Chapter 5
Piecewise Constant Modeling of Sequential Data Using Reversible Jump Markov Chain Monte Carlo

Marko Salmenkivi and Heikki Mannila

Summary

We describe the use of reversible jump Markov chain Monte Carlo (RJMCMC) methods for finding piecewise constant descriptions of sequential data. The method provides posterior distributions on the number of segments in the data and thus gives a much broader view on the potential data than do methods (such as dynamic programming) that aim only at finding a single optimal solution. On the other hand, MCMC methods can be more difficult to implement than discrete optimization techniques, and monitoring convergence of the simulations is not trivial. We illustrate the methods by modeling the GC content and distribution of occurrences of ORFs and SNPs along the human genomes. We show how the simple models can be extended by modeling the influence of GC content on the intensity of ORF occurrence.

5.1 Introduction

Sequential data occur frequently in biological applications. At least three different types of sequential data can be distinguished: strings, sequences of events, and time series. A string is simply a sequence of symbols from some alphabet Σ (typically Σ is assumed to be finite). In genomic applications, the alphabet is typically the four-letter DNA alphabet. A sequence of events over alphabet Σ is a collection of pairs (e, t), where $e \in \Sigma$ is the event and t is the occurrence time (or position) of the event. As an example, if we are interested in the occurrences of certain specific words $w_1, \ldots, w_k$ in the genome, we can model the occurrences as a sequence of events consisting of pairs (w_i, t), where t is the position in the sequence of w_i. A time series consists also of pairs (e, t), where t is the occurrence, or measurement time,

but e is a possibly many-dimensional value. For example, we can consider the frequency of all two-letter words in overlapping windows of some length in the genome to obtain a time series with dimension 16.

The process that creates the sequential data can often be assumed to have several hidden states. For example, a genomic sequence could contain segments stemming from different sources. This leads to the question of verifying whether there are different segments, and if there are, finding the change points between the segments.

A natural way of modeling the sources and transitions between them is to use *piecewise constant functions*. Change points of a piecewise constant function can be interpreted as modeling the transitions between hidden sources. Function values in each piece correspond to the relatively stable behavior between the transitions.

In the case of time series data, a common choice for modeling is to use some function $\alpha(t)$, which determines the value of the time series at time t, except for random error. The error is assumed to be normally distributed with zero mean, which leads to the loglikelihood being proportional to the sum of squared distances between the observations and the model predictions. In piecewise representations of $\alpha(t)$, we obtain the total loglikelihood as the sum of the loglikelihoods in each piece.

As we use piecewise constant functions, the function $\alpha(t)$ has the following form:

$$\alpha(t) = \begin{cases} \alpha_1 \text{ if } S_s \leq t < c_1 \\ \alpha_2 \text{ if } c_1 \leq t < c_2 \\ \vdots \quad \vdots \\ \alpha_i \text{ if } c_{i-1} \leq t \leq S_e \\ 0 \quad \text{elsewhere} \end{cases}$$

Here $\{S_s, S_e\} \in \mathbf{R}$ are the start and end points of the sequence, the values $\{\alpha_1, ..., \alpha_i\} \in \mathbf{R}^+$ are the function values in i pieces, and $\{c_1, ..., c_{i-1}\} \in [S_s, S_e]$ are the *change points* of the function.

Figure 5.1 shows an example of a piecewise constant description $\alpha(t)$ of a time series. Measurements are indicated by the values at positions $t_1, \ldots t_7$, and they are illustrated by the filled bars.

Dynamic programming methods can be used to find the best-fitting piecewise constant function in time $\mathcal{O}(n^2k)$, for n observations and k segments [36, 268]. The problem with the dynamic programming methods is that, as maximum likelihood methods, they always provide a segmentation with a given number of segments, whether the data support one or not. This can lead to spurious or downright misleading results, unless care is taken to control carefully for the significance of the output.

In this chapter we give an introduction to Bayesian modeling of sequential data and the reversible jump Markov chain Monte Carlo