In: Bayesian Inference and Maximum Entropy Methods in Science and Engineering,
G. Erickson, Y. Zhai (eds.).
American Institute of Physics 2004, 295-302.

Bayesian model comparison using Gauss approximation on multicomponent mass spectra from CH_4 plasma

H.D. Kang and V. Dose

Centre for Interdisciplinary Plasma Science, Max-Planck-Institut für Plasmaphysik, EURATOM Association, Boltzmannstrasse 2, D-85748 Garching bei München, Germany

Abstract. We performed Bayesian model comparison on mass spectra from CH_4 rf process plasmas to detect radicals produced in the plasma. The key ingredient for its implementation is the high-dimensional evidence integral. We apply Gauss approximation to evaluate the evidence. The results were compared with those calculated by the thermodynamic integration method using Markov Chain Monte Carlo technique. In spite of very large difference in the computation time between two methods a very good agreement was obtained. Alternatively, a Monte Carlo integration method based on the approximated Gaussian posterior density is presented. Its applicability to the problem of mass spectrometry is discussed.

INTRODUCTION

Radicals from low temperature process plasma play a decisive role in the plasma-surface reaction and are therefore of fundamental interest in many industrial applications. Mass spectrometry is a unique tool to provide quantitative information of the radical flux density with a lateral resolution. Radicals in the plasma however appear to a very tiny extent compared to the hundreds times denser feed gas and thus the extraction of their information from mass signals of the mixture gases constitutes challenging problem. Bayesian probability theory suitably tackles such problems. It was successfully applied to the decomposition of multicomponent mass spectra.[1, 2] In particular, its ability to handle unknown mixtures and to identify their components was proved in the case of CH_4 rf plasmas through Bayesian model comparison.[3]

The key quantity for Bayesian model comparison is the probability of the data given a particular model equation and consists of a high-dimensional integral, the so-called evidence integral. The integrand of this integral can take a complicated form depending on the prior probabilities and the model equations, and thus a general analytical treatment is impossible. A traditional numerical integration is also not conceivable due to the huge multivariate density. A reliable method for computing the evidence uses the thermodynamic integration method (TDI) borrowed from statistical physics.[4, 5] In combination with the Markov chain Monte Carlo (MCMC) technique it provided promising results in challenging problems.[3, 4] But this method is computationally very demanding.

In this work we apply an analytical approach using Gauss approximation, also known as 'steepest descent' approximation, to evaluate the evidence integral for the case of mass spectrometry from CH_4 rf process plasmas. Though our sampling distribution contains a

fourth order term of the cracking coefficients **C** and species concentrations **X** it is exactly Gaussian given either **C** or **X**, which encourages the use of the approximative tool. Much care was devoted to the choice of adequate priors since they should be simple enough for the analytical treatment and strong enough to handle large numbers of unknowns which are typical in mass spectrometry. Alternatively, we employed the approximated Gaussian posterior density p_G to carry out the Monte Carlo evaluation of the integrals, which becomes possible through the replacement of the huge multivariate density by p_G. The results are compared with those obtained by TDI.

MODEL

We assume a linear response of the mass spectrometer and model the mass signal vector $\mathbf{d}_j$ of measurement j as

$$\mathbf{d}_j = \mathbf{C}\mathbf{x}_j + \boldsymbol{\varepsilon}_j. \tag{1}$$

$\mathbf{x}_j$ is the vector of species concentrations[6] and $\boldsymbol{\varepsilon}_j$ the error vector associated with $\mathbf{d}_j$. **C** is the cracking matrix which results from the fragmentation of species in the ionization source. The cracking column vectors and the data vector are normalized to sum up to one, which implies the same sum norm for the concentrations. The number of column vectors in **C** as well as the dimension of the concentration vectors corresponds to the number of species incorporated into the model which is unknown at the moment. The determination of a set of species which best describes the mass spectra measured is the central topic of our model comparison.

The probability for a particular model having E species is given in terms of the data **D** and variances **S** by Bayes' theorem[7]

$$p(E|\mathbf{D},\mathbf{S},I) = \frac{p(E|I)\ p(\mathbf{D}|E,\mathbf{S},I)}{p(\mathbf{D}|\mathbf{S},I)}. \tag{2}$$

For reasons of convenience we use the notation $\mathbf{D} \equiv \{\mathbf{d}_j\}$ and $\mathbf{X} \equiv \{\mathbf{x}_j\}$. **S** denotes the ensemble of diagonal matrices $\mathbf{S}_j$ with the components $\left(\mathbf{S}_j^{-2}\right)_{ii} = 1/s_{ij}^2$, where s_{ij} is the measurement error of the i-th mass channel in measurement j. For the prior probability $p(E|I)$ we choose a constant $p(E|I)$=1/E_{max}. E_{max} is the maximal number of components in the model. $p(\mathbf{D}|\mathbf{S},I)$ is the normalization factor. The marginal likelihood $p(\mathbf{D}|E,\mathbf{S},I)$ can be obtained by applying the marginalization rule and Bayes' theorem as

$$p(\mathbf{D}|E,\mathbf{S},I) = \int d\mathbf{C}d\mathbf{X}\ p(\mathbf{C}|E,I)\ p(\mathbf{X}|E,I)\ p(\mathbf{D}|\mathbf{C},\mathbf{X},E,\mathbf{S},I) \equiv T. \tag{3}$$

This is the evidence integral which involves the integration of the likelihood multiplied by the prior over the parameter space. Assuming the normal distribution of the error $\boldsymbol{\varepsilon}_j$ the sampling distribution (likelihood) $p(\mathbf{D}|\mathbf{C},\mathbf{X},E,\mathbf{S},I)$ is Gaussian in the probabilistic term[7]

$$p\left(\mathbf{D}|\mathbf{C},\mathbf{X},E,\mathbf{S},I\right) = \prod_j \frac{1}{\prod_i s_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\mathbf{d}_j - \mathbf{C}\mathbf{x}_j\right)^T \mathbf{S}_j^{-2}\left(\mathbf{d}_j - \mathbf{C}\mathbf{x}_j\right)\right\}. \tag{4}$$

$p(\mathbf{C}|E,I)$ and $p(\mathbf{X}|E,I)$ are the prior probabilities for the cracking patterns and compositions, respectively.

PRIOR

Cracking patterns (CP) of stable molecules are listed in the literature as point estimates.[8] For radicals for which there are no such values available they can be estimated from the tabulated pattern of the next heavier stable molecule as a very coarse estimate. This is indeed vague prior knowledge, but if the data are sufficiently informative they will overrule the prior information completely. For the composition vectors we can also make a very rough estimation from the experience that in a CH_4 plasma the main neutral constituents are CH_4($\sim$90%) and H_2($\sim$10%) and that the reaction products like radicals and C_2-molecules may appear in a fractional percent range ($\sim$0.1%).

According to the principle of maximum entropy these point estimates can be coded into exponential prior. In our model there is, however, a further crucial information that both cracking coefficients and concentrations are normalized and confined in the interval $[0,1]$. This can be employed to determine the implied variance θ_V of the model parameters. θ_V allows forming the more informative Gaussian prior which has a maximum in the allowed parameter range. The use of Gaussian instead of exponential prior is vital for treating problems associated with a large number of unknown parameters, since the posterior obtained using this prior gives a solution even if the number of data is smaller than the number of unknowns or even in the case of no data.

The exponential prior of the true, unknown parameter θ in terms of the tabulated value θ_0 reads

$$p(\theta|\theta_0) = \frac{1}{Z} e^{-\lambda\theta}, \tag{5}$$

where $Z = (1-e^{-\lambda})/\lambda$ is the normalization factor. λ depends on the point estimate θ_0 and can be determined from the requirement $\langle\theta\rangle = \theta_0$. The implied variance θ_V to the parameter θ is given by

$$\theta_V = \langle\theta^2\rangle - \theta_0^2. \tag{6}$$

The second moment $\langle\theta^2\rangle$ on the support of $[0,1]$ can be easily obtained over the second derivative of $Z(\lambda)$ as

$$\begin{aligned} \langle\theta^2\rangle &= \frac{1}{Z}\int_0^1 \theta^2\, p(\theta)\, d\theta = \frac{1}{Z}\frac{d^2}{d\lambda^2} Z(\lambda) \\ &= \frac{1}{Z}\left(\frac{2(1-e^{-\lambda})}{\lambda^3} - \frac{2e^{-\lambda}}{\lambda^2} - \frac{e^{-\lambda}}{\lambda}\right). \end{aligned} \tag{7}$$

The Gaussian prior with the variance θ_V generally appears to be rigid. The point estimates therefore should be chosen with care, especially for the CP of radicals.

To make our analysis more reliable, calibration measurements $\mathbf{d}_i^*$ for the stable species can be used as further prior information in form of the posterior $p(\mathbf{c}_i|\mathbf{d}_i^*,I)$ on their

cracking pattern $\mathbf{c}_i$, which results in a very narrow Gaussian prior. With the calibration data $\mathbf{d}_i^*$, the data error $\mathbf{S}_i^*$ and the literature values $\mathbf{c}_0$ the posterior on $\mathbf{c}_i$ reads

$$\begin{aligned} & p(\mathbf{c}_i|\mathbf{d}_i^*, \mathbf{S}_i^*, \mathbf{c}_0) \\ \sim \;& p(\mathbf{c}_i|\mathbf{c}_0)\; p(\mathbf{d}_i^*|\mathbf{c}_i, \mathbf{S}_i^*, \mathbf{c}_0) \\ \sim \;& \prod_m^M \exp(-\lambda_{mi} c_{mi}) \exp\left\{ -\frac{1}{2} \frac{(d_{mi}^* - c_{mi})^2}{s_{mi}^{*2}} \right\}, \end{aligned} \tag{8}$$

where M is the total number of mass channels. This is again a Gaussian with the variance $\mathbf{s}_{mi}^{*2}$. We determine the mode $\mathbf{c}_i^0$ subject to the condition that the components of $\mathbf{c}_i$ sum up to one. The functional ψ to be minimized is then

$$\psi = \sum_m^M \lambda_{mi} c_{mi} + \frac{1}{2} \sum_m^M \frac{(d_{mi}^* - c_{mi})^2}{s_{mi}^{*2}} - \mu(\sum_m^M c_{mi} - 1), \tag{9}$$

yielding the decoupled equations

$$\begin{aligned} c_{mi} - s_{mi}^{*2}\mu &= d_{mi}^* - \lambda_{mi} s_{mi}^{*2} \\ \sum_m^M c_{mi} &= 1. \end{aligned} \tag{10}$$

These lead to the solution

$$c_{mi}^0 = d_{mi}^* - \lambda_{mi} s_{mi}^{*2} + \frac{s_{mi}^{*2}}{\sum_m^M s_{mi}^{*2}} \left\{ 1 - \sum_m^M (d_{mi}^* - \lambda_{mi} s_{mi}^{*2}) \right\} \tag{11}$$

to form the Gaussian

$$p(\mathbf{c}_i|\mathbf{d}_i^*, \mathbf{S}_i^*, \mathbf{c}_0) \sim \exp\left\{ -\frac{1}{2} \sum_m^M \frac{(c_{mi} - c_{mi}^0)^2}{s_{mi}^{*2}} \right\}. \tag{12}$$

We introduce the parameter vector $\boldsymbol{\theta}$ which includes all non-zero cracking coefficients and concentrations. The global prior probability on $\boldsymbol{\theta}$ is then given by

$$p(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \prod_{k=1}^R \frac{1}{Z_k} \exp\left\{ -\frac{1}{2} \frac{(\theta_k - \theta_{0k})^2}{\theta_{Vk}} \right\}. \tag{13}$$

The restriction $0 < \theta_k < 1$ forces the corresponding normalization

$$Z_k = \int_0^1 \exp\left\{ -\frac{1}{2} \frac{(\theta_k - \theta_{0k})^2}{\theta_{Vk}} \right\}, \tag{14}$$

which was determined numerically.

EVIDENCE INTEGRAL: GAUSS APPROXIMATION

Using the prior and the likelihood determined above the integral T in Eq. (3) can be written as

$$T = \frac{1}{Z_P Z_L} \int e^{-\phi(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{15}$$

with

$$\phi(\boldsymbol{\theta}) = \sum_{k=1} \frac{1}{2} \frac{(\theta_k - \theta_{0k})^2}{\theta_{Vk}} + \sum_{i,j} \frac{1}{2} \frac{(d_{ij} - \sum_l c_{il} x_{lj})^2}{s_{ij}^2}. \tag{16}$$

Z_P and Z_L include all normalization factors in the prior and the likelihood, respectively. The Taylor-expansion of $\phi(\boldsymbol{\theta})$ up to the second order around the mode $\boldsymbol{\theta}_m$ yields

$$\phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}_m) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_m)^T \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_m) \tag{17}$$

where $\mathbf{H} = \nabla\nabla^T \phi(\boldsymbol{\theta})$ is the Hessian matrix. Note that the gradient of ϕ vanishes at the mode. The Gauss-approximated integral T_G is then written as

$$T \approx T_G = \frac{1}{Z_P Z_L} e^{-\phi(\boldsymbol{\theta}_m)} \int e^{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_m)^T \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_m)} d\boldsymbol{\theta}. \tag{18}$$

This integral can be performed analytically. We obtain

$$T_G = \frac{1}{Z_P Z_L} e^{-\phi(\boldsymbol{\theta}_m)} \frac{(2\pi)^{N/2}}{\sqrt{|\det \mathbf{H}|}}. \tag{19}$$

The determinant of the high-dimensional Hessian was calculated by Cholesky decomposition.[9] For the determination of the $\boldsymbol{\theta}_m$ we require the vanishing gradient from Eq (17)

$$\nabla\phi(\boldsymbol{\theta}) = \nabla\phi|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n} + \mathbf{H}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n} (\boldsymbol{\theta} - \boldsymbol{\theta}_n) = 0 \tag{20}$$

at an arbitrary point $\boldsymbol{\theta} = \boldsymbol{\theta}_n$. $\boldsymbol{\theta}_m$ then can be assigned iteratively using

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathbf{H}^{-1}(\boldsymbol{\theta}_n) \nabla\phi(\boldsymbol{\theta}_n). \tag{21}$$

$\boldsymbol{\theta}_m$, in addition, has to meet the normalization condition $\sum_i c_{ij} = 1$. This was considered through the Gaussian

$$\exp\left\{ -\frac{1}{2} \frac{(\sum_i c_{ij} - 1)^2}{\rho^2} \right\}, \tag{22}$$

which was added in $\phi(\boldsymbol{\theta})$. We allowed a tolerance of 1% ($\rho = 0.01$). Note that this condition is merely used to find the mode and should not affect the integral T_G.

It is important to point out that the integration limit of the analytical solution in Eq. (19) was $(-\infty, \infty)$, while our parameter space is limited to $[0, 1]$. A correction can

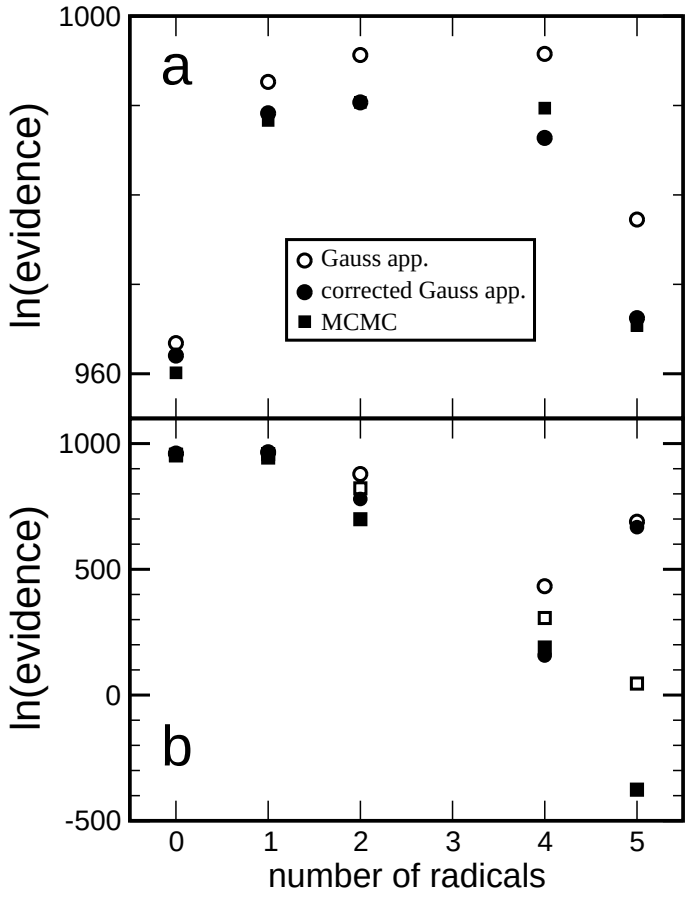

FIGURE 1. Logarithmic evidences calculated by (a) Gauss approximation, thermodynamic integration and (b) Monte Carlo integration using Gaussian posterior density. The x axis shows the number of radicals incorporated into each model, which were taken in the order, C_2H_5, CH_3, C_2H_3, CH_2, CH. All models, in addition, involve 5 stable species, H_2, CH_4, C_2H_2, C_2H_4, C_2H_6. In (b) random samples were taken from the Gauss (squares) and hyperbolic Cauchy (circles) distribution. Filled and open symbols refer to 10^7 and 10^8 sample numbers, respectively.

easily be done by Monte Carlo technique: multivariate samples are drawn from the normalized posterior density. The ratio of the number of draws N^* belonging to the parameter space to the total number of draws N provides the needed correction factor, resulting in the corrected integral T_{corr}

$$T_{corr} = T_G \, \frac{N^*}{N}. \tag{23}$$

The calculated evidences are shown in Fig. 1a along with those computed by TDI, which is described elsewhere in detail.[3] The number of radicals in the x-axis corresponds to different models considered. The agreement between two methods (filled symbols) is excellent for all models, even in terms of the absolute scale. This is a very remarkable result in view of the strongly varying computation time of minutes for the Gaussian approximation and days for TDI. Both methods also agree in the choice of

the best model which involves two radicals CH_3 and C_2H_5. The uncorrected evidences clearly deviate from those of MCMC, indicating that our posterior significantly stretches out of the allowed parameter space. It is also shown that the correction factor becomes larger the more complicated models are. The evidences without correction for finite integration supports would have chosen the model 4 as the best one.

EVIDENCE INTEGRAL: MONTE CARLO INTEGRATION

The Gaussian approximated posterior density p_G provides an interesting starting point for a numerical evaluation of the evidence. Using p_G the evidence integral T can be rewritten as

$$T = \int \frac{p_0}{p_G}\, p_G\, d\boldsymbol{\theta}, \tag{24}$$

where p_0 is the true posterior. The integral T is then obtained by averaging the ratio p_0/p_G over the multivariate normalized Gaussian p_G. The random vector $\boldsymbol{\theta}$ from the multivariate Gaussian can be drawn after an orthonormal transformation of the exponent

$$(\boldsymbol{\theta}-\boldsymbol{\theta}_m)^T\mathbf{H}(\boldsymbol{\theta}-\boldsymbol{\theta}_m) = \mathbf{w}^T\mathbf{w} \tag{25}$$

using

$$\mathbf{w} = \mathbf{H}^{1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}_m) \rightarrow \boldsymbol{\theta} = \mathbf{H}^{-1/2}\mathbf{w}+\boldsymbol{\theta}_m. \tag{26}$$

The diagonalization makes the MC integral simple and fast since the problem of drawing an n-dimensional random sample is thereby reduced to the problem of drawing n independent random samples from a unit-variance Gaussian density. The square root of the positive definite real symmetric matrix can be taken again by Cholesky decomposition.[9] Considering the Jacobian $\partial\boldsymbol{\theta}/\partial\mathbf{w} = |\mathbf{H}^{-1/2}|$ the evidence integral reads

$$T = \int \frac{p_0}{p_G} p_G(w) \left|\frac{\partial\boldsymbol{\theta}}{\partial\mathbf{w}}\right| d\mathbf{w} \tag{27}$$

$$= \frac{1}{\sqrt{\det\mathbf{H}}}\frac{1}{Q}\sum_q^Q \frac{p_0(\boldsymbol{\theta}_q)}{p_G(\boldsymbol{\theta}_q)}. \tag{28}$$

Despite the radical shrinkage of the sampling density to p_G the numerical demand may still be high, it is related to how well the Gaussian posterior fits the true one. Another problem when using a Gaussian density arises from the fact that it may decay faster than the true posterior density. The ratio p_0/p_G may then become very large in some cases. Though such draws are rare due to the low sampling probability compensating the large contribution to the integral, a single large draw may become a problem in the Monte Carlo runs of finite length. To overcome this difficulty heavy-tail functions like Cauchy or hyperbolic Cauchy distribution could be used since we are free to replace the Gaussian by another similar function after the diagonalizing transformation.

Fig. 1b shows the results for the sampling number of 10^7 and 10^8. For models having no radicals or only one the evidences are roughly comparable with those in Fig. 1a. For

more complicated models, however, they are far from being acceptable, which indicates that the multivariate Gaussian posterior deviates from the true one. On the other hand this also shows that the sampling density still covers a large part of the parameter space due to the high dimensionality and cannot be fully scanned in a reasonable CPU time, because even in the case of strong deviation of the Gaussian and the true posterior a long enough Monte Carlo run would give a right answer. The use of hyperbolic Cauchy function (circles) instead of Gaussian (squares) for drawing random samples improves the results only slightly. Sampling distributions with much heavier tails like Lorentzian cannot heal this drawback since the higher probability for void draws in turn requires longer Monte Carlo runs.

SUMMARY

We employed the Gauss approximation to implement Bayesian model comparison on mass spectra from CH_4 rf plasmas. Our algorithm provides evidence integrals which are in excellent agreement with those calculated by the thermodynamic integration method for low-dimensional problems and it reduces the computation time by nearly three orders of magnitude. This method, however, confronts the obstacle that our Hesse matrix is not always positive-definite due to the fourth order terms in the likelihood. In this case the Cholesky decomposition does not work. The major problem is obviously to find a proper starting point $\boldsymbol{\theta}_1$ for which the Hesse matrix is positive definite. This problem becomes more severe the larger the parameter space is and consequently limits the use of the Gauss approximation. The Monte Carlo evaluation of the evidence integrals using the Gauss-approximated normalized density could not be accomplished in an acceptable CPU time, again due to the high dimensionality. For small-size models with up to $\sim$50 parameters these difficulties do not appear and both analytical and numerical integration work trouble-free.

REFERENCES

1. H. D. Kang, R. Preuss, T. Schwarz-Selinger, and V. Dose, J. Mass Spec. **37**, 748 (2002).
2. T. Schwarz-Selinger, R. Preuss, V. Dose, and W. von der Linden, J. Mass Spec. **36**, 866 (2001).
3. H. D. Kang and V. Dose, J. Vac. Sci. Technol. A, accepted (2003).
4. W. von der Linden, R. Preuss, and V. Dose, in *Maximum Entropy and Bayesian Methods*, edited by W. von der Linden, V. Dose, R. Fischer, and R. Preuss (Kluwer Academic Publishers, 1999), p. 319.
5. J. Skilling, in *Maximum Entropy and Bayesian Methods*, edited by P. Fougère (Kluwer Academic Publishers, 1990), p. 341.
6. For the absolute concentrations the mass signal has to be scaled by a sensitivity factor which depends on particular molecules.
7. D. Sivia, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, 1996).
8. A. Cornu and R. Massot, *Compilation of Mass Spectral Data* (Heyden London, 1979).
9. W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in Fortan* (Cambridge University Press, Cambridge, 1992).