Appeared in W. Gaul, O. Opitz, and M. Schader, eds., Data Analysis, Scientific Modeling and Practical Application. Springer, Heidelberg, 2000, pp. 41–50

Classification and clustering of objects with variants

Gunter Ritter[1]

Universität Passau
Fakultät für Mathematik und Informatik
94 030 Passau
ritter@fim.uni-passau.de

Abstract: The method of *variants* has proved a powerful method for reducing the error rate in Bayesian pattern recognition. The method serves to recover from ambiguities often not avoidable during the early stage of processing. Applications of this method to object identification, supervised classification, and clustering are discussed.

Key words: Bayesian pattern recognition, Principle of Least Commitment, method of variants, supervised classification, clustering

1 Introduction

Pattern recognition is the classification of an *object* in one of a number of classes. This is opposed to discriminant analysis or statistical clustering both of which deal with *feature data* rather than with more complex objects. By an object we mean a complex entity that is not further decomposed during the recognition process and is the basis for extraction of *one* set of features. The objects may be optical or acoustical or of another nature, such as textual documents or social and economic agents. Their formal representations may need discrete, continuous or mixed data structures.

Classification of an object is essentially a stepwise process of complexity reduction; it is usually performed in a number of consecutive steps:

- Identification of the *standard representation* of the object.
- Extraction of *features* from the standard representation.
- Supervised or unsuprevised *classification* of the feature set.

Each of these steps means a *coarsening* of the information describing the objects and the transition to a *partition* in the whole population. The first step

[1] Research supported by Deutsche Forschungsgemeinschaft, Ri477/4

is the most critical and at the same time the least understood. It means the application of an equivalence relation inherent to the structure of the objects. As an example, congruent geometric figures are often (but not always, cf. the cyphers 9 and 6) considered as being equivalent. In many cases, the algorithm for feature extraction cannot deal with any member of an equivalence class but needs a standard representation as its input. The rules for producing a standard representation are often *ambiguous*. Here are two examples.

The standard representation of a spoken word is a sequence of phonemes. It is not unambigously clear at the beginning of the recognition process at which point in time one phoneme ends and the next one begins. However, it is very important to identify these points since they allow to extract the features of the phonemes to some precision. A similar problem arises when features are extracted from geometrical objects. Feature extraction from figures such as the triangles in Fig. 1 often requires previous recognition of orientation and shape. It is only then that the standard representation of a

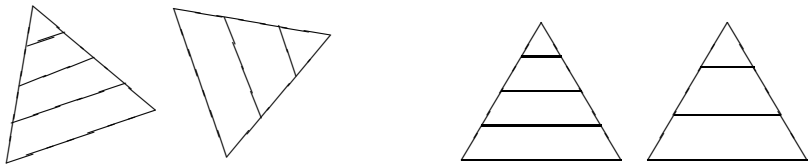

Fig. 1. Two geometric figures (left) and their standard representations (right). Internal features are number and location of stripes; they can be easily determined by vertical cuts if the object is present in its standard representation. A means for producing a standard representation of the object is the analysis of its contour and the dominant points of the contour (vertices in this case). This will, however, yield ambiguous information here since the figures are equilateral triangles.

figure is identified and features can be safely extracted. In both cases, the rules do not describe a *mapping* from object space to representation space, but rather a *relation* between the two spaces. As a cure, one may process *all* reasonable representations. We will deal with the problem of selecting the standard representation and, hence, extracting the correct features, in Sect. 2. The remaining two sections are devoted to supervised classification, Sect. 3, and clustering, Sect. 4, under such ambiguities.

2 Variants as a response to the Principle of Least Commitment

2.1 Variants and variant selection

As explained in the introduction, a necessary prerequisite for feature measurement is the correct *standard representation* of the given object. Obtaining

this representation is often not possible at the early stage of processing since it may need the yet unknown internal features, cf. Fig. 1; it is only at the end of the whole pattern recognition process that the necessary information is available. In other words, we face the vicious circle that feature extraction needs the standard representation and establishing the standard representation may need internal features of the object. A wrong decision in this first step would be disastrous for feature extraction and for the remaining steps since features extracted under wrong assumptions are meaningless. Marr [3] postulates the *Principle of Least Commitment*: never take a decision in a process of recognition or identification unless this decision is safe. Otherwise, it may later have to be undone. A way out of this vicious circle is application of the recently proposed method of *variants* [4], [5]: It processes all reasonable representations (*variants*) of the object taking the decision on the correct ("regular") variant only when sufficient information is accumulated. Here, we assume that there is exactly one regular variant.

Several problems arise from the consideration of variants: The most basic is *selection* of the regular variant from the set of variants. Two others are *classification* of an object into one of several given classes and clustering of objects in the presence of variants of the objects. An application of the Bayesian paradigm immediately shows that all require the *joint* distribution of all variants, cf. Eq. (1). This, however, is often unknown. Therefore, in [4] and [5], algorithms for selection and classification were designed which need essentially the statistical model of the regular variant alone.

Variant selection must not be confused with (supervised) *classification*, *hypothesis testing*, or *goodness-of-fit tests*. In some sense, variant selection is even opposed to classification since, in the former case, several observations (variants) compete for one statistical model and in the latter case, several statistical models compete for an observation. A test of hypotheses needs statistical models of both hypotheses and is in this sense similar to classification. Finally, a goodness-of-fit test compares two models with each other, one of them given by realizations.

2.2 The Simple Selector and its optimality

A Bayesian framework for the selection problem is as follows. Let (Ω, P) denote a probability space, let E be a Polish state space with its Borel σ–algebra $\mathcal{B}(E)$, and let $Z_i : (\Omega, P) \to E$, $i \in 1..b$, be $b \geq 1$ variants of an object, Z_1 being the regular one. We observe a realization $\mathbf{x} = (x_1, \cdots, x_b) \in E^b$ of $X = (X_1, \cdots, X_b) = (Z_{T(1)}, \cdots, Z_{T(b)}) = Z_T$, a random permutation $T : \Omega \to \mathcal{S}_b$ of the b-tuple $Z = (Z_1, \cdots, Z_b)$. The task is estimation of the unknown random position $H : \Omega \to 1..b$ of the regular variant, i.e., the position H so that x_H emanates from Z_1. Clearly, we have $T(H) = 1$, $H = T^{-1}(1)$ and the assertions $T(h) = 1$ and $H = h$ are synonymous. We assume that the random permutation T is independent of Z.

The related MAP model is the quadruple $(X, (P_{Z_\pi})_{\pi \in \mathcal{S}_b}, \mathcal{D}, G)$. Here, the parameter set is the symmetric group $\mathcal{S}_b$, the decision set $\mathcal{D}$ is the interval $1..b$, and the gain function $G : \mathcal{S}_b \times (1..b) \to \mathbf{R}$ is

$$G(\pi, h) = \begin{cases} 1, & \text{if} \quad \pi(h) = 1, \\ 0, & \text{otherwise.} \end{cases}$$

For all $h \in 1..b$, let $q_h := P[T(h) = 1]$ denote the prior probability of the regular variant to occupy site h. Without loss of generality we assume $q_h > 0$ for all $h \in 1..b$. We call any estimator $S : E^b \to 1..b$ of the regular variant a *selector*.

2.3 Lemma

Let μ be some σ-finite measure on E^b such that P_{Z_T} is absolutely continuous with respect to μ. The *Bayesian selector* BS for the statistical model at hand is

$$BS(\mathbf{x}) = \text{argmax}_{h \in 1..b} \frac{P[Z_T \in d\mathbf{x}/T(h) = 1]}{\mu(d\mathbf{x})} \, q_h, \tag{1}$$

P_{Z_T}–a.a. $\mathbf{x} \in E^b$.

Proof. The *Bayesian selector* BS for the above statistical model is defined by

$$\begin{aligned} BS(\mathbf{x}) &= \text{argmax}_{h \in 1..b} E[G(T, h)/X = \mathbf{x}] \\ &= \text{argmax}_{h \in 1..b} P[T(h) = 1/Z_T = \mathbf{x}]. \end{aligned}$$

This set is nonempty and uniquely defined for P_{Z_T}-a.a. $\mathbf{x} \in E^b$. By Bayes' formula, $P[T(h) = 1/Z_T = \mathbf{x}]$ equals the density

$$\frac{P[Z_T \in d\mathbf{x}, T(h) = 1]}{P[Z_T \in d\mathbf{x}]} = \frac{P[Z_T \in d\mathbf{x}, T(h) = 1]}{\mu(d\mathbf{x})} \Big/ \frac{P[Z_T \in d\mathbf{x}]}{\mu(d\mathbf{x})}.$$

This implies the claim. □

This selector needs information on the *joint* distribution of *all* variants. There is usually one correct (regular) variant, but there may be many unknown causes for incorrect, spurious variants. In these cases, complete knowledge about all variants, let alone their joint distribution, will not be available. This is one reason why the simple selector was proposed. For its description we need some more notation. Let ρ be some σ–finite reference measure on E and let $f^\rho_{Z_1}$ be the density of the random features $Z_1 \colon \Omega \to E$ of the regular variant. We wish to select the position of the regular variant with the aid of $f^\rho_{Z_1}$ and q_h. It is tempting to choose an index h for which the quantity $f^\rho_{Z_1}(x_h)\, q_h$ is maximal, i.e., to use the

Simple (Variant) Selector associated with the reference measure ρ, [4], [5]:

$$SS^{\rho}(x_1, \cdots, x_b) = \text{argmax}_{h \in 1..b} f^{\rho}_{Z_1}(x_h)\, q_h.$$

Note that the Simple Selector depends on the reference measure ρ. This is contrary to the Bayesian selector (1) which does not depend on the measure μ on E^b. It is easy to see that the set $\{h \in 1..b / f^{\rho}_{Z_1}(x_h)\, q_h > 0\}$ is nonempty for P_{Z_T}-a.a. $\mathbf{x} \in E^b$. Therefore, $\max_{h \in 1..b} f^{\rho}_{Z_1}(x_h)\, q_h$ is strictly positive for P_{Z_T}–a.a. $\mathbf{x} \in E^b$.

The Simple Selector may, however, be grossly misleading. Nevertheless, one can give conditions which guarantee its optimality, i.e., conditions ensuring that the Bayesian selector BS just depends on the reduced set of quantities. Moreover, there do exist interesting situations where these conditions are satisfied. The following general sufficient condition for optimality of the Simple Selector appears in [5]. We need some preliminaries. Conditioning on the event $Z_1 = x$ is defined for P_{Z_1}-a.a. $x \in E$, only. However, since E is Polish, the conditional distribution $P[Z_{\hat{1}} \in d\mathbf{y} / Z_1 = x]$ may be extended to a Markovian kernel K on $E \times E^{b-1}$ such that $K(x, d\mathbf{y}) = P[Z_{\hat{1}} \in d\mathbf{y} / Z_1 = x]$ for P_{Z_1}–a.a. $x \in E$. Here, $\mathbf{x}_{\hat{h}}$ stands for $(x_1, \cdots, x_{h-1}, x_{h+1}, \cdots, x_b) \in E^{b-1}$, $\mathbf{x} = (x_1, \cdots, x_b) \in E^b$, $h \in 1..b$. Let us say that a selector is optimal if it equals the Bayesian Selector for P_{Z_T}–a.a. $\mathbf{x} \in E^b$.

2.4 Theorem

If P_{Z_1} is absolutely continuous with respect to ρ and if $\rho \otimes K$ is exchangeable then the Simple Selector SS^{ρ} is optimal. Moreover, $f^{\rho}_{Z_1} = dP_Z / d(\rho \otimes K)$.

Proof. First, exchangeability of $\rho \otimes K$ implies exchangeability of $K(x, \cdot)$ for ρ–a.a. $x \in E$. Hence, we have for all h and $\pi \in \mathcal{S}_b$ such that $\pi h = 1$

$$\begin{aligned} &P[Z_{\pi} \in d\mathbf{x}] = P[Z_{\pi h} \in dx_h, Z_{\pi \hat{h}} \in d\mathbf{x}_{\hat{h}}] = P[Z_1 \in dx_h, Z_{\hat{1}} \in d\mathbf{x}_{\pi^{-1}\hat{1}}] \\ &= P_{Z_1}(dx_h) K(x_h, d\mathbf{x}_{\pi^{-1}\hat{1}}) = P_{Z_1}(dx_h) K(x_h, d\mathbf{x}_{\hat{h}}) \\ &= P_{Z_1} \otimes K(dx_h, d\mathbf{x}_{\hat{h}}). \end{aligned}$$

From statistical independence of Z and T it follows

$$P[Z_T \in d\mathbf{x} / Th = 1] = P_{Z_1} \otimes K(dx_h, d\mathbf{x}_{\hat{h}}). \tag{2}$$

Using again exchangeability of $\rho \otimes K$, we obtain

$$\frac{P[Z_T \in d\mathbf{x} / Th = 1]}{\rho \otimes K(d\mathbf{x})} = \frac{P_{Z_1} \otimes K(dx_h, d\mathbf{x}_{\hat{h}})}{\rho \otimes K(dx_h, d\mathbf{x}_{\hat{h}})} = f^{\rho}_{Z_1}(x_h)$$

for $\rho \otimes K$–a.a. and, hence, P_{Z_T}–a.a $\mathbf{x} \in E^b$ and the first claim follows from (1).

The second claim follows from the definition of K:

$$P_Z(d\mathbf{x}) = P[Z_{\hat{1}} \in d\mathbf{x}_{\hat{1}}/Z_1 = x_1]P_{Z_1}(dx_1) = P_{Z_1} \otimes K(d\mathbf{x})$$
$$= f^{\rho}_{Z_1}(x_1)(\rho \otimes K)(d\mathbf{x}). \qquad \square$$

This theorem entails a number of corollaries, cf. [4], [5].

2.5 Corollary

Let all b variants be independent, let $Z_2, \cdots, Z_b$ be identically distributed, and suppose that P_{Z_1} is absolutely continuous with respect to P_{Z_2}. Then the Simple Selector SS^{ρ} with $\rho = P_{Z_2}$ is optimal.

Proof. We may put $K(x, \cdot) := (P_{Z_2})^{\otimes(b-1)}$, $x \in E$. The claim follows from Theorem 2.4 since $\rho \otimes K$ is a product measure with equal factors and, hence, exchangeable. $\square$

2.6 Corollary

Let $b = 2$ and assume that the Markovian kernel K has the reversible measure ρ. If P_{Z_1} is absolutely continuous with respect to ρ then the Simple Selector SS^{ρ} is optimal.

Proof. Indeed, if $b = 2$ then exchangeability of $\rho \otimes K$ is just reversibility of ρ. $\square$

Any measurable function $\varphi : E \to E$ induces the Markov kernel $K(x, \cdot) = \delta_{\varphi(x)}$ on $E \times E$. If φ is involutive then $\mu + \mu_{\varphi}$ is φ–invariant and, hence, reversible with respect to any nonzero measure μ on E. In this case, we have $Z_2 = \varphi(Z_1)$ and the elements in $E \times E$ *observable* for X are of the form $(x, \varphi(x))$. There is the following corollary.

2.7 Corollary

Let $b = 2$, let $Z_2 = \varphi(Z_1)$ for some measurable involution $\varphi : E \to E$, and let ρ be a φ–invariant reference measure such that P_{Z_1} is absolutely continuous with respect to ρ. Then the Simple Selector SS^{ρ} is optimal.

3 Supervised classification with variants

Variant selection can be combined with (supervised) classification. Suppose that an object having several variants $x_1, \cdots, x_b$ is to be assigned to one of several given classes $j \in 1..n$. Let the regular variant $Z_{j,1}$ of class j be

absolutely continuous relative to some reference measure ρ on E and let the density function be designated by $f^{\rho}_{Z_{j,1}}$. Let $q_{j,h}$ be the prior probability for j to be the correct class and for position h to be that of the regular variant. There are two related problems, namely optimal assignment of the object to its class *with* and *without* simultaneous selection of the regular variant. The following classifier, designed for the former case, is due to M.T. Gallegos and the author.

Simple Classifier–Selector. The *Simple Classifier–Selector* associated with the reference measure ρ is defined as

$$SC^{\rho}(\mathbf{x}) = \mathrm{argmax}_j \max_{h \in 1..b} f^{\rho}_{Z_{j,1}}(x_h)\, q_{j,h}, \quad \mathbf{x} = (x_1, \ldots, x_b), \tag{3}$$

and the estimate of the position of the regular variant is the maximal $h \in 1..b$.

The Simple Classifier–Selector uses all variants of the object and decides at the same time on the class and the regular variant. The point is that the Simple Classifier–Selector needs the densities $f^{\rho}_{Z_{j,1}}$ of the *regular* variants $Z_{j,1}$ of all classes j, only, and not the densities of the irregular variants for estimating the class of $\mathbf{x}$. Yet, it can be shown that it equals the Bayesian estimator given the whole statistical information if the joint distribution of all variants satisfies one of the the conditions stated in Corollaries 2.5–2.7. If the class, only, is to be estimated then maximization over $1 \leq h \leq b$ in (3) is replaced with summation. The resulting classifier is called the *Simple Classifier*. It implicitly uses the regular variant for classification.

3.1 Applications of the Simple Classifier

Recently, a constrained version of the Simple Cassifier was successfully applied in various contexts to the "automatic classification of chromosomes" [6–8]. Feature extraction from the oblong metaphase chromosomes under a light microscope needs their correct polarities. These are not a priori given, a situation giving rise to considering two variants for each chromosome, one feature set for each polarity. After collecting information at a higher level, the Simple Classifier implicitly selects the most prospective of them using it as its basis for classification. The resulting "polarity free" classification method reduces the error rate by about 25% [6].

Some methods of feature measurement on chromosomes require the extraction of *longitudinal axes* along the chromosomes. These define suitable standard representations in the sense of the introduction. In the case of a severely bent, badly shaped, or small chromosome the axis (and, hence, the shape) is not easily determined and a way of handling this ambiguity is the simultaneous consideration of various possible axes [7], [8]. Variants thus help to attain the presently worldwide lowest error rate of 0.6% in this field. Applications to automatic image, document, and speech processing also lend themselves.

4 Clustering with variants

Clustering differs from supervised classification in that the class conditional distributions are unknown. Like supervised classification, clustering of objects can be treated in the presence of variants as well.

4.1 Explanation and Notation

Let

$$X_1 = (X_{1,1}, \ldots, X_{1,b}), \ldots, X_m = (X_{m,1}, \ldots, X_{m,b})$$

be m objects to be clustered in an a priori given number of classes, each object being observed by way of $b \geq 1$ variants. A statistical model of this situation uses the following table of $n \cdot m$ random variables $Z_j^{(i)} = (Z_{j,1}^{(i)}, \ldots, Z_{j,b}^{(i)})$: $\Omega \to E^b$, $i \in 1..m$, $j \in 1..n$.

$$\begin{array}{cccc}
(Z_{1,1}^{(1)}, \ldots, Z_{1,b}^{(1)}), & \ldots, & (Z_{1,1}^{(i)}, \ldots, Z_{1,b}^{(i)}), & \ldots, (Z_{1,1}^{(m)}, \ldots, Z_{1,b}^{(m)}), \\
\vdots & \vdots & \vdots & \vdots \\
(Z_{j,1}^{(1)}, \ldots, Z_{j,b}^{(1)}), & \ldots, & (Z_{j,1}^{(i)}, \ldots, Z_{j,b}^{(i)}), & \ldots, (Z_{j,1}^{(m)}, \ldots, Z_{j,b}^{(m)}), \\
\vdots & \vdots & \vdots & \vdots \\
(Z_{n,1}^{(1)}, \ldots, Z_{n,b}^{(1)}), & \ldots, & (Z_{n,1}^{(i)}, \ldots, Z_{n,b}^{(i)}), & \ldots, (Z_{n,1}^{(m)}, \ldots, Z_{n,b}^{(m)}).
\end{array}$$

Each of the m joint random variables $Z_j^{(i)}$ in the jth line represents the jth class; their distributions are equal. For all $i \in 1..m$, $Z_{j,1}^{(i)}$ represents the regular and $Z_{j,k}^{(i)}$ the kth variant of the generic object of class $j \in 1..n$. Let $L\colon \Omega \to (1..n)^m$ stand for an unknown assignment of the m objects to the n classes and let $T_i\colon \Omega \to \mathcal{S}_b$ be an unknown permutation of the b variants of object $i \in 1..m$. We may assume that the random variables L, $T_1, \ldots, T_m$, and all $Z_j^{(i)}$ are independent of each other. Variants of *one* object may (and will in general) be statistically dependent. Writing $Z_{j,\pi}^{(i)} = (Z_{j,\pi 1}^{(i)}, \ldots, Z_{j,\pi b}^{(i)})$ for π in the symmetric group $\mathcal{S}_b$, we observe the random choice $X_1 = Z_{L(1),T_1}^{(1)}, \ldots, X_m = Z_{L(m),T_m}^{(m)}$ in the above table; i.e., object X_i is the entry in line $L(i)$ and column i, randomly permuted according to $T_i\colon \Omega \to \mathcal{S}_b$.

Besides the clustering, we wish to estimate the regular variant of each object. This amounts to estimating the labels $L(i)$ and the sites $H_i = T_i^{-1}1$, $i \in 1..m$. The case $b = 1$ corresponds to the classical case, cf. H.H. Bock [1], [2]; here, the maximum likelihood paradigm is a popular method of estimation. One chooses a parametric model with parameter set Θ for the class-conditional distributions, a suitable reference measure ρ on E and defines the densities $f^\rho(\theta, x)$, $\theta \in \Theta$, $x \in E$. With the abbreviation $X = (X_1, \ldots, X_m)$,

the ML–estimate of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \Theta^n$ and $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_m) \in (1..n)^m$ given the observation $\mathbf{x} = (x^{(1)}, \ldots, x^{(m)}) \in E^m$ is

$$\mathrm{argmax}_{\boldsymbol{\theta},\boldsymbol{\ell}} \frac{P^{\boldsymbol{\theta}}[X \in d\mathbf{x}/L = \boldsymbol{\ell}]}{\rho(dx^{(1)}) \ldots \rho(dx^{(m)})} = \mathrm{argmax}_{\boldsymbol{\theta},\boldsymbol{\ell}} \prod_{j=1}^{n} \prod_{i:\ell_i=j} f^{\rho}(\theta_j, x^{(i)}).$$

In the presence of variants, we propose a mixture of an ML–estimator for the distributional parameters and the cluster assignment and an MAP–estimator for the positions $\mathbf{h} \in (1..b)^m$ of the regular variants. With the notation $T = (T_1, \ldots, T_m)$, the relative clustering criterion is, thus,

$$\mathrm{argmax}_{\boldsymbol{\theta},\boldsymbol{\ell},\mathbf{h}} \frac{P^{\boldsymbol{\theta}}[X \in d\mathbf{x}, T\mathbf{h} = \mathbf{1}/L = \boldsymbol{\ell}]}{\nu(d\mathbf{x}^{(1)}) \ldots \nu(d\mathbf{x}^{(m)})}; \tag{4}$$

here, ν is some reference measure on E^b and $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}) \in E^{mb}$. The symbol $\mathbf{1}$ stands for the constant vector $(1, \ldots, 1)$.

Simple Clustering Criterion. Let us define the *Simple Clustering Criterion* associated with the reference measure ρ on E as

$$\mathrm{argmax}_{\boldsymbol{\theta},\boldsymbol{\ell},\mathbf{h}} \prod_{j=1}^{n} \prod_{i:\ell_i=j} \left(f^{\rho}(\theta_j, x_{h_i}^{(i)})\ q_{h_i}^{(i)} \right). \tag{5}$$

There is the following theorem on its optimality; the condition (6) appearing there means that the mechanism generating the irregular variants from the regular ones is class independent.

4.2 Theorem

Let there be a Markov kernel K on $E \times E^{b-1}$ such that

$$P^{\theta}[Z_{j,\hat{1}} \in d\mathbf{y}/Z_{j,1} = x] = K(x, d\mathbf{y}) \tag{6}$$

for all $j \in 1..n$, $\theta \in \Theta$, and $P_{Z_{j,1}}$–a.a. $x \in E$. Assume that there exists a σ–finite measure ρ on E such that $P_{Z_{j,1}}$ is absolutely continuous with respect to ρ for all j and such that the product $\rho \otimes K$ is exchangeable on E^b. Then the Simple Clustering Criterion (5) equals the clustering criterion (4) for $(\rho \otimes K)^{\otimes m}$–a.a. $\mathbf{x} \in E^{mb}$.

Proof. Let us first compute

$$\begin{aligned}
&P^{\boldsymbol{\theta}}[X \in d\mathbf{x}/L = \boldsymbol{\ell}, T\mathbf{h} = \mathbf{1}] \\
&= P^{\boldsymbol{\theta}}[X_1 \in d\mathbf{x}^{(1)}, \ldots, X_m \in d\mathbf{x}^{(m)}/L = \boldsymbol{\ell}, T\mathbf{h} = \mathbf{1}] \\
&= P^{\boldsymbol{\theta}}[Z^{(1)}_{L(1),T_1} \in d\mathbf{x}^{(1)}, \ldots, Z^{(m)}_{L(m),T_m} \in d\mathbf{x}^{(m)}/L = \boldsymbol{\ell}, T\mathbf{h} = \mathbf{1}] \\
&= P^{\boldsymbol{\theta}}[Z^{(1)}_{\ell_1,T_1} \in d\mathbf{x}^{(1)}, \ldots, Z^{(m)}_{\ell_m,T_m} \in d\mathbf{x}^{(m)}/L = \boldsymbol{\ell}, T\mathbf{h} = \mathbf{1}] \\
&= P^{\boldsymbol{\theta}}[Z^{(1)}_{\ell_1,T_1} \in d\mathbf{x}^{(1)}, \ldots, Z^{(m)}_{\ell_m,T_m} \in d\mathbf{x}^{(m)}/T_1h_1 = 1, \ldots, T_mh_m = 1] \\
&= P^{\boldsymbol{\theta}}[Z^{(1)}_{\ell_1,T_1} \in d\mathbf{x}^{(1)}/T_1h_1 = 1] \ldots P^{\boldsymbol{\theta}}[Z^{(m)}_{\ell_m,T_m} \in d\mathbf{x}^{(m)}/T_mh_m = 1].
\end{aligned}$$

The assumption of exchangeability, the condition (6), and (2) together imply for all i, j, and $h \in 1..b$ the equality

$$P^{\boldsymbol{\theta}}[Z_{j,T_i} \in d\mathbf{y}/T_i h = 1] = P^{\boldsymbol{\theta}}_{Z_{j,1}} \otimes K(dy_h, d\mathbf{y}_{\hat{h}}), \qquad \mathbf{y} \in E^b.$$

From the above, it follows

$$\begin{aligned} &P^{\boldsymbol{\theta}}[X \in d\mathbf{x}, T\mathbf{h} = \mathbf{1}/L = \boldsymbol{\ell}] \\ &= P^{\boldsymbol{\theta}}_{Z_{\ell_1,1}} \otimes K(dx^{(1)}_{h_1}, d\mathbf{x}^{(1)}_{\widehat{h_1}}) \cdots P^{\boldsymbol{\theta}}_{Z_{\ell_m,1}} \otimes K(dx^{(m)}_{h_m}, d\mathbf{x}^{(m)}_{\widehat{h_m}})\ q^{(1)}_{h_1} \cdots q^{(m)}_{h_m}. \end{aligned}$$

Again by exchangeability, the clustering criterion (4) with $\nu = \rho \otimes K$ now assumes the form

$$\begin{aligned} &\frac{P^{\boldsymbol{\theta}}[X \in d\mathbf{x}, T\mathbf{h} = \mathbf{1}/L = \boldsymbol{\ell}]}{\rho \otimes K(d\mathbf{x}^{(1)}) \cdots \rho \otimes K(d\mathbf{x}^{(m)})} = \prod_{i=1}^{m} \left(\frac{P^{\boldsymbol{\theta}}_{Z_{\ell_i,1}} \otimes K(dx^{(i)}_{h_i}, d\mathbf{x}^{(i)}_{\widehat{h_i}})}{\rho \otimes K(d\mathbf{x}^{(i)})}\ q^{(i)}_{h_i} \right) \\ &= \prod_{i=1}^{m} \left(\frac{P^{\boldsymbol{\theta}}_{Z_{\ell_i,1}} \otimes K(dx^{(i)}_{h_i}, d\mathbf{x}^{(i)}_{\widehat{h_i}})}{\rho \otimes K(dx^{(i)}_{h_i}, d\mathbf{x}^{(i)}_{\widehat{h_i}})}\ q^{(i)}_{h_i} \right) = \prod_{i=1}^{m} \left(\frac{P^{\boldsymbol{\theta}}_{Z_{\ell_i,1}}(dx^{(i)}_{h_i})}{\rho(dx^{(i)}_{h_i})}\ q^{(i)}_{h_i} \right) \\ &= \prod_{i=1}^{m} \left(f^{\rho}(\theta_{\ell_i}, x^{(i)}_{h_i})\ q^{(i)}_{h_i} \right) = \prod_{j=1}^{n} \prod_{i:\ell_i=j} \left(f^{\rho}(\theta_j, x^{(i)}_{h_i})\ q^{(i)}_{h_i} \right). \end{aligned}$$

The third equality in this chain is true for $(\rho \otimes K)^{\otimes m}$–a.a. $\mathbf{x} \in E^{mb}$. This concludes the proof. □

References

1.Bock, H.H., Automatische Klassifikation. Vandenhoek & Ruprecht, Göttingen, 1974

2.Bock, H.H., Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis* 23 (1996) 5–28

3.Marr, D., Vision. Freeman, San Francisco, 1982

4.Ritter, G., M.T. Gallegos, A Bayesian approach to object identification in pattern recognition. In *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, 2000

5.Ritter, G., M.T. Gallegos, Bayesian object identification: variants. To appear in *J. Multivariate Analysis* (2002)

6.Ritter, G., Ch. Pesch, Polarity–free automatic classification of chromosomes. To appear in *Computational Statistics and Data Analysis* (2001)

7.Ritter, G., G. Schreib, Profile and feature extraction from chromosomes. In *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, 2000

8.Ritter, G., G. Schreib, Using dominant points and variants for profile extraction from chromosomes. To appear in *Pattern Recognition* (2001)