Lambda_j q_k(d)}$$

Limit $d \to 0$ For $d = 0$ we have

$$I_{ii} = \frac{\Lambda_i^2}{\Lambda_i + \Lambda_j} S(0)$$
$$I_{ij} = \frac{\Lambda_i \Lambda_j}{\Lambda_i + \Lambda_j} S(0)$$

where $S(d) = \sum_{k=1}^{K} \frac{(q'_k)^2}{q_k + q_k(d)}$.

Numerator p in Eq.(C.5)

$$p = I_{11} + I_{22} + 2I_{12} = \frac{S(0)}{\Lambda_1 + \Lambda_2}(\Lambda_1^2 + \Lambda_2^2 + 2\Lambda_1\Lambda_2) = \frac{S(0)}{\Lambda_1 + \Lambda_2}(\Lambda_1 + \Lambda_2)^2$$

is non-zero for any Λ_1, Λ_2.

The denominator in Eq.(C.5)

$$r = \det[\boldsymbol{I}(\theta)] = I_{11}I_{22} - I_{12}^2 = \frac{S^2(0)}{(\Lambda_1 + \Lambda_2)^2}\left(\Lambda_1^2\Lambda_2^2 - (\Lambda_1\Lambda_2)^2\right) \ \equiv \ 0 \text{ for any } \Lambda_i$$

$\boldsymbol{I}(\theta)$ is therefore a singular matrix for $d = 0$ and inversion $\boldsymbol{I}^{-1}(\theta)$ does not exist.

However, for the limit $d \to 0$ and $p \neq 0$, $r \to 0$ and $\text{var}(d \to 0) = \frac{p}{r} \to \infty$.

Limit $d \to \infty$ The cross term $I_{ij} = 0,\ i \neq j$ and we get f

$$\mathrm{var}(d) = \frac{1}{I_{11}} + \frac{1}{I_{22}}$$

and

$$I_{ii} = \sum_{k=1}^{K} \frac{(\Lambda_i q_k')^2}{\Lambda_i q_k + \Lambda_j q_k(d)} = \Lambda_i \sum_{k=1}^{K} \frac{(q_k')^2}{q_k} = 2\Lambda_i S(0)$$

as the PSF $q(x)$ (and also $q'(x)$) have a finite support, if d is big, $q(x-d)$ is outside the support of the $q'(x)$. They have no overlap so it doesn't have any effect in the denominator.

For non-pixelised version, $\Lambda_1 = \Lambda_2 = \Lambda$ and for Gaussian approximation of the PSF ($q(x-a) \propto \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$ (with $\sigma = \frac{\sqrt{2}}{2\pi}\frac{\lambda}{NA}$ [Zhang et al. 2007]) we have $q'(x) = \frac{1}{\sigma^2} x q(x)$ and and for $\Lambda_1 = \Lambda_2 = \Lambda$ and for Gaussian approximation of the PSF $S(0) = \frac{1}{2\sigma^2}$:

$$I(d \to \infty) = \frac{\Lambda}{\sigma^2}$$
$$\mathrm{var}(d \to \infty) = \frac{\sigma^2}{\Lambda}$$

Limit $\Lambda_i = 0,\ \Lambda_j \neq 0$ then $I_{ii} \equiv 0$ and $I_{ij} \equiv 0$ and so $\det(\boldsymbol{I}(\theta)) \equiv 0$, and matrix is singular. In the limit $\Lambda_i \to 0$ the variance (C.5) $\mathrm{var}(d) \to \infty$.

Time distribution of the intensities (blinking)

For likelihood dependent on parameter Λ_t (T different time slices)

$$l_T(d, \Lambda) = \prod_{k=1}^{K} \prod_{t=1}^{T} p(n_k | d, \Lambda_t) p(\Lambda_t)$$

$$\mathcal{L}_T(d, \Lambda) = \sum_{k=1}^{K} \sum_{t=1}^{T} \left[\log\left(l_k(d, \Lambda_t)\right) + \log\left(p(\Lambda_t)\right)\right]$$

as $p(\Lambda)$ is not dependent on d then

$$\frac{\partial^2 \mathcal{L}_T(d, \Lambda)}{\partial d^2} = \sum_{t=1}^{T} \frac{\partial^2 \mathcal{L}(d, \Lambda_t)}{\partial d^2}$$

but in the expectation equation Eq.(C.2) the time dependence appears as

$$I_T(\theta) = -\mathbb{E}_T\left[\sum_{t=1}^{T} \frac{\partial^2 \mathcal{L}(d, \Lambda_t)}{\partial d^2}\right] = \sum_{t=1}^{T} -\mathbb{E}_T\left[\frac{\partial^2 \mathcal{L}(d, \Lambda_t)}{\partial d^2}\right] = \sum_{t=1}^{T} \mathbb{E}_T\left[\left(\frac{\partial \mathcal{L}(d, \Lambda_t)}{\partial d}\right)^2\right]$$
$$= \sum_{t=1}^{T} \int_{\Lambda_t} p(\Lambda_t) I(\theta) d\Lambda_t = \sum_{t,k} \int_{\Lambda_t} p(\Lambda_t) \frac{1}{\lambda_k(\Lambda_t)} \left(\frac{\partial \lambda_k(\Lambda_t)}{\partial d}\right)^2 d\Lambda_t$$

Time distribution of the intensities - integrating out Λ

$$l_k(d) = \int_\Lambda l_k(d,\Lambda)d\Lambda = \int_\Lambda p(n_k|d,\Lambda)p(\Lambda)d\Lambda$$

for four state model of two sources: $\{(\Lambda_1,0),(0,\Lambda_2),(\Lambda_1,\Lambda_2),(0,0)\}$: $\lambda^1 = \Lambda_1 q(x-c_1)$, $\lambda^2 = \Lambda_2 q(x-c_2)$, $\lambda^3 = +\Lambda_1 q(x-c_1) + \Lambda_2 q(x-c_2)$, $\lambda^4 = 0$ with uniform distribution over these states

$$l_k(\theta) = \frac{1}{4}\sum_{i=1}^{4} \mathrm{Po}(\lambda_k^i)$$

derivatives

$$\frac{\partial l_k}{\partial c_p} = \frac{1}{4}\sum_i \frac{\partial \mathrm{Po}(\lambda_k^i)}{\partial c_p} = \frac{1}{4}\sum_i \left(\mathrm{Po}(\lambda_k^i)\frac{(n_k-\lambda_k^i)}{\lambda_k^i}\frac{\partial \lambda_k^i}{\partial c_p}\right)$$

The Fisher information matrix diagonal entries:

$$\begin{aligned}
I_{pp}(\theta) &= \mathbb{E}\left[\left(\sum_{k=1}^{N}\frac{1}{l_k}\frac{\partial l_k}{\partial c_p}\right)^2\right] \\
&= \mathbb{E}\left[\left\{\sum_{k=1}^{N}\left(\frac{1}{\sum_{j=1}^4 \mathrm{Po}(\lambda_k^j)}\frac{\partial \sum_{i=1}^4 \mathrm{Po}(\lambda_k^i)}{\partial c_p}\right)\right\}\left\{\sum_{l=1}^{N}\left(\frac{1}{\sum_{j=1}^4 \mathrm{Po}(\lambda_l^j)}\frac{\partial \sum_{i=1}^4 \mathrm{Po}(\lambda_l^i)}{\partial c_p}\right)\right\}\right] \\
&= \sum_{k=1}^{N}\mathbb{E}_k\left[\frac{\left(\sum_{i=1}^4 \frac{\partial \mathrm{Po}(\lambda_k^i)}{\partial c_p}\right)^2}{\left(\sum_{j=1}^4 \mathrm{Po}(\lambda_k^j)\right)^2}\right] \qquad \text{(C.7)}
\end{aligned}$$

as the cross terms $(k,\, l)$ in the sum (2nd row) are zeros:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\sum_{i=1}^4 \frac{\partial \mathrm{Po}(\lambda_k^i)}{\partial c_p}}{\sum_{j=1}^4 \mathrm{Po}(\lambda_k^j)}\right)\left(\frac{\sum_{i=1}^4 \frac{\partial \mathrm{Po}(\lambda_l^i)}{\partial c_p}}{\sum_{j=1}^4 \mathrm{Po}(\lambda_l^j)}\right)\right] &= \mathbb{E}_k\left[\frac{\sum_{i=1}^4 \frac{\partial \mathrm{Po}(\lambda_k^i)}{\partial c_p}}{\sum_{j=1}^4 \mathrm{Po}(\lambda_k^j)}\right]\mathbb{E}_l\left[\frac{\sum_{i=1}^4 \frac{\partial \mathrm{Po}(\lambda_l^i)}{\partial c_p}}{\sum_{j=1}^4 \mathrm{Po}(\lambda_l^j)}\right] \\
&= \sum_{i=1}^4 \frac{\partial}{\partial c_p}\left(\sum_{n_k\geq 0}\mathrm{Po}(\lambda_k^i)\right)\sum_{i=1}^4 \frac{\partial}{\partial c_p}\left(\sum_{n_k\geq 0}\mathrm{Po}(\lambda_l^i)\right) \\
&= 0
\end{aligned}$$

Expressing the derivatives and the expectation from Eq.(C.7):

$$\begin{aligned}
I_{pp}(\theta) &= \sum_{k=1}^{N}\mathbb{E}_k\left[\left\{\frac{\sum_{i=1}^4\left(\mathrm{Po}(n_k;\lambda_k^i)\frac{(n_k-\lambda_k^i)}{\lambda_k^i}\frac{\partial \lambda_k^i}{\partial c_p}\right)}{\sum_{j=1}^4 \mathrm{Po}(n_k;\lambda_k^j)}\right\}^2\right] \\
&= \frac{1}{4}\sum_{k=1}^{N}\sum_{n_k\geq 0}\frac{\left\{\sum_{i=1}^4\left(\mathrm{Po}(n_k;\lambda_k^i)\frac{(n_k-\lambda_k^i)}{\lambda_k^i}\frac{\partial \lambda_k^i}{\partial c_p}\right)\right\}^2}{\sum_{j=1}^4 \mathrm{Po}(n_k;\lambda_k^j)}
\end{aligned}$$

For the four states model we have $\lambda^3(c_1,c_2) = \lambda^1(c_1) + \lambda^2(c_2)$ and so $\frac{\partial \lambda^3}{\partial c_p} = \frac{\partial \lambda^p}{\partial c_p}$ and $\frac{\partial \lambda^j}{\partial c_p} = 0,\ i \neq j$ for $p = \{1,2\},\ j = \{1,2,4\}$; so

$$I_{pp}(\theta) = \sum_{k=1}^{N}\left(\frac{\partial \lambda_k^p}{\partial c_p}\right)^2 \mathbb{E}_k\left[\left\{\frac{\sum_{i=\{p,3\}}\left(\mathrm{Po}(n_k;\lambda_k^i)\frac{(n_k-\lambda_k^i)}{\lambda_k^i}\right)}{\sum_{j=1}^4 \mathrm{Po}(n_k;\lambda_k^j)}\right\}^2\right]$$

The Fisher information matrix off-diagonal entries:

$$\begin{aligned} I_{pq}(\theta) &= \sum_{k=1}^{N} \mathbb{E}_k \left[\frac{\left(\sum_{i=1}^4 \frac{\partial \mathrm{Po}(\lambda_k^i)}{\partial c_p}\right)\left(\sum_{l=1}^4 \frac{\partial \mathrm{Po}(\lambda_k^l)}{\partial c_q}\right)}{\left(\sum_{j=1}^4 \mathrm{Po}(\lambda_k^j)\right)^2} \right] \\ &= \sum_{k=1}^{N} \left(\frac{\partial \lambda_k^p}{\partial c_p}\right)\left(\frac{\partial \lambda_k^q}{\partial c_q}\right) \mathbb{E}_k \left[\frac{\left(\sum_{i=\{p,3\}} \mathrm{Po}(n_k;\lambda_k^i)\frac{(n_k-\lambda_k^i)}{\lambda_k^i}\right)\left(\sum_{i=\{q,3\}} \mathrm{Po}(n_k;\lambda_k^i)\frac{(n_k-\lambda_k^i)}{\lambda_k^i}\right)}{\left(\sum_{j=1}^4 \mathrm{Po}(n_k;\lambda_k^j)\right)^2} \right] \end{aligned} \tag{C.8}$$

Limit $d \to 0$ When $c^1 = c^2$ then $\lambda^1 = \lambda^2$ and $\frac{\partial \mathrm{Po}(\lambda^1)}{\partial c^1} = \frac{\partial \mathrm{Po}(\lambda^2)}{\partial c^2}$. Then all entries in I_{pq} are equal and the matrix is singular. For the limit $d \to 0$ the determianat $\det(\boldsymbol{I}) \to 0$ and the variance $\mathrm{var}(d) \to \infty$.

Limit $d \to \infty$ Sources are far apart and λ^1 and λ^2 do not have a common overlap. For k' where $\lambda_{k'}^1 > 0$, $\lambda_{k'}^2 \equiv 0$ and $\mathrm{Po}(n_{k'}, \lambda_{k'}^3) = \mathrm{Po}(n_{k'}, \lambda_{k'}^1) + \mathrm{Po}(n_{k'}, \lambda_{k'}^2 = 0) = \mathrm{Po}(n_{k'}, \lambda_{k'}^1) + 1$. Also $\frac{\partial \lambda^p}{\partial c_q} = 0$, $p \neq q$. From Eq.(C.7) the diagonal elements

$$\begin{aligned} I_{pp} &= \sum_{k=1}^{N} \mathbb{E}_k \left[\frac{\left(2\frac{\partial \mathrm{Po}(\lambda_k^p)}{\partial c_p}\right)^2}{(2\mathrm{Po}(\lambda_k^p) + 2\mathrm{Po}(\lambda_k^q))^2} \right] \\ &= \sum_{k=1}^{N} \mathbb{E}_k \left[\frac{\left(\mathrm{Po}(\lambda_k^p)\frac{(n_k-\lambda_k^p)^2}{\lambda_k^p}\frac{\partial \lambda_k^p}{\partial c_p}\right)^2}{(\mathrm{Po}(\lambda_k^p) + 1)^2} \right] \\ &= \sum_{k=1}^{N} \left(\frac{1}{\lambda_k^p}\frac{\partial \lambda_k^p}{\partial c_p}\right)^2 \mathbb{E}_k \left[(n_k - \lambda_k^p)^2 \left(\frac{\mathrm{Po}(\lambda_k^p)}{\mathrm{Po}(\lambda_k^p) + 1}\right)^2 \right] \end{aligned}$$

For large λ_k^p the second term in the expectation is approximately one: $\frac{\mathrm{Po}(\lambda_k^p)}{\mathrm{Po}(\lambda_k^p)+1} = 1 - \frac{1}{1+\mathrm{Po}(\lambda_k^p)} \approx 1$

$$I_{pp} \approx \frac{1}{2}\sum_{k=1}^{N} \frac{1}{\lambda_k^p}\left(\frac{\partial \lambda_k^p}{\partial c_p}\right)^2 \tag{C.9}$$

which is the Eq.(C.3) (up to the factor 2). As the the term is upper bounded by one: $\frac{\mathrm{Po}(\lambda_k^p)}{\mathrm{Po}(\lambda_k^p)+1} = 1 - \frac{1}{1+\mathrm{Po}(\lambda_k^p)} < 1$ the terms I_{pp} will be slightly smaller then the approximation Eq.(C.9):

$$I_{pp} = \frac{1}{2}\sum_{k=1}^{N} \frac{1}{\lambda_k^p}\left(\frac{\partial \lambda_k^p}{\partial c_p}\right)^2 - \epsilon \tag{C.10}$$

The off-diagonal entries:

$$I_{pq} = 0$$

as

$$\frac{\partial \mathrm{Po}(\lambda^p)}{\partial c_p}\frac{\partial \mathrm{Po}(\lambda^q)}{\partial c_q} = 0$$

because $\lambda^p(x)$ and $\lambda^q(x)$ do not have a common support. and the

Therefore

$$\begin{aligned}\mathrm{var}(d) &= \frac{1}{I_{11}} + \frac{1}{I_{22}} \\ &= 2\left(\frac{1}{I_{11}^{\mathrm{static}} - \epsilon} + \frac{1}{I_{22}^{\mathrm{static}} - \epsilon}\right) \\ &> 2\mathrm{var}(d^{\mathrm{static}}) \qquad \text{(C.11)}\end{aligned}$$

where I^{static} and $\mathrm{var}^{\mathrm{static}}$ correspond to the Fisher information matrix Eq.(C.3) and the variance Eq.(C.5) of the static case. The factor of 2 stems from the fact that the total number of photons is double in the static case compared to the blinking model.