ON THE SELECTION OF MARKOV RANDOM FIELD TEXTURE MODELS

CHUANSHU JI AND LYNNE SEYMOUR

University of North Carolina at Chapel Hill

ABSTRACT. The problem of selecting pair-potentials of finite range for Gibbs random fields is considered as an important step in modelling multi-textured images. In a decision theoretic set-up, the Bayesian procedure is approximated by using Laplace's method for asymptotic expansion of integrals. Certain frequentist properties of the selection procedure are investigated. In particular, its consistency is justified regardless of phase transition of the Gibbs random fields.

1. INTRODUCTION

Other than Grenander's fundamental work [cf. Grenander (1989)], the breakthroughs of statistical image analysis in the 1980's are Geman and Geman (1984) and Besag (1986). Their basic approach is to use Gibbs random fields (GRFs) — originally used as models in statistical mechanics — to model the pixel intensities and other relevant scene attributes to capture the tendencies and constraints that charaterize the image of interest. In a Bayesian set-up, GRFs play the role of the prior distribution on possible images, with its mean or mode(s) being regarded as the true picture. Due to degradation, only partial or corrupted observations are taken, and then the corresponding posterior mean or mode(s) based on these observations are taken as the estimates of the true scene. Two well-known algorithms developed in these works, called stochastic relaxation (SR) by Geman and Geman

1980 *Mathematics Subject Classification* (1985 *Revision*). primary 60G60, 62F99; secondary 62M30, 68U10, 82A25.

Key words and phrases. Markov random fields, Gibbs random fields, model selection, BIC, phase transition, image analysis, texture segmentation.

(1984) and iterated conditional mode (ICM) by Besag (1986), are used extensively in various imaging problems, such as image restoration, boundary detection, texture segmentation, positron emission tomography (PET), etc.

Since these algorithms are based on a fully specified energy function for a GRF, specifying the energy function (or equivalently, the potential) at the modelling stage becomes a crucial step. This step may include parameter estimation and model selection: the former is an inference problem of estimating unknown parameters contained in the energy function for the GRF; the latter is a multiple decision problem of choosing an energy function from a finite set of candidates that may induce the GRF as a true model.

The need for parameter estimation and model selection varies in different imaging problems. A typical example for such a need is texture segmentation [cf. Geman and Graffigne (1986)]. This involves estimating the coupling coefficients for each pair of pixels and choosing the neighborhood system for a Markov random field (MRF) — a special case of the GRF induced by potentials of finite range. The ability of a MRF texture model to label the texture type of each pixel in the scene or to segment the picture into several single-textured regions is critically determined by the parameter estimation and model selection.

In parameter estimation, maximum likelihood estimates (MLEs) and maximum pseudo-likelihood estimates (MPLEs) are usually considered. MLEs are computionally intractable in their basic forms due to large dimensionality of image data. Some work has been done to remedy this drawback. MPLEs are much easier to compute and are extensively used. Since only MPLEs are relevant to our work in this paper, we refer to Geman and Graffigne (1986), Gidas (1986) and Comets (1989) for the

results on consistency and some other asymptotic properties of MPLEs.

Although model selection is extensively studied in regression and time series, little has been done in the context of random fields in image analysis, except Kashyap and Chellappa (1983) for Gaussian random fields, and Smith and Miller (1990) for MRFs based on Rissanen's stochastic complexity [cf. Rissanen (1984)]. Hence the work in this paper appears to be the first theoretical study in this direction.

Schwarz (1978) gave a BIC — a Bayesian modification of Akaike's AIC — for choosing, among a finite number of exponential families, the "best" model to fit iid data. Woodroofe (1982) discussed the connection between model selection and the arcsine law for random walks and proved the consistency of Schwarz' procedure as a by-product. Haughton (1988) extended Schwarz' result to exponential families with their natural parameter spaces being submanifolds. Meanwhile, consistency of selection procedures was studied for various time series models in several papers, including Hannan (1980) and Shibata (1980), among others. A main assumption in the aforementioned papers is that the observations either are iid or satisfy certain mixing conditions. No results are available on selection of models with long-range dependence. However, as is pointed out in Ripley (1988), in (one-dimensional) time series analysis, short-range dependence is the norm and special models are needed to demonstrate long-range dependence; while in (multi-dimensional) spatial problems, long-range dependence appears inevitable. In particular, long-range dependence seems to be quite natural in many imaging problems. For instance, when GRFs on the two-dimensional integer lattice $\mathbb{Z}^2$ are used as texture models, the geometry of many texture types often indicates that the conditional distribution of the grey-level intensity X_o of the pixel o (the origin of $\mathbb{Z}^2$) given a configuration outside

the $n \times n$ symmetric square Λ_n centered at o need not eventually be equal to the corresponding unconditional distribution of X_o as n tends to infinity. Such a non-vanishing effect of the distant external conditions is interpreted as the phenomenon of *phase transition* in statistical mechanics, which creates a rich class of statistical models with long-range dependence. An important result in this paper is that Schwarz' BIC in the context of MRF texture models is consistent regardless of phase transition.

This paper is organized as follows. In Section 2 we give an introductory background of GRFs and a Bayesian formulation of the problem of choosing potentials for GRFs. An approximation of the Bayesian selection procedure — the BIC — is derived in Section 3 by using the traditional Laplace's method for asymptotic expansion of integrals. In Section 4 we investigate the asymptotics of MPLEs, including the mean square errors and certain moderate deviation probabilities, which are the key to the consistency of the BIC. Finally in Section 5 we make some comments on related future research.

2. MRFs AND RELATED MODEL SELECTION PROBLEMS

2.1. GRFs induced by pair-potentials of finite range.

In this section, we will use terms from statistical mechanics and image analysis freely unless further explanations are needed. Also, we only give a definition of GRFs that is convenient for the model selection problem considered in this paper. For more general definitions of GRFs, we refer to Ruelle (1978) and Georgii (1988).

With each pixel site $i \in \mathbb{Z}^2$, we associate a random variable X_i taking values in a finite set $S = \{\pm 1, \ldots, \pm r\}$, where $r \in \mathbb{N}$. Let $\Omega = S^{\mathbb{Z}^2}$ be the configuration

space and each $x \in \Omega$ be a realization of a random field X. Here $x = (x_i, i \in \mathbb{Z}^2)$ and $X = (X_i, i \in \mathbb{Z}^2)$, where x_i represents the grey-level of the pixel i. For every $\Lambda \subset \mathbb{Z}^2$, the subconfiguration space is $\Omega_\Lambda = S^\Lambda$, and we write $x_\Lambda = (x_i, i \in \Lambda) \in \Omega_\Lambda$ and $X_\Lambda = (X_i, i \in \Lambda)$. For every $i \in \mathbb{Z}^2$, we also write ${}_ix = (x_j, j \neq i)$ and ${}_iX = (X_j, j \neq i)$.

A set $U = \{hU_1(x_i), \quad \beta_{ij}U_2(x_i, x_j) : \quad x_i, x_j \in S; \quad i, j \in \mathbb{Z}^2\}$ is called a pair-potential of range $R \in \mathbb{N}$ if

(i) $U_1 : S \to \mathbb{R}$ and $U_2 : S \times S \to \mathbb{R}$ are two known functions;

(ii) $h \in \mathbb{R}$, called the external field coefficient, is an unknown parameter;

(iii) $\beta_{ij} \in \mathbb{R}, i, j \in \mathbb{Z}^2$, called the coupling coefficients, are also unknown parameters satisfying $\beta_{ij} = \beta_{ji}\ \forall i, j$ (symmetry); $\beta_{ij} = \beta_{o,j-i}\ \forall i, j$ (translation invariance); and $\beta_{ij} = 0\ \forall i, j$ with $\|i - j\| > R$, where the norm $\|\cdot\|$ on $\mathbb{Z}^2$ is defined by $\|i\| = |i_1| + |i_2|\ \forall i = (i_1, i_2) \in \mathbb{Z}^2$.

For every finite $\Lambda \subset \mathbb{Z}^2$ and every $x \in \Omega$, define the energy on Λ associated with x by

$$H_\Lambda(x) = -h \sum_{i \in \Lambda} U_1(x_i) - \frac{1}{2} \sum_{i,j \in \Lambda} \beta_{ij} U_2(x_i, x_j) - \sum_{\substack{i \in \Lambda \\ j \in \Lambda^c}} \beta_{ij} U_2(x_i, x_j). \tag{2.1}$$

The finite-volume Gibbs measure on Ω_Λ with respect to the external condition $y \in \Omega_{\Lambda^c}$ is given by

$$p_{\Lambda,y}(x_\Lambda) = \frac{\exp[-H_\Lambda(x_\Lambda \oplus y)]}{Z_{\Lambda,y}}, \quad x_\Lambda \in \Omega_\Lambda, \tag{2.2}$$

where the combined configuration $x_\Lambda \oplus y$ agrees with x_Λ on Λ, and y on Λ^c, and the normalizing factor

$$Z_{\Lambda,y} = \sum_{x_\Lambda \in \Omega_\Lambda} \exp[-H_\Lambda(x_\Lambda \oplus y)] \tag{2.3}$$

is called the partition function on Λ given $y \in \Omega_{\Lambda^c}$.

In particular, when Λ is a singleton $\{i\}$, the corresponding single-site Gibbs measure at i is called the local characteristic at i. It is known that the family of all local characteristics induced by a potential U determines the set of all finite-volume Gibbs measures induced by U.

Example 2.1. Let $U_1(x_i) = x_i$, $U_2(x_i, x_j) = x_i x_j$ and $S = \{-1, 1\}$. This corresponds to the pair-potential for the general Ising models (GIMs). Specifically, if $\beta_{ij} = \beta > 0$ whenever $\|i - j\| = 1$, and $\beta_{ij} = 0$ otherwise, then we have the two-dimensional Ising model.

Example 2.2. Let $U_1(x_i) \equiv 0$, and $U_2(x_i, x_j) = 1/[1 + \sigma(x_i - x_j)^2]$, where σ is a constant. This corresponds to the pair-potential for the texture models described in Geman and Graffigne (1986).

Let $\mathcal{G}(U)$ be the set of infinite-volume GRFs on Ω induced by U, so that $P \in \mathcal{G}(U)$ if for every finite $\Lambda \subset \mathbb{Z}^2$ and every $y \in \Omega_{\Lambda^c}$,

$$P(X_\Lambda = x_\Lambda | X_{\Lambda^c} = y) = p_{\Lambda,y}(x_\Lambda), \quad x_\Lambda \in \Omega_\Lambda. \tag{2.4}$$

P is said to be stationary if

$$P(X_{\Lambda+i} = x_\Lambda) = P(X_\Lambda = x_\Lambda) \quad \forall \text{ finite } \Lambda \subset \mathbb{Z}^2, i \in \mathbb{Z}^2, x_\Lambda \in \Omega_\Lambda, \tag{2.5}$$

where $X_{\Lambda+i} = (X_{j+i}, j \in \Lambda)$. Note that although the pair-potentials considered in this paper are translation invariant, they need not induce stationary GRFs (a phenomenon called symmetry breaking).

Meanwhile, under our assumptions on U the set $\mathcal{G}(U)$ is always nonempty, but need not be a singleton (a phenomenon called phase transition). In general, $\mathcal{G}(U)$ is a convex, compact Choquet simplex.

2.2. Formulation of the problem to choose potentials in MRF texture models.

Model selection is presumably a general issue in every imaging problem. However, the particular set-up in this paper is motivated by texture discrimination (closely related to segmentation) considered in Geman and Graffigne (1986), and Geman, Geman and Graffigne (1987). Suppose the data is a grey-level image consisting of several textured regions, whose maximum number and possible types are known. The goal is to classify the pixels. A simplified preliminary step is texture identification, in which a training sample from each texture is used to specify the MRF single-texture model, which is parametrized by the unknown coupling coefficients for some pairs of pixels. The specification of the MRFs is performed in this paper via selection of the potentials that induce the MRFs. More detailed discussions on texture discrimination can be found in Geman and Graffigne (1986), and Geman, Geman and Graffigne (1987).

Specification of a pair-potential consists of two interconnected parts: determination of the dimension of the parameter contained in it, and choice of the neighborhood system for the MRF. A neighborhood system $\mathcal{N}$ is a collection $\{N_{(i)} : i \in \mathbb{Z}^2\}$, where $N_{(i)} \subset \mathbb{Z}^2$ is called the set of neighbors of $i \in \mathbb{Z}^2$ satisfying $i \notin N_{(i)}$ and $i \in N_{(j)} \Leftrightarrow j \in N_{(i)}$ $\forall i, j \in \mathbb{Z}^2$. In connection with Section 2.1, every $P \in \mathcal{G}(U)$ is a MRF with respect to a neighborhood system $\mathcal{N}$ in the sense that

$$(2.6) \qquad P(X_i = x_i|_i X = {}_i x) = P(X_i = x_i | X_j = x_j, j \in N_{(i)}) \quad \forall\, i \in \mathbb{Z}^2,$$

where every $N_{(i)}$ is a subset of $\{j \in \mathbb{Z}^2 : \|i - j\| \leq R\}$.

Now let $\Theta = \mathbb{R}^K$ be the parameter space of interest, which is decomposed as a

disjoint union of several subspaces; i.e.

$$\Theta = \bigcup_{m=0}^{M} \Theta_m, \quad \Theta_{m_1} \cap \Theta_{m_2} = \emptyset \ \ \forall m_1 \neq m_2, \tag{2.7}$$

where each Θ_m corresponds to a candidate model (a pair-potential) parametrized by an element $\theta \in \mathbb{R}^{k_m}$. Here we assume every closure $\overline{\Theta}_m$ is an k_m-dimensional linear subspace of $\mathbb{R}^K$, $m = 0, 1, \ldots, M$. In particular, Θ_0 corresponds to the fully specified model with no unknown parameter. The components of a generic element $\theta \in \Theta$ are those coefficients h and β_{ij}'s given in Section 2.1.

Denote the set of all candidate models by $\mathcal{M} = \{0, 1, \ldots, M\}$, and let $\mathcal{N}_m$ be the neighborhood system for the model $m \in \mathcal{M}$. Notice that two models $m_1 \neq m_2$ are distinct if either: their parameter spaces have different dimensions (i.e., $k_{m_1} \neq k_{m_2}$); they are associated with different neighborhood systems (i.e., $\mathcal{N}_{m_1} \neq \mathcal{N}_{m_2}$); or maybe both. Note that $\mathcal{M}$ consists of mutually exclusive models.

The following three simple examples may help to illustrate the ideas. For simplicity, we take $U_1(x_i) \equiv 0$, $U_2(x_i, x_j) = x_i x_j$ and $S = \{-1, 1\}$ in all of them.

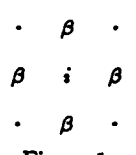

Figure 1
(Model m_1)

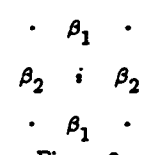

Figure 2
(Model m_2)

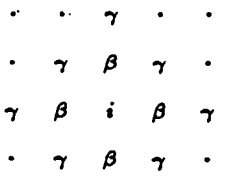

Figure 3
(Model m_3)

Example 2.3. (Fig. 1) Let

$$\beta_{ij} = \begin{cases} \beta & \text{if } \|i-j\| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Example 2.4. (Fig. 2) Denote $i = (i_1, i_2)$ for each $i \in \mathbb{Z}^2$ and let

$$\beta_{ij} = \begin{cases} \beta_1 & \text{if } \|i-j\| = 1 \text{ and } i_1 = j_1, \\ \beta_2 & \text{if } \|i-j\| = 1 \text{ and } i_2 = j_2, \\ 0 & \text{otherwise.} \end{cases}$$

Example 2.5. (Fig. 3) Let

$$\beta_{ij} = \begin{cases} \beta & \text{if } \|i-j\| = 1, \\ \gamma & \text{if } \|i-j\| = 2. \\ 0 & \text{otherwise.} \end{cases}$$

Denote the models in Fig. 1, 2 and 3 by m_1, m_2 and m_3 respectively. Notice that $\mathcal{N}_{m_1} = \mathcal{N}_{m_2}$, in which each site has four nearest neighbors; while $\mathcal{N}_{m_3}$ has an expanded gragh, in which the twelve neighbors of each site form two layers. On the other hand, the parameter space for m_1 is one-dimensional, while those for m_2 and m_3 are two-dimensional. In texture identification, m_1 or m_3 may be used to describe the horizontal cross-section of a tree, while m_2 may be more suitable for the vertical cross-section of the tree.

In general, starting from $\theta \in \Theta$, a different model can be obtained either by equating some components β_{ij} in θ (e.g. letting $\beta_1 = \beta_2 \triangleq \beta$ in m_2 in Example 2.4 to obtain m_1 in Example 2.3), or by letting $\beta_{ij} = 0$ for some pairs (i, j) (e.g. letting $\gamma = 0$ in m_3 in Example 2.5 to obtain m_1 in Example 2.3).

To characterize the original belief on the candidate models, we define a prior π on Θ as a convex mixture of mutually singular probability measures on the subspaces Θ_m, $m = 0, 1, \ldots, M$:

$$\pi = \sum_{m=0}^{M} \alpha_m \pi_m, \tag{2.8}$$

where $\alpha_m > 0$, $m = 0, 1, \ldots, M$ are constants with $\sum_{m=0}^{M} \alpha_m = 1$, and every π_m is a probability measure supported on the closure $\overline{\Theta}_m$ with a density μ_m, $m = 0, 1, \ldots, M$.

Let $\Lambda_n \subset \mathbb{Z}^2$ be the $n \times n$ symmetric square centered at the origin. Suppose the data x_{Λ_n}, denoted by $x(n)$ in what follows, is a realization of $X_{\Lambda_n} \triangleq X(n)$, where X has a distribution $P \in \mathcal{G}(U)$. Sometimes we write P_θ for P to indicate the parametrization. Now define the pseudo-likelihood function based on $x(n)$ as follows.

First extend $x(n)$ to a periodic configuration $\widetilde{x}(n) \in \Omega$ by periodization (or tiling), then define the pseudo-likelihood:

$$\mathcal{PL}(x(n), \theta) = \prod_{i \in \Lambda_n} P_\theta(X_i = x_i | {}_iX = {}_i\widetilde{x}(n)). \tag{2.9}$$

Any θ which maximizes the pseudo-likelihood is called a MPLE (based on $x(n)$). More about MPLEs will be discussed in Section 3 and 4.

The posterior distribution on Θ given $x(n)$ is then given by

$$\Pi_{x(n)}(A) = \frac{\int_A \mathcal{PL}(x(n), \theta) d\pi(\theta)}{\int_\Theta \mathcal{PL}(x(n), \theta) d\pi(\theta)} \quad \forall \text{ measurable } A \subset \Theta. \tag{2.10}$$

Denote the decision to select a model m based on $x(n)$ by $d(x(n)) = m$, where $d : \Omega_{\Lambda_n} \to \mathcal{M}$ is the decision function. If we impose 0-1 loss, the posterior Bayes

risk of $d(\cdot)$ given $x(n)$ is therefore

$$\mathcal{R}(x(n), d) = \int_{\Theta} 1_{\{\theta \notin \Theta_{d(x(n))}\}} d\Pi_{x(n)}(\theta) = 1 - \Pi_{x(n)}(\Theta_{d(x(n))}). \tag{2.11}$$

Hence the Bayesian solution to the model selection propblem is to choose a model m^* with $\Pi_{x(n)}(\Theta_{m^*}) \geq \Pi_{x(n)}(\Theta_m) \quad \forall m \in \mathcal{M}$. In Section 3 we will derive an implementable BIC as an approximation of the Bayesian solution.

3. The BIC for selecting pair-potentials

In this section we extend Schwarz' BIC to the MRF texture models based on the fact that the pseudo-likelihood function given in (2.9) and its analogous expressions restricted to the subspaces $\overline{\Theta}_m$, $m = 0, 1, \ldots, M$, can be written as exponential families. For general theory of exponential families, we refer to Barndorff-Nielsen (1978) and Brown (1986).

First notice that $\mathcal{PL}(X(n), \theta)$, $\theta \in \Theta$ is a full K-dimensional (standard) exponential family. Since any exponential family which is not minimal can always be reduced to a minimal exponential family through sufficiency, reparametrization and proper choice of the reference measure, we can let $\mathcal{PL}_m(X(n), \theta)$, $\theta \in \overline{\Theta}_m$ be the corresponding k_m-dimensional minimal exponential family restricted to $\overline{\Theta}_m$, $m = 1, \ldots, M$. We will use the same notation θ for the new parameters after reparametrization unless further distinction is needed.

More specifically, we write

$$\mathcal{PL}(X(n), \theta) = \exp\{|\Lambda_n|[\theta^t \cdot Y(n) - b(\theta, n)]\}, \quad \theta \in \Theta, \tag{3.1}$$

where θ^t is the transpose of the (column) vector θ, $|\Lambda_n|$ is the cardinality of Λ_n,

and the sufficient statistic $Y(n)$ is an K-dimensional vector with the components

$$\begin{cases} \dfrac{1}{|\Lambda_n|}\sum_{i\in\Lambda_n} U_1(X_i) & \text{corresponding to } h \text{ in } \theta, \\ \dfrac{1}{|\Lambda_n|}\sum_{i\in\Lambda_n} U_2(X_i, \widetilde{X}_j(n)) & \text{corresponding to } \beta_{ij} \text{ in } \theta, \end{cases}$$

and the cumulant generating function $b(\theta, n)$ (depending on $X(n)$ also) is given by

$$b(\theta, n) = \frac{1}{|\Lambda_n|}\sum_{i\in\Lambda_n} \log \sum_{X_i\in S} \exp[\theta^t \cdot Y(n)].$$

By the same token, for each $m \in \mathcal{M}$ we can also write

$$\mathcal{PL}_m(X_n, \theta) = \exp\{|\Lambda_n|[\theta^t \cdot Y_m(n) - b_m(\theta, n)]\}, \quad \theta \in \overline{\Theta}_m, \tag{3.2}$$

with the corresponding sufficient statistic $Y_m(n)$ and cumulant generating function $b_m(\theta, n)$.

The BIC is closely related the MPLEs as follows. For each $m \in \mathcal{M}$, let $\widehat{\theta}_m$ be a MPLE restricted to $\overline{\Theta}_m$. Define the index

$$Q_m(n) = |\Lambda_n|[\widehat{\theta}_m^t \cdot Y_m(n) - b_m(\widehat{\theta}_m, n)] - \frac{k_m}{2}\log|\Lambda_n|. \tag{3.3}$$

Then the BIC selects the model of largest index.

To make the argument rigorous, we need the following assumptions.

(A1) (Bounds on the densities in the decomposition of prior π) There exist positive constants c' and C' such that for every $m \in \mathcal{M}$

$$c' \le \mu_m(\theta) \le C' \quad \forall \theta \in \overline{\Theta}_m.$$

(A2) (Identifiability conditions) There are several ways to impose identifiability. Referring to (2.2), let $p_o(x;\theta) \triangleq p_{\{o\},o x}(x_o)$, $x \in \Omega$ be the local characteristic at

the origin parametrized by θ. Then identifiability means that $p_o(x;\theta) \neq p_o(x;\theta')$ for some $x \in \Omega$ whenever $\theta \neq \theta'$. The other alternative is to impose conditions on potentials, which is related to the notion of equivalence of potentials [cf. Georgii (1988) and Gidas (1988)]. A sufficient condition is given in Ji (1990), in terms of double extremal points [cf. (A4) in Ji (1990)].

For $P \in \mathcal{G}(U)$, the following "conditional independence lemma" is simply due to the fact that the potential U has a range R.

Lemma 3.1. *Let $B_1, \ldots, B_L$ be bounded regions in $\mathbb{Z}^2$, and $\mathcal{B} = \mathbb{Z}^2 \backslash (\bigcup_{l=1}^{L} B_l)$. If the distances between B_l and $B_{l'}$ are greater than R for all $l \neq l'$, then for any bounded measurable functions $f_l : \Omega_{B_l} \to \mathbb{R}$, $l = 1, \ldots, L$, we have*

$$E\left\{ \prod_{l=1}^{L} f_l(X_{B_l}) \,\middle|\, x_{\mathcal{B}} \right\} = \left\{ \prod_{l=1}^{L} E[f_l(X_{B_l}) \,|x_{\mathcal{B}}] \right\}$$

uniformly $\forall\, x_{\mathcal{B}} \in \Omega_{\mathcal{B}}$, where $E(\cdot|x_{\mathcal{B}})$ is the conditional expectation with respect to $P(\cdot|x_{\mathcal{B}})$.

Now for each $i \in \Lambda_n$ let $\Lambda_{i,2R}$ be the $2R \times 2R$ square lattice centered at i. In particular, $\Lambda_{o,2R} = \Lambda_{2R}$. Denote generic elements of S and $\Omega_{\Lambda_{2R}\backslash\{o\}}$ by ζ and η respectively, hence $\zeta \oplus \eta \in \Omega_{\Lambda_{2R}}$. Define

$$\begin{cases} T_i(\zeta \oplus \eta) = 1_{\{\tilde{X}_{\Lambda_{i,2R}} = \zeta \oplus \eta\}}, \\ \quad T_i(\eta) = 1_{\{\tilde{X}_{\Lambda_{i,2R}\backslash\{i\}} = \eta\}}, \quad i \in \Lambda_n; \end{cases} \tag{3.4}$$

and

$$\begin{cases} N_n(\zeta \oplus \eta) = \sum_{i \in \Lambda_n} T_i(\zeta \oplus \eta), \\ \quad N_n(\eta) = \sum_{i \in \Lambda_n} T_i(\eta). \end{cases} \tag{3.5}$$

Lemma 3.2. *For $P \in \mathcal{G}(U)$, there exists a constant $\lambda_1 > 0$, such that*

$$P\left(\frac{N_n(\zeta \oplus \eta)}{|\Lambda_n|} < \lambda_1\right) \leq C_1 \exp(-a_1 n)$$

for some constants $C_1 > 0$, $a_1 > 0$, and uniformly for all large n and all $\zeta \oplus \eta$.

Proof. Partition Λ_n as a union of disjoint tiles: $\Lambda_n = \bigcup_{l=1}^{L} D_l$, so that each tile D_l is a $3R \times 3R$ square lattice. Here we suppose n divides $3R$ without loss of generality. Therefore $L = (\frac{n}{3R})^2$. On the other hand, we can also write the decomposition $\Lambda_n = \bigcup_{k=1}^{(3R)^2} G_k$, where every G_k contains exactly L pixels with the same relative positions in the L tiles. For instance, one G_k may consist of the centers of the L tiles. Another G_k may consist of all upper-left corners of the L tiles.

For a fixed k, let $\mathcal{B}_k = \mathbb{Z}^2 \setminus \bigcup_{i \in G_k} \Lambda_{i,2R}$ be the "corridor". By Lemma 3.1, we have

$$\begin{aligned} E\left\{\exp\left[-\frac{1}{n}\sum_{i\in G_k} T_i(\zeta \oplus \eta)\right]\right\} &= E\left(E\left\{\exp\left[-\frac{1}{n}\sum_{i\in G_k} T_i(\zeta \oplus \eta)\right] \,\middle|\, X_{\mathcal{B}_k}\right\}\right) \\ &= E\left(\prod_{i\in G_k} E\{\exp[-\frac{1}{n}T_i(\zeta \oplus \eta)] \mid X_{\mathcal{B}_k}\}\right). \end{aligned}$$

Note that for every $x_{\mathcal{B}_k}$ we have the Taylor expansion

$$\begin{aligned} E\{\exp[-\frac{1}{n}T_i(\zeta \oplus \eta)] \mid x_{\mathcal{B}_k}\} &= 1 - \frac{1}{n}E[T_i(\zeta \oplus \eta) \mid x_{\mathcal{B}_k}] + o(\frac{1}{n}) \\ &\leq 1 - \frac{a'}{n} \end{aligned}$$

for some $a' > 0$ and all large n. Therefore,

$$E\left\{\exp\left[-\frac{1}{n}\sum_{i\in G_k} T_i(\zeta \oplus \eta)\right]\right\} \leq \exp(-a'L/n) = \exp(-a''n), \quad \text{for some } a'' > 0.$$

So for every $\zeta \oplus \eta$ we have

$$P\left(\frac{1}{|\Lambda_n|}\sum_{i\in G_k} T_i(\zeta \oplus \eta) < \lambda'\right) \leq \exp[-(a'' - \lambda')n]$$

for some $0 < \lambda' < a''$.

Since $N_n(\zeta \oplus \eta) = \sum_{k=1}^{(3R)^2} \sum_{i \in G_k} T_i(\zeta \oplus \eta)$, and there are only a finite number of possible $\zeta \oplus \eta$, Lemma 3.2 follows easily. □

Define the event

$$\mathcal{E}_n = \{x(n) \in \Omega_{\Lambda_n} : \quad \frac{N_n(\zeta \oplus \eta)}{|\Lambda_n|} \geq \lambda_1 \ \ \forall\, \zeta \oplus \eta \in \Omega_{\Lambda_{2R}}\}.$$

We will concentrate on $\mathcal{E}_n$ since its complement is essentially negligible based on Lemma 3.2.

Lemma 3.3. *There exist positive constants c and C, such that*

$$-C \leq v^t \cdot \nabla^2 b_m(\theta, n) \cdot v \leq -c$$

uniformly for all $\theta \in \overline{\Theta}_m$, $v \in \mathbb{R}^{k_m}$ with $\|v\| = 1$, $m \in \mathcal{M}$ and $x(n) \in \mathcal{E}_n$, where $\nabla^2 b_m(\theta, n)$ is the Hessian matrix of $b_m(\theta, n)$ with respect to θ.

Proof. Let

$$K_m(\theta, n) = \theta^t \cdot Y_m(n) - b_m(\theta, n), \quad \theta \in \overline{\Theta}_m. \tag{3.6}$$

By (3.2), (3.4) amd (3.5), we have

$$K_m(\theta, n) = \sum_{\eta} \frac{N_n(\eta)}{|\Lambda_n|} \sum_{\zeta} \frac{N_n(\zeta \oplus \eta)}{N_n(\eta)} \log p_o^{(m)}(\zeta|\eta; \theta), \tag{3.7}$$

where $p_o^{(m)}(\zeta|\eta; \theta)$ is the local characteristic at o with respect to $\theta \in \overline{\Theta}_m$. Write

$$p_o^{(m)}(\zeta|\eta; \theta) = \frac{\exp[\theta^t \cdot \phi_m(\zeta \oplus \eta)]}{\sum_{\zeta'} \exp[\theta^t \cdot \phi_m(\zeta' \oplus \eta)]}, \quad \theta \in \overline{\Theta}_m \tag{3.8}$$

with an appropriate vector-valued function ϕ_m. By routine calculation, for $v \in \mathbb{R}^{k_m}$ and $\theta \in \overline{\Theta}_m$ we have

$$v^t \cdot \nabla^2 b_m(\theta, n) \cdot v = v^t \cdot \nabla^2 K_m(\theta, n) \cdot v$$

$$(3.9) \quad = -\sum_{\eta} \frac{N_n(\eta)}{|\Lambda_n|} E_\theta^{(m)}\{ (v^t \cdot [\phi_m(X_o \oplus \eta) - E_\theta^{(m)}(\phi_m(X_o \oplus \eta)|\eta)])^2 \mid \eta \},$$

where $E_\theta^{(m)}(\cdot |\eta)$ is the conditional expectation with respect to $p_o^{(m)}(\cdot |\eta; \theta)$. Note that $\frac{N_n(\eta)}{|\Lambda_n|} \geq \lambda_1$ on $\mathcal{E}_n$ for all η. Therefore, Lemma 3.3 follows from the identifiability condition (A2). □

Lemma 3.4. *For every $m \in \mathcal{M}$ and $x(n) \in \mathcal{E}_n$, there exists a unique MPLE $\widehat{\theta}_m$ on $\overline{\Theta}_m$.*

Proof. On the event $\mathcal{E}_n$, $K_m(\theta, n)$ is strictly concave in $\theta \in \overline{\Theta}_m$ (Lemma 3.3). So Lemma 3.4 follows from Theorem 5.5 in Brown (1986). □

The next theorem is the derivation of BIC characterized by (3.3). Since the argument is a standard application of Laplace's method, we only give a sketch.

Theorem 3.1. *On the event $\mathcal{E}_n$ with large n, the Bayesian selection procedure is equivalent to choosing the model m with largest index $Q_m(n)$ given in* (3.3).

Proof. It suffices to show that for every $m \in \mathcal{M}$, the asymptotic expansion

$$(3.10) \qquad \log \int_{\overline{\Theta}_m} \mathcal{PL}_m(x(n), \theta)\ \mu_m(\theta) d\theta = Q_m(n) + O(1)$$

holds uniformly for all large n and all $x(n) \in \mathcal{E}_n$.

First by Taylor expansion we have

$$K_m(\theta, n) - K_m(\widehat{\theta}_m, n)$$

$$(3.11) \quad = (\theta - \widehat{\theta}_m)^t \cdot [Y_m(n) - \nabla b_m(\widehat{\theta}_m, n)] - \frac{1}{2}(\theta - \widehat{\theta}_m)^t \cdot \nabla^2 b_m(\omega, n) \cdot (\theta - \widehat{\theta}_m)$$

for $\theta \in \overline{\Theta}_m$, where $\nabla b_m(\theta, n)$ is the gradient vector of $b_m(\theta, n)$ with respect to θ, and $\omega \in \overline{\Theta}_m$ satisfying $\|\omega - \widehat{\theta}_m\| \leq \|\theta - \widehat{\theta}_m\|$. It follows from Lemma 3.3 that

$$-\frac{C}{2} \leq \frac{K_m(\theta, n) - K_m(\widehat{\theta}_m, n)}{\|\theta - \widehat{\theta}_m\|^2} \leq -\frac{c}{2} \quad \forall \theta \in \overline{\Theta}_m. \tag{3.12}$$

Secondly, for $\delta > 0$ decompose $\overline{\Theta}_m = \mathcal{V}_\delta(\widehat{\theta}_m) \cup [\overline{\Theta}_m \backslash \mathcal{V}_\delta(\widehat{\theta}_m)]$, where $\mathcal{V}_\delta(\widehat{\theta}_m) = \{\theta \in \overline{\Theta}_m : \|\theta - \widehat{\theta}_m\| \leq \delta\}$. Notice that

$$\int_{\overline{\Theta}_m \backslash \mathcal{V}_\delta(\widehat{\theta}_m)} \mathcal{PL}_m(x(n), \theta)\, \mu_m(\theta) d\theta \leq \exp[|\Lambda_n| K_m(\widehat{\theta}_m, n)]\, \exp(-a_2 |\Lambda_n|)$$

for some $a_2 > 0$. And

$$\begin{aligned} &\int_{\mathcal{V}_\delta(\widehat{\theta}_m)} \mathcal{PL}_m(x(n), \theta)\, \mu_m(\theta) d\theta \\ &= \exp[|\Lambda_n| K_m(\widehat{\theta}_m, n)] \int_{\mathcal{V}_\delta(\widehat{\theta}_m)} \exp\{|\Lambda_n| [K_m(\theta, n) - K_m(\widehat{\theta}_m, n)]\}\, \mu_m(\theta) d\theta. \end{aligned}$$

Using the upper and lower bounds given in (3.12) and the condition (A1), we derive

$$\int_{\mathcal{V}_\delta(\widehat{\theta}_m)} \exp\{|\Lambda_n| [K_m(\theta, n) - K_m(\widehat{\theta}_m, n)]\}\, \mu_m(\theta) d\theta = O\left(\frac{1}{|\Lambda_n|^{\frac{k_m}{2}}}\right).$$

Therefore, (3.10) follows. □

Remark. The expansion given in (3.3) has two terms. A more accurate expansion with three terms can be derived without too much trouble which is similar to the result in Haughton (1988), Theorem 2.3. For the study of frequentist properties of BIC carried out in Section 4, our two-term expansion is good enough.

4. Some asymptotics of MPLEs and the consistency of BIC

Although the BIC given in (3.3) is an approximation of the Bayesian selection procedure based on a decision-theoretic set-up, the expression of each index $Q_m(n)$ in (3.3) depends on the prior information only through $\overline{\Theta}_m$ — the support of μ_m.

(A three-term expansion will involve more prior information in terms of α_m and μ_m.) On the other hand, assuming only one of the available models is correct, a frequentist may want to know whether the selected model will converge to the true model as n tends to infinity. To answer this question, we assume in Section 4 that the data $X(n)$ is generated by $P_\theta \in \mathcal{G}(U)$ with $\theta \in \Theta_m$ for some $m \in \mathcal{M}$.

Definition 4.1. A selection procedure $d(\cdot)$ is said to be *consistent* if

$$\lim_{n\to\infty} P_\theta\left(d(X(n)) = m\right) = 1.$$

The main result in Section 4 is:

Theorem 4.1. *The BIC given in* (3.3) *is consistent regardless of phase transition.*

To prove Theorem 4.1, we first fix a true model $m \in \mathcal{M}$ and then write $\mathcal{M}\backslash\{m\} = \mathcal{M}_1(m) \bigcup \mathcal{M}_2(m)$, where

$$\begin{cases} \mathcal{M}_1(m) = \{m' \in \mathcal{M} : \ \overline{\Theta}_m \nsubseteq \overline{\Theta}_{m'}\}; \\ \mathcal{M}_2(m) = \{m' \in \mathcal{M} : \ \overline{\Theta}_m \subsetneq \overline{\Theta}_{m'}\}. \end{cases} \tag{4.1}$$

We need to show that for $\theta \in \Theta_m$, both $P_\theta(d(X(n)) \in \mathcal{M}_1(m))$ and $P_\theta(d(X(n)) \in \mathcal{M}_2(m))$ tend to zero at appropriate rates as n tends to infinity.

Call $d(x(n)) \in \mathcal{M}_1(m)$ and $d(x(n)) \in \mathcal{M}_2(m)$ Case 1 and Case 2 respectively. Refer to Examples 2.3, 2.4 and 2.5 for the following clarification.

Case 1: choose m_1 when either m_2 or m_3 is true (underparametrization); choose m_2 when m_3 is true, or vice versa (incorrect specification of a neighborhood system).

Case 2: choose either m_2 or m_3 when m_1 is true (overparametrization).

In the investigations of both Case 1 and Case 2, the tight connection between parameter estimation and model selection is exploited. First of all, Case 1 is easy to

handle based on the consistency of MPLEs. The following lemma is a strengthened version of the theorem on page 1512, Geman and Graffigne (1986). The argument there essentially carries over, hence the proof is omitted. We also refer to Theorem 3.2 in Ji (1990) for a stronger result.

Lemma 4.1. *For every $\epsilon > 0$, there exists $a_3 > 0$, such that when $\theta \in \Theta_m$,*

$$P_\theta(\|\widehat{\theta}_m - \theta\| > \epsilon) \leq \exp(-n^{a_3})$$

uniformly for all large n.

Lemma 4.2. *There exists $a_4 > 0$, such that when $\theta \in \Theta_m$,*

$$P_\theta(d(X(n)) \in \mathcal{M}_1(m)) \leq \exp(-n^{a_4})$$

uniformly for all large n.

Proof. First, notice that there exists $\epsilon > 0$, such that $\|\theta' - \theta\|^2 \geq 2\epsilon$ for all $\theta' \in \overline{\Theta}_{m'}$ and all $m' \in \mathcal{M}_1(m)$.

Second, without loss of generality let $M \in \mathcal{M}$ be the "largest" model, i.e. $\overline{\Theta}_M = \Theta$.

Fix $m' \in \mathcal{M}_1(m)$ and define

$$\mathcal{D}_n = \{x(n) \in \Omega_{\Lambda_n} : \|\widehat{\theta}_m - \theta\| + \|\widehat{\theta}_M - \theta\| \leq \epsilon\}.$$

Then $P_\theta(\mathcal{D}_n^c) \leq \exp(-n^{a_5})$ for some $a_5 > 0$ and for all large n by Lemma 4.1.

On the event $\mathcal{E}_n \bigcap \mathcal{D}_n$, we have

$$\begin{aligned}
& K_{m'}(\widehat{\theta}_{m'}, n) - K_m(\widehat{\theta}_m, n) \\
\leq & \left[K_{m'}(\widehat{\theta}_{m'}, n) - K_M(\widehat{\theta}_M, n)\right] + \left[K_M(\widehat{\theta}_M, n) - K_m(\widehat{\theta}_m, n)\right] \\
\leq & -\frac{c}{2}\|\widehat{\theta}_{m'} - \widehat{\theta}_M\|^2 + \frac{C}{2}\|\widehat{\theta}_M - \widehat{\theta}_m\|^2 \\
\leq & -\frac{c}{2}\epsilon + \frac{C}{2}\epsilon^2 \\
\leq & -a_6
\end{aligned}$$

for some $a_6 > 0$ provided ϵ is sufficiently small. Therefore, there exists $a_7 > 0$, such that

$$Q_m(n) - Q_{m'}(n) \geq a_7|\Lambda_n| \tag{4.2}$$

holds on $\mathcal{E}_n \bigcap \mathcal{D}_n$ for all large n. Hence Lemma 4.2 follows. □

Case 2 is much more subtle. Notice that in the expression of the index $Q_m(n)$ for each m, the first term $|\Lambda_n|K_m(\widehat{\theta}_m, n)$ represents the fit of the model, while the second term $\frac{k_n}{2}\log|\Lambda_n|$ is the penalty for overparametrization. In Case 1 the first term dominates the second term. However in Case 2 the "parsimony principle" prevails against overfitting. This can only be realized by carefully investigating certain asymptotic orders for the MPLEs, which are stated in the next two lemmas. To the best of our knowledge, the results contained in Lemma 4.3 and Lemma 4.4 have not appeared anywhere else. They are derived by using the conditioning argument developed in Geman and Graffigne (1986), and in Ji (1990) to overcome the difficulties caused by phase transition.

Lemma 4.3. *(Restricted mean square errors of MPLEs) For every $\theta \in \Theta_m$ and $m \in \mathcal{M}$, we have*

$$E_\theta(\|\widehat{\theta}_m - \theta\|^2\ 1_{\mathcal{E}_n}) = O\left(\frac{1}{|\Lambda_n|}\right), \quad as\ n \to \infty;$$

Proof. Note that

$$\begin{aligned}\nabla K_m(\theta, n) =& \nabla K_m(\widehat{\theta}_m, n) + \nabla^2 K_m(\omega', n) \cdot (\theta - \widehat{\theta}_m) \\ =& \nabla^2 b_m(\omega', n) \cdot (\theta - \widehat{\theta}_m)\end{aligned} \tag{4.3}$$

for some $\omega' \in \overline{\Theta}_m$ with $\|\omega' - \widehat{\theta}_m\| \leq \|\theta - \widehat{\theta}_m\|$. By Lemma 3.3 we obtain

$$\|\widehat{\theta}_m - \theta\|^2 \leq C^2 \|\nabla K_m(\theta, n)\|^2 \quad \text{on } \mathcal{E}_n. \tag{4.4}$$

Therefore, it suffices to show that

$$E_\theta(\|\nabla K_m(\theta, n)\|^2) = O\left(\frac{1}{|\Lambda_n|}\right). \tag{4.5}$$

By (3.4), (3.5) and (3.7) we rewrite $K_m(\theta, n)$ as

$$K_m(\theta, n) = \frac{1}{|\Lambda_n|} \sum_{i \in \Lambda_n} \left(\sum_{\zeta \oplus \eta} T_i(\zeta \oplus \eta) \log p_o^{(m)}(\zeta | \eta; \theta) \right).$$

Then using the notation in (3.8) and (3.9) we obtain

$$\begin{aligned}\nabla K_m(\theta, n) &= \frac{1}{|\Lambda_n|} \sum_{i \in \Lambda_n} \widetilde{W}_i \quad \text{with} \\ \widetilde{W}_i &= \sum_{\eta} T_i(\eta) \left\{ \phi_m(X_i \oplus \eta) - E_\theta^{(m)}[\phi_m(X_i \oplus \eta) | \eta] \right\}.\end{aligned} \tag{4.6}$$

Let W_i be a generic component of $\widetilde{W}_i$. Then it suffices to show that

$$E_\theta \left[\left(\sum_{i \in \Lambda_n} W_i \right)^2 \right] = O(|\Lambda_n|). \tag{4.7}$$

Using the decompositions of Λ_n in Lemma 3.2, it suffices to show that

$$E_\theta \left[\left(\sum_{i \in G_k} W_i \right)^2 \right] = O(|\Lambda_n|) \quad \text{for each } G_k. \tag{4.8}$$

Since $E_\theta(W_i | x_{\mathcal{B}_k}) = 0$ for every $x_{\mathcal{B}_k}$, it follows from Lemma 3.1 that

$$E_\theta \left[\left(\sum_{i \in G_k} W_i \right)^2 \middle| x_{\mathcal{B}_k} \right] = \sum_{i \in G_k} E_\theta(W_i^2 | x_{\mathcal{B}_k}).$$

Then (4.8) follows since all W_i are uniformly bounded. □

Lemma 4.4. *(Moderate deviation probabilities of MPLEs) For every $\epsilon \in (0,1)$, there exists $a_8 \in (0,1)$ such that*

$$P_\theta\left(|\Lambda_n|\ \|\widehat{\theta}_m - \theta\|^2 > \epsilon \log n\right) = O\left(\frac{1}{n^{a_8}}\right), \quad as\ n \to \infty.$$

Proof. By (4.4) and Lemma 3.2, it suffices to show that for every $\epsilon \in (0,1)$,

$$P_\theta(|\Lambda_n|\ \|\nabla K_m(\theta, n)\|^2 > \epsilon \log n) = O\left(\frac{1}{n^{a_8}}\right). \tag{4.9}$$

And componentwise, it suffices to show that for every $\epsilon \in (0,1)$,

$$P_\theta\left(\left|\sum_{i \in G_k} W_i\right| > \epsilon \tau_n\right) = O\left(\frac{1}{n^{a_8}}\right), \tag{4.10}$$

where $\tau_n = \sqrt{|\Lambda_n| \log n}$.

By Taylor expansion, for a small $\rho > 0$ we have

$$E_\theta\left[\exp(\rho W_i/\sqrt{\tau_n})\ |x_{\mathcal{B}_k}\right] = 1 + \frac{\rho^2}{\tau_n} O(1). \tag{4.11}$$

Hence by Lemma 3.1 we have

$$\begin{aligned} E_\theta\left[\exp\left(\sum_{i \in G_k} \rho W_i\right) \Big/ \sqrt{\tau_n}\ \Big|\ x_{\mathcal{B}_k}\right] &= \prod_{i \in G_k} E_\theta\left[\exp(\rho W_i/\sqrt{\tau_n})\ |x_{\mathcal{B}_k}\right] \\ &\le \exp\left(a_9 n^2 \rho^2/\tau_n\right) \end{aligned} \tag{4.12}$$

for some $a_9 > 0$. Therefore, by the Markov inequality we obtain

$$\begin{aligned} P_\theta\left(\sum_{i \in G_k} W_i > \epsilon \tau_n\right) &\le \exp(-\rho\epsilon\sqrt{\tau_n})\ E_\theta\left[\exp\left(\sum_{i \in G_k} \rho W_i/\sqrt{\tau_n}\right)\right] \\ &= \exp(-\rho\epsilon\sqrt{\tau_n})\ E_\theta\left\{\prod_{i \in G_k} E_\theta\left[\exp(\rho W_i/\sqrt{\tau_n})\ |X_{\mathcal{B}_k}\right]\right\} \\ &\le \exp\{-\rho\epsilon\sqrt{\tau_n}\ [1 - a_9 n^2 \rho/(\epsilon \tau_n^{\frac{3}{2}})]\} \\ &= O\left(\frac{1}{n^{a_8}}\right) \end{aligned}$$

provided we set $\rho = \nu n^{-\frac{1}{2}} (\log n)^{\frac{3}{4}}$ and $a_8 = \nu\epsilon$, where $\nu \in (0,\ \epsilon/a_9)$ is a small constant.

On the other hand, $P_\theta(-\sum_{i \in G_k} W_i > \epsilon \tau_n) = O\,(1/n^{a_8})$ can be derived in the same way. Hence (4.10) follows. □

Lemma 4.5. *For $\theta \in \Theta_m$, we have*

$$P_\theta(d(X(n)) \in \mathcal{M}_2(m)) = O\left(\frac{1}{n^{a_8}}\right),$$

where $a_8 \in (0,1)$ is given in Lemma 4.4.

Proof. For any $m'' \in \mathcal{M}_2(m)$, note that on $\mathcal{E}_n$ we have

$$K_m(\widehat{\theta}_m, n) - K_{m''}(\widehat{\theta}_{m''}, n) \geq -\frac{C}{2}(\|\widehat{\theta}_{m''} - \theta\|^2 + \|\widehat{\theta}_m - \theta\|^2).$$

By Lemma 4.4 there exists a set $\mathcal{J}_n$ with $P_\theta(\mathcal{J}_n^c) = O\left(\frac{1}{n^{a_8}}\right)$ such that on $\mathcal{J}_n$ we have

$$Q_m(n) - Q_{m''}(n) \geq a_{10} \log n,$$

where $a_{10} = k_{m''} - k_m - C\epsilon > 0$ provided ϵ is sufficiently small. Hence Lemma 4.5 follows. □

Now we can complete the proof of Theorem 4.1 easily.

Proof of Theorem 4.1. There are two ways to prove Theorem 4.1: the first one is by Lemma 4.2, Lemma 4.3 and Chebyshev inequality; the second one is by Lemma 4.2 and Lemma 4.5. Note that the decay rate of the probability of choosing an incorrect model produced by the first method can only be of the order $1/\log n$. So the second method gives a stronger result. □

Remark. The consistency given in Definition 4.1 is in the weak sense. A selection procedure $d(\cdot)$ is said to be strongly consistent if for every $\theta \in \Theta_m$, we have $d(X(n)) \to m$ with P_θ-probability one as n tends to infinity. However, Schwarz' BIC does not seem to be strongly consistent. This can roughly be viewed in the following way. The moderate deviation probabilities in Lemma 4.4 are not summable in n, thus the Borel-Cantelli lemma does not apply. This is supported by the

exact order for moderate deviation probabilities in the iid case given in Rubin and Sethuraman (1965).

5. Concluding remarks

There are several unresolved issues that we would like to draw to attention. We are currently studying them and will report the progress in another manuscript.

This paper has focused on BIC and its asymptotic properties. A natural question is: what about AIC? In the iid case, it was pointed out in Woodroofe (1982) that AIC is superior to BIC asymptotically when the dimensionality of the parameter tends to infinity at an appropriate rate as the sample size tends to infinity. We expect that a similar result will hold for the MRF texture models if we either let the range of the potential $R = R_n \to \infty$, or equivalently let the dimensionality of the parameter $K = K_n \to \infty$, as $n \to \infty$. More elegant asymptotics are needed to accomplish this. In this aspect, the method in Ji (1990) may be helpful.

Another important question is: what is the asymptotic distribution for each index $Q_m(n)$? Such a need is shown in Woodroofe (1982), in which the distribution of the number of superfluous parameters contained in the selected model is found. One can use this result to make numerical comparisons between different models. However, the situation with which we are dealing is much more complicated than the iid case in Woodroofe (1982). The arcsine law does not apply. Moreover, the central limit theorem for GRFs generally fails under phase transition. A possible attempt at a solution would be to investigate the convergence rate of the asymptotic normality for the MPLEs under some reasonable conditions, and then to generalize Woodroofe's result.

Although the results in this paper hold regadless of phase transition, they are limited to the study of single-texture identification. For multi-texture models, the assumption of translation invariant potentials is violated. Heuristically, one could still use the BIC given in (3.3) to select a potential for each single-textured region and then adopt the combined potential for the whole random field. Theoretical justification seems to be quite challenging.

Acknowledgement. Grant support from ONR (N00014-89-J-1760) is greatly acknowledged. The authors particularly thank Stuart Geman for helpful comments and suggestions, and also Richard Smith, Adrian Raftery, Kurt Smith and Michael Miller for providing references.

REFERENCES

(1) Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

(2) Besag, J. (1986). On the statistical analysis of dirty pictures (with dicussion). *J. Roy. Stat. Soc. Ser. B* **6**, 259-302.

(3) Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families.* IMS, Hayward, CA.

(4) Comets, F. (1989). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. Preprint, Univ. of Paris-X.

(5) Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and Bayesian restoration of images. *IEEE Trans. PAMI* **6**, 721-741.

(6) Geman, D., Geman, S. and Graffigne, C. (1987). Locating texture and object boundaries. *Pattern Recognition Theory and Applications* (P. A.

Devijver and J. Kittler, ed.), Springer, New York.

(7) Geman, S. and Graffigne, C. (1986). Markov random field image models and their applications to computer vision. *Proceedings of the International Congress of Mathematicians* (A. M. Gleason, ed.), AMS, Providence, RI.

(8) Georgii, H. O. (1988). *Gibbs Measures and Phase Transitions.* Walter de Gruyter, Berlin-New York.

(9) Gidas, B. (1986). Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions. *Proceedings of the Workshop on Stochastic Differential Systems with Applications in Electrical/Computer Engineering, Control Theory and Operations Research.* IMS, Springer, New York.

(10) Gidas, B. (1988). Parameter estimation for Gibbs distributions. Preprint, Brown Univ.

(11) Grenander, U. (1989). Advanced pattern theory. *Ann. Stat.* **17**, 1-30.

(12) Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Stat.* **8**, 1071-1081.

(13) Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Stat.* **16**, 342-355.

(14) Ji, C. (1990). Sieve estimators for pair-interaction potentials and local characteristics in Gibbs random fields. Tech. Report #2037, Dept. of Stat., UNC-Chapel Hill.

(15) Kashyap, R. and Chellappa, R. (1983). Estimation and choice of neighbors in spatial-interaction models of images. *IEEE Trans. IT* **29**, 60-72.

(16) Ripley, B. D. (1988). *Statistical Inference for Spatial Processes.* Cambridge

Univ. Press, Cambridge-New York.

(17) Rissanen, J. (1984). Stochastic complexity. *J. Roy. Stat. Soc. Ser. B* **49**, 223-239.

(18) Rubin, H. and Sethuraman, J. (1965). Probabilities of moderate deviations. *Sankhyā, Ser. A* **27**, 325-346.

(19) Ruelle, D. (1978). *Thermodynamic Formalism.* Addison-Wesley, Reading.

(20) Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461-464.

(21) Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Stat.* **8**, 147-164.

(22) Smith, K. R. and Miller, M. I. (1990). A Bayesian approach incorporating Rissanen complexity for learning Markov random field texture models. *IEEE (?)*, 2317-2320.

(23) Woodroofe, M. (1982). On model selection and the arcsine laws. *Ann. Stat.* **10**, 1182-1194.

DEPARTMENT OF STATISTICS, UNIVERSITY OF NORTH CAROLINA, CHAPEL HILL, NC 27599-3260, USA