On Estimation of Basic Neighborhood of Markov Random Fields

Zsolt Talata*

Abstract

Markov random fields on the d-dimensional integer lattice with finite state space are considered, and the problem of estimation of the basic neighborhood from a single realization observed in a finite region is addressed. The Optimal Likelihood Ratio (OLR) estimator is introduced. Its nearly linear computation complexity is shown, and a bound on the probability of the estimation error is proved that implies strong consistency.

Key Words: Markov random field; likelihood ratio; Gibbs measure; model selection; information criterion

1. Introduction

This paper considers spatial processes on $\mathbb{Z}^d$ with values from a finite set A. Modeling interactions in such random fields by Gibbs fields was primarily motivated by statistical physics [9], but it proved to be an efficient approach in various other areas [17, 18]. Gibbs fields involve a potential function to determine the specification, the conditional distribution at sites in a region given the values at all other sites. Parameter estimation of Gibbs fields addresses the estimation of coefficients in the potential function from a sample, a *single* realization of the random field observed in a finite region, assuming that the interaction structure is known [15, 16].

If the range of the interactions is finite, then Gibbs fields are equivalent to Markov random fields. Markov random fields are characterized by the property that the conditional distribution at sites in a region given the values at all other sites is determined by a finite neighborhood of the region. The equivalence, first proved in [1], implies that the structure of the potential coincides with this Markov neighborhood. The interactions are of finite range in many applications that simplifies the estimation of the parametrized Gibbs distribution [7, 8, 13, 21].

It can be shown [14] that the specification, the conditional distribution at sites in regions, is determined by the one-point specification, the conditional distribution at single sites. The basic neighborhood is the region around a site that determines the conditional distribution at the site. This neighborhood may vary with the values at the surrounding sites, but if the random field is Markov, then for all configurations it is a subset of a finite region called the basic neighborhood of the Markov random field. Therefore, instead of using the potential, the random field can be parametrized by the one-point specification, that requires less information about the interaction structure and still includes finite number of parameters for Markov random fields.

In many applications, such as in image analysis [12] and in pattern recognition [10], there is usually no prior information on the interaction structure. Then the natural way to construct a Gibbs model is to find the one-point specification; often the nearest neighbor potential is used [11], that corresponds to the Markov random field model with basic

*Department of Mathematics, University of Kansas, 1460 Jayhawk Boulevard, Lawrence, KS 66045

neighborhood of radius one. Using a larger radius or basic neighborhood leads to a *model selection* problem. Focusing on the one-point specification is also motivated by the fact that the likelihood cannot be calculated explicitly so Besag's pseudo-likelihood [2] is popular, which is formulated in terms of the one-point specification. The maximum pseudo-likelihood (MPL) estimator of the one-point conditional probabilities is tractable and has a form analog to the maximum likelihood estimator of the transition probabilities of Markov chain models of time series [23, 3].

For order estimation of Markov chains, the minimum description length (MDL) principle [22] proved to be an efficient approach, in particular the information criteria provided by the normalized maximum likelihood (NML) code length [24] and the Krichevsky-Trofimov (KT) code length [20]. Both information criteria are asymptotically equivalent (for any Markov order) to the Bayesian information criterion (BIC), which is a penalized maximum likelihood. Therefore, it is reasonable to consider a penalized MPL to estimate the basic neighborhood of Markov random fields. Indeed, Smith and Miller formulated this conjecture [23] and Ji and Seymour proved the weak consistency of the estimator assuming the prior information of a finite set of possible basic neighborhoods [19]. Csiszár and Talata proved the strong consistency of the estimator without prior information [6]. The practical relevance of the penalized MPL estimator of the basic neighborhood is limited by its computational complexity, as it requires to calculate a score for each possible basic neighborhood. However, the basic neighborhood estimation with feasible computational complexity has been *an open problem* [6]. The method presented in this paper provides a solution to this problem.

In this paper a new method is introduced to estimate the basic neighborhood of Markov random fields. Some of the subsequent results were also presented at the IEEE International Symposium on Information Theory, Honolulu, Hawaii, June 2014. The complete proofs of all of the results given in this paper are contained in [25]. The *optimal likelihood ratio (OLR) estimator* is defined recursively along a sequence of extending lattice regions and uses the property that MPL increases at a larger rate in case of underestimation than in case of overestimation. The OLR estimator can be computed in $\mathcal{O}(|\Lambda_n| \log^{1/d} |\Lambda_n|)$ time, where $|\Lambda_n|$ is the sample size, the number of observed sites. The computational complexity could be strictly linear $O(|\Lambda_n|)$ if the rate at which the radius of the hypothetical basic neighborhoods may increase with the sample size was not calculated from the sample but it was a deterministic function as in [6]. An explicit bound is proved on the probability that the OLR estimator does not identify the basic neighborhood correctly, which implies the *strong consistency* of the estimator. Stationarity of the random field is not assumed and phase transition does not affect the results.

2. Markov Random Fields and the Main Results

The elements i of the d-dimensional lattice $\mathbb{Z}^d$ are called sites. A *random field* is a set of random variables indexed by the sites, $X = \{X(i) : i \in \mathbb{Z}^d\}$, where each random variable $X(i)$ takes values from a finite set A. $\|i\|$ denotes the maximum norm of the site $i \in \mathbb{Z}^d$, and $|\,.\,|$ denotes the cardinality of a set. The ball with radius r, that is the cube with sides of length $2r + 1$, is $\Delta_r = \{i \in \mathbb{Z}^d : \|i\| \leq r\}$.

The subset of the random variables with indices in a region $\Gamma \subseteq \mathbb{Z}^d$ of the lattice is denoted by $X(\Gamma)$. A block $g = \{a_i \in A : i \in \Gamma\}$ has the shape $\mathcal{S}(g) = \Gamma$ and the i'th coordinate $g(i) = a_i$; leaving out the j'th coordinate it is $g_{\setminus j} = \{a_i \in A : i \in \Gamma \setminus \{j\}\}$. Its

radius is $r(g) = r(\Gamma) = \sup_{i\in\Gamma} \|i\|$, and it is finite if $|g| = |\Gamma| < \infty$, and infinite otherwise. Denote $g|_\Phi$ the block g truncated to the region $\Phi \subseteq \mathbb{Z}^d$: $g|_\Phi = \{g(i) : i \in \Gamma \cap \Phi\}$. For $\Phi = \Delta_K$, $K > 0$, we write $g|_K = g|_{\Delta_K}$. Two blocks $g_1 \in A^{\Gamma_1}$ and $g_2 \in A^{\Gamma_2}$ are compatible, denoted by $g_1 \sim g_2$, if $g_1(i) = g_2(i)$ for all $i \in \Gamma_1 \cap \Gamma_2$. The set operations $g_1 \cup g_2$, $g_1 \cap g_2$ and $g_1 \backslash g_2$ are well-defined for compatible blocks. g_1 is smaller than g_2 if $g_1 \sim g_2$ and $g_1 \subset g_2$. The translate of a block g and a region Γ when the origin 0 is translated to the site i is denoted by g^i and Γ^i, respectively.

The distribution of the random field, the joint distribution of the random var.'s $X(i)$, is denoted by Q. Its Γ-marginal is

$$Q(g) = \Pr\{X(\Gamma) = g\}, \quad g \in A^\Gamma.$$

We assume that its finite-dimensional marginals are positive: $Q(g) > 0$ for all $|g| < \infty$. Under this standard assumption the following conditional probabilities are well-defined.

$$Q(g|f) = \Pr\{X(\Gamma) = g \,|X(\Phi) = f\}, \quad f \in A^\Phi, |f| < \infty.$$

The random field is a Markov random field if the uniform region Γ^i around the sites i in which the one-point conditional probabilities actually depend on the values can be reduced to Γ_0^i, the basic neighborhood, but not further.

Definition 1. *The random field is a* Markov random field *if there exists a unique finite region $\Gamma_0 \subset \mathbb{Z}^d$, called* basic neighborhood*, such that*

(i) *for any regions $\Gamma \supseteq \Gamma_0$ and $\Gamma' \subset \Gamma$ with $(\Gamma\backslash\Gamma') \cap \Gamma_0 = \emptyset$, for all blocks $g \in A^\Gamma$*

$$Q\left(b\,|g|_{\Gamma'}\right) = Q\left(b\,|g\right) \quad \text{for all } b \in A^{\{0\}}, \tag{1}$$

(ii) *for any regions $\Gamma \supseteq \Gamma_0$ and $\Gamma' \subset \Gamma$ with $(\Gamma\backslash\Gamma') \cap \Gamma_0 \neq \emptyset$, there exists a block $g \in A^\Gamma$ such that*

$$Q\left(b\,|g|_{\Gamma'}\right) \neq Q\left(b\,|g\right) \quad \text{for some } b \in A^{\{0\}}. \tag{2}$$

Moreover, the one-point conditional probabilities $Q(b|g)$, $b \in A^{\{0\}}$, $g \in A^{\Gamma_0}$, are translation invariant, $Q(b|g) = Q(b^i|g^i)$ for all $i \in \mathbb{Z}^d$.

Denote $q_{\min} = \min\{Q(b|g) : b \in A^{\{0\}}, g \in A^{\Gamma_0}\}$ and note that $q_{\min} > 0$. The following lemma shows that the one-point conditional probabilities determine the finite-block conditional probabilities. This generalization of the Markov property is proved in [6].

Lemma 1. *For any blocks $f \in A^\Phi$ and $\bar{f} \in A^{\mathbb{Z}^d\backslash\Phi}$*

$$\begin{aligned}&\Pr\left\{X(\Phi) = f \,\middle|X(j) = \bar{f}(j), j \in \mathbb{Z}^d\backslash\Phi\right\}\\ &= \Pr\left\{X(\Phi) = f \,\middle|X(j) = \bar{f}(j), j \in \bigcup_{i\in\Phi} \Gamma_0^i\backslash\Phi\right\},\end{aligned}$$

where the conditional probabilities are translation invariant and determined by the collection $Q(b|g)$, $b \in A^{\{0\}}$, $g \in A^{\Gamma_0}$.

The concept of Markov random field is equivalent to that of a Gibbs field [14]. Here, the interaction range is finite as the one-point specification is the collection

$$Q_{\Gamma_0} = \{Q(b|g) : b \in A^{\{0\}}, g \in A^{\Gamma_0}\}.$$

Given the one-point specification Q_{Γ_0}, the distribution of the random field is not necessarily unique, Q is one of them. These distributions are called Gibbs distributions and they are not necessarily translation invariant.

In this paper, we address statistical estimation of the basic neighborhood Γ_0 from a single realization $x \in A^{\mathbb{Z}^d}$ of the random field observed in a finite region Λ_n, $n \in \mathbb{N}$. Thus, the n'th random sample is $X_n = X(\Lambda_n)$ and the sample size is $|\Lambda_n|$.

Let $N_n^{k,D}(h)$, where $k \in \Delta_{2D}$, denote the number of the following non-overlapping occurrences of the cube $h \in A^{\Delta_D}$ in the sample X_n

$$\#\left\{i \in \mathbb{Z}^d : h^i \subset X_n, i = k + (4D+1)v \text{ for some } v \in \mathbb{Z}^d\right\}.$$

Above those cubes are considered whose centers are in $(4D+1)$ distance, that is, whose centers are in the k'th sieve $\Lambda_n^{k,D}$. The larger the distance D, the smaller the counts $N_n^{k,D}(h)$. The largest D will be used that provides sufficient numbers of occurrences.

$$\max\left\{D : N_n^{k,D}(h) \geq \log^3 |\Lambda_n| \text{ for all } h \in A^{\Delta_D}, k \in \Delta_{2D}\right\}.$$

Then the number of occurrences of a block g with $r(g) \leq D(X_n)$ in the sample X_n is defined as

$$N_n(g) = \sum_{k \in \Delta_{2D(X_n)}} N_n^k(g)$$

where $N_n^k(g)$ is the number of occurrences of the block g in the sample X_n with center in the k'th sieve $\Lambda_n^k = \Lambda_n^{k,D(X_n)}$:

$$N_n^k(g) = \sum_{f:\, \mathcal{S}(g \cup f) = \Delta_{D(X_n)}} N_n^{k,D(X_n)}(g \cup f).$$

Clearly, for any $\Gamma \subset \mathbb{Z}^d$ with $r(\Gamma) \leq D(X_n)$ and $g \in A^\Gamma$

$$\sum_{b \in A^{\{0\}}} N_n(g \cup b) = N_n(g) \quad \text{and} \quad \sum_{g \in A^\Gamma} N_n(g) \leq |\Lambda_n|.$$

The shape Λ_n of the sample region is arbitrary but it must satisfy that it is increasing in n and that the total number of centers considered for the number of occurrences is asymptotically equivalent to the number of sites in sample region:

$$|\Lambda_n| \Big/ \sum_{k \in \Delta_{2D(X_n)}} |\Lambda_n^k| \to 1 \quad \text{as } n \to \infty.$$

The following observation follows from the bound derived in the proof of Lemma 3.1 in [6].

Lemma 2. *For some absolute constant $\nu > 0$*

$$\Pr\left\{D(X_n) \geq C \log^{1/d} |\Lambda_n|\right\} \geq 1 - e^{-|\Lambda_n|^\nu}, \quad n \geq n_0,$$

where $C > 0$ depends only on q_{min}.

Given the sample X_n, the *pseudo-likelihood* function of a region $\Gamma \subset \mathbb{Z}^d$ with $r(\Gamma) \leq D(X_n)$, regarding Γ as the basic neighborhood of the random field with a hypothetical distribution Q', is

$$\mathrm{PL}_{X_n}(\Gamma, Q') = \prod_{i:\, \Delta^i_{D(X_n)} \subseteq \Lambda_n} Q'(X_n(i) |\, X_n(j), j \in \Gamma^i) = \prod_{g \in A^\Gamma,\, b \in A^{\{0\}}} Q'(b|g)^{N_n(g \cup b)}.$$

Although a probability measure Q' does not necessarily exist for any $Q'_\Gamma = \{Q'(b|g) : b \in A^{\{0\}}, g \in A^\Gamma\}$ satisfying $\sum_{b \in A^{\{0\}}} Q'(b|g) = 1$, the maximum of the pseudo-likelihood over all such Q'_Γ is called *maximum pseudo-likelihood*, and it has the explicit form

$$\mathrm{MPL}_{X_n}(\Gamma) = \prod_{g \in A^\Gamma} \mathrm{MPL}'_{X_n}(g)\,,$$

where

$$\mathrm{MPL}'_{X_n}(g) = \prod_{b \in A^{\{0\}}} \left(\frac{N_n(g \cup b)}{N_n(g)} \right)^{N_n(g \cup b)}. \tag{3}$$

In this paper we propose the following estimator of the basic neighborhood Γ_0.

Definition 2. *Given the sample X_n, the* optimal likelihood ratio *(OLR) estimator $\widehat{\Gamma}_{OLR}(X_n)$ of the basic neighborhood Γ_0 is defined recursively with respect to the region $\Theta \subset \mathbb{Z}^d$ as follows.*

(i) $\Gamma^{\{0\}} = \Delta_{D(X_n)/2} \backslash \{0\}$

(ii) Given Γ^Θ with $\Delta_{R-1} \subseteq \Theta \subset \Delta_R$ for some $R \geq 1$, fix $i \in \Delta_R \backslash \Theta$. If

$$\prod_{a \in A^{\{i\}}} \mathrm{MPL}'_{X_n}\big(g_{\backslash i} \cup a\big) \Big/ \mathrm{MPL}'_{X_n}\big(g_{\backslash i}\big) \leq e^{N_n(g_{\backslash i})^{3/4}} \tag{4}$$

for all $g_{\backslash i} \in A^{\Gamma^\Theta \cup (\Delta_{D(X_n)} \backslash \Theta \backslash \{i\})}$, then let $\Gamma^{\Theta \cup \{i\}} = \Gamma^\Theta \backslash \{i\}$;
and otherwise let $\Gamma^{\Theta \cup \{i\}} = \Gamma^\Theta$.

(iii) $\widehat{\Gamma}_{OLR}(X_n) = \Gamma^{\Delta_{D(X_n)/2}}$

The proposed estimator has the following bounds on error probability and computational complexity.

Theorem 1. *Given a Markov random field, the estimation error of the OLR estimator of the basic neighborhood is controlled as*

$$\Pr\left\{ \widehat{\Gamma}_{OLR}(X_n) \neq \Gamma_0 \right\} \leq 4 |\Lambda_n|^{-(1/6)\log^{3/2}|\Lambda_n|}, \quad n \geq n_0.$$

Proof. Theorem 1 is proved in [25]. □

Corollary 1. *Given a Markov random field, the OLR estimator of the basic neighborhood satisfies*

$$\widehat{\Gamma}_{OLR}(X_n) = \Gamma_0$$

if n is sufficiently large, with probability 1.

Proof. Corollary 1 follows from an application of Borel-Cantelli's lemma to Theorem 1. □

Corollary 2. *Given a Markov random field, the OLR estimator of the one-point specification,*

$$\widehat{Q}_{\widehat{\Gamma}_{OLR}(X_n)} = \left\{ \frac{N_n(g \cup b)}{N_n(g)} : b \in A^{\{0\}}, g \in A^{\widehat{\Gamma}_{OLR}(X_n)} \right\}$$

converges to the one-point specification Q_{Γ_0} as $n \to \infty$, with probability 1.

Proof. It follows from Corollary 1 and Theorem 3 below. □

Theorem 2. *The number of computations needed to determine the OLR estimator of the basic neighborhood for a given sample X_n is $\mathcal{O}(|\Lambda_n| \log^{1/d} |\Lambda_n|)$, and this can be achieved storing $O(|\Lambda_n|)$ data.*

Proof. Theorem 2 is proved in [25]. □

Remark 1. As the proof of Theorem 2 shows, finding $D(X_n)$ requires $\mathcal{O}(|\Lambda_n| \log^{1/d} |\Lambda_n|)$ computations and after that calculating the OLR estimator requires only $O(|\Lambda_n|)$ computations. The former computations could be avoided by setting $D(X_n) = D(|\Lambda_n|) = o(\log^{1/d} |\Lambda_n|)$, see Lemma 2. Then the computational complexity of the OLR estimator would be strictly linear $O(|\Lambda_n|)$, however, such $D(X_n)$ would not be practical.

The estimation error proof relies upon the following concentration inequality for Markov random fields, whose proof is included in [25].

Theorem 3. *Given a Markov random field and $0 < \eta < 1/2$, with probability at least $1 - 2|\Lambda_n|^{-(1/6)\log^{3-6\eta}|\Lambda_n|}$, simultaneously for all $i \in \Delta_{D(X_n)/2}$, all $b \in A^{\{i\}}$, all blocks g with $\Delta_{D(X_n)} \backslash \{i\} \supseteq \mathcal{S}(g) \supseteq (\Delta_{D(X_n)} \backslash \Delta_{D(X_n)/2})$,*

$$\left| \frac{N_n(g \cup b)}{N_n(g)} - Q(b|g) \right| < \left(\frac{2^d}{\log |A|} \right)^{\eta} \frac{\log^{\eta} |\Lambda_n|}{N_n(g)^{\eta}},$$

if n is sufficiently large.

Remark 2. The proof uses Lemma 2 and implies that $D(X_n) \geq C \log^{1/d} |\Lambda_n|$, $C > 0$, also holds over the event where the claimed inequality holds.

Acknowledgment

Research was supported in part by the NSF under Grants DMS 1407819 and DMS 0906929, and by the ARO under Grant MSD 65386-MA-II.

References

[1] AVERINTSEV, F. (1970). On a method of describing discrete parameter random fields. *Problemy Peredachi Informatsii* **6** 100–109.

[2] BESAG, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24** 179–195.

[3] CSISZÁR, I. and SHIELDS, P.C. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.* **28** 1601–1619.

[4] CSISZÁR, I. (2002). Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory* **48** 1616–1628.

[5] CSISZÁR, I. and TALATA, ZS. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory* **52** 1007–1016.

[6] CSISZÁR, I. and TALATA, ZS. (2006). Consistent estimation of the basic neighborhood of Markov random fields. *Ann. Statist.* **34** 123–145.

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

[8] CRESSIE, N.A.C. (1993). *Statistics for spatial data.* Wiley, New York.

[9] DOBRUSHIN, R.L. (1968). The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory Probab. Appl.* **13** 197–224.

[10] ELLIOTT, H. and DERIN, H. (1987). Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Trans. Pattern. Anal.* **9** 39–55.

[11] GEMAN, D. (1990). *Random fields and inverse problems in imaging.* Lec. Notes Math. 1427, Springer, Berlin.

[12] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern. Anal.* **6** 721–741.

[13] GEMAN, S. and GRAFFIGNE, C. (1987). Markov random fields image models and their applications to computer vision. In *Proceedings of the International Congress Mathematicians* (A. M. Gleason, ed.) **2** 1496–1517. Amer. Math. Soc., Providence, R.I.

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

[15] GIDAS, B. (1988). Consistency of maximum likelihood and pseudolikelihood estimators for Gibbs distributions. *Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol. Math. Appl.* **10** 129–145.

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

[17] GRENANDER, U. and MILLER, M.I. (2007). *Pattern theory: from representation to inference.* Oxford University Press, Oxford.

[18] GUYON, X. (1995). *Random fields on a network.* Springer, New York.

[19] JI, C. and SEYMOUR, L. (1996). A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *Ann. Appl. Probab.* **6** 423–443.

[20] KRICHEVSKY, R.E. and TROFIMOV, V.K. (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory* **27** 199–207.

[21] PICKARD, D.K. (1987). Inference for discrete Markov field: The simplest non-trivial case. *J. Amer. Statist. Assoc.* **82** 90–96.

[22] RISSANEN, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14** 1080–1100.

[23] SMITH, K.R. and MILLER, M. (1990). A Bayesian approach incorporating Rissanen complexity for learning Markov random field texture models. *International Conference on Acoustics, Speech, and Signal Processing,* **4** 2317–2320.

[24] STARKOV, J.M. (1977). Coding of discrete sources with unknown statistics. In *Topics in Information Theory (Second Colloq., Keszthely, 1975). Colloq. Math. Soc. János Bolyai* **16** 559–574. North-Holland, Amsterdam.

[25] TALATA, ZS. (2014). Markov Neighborhood Estimation with Linear Complexity for Random Fields. *Manuscript,* 17 pages.