Visualizations for Assessing Convergence and Mixing of MCMC

Jarkko Venna, Samuel Kaski, and Jaakko Peltonen

Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland
{jarkko.venna, samuel.kaski, jaakko.peltonen}@hut.fi

Abstract. Bayesian inference often requires approximating the posterior distribution with Markov Chain Monte Carlo (MCMC) sampling. A central problem with MCMC is how to detect whether the simulation has converged. The samples come from the true posterior distribution only after convergence. A common solution is to start several simulations from different starting points, and measure overlap of the different chains. We point out that Linear Discriminant Analysis (LDA) minimizes the overlap measured by the usual multivariate overlap measure. Hence, LDA is a justified method for visualizing convergence. However, LDA makes restrictive assumptions about the distributions of the chains and their relationships. These restrictions can be relaxed by a recently introduced extension.

1 Introduction

Probabilistic generative modeling is one of the theoretical foundations of current mainstream machine learning and data analysis. Bayesian inference makes very accurate but computationally intensive predictions possible, and gives rigorous methods for model selection and complexity control. In a nutshell, the uncertainty in the data is converted into uncertainty of the model parameters in the form of a distribution. Inference of parameter values and of predictions is then done based on this distribution.

Bayesian inference is potentially very powerful but closed-form solutions are seldom available. Inference has to be based on either sophisticated approximation methods or simulations with Markov Chain Monte Carlo (MCMC) [1] sampling. MCMC sampling is a very versatile yet computationally intensive procedure. The main practical problem of MCMC is how to assess whether the simulation has converged. The resulting samples come from the true distribution only after convergence.

There are several strategies for monitoring convergence [2]. Often in practice convergence is assessed by starting the simulation from several different initial conditions, and by monitoring when the different simulation chains become sufficiently mixed together. The mixing can be monitored visually on scatter plots of the MCMC samples against all pairs of variables, which is of course feasible only for models with few parameters. An alternative is to measure convergence quantitatively; measures of the overlap of the different sampling chains have been proposed by Brooks and Gelman [3]. The measures have the problem that rules of thumb are required for deciding whether the simulation has converged or not, and hence they are often complemented with visualizations. The other advantage of visualizations is that they are useful also for analyzing reasons of convergence problems.

It turns out that the main multivariate convergence measure equals the cost function of a one-dimensional LDA (for a definition of LDA see [4]), a method that discriminates between data classes. Here the classes are the different sampling chains. Our first main result or suggestion is to use LDA for visual evaluation of convergence. It has a rigorous criterion for visualizing convergence and complements the existing quantitative measures. Our second main result is an extension of the LDA visualization by applying a less restrictive measure of the overlap of the chains, resulting in a connection with a recent extension of discriminant analysis.

2 Bayesian modeling in a nutshell

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

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

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

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

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

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

In practice the posterior distribution is usually not known in closed form and has to be approximated. A common method for approximation is MCMC sampling. MCMC generates samples x_t that are distributed proportionally to the posterior distribution. These samples can be used to estimate any statistic of the distribution and integrals over the posterior get approximated with sums over samples.

3 Monitoring convergence using multiple sequences

3.1 Measuring convergence

One of the most common methods for monitoring MCMC convergence is the potential scale reduction factor (PSRF) proposed by Gelman and Rubin [5]. Multiple MCMC sequences are started from different (overdispersed) initial points and compared. At convergence the chains should come from the same distribution, which is assessed by comparing the variance and mean of each chain to the variance and mean of the combined chain.

The PSRF is defined for one-dimensional data as follows. A number (m) of parallel chains are started, with $2n$ samples each. Only the last n potentially better converged samples from each chain are used. The between-chain variance B/n and pooled within-chain variance W are defined by

$$\frac{B}{n} = \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{x}_{j\cdot} - \bar{x}_{\cdot\cdot}\right)^2 \quad \text{and} \tag{3}$$

$$W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=1}^{n}\left(x_{jt} - \bar{x}_{j\cdot}\right)^2 \ , \tag{4}$$

where $\bar{x}_{j\cdot}$ is the mean of the samples in chain j and $\bar{x}_{\cdot\cdot}$ is the mean of the combined chains.

By taking the sampling variability of the combined mean into account we get a pooled estimate for the posterior variance

$$\hat{V} = \frac{n-1}{n}W + \left(1 + \frac{1}{m}\right)\frac{B}{n}. \tag{5}$$

Finally an estimate $\hat{R}$ of PSRF is obtained by dividing the pooled posterior variance estimate with the pooled within chain variance,

$$\hat{R} = \frac{\hat{V}}{W} \ . \tag{6}$$

If the chains have converged, the PSRF is close to one, which makes it a useful indicator of convergence. It is not a perfect indicator, however, since it does not guarantee convergence. The chains might not have traveled the whole state space yet and might discover possible new areas of high probability. Additionally, it does not take higher-order moments into account, only the mean and variance, and it is applicable to only one variable at a time.

Brooks and Gelman [3] have extended the PSRF to a multivariate version, MPSRF. It is defined, similarly to the univariate PSRF, in terms of the estimate of the posterior covariance matrix $\hat{\mathbf{V}}$, which we get from (5) by replacing the scalar variances B/n and W with the covariance matrices

$$\frac{\mathbf{B}}{n} = \frac{1}{m-1} \sum_{j=1}^{m} \left(\bar{\mathbf{x}}_{j\cdot} - \bar{\mathbf{x}}_{\cdot\cdot}\right)\left(\bar{\mathbf{x}}_{j\cdot} - \bar{\mathbf{x}}_{\cdot\cdot}\right)^T \text{ and} \tag{7}$$

$$\mathbf{W} = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{t=1}^{n} \left(\mathbf{x}_{jt} - \bar{\mathbf{x}}_{j\cdot}\right)\left(\mathbf{x}_{jt} - \bar{\mathbf{x}}_{j\cdot}\right)^T . \tag{8}$$

In the multivariate case the comparison of within-chain variance to the pooled variance requires comparing the matrices. Brooks and Gelman chose to summarize the comparison by a maximum root statistic which gives the maximum scale reduction factor of any linear projection of $\mathbf{x}$. The estimate $\hat{R}^p$ of MPSRF is defined by

$$\hat{R}^p = \max_{\mathbf{a}} \frac{\mathbf{a}^T \hat{\mathbf{V}} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}} \tag{9}$$

$$= \frac{n-1}{n} + \left(\frac{m+1}{m}\right) \max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}/n}{\mathbf{a}^T \mathbf{W} \mathbf{a}} \tag{10}$$

$$= \frac{n-1}{n} + \left(\frac{m+1}{m}\right) \lambda_1, \tag{11}$$

where the λ_1 is the largest eigenvalue of the matrix $\mathbf{W}^{-1}\mathbf{B}/n$.

This criterion is very closely related to linear discriminant analysis (LDA). The goal of (a one-dimensional) LDA is to find the linear transformation $y = \mathbf{a}^T \mathbf{x}$ that maximizes the variance between classes, relative to the variance within classes. More formally, LDA solves the problem

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{B}_{ss} \mathbf{a}}{\mathbf{a}^T \mathbf{W}_{ss} \mathbf{a}} , \tag{12}$$

where $\mathbf{B}_{ss}$ and $\mathbf{W}_{ss}$ are the between and within sum of squares and cross products (SSCP) matrices which differ only by a constant scale from the corresponding covariance matrices. This is a generalized eigenvalue problem, and its solution $\mathbf{a}$ is the eigenvector corresponding to the largest eigenvalue of $\mathbf{W}_{ss}^{-1}\mathbf{B}_{ss}$.

Hence, disregarding the constants, MPSRF equals the cost function of (a one-dimensional) LDA. In other words, optimizing the LDA is equivalent to choosing the component that best detects convergence, in the sense of MPSRF. Monitoring convergence by MPSRF or by the LDA cost function is equivalent; if the chains can be discriminated, then they have not converged.

3.2 Visualizing convergence

Current practice. It is common practice to complement the convergence measures by visualizations of the MCMC chains. Visualizations are useful especially when analyzing reasons of convergence problems. Convergence measures can only tell that the simulations did not convergence, not why they did not.

MCMC chains have traditionally been visualized in three ways. Each variable in the chain can be plotted as a separate time series, or alternatively the marginal distributions can be visualized as histograms. The third option is a scatter or contour plot of two parameters at a time, possibly showing the trajectory of the chain on the projection. The obvious problem with these visualizations is that they do not scale up to large models with lots of parameters. The number of displays would be large, and it would be hard to grasp the underlying high-dimensional relationships of the chains based on the component-wise displays.

Some new methods have been suggested. For three dimensional distributions advanced computer graphics methods can be used to visualize the shape of the distribution [6]. Alternatively, if the outputs of the models can be visualized in an intuitive way, the chain can be visualized by animating the outputs of models corresponding to successive MCMC samples [7]. These visualizations are, however, applicable only to special models.

A principled way of visualizing convergence. The worst problem with the straightforward visualization methods is that they lack the means to focus on visualizing variables or dimensions that are relevant for convergence. This worsens the problems caused by the required large number of plots.

In the previous Section (3.1) it was noted that the MPSRF measure of MCMC convergence (10) is closely related to linear discriminant analysis (LDA). We will use this connection to justify the use of LDA to visualize the convergence of the MCMC sampler.

In summary, LDA finds a projection that best separates the classes in the sense of maximizing the between-class variation relative to within-class variation. For a one-dimensional projection this was shown to be equivalent to choosing MPSRF as the criterion for the projection.

There is no reason to confine the visualization to be one-dimensional. LDA chooses the second direction or projection axis to be the eigenvector corresponding to the second largest eigenvalue, etc. A K-dimensional LDA then maximizes $\sum_{k=1}^{K} \lambda_k$, the relative between-chain variance representable by the K directions together. This criterion could actually be used as an alternative convergence criterion to MPSRF; it takes directly into account deviation in several directions instead of only the dominant one.

When LDA is used to visualize MCMC convergence we in effect try to find a linear transformation that visualizes the convergence problems as clearly as possible, in the sense of the (extended) MPSRF measure.

3.3 Informative components

Brooks and Gelman [3] noted that any statistic calculated from the separate chains should be equal to the one calculated from the combined chain when the chains have reached convergence, as the distributions should then be the same. The LDA connection above resulted from comparing means and variances. We propose that instead of comparing a statistic, a more general measure would result from comparing the distributions themselves. A natural measure is the mutual information between the distributions and the chain index. The difference between this and the LDA (MPSRF) criterion is discussed below.

Problems with LDA. LDA assumes that each class is normally distributed with the same covariance matrix in each class. If the assumptions are correct, LDA discriminates between two classes optimally. This does not hold in general, however, in particular not before MCMC convergence for small data.

Another problem surfaces when generalizing LDA to several classes. The objective considers only pairwise divergences between classes, and no longer corresponds to optimal discrimination. See the Appendix for details.

To address the above problems, we suggest to complement LDA-based analysis with a generalization of LDA. The projection is linear but the assumptions about the distribution of data are relaxed.

Relevant component analysis. A recent method for finding *informative* or *relevant* components directly maximizes their class-prediction power [8]. Formally, the conditional (log) likelihood

$$L = \sum_{(\mathbf{x},c)} \log p(c|\mathbf{W}^T\mathbf{x}) \tag{13}$$

of classes is maximized within the subspace formed by the components. Here $\mathbf{x}$ is the sample, c is its class, and $\mathbf{W}$ is the (orthogonal) projection matrix whose columns are the component directions. The optimal projection is specific to the number of components sought. The well-defined objective for finite data, the likelihood, is asymptotically equivalent to the mutual information between components and classes. The task of finding such components was coined *relevant component analysis* (RCA). A sketch of the connection between LDA and RCA is presented in the Appendix.

In this paper the c are the different chains, and RCA maximizes the (log) likelihood of correctly guessing which MCMC chain each sample is from. For converged chains one cannot (asymptotically) do better than a random guess; hence, large likelihood indicates non-convergence which can be assessed visually from the RCA projection.

With finite data, we do not know the exact densities $p(c|\mathbf{W}^T\mathbf{x})$, but we can optimize the projection parameters by using a nonparametric estimate $\hat{p}(c|\mathbf{W}^T\mathbf{x})$ in the projection space. Since this estimate is non-parametric, RCA makes no distributional assumptions. For details on RCA and its optimization, see [8]. Technically, we replaced the stochastic gradient in [8] by conjugate (batch) gradient optimization.

The main justification for using RCA here is that it maximizes a flexible measure of separation of the classes. It remains an empirical question of how much the RCA improves the LDA-based visualizations. In Section 4.2 we apply both methods to assess convergence in a relatively simple task.

4 Analysis of a MCMC run

To demonstrate visual analysis of a MCMC sampler we have chosen a data set that contains reaction times for schizophrenics and nonschizophrenics. The model and the problem are described in the book Bayesian Data Analysis [9] (Example 16.4, p.426) and were also used to illustrate the use of the PSRF measure in the original article [5].

The data consist of (log) reaction time measurements from 11 nonschizophrenics and 6 schizophrenics. Each person had their reaction time measured 30 times. It is believed that schizophrenics suffer from attentional deficit on some measurements as well as an overall motor reflex retardation.

For the nonschitzophrenics the reaction time is modeled as a random-effects model with a distinct mean α_j for each person and a common variance σ_y^2. The reaction times for the schizophrenics are modeled with a two-component Gaussian mixture. With probability $(1-\lambda)$ there is no attention lapse and the response time has mean α_j and variance σ_y^2. With probability λ there is a delay and the response time has mean $\alpha_j+\tau$ and the same variance σ_y^2. To address the question about the amount of motor reflex retardation a hierarchical population model is devised. The means of the reaction times α_j are modeled to be normally distributed with a mean μ for the nonschizophrenics and a mean $\mu+\beta$ for the schizophrenics. The model can be expressed as

$$y_{ij}|\alpha_j, \zeta_{ij}, \boldsymbol{\phi} \sim \mathrm{N}(\alpha_j + \tau\zeta_{ij}, \sigma_y^2), \tag{14}$$

$$\alpha_j|\zeta_{ij}, \boldsymbol{\phi} \sim \mathrm{N}(\mu + \beta S_j, \sigma_\alpha^2), \tag{15}$$

$$\zeta_{ij}|\boldsymbol{\phi} \sim \mathrm{Bernoulli}(\lambda S_j), \tag{16}$$

where $\boldsymbol{\phi} = (\sigma_\alpha^2, \beta, \lambda, \tau, \mu, \sigma_y^2)$ contains the hyperparameters and y_{ij} is the response i from person j. The term S_j is an indicator that equals 1 for schizophrenics and 0 for nonschizophrenics, and ζ_{ij} is an unobserved indicator that equals 1 if the observation arose from a delayed response and 0 otherwise.

The hyperparameters in $\boldsymbol{\phi}$ are assigned a noninformative uniform prior density. Additionally, τ, σ_α^2 and σ_y^2 are restricted to be positive. The mixture parameter λ is further restricted to the interval $[0.001, 0.999]$. As all necessary conditional distributions were readily available, Gibbs sampling, a form of MCMC, was used.

Ten chains of 1500 samples each were generated from random starting positions. The MPSRF measure showed that the sampling had not converged. Calculating the univariate PSRF measures for the 23 variables we were interested in (all except the indicator variables ζ_{ij}) showed that several variables had not converged. At this point we still had no idea what had gone wrong with the sampler, or was the convergence just slow.

4.1 Visualization with LDA

Gaining insight on the problem. In order to better understand the behavior of the chains we visualized a part of the simulation, samples $[200, 600]$ around the point 350 after which the MPSRF measure seemed to have stabilized at a high value.

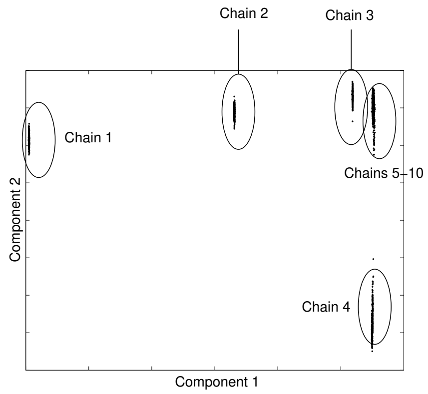

Fig. 1. Two-dimensional LDA projection of all samples from the interval $[200, 600]$. The ellipses have been drawn by hand to mark the chains.

It is clear from the LDA projection (Fig. 1) that there are five distinct clusters in the sample set. By color coding (not shown) the different chains with different colors it was easy to identify the chains. Six of the chains were clustered together and the other four formed a separate cluster each. Three of the chains were separated from the main cluster on discriminative component 1 and one on component 2. We additionally checked whether any of the separate chains could still be moving toward the common cluster, by color coding based on sampling time. There was no visible hint of that.

Verifying the findings. A further study showed that four of the chains had ended up in a degenerate part of the parameter space, that is, in a part where the mixture model has collapsed to a one-component model, already very soon after the initialization. For three of these chains (chains 1, 2, and 3) the probability of a sample being generated by a delayed mixture component was so low that no samples were assigned to it. This was apparent already by a quick look at the one-dimensional time series plots of these chains. The delay parameter τ had not changed at all from the starting position.

The reason for the fourth chain appearing separated is the reverse. Nearly all samples came from the mixture component representing delayed measurements, and hence the β and τ could not be identified separately. It was harder to diagnose the problem with this chain because the time series plots looked normal. The LDA visualization in Figure 1 helped to quickly identify the problem areas.

Checking the behavior of the sampler near convergence. At this point we could have modified our model or our sampler to remove the problems. If there are a

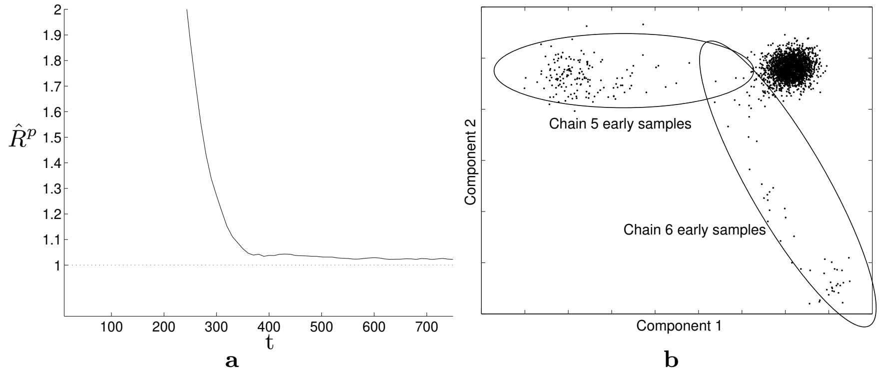

Fig. 2. a) MPSRF measure calculated from the nondegenerate chains (5-10). **b**) LDA projection of the nondegenerate samples from the interval $[200, 600]$. The ellipses have been drawn by hand to mark early samples from chains 5 and 6. The samples can be visualized by a time-based color code.

sufficient number of chains, a rapid alternative is to discard the degenerate ones. We computed the MPSRF measure again for the remaining chains. It is clear from Figure 2**a** that this time convergence has been reached after about 350 samples. For a demonstration we created a new LDA projection showing only the nondegenerate chains. In Figure 2**b** we can see that there are two 'tails' from chains which are moving toward the common distribution. By color coding the samples based on time we verified that the samples were indeed early samples and that the two chains became combined with the other chains after the early samples. Thus we could conjecture that the simulation had converged this time.

4.2 Visualization with RCA

We finally compare qualitatively the less restrictive RCA projection with LDA to verify that it gives the same or better insights on convergence.

From the two-dimensional RCA projection of all samples from the interval $[200, 600]$ (Fig. 3) we can see that RCA has discovered the same five clusters as LDA. Four of the clusters are composed of a single chain each, and the last consisted of six chains. In addition, RCA has found the two 'tails' of samples, generated by two chains converging toward the multi-chain cluster. These are the same 'tails' that were found using LDA on the nondegenerate chains (Fig. 2**b**).

Chains 1 and 2 are far from the others in both the LDA and the RCA visualizations. However, the LDA visualization kept the chain 4 far apart as well, whereas RCA placed it closer to chains 5-10 and instead separated the 'tails' of early samples of chains 5 and 6. Since the chain 4 can still be discriminated well, this yields a more informative projection.

In conclusion, RCA visualization displayed all the discovered convergence properties in a single two-dimensional visualization. No additional studies were

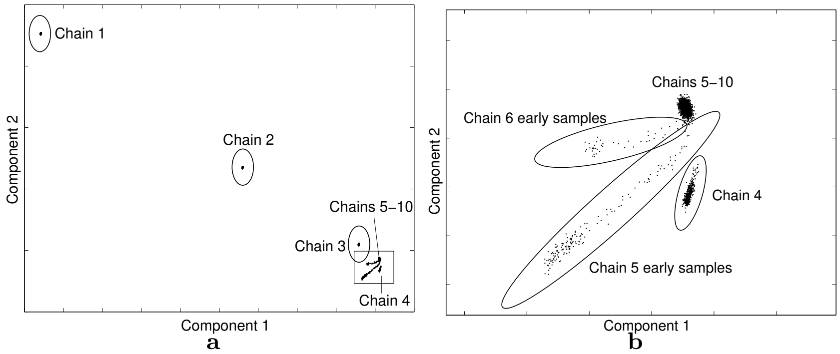

Fig. 3. a) 2D RCA projection of all samples from the interval $[200, 600]$. **b**) Enlarged view of the box in lower right corner of **a**.

required as with LDA. (A visualization corresponding to Figure 2**b** was computed just in case, and revealed only the same properties.)

5 Discussion

We have shown how to create visualizations for MCMC convergence analysis with linear discriminant analysis (LDA). Problems can be identified quickly using only a few visualizations. Justification for LDA comes from its connection to a common convergence measure: Its goal is to separate the different simulation chains, and if it is successful the simulation has not converged. This was demonstrated in a case study.

It is straightforward to extend the black-and-white visualizations of this Proceedings with color coding. If the different chains are colored differently it is easy to distinguish them in the figures. Coloring samples with shades that change as a function of time brings visible the evolution of the chains during sampling. Further possibilities for extensions are coloring according to the likelihood of the sampled models, or coloring according to the prior or posterior density of the samples. This would clearly show how much the posterior differs from the prior, for example.

If more details about the behavior of the sampler are of interest, some more technical measures like acceptance ratio or autocorrelation within a window around the sample could be visualized by the color code. This could possibly identify areas where the sampler is performing poorly. These ideas could be combined in an interactive visualization tool aimed at easy exploratory analysis of the behavior of a MCMC sampler.

Even though LDA can be used for principled visualizations of MCMC chains, it is based on assumptions that often do not hold. It assumes normally distributed chains, which usually does not hold, and that the covariance matrices of the chains are the same, which holds only after convergence. A new method, RCA, is based on a more flexible measure of the overlap of the simulation chains: The likelihood of predicting the chains, which asymptotically becomes the mutual information. These theoretical connections justify the use of the RCA, and it was demonstrated to work better than LDA in a small case study.

Finally, the objective function of RCA could additionally serve as a measure of convergence, when compared with a naive estimate that simply predicts the overall chain proportions. If the values are different, MCMC has not converged.

Acknowledgments

This work was supported by the Academy of Finland, grants 1164349 and 52123.

References

1. W.R.Gilks, S. Richardson, and D.J.Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics. Chapman & Hall/CRC, Boca Raton, Florida, 1995.
2. Stephen Brooks and Andrew Gelman. Some issues in monitoring convergence of iterative simulations. In *Proceedings of the Section on Statistical Computing*. ASA, 1998.
3. Stephen Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–456, 1998.
4. Neil H. Timm. *Applied Multivariate Analysis*. Springer Texts in Statistics. Springer-Verlag, New York, 2002.
5. Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.
6. Edward J. Wegman and Qiang Luo. On methods of computer graphics for visualizing densities. *Journal of Computational and Graphical Statistics*, 11:137–162, 2002.
7. Nicole A. Lazar and Joseph B. Kadane. Movies for the visualization of MCMC output. *Journal of Computational and Graphical Statistics*, 11:836–874, 2002.
8. Samuel Kaski and Jaakko Peltonen. Informative Discriminant Analysis. In *Proceedings of ICML-2003, The Twentieth International Conference on Machine Learning*, 2003. In press.
9. Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, Florida, 1995.
10. S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, San Diego, CA, 1999.

Appendix: Connection between LDA and RCA

Reformulating LDA. For simplicity, consider only the first LDA component $\mathbf{a}$. Denote $\sigma_{\mathbf{a}}^2 = \mathbf{a}^T \mathbf{W}_{ss} \mathbf{a}/N$, where N is the total number of samples. The LDA objective equals the variance of class centers along the projection direction, relative to the within-class variance:

$$\frac{\mathbf{a}^T\mathbf{B}_{ss}\mathbf{a}}{\mathbf{a}^T\mathbf{W}_{ss}\mathbf{a}} = \frac{1}{N\sigma_{\mathbf{a}}^2}\mathbf{a}^T\mathbf{B}_{ss}\mathbf{a} = \sum_c \frac{n_c}{N}\frac{(\mathbf{a}^T(\bar{\mathbf{x}}_{c.} - \bar{\mathbf{x}}_{..}))^2}{\sigma_{\mathbf{a}}^2} \ . \tag{17}$$

Since, for a scalar variable x, $E_{x_1,x_2}[(x_1-x_2)^2] = 2E[x^2] - 2(E[x])^2 = 2E[(x - E[x])^2]$, the objective further equals (up to a constant multiplier) the weighted sum of squared distances between class pairs:

$$\frac{2}{N\sigma_{\mathbf{a}}^2}\mathbf{a}^T\mathbf{B}_{ss}\mathbf{a} = \sum_{c_1,c_2} \frac{n_{c_1}n_{c_2}}{N^2}\frac{(\mathbf{a}^T(\bar{\mathbf{x}}_{c_1.} - \bar{\mathbf{x}}_{c_2.}))^2}{\sigma_{\mathbf{a}}^2} \ . \tag{18}$$

Since $\mathbf{a}^T\mathbf{a} = 1$, each Gaussian class has a variance of $\sigma_{\mathbf{a}}^2$ along the projection dimension. Then, for each pair of classes c_1 and c_2, the rightmost term equals the squared *Mahalanobis distance* of the projected class centers along the projection. This in turn equals the following *symmetrized Kullback-Leibler divergence* between the distributions along the projection [10]:

$$\frac{1}{\sigma_{\mathbf{a}}^2}(\mathbf{a}^T(\bar{\mathbf{x}}_{c_1.} - \bar{\mathbf{x}}_{c_2.}))^2 = D_{KL}(p(\mathbf{a}^T\mathbf{x}|c_1), p(\mathbf{a}^T\mathbf{x}|c_2)) + D_{KL}(p(\mathbf{a}^T\mathbf{x}|c_2), p(\mathbf{a}^T\mathbf{x}|c_1)) \tag{19}$$

LDA thus maximizes a sum of symmetrized Kullback-Leibler divergences between the classes along the projection, weighted by the fractions $n_{c_1}n_{c_2}/N^2$.

Improving the cost function. Optimizing the above objective (18) does not result in optimal discrimination. We will improve it in two steps. First, for each class pair (c_1, c_2), replace the symmetrization in (19) with the *Jensen-Shannon divergence*. This helps to reinterpret the objective in a form that can be easily generalized. For brevity, denote $y = \mathbf{a}^T\mathbf{x}$, denote the proportions of the class prior probabilities by $p_{c_1} = p(c_1)/(p(c_1)+p(c_2))$ and $p_{c_2} = p(c_2)/(p(c_1)+p(c_2))$, and set $q(y) = p_{c_1}p(y|c_1) + p_{c_2}p(y|c_2) = p(y|c_1 \vee c_2)$, where $c_1 \vee c_2$ referes to the distribution containing only clases c_1 and c_2. The Jensen-Shannon divergence is

$$\begin{aligned} D_{JS}(p(y,c_1), p(y,c_2)) &= p_{c_1}D_{KL}\left(p(y|c_1), q(y)\right) + p_{c_2}D_{KL}\left(p(y|c_2), q(y)\right) \\ &= p_{c_1}\int p(y|c_1)\log\frac{p(y|c_1)}{q(y)}dy + p_{c_2}\int p(y|c_2)\log\frac{p(y|c_2)}{q(y)}dy \\ &= \int \sum_{c=c_1,c_2} p(y|c)p_c \log\frac{p(y|c)}{q(y)}dy = I(y, c|c_1 \vee c_2) \ . \end{aligned} \tag{20}$$

LDA then finds (roughly, due to the different symmetrization) the direction that maximizes the sum of pairwise mutual informations between classes, weighted by the class proportions. This suggests the natural extension to consider more than just pairwise class interactions, and maximize the complete mutual information $I(c, y)$ between classes and projected data. It can be shown that as the amount of data grows, the likelihood objective of RCA asymptotically equals $I(c, y)$, up to a constant. RCA is then a finite-data implementation of an LDA extension.