Winkler, Kempe, Liebscher, Wittich:

Parsimonious Segmentation of Time Series' by Potts Models

Sonderforschungsbereich 386, Paper 348 (2003)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Parsimonious Segmentation of Time Series' by Potts Models

Gerhard Winkler[1], Angela Kempe[2], Volkmar Liebscher[1], and Olaf Wittich[1]

[1] Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
D-85758 Neuherberg/München, Germany
[2] Graduate Programme Applied Algorithmic Mathematics
Center for Mathematical Sciences, Munich University of Technology
D-85747 Garching

Abstract. Typical problems in the analysis of data sets like time-series or images crucially rely on the extraction of primitive features based on segmentation. Variational approaches are a popular and convenient framework in which such problems can be studied. We focus on Potts models as simple nontrivial instances. The discussion proceeds along two data sets from brain mapping and functional genomics.

1 Introduction

The purpose of the present note is twofold: We want to give an elementary introduction to variational approaches to the analysis of one- and multi-dimensional data, and further to illustrate by way of simple data sets and statistical models what we mean by parsimonious statistics.

We will briefly discuss a particularly simple parsimonious approach to the statistical analysis of real-world data sets from life-sciences. Frequently, there is little or no ground truth, and the stochastic mechanism generating (noisy) data is essentially unknown. The only way to associate data to some hidden real event is to verify or falsify rough and basic criteria which characterize the event in question. Such criteria frequently are based on primitive signal features. In images these may be boundaries between regions of different intensity or texture, in time series they may be morphological features like modes or 'ups and downs', domains of monotony, or plateaus where the signal is constant. We start the discussion with two one-dimensional data sets, one from brain mapping and one from functional genomics. We expect that in these examples the observation period can be partitioned into intervals where the underlying signal can reasonably be represented by a constant. This is a primitive morphological feature, and the resulting step functions allow sound biological interpretations.

To extract piecewise constant 'regressions' from data, we adopt the simplest variational approach based on the Potts model. It is well known to physicists as the straightforward generalization of the Ising model for binary spins to multiple states. For a detailed discussion see Winkler (2003).

2 Two Data Sets from Life Sciences

In order to introduce and illustrate the concept, we present two sets of data. The first one consists of time series from functional magnetic resonance imaging (fMRI) of the human brain, and the second one of melting or fractionation curves for spots on a cDNA microchip.

Example 1 (fMRI Brain Data: Identification of Response Regions). The final aim is to identify regions of increased activity in the human brain in response to outer stimuli. Typically such stimuli are boxcar shaped as indicated in Fig. 1. They may represent 'light or sound on and off', i.e. visual or acoustic stimuli, or tactile ones like finger tipping on a desk. Functional magnetic resonance imaging (fMRI) exploits the BOLD effect which basically is a change of paramagnetic properties caused by an increase of blood flow in response to the demand of activated neurons for more oxygen. The degradation mecha-

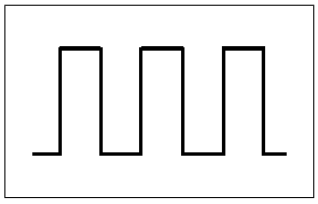

Fig. 1. A box car shaped signal representing 'on-off' stimuli in fMRI brain mapping.

nism along the path '(complex) eye - (highly complex) brain - (complicated) measuring device' is only partially known. Moreover, measurement is indirect, since the recorded BOLD effect is a physiological quantity related to increase of blood flow and not a direct function of cortical activation. Hence a parsimonious approach based on significant plateaus should be appropriate.

Example 2 (Fractionation Curves from Gene Expression). The aim of this experiment is to explore the structure of unknown genes. To this end, single stranded sections of *known* cDNA are put on spots of microchips, which typically consist of about 20.000 spots. Each section is a finite sequence of four nucleic acids, which are coded by the letters A(denin), C(ytosin), G(uanin), and T(hymin). If further nucleic acids are added then they tend to bind to the known nucleic acids where T binds to A, and G binds to C.
Hence sections of single stranded *unknown* cDNA tend to pair with DNA of similar sequence. The binding energy is maximal for perfect matches like

A C T A C A G T A C C C A
T G A T G T C A T G G G T

and such a perfect match means high stability. With perfect match the unknown sequence could be identified perfectly. A main problem is *cross-hybridisation*, which means that DNA sections pair with DNA of similar - but not precisely equal - sequence, for example

A C T A C A G T A C C C A
T G A T T T C A T G A G T

Perfect match and mismatch are illustrated in Figure 2. A new and innovative experiment provides data which hopefully will allow to identify mismatch dissociation at low stringency. It is called 'Specificity Assessment From Frac-

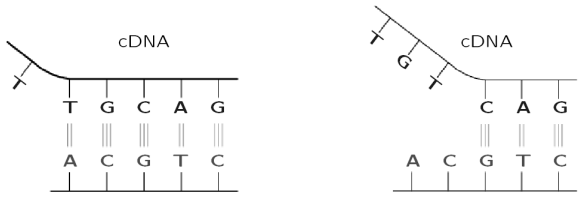

Fig. 2. Specific and unspecific hybridization

tionation Experiments' or in short-hand notation 'SAFE', see Drobyshev et al.(2003). It is plausible that 'the melting temperature' of double stranded DNA depends on length and contents of specific sequences. It is also plausible that increasing temperature has similar effects as increasing washing stringencies with *formamide* solutions. Both decrease the binding energies and thus cross-hybridisation. This is the basis for the measurement of specific and cross-hybridisation. In the experiment, the chips are washed repeatedly (29 times) with formamide solutions of increasing concentration, and fractionation curves like in Fig. 3 are recorded. The aim of the statistical analysis is

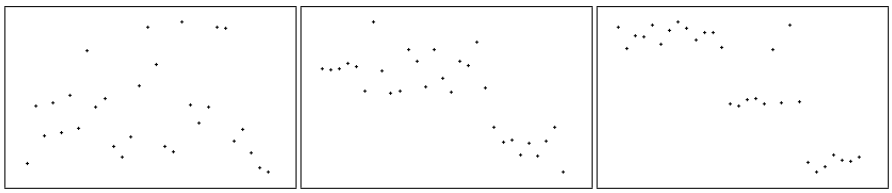

Fig. 3. Typical fractionation curves of single spots: lousy, fairly good, intermediate. (Data from Drobyshev et al. (2003))

to identify locations and heights of abrupt decreases, since they indicate that a certain type of cross-hybridizing cDNA was washed away.

In view of such data, one may doubt about too 'specific' methods or too detailed models for their analysis, and in fact we do so. A way out of this misery is to try a parsimonious approach as indicated above, see Davies (1995). There are attempts like in Davies and Kovac (2001), who adopt the taut string algorithm and its relatives for certain types of data. We tried a parsimonious variational approach called the *Potts Model* for our data.

3 The Potts Model: Rigorous Results

The relevant features in Example 1 are successions of high and low plateaus, and in Example 2 the positions of rapid decreases and their height. Therefore we try to fit piecewise constant functions to our data.

The *Potts functional* is defined by

$$x = (x_1, \ldots, x_n) \longmapsto P_{\gamma,y}(x) = \gamma |J(x)| + \sum_{k=1}^{n} \big(x_k - y_k\big)^2, \tag{1}$$

where $y = (y_1, \ldots, y_n)$ denotes real (fixed) data, and $J(x)$ is the set of time points k where $x_k \neq x_{k+1}$, $k = 1, \ldots, n-1$. $|J(x)|$ denotes the cardinality of $J(x)$. The second term rates fidelity of the signal x to data y, and the first one penalizes undesired properties of x. Thus the functional is a penalized likelihood function.

A minute of contemplation reveals that there are three elementary concepts combined in this model:

(i) A notion of a 'jump' or 'break': In the Potts model such a jump is present, where the values of the signal x in two subsequent time points differ from each other.

(ii) A notion of smoothness: this concerns the behaviour of the signal between two subsequent jumps. It is a consequence of *(i)* that in the Potts model a signal is constant there.

(iii) A notion of fidelity to data, i.e. some measure of distance between data y and the signal x.

Note that *(ii)* is a rather strict notion of smoothness: the signal on a discrete interval is 'smooth' only if it is constant. The first term penalizes the number of jumps irrespective of their size and the parameter $\gamma > 0$ controls the degree of smoothness.

Given data y, a 'filter output' $T_\gamma(y)$ is defined as a signal which minimizes $P_{\gamma,y}$. In general, it is not unique, but fortunately the following result from the forthcoming thesis Kempe (2003) guarantees uniqueness almost surely:

Theorem 1. *Suppose that the law of data y admits a Lebesgue density. Then for almost all y the functional $P_{\gamma,y}$ in (1) has one and only one minimizer.*

If the hypotheses hold a *filter* is defined uniquely for almost all y by the signal

$$T_\gamma(y) = \operatorname*{argmin}_x P_{\gamma,y}(x).$$

We are going now to report essential properties of the filter. It is crucial that the range of hyperparameters γ can be partitioned into intervals, on which the estimate does not change as shown in Kempe (2003). Dependence on hyperparameters is illustrated in Fig. 4.

Theorem 2. *For almost all data y the following is true: There are an integer k and hyperparameters $\infty = \gamma_0 > \gamma_1 > \cdots > \gamma_k > \gamma_{k+1} = 0$ such that $T_\gamma(y)$ is unique for all $\gamma \in (\gamma_{j+1}, \gamma_j)$. Moreover, it is the same time-series for all $\gamma \in (\gamma_{j+1}, \gamma_j)$. $T_\gamma(y)$ is a constant signal for each $\gamma > \gamma_1$, and $T_\gamma(y) = y$ for $\gamma < \gamma_k$. The number of jumps of $T_\gamma(y)$ on the intervals (γ_{j+1}, γ_j) increases in j. For each $0 < i \leq k$, the functional $P_{\gamma_i, y}$ has precisely the two minimizers belonging to the γ-intervals adjacent to γ_i.*

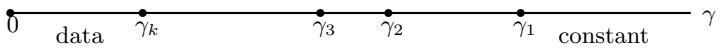

Fig. 4. Intervals on which estimates do not change.

Significant simulations can only be carried out with an exact algorithm for the computation of the minimizer. This rules out stochastic algorithms like simulated annealing. For one dimension, an algorithm based on ideas from dynamic programming was presented in Winkler and Liebscher (2002).

Theorem 3. *There is an algorithm to compute a minimizer of $P_{\gamma,y}$ in time complexity $O(n^3)$ for all $\gamma \in \mathbb{R}$ simultaneously.*

The filter has some more pleasant properties. In particular, the iteration of the filter stops after one step More precisely, a repeated application returns the same signal as a single one, or in other words the filter is *idempotent* in the sense that $T_\gamma \circ T_\gamma = T_\gamma$. This implies that T_γ is a morphological filter in the sense of Serra (1982, 1988), see Winkler and Liebscher (2002):

Theorem 4. *The Potts filter is idempotent.*

The filter has continuity or consistency properties like the following one:

Theorem 5 (V. Liebscher, O. Wittich, unpublished). *Let $\gamma \in \mathbb{R}$ and $y^\infty \in \mathbb{R}^n$. Suppose that $T_\gamma(y^\infty)$ is unique. If y^∞ is degraded by random noise ε_n according to $Y^n = y^\infty + \varepsilon_n$, and noise ε_n tends to zero in probability, then $T_\gamma(Y^n)$ tends to $T_\gamma(y^\infty)$ in probability.*

We are not interested in a 'restoration' but in feature extraction. This is reflected by the theorem since we recover $T_\gamma(y^\infty)$ - and not y^∞ - in the limit.

4 Back to Data

Scanning the filter outputs $T_\gamma(y)$ along the hyperparameter γ in view of Theorem 2, is illustrated in the Figures 5 and 6 for the brain and gene data. Visual inspection of the plots reveals clearly what the desired outputs of the method are. On the other hand, we did not yet find an overall satisfying unsupervised method for the identification of an appropriate hyperparameter γ.

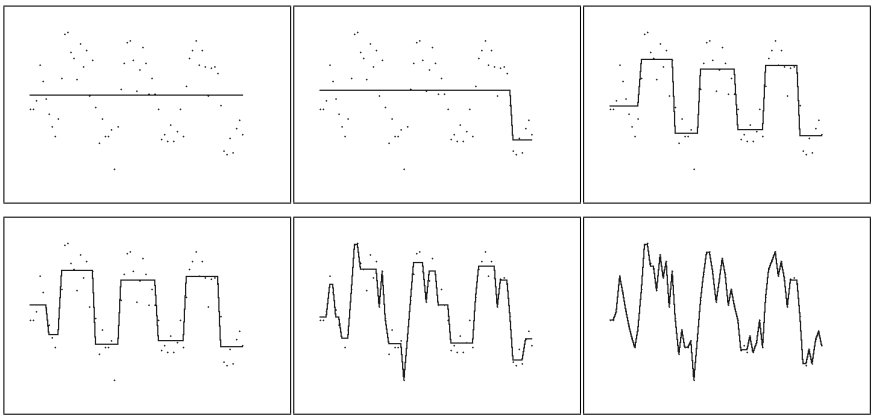

Fig. 5. Some steps of a scan through $T_\gamma(y)$ along decreasing hyperparameters γ for fMRI brain data. Dots indicate data y. Upper right is desired. (Data from D. Auer)

This problem is crucial for such and many similar approaches. For example, there is a host of model selection criteria like the classical ones from Akaike (1974) and Schwarz (1978). Cavanaugh (1997) develops exact criteria which can be adapted to our case, see Kempe (2003). Unfortunately, estimators based on these methods basically return data, as shown in Fig. 7. Therefore we watched out for a criterion which is stable under changes of the hyperparameter γ and of data y. Our first naive idea was to choose $T_{\gamma^*}(y)$ with γ^* from the longest interval of γ-values according to Theorem 2. For the brain data, this simple method outruled the classical criteria. Its results for brain data are contrasted to the classical criteria in Fig. 7.

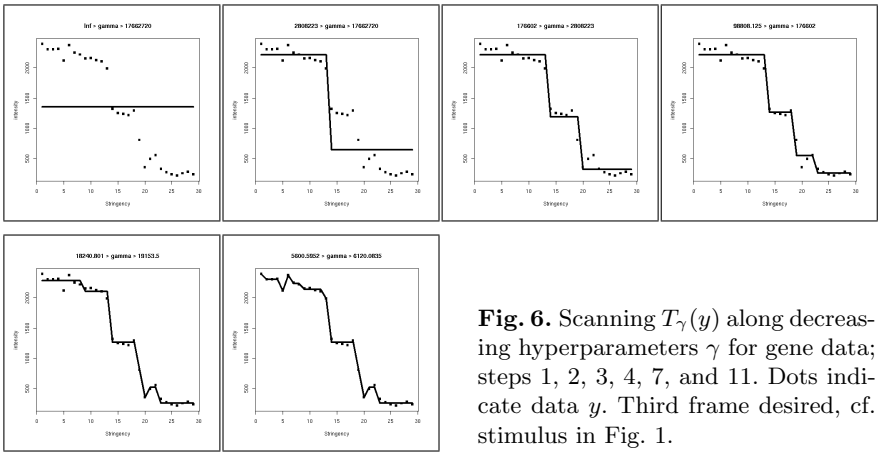

Fig. 6. Scanning $T_\gamma(y)$ along decreasing hyperparameters γ for gene data; steps 1, 2, 3, 4, 7, and 11. Dots indicate data y. Third frame desired, cf. stimulus in Fig. 1.

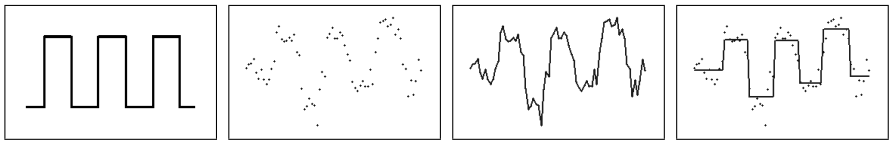

Fig. 7. Stimulus, data, $T_\gamma(y)$ for hyperparameter from Akaike's and Schwarz' information criterion and longest interval criterion: brain data. The latter gives a decent estimate whereas the former basically return data.

For the gene data, we have the additional restriction that the 'true' signal should be decreasing. Therefore, we modified our strategy to choose γ from the leftmost γ-interval on which $T_\gamma(y)$ decreases. Fig. 8 displays some of these estimates.

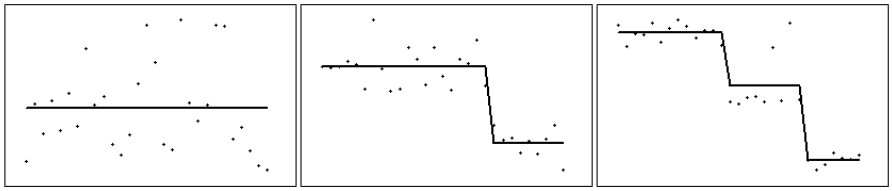

Fig. 8. Estimate for leftmost γ-interval with decreasing $T_\gamma(y)$: gene data.

We obtained partial results about such 'estimators' of intervals, but a satisfying rigorous justification is still missing. This is work in progress.

5 Summary and Outlook

The Potts functional discussed above is a simple instance of a family of similar functionals. There are modified and more complicated penalties, or other data terms, for example with the sum of squares replaced by absolute deviations. The functionals may live on signals with discrete or continuous time.

For example, the Blake-Zisserman functional

$$x \longmapsto BZ(x) = \sum_{k=1}^{n-1} \min\{(x_{k+1} - x_k)^2/\mu^2, \nu\} + \sum_{k=1}^{n} \big(x_k - y_k\big)^2$$

considers a deviation $\delta = |x_{i+1} - x_i|$ as jump if $\delta > \nu^{1/2}\mu$. Between subsequent jumps it returns a signal which is smooth in the L^2-sense. A comprehensive treatment is Blake and Zisserman (1987). This functional was constructed as a discrete approximation of the Mumford-Shah functional

$$E_{\mu,\nu,g}(f_S) := \nu|S| + \sum_{k=1}^{n+1} \int_{J_k(S)} \big(f_k(x) - g(x)\big)^2 + \frac{1}{\mu^2} \, |f_k'(x)|^2 \, dx, \qquad (2)$$

where time varies over a continuous time interval and the functions f are combined of pieces from functions in Sobolev spaces; it was introduced in Mumford and Shah (1989). A discussion of such functionals in the spirit of the above considerations is work in progress.

ACKNOWLEDGEMENT. All simulations were performed by means of the software package ANTS IN FIELDS developed by Friedrich (2003a) in cooperation with the University of Heidelberg. The CD-ROM is attached to Winkler (2003). For a free download see Friedrich (2003b). We are also indebted to J. Beckers, A. Drobyshev, and D. Auer for providing data and introducing us to their subjects.

References

AKAIKE, H. (1974):A new look at the statistical model idenfication. *IEEE Transactions on Automatic Control, 19(6):716–723.*

BLAKE, A. and ZISSERMAN, A. (1987): *Visual Reconstruction.* The MIT Press Series in Artificial Intelligence, MIT Press, Massachusetts, USA.

CAVANAUGH, J.E. (1997): Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters, 33:201–208.*

DAVIES, P.L. (1995): Data features. *J. of the Netherlands Society for Statistics and Operations Research, 49:2,185–245.*

DAVIES, P.L. and KOVAC, A. (2001): Local extremes, runs, strings and multiresolution. *Ann. Stat., 29, 1–65.*

DROBYSHEV, A. L., MACHKA, CHR., HORSCH, M., SELTMANN, M., LIEBSCHER, V., HRABÉ DE ANGELIS, V. and BECKERS, J. (2003): Specificity assessment from fractionation experiments, (SAFE): a novel method to evaluate microarray probe specificity based on hybridization stringencies. *Nucleic Acids Res. 31:2, 1-10.*

FRIEDRICH, F. (2003a): *Stochastic Simulation and Bayesian Inference for Gibbs fields.* CD-ROM, Springer Verlag, Heidelberg, New York.

FRIEDRICH, F. (2003b): ANTSINFIELDS*: Stochastic simulation and Bayesian inference for Gibbs fields*, URL: http://www.AntsInFields.de.

KEMPE, A. (2003), *Statistical analysis of the Potts model and applications in biomedical imaging*, Thesis, Institute of Biomathematics and Biometry, National Research Center for Environment and Health Munich, Germany.

MUMFORD, D. and SHAH, J. (1989): Optimal approximation by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math. 42, 577-685.*

SCHWARZ, G. (1978): Estimating the dimension of a model. *The Annals of Statistics, 6(2):461–464.*

SERRA, J. (1982, 1988): *Image analysis and mathematical morphology. Vol. I, II.* Acad. Press, London.

WINKLER, G. (2003):. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods. A Mathematical Introduction.* volume 27 of *Applications of Mathematics*, Springer Verlag, Berlin, Heidelberg, New York, second edition.

WINKLER, G. and LIEBSCHER, V. (2002): Smoothers for Discontinuous Signals. *J. Nonpar. Statist. 14:1-2, 203–222.*