Visualizing high-dimensional posterior distributions in Bayesian modeling

Jarkko Venna and Samuel Kaski
Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 9800, FIN-02015 HUT, Finland
{*jarkko.venna,samuel.kaski*}*@hut.fi*

Abstract— **In Bayesian modeling inference is based on the posterior distribution of the model parameters. The closed-form solution is seldom known and samples of the posterior have to be computed with Markov Chain Monte Carlo (MCMC) methods. The problem is that for large models the samples are high-dimensional, and it is hard to piece together properties of the posterior. Our proposal is to use a non-linear dimensionality reduction method for visualizing sampled posterior distributions; the current method is based on Self-Organizing Maps and Fisher metrics. Sample applications have been included on studies of convergence and mixing of the MCMC simulations, of shape (e.g. multimodality) of the posterior, and of sensitivity of the model.**

I. Introduction

Bayesian modeling is a principled way of incorporating prior knowledge into data analysis. The result is a distribution of models; uncertainty in the data is transformed into uncertainty of the model parameters. This posterior distribution is therefore of great interest, but in most cases its analytical form is not known. In practice the only possibility often is to resort to sampling methods like MCMC [1] for inference.

The result of sampling is a high-dimensional cloud of points in the parameter space. Based on this cloud we would like to answer questions such as is the posterior mean or mode a valid summary of the distribution, or is the distribution possibly multimodal. In MCMC sampling there are further questions: Has the chain converged to the stationary distribution? Are there regions in the parameter space where the sampler performed poorly?

We might also be interested in knowing how robust the model is to changes in parameter values. Robustness is reflected in the magnitude of changes in the model output, caused by changes in the data. By visualizing the posterior distribution we hope to gain a quick preliminary overview to its properties.

II. Bayesian modeling in a nutshell

In Bayesian modeling the relationship between the data $\mathbf{y}$ and the parameters $\boldsymbol{\theta}$ of the model is defined by the *likelihood* $p(\mathbf{y}|\boldsymbol{\theta})$. The knowledge that we possess about the parameter values prior to having observed the data is given by the *prior distribution* $p(\boldsymbol{\theta})$. By combining these we get the *posterior distribution* that summarizes the state of our knowledge of the parameters after we have observed the data. The posterior can be calculated from the prior and likelihood with the Bayes formula

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \quad (1)$$

where $\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is a normalizing term.

While in maximum likelihood estimation a single parameter value is sought, in Bayesian data analysis the result is the whole parameter distribution. This allows us to take into account the uncertainty about the parameter values in inference. A Bayesian model can be used to predict new values $\tilde{\mathbf{y}}$ by using the posterior predictive distribution

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \int p(\tilde{\mathbf{y}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \quad (2)$$

where the uncertainty of the parameter values has been taken into account by integrating over the posterior distribution.

III. Visualizing the posterior distribution

Visualization is a fast way of presenting an overview of data. Posterior distributions have been traditionally visualized by plotting the marginal distributions. The joint marginal distribution of two variables can be plotted using a scatter or contour plot. For three dimensional distributions advanced computer graphics methods visualize the shape [2]. However, it is easy to find examples where the marginal distributions or even low-dimensional linear projections of a high-dimensional posterior distribution can give a misleading picture of the shape of the distribution (Fig. 1).

Traditional visualizations of the posterior distribution depend on the parameterization of the model. If the parameterization is changed the visualization changes even though the resulting model does not change. A related problem is that a straightforward visualization of the posterior does not reflect the effect the parameters have on the model. A large change of parameter values might cause only a small change in the model output if the model is "stiff" in that part of the parameter space. These problems can be solved by choosing a proper metric.

A. *The metric of the model family*

Many visualization methods depend either explicitly or implicitly on the metric of the data space. In modeling the

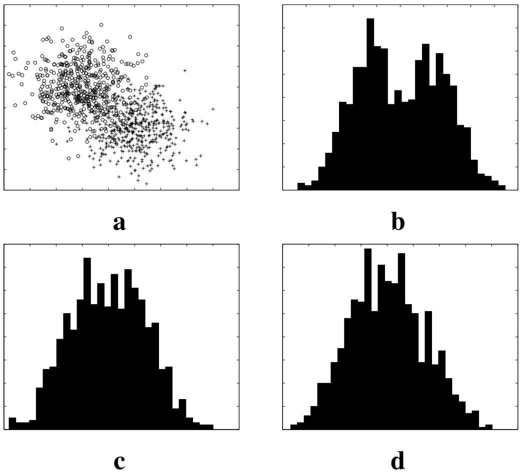

Fig. 1. It is sometimes hard to notice multimodality from marginal distributions. (**a**) Plot of two 2D Gaussians. (**b**) Histogram in the direction of maximum separation. (**c**) Histogram of the marginal density on horizontal axis. (**d**) Histogram of the marginal density on vertical axis.

ultimate goal often is not to study the model parameters *per se*, but to create models that can generate or predict new values. Our goal is to visualize the posterior in a way that reveals how the posterior predictive distribution behaves.

The difference of two distributions can be measured with the Kullback-Leibler divergence D_{KL}. If the distributions are close-by the divergence is

$$D_{KL}(p(\tilde{\mathbf{y}}|\boldsymbol{\theta})||p(\tilde{\mathbf{y}}|\boldsymbol{\theta}+d\boldsymbol{\theta}) = d\boldsymbol{\theta}^T\mathbf{J}(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{3}$$

where $\mathbf{J}$ is the Fisher information matrix

$$\mathbf{J}(\boldsymbol{\theta}) = -E_{p(\tilde{\mathbf{y}}|\boldsymbol{\theta})}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(\tilde{\mathbf{y}}|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(\tilde{\mathbf{y}}|\boldsymbol{\theta})\right)^T\right]. \tag{4}$$

This defines the so called *Fisher* or *information metric* [3], [4]to the parameter space. Distances between two far-away points are defined as path integrals along the shortest path between the points. In this paper we have used local approximations of distances witch are sensible for methods such as the SOM that rely on local distances. The distance is calculated based on the Fisher matrix at either one end point or both.

There may be points in the parameter space where the Fisher information matrix is singular (see for example [5]). In such cases the distance between two different points is zero and the Fisher metric is not a proper metric. However the points can be considered equivalent in terms of the model as the change in parameters does not cause any change in the output. Thus the Fisher metric can be considered as a metric on equivalent classes of points. The singularities might also cause problems in the learning algorithm used. This can be solved by regularizing the Fisher metric by adding a small amount of Euclidean metric to make the distance non-singular.

The Fisher metric has two useful properties. First, it defines the distance between models as the change in the predicted distribution. Thus visualizations based on it emphasize the differences that are relevant to predictions. Second, it is invariant to changes in the parameterization.

B. Visualization method

For visualization the posterior distribution has to be presented in a low-dimensional space. One way of doing this is to project the samples to two dimensions in such a way that the important properties of the distribution are preserved. Some suitable methods are multidimensional scaling (MDS) [6], and the self-organizing map (SOM) [7], and their variants. Some other distance-based projection methods could be used as well. In this paper we have used the SOM with the metric (3). The first reason for choosing the SOM is that it mainly relies on local distances and hence the distance approximation suits it. the second reason is that the number of samples is usually large and the MDS-based methods do not scale up computationally as well.

The SOM consists of a regular grid of *units*. Each unit contains a model vector $\mathbf{m}_i \in \mathbb{R}^n$, where n is the dimensionality of the data. The sequential SOM algorithm iterates two steps. First, for each sample $\boldsymbol{\theta} \in \mathbb{R}^n$, the best matching model vector $\mathbf{m}_c$ is sought using the equation

$$c = \operatorname*{argmin}_i\{d(\boldsymbol{\theta}, \mathbf{m}_i)\}, \tag{5}$$

where $d(\boldsymbol{\theta}, \mathbf{m}_i)$ is the distance between the sample $\boldsymbol{\theta}$ and the model vector $\mathbf{m}_i$. When the best matching model vector is found, the model vectors are updated by

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)]. \tag{6}$$

Here $t = 0, 1, 2, \ldots$ is the number of the iteration. The function $h_{ci}(t) = h(\|\mathbf{r}_c - \mathbf{r}_i\|; t)$, where $\mathbf{r}_i$ and $\mathbf{r}_c$ are the location vectors on the SOM grid for units i and c, is the neighborhood function. The neighborhood function controls how each unit is updated in each step. The update rule (6) results from the natural gradient and is the same as in the basic SOM algorithm. For further details on learning metrics for SOMs and for SOMs in Fisher metrics see [8], [9], [10].

IV. Applications: Toy demo

To demonstrate the effect of parameterization on the visualization we created two 2-dimensional Gaussian distributions that simulate the posterior distribution of the same Gaussian model. One of these distributions was reparameterized so that the scale of the parameter on the vertical axis became five times larger than the horizontal. One-dimensional SOMs were computed for both of these data sets with both the Euclidean and the Fisher metric.

The SOM in the Fisher metric behaved similarly in both the scaled and the unscaled case (Fig. 2); it is invariant to the change in parameterization. The SOM in the Euclidean metric, on the other hand, focuses on modeling the variation in the stretched vertical dimension.

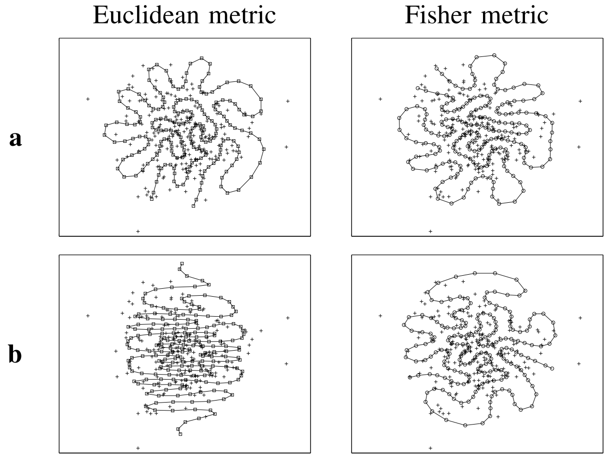

Fig. 2. Demonstration of the Fisher metric: (**a**) 1D SOM computed of a symmetric normal distribution. (**b**) The vertical dimension has been reparameterized to be 5 times larger than in **a**. In the visualization the axes have been re-scaled back to the original.

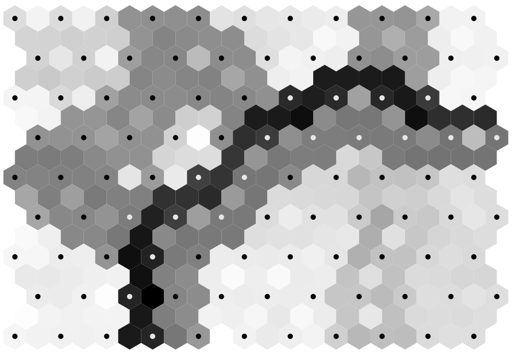

Fig. 3. Multimodality of the posterior visualized by a U-matrix of the SOM in the regularized Fisher metric (logarithmic scale). Units of the SOM are marked with dots. Dark color between between units corresponds to large change in the model and a light color to a small change

V. Applications: Bayesian MLP

To demonstrate the visualization method we created a simple Bayesian Multi-Layer Perceptron (MLP) having one hidden layer with three hidden units, and one output unit. The activation function was the hyperbolic tangent in the hidden layer and linear in the output layer. A Gaussian noise model was assumed. The data consisted of 100 points generated from the $\sin(2x)/x$ function in the interval $[0.001, 2\pi]$. Six chains of 6000 samples were generated using Hybrid Monte Carlo sampling [11], and the first 500 samples of each chain were discarded.

A SOM was computed of the combined data from all the chains. The Fisher matrix of an MLP network, and its metric, is conditional on the input data. Hence the optimal metric changes with the input data. Since we want a summary of the behavior of the model, we chose expected metric (the expectation of the Fisher matrix over the input data). Thus the visualizations illustrate the average change in the model when parameters change. Finally the metric of the SOM was regularized by taking the average of the squared Euclidean and the Fisher metric.

The shape of the posterior distribution seems to be multimodal (Fig. 3). A dark stripe partitions the posterior into two clusters, and the lighter gray stripes further into several sub-regions. By studying the parameter values in the different parts of the SOM it was verified that these subregions indeed do have different parameter values and are not artifacts. These regions will be studied further below.

We first studied convergence of the MCMC sampling by visualizing whether the different sampling chains become mixed together. If the sampler has converged, each unit of the SOM should receive an approximately equal number of samples from each chain. It is clear that the sampler has not yet reached convergence. The samples from each chain have stayed mostly in small separate parts of the parameter space

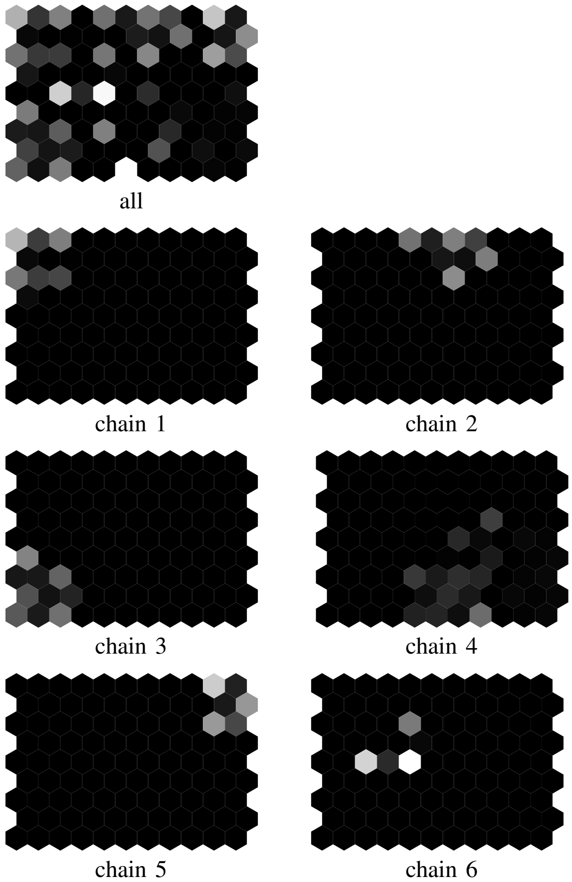

Fig. 4. Top left: Total number of MCMC samples in each SOM unit shown by gray shades (black: zero, white: maximal). The rest: distribution of the samples from each of the 6 chains coded with the same gray scale (black: zero, white: maximal)

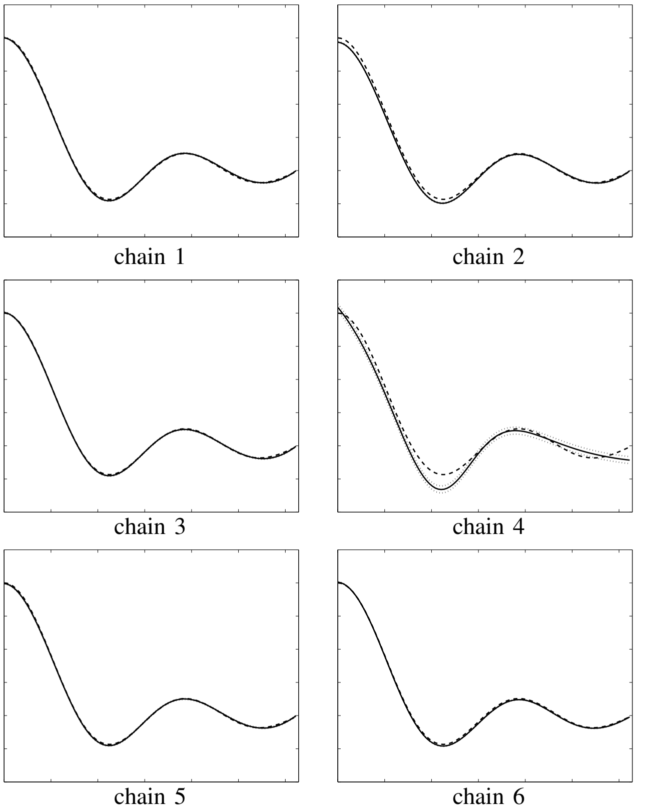

Fig. 5. The output of the MLP at the SOM unit that got the most hits from the chain, one image per chain. *Solid line*: Output of the MLP. *Dotted line*: Estimated standard deviation of the output. *Dashed line*: Target function.

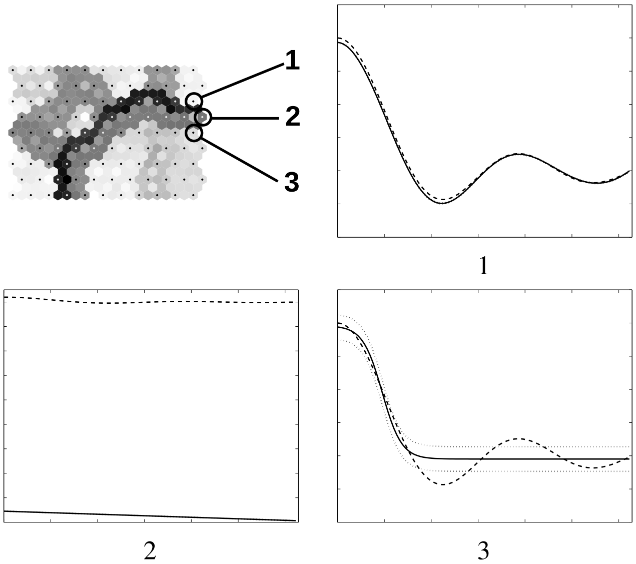

Fig. 6. The change in the MLP output when moving along the three units on the right edge of the SOM (Top left). *Solid line*: Output of the MLP. *Dotted line*: Estimated standard deviation of the output. *Dashed line*: Target. function. Note that the scale of the vertical axis in image 2 differs from the others, but the dashed line is the same.

(Fig. 4). The parts seem to roughly correspond to the modes revealed by the U-matrix.

To better understand the result we studied the predictions the models in each mode produce. The outputs of the MLPs with a representative parameter value of each mode (the unit receiving most samples) are shown in Figure 5. With the slight exception of chains 2 and 4 the outputs of the MLPs are very similar. Hence, the multimodality does not seem to affect the posterior predictive distribution.

In this model the multimodality arises from the non-identifiability of the MLP. The order of the hidden layer units can be changed without changing the output.

Finally, the black stripe of the U-matrix is studied in Figure 6. The outputs of the models on either side of the stripe (1 and 3 in Fig. 6) are very similar to each other when compared to the one in the middle (2 in the figure). Also the parameter values on different sides of the stripe are very different, which further supports the interpretation that the modes probably result from different permutations of the hidden units.

VI. Discussion

We have introduced a method for visualizing posterior distributions based on MCMC samples of them. The visualizations are invariant to parameterizations and hence focus on those differences in models that cause differences in predictions. The visualizations were demonstrated to be applicable to detecting multimodality in the posteriors, and convergence and mixing of the MCMC chains. It would additionally be possible to study sensitivity to changes in the data and the prior with them.

Acknowledgments

This work was supported by the Academy of Finland, in part by the grant 1164349.

References

[1] W.R.Gilks, S. Richardson, and D.J.Spiegelhalter, *Markov Chain Monte Carlo in Practice*, ser. Interdisciplinary Statistics. Boca Raton, Florida: Chapman & Hall/CRC, 1995.

[2] E. J. Wegman and Q. Luo, "On methods of computer graphics for visualizing densities," *Journal of Computational and Graphical Statistics*, vol. 11, no. 1, pp. 137–162, 2002.

[3] R. E. Kass and P. W. Vos, *Geometrical Foundations of Asymptotic Inference*. New York: Wiley, 1997.

[4] S. Amari, *Differential-Geometrical Methods in Statistics*. New York: Springer, 1990.

[5] S. Watanabe, "Algebraic analysis for nonidentifiable learning machines," *Neural Computation*, vol. 13, pp. 899–933, 2001.

[6] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec 1952.

[7] T. Kohonen, *Self-Organizing Maps*. Berlin: Springer, 1995, (Third, extended edition 2001).

[8] S. Kaski, J. Sinkkonen, and J. Peltonen, "Bankruptcy analysis with self-organizing maps in learning metrics," *IEEE Transactions on Neural Networks*, vol. 12, pp. 936–947, 2001.

[9] J. Peltonen, A. Klami, and S. Kaski, "Learning more accurate metrics for self-organizing maps," in *Artificial Neural Networks - ICANN 2002*, J. R. Dorronsoro, Ed. Berlin: Springer, 2002, pp. 999–1004.

[10] S. Kaski and J. Sinkkonen, "Principle of learning metrics for data analysis," *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, Special Issue on Data Mining and Biomedical Applications of Neural Networks*, Accepted for publication.

[11] R. M. Neal, "Bayesian training of backpropagation networks by the hybrid monte carlo method," Dept. of Computer Science, University of Toronto, Tech. Rep. CRG-TR-92-1, April 1992.