Bayesian inference and mathematical imaging.
Part I: Bayesian analysis and decision theory.

Dr. Marcelo Pereyra
http://www.macs.hw.ac.uk/~mp71/

Maxwell Institute for Mathematical Sciences, Heriot-Watt University

January 2019, CIRM, Marseille.

Outline

1 Imaging inverse problems

2 Bayesian modelling

3 Bayesian inference

4 Conclusion

Imaging inverse problems

- We are interested in an unknown image $x \in \mathbb{R}^d$.
- We measure y, related to x by some mathematical model.
- For example, in many imaging problems

$$y = Ax + w,$$

for some linear operator A, and additive noise w.

Regularisation

- If $A^\top A$ is ill-conditioned or rank deficient, then y does not allow reliably recovering x without additional information.
- In words, there is significant uncertainty about x given y.
- To reduce uncertainty and deliver meaningful results we need to regularise the estimation problem $y \to x$ to make it well-posed.

Regularisation

- For example, given some penalty function $h : \mathbb{R}^d \to [0, \infty)$ promoting expected properties of x, we could construct the estimator

$$\hat{x} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \|y - Ax\|_2^2 + \theta h(x) \,, \tag{1}$$

  that combines the data fidelity term $\|y - Ax\|_2^2$ and the penalty h, where the "regularisation" parameter $\theta > 0$ controls the balance.

- When h is convex and l.s.c., $\hat{x}$ can be computed efficiently by (proximal) convex optimisation (Chambolle and Pock, 2016).

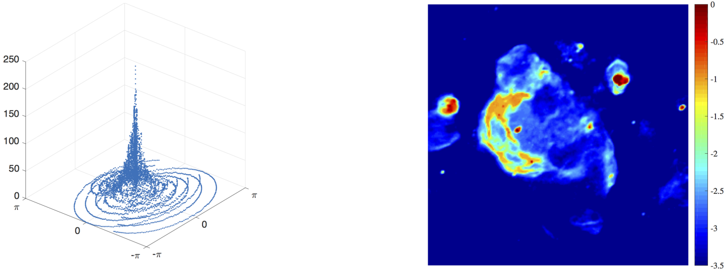
- Other data fidelity terms could be considered too (e.g., $\|y - Ax\|_1$).

Illustrative example: astronomical image reconstruction

Recover $x \in \mathbb{R}^d$ from low-dimensional observation $y = M\mathcal{F}x + w$, where $\mathcal{F}$ is the cont. Fourier transform, $M \in \mathbb{C}^{m\times d}$ is a mask. We use the estimator

$$\hat{x} = \underset{x\in\mathbb{R}_+}{\operatorname{argmin}} \|y - M\mathcal{F}x\|_2^2 + \theta\|\Psi x\|_1 \,, \tag{2}$$

where Ψ is a specialised wavelet dictionary and $\theta > 0$ is some parameter.



$|y|$ $\hat{x}$

Figure : Radio-interferometric image reconstruction of the `W28 supernova`.

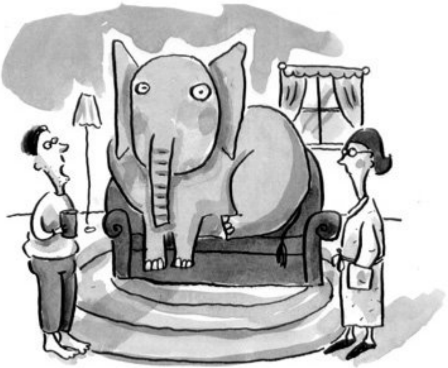
Illustrative example: astronomical image reconstruction

- Modern convex optimisation can compute $\hat{x}$ very efficiently...
- With parallelised and distributed algorithms...
- With theoretical convergence guarantees..
- And GPU implementations...
- Also non-convex extensions...

Also, if we had abundant training data (we do not) we could learn a neural network to recover x from y. Or alternatively learn $y \rightarrow \hat{x}$ for efficiency.

So the problem is quite solved, right?

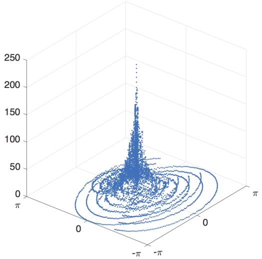
Not really...



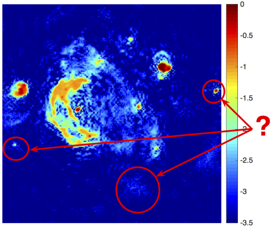
What elephant?

Tomassi

Elephant 1: what is the uncertainty about $\hat{x}$?



$|y|$



$\hat{x}_{MAP}$

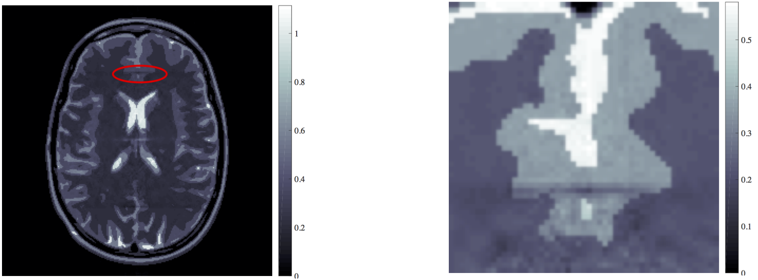
How confident are we about all these structures in the image?

What is the error in their intensity, position, spectral properties?

Using $\hat{x}_{MAP}$ to derive physical quantities? what error bars should we put...

Illustrative example: magnetic resonance imaging

We use very similar techniques to produce magnetic resonance images...



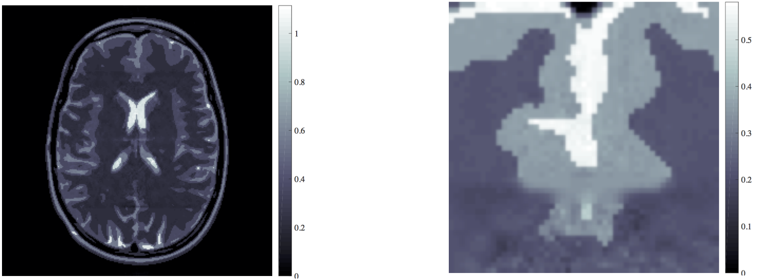
$\hat{x}$ $\hat{x}$ (zoom)

Figure : Magnetic resonance imaging of brain lession.

How can we quantify our uncertainty about the brain lesion in the image?

Illustrative example: magnetic resonance imaging

What about this other solution to the problem, with no lesion?



$\hat{x}'$

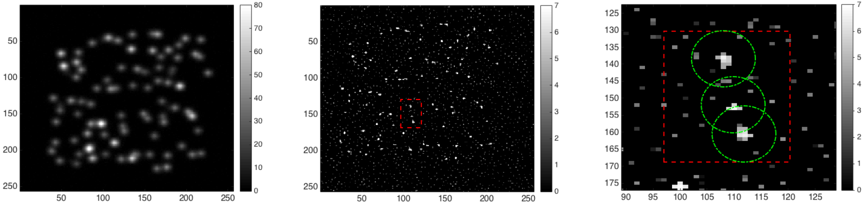
$\hat{x}'$ (zoom)

Figure : Magnetic resonance imaging of brain lesion.

Do we have any arguments to reject this solution?

Elephant 1: what is the uncertainty about $\hat{x}$?

Another example related to sparse super-resolution in live-cell microscopy



y $\hat{x}_{MAP}$ $\hat{x}_{MAP}$ (zoom)

Figure : Live-cell microscopy data (Zhu et al., 2012).

The image is sharpened to enhance molecule position measurements, but what is the precision of the procedure?

Elephant 2: multiple and partially unknown models

We usually have several alternative models/cost functions to recover x

$$\begin{aligned}\hat{x}_1 &= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \|y - A_1 x\|_2^2 + \theta_1 h_1(x), \\ \hat{x}_2 &= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \|y - A_2 x\|_2^2 + \theta_1 h_2(x),\end{aligned} \tag{3}$$

How can we compare them without ground truth available?

Can we use several models simultaneously?

Elephant 2: multiple and partially unknown models

Some of the model parameters might also be unknown; e.g., $\theta \in \mathbb{R}^+$ in

$$\hat{x}_1 = \operatorname*{argmin}_{x \in \mathbb{R}^d} \|y - A_1 x\|_2^2 + \theta h_1(x). \tag{4}$$

Then θ parametrises a class of models for $y \to x$.

How can we select θ without using ground truth?

Could we use all models simultaneously?

Other elephants in the neighbourhood...

Why do we formulate solutions to imaging problems as penalised data-fitting problems? Are there other relevant formulations?

Suppose we had a specific meaningful way of measuring estimation error, reflecting specific aspects of our problem or the application considered. What would be the optimal estimator then?

Other elephants in the neighbourhood...

Should solutions to imaging problems always be images? even when their purpose is to inform decision making, scientific conclusions?

What other mathematical objects could we construct to represent solutions to imaging problems? e.g., could curves (videos) be interesting solutions?

Outline

1. Imaging inverse problems
2. Bayesian modelling
3. Bayesian inference
4. Conclusion

Introduction

Bayesian statistics is a mathematical framework for deriving inferences about x, from some observed data y and prior knowledge available.

Adopting a subjective probability, we represent x as a random quantity, and model our knowledge about x by using probability distributions.

To derive inferences about x from y we postulate a joint statistical model $p(x, y)$; typically specified via the decomposition $p(x, y) = p(y|x)p(x)$.

Introduction

The decomposition $p(x,y) = p(y|x)p(x)$ has two key ingredients:

The likelihood function: the conditional distribution $p(y|x)$ that models the data observation process (forward model).

The prior function: the marginal distribution $p(x) = \int p(x,y)\mathrm{d}x$ that models our knowledge about x "before observing y".

For example, for $y = Ax + w$, with $w \sim \mathcal{N}(0, \sigma^2\mathbb{I})$, we have

$$y \sim \mathcal{N}(Ax, \sigma^2\mathbb{I}),$$

or equivalently

$$p(y|x) \propto \exp\{-\|y - Ax\|_2^2/2\sigma^2\}.$$

Prior distribution

The prior distribution is often of the form:

$$p(x) \propto \mathrm{e}^{-\theta h(x)} 1_{\Omega}(x)\,,$$

for some $h : \mathbb{R}^d \to \mathbb{R}^m$, $\theta \in \mathbb{R}^m$, and constraint set Ω.

This prior is essentially specifying the expectation

$$\mathrm{E}(h) = \int_{\Omega} h(x) p(x) \mathrm{d}x = \nabla_{\theta} \log C(\theta)$$

where

$$C(\theta) = \int_{\Omega} \mathrm{e}^{-\theta h(x)} \mathrm{d}x\,.$$

Prior distribution

For example, priors of the form

$$p(x) \propto \mathrm{e}^{-\theta\|\Psi x\|_{\dagger}},$$

for some dictionary $\Psi \in \mathbb{R}^{d\times p}$ and norm $\|\cdot\|_{\dagger}$, are encoding

$$\mathrm{E}(\|\Psi x\|_{\dagger}) = \frac{d}{\theta}.$$

See Pereyra et al. (2015) for more details and other examples.

Posterior distribution

Given an observation y, we naturally base our inferences on the posterior distribution $p(x|y)$.

We derive $p(x|y)$ from the likelihood $p(y|x)$ and the prior $p(x)$ by using

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

where $p(y) = \int p(y|x)p(x)\mathrm{d}x$ measures model-fit-to-data.

The conditional $p(x|y)$ models our knowledge about x after observing y.

Observe that $p(x)$ may significantly regularise the estimation problem and address identifiability issues in $p(y|x)$ (e.g., when A is rank deficient).

Maximum-a-posteriori (MAP) estimation

The predominant Bayesian approach in imaging to extract a solution from $p(x|y)$ is MAP estimation

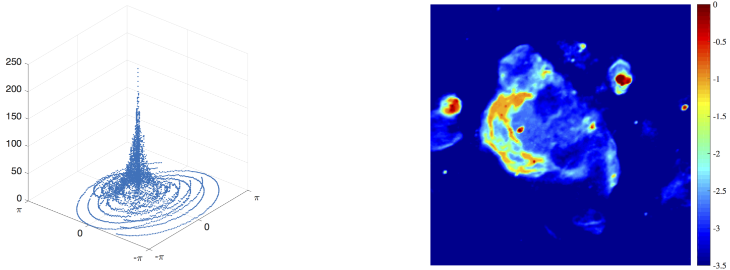
$$\begin{aligned} \hat{x}_{MAP} &= \operatorname*{argmax}_{x\in\mathbb{R}^d} p(x|y), \\ &= \operatorname*{argmin}_{x\in\mathbb{R}^d} -\log p(y|x) - \log p(x) + \log p(y). \end{aligned} \tag{5}$$

When $p(x|y)$ is log-concave, then $\hat{x}_{MAP}$ is a convex optimisation problem and can be efficiently solved (Chambolle and Pock, 2016).

Illustrative example: astronomical image reconstruction

Recover $x \in \mathbb{R}^d$ from low-dimensional degraded observation $y = M\mathcal{F}x + w$, where $\mathcal{F}$ is the continuous Fourier transform, $M \in \mathbb{C}^{m \times d}$ is a measurement operator and w is Gaussian noise. We use the model

$$p(x|y) \propto \exp\left(-\|y - M\mathcal{F}x\|^2/2\sigma^2 - \theta\|\Psi x\|_1\right)\mathbf{1}_{\mathbb{R}^n_+}(x). \tag{6}$$



y $\hat{x}_{MAP}$

Figure : Radio-interferometric image reconstruction of the `W28 supernova`.

Outline

1. Imaging inverse problems

2. Bayesian modelling

3. Bayesian inference

4. Conclusion

Bayesian decision theory

Given the following elements defining a decision problem:

1. Decision space Δ
2. Loss function $L(\delta, x) : \Delta \times \mathbb{R}^d \to \mathbb{R}$ quantifying the loss (or profit) related to taking action $\delta \in \Delta$ when the truth is $x \in \mathbb{R}^d$.
3. A model $p(x)$ representing probabilities for x.

What is the optimal decision $\delta^* \in \Delta$ when x is unknown?

Bayesian decision theory

Given the following elements defining a decision problem:

1. Decision space Δ
2. Loss function $L(\delta, x) : \Delta \times \mathbb{R}^d \to \mathbb{R}$ quantifying the loss (or profit) related to taking action $\delta \in \Delta$ when the truth is $x \in \mathbb{R}^d$.
3. A probability model $p(x)$ representing knowledge about x.

According to Bayesian decision theory (Robert, 2001), the optimal decision under uncertainty is

$$\delta^* = \underset{\delta \in \Delta}{\operatorname{argmin}} \, \mathrm{E}\{L(\delta, x)|y\} = \underset{\delta \in \Delta}{\operatorname{argmin}} \int L(\delta, x) p(x) \mathrm{d}x.$$

Bayesian point estimators

Bayesian point estimators arise from the decision "what point $\hat{x} \in \mathbb{R}^d$ summarises $x|y$ best?". The optimal decision under uncertainty is

$$\hat{x}_L = \operatorname*{argmin}_{u \in \mathbb{R}^d} \mathrm{E}\{L(u,x)|y\} = \operatorname*{argmin}_{u \in \mathbb{R}^d} \int L(u,x)p(x|y)\mathrm{d}x$$

where the loss $L(u,x)$ measures the "dissimilarity" between u and x.

General desiderata:

1. $L(u,x) \geq 0, \; \forall \, u, x \in \mathbb{R}^d$,
2. $L(u,x) = 0 \iff u = x$,
3. L strictly convex w.r.t. its first argument (for estimator uniqueness).

Bayesian point estimators - MMSE estimation

Example: the squared Euclidean distance $L(u,x) = \|u - x\|^2$ defines the so-called minimum mean squared error estimator.

$$\hat{x}_{MMSE} = \underset{u \in \mathbb{R}^d}{\operatorname{argmin}} \int \|u - x\|_2^2 \, p(x|y) \mathrm{d}x \, .$$

By differentiating w.r.t. to u and equating to zero we obtain that

$$\int (\hat{x}_{MMSE} - x) p(x|y) \mathrm{d}x = 0 \implies \hat{x}_{MMSE} \int p(x|y) \mathrm{d}x = \int x p(x|y) \mathrm{d}x \, ,$$

hence $\hat{x}_{MMSE} = \mathrm{E}\{x|y\}$ (recall that $\int p(x|y) \mathrm{d}x = 1$).

Bayesian point estimators - MAP estimation

The predominant Bayesian approach in imaging is MAP estimation

$$\begin{aligned}\hat{x}_{MAP} &= \underset{x\in\mathbb{R}^d}{\operatorname{argmax}}\, p(x|y), \\ &= \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} - \log p(y|x) - \log p(x).\end{aligned} \tag{7}$$

In many problems $\hat{x}_{MAP}$ can be computed efficiently by optimisation (Chambolle and Pock, 2016).

What is the loss function L associated with this estimator?

Posterior credible regions

Where does the posterior probability mass of x lie?

A set C_α is a posterior credible region of confidence level $(1-\alpha)\%$ if

$$\mathrm{P}\left[x \in C_\alpha | y\right] = 1 - \alpha.$$

For any $\alpha \in (0,1)$ there are infinitely many regions of the parameter space that verify this property.

The *highest posterior density* (HPD) region is decision-theoretically optimal in a compactness sense

$$C_\alpha^* = \{x : \phi(x) \leq \gamma_\alpha\}$$

with $\gamma_\alpha \in \mathbb{R}$ chosen such that $\int_{C_\alpha^*} p(x|y)\mathrm{d}x = 1 - \alpha$ holds.

Hypothesis testing

Hypothesis test split the solution space in two meaningful regions, e.g.,

$$\mathrm{H}_0 : x \in \mathcal{S}$$
$$\mathrm{H}_1 : x \notin \mathcal{S}$$

where $\mathcal{S} \subset \mathbb{R}^d$ contains all solutions with some characteristic of interest.

We can then assess the degree of support for H_0 vs. H_1 by computing

$$P(H_0|y) = \int_{\mathcal{S}} p(x|y)\mathrm{d}x, \quad P(H_1|y) = 1 - P(H_0|y)\,.$$

We can also reject H_0 in favour of H_1 with significance $\alpha \in [0,1]$ if

$$\mathsf{P}(\mathrm{H}_0|y) \leq \alpha.$$

Bayesian model selection

The Bayesian framework provides theory for comparing models objectively.

Given K alternative models $\{\mathcal{M}_j\}_{j=1}^K$ with posterior densities

$$\mathcal{M}_j : \quad p_j(x|y) = p_j(y|x)p_j(x))/p_j(y)\,,$$

we compute the (marginal) posterior probability of each model, i.e.,

$$p(\mathcal{M}_j|y) \propto p(y|\mathcal{M}_j)p(\mathcal{M}_j) \tag{8}$$

where $p(y|\mathcal{M}_j) \triangleq p_j(y) = \int p_j(y|x)p_j(x)\mathrm{d}x$ measures model-fit-to-data.

We then select for our inferences the "best" model, i.e.,

$$\mathcal{M}^* = \underset{j\in\{1,\dots,K\}}{\operatorname{argmax}}\, p(\mathcal{M}_j|y).$$

Bayesian model calibration

Alternatively, given a continuous class of models $\{\mathcal{M}_\theta, \theta \in \theta\}$ with

$$\mathcal{M}_{\boldsymbol{\theta}} : \quad p(x|y,\theta) = \frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)},$$

we compute the (marginal) posterior

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y) \tag{9}$$

where $p(y|\theta) = \int p(y|x,\theta)p(x|\theta)\mathrm{d}x$ measures model-fit-to-data.

We then calibrate our model with the "best" value of θ, i.e.,

$$\hat{\theta}_{MAP} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, p(\theta|y).$$

Bayesian model averaging

We can also use all models simultaneously!

Given K alternative models $\{\mathcal{M}_j\}_{j=1}^K$ with posterior densities

$$\mathcal{M}_j : \quad p_j(x|y) = p_j(y|x)p_j(x))/p_j(y)\,,$$

we marginalise w.r.t. the model selector j, i.e.,

$$p(x|y) = \sum_{j=1}^{K} p(x, \mathcal{M}_j|y) = \sum_{j=1}^{K} p(x|y, \mathcal{M}_j)p(\mathcal{M}_j|y) \tag{10}$$

where the posterior probabilities $p(\mathcal{M}_j|y)$ control the relative importance of each model.

Bayesian model calibration

Similarly, given a continuous class of models $\{\mathcal{M}_\theta, \theta \in \Theta\}$ with

$$\mathcal{M}_{\boldsymbol{\theta}} : \quad p(x|y,\theta) = \frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)},$$

and a prior $p(\theta)$, we marginalise θ

$$\begin{aligned} p(x|y) &= \int_\Theta p(x,\theta|y)\mathrm{d}\theta, \\ &= \int_\Theta p(x|y,\theta)p(\theta|y)\mathrm{d}\theta \end{aligned}$$

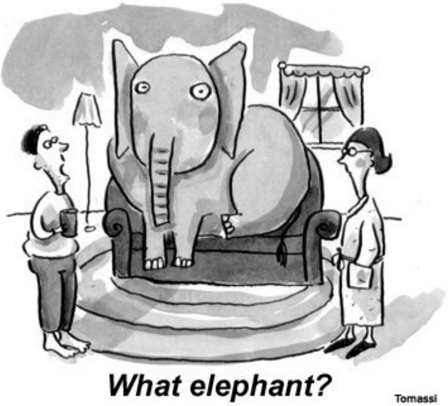
where again $p(\theta|y)$ controls the relative contribution of each model.

Summary

- The Bayesian statistical paradigm provides a power mathematical framework to solve imaging problems...
- It allows deriving optimal estimators for x..
- As well as quantifying the uncertainty in the solutions delivered...
- It supports hypothesis tests to inform decisions and conclusions...
- And allows operating with partially unknown models...
- And with several competing models...

So the problem is quite solved, right?

Not really...



How do we compute all these probabilities?

Outline

1 Imaging inverse problems

2 Bayesian modelling

3 Bayesian inference

4 Conclusion

Conclusion

There are several mathematical frameworks to solve imaging problems.

Variational approaches offer excellent point estimation results and efficient computer algorithms with theoretical guarantees.

However, they struggle to go beyond point estimation.

Bayesian statistics provide a powerful framework to formulate more advanced inferences and deal with other complications.

However, computing probabilities is challenges and they struggle to scale to large problems as a result.

Conclusion

In the next lecture...
We will explore ways of accelerating Bayesian inference methods by integrating modern stochastic and variational approaches at algorithmic, methodological, and theoretical levels.

Thank you!

Bibliography:

Cai, X., Pereyra, M., and McEwen, J. D. (2017). Uncertainty quantification for radio interferometric imaging II: MAP estimation. *ArXiv e-prints*.

Chambolle, A. and Pock, T. (2016). An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319.

Deledalle, C.-A., Vaiter, S., Fadili, J., and Peyré, G. (2014). Stein unbiased gradient estimator of the risk (sugar) for multiple parameter selection. *SIAM Journal on Imaging Sciences*, 7(4):2448–2487.

Durmus, A., Moulines, E., and Pereyra, M. (2018). Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM J. Imaging Sci.*, 11(1):473–506.

Fernandez-Vidal, A. and Pereyra, M. (2018). Maximum likelihood estimation of regularisation parameters. In *Proc. IEEE ICIP 2018*.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862.

Moreau, J.-J. (1962). Fonctions convexes duales et points proximaux dans un espace Hilbertien. *C. R. Acad. Sci. Paris Sér. A Math.*, 255:2897–2899.

Pereyra, M. (2015). Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*. open access paper, http://dx.doi.org/10.1007/s11222-015-9567-4.

Pereyra, M., Bioucas-Dias, J., and Figueiredo, M. (2015). Maximum-a-posteriori estimation with unknown regularisation parameters. In *Proc. Europ. Signal Process. Conf. (EUSIPCO) 2015*.

Robert, C. P. (2001). *The Bayesian Choice (second edition)*. Springer Verlag, New-York.

Zhu, L., Zhang, W., Elnatan, D., and Huang, B. (2012). Faster STORM using compressed sensing. *Nat. Meth.*, 9(7):721–723.