The Annals of Applied Probability
1996, Vol. 6, No. 2, 423–443

A CONSISTENT MODEL SELECTION PROCEDURE FOR MARKOV RANDOM FIELDS BASED ON PENALIZED PSEUDOLIKELIHOOD

BY CHUANSHU JI[1] AND LYNNE SEYMOUR

University of North Carolina and University of Georgia

Motivated by applications in texture synthesis, we propose a model selection procedure for Markov random fields based on penalized pseudolikelihood. The procedure is shown to be consistent for choosing the true model, even for Gibbs random fields with phase transitions. As a by-product, rates for the restricted mean-square error and moderate deviation probabilities are derived for the maximum pseudolikelihood estimator. Some simulation results are presented for the selection procedure.

1. Introduction. Markov random fields are widely used as models in statistical image analysis [cf. Karr (1991) and Rosenfeld (1993)]. Since Hassner and Sklansky (1980) and Cross and Jain (1983) first used isotropic Markov random fields to generate synthetic textures, others have explored different types of Markov random fields for texture synthesis. How does one choose a model from a collection of Markov random fields such that its typical sample resembles an observed texture? In this paper we present a model selection procedure based on penalized pseudolikelihood for Markov random fields in the form of an exponential family. It is shown that, asymptotically, this procedure chooses the correct model under very general conditions.

Little has been done to address selection of Markov random fields. Kashyap and Chellappa (1983) first proposed a method of selection based on linear combinations of gray levels plus Gaussian noise. Smith and Miller (1990) proposed a selection procedure which is based on the stochastic complexity of Rissanen (1984) and is similar to the one presented here. Seymour and Ji (1996) derived two Bayesian selection criteria [Akaike (1978); Schwarz (1978)]. The first is based on the maximum likelihood estimate; it is of theoretical interest, but is intractable for random fields. The other criterion uses the Markov chain Monte Carlo approximation to the likelihood developed by Geyer and Thompson (1992). Although Markov chain Monte Carlo criterion is viable, it is difficult to implement for images and requires that the random field exhibit weak spatial dependence.

In Section 2, the required random field framework is briefly introduced. Section 3 gives the formulation of the model selection problem and the main

Received November 1994; revised February 1996.
[1]Research supported in part by ONR Grant N00014-89-J-1760 and NSF Grant DMS-93-10322.
AMS 1991 *subject classification*. Primary 62M40; secondary 62F12, 68U10.
Key words and phrases. Markov random fields, Gibbs random fields, model selection, pseudolikelihood, texture synthesis, image analysis.

results. The selection procedure presented is based on the maximum pseudolikelihood estimate of Besag (1974). Although the criterion is similar to the Bayes criteria discussed above, it is not a Bayesian criterion. Even so, it has distinct advantages over the Bayes criteria; our criterion is much easier to compute and asymptotically it is shown to give a consistent choice of model, whether the spatial dependence is weak or strong. Since the spatial dependence involved in texture modelling may vary from short range to long range, the pseudolikelihood procedure has more extensive applications.

Because model selection and parameter estimation are closely related, the maximum pseudolikelihood parameter estimate and existing asymptotic results for this estimate are discussed in Section 4. In addition, two new lemmas are proven which provide rates for the restricted mean-square error and moderate deviation probabilities for the maximum pseudolikelihood estimate.

Section 5 presents and discusses some highlights of a simulation study of model selection via pseudolikelihood. Some concluding remarks are made in Section 6. All technical proofs are in the Appendix.

2. Random field framework. We consider Gibbs random fields induced by translation-invariant pair-potentials of finite range. The extensions to other finite-range translation-invariant potentials is straightforward but involves heavy notation. For a more general discussion of Gibbs random fields, see Georgii (1988).

With each site $i \in \mathbb{Z}^2$, associate a random variable X_i taking values in a finite set S. Then $X = \{X_i,\ i \in \mathbb{Z}^2\}$ is a random field with configuration space $\Omega = S^{\mathbb{Z}^2}$. Let $x = \{x_i,\ i \in \mathbb{Z}^2\} \in \Omega$ denote a realization of X. For a region $\Lambda \subset \mathbb{Z}^2$, the subconfiguration space is given by $\Omega_\Lambda = S^\Lambda$, so write $X_\Lambda = \{X_i,\ i \in \Lambda\}$ for the random field on Λ and $x_\Lambda = \{x_i,\ i \in \Lambda\} \in \Omega_\Lambda$ for a realization of X_Λ.

Let the potential $U = \{hU_1(x_o), \beta_j U_2(x_o, x_j)\colon\ x_o, x_j \in S;\ j \in \mathbb{Z}^2\}$, with o representing the origin, be a collection of functions such that $U_1\colon S \to \mathbb{R}$ and $U_2\colon S \times S \to \mathbb{R}$ are known and $U_2(s,t) = U_2(t,s)$. The term $hU_1(\cdot)$ (though not usually employed) may be used to model large-scale spatial trends, where $h \in \mathbb{R}$ (the external field coefficient) is an unknown parameter. The term $\beta_j U_2(\cdot,\cdot)$ is a pair-potential of range $R > 0$: the parameters $\beta_j \in \mathbb{R}$, $j \in \mathbb{Z}^2$ (the coupling coefficients) are also unknown and are such that $\beta_j = \beta_{-j}\ \forall\ j$ and $\beta_j = 0\ \forall\ j$ with $|j| > R$, where $|\cdot|$ is a norm on $\mathbb{Z}^2$. In particular, $\beta_o = 0$. Let θ denote the vector parameter with components being the external field and coupling coefficients.

The following examples are just two of the potentials that have been used for modelling with Markov random fields.

EXAMPLE 1. Consider the general Ising models, where $U_1(x_i) = x_i$, $U_2(x_i, x_j) = x_i x_j$ and $S = \{-1, 1\}$. If $\beta_j = \beta > 0$ for $|j| = 1$ and $\beta_j = 0$ otherwise, then we have the well-known two-dimensional Ising model.

EXAMPLE 2. Let $U_1(x_i) \equiv 0$ and $U_2(x_i, x_j) = 1/[1 + \sigma(x_i - x_j)^2]$, where $\sigma > 0$ is a constant. This potential is used in Geman and Graffigne (1986).

A Gibbs measure (Gibbs random field) induced by a potential U is a probability measure P on Ω such that for every $x \in \Omega$ and any finite $\Lambda \in \mathbb{Z}^2$,

$$P(X_\Lambda = x_\Lambda | X_{\Lambda^c} = x_{\Lambda^c}) = \frac{\exp[-H_\Lambda(x)]}{\mathscr{Z}_\Lambda}, \tag{2.1}$$

where the *energy* associated with x on Λ is given by

$$H_\Lambda(x) = -h \sum_{i \in \Lambda} U_1(x_i) - \tfrac{1}{2} \sum_{\substack{i, j \in \Lambda \\ 0 < |j-i| \le R}} \beta_{j-i} U_2(x_i, x_j) - \sum_{\substack{i \in \Lambda \\ j \notin \Lambda \\ |j-i| \le R}} \beta_{j-i} U_2(x_i, x_j)$$

and the normalizing factor, called the *partition function*, is given by

$$\mathscr{Z}_\Lambda = \mathscr{Z}_\Lambda(x_{\Lambda^c}) = \sum_{x_\Lambda} \exp[-H_\Lambda(x)].$$

The conditional probabilities $\{P(X_i = x_i | {}_iX =_i x),\ x \in \Omega\}$, are called the *local characteristics* at site $i \in \mathbb{Z}^2$, where ${}_iX = \{X_j,\ j \ne i\}$ and ${}_ix = \{x_j,\ j \ne i\}$. Indeed, the left-hand side of (2.1) is determined by the local characteristics at all $i \in \Lambda$ [Geman (1991)].

Under our assumptions on U, the set $\mathscr{G}(U)$ of Gibbs random fields induced by U is always non-empty, but need not be a singleton (in which case there are *phase transitions* of the Gibbs random field and in which case the random field exhibits spatial long-range dependence).

A neighborhood system $\mathfrak{N}$ is a collection $\{\mathscr{N}(i): i \in \mathbb{Z}^2\}$, where $\mathscr{N}(i) \subset \mathbb{Z}^2$ is the set of neighbors of $i \in \mathbb{Z}^2$ satisfying $i \notin \mathscr{N}(i)$ and $i \in \mathscr{N}(j) \Leftrightarrow j \in \mathscr{N}(i)$ $\forall\ i, j \in \mathbb{Z}^2$. Define the boundary of a finite region $\Lambda \subset \mathbb{Z}^2$ by $\partial\Lambda = (\bigcup_{i \in \Lambda} \mathscr{N}(i)) \setminus \Lambda$. Then every $P \in \mathscr{G}(U)$ is a Markov random field with respect to a neighborhood system $\mathfrak{N}$ in the sense that for every $x \in \Omega$ and any finite $\Lambda \subset \mathbb{Z}^2$,

$$P(X_\Lambda = x_\Lambda | X_{\Lambda^c} = x_{\Lambda^c}) = P(X_\Lambda = x_\Lambda | x_{\partial\Lambda} = x_{\partial\Lambda})$$

with $\mathscr{N}(i) = \{j \in \mathbb{Z}^2: \beta_{j-i} \ne 0\}$ for every i. In fact, a Markov random field on a finite lattice has a Gibbs representation [Hammersley–Clifford theorem in Geman (1991)].

We will be referring to the following examples, which illustrate the similarities and differences in specifying both the neighborhood and the parameter dimension. For these examples, let $U_1(x_i) \equiv 0$, $U_2(x_i, x_j) = x_i x_j$ and $S = \{-1, 1\}$.

EXAMPLE 3. The neighborhood system depicted in Figure 1, denoted $m1$, is for the two-dimensional Ising model. Each site i has four nearest neighbors. The same coupling coefficient β is imposed for each pair (i, j), $j \in \mathscr{N}(i)$.

$$\begin{array}{ccc} & \beta & \\ \beta & i & \beta \\ & \beta & \end{array}$$

FIG. 1. *Model* $m1$.

$$\begin{array}{ccc} & \beta_V & \\ \beta_H & i & \beta_H \\ & \beta_V & \end{array}$$

FIG. 2. *Model* $m2$.

$$\begin{array}{ccccc} & & \gamma & & \\ & \gamma & \beta & \gamma & \\ \gamma & \beta & i & \beta & \gamma \\ & \gamma & \beta & \gamma & \\ & & \gamma & & \end{array}$$

FIG. 3. *Model* $m3$.

EXAMPLE 4. For the model in Figure 2, denoted $m2$, every site i again has four nearest neighbors. However, two parameters β_V and β_H are used for "vertical pair" and "horizontal pair" interactions, respectively.

EXAMPLE 5. For the model in Figure 3, denoted $m3$, each site i has 12 neighbors that can be subdivided into two layers. The parameters β and γ are associated with the inner layer and the outer layer, respectively.

Write $\{p_i(x;\theta),\ x \in \Omega,\ i \in \mathbb{Z}^2\}$ for the local characteristics with parameter θ.

DEFINITION 1. The parameter θ is said to be *identifiable* if $p_o(x;\theta) \neq p_o(x;\theta')$ for some $x \in \Omega$ whenever $\theta \neq \theta'$.

REMARK. Identifiability may also be imposed via conditions on the potentials [Georgii (1988); Gidas (1993)] or by conditions on $\mathcal{G}(U)$ [Cométs (1992)].

For an $n \times n$ square lattice $\Lambda(n)$, let $x_{\Lambda(n)} = x(n)$ denote a single realization of $X_{\Lambda(n)} = X(n)$, where X has a distribution $P \in \mathcal{G}(U)$. Write P_θ for P to indicate the parameterization and write $E_\theta(\cdot)$ for the expectation with respect to P_θ. Extend the observation $x(n)$ to a configuration $\tilde{x}$ on $\mathbb{Z}^2$ by periodization, or *tiling* [toroidal edge correction; Ripley (1981)], as illustrated in Figure 4. Correspondingly, let $\tilde{X}$ denote the periodic random field based on $X_{\Lambda(n)}$.

Define the *pseudolikelihood function* [Besag (1974)], a product of the local characteristics of the sites of $\Lambda(n)$, as

$$\mathcal{PL}(x(n),\theta) = \prod_{i \in \Lambda(n)} P_\theta\big(X_i = \tilde{x}_i |_i X =_i \tilde{x}\big).$$

Any measurable function of $x(n)$ which maximizes $\mathcal{PL}(x(n),\cdot)$ is called a maximum pseudolikelihood estimate of θ based on $x(n)$. We denote this estimate by $\tilde{\theta}$.

There are several motivating factors for using the maximum pseudolikelihood estimate of θ. A practical one is that the local characteristics are quickly

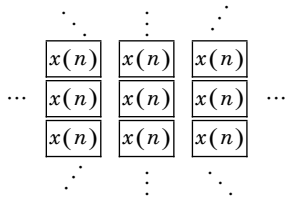

FIG. 4. *Tiling*.

and easily computed. An intuitive one is that the local geometry of an image may be reasonably summarized by the local characteristics. A theoretical one is that its existence, uniqueness and consistency have been proven by Geman and Graffigne (1986); independently, Gidas (1988) and Cométs (1992) have established its consistency.

3. The model selection problem and a consistency result. Specification of a potential (i.e., selecting a Markov random field model) consists of the interconnected parts of specifying both the neighborhood system $\mathfrak{N}$ and the dimension of the parameter θ.

Let $\Theta = \mathbb{R}^K$ be the parameter space of interest, decomposed as the disjoint union $\Theta = \bigcup_{m=0}^{M} \Theta_M$, $\Theta_m \cap \Theta_{m'} = \varnothing \ \forall\ m \neq m'$, where each Θ_m corresponds to a candidate model (i.e., a potential) parameterized by an element of $\mathbb{R}^{k_m}$. We assume that every closure $\overline{\Theta}_m$ is a k_m-dimensional linear subspace of $\mathbb{R}^K$, $m = 0, 1, \ldots, M$ [cf. Schwarz (1978)]. In particular, Θ_0 corresponds to the completely specified model with no unknown parameter. Denote the set of all candidate models by $\mathscr{M} = \{0, 1, \ldots, M\}$ and let $\mathfrak{N}_m$ be the neighborhood system for the model $m \in \mathscr{M}$.

In Examples 3, 4 and 5, one sees that $\mathfrak{N}_{m1} = \mathfrak{N}_{m2} \neq \mathfrak{N}_{m3}$ and that $k_{m1} = 1$, while $k_{m2} = k_{m3} = 2$. Several synthetic textures generated from $m1$, $m2$ and $m3$ by the Gibbs sampler [Geman and Geman (1984)] are shown in Figures 5–10. The coupling coefficients are assigned different values to produce different imaginary patterns of both weak and strong spatial dependence: "sands" (Figure 5), "clouds" (Figure 6), "wood grain" (Figure 8) and "wall papers" (Figures 7, 9 and 10). Note that samples from such simple models are far from resembling real textures.

In general, starting from $\theta \in \Theta$, a different model can be obtained either by equating some components in θ (e.g., letting $\beta_V = \beta_H \triangleq \beta$ in $m2$ to obtain $m1$) or by setting some components to zero (e.g., letting $\gamma = 0$ in $m3$ to obtain $m1$). In this way the pseudolikelihood, when written in the form of an exponential family, may be reduced to its minimal form [cf. Barndorff-Nielsen (1978); Brown (1986)].

For each $m \in \mathscr{M}$, let $\tilde{\theta}_m$ be the maximum pseudolikelihood estimate restricted to $\overline{\Theta}_m$. Let $\mathscr{PL}_m(\cdot, \cdot)$ denote the pseudolikelihood for model $m \in \mathscr{M}$ in minimal form: $\mathscr{PL}_m(x(n), \theta) = \exp\{[\Lambda(n)][\theta^{\mathsf{T}} V_m - g_m(\theta)]\}$, where V_m and $g_m(\cdot)$ denote functions analogous to the sufficient statistic and cumulant

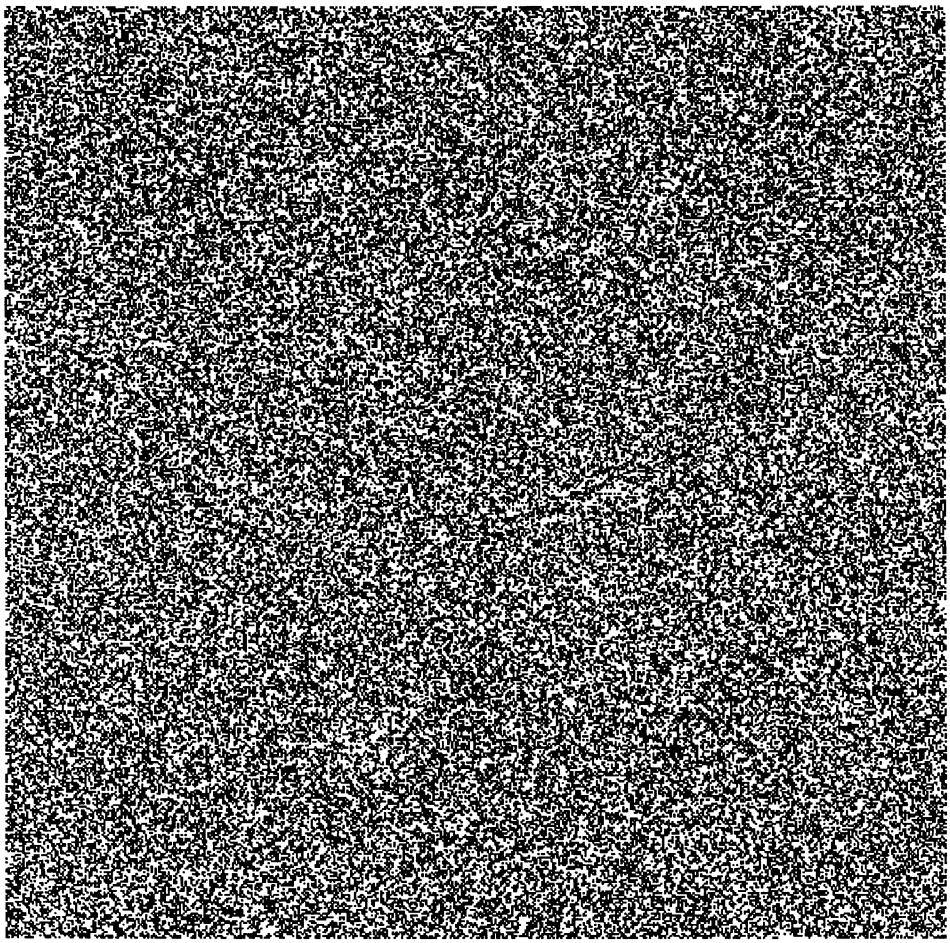

FIG. 5. *Model* $m1$, $\beta = 0.1$.

generating function, respectively. Define the information criterion as

$$Q_m = \sup_{\vartheta \in \overline{\Theta}_m} \log \mathscr{PL}_m(x(n), \vartheta) - \frac{k_m}{2} \log|\Lambda(n)|.$$

Then the pseudolikelihood selection procedure is to choose the model $\hat{m} \in \mathscr{M}$ which maximizes Q_m.

Decompose the collection of candidate models as $\mathscr{M} = \mathscr{M}_1(\pi) \cup \{\pi\} \cup \mathscr{M}_2(\pi)$, where $\pi \in \mathscr{M}$ is the true model, $\theta \in \Theta_\pi$ is the true parameter which is assumed to be identifiable (see Definition 1), $\mathscr{M}_1(\pi) = \{m \in \mathscr{M}: \theta \notin \overline{\Theta}_m\}$ and $\mathscr{M}_2(\pi) = \{m \in \mathscr{M}: \overline{\Theta}_\pi \subset \overline{\Theta}_m\}$. Here $\mathscr{M}_1(\pi)$ corresponds to an underparameterized choice of model or to an incorrect specification of neighborhood system (different neighborhoods will correspond to different subspaces which may

FIG. 6. *Model* $m1$, $\beta = 1.0$.

have the same dimension), while $\mathscr{M}_2(\pi)$ corresponds to an overparameterized choice. Note particularly that $\overline{\Theta}_\pi$ is a *proper* subset of $\overline{\Theta}_m$ if $m \in \mathscr{M}_2(\pi)$ and that our decomposition of $\mathscr{M}$ leaves out no choice of model, since we have decomposed the parameter space Θ into a *disjoint* union of subspaces Θ_m, $m \in \mathscr{M}$, earlier in this section.

Denote the selection procedure which chooses a model $\hat{m}$ based on $x(n)$ by $\hat{m} = d(x(n))$, where d: $\Omega_{\Lambda(n)} \to \mathscr{M}$ denotes the decision function.

DEFINITION 2. A selection procedure $d(\cdot)$ is said to be consistent if $\lim_{n\to\infty} P_\theta(d(X(n)) = \pi) = 1$, where $X(n)$ is a sample from P_θ, $\theta \in \Theta_\pi$, $\pi \in \mathscr{M}$.

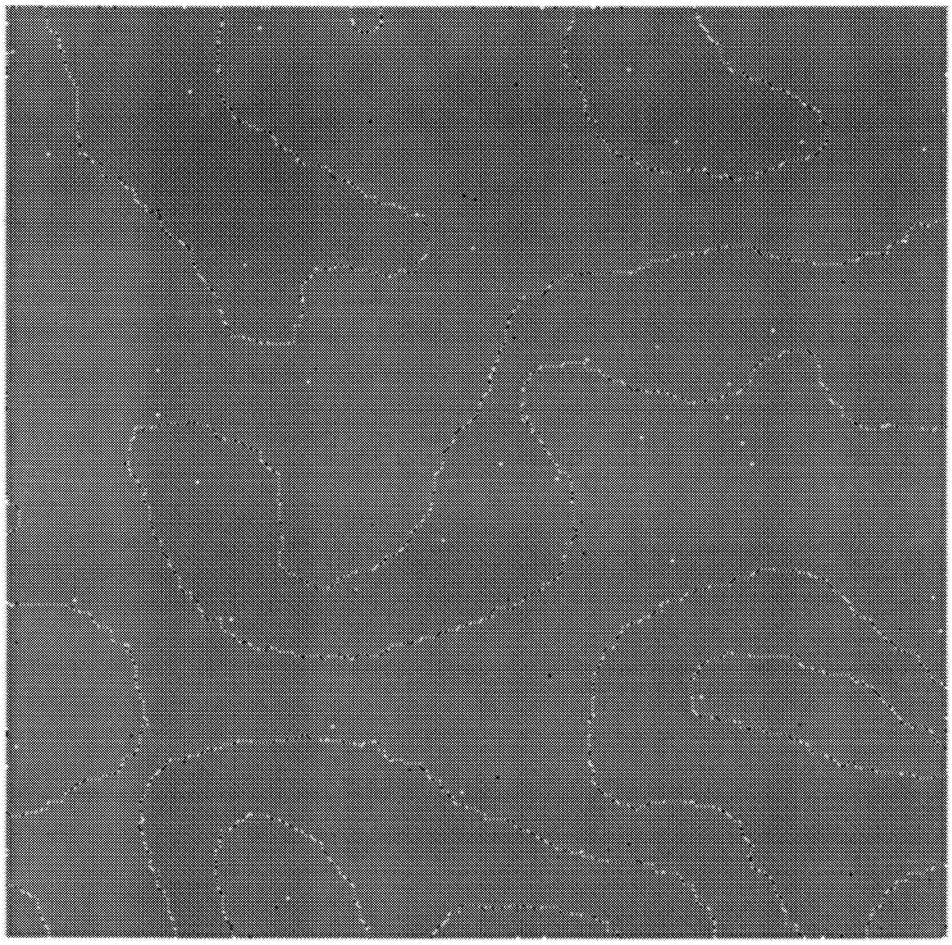

FIG. 7. *Model* $m1$, $\beta = -1.0$.

The following two propositions give decay rates for the probabilities of choosing an incorrect model in $\mathscr{M}_1(\pi)$ and in $\mathscr{M}_2(\pi)$, respectively.

PROPOSITION 1. *There exists* $c > 0$ *such that* $P_\theta(\hat{m} \in \mathscr{M}_1(\pi)) \leq \exp(-n^c)$ *for sufficiently large* n.

PROPOSITION 2. *There exists* $\alpha > 0$ *such that* $P_\theta(\hat{m} \in \mathscr{M}_2(\pi)) = O(n^{-\alpha})$ *as* $n \to \infty$.

The following theorem is an immediate consequence of Propositions 1 and 2.

THEOREM 1. *The selection procedure based on* Q_m *is consistent*.

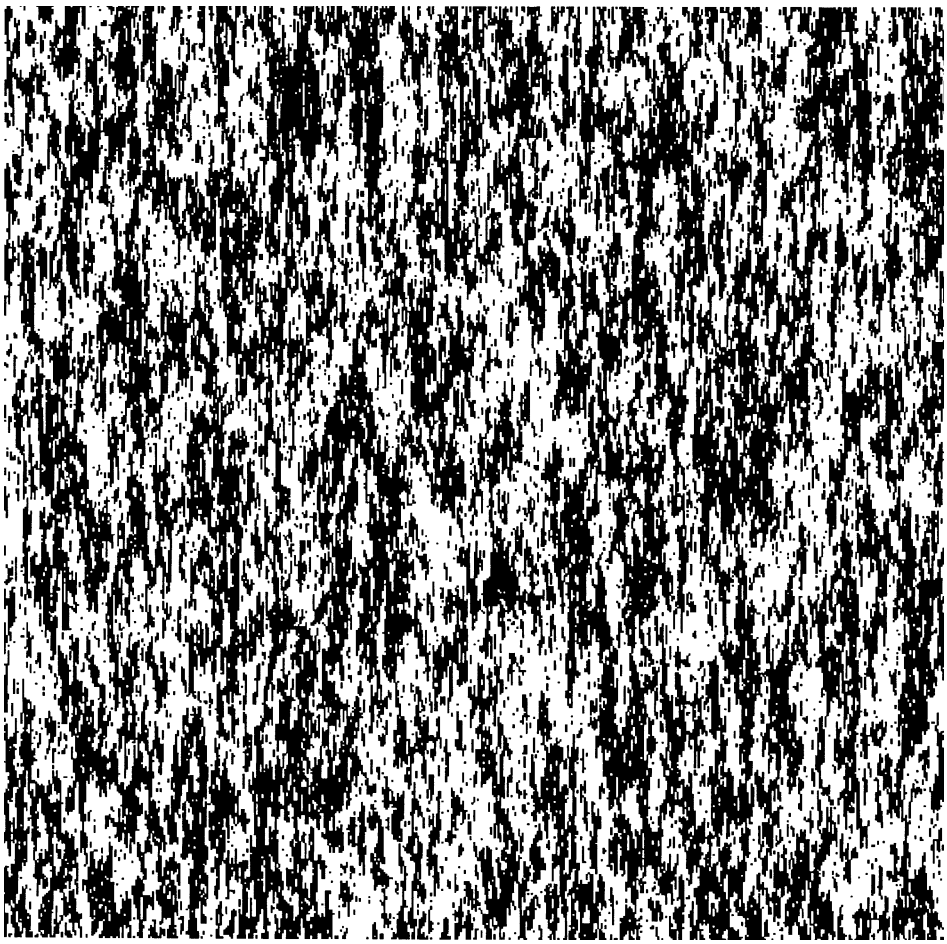

FIG. 8. *Model m*2, $\beta_1 = 1.0$, $\beta_2 = 0.1$.

4. Some properties of the maximum pseudolikelihood estimator. Because parameter estimation is such an important part of model selection, some asymptotic properties of the maximum pseudolikelihood estimator are discussed in this section. In particular, Lemmas 3 and 4 in this section provide some asymptotic orders of consistency for the maximum pseudolikelihood estimator; these are crucial in proving the consistency for the selection procedure.

Fix a model $m \in \mathscr{M}$ and a parameter $\theta \in \Theta_m$ and suppress the notation indicating the model in this section and in the corresponding proofs in the Appendix. Recall the pseudolikelihood in exponential family form: the "sufficient statistic" is given by

$$V = \frac{1}{|\Lambda(n)|} \sum_{i \in \Lambda(n)} Z\left(\tilde{x}_i, \tilde{x}_{\mathscr{N}(i)}\right)$$

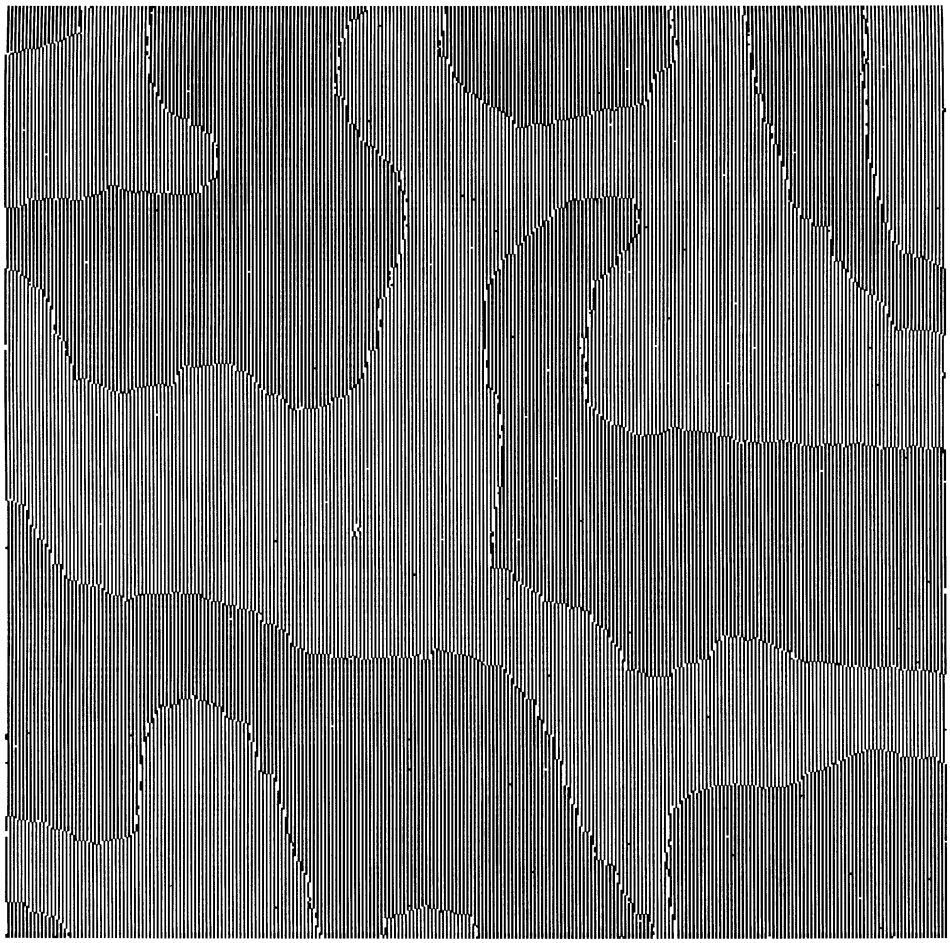

FIG. 9. *Model m*2, $\beta_1 = 1.0$, $\beta_2 = -1.0$.

for some function $Z(\cdot, \cdot)$ of the appropriate potentials, and the "cumulant generating function" is given by

$$g(\theta) = \frac{1}{|\Lambda(n)|} \sum_{i \in \Lambda(n)} \log \sum_{s \in S} \exp\{\theta^T Z(s, \tilde{x}_{\mathcal{N}(i)})\}.$$

The gradient of $g(\vartheta)$ with respect to ϑ is given by

$$\nabla g(\vartheta) = \frac{1}{|\Lambda(n)|} \sum_{i \in \Lambda(n)} E_{\vartheta}(Z|\tilde{x}_{\mathcal{N}(i)}),$$

where ϑ specifically denotes a variable.

For each $i \in \Lambda(n)$, let $\Lambda(i, R)$ be the $(2R + 1) \times (2R + 1)$ square lattice centered at i, where R is the range of the Gibbs distribution. In particular,

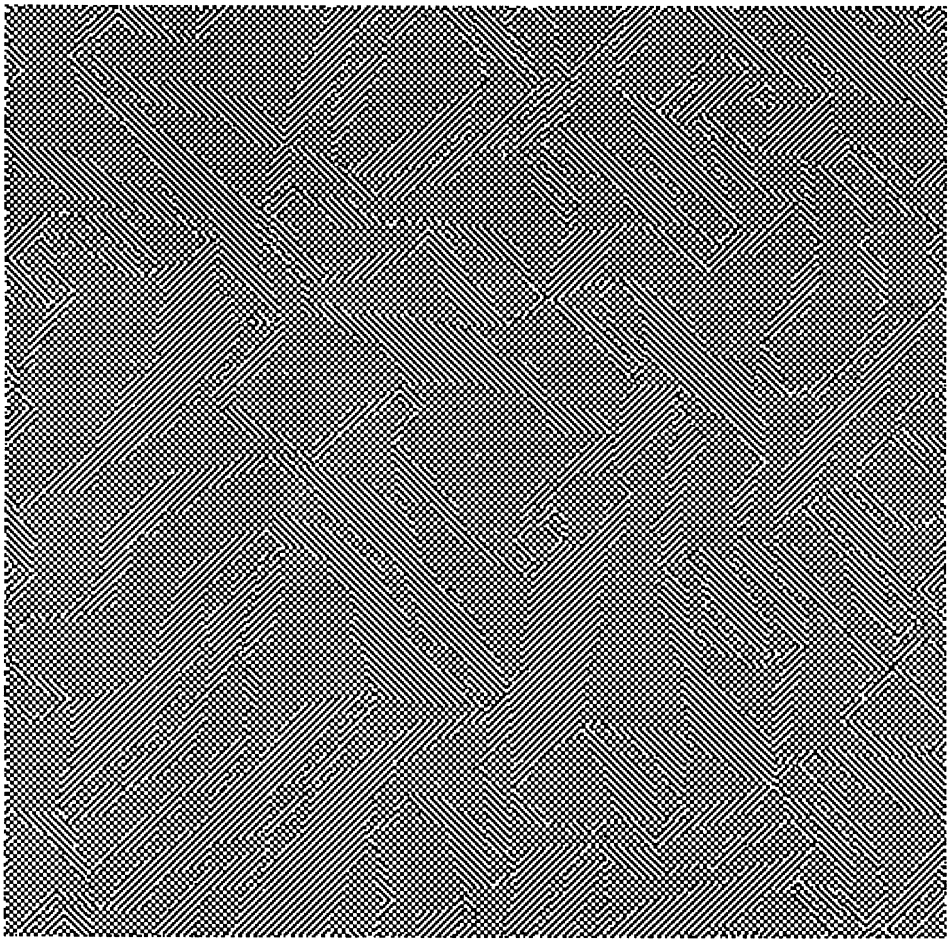

FIG. 10. *Model* $m3$, $\beta = 1.0$, $\gamma = -1.0$.

denote $\Lambda(o, R) = \Lambda(2R + 1)$. Let $\xi \in S$ and $\eta \in \Omega_{\Lambda(o,R)\setminus\{o\}}$, so that the combined configuration is $\xi \oplus \eta \in \Omega_{\Lambda(2R+1)}$. Define

$$\mathbb{I}_i(\xi \oplus \eta) = 1_{\{\tilde{X}_{\Lambda(i,R)} = \xi \oplus \eta\}},$$
$$\mathbb{I}_i(\eta) = 1_{\{\tilde{X}_{\Lambda(i,R)\setminus\{i\}} = \eta\}} \quad \text{for } i \in \Lambda(n)$$

and

$$N_n(\xi \oplus \eta) = \sum_{i \in \Lambda(n)} \mathbb{I}_i(\xi \oplus \eta),$$
$$N_n(\eta) = \sum_{i \in \Lambda(n)} \mathbb{I}_i(\eta).$$

Define the event

$$\mathscr{A}(n) = \left\{ x(n) \in \Omega_{\Lambda(n)} \colon \frac{N_n(\xi \oplus \eta)}{|\Lambda(n)|} \geq \lambda \quad \forall\, \xi \oplus \eta \in \Omega_{\Lambda(2R+1)} \right\}$$

on which the empirical probabilities for all configurations in $\Omega_{\Lambda(n)}$ are bounded away from 0. The complement of this set is negligible for large n.

LEMMA 1. *There exist positive constants λ, c and C such that*

$$P_\theta\left(\frac{N_n(\xi \oplus \eta)}{|\Lambda(n)|} < \lambda \right) \leq C \exp(-cn)$$

for all large n and all $\xi \oplus \eta \in \Omega_{\Lambda(2R+1)}$.

Hence, the following lemma is restricted to $\mathscr{A}(n)$.

LEMMA 2. *There exist $c, C > 0$ such that $c \leq v^T \nabla^2 g(\vartheta) v \leq C$ for all unit vectors $v \in \mathbb{R}^{k_m}$, all $\vartheta \in \Theta_m$ in a neighborhood of θ, all $x(n) \in \mathscr{A}(n)$ and all large n.*

REMARK. The following is a simple argument for the existence and uniqueness of the maximum pseudolikelihood estimator. The "pseudolikelihood" equation is given by $V = \nabla g(\vartheta)$. Now, for all $\vartheta \in \mathbb{R}^{k_m}$, it can be shown that $E_\vartheta[V - \nabla g(\vartheta)] = 0$, so that by Theorem 14.A8 of Georgii (1988), we have

$$\lim_{n \to \infty} [V - \nabla g(\theta)] = 0, \quad P_\theta\text{-a.s.} \tag{4.1}$$

By Lemma 2, there exists a small neighborhood of θ, say $\mathscr{O}$, on which $\nabla g(\cdot)$ is a homeomorphism. Then, for large n, we have $V \in \nabla g(\mathscr{O})$ by (4.1). Thus there exists $\omega \in \mathscr{O}$ satisfying the pseudolikelihood equation, $V = \nabla g(\omega)$, P_θ-a.s. Since $g(\cdot)$ is globally convex [see (A.1) in the proof of Lemma 2 in the Appendix] and locally strictly convex by Lemma 2, the solution ω is the unique maximum pseudolikelihood estimate $\tilde{\theta}$.

The next two lemmas provide asymptotic orders for the restricted mean squared error and moderate deviation probabilities for the maximum pseudolikelihood estimate.

LEMMA 3. $E_\theta\{\|\tilde{\theta} - \theta\|^2 1_{\mathscr{A}(n)}\} = O(|\Lambda(n)|^{-1})$ *as* $n \to \infty$.

REMARK. Theorem 1 may also be proven by Proposition 1, Lemma 3 and the Chebyshev inequality. However, the decay rate of the probability of choosing an incorrect model produced by this method can only be of the order $1/\log n$—a rate inferior to the one inferred by using Propositions 1 and 2.

LEMMA 4. *For every* $\varepsilon > 0$ *there exists* $\alpha > 0$ *such that*

$$P_\theta\big(|\Lambda(n)|\|\tilde{\theta} - \theta\|^2 > \varepsilon \log n\big) = O(n^{-\alpha})$$

as $n \to \infty$.

REMARK. It is noteworthy that the constant α in Lemma 4, which is the same α as in Proposition 2, cannot be made greater than 1 in general. This precludes the use of the Borel–Cantelli lemma in an effort to prove the strong consistency of the pseudolikelihood selection procedure. [A procedure $d(\cdot)$ is said to be *strongly consistent* if $d(X(n)) \to \pi$, P_θ-a.s. as $n \to \infty$, where $X(n)$ is a sample from P_θ, $\theta \in \Theta_\pi$, $\pi \in \mathcal{M}$.] This observation is supported by the exact order of moderate deviation probabilities in the i.i.d. case given in Rubin and Sethuraman (1965).

5. Some simulation results. Although there is an extensive literature in various Markov chain simulation algorithms, we use the Gibbs sampler [Geman and Geman (1984)] for simulating textures. In our simulation studies, we have used the three models $m1$, $m2$ and $m3$ which were introduced in Section 3. For convenience, we have omitted including a "largest" model among the candidates (cf. Proof of Proposition 1 in the Appendix). We study the pseudolikelihood procedure for 500×500 random fields with the neighborhood interactions varying from weak to strong. In the tables we present, β_1 corresponds to β for $m1$ and $m3$ and β_V for $m2$, while β_2 corresponds to β_H for $m2$ and γ for $m3$. The symbol $**$ indicates the chosen model.

The pseudolikelihood procedure seems to work well in all cases, given the similarity of the candidate models. When neighborhood interactions are weak, as in Table 1, there are no phase transitions and an identifying structure cannot be discerned in a realization. In such cases, the pseudolikelihood procedure tends to overparametrize—in fact, the values of Q_m do not vary much among the models. Also, the sample from the true model was practically indistinguishable from a sample from $m1$ with no phase transitions (indeed, $m1$ is a special case of both). On the other hand, when neighborhood interactions are stronger, phase transitions are possible. The procedure still tends to overparametrize when the true model is $m1$ (seen in Table 2), but the chosen model is close to the true model and Q_m again does not vary much. As seen in Table 3, the pseudolikelihood procedure worked extremely well when structures unique to the model are easily discernible, making a clear (i.e., one value of Q_m is much larger than the others) and correct choice over all other candidates.

REMARK. These same phenomena may be observed for $m3$ [Seymour (1993)].

6. Concluding remarks. The model selection procedure proposed in this paper has a similar expression to that of the Bayesian information criterion [Schwarz (1978); Akaike (1978)] with the likelihood replaced by the

TABLE 1
Weak neighborhood interactions

Model	$\tilde{\beta}_1$	$\tilde{\beta}_2$	Q_m
	$\beta = 0.1$		
1	0.09880	—	−167024
2	0.10120	0.09651	−167028
**3	0.09921	−0.00108	−165694
	$\beta_V = 0.01$, $\beta_H = 0.1$		
1	0.05414	—	−170461
2	0.01175	0.09568	−169606
**3	0.05403	−0.00153	−169099
	$\beta_V = 0.1$, $\beta_H = 0.01$		
1	0.05557	—	−170388
2	0.10343	0.00657	−169255
**3	0.05554	−0.00229	−169032

TABLE 2
*Strong neighborhood interactions m*1

Model	$\tilde{\beta}_1$	$\tilde{\beta}_2$	Q_m
	$\beta = 1$		
1	1.0176	—	−2523
2	1.0325	1.0027	−2529
**3	1.0676	−0.02667	−2485
	$\beta = -1$		
1	−1.0242	—	−2702
2	−1.0053	−1.0425	−2708
**3	−0.97953	−0.02614	−2678
	$\beta = 2$		
1	1.9805	—	−1206
2	6.3686	1.7551	−1210
**3	1.9371	0.04030	−1201

pseudolikelihood in the first term and the same penalty for overparameterization in the second term. A similar modification of Akaike's information criterion [Akaike (1974)] may be considered. In the i.i.d. case, Woodroofe (1982) pointed out that Akaike's criterion is superior to the Bayesian criterion 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 Markov random field texture models if we let the range of the potential $R = R_n \to \infty$; however, more delicate asymptotics are needed to accomplish this, and the result in Ji (1990) may be helpful.

TABLE 3
Strong neighborhood interactions m2

Model	$\tilde{\beta}_1$	$\tilde{\beta}_2$	Q_m
	$\beta_V = 0.1, \beta_H = 1$		
1	0.54126	—	−64924
**2	0.09975	1.0074	−44670
3	0.74073	−0.12326	−60873
	$\beta_V = 1, \beta_H = 0.1$		
1	0.54077	—	−65320
**2	1.0010	0.10003	−45135
3	0.73140	−0.11896	−61489
	$\beta_V = 1, \beta_H = -1$		
1	−0.01349	—	−171907
**2	0.97933	−1.0521	−2379
3	−0.01865	−0.18243	−169900
	$\beta_V = -1, \beta_H = 1$		
1	−0.02736	—	−171902
**2	−0.98039	1.0469	−2398
3	−0.03704	−0.19437	−169833

The asymptotic distributions for the indices Q_m, $m \in \mathcal{M}$, may also be investigated. The need for such was demonstrated in Woodroofe (1982), in which the distribution of the number of superfluous parameters contained in the selected model was found. Such a result could be used to make numerical comparisons between different models. The derivation may not be too difficult under Dobrushin's uniqueness condition for Gibbs random fields [Georgii (1988)]. However, the derivation is very challenging under the assumptions we have made in this paper due to the lack of a central limit theorem for Gibbs random fields under phase transitions.

Extensive simulation studies are still being done for real texture synthesis, and many issues remain open. A rich class of candidate potentials is required for using Markov random field models for texture synthesis. Our approach in this paper has been to consider a great variety of neighborhood systems. The recent approach of Künsch, Geman and Kehagias (1995) is to code each site variable in a complex manner while restricting to the nearest neighbors. Both of these approaches involve extremely intense computation. Current research still has yet to achieve the ideal of a convenient statistical method for replicating real textures.

APPENDIX

LEMMA A.1. *Let R be the range of the Gibbs random field. Let $\mathcal{B}(1), \ldots, \mathcal{B}(T)$ be bounded regions in $\mathbb{Z}^2$, $T \in \mathbb{N}$, with the distances between $\mathcal{B}(t)$ and $\mathcal{B}(t')$ greater than R for all $t \neq t'$. Also, let $\mathcal{C} = \mathbb{Z}^2 \setminus (\bigcup_{t=1}^{T} \mathcal{B}(t))$ be the*

corridor between these regions. Then for any collection of bounded measurable functions $f_t\colon \Omega_{\mathscr{B}(t)} \to \mathbb{R}$, $t = 1, \ldots, T$, *we have*

$$E_\theta\left\{ \prod_{t=1}^{T} f_t(X_{\mathscr{B}(t)}) \middle| x_{\mathscr{C}} \right\} = \prod_{t=1}^{T} E_\theta\left[f_t(X_{\mathscr{B}(t)}) \middle| x_{\mathscr{C}} \right]$$

uniformly for all corridor configurations $x_{\mathscr{C}} \in \Omega_{\mathscr{C}}$, *where* $E_\theta(\cdot|x_{\mathscr{C}})$ *is the conditional expectation with respect to* $P_\theta(\cdot|x_{\mathscr{C}})$.

PROOF. This result follows from the Markov property of X. □

PROOF OF LEMMA 1. Assume without loss of generality that $(3R + 1)$ divides n. Partition $\Lambda(n)$ as a union of disjoint tiles $\Lambda(n) = \bigcup_{t=1}^{T} D(t)$, so that each tile $D(t)$ is a $(3R + 1) \times (3R + 1)$ square lattice. Then $T = [n/(3R + 1)]^2$.

Also, write the decomposition $\Lambda(n) = \bigcup_{k=1}^{(3R+1)^2} G(k)$, where every $G(k)$ contains exactly T sites with the same relative positions in the disjoint tiles $D(t)$, $t = 1, \ldots, T$. For instance, one $G(k)$ may consist of the centers of the T tiles, while another $G(k)$ may consist of all upper left corners of the T tiles. Therefore, $N_n(\xi \oplus \eta) = \sum_{k=1}^{(3R+1)^2} \sum_{i \in G(k)} \mathbb{I}_i(\xi \oplus \eta)$ and for every $\xi \oplus \eta \in \Omega_{\Lambda(2R)}$ we have

$$P_\theta\left(\frac{N_n(\xi \oplus \eta)}{|\Lambda(n)|} < \lambda \right) \le \exp(-\lambda n) \sum_{k=1}^{(3R+1)^2} E_\theta\left[\exp\left\{ -\frac{1}{n} \sum_{i \in G(k)} \mathbb{I}_i(\xi \oplus \eta) \right\} \right].$$

For a fixed index k, let $\mathscr{C}(k) = \mathbb{Z}^2 \setminus (\bigcup_{i \in G(k)} \Lambda(i, R))$ be the corridor dividing the regions $\Lambda(i, R)$, $i \in G(k)$. Then

$$E_\theta\left[\exp\left\{ -\frac{1}{n} \mathbb{I}_i(\xi \oplus \eta) \right\} \middle| x_{\mathscr{C}(k)} \right] \le 1 - \frac{c_1}{n}$$

for some $c_1 > 0$ and all large n. Therefore, employing Lemma A.1,

$$P_\theta\left(\frac{N_n(\xi \oplus \eta)}{|\Lambda(n)|} < \lambda \right) \le C \exp(-cn)$$

for some $C > 0$ and some $c > 0$. □

PROOF OF LEMMA 2. Define

$$K(\vartheta, n) = \vartheta^T V - g(\vartheta) = |\Lambda(n)|^{-1} \log \mathscr{PL}(x(n), \vartheta).$$

Let $p_o(\xi|\eta; \vartheta)$ be the local characteristic at the origin, where $\xi \oplus \eta \in \Omega_{\Lambda(2R+1)}$, $\xi \in S$ is the value at the origin o and $\eta \in \Omega_{\Lambda(2R+1)\setminus\{o\}}$. Then, via translation invariance, we may write

$$K(\vartheta, n) = \sum_{\eta} \frac{N_n(\eta)}{|\Lambda(n)|} \sum_{\xi} \frac{N_n(\xi \oplus \eta)}{N_n(\eta)} \log p_o(\xi|\eta; \vartheta).$$

Write the local characteristic at the origin in exponential family form

$$p_o(\xi|\eta;\vartheta) = \frac{\exp\{\vartheta^T\phi(\xi\oplus\eta)\}}{\sum_s \exp\{\vartheta^T\phi(s\oplus\eta)\}},$$

where $\vartheta \in \mathbb{R}^{k_m}$ and $\phi(\cdot)$ is an appropriate vector-valued function. Because $\sum_\xi N_n(\xi\oplus\eta) = N_n(\eta)$ for fixed η, we have

$$-\nabla^2 K(\vartheta,n) = \sum_\eta \frac{N_n(\eta)}{|\Lambda(n)|} E_\vartheta\Big[\big(\phi(X_o\oplus\eta) - E_\vartheta[\phi(X_o\oplus\eta)|\eta]\big) \times\big(\phi(X_o\oplus\eta) - E_\vartheta[\phi(X_o\oplus\eta)|\eta]\big)^T\Big|\eta\Big],$$

where $E\vartheta(\cdot|\eta)$ is the conditional expectation with respect to $p_o(\cdot|\eta;\vartheta)$. Hence for $v\in\mathbb{R}^{k_m}$, we have

$$\begin{aligned} &v^T\nabla^2 g(\vartheta)v \\ &\quad= -v^T\nabla^2 K(\vartheta,n)v \\ &\quad= \sum_\eta \frac{N_n(\eta)}{|\Lambda(n)|} E_\vartheta\Big\{\big[v^T\big(\phi(X_o\oplus\eta) - E_\vartheta[\phi(X_o\oplus\eta)|\eta]\big)\big]^2\Big|\eta\Big\}. \end{aligned} \tag{A.1}$$

Since this expectation is bounded and there are only finitely many $\eta\in\Omega_{\Lambda(2R+1)\setminus\{o\}}$, we see that $v^T\nabla^2 g(\vartheta)v \le C$ for some $C>0$.

On the other hand, the identifiability of θ guarantees that the "outside" expectation is strictly positive for at least one of the η-configurations. Also, for fixed $\xi\in S$ and $\eta\in\Omega_{\Lambda(2R+1)\setminus\{o\}}$, it should be clear that $N_n(\eta)\ge N_n(\xi\oplus\eta)$. Then on $\mathscr{A}(n)$,

$$\frac{N_n(\eta)}{|\Lambda(n)|} \ge \frac{N_n(\xi\oplus\eta)}{|\Lambda(n)|} \ge \lambda > 0$$

for all η-configurations. Therefore, $v^T\nabla^2 g(\vartheta)v \ge c$ for some $c>0$. □

PROOF OF LEMMA 3. The Taylor expansion of $\nabla K(\vartheta,n)$ about $\tilde\theta$ gives

$$\begin{aligned} \nabla K(\theta,n) &= \nabla K(\tilde\theta,n) + \nabla^2 K(\vartheta',n)(\theta-\tilde\theta) \\ &= -\nabla^2 g(\vartheta')(\theta-\tilde\theta) \end{aligned}$$

for some ϑ' satisfying $\|\vartheta'-\tilde\theta\| \le \|\theta-\tilde\theta\|$. By Lemma 2,

$$\|\tilde\theta-\theta\|^2 \le C\|\nabla K(\theta,n)\|^2 \tag{A.2}$$

on $\mathscr{A}(n)$ for some $C>0$. Thus $E_\theta\{\|\tilde\theta-\theta\|^2\mathbf{1}_{\mathscr{A}(n)}\} \le CE_\theta\{\|\nabla K(\theta,n)\|^2\}$. Rewrite $K(\theta,n)$ as

$$K(\theta,n) = \frac{1}{|\Lambda(n)|}\sum_{i\in\Lambda(n)}\Big(\sum_{\xi\oplus\eta}\mathbb{I}_i(\xi\oplus\eta)\log p_o(\xi|\eta;\theta)\Big)$$

so that

$$\nabla K(\theta, n) = \frac{1}{|\Lambda(n)|} \sum_{i \in \Lambda(n)} W_i, \tag{A.3}$$

where W_i is the vector

$$W_i = \sum_{\eta} \mathbb{I}_i(\eta)\big(\phi(X_i \oplus \eta) - E_\theta[\phi(X_i \oplus \eta)|\eta]\big).$$

For each $i \in \Lambda(n)$, all of the components of W_i are bounded. Let w_i denote an arbitrary component of W_i. Then it is sufficient to show

$$E_\theta\left\{\left(\sum_{i \in \Lambda(n)} w_i\right)^2\right\} = O(|\Lambda(n)|).$$

Using the decompositions of $\Lambda(n)$ from the proof of Lemma 1, it is enough to show

$$E_\theta\left\{\left(\sum_{i \in G(k)} w_i\right)^2\right\} = O(|\Lambda(n)|)$$

for each $G(k)$. Let $\mathscr{C}(k)$ be a corridor as constructed in the proof of Lemma 1. Then $E_\theta(W_i|x_{\mathscr{C}(k)}) = 0$ for $i \in G_k$ and every configuration $x_{\mathscr{C}(k)}$. Since the elements w_i are bounded, Lemma A.1 gives

$$E_\theta\left\{\left(\sum_{i \in G(k)} w_i\right)^2\right\} \le C|G_k|$$

for some $C > 0$. Because $|G(k)| = |\Lambda(n)|(3R + 1)^{-2}$, the result clearly follows. □

PROOF OF LEMMA 4. By (A.2) and Lemma 1, it suffices to show that for every $\varepsilon > 0$,

$$P_\theta\big(|\Lambda(n)|\|\nabla K(\theta, n)\|^2 > \varepsilon \log n\big) = O(n^{-\alpha}).$$

Using the notation in the proof of Lemma 3, it is enough to show that for every $\varepsilon > 0$,

$$P_\theta\left(\left|\sum_{i \in G(k)} w_i\right| > \varepsilon\tau_n\right) = O(n^{-\alpha}) \tag{A.4}$$

for some $\alpha > 0$, where $\tau_n = \sqrt{|\Lambda(n)|\log n}$.

Consider the two cases for the absolute value, studying first the positive case. For $\rho > 0$ (to be specified),

$$P_\theta\Big(\sum_{i \in G(k)} w_i > \varepsilon\tau_n\Big) \le \exp\big(-\rho\varepsilon\sqrt{\tau_n}\big) E_\theta\left[\exp\left(\sum_{i \in G(k)} \frac{\rho w_i}{\sqrt{\tau_n}}\right)\right].$$

Using Lemma A.1 and the construction of the corridor $\mathscr{C}(k)$ in the proof of Lemma 1,

$$E_\theta\left[\exp\left(\sum_{i\in G(k)}\frac{\rho w_i}{\sqrt{\tau_n}}\right)\right] = E_\theta\left\{\prod_{i\in G(k)} E_\theta\left[\exp\left(\frac{\rho w_i}{\sqrt{\tau_n}}\right)\middle| X_{\mathscr{C}(k)}\right]\right\}.$$

From the proof of Lemma 3, $E_\theta(w_i|x_{\mathscr{C}(k)}) = 0$ for every corridor configuration $x_{\mathscr{C}(k)}$, so that

$$P_\theta\left(\sum_{i\in G(k)} w_i > \varepsilon\tau_n\right) \le \exp\left(-\rho\varepsilon\sqrt{\tau_n}\right)\exp\left(\frac{\alpha'\rho^2|G(k)|}{\tau_n}\right)$$

for some $\alpha' > 0$. Let $\alpha'' = \alpha'/(3R+1)^2$ and recall that $|G(k)| = |\Lambda(n)|/(3R+1)^2$. Set $\rho = \varepsilon n^{-1/2}(\log n)^{3/4}(2\alpha'')^{-1}$ and note that $|\Lambda(n)| = n^2$. Then

$$P_\theta\left(\sum_{i\in G(k)} w_i > \varepsilon\tau_n\right) = O(n^{-\alpha})$$

with $\alpha = \varepsilon^2/4\alpha''$.

For the negative case,

$$P_\theta\left(-\sum_{i\in G(k)} w_i > \varepsilon\tau_n\right) = O(n^{-\alpha})$$

can be derived in the same way. Hence (A.4) follows. □

PROOF OF PROPOSITION 1. Note that for all $\vartheta \in \overline{\Theta}_m$ and all $m \in \mathscr{M}_1(\pi)$, there exists $\varepsilon_1 > 0$ such that $\|\theta - \vartheta\| \ge 3\varepsilon_1$. Let M be the "largest" model (i.e., the model which can be reduced to any of the other candidate models), so that $\overline{\Theta}_m \subseteq \overline{\Theta}_M$ for all $m \in \mathscr{M}$. Note here that $\overline{\Theta}_M = \Theta = \mathbb{R}^K$ and that we may write $\tilde{\theta}$ for $\tilde{\theta}_M$ since $\tilde{\theta}_M$ is a "global" maximum pseudolikelihood estimate over the set $\mathscr{M}$.

Let $\mathscr{D}(n) = \{x(n) \in \Omega_{\Lambda(n)}\colon \|\tilde{\theta} - \theta\| \le \varepsilon_1\}$. Then, applying Lemma 1 to $P_\theta(\mathscr{A}(n)^c)$ and the exponential consistency of $\tilde{\theta}$ [Cométs (1992)] to $P_\theta(\mathscr{D}(n)^c)$, we have $P_\theta(\{\mathscr{A}(n)\cap\mathscr{D}(n)\}^c) \le \exp(-n^{a_1})$ for some $a_1 > 0$ and for all large n. Hence we restrict our attention to $\mathscr{A}(n)\cap\mathscr{D}(n)$.

Let $\hat{m} \in \mathscr{M}_1(\pi)$, where $\hat{m}$ denotes the chosen model. Recall from the proof of Lemma 2 that $K(\vartheta, n) = \vartheta^\top V - g(\vartheta) = \vartheta^\top V_M - g_M(\vartheta)$. Recall also that $K(\vartheta, n)$ is globally concave and locally strictly concave so that $\tilde{\theta}$ is the unique maximum pseudolikelihood estimate. Since the true parameter θ is some positive distance away from $\overline{\Theta}_m$, there exists $\delta > 0$ such that $\sup_{\vartheta\in\overline{\Theta}_m} K(\vartheta, n) \le K(\tilde{\theta}, n) - \delta$ for all large n, P_θ-a.s. Then $Q_M - Q_m \ge a_2|\Lambda(n)|$ for some $a_2 > 0$ and for all $m \in \mathscr{M}_1(\pi)$ so that

$$P_\theta(\hat{m} \in \mathscr{M}_1(\pi)) \le P_\theta(Q_{\hat{m}} - Q_M > 0) \le \exp(-n^c)$$

as $n \to \infty$ for some $c > 0$. □

PROOF OF PROPOSITION 2. Let $\mathscr{F}(n) = \{x(n) \in \Omega_{\Lambda(n)}\colon\ |\Lambda(n)|\|\tilde{\theta}_m - \theta\|^2 \leq \varepsilon \log n \text{ for all } m \in \mathscr{M}\}$, so that by Lemma 4, $P_\theta(\mathscr{F}(n)^c) = O(n^{-\alpha})$ for some $\alpha > 0$. Hence we restrict our attention to $\mathscr{A}(n) \cap \mathscr{F}(n)$.

On $\mathscr{A}(n)$, for a chosen model $\hat{m} \in \mathscr{M}_2(\pi)$, we have

$$Q_\pi - Q_{\hat{m}} = |\Lambda(n)|\left[K_\pi\left(\tilde{\theta}_\pi, n\right) - K_{\hat{m}}\left(\tilde{\theta}_{\hat{m}}, n\right)\right] - (k_\pi - k_{\hat{m}})\log n.$$

Hence, we have

$$K_\pi\left(\tilde{\theta}_\pi, n\right) - K_{\hat{m}}\left(\tilde{\theta}_{\hat{m}}, n\right) \geq -C\left(\|\tilde{\theta}_\pi - \theta\|^2 + \|\tilde{\theta}_{\hat{m}} - \theta\|^2\right)$$

for some $C > 0$, so that

$$Q_\pi - Q_{\hat{m}} \geq |\Lambda(n)|\left[-C\left(\|\tilde{\theta}_{\hat{m}} - \theta\|^2 + \|\tilde{\theta}_\pi - \theta\|^2\right)\right] - (k_\pi - k_{\hat{m}})\log n.$$

Then on $\mathscr{A}(n) \cap \mathscr{F}(n)$,

$$Q_\pi - Q_{\hat{m}} \geq a_1 \log n,$$

where $a_1 = k_{\hat{m}} - k_\pi - 2C\varepsilon > 0$, provided ε is sufficiently small. Hence,

$$P_\theta(\hat{m} \in \mathscr{M}_2(\pi)) \leq P_\theta(Q_{\hat{m}} - Q_\pi > 0) = O(n^{-\alpha})$$

as $n \to \infty$ for some $\alpha > 0$. □

Acknowledgments. We are grateful to Stuart Geman and Basilis Gidas for stimulating discussions and to Richard Smith for helpful references and suggestions.

REFERENCES

AKAIKE, H. (1974). A new look at statistical model identification. *IEEE Trans. Automat. Control* **10** 716–723.

AKAIKE, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30** 9–14.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **6** 192–236.

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. IMS, Hayward, CA.

COMÉTS, F. (1992). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *Ann. Statist.* **20** 455–468.

CROSS, G. and JAIN, A. (1983). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5** 25–39.

GEMAN, D. (1991). *Random Fields and Inverse Problems in Imaging. Lecture Notes in Math.* **1427** 113–193. Springer, New York.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

GEMAN, S. and C. GRAFFIGNE (1986). Markov random field image models and their applications to computer vision. *Proceedings of the International Congress of Mathematicians* 1496–1517. Berkeley, CA.

GEORGII, H. O. (1988). *Gibbs Measures and Phase Transitions*. de Gruyter, Berlin.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B* **54** 657–699.

GIDAS, B. (1988). Consistency of maximum likelihood and maximum pseudo-likelihood estimators for Gibbs distributions. In *Stochastic Differential Systems*, *Stochastic Control Theory and Applications* (W. Fleming and P.-L. Lyons, eds.) 129–146. Springer, New York.

GIDAS, B. (1993). Parameter estimation for Gibbs distributions from fully observed data. In *Markov Random Fields*: *Theory and Applications* (R. Chellappa and A. Jain, eds.) 471–498. Academic Press, New York.

HASSNER, M. and SKLANSKY, J. (1980). The use of Markov random fields as models of texture. *Computer Graphics and Image Processing* **12** 357–370.

JI, C. (1990). Sieve estimators for pair-interaction potentials and local characteristics in Gibbs random fields. Technical Report 2037, Dept. Statistics, Univ. North Carolina.

KARR, A. (1991). Statistical models and methods in image analysis: a survey. In *Inference for Stochastic Processes* (I. V. Basawa and N. U. Prabhu, eds.). Dekker, New York.

KASHYAP, R. and CHELLAPPA, R. (1983). Estimation and choice of neighbors in spatial interaction models of images. *IEEE Trans. Inform. Theory* **29** 60–72.

KÜNSCH, H., GEMAN, S. and KEHAGIAS, A. (1995). Hidden Markov random fields. *Ann. Appl. Probab.* **5** 577–602.

RIPLEY, B. D. (1981). *Spatial Statistics*. Wiley, New York.

RISSANEN, J. (1984). Stochastic complexity. *J. Roy. Statist. Soc. Ser. B* **49** 223–239.

ROSENFELD, A. (1993). Image modeling during the 1980's: a brief overview. In *Markov Random Fields*: *Theory and Applications* (R. Chellappa and A. Jain, eds.). Academic Press, New York.

RUBIN, H. and SETHURAMAN, J. (1965). Probabilities of moderate deviations. *Sankhyā Ser. A* **27** 325–346.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SEYMOUR, L. (1993). Parameter estimation and model selection in image analysis using Gibbs–Markov random fields. Ph.D. dissertation, Dept. Statistics, Univ. North Carolina.

SEYMOUR, L. and JI, C. (1996). Approximate Bayes model selection criteria for Markov random fields. *J. Statist. Plann. Inference* **51** 75–97.

SMITH, K. and MILLER, M. (1990). A Bayesian approach incorporating Rissanen complexity for learning Markov random field texture models. *Proceedings of the 15th International Conference on Acoustics, Speech, and Signal Processing* **4** 2317–2320.

WOODROOFE, M. (1982). On model selection and the arcsine laws. *Ann. Statist.* **10** 1182–1194.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260
E-MAIL: cji@stat.unc.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602-1952
E-MAIL: seymour@rolf.stat.uga.edu