Analysis of Mixtures in Physical Spectra

R. FISCHER and V. DOSE
Max-Planck-Institut fuer Plasmaphysik and
EURATOM Association, Germany

Abstract: A reversible jump Markov chain Monte Carlo technique is applied to estimate the number and parameters of peaks in ubiquitous physical problems in the framework of Bayesian probability theory. For measured physical spectra often only the functional form of the structures is known but the number of the peaks and the parameters are unknown. The full joint posterior distribution for all parameters is sampled for estimating the number of components supported by the significant information in the noisy data and for estimating the unknown parameters for the most probable number of components. The method is applied to the classical Old Faithful density estimation problem and the physical problem of resolution enhancement and spectral decomposition of high-resolution electron energy loss data.

Keywords: REVERSIBLE JUMP, MARKOV CHAIN MONTE CARLO, MIXTURES, LINEAR PHYSICAL MODELS.

1. INTRODUCTION

Measured spectra in physics often comprise an unknown number of structures of known parametric family. The structures are commonly blurred with an apparatus transfer function and superposed with a smooth background. The parameters of the peaks are usually estimated for different numbers of components separately, and various criteria are employed to infer the number of components. The criteria used comprise simple significance tests like the data misfit as well as non-Bayesian methods such as the Akaike information criterion. The intention of the authors is to apply a recently proposed method of *reversible jump* Markov chain Monte Carlo (RJMCMC) by Green (1995) and Richardson *et al.* (1997) to physical problems where the number of components of a mixture of distributions is jointly estimated with the parameters of the components. The inference is done in the Bayesian framework using MCMC for sampling the joint posterior distribution.

The parametric distributions commonly used in physics are Gaussian distributions for apparatus functions, Lorentzian distributions for states decaying with a characteristic life time, Voigt profiles (convolution of Gaussian with Lorentzian) comprising both, life time broadening and apparatus broadening effects, and asymmetric variants of these distributions encompassing many-particle effects. The present paper addresses two examples

where Gaussian components are suitable. The other parametric families will be published in a forthcoming paper with the corresponding physical examples. In addition to the mixture of the blurred components the measured spectra may be superposed by a smooth background.

Deconvolution of an apparatus function is an ill-posed problem which only recently has been solved successfully with an adaptive kernel method using an entropic prior (Fischer, 1998). The adaptive kernel method uses a sufficiently large number of, e.g., Gaussian kernels with fixed, densely distributed mean values and variable variances. The amplitudes and variances are marginalized within the Bayesian probability theory. The effective degrees of freedom of the form-free, de-blurred reconstruction depend on the variance distribution of the kernels which constitute correlations in the reconstruction. The drawback of the adaptive kernel method is that the number of kernels (mixture components) is large and fixed. Problems where the number of components are of interest can not be tackled. The RJMCMC method for mixtures provides a useful generalization of the adaptive kernel method.

The object of this work is to extend the principle ideas of the RJMCMC method developed for good-natured Gaussian components to parametric families most commonly used in physics including blurring transfer functions and additive smooth background. In Section 2 we summarize the model for the classical density estimation problem using Gaussian mixtures in the framework given by Richardson *et al.* (1997), and describe a common model used for physical problems. Section 3 summarizes the RJMCMC approach. In Section 4 we show the results for the classical density estimation problem of the Old Faithful geyser which was compared to the results of the adaptive kernel method. Though the Old Faithful geyser does not comprises a physical problem, it represents a class of physical problems which deals with density estimation of scattered events such as the reconstruction of blurred images in astrophysics from the measurement of the distribution of photons. In section 5 the physical problem of analyzing a data set of high-resolution electron energy loss spectroscopy (HREELS) is shown. We conclude with a summary and an outlook.

2. MIXTURE MODELS

In general, the classical Gaussian mixture approach introduced by Richardson *et al.* (1997) is different from mixture models describing physical problems.

2.1 *Gaussian mixture model*

For the classical approach the observations x_i are assumed to be independently drawn from the likelihood pdf

$$p(x_i \mid \boldsymbol{w}, \boldsymbol{\theta}, E) = \sum_{k=1}^{E} w_k f(x_i, \theta_k) \tag{1}$$

$$f(x_i, \theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right) \tag{2}$$

where x_i are the observations, $i = 1, \ldots, N_{data}$, E is the number of mixture components (*expansion order*), w_k are the component weights (sum up to 1), $k = 1, \ldots, E$, and $f(x_i, \theta_k)$ are Gaussian mixture components with parameters $\theta_k = (\mu_k, \sigma_k)$.

An identifiability problem arises from the invariance of the likelihood pdf (1) under permutation of the indices k. We adopt the idea of Richardson *et al.* (1997) to impose an identifying ordering constraint on the parameters μ_k, $\mu_1 < \mu_2 < \cdots < \mu_E$. This constraint is arbitrary since we can also impose ordering constraints on w_k or σ_k. The reason for reducing the parameter space is to allow for more efficient exploration of the multi-modal probability space, and, hopefully, to identify individual components and estimate their parameters from the MC sample. Unfortunately, the ordering constraint may not be sufficient to efficiently explore the posterior pdf and to eliminate all problems in interpretation of the sample. As Celeux *et al.* (2000) pointed out the truncation of the multi-modal posterior space does not necessarily account for the geometry and shape of the unrestricted posterior distribution. All applications have to be carefully analyzed for effects arising from the ordering constraint.

According to Richardson *et al.* (1997) we introduce a group label z_i, which indicates the identity or label of the component from which each observation x_i is drawn. For known allocation variable z_i the likelihood is

$$p(x_i \mid z_i = k, \boldsymbol{\theta}) \quad = \quad f(x_i, \theta_k) \tag{3}$$

where the z_i are supposed to be independently drawn from

$$p(z_i = k \mid \boldsymbol{w}) \quad = \quad w_k \qquad \text{for} \quad k = 1, \ldots, E. \tag{4}$$

The joint posterior pdf is

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}, E) \quad = \quad p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}, E) p(\boldsymbol{z} \mid \boldsymbol{w}, E) p(\boldsymbol{w} \mid E) p(\boldsymbol{\theta} \mid E) p(E) \tag{5}$$

with the conjugate priors

$$p(\boldsymbol{w} \mid E, \boldsymbol{\delta}) \quad = \quad \text{Dirichlet}(\boldsymbol{w} \mid \boldsymbol{\delta}) \tag{6}$$

$$p(\boldsymbol{\theta} \mid E, \xi, \kappa, \alpha, \beta) \quad = \quad \prod_{k=1}^{E} N(\mu_k \mid \xi, \kappa) \Gamma(\sigma_k^{-2} \mid \alpha, \beta) \quad . \tag{7}$$

The prior for E is flat. For details of the priors, the hierarchical model, and the Gibbs sampling of the parameters see Richardson *et al.* (1997).

2.2 *Physical mixture model*

For ubiquitous linear physical problems the mixture is written as

$$D(x_i) \quad = \quad \sum_j A(x_i - y_j) \left(s \sum_{k=1}^{E} w_k f(y_j, \theta_k) + b_j(\boldsymbol{\eta}) \right) \tag{8}$$

$$d(x_i) \quad = \quad D(x_i) + \epsilon_i \tag{9}$$

where $d_i = d(x_i)$ is the measured data (intensity) at coordinate $x_i, i = 1, \ldots, N_{data}$, D_i is the model data, and ϵ_i is the (statistical) uncertainty associated with the measurement process. We assume mean values of $\langle \epsilon_i \rangle = 0$, $\langle \epsilon_i^2 \rangle = \Sigma_i^2$. $A_{ij} = A(x_i - y_j)$ is the blurring matrix accounting for the apparatus transfer function. The y_j are chosen fine enough to

allow to fit the data within the noise level. This is often achieved by setting $\boldsymbol{y} = \boldsymbol{x}$. Knowing the apparatus transfer function precisely is vital for a useful deconvolution of the data. The best method is to measure the transfer function and to combine the measurement uncertainties of the transfer function and the data to be deconvolved as shown in Dose *et al.* (1998). Most commonly, the transfer function is assumed to be described by a function with parameters incorporated into the fitting routine. In many applications the transfer function is well approximated by a Gaussian, whereas for the present HREELS application the asymmetric transfer function is separately measured.

E is the number of mixture components. Please note we use a slightly different notation from Richardson *et al.* (1997) where the number of mixtures is denoted by k. s is a scale factor accounting for the total intensity of the measured spectrum without background, which is not necessarily normalized to one. $b(\boldsymbol{\eta})$ is the background with parameters $\boldsymbol{\eta}$. Smooth backgrounds are commonly described by polynomials or spline functions. Fischer *et al.* (2000) proposed to discriminate a smooth background from the useful signal by introducing an additional expansion order for modeling only the background. For the present HREELS application it is sufficient to use a constant background $b_j(\boldsymbol{\eta}) = \eta_0$.

The likelihood probability density for the present counting experiment is a Poisson distribution,

$$p(\boldsymbol{d} \mid \boldsymbol{D}) \quad = \quad \prod_{i=1}^{N_{data}} \frac{D_i^{d_i}}{d_i!} \mathrm{e}^{-D_i} \tag{10}$$

which is well approximated for large counts with a Gaussian

$$p(\boldsymbol{d} \mid \boldsymbol{D}, \Sigma) \quad = \quad \prod_{i=1}^{N_{data}} \frac{1}{\sqrt{2\pi}\Sigma_i} \exp\left(-\frac{(d_i - D_i)^2}{2\Sigma_i^2}\right) \tag{11}$$

with variance $\Sigma_i^2 \approx d_i$. Since conjugate priors are no longer used, the algorithm can easily be modified for the Poisson likelihood. Nevertheless, for the HREELS application we use the Gaussian likelihood since this allows easy incorporation of the transfer function measurement uncertainty (Dose *et al.* , 1998).

For the HREELS application the parameters are E, s, w_k, θ_k, and η_0. The augmentation concept of introducing an allocation variable is no longer applicable because the intensities d_i at coordinate x_i can not be identified with a single component. Therefore, the full conditional pdfs necessary for Gibbs sampling can not be calculated analytically. Conjugate priors are no longer necessary to keep things simple, and we change to more sensible priors. As an exception, for a counting experiment with data binning the allocation concept can still be applied by augmenting the data vector from N_{data} intensities d_i at coordinates x_i to $\sum_i d_i$ events with degenerated coordinates. The likelihood (1) has to be replaced by a multinomial distribution. The drawback is that the number of events and, hence, the number of allocation variables in a typical physical experiment is too large to be tackled numerically.

The priors for the mean values are chosen to be flat on the measurement interval to avoid mixtures unrestricted by the data. It is easily seen that E is estimated to be too large when the data do not constrain the parameter space. Components located outside the interval with a variance small enough to prevent significant overlap with the data are not

restricted by the likelihood. Due to a flat prior on E, marginalization of the parameters of those extra components does not result in penalizing Ockham (Bayes) factors.

The prior for s is exponential with a mean according to the overall intensity. The prior for the background parameter η_0 is chosen constant on a positive interval.

3. MCMC

The types of MCMC moves divide into the two groups of moves within the parameter space and moves between different parameter spaces. The within parameter space moves depend on the model used and are chosen to be Gibbs moves for the Gaussian mixture model with data augmentation and Metropolis-Hastings moves for the ubiquitous physical model. The Gibbs moves for the Gaussian model are shown in detail in Richardson *et al.* (1997) where the use of conjugate priors allow for the calculation of the full conditional posterior distributions of all parameters.

The moves between different parameter spaces are provided by the reversible jump method proposed by Green (1995) and applied to normal mixtures by Richardson *et al.* (1997). The two types of reversible jump moves are split/combine moves of adjacent components and birth/death moves of *empty* components. The parameters of the split/combine moves are assigned by matching the 0^{th}, 1^{st}, and 2^{nd} moments of the single component to those of the splitted components. A birth move comprises the generation of a new component with random location, amplitude and variance, preferentially from the prior distributions. A death move is simply deleting an existing component and re-scaling the amplitudes. Details can be found in Richardson *et al.* (1997).

4. APPLICATION: OLD FAITHFUL GEYSER

A first example is the textbook density-estimation problem of the eruptions of the Old Faithful geyser (Silver, 1994). Figure 1 shows the raw data for the duration of 109 eruptions displayed as a scatter plot and as a histogram using 109 (!) bins. The mean of the posterior pdf is shown as solid line. The dashed lines correspond to an asymmetric $\pm$ one standard deviation above/below the mean value.

The posterior pdf of the number of components is shown in the left panel of Figure 2. The most probable number of components is 4. The right panel shows the changes of the number of components E against the number of sweeps. The MCMC sample mixes well over E. Small as well as large values of E are short-lived.

For parameter estimation the MCMC sample for a fixed number of components E has to be analyzed. In Figure 3 typical traces of the parameters $\boldsymbol{w}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ are shown for the most probable number of components 4. The ordering constraint is $\mu_1 < \mu_2 < \cdots < \mu_E$. The trace is taken from a run of 10^6 sweeps, which included about 360 000 visits to $E = 4$. For the sake of clarity only every 200 visit is plotted. The pattern of the 4 components look quite different. The mean values of the first and third components (lowest mean at about 2 minutes, full circle and mean at about 4 minutes, downward triangle) are well determined, while the second mean value (upward triangle) is a little fuzzy. The first and third components have also well determined amplitudes and standard deviations. The fourth component (highest mean, open circles) exhibits switching from a mean value of about 6.2 minutes to a mean of about 4.5 minutes. At a duration of about 6.2 minutes there

is only the sparse information of two measured eruptions. At about 4.5 minutes there is a barely significant peak split-off from the main feature at about 4 minutes. The presence of both barely significant features results in a significant fourth component which switches between the two peaks. Please note the correlation of the three parameters w, μ, and σ when switching occurs. In addition, the switching behavior between both modes clearly show the good convergence of the MCMC sample.

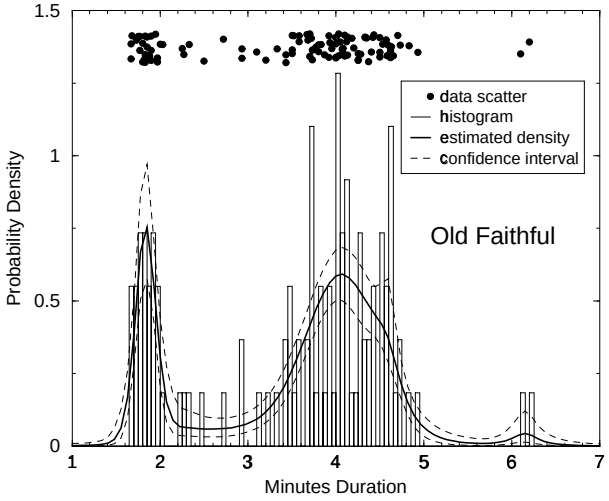

Figure 1. *Density estimation of 109 eruption durations of the Old Faithful geyser*

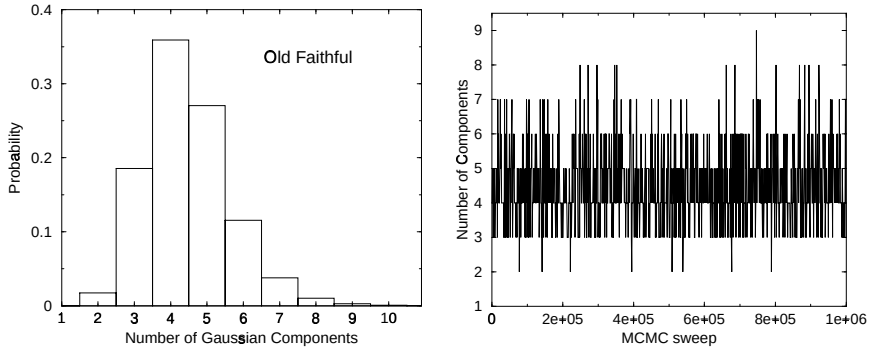

Figure 2. *Posterior probability and mixture of number of components for the Old Faithful geyser*

The mean of the posterior pdf as shown in Figure 1 has to be compared to previous analysis with the adaptive kernel method (Fischer, 1999). Both methods yield density estimates which compare quite well within the confidence interval. The results are insensitive to the methods used as long as methods with adaptive flexibility in the parameter space are applied. The advantage of the RJMCMC method is that it allows to estimate the parameters of the components which is not possible with the adaptive kernel method.

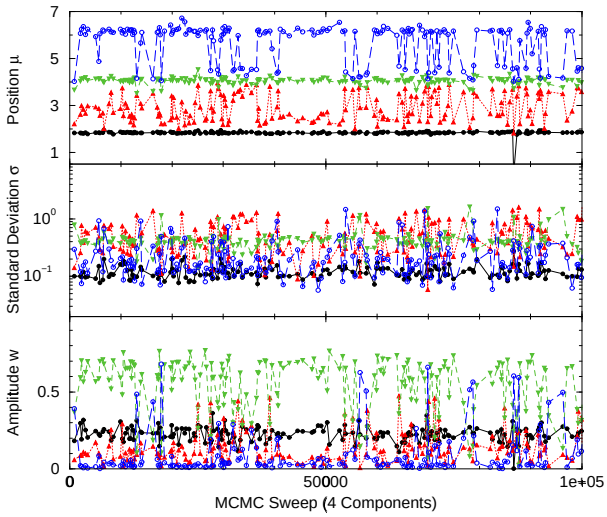

Figure 3. *Trace of Gaussian parameter estimates against visits to 4 components for the Old Faithful geyser*

5. APPLICATION: HREELS

A typical physical application for RJMCMC is the deconvolution of an apparatus broadening function from a blurred and noisy spectrum, the estimation of the background, and the estimation of the number and the parameters of the components underlying the spectra. The goal is to infer only the mixture components which are significantly supported by the data and to avoid the reconstruction of spurious structures which are due to the noise.

High-resolution electron energy loss spectroscopy (HREELS) is one of the most prominent techniques used to probe the vibrations of atoms and molecules adsorbed on solid surfaces. Mono-energetic electrons are inelastically scattered from a surface. An interpretation of the energy loss intensities of the scattered electrons (vibrational frequencies of adsorbed species) may give information on adsorption sites, coverage, strength of the surface bond and the degree of association of species adsorbed on the surface. The interpretation is more complicated if neighboring loss peaks overlap. Recently von der Linden (1997) showed how to improve the resolution in HREELS using the maximum-entropy concept. In the present work, the number and the parameters of the loss peaks are estimated in addition to the achieved resolution enhancement.

In the left panel of Figure 4 HREELS data (dots) of CO-molecules adsorbed on a $Pt_x Ni_{1-x}(111)$ single-crystal surface are shown. The specular reflected elastic electron energy distribution (dot-dashed line) provides directly the apparatus transfer function which blurres the loss peaks. Instead of using more sophisticated, and, hence, more expensive electron spectrometers for resolution enhancement, the apparatus function can be deconvolved post-measurement from the data. Please note the asymmetric transfer function which can not be quantified with simple analytical functions. The deconvolved energy loss

distribution (solid line) is the mean of the posterior pdf of the mixture with marginalized E.

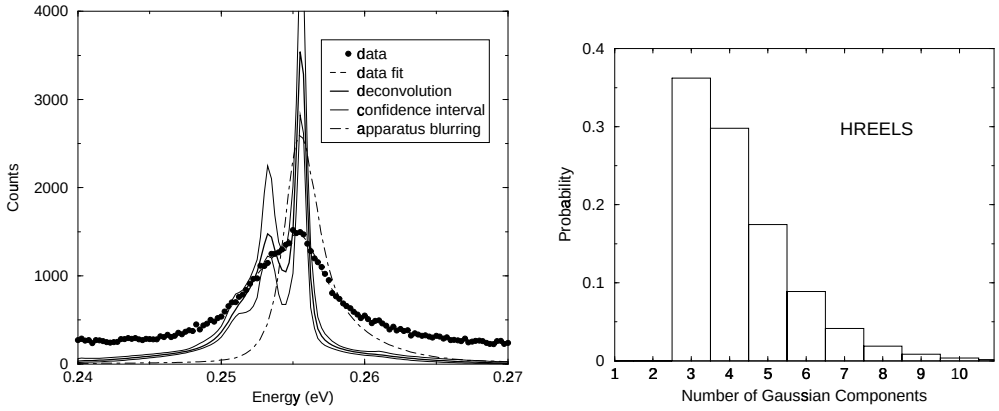

Figure 4. *Left: Deconvolution, spectral decomposition, and background subtraction applied to data from HREELS. The apparatus transfer function (dot-dashed) is deconvolved from the data (dots). Right: Posterior probability of the number of components.*

The posterior probability distribution of the number of components is shown in the right panel of Figure 4. The most probable number of components is 3.

The left panel of Figure 5 shows the changes of the number of components E against the number of sweeps. The mixing is not as good as with the Old Faithful example because the Gaussian mixture components are convolved with the asymmetric apparatus transfer function. The combine/split moves with the moment matching condition is sub-optimal for the HREELS application.

In the right panel of Figure 5 typical traces of the Gaussian parameters $\boldsymbol{w}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ are shown for the most probable number of components 3. About 36% of the sweeps visit to $E = 3$. The three parameters of the most prominent, third component (largest mean at about 0.255 eV, cross) are well determined. The mean values of the first and second components overlap. The standard deviations and the amplitudes show a distinct switching behavior due to a bi-modality in the posterior which is well sampled. The question arises if the ordering constraint applied to the mean values is reasonable for this application. The mean values for w_1 and w_2 (σ_1 and σ_2) are almost the same and the variances of these estimates are by far too large. This switching behavior clearly shows up the limits of ordering constraints. For inferring the parameters new techniques have to be found to allocate the components to the individual loss peaks. Celeux *et al.* (2000) suggest to discard the ordering constraints and to sample the multi-modal posterior pdf. The MCMC sample is finally classified using clustering tools. Due to the increased complexity of sampling a multi-modal distribution this task is far beyond the scope of this paper.

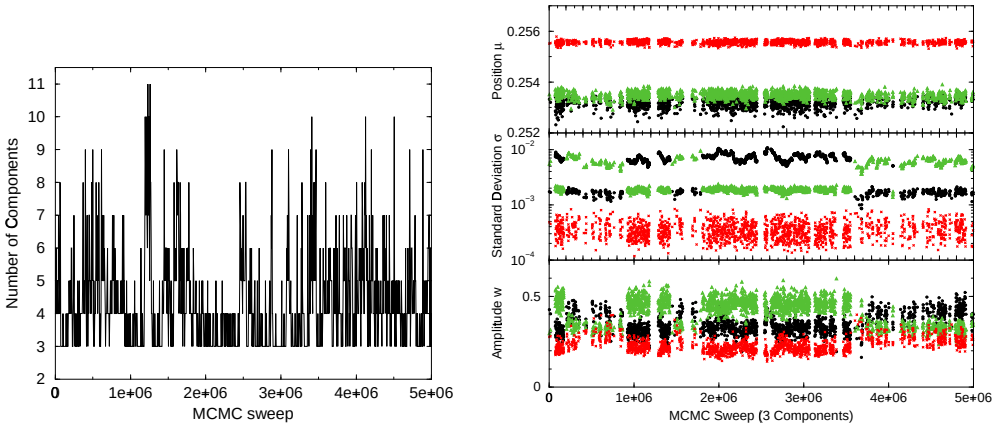

Figure 5. *Left: Mixture of number of components. Right: Trace of Gaussian parameter estimates against visits to 3 components.*

The fit to the data is best shown with the posterior distribution of the deviances $-2\log p(\boldsymbol{d}\,|\,E,\boldsymbol{\theta})$ for increasing E, which is the familiar χ^2-misfit for a Gaussian likelihood distribution. The left panel of Figure 6 depicts normalized distributions for the unconditioned deviance (thick line) and the deviances conditional on $E = 3-7$, which overlap substantially. The misfit of the data is in accordance with the number of data $N_{\rm data} = 121$. The scale parameter and the background parameters of the physical model are simultaneously sampled along with the Gaussian parameters and the number of mixture components. The posterior pdf of the scale parameter and of the constant background parameter η_0 are depicted in the middle and right panels, respectively. The distribution for the constant background is asymmetric with a steep decrease to larger values which is determined by the noise level of the data. Small values of η_0 are counterbalanced by very broad Gaussian components. The smooth decrease of the posterior pdf with the decrease of η_0 is due to the weak effect of Ockham's (Bayes) factor penalizing more complex models. The same arguments hold also for the posterior pdf of the scale s.

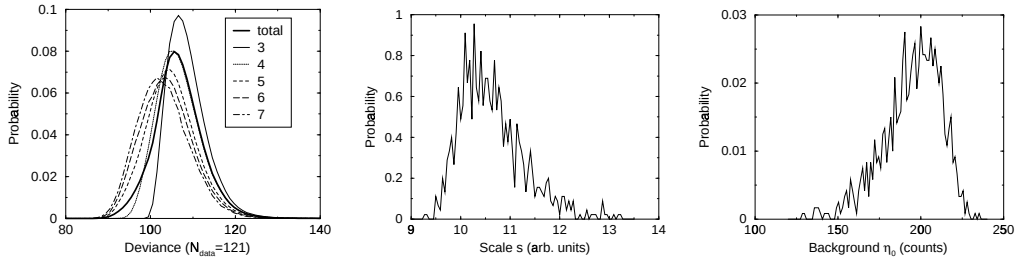

Figure 6. *Posterior probability densities of deviances, scale parameter, and background parameter*

6. SUMMARY

The RJMCMC method was applied to the density estimation problem of the Old Faithful geyser using the classical Gaussian mixture algorithm and to a physical deconvolution problem where the background as well as the number of mixture components and their parameters are estimated simultaneously. Sampling from the unconstrained multi-modal posterior pdf seems to be advantageous despite the cumbersome techniques to explore the parameter space.

In a subsequent paper the method will be applied to various physical problems where Lorentzian, Voigt, and asymmetric mixtures have to be used. A major problem arises due to slow mixing of the chain with concomitant bad convergence behavior. More elaborated matching techniques for the split/combine moves have to be found for allowing routine work on analyzing physical problems.

REFERENCES

Castelloe, J. (1999). Reversible jump Markov chain Monte Carlo analysis of spatial Poisson cluster processes with bivariate normal displacement. Proceedings of Computing Science and Statistics, 31st Symposium on the Interface, 1999.

Celeux, G., *et al.* (2000). Computational and inferential difficulties with mixture posterior distributions. to appear in *J. Amer. Statist. Assoc.* .

Dose, V., *et al.* (1998). Deconvolution based on experimentally determined apparatus functions. *Maximum Entropy and Bayesian Methods*. (G. Erickson, ed.). Dordrecht: Kluwer Academic, 147-152.

Fischer, R., *et al.* (1998). Energy resolution enhancement in ion beam experiments with Bayesian probability theory. *Nucl. Inst. Meth. B* **136-138**, 1140.

Fischer, R. (1999). The Adaptive Resolution Concept in Form-Free Distribution Estimation. *Proceedings of the Workshop on Physics and Computer Science*. (W. Kluge, ed.). Department of Computer Science, Christian-Albrechts-University, Kiel, Germany.

Fischer, R., *et al.* (2000). Background estimation in experimental spectra. *Phys. Rev. E* **61**, 1152.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

von der Linden, W., *et al.* (1997). Improved resolution in HREELS using maximum-entropy deconvolution: CO on Pt_xNi_{1-x}(111). *J. Electron Spectrosc. Relat. Phenom.* **83**, 1-7.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. B* **59**, 731–792. Richardson, S. and Green, P. J. (1998). Correction to "On Bayesian analysis of mixtures with an unknown number of components". *J. Roy. Statist. Soc. B* **60**, U3.

Silver, R. and Martz, H. (1994). Applications of quantum entropy to statistics. *J. Amer. Statist. Assoc. , 1994 Proceedings Toronto*, 61–70.