Weighted Maximum Posterior Marginals for Random Fields Using an Ensemble of Conditional Densities From Multiple Markov Chain Monte Carlo Simulations

James Peter Monaco*, *Member, IEEE*, and Anant Madabhushi, *Senior Member, IEEE*

***Abstract*—The ability of classification systems to adjust their performance (sensitivity/specificity) is essential for tasks in which certain errors are more significant than others. For example, mislabeling cancerous lesions as benign is typically more detrimental than mislabeling benign lesions as cancerous. Unfortunately, methods for modifying the performance of Markov random field (MRF) based classifiers are noticeably absent from the literature, and thus most such systems restrict their performance to a single, static operating point (a paired sensitivity/specificity). To address this deficiency we present weighted maximum posterior marginals (WMPM) estimation, an extension of maximum posterior marginals (MPM) estimation. Whereas the MPM cost function penalizes each error equally, the WMPM cost function allows misclassifications associated with certain classes to be weighted more heavily than others. This creates a preference for specific classes, and consequently a means for adjusting classifier performance. Realizing WMPM estimation (like MPM estimation) requires estimates of the posterior marginal distributions. The most prevalent means for estimating these—proposed by Marroquin *et al.*—utilizes a Markov chain Monte Carlo (MCMC) method. Though Marroquin's method (M-MCMC) yields estimates that are sufficiently accurate for MPM estimation, they are inadequate for WMPM. To more accurately estimate the posterior marginals we present an equally simple, but more effective extension of the MCMC method (E-MCMC). Assuming an identical number of iterations, E-MCMC as compared to M-MCMC yields estimates with higher fidelity, thereby 1) allowing a far greater number and diversity of operating points and 2) improving overall classifier performance. To illustrate the utility of WMPM and compare the efficacies of M-MCMC and E-MCMC, we integrate them into our MRF-based classification system for detecting cancerous glands in (whole-mount or quarter) histological sections of the prostate.**

***Index Terms*—Histology, Markov Chain Monte Carlo, Markov random fields, maximum posterior marginals, prostate cancer, Rao-Blackwellized estimator.**

I. Introduction

MANY estimation tasks require classification systems capable of modeling the inherent dependencies among objects (sites). In the context of medical imaging, these objects, for example, could be calcifications in a mammogram or the pixels of a magnetic resonance (MR) image [1]. Within a Bayesian framework each site is a random variable, and the collection of these random variables (under minor assumptions) is a Markov random field (MRF) [2]. Because of their ability to model statistical dependencies among variables, MRFs have proven invaluable in a variety of computer vision and image processing tasks such as segmentation [3]–[8], denoising [9], [10], and texture synthesis [11], [12]. In [13] and [14] we used probabilistic pairwise Markov models (PPMMs), a novel type of Markov model, to detect cancer on MR images and digitized histological sections of the prostate. PPMMs formulate Markov priors in terms of probability densities, instead of the typical potential functions, facilitating the creation of more sophisticated priors.

In addition to modeling inter-variable dependencies, classifiers often require the ability to adjust their performance (i.e., sensitivity/specificity) with respect to specific classes. This ability is essential for tasks in which certain types of errors are more significant than others. Such tasks are particularly pervasive in medical imaging. For example, in the context of mammography, mislabeling cancerous lesions as benign is typically more detrimental than mislabeling benign lesions as cancerous; consequently, commercial computer-aided detection systems for identifying mammographic abnormalities are typically adjusted to the highest detection sensitivity that incurs no more than one false positive per image [15], [16].

For univariate Bayesian systems—or multivariate systems with statistically independent variables—the methodology for modifying classifier performance is well-established [17]: appropriately weight (scale) the *a posteriori* probability associated with each class, and then select the class with the greatest weighted probability. In the two-class case this reduces to the familiar thresholding. Unfortunately, analogous methods compatible with random fields are noticeably absent from the literature. Consequently, most MRF-based classification systems restrict their performance to a single, static operating point (i.e., a paired sensitivity/specificity).

To address this deficiency we present weighted maximum posterior marginals (WMPM) estimation, an extension of

Manuscript received December 07, 2010; revised February 03, 2011; accepted February 08, 2011. Date of publication February 17, 2011; date of current version June 29, 2011. This work was supported in part by the Wallace H. Coulter Foundation, in part by the National Cancer Institute under Grant R01CA136535-01, Grant R01CA14077201, and Grant R03CA143991-01), and in part by the Cancer Institute of New Jersey. Both authors are major shareholders of IbRiS Inc. *Asterisk indicates corresponding author.*

*J. P. Monaco is with the Department of Biomedical Engineering, Rutgers University, Piscataway, NJ 08854 USA (e-mail: jpmonaco@rci.rutgers.edu).

A. Madabhushi is with the Department of Biomedical Engineering, Rutgers University, Piscataway, NJ 08854 USA (e-mail: anantm@rci.rutgers.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2011.2114896

maximum posterior marginals (MPM) estimation [18] that provides a means for adjusting classifier performance. Marroquin *et al.* [18] introduced MPM as an alternative to maximum *a posteriori* (MAP) estimation [2] because of its superior performance in the presence of high noise [18], [19]. Like all Bayesian schemes, the MPM estimation criterion is derived by minimizing the expected value of a specified cost function. The MPM cost function counts the total number of misclassifications, penalizing each error equally. In this work we generalize the MPM cost function, allowing misclassifications of certain classes to be weighted more heavily than others. This creates a natural preference for specific classes, and consequently a means for adjusting classifier performance.

Performing WMPM estimation (like MPM estimation) requires estimates of the so-called posterior marginal distributions. The most prevalent means for estimating them—proposed by Marroquin *et al.* [18] and recently employed in [19]–[21]—utilizes a Markov chain Monte Carlo (MCMC) method. Marroquin's method, which we will henceforth refer to as M-MCMC, employs the Metropolis–Hastings algorithm [22], [23] (or a specific instance such as the Gibbs Sampler [2]) to construct a Markov chain of the MRF that converges to a prescribed probability distribution. Using samples from this chain, the algorithm performs a Monte Carlo estimation of each site's posterior marginal by recording the fraction of iterations in which the site under consideration assumes a specified class label. That is, M-MCMC performs density estimation via histogramming.

Though M-MCMC yields estimates that are sufficiently accurate for MPM estimation, they are inadequate for WMPM estimation. By inadequate we mean the following: changes in the class-specific weights (inherent in the WMPM cost function) often do not produce the expected changes in classifier performance. From a practical perspective, this manifests as a severe reduction in the number of attainable operating points. For example, if the goal were to set classifier sensitivity with respect to a specified class to 90%, we might find that the corresponding operating point—or any operating point close to it—would not exist. The reasons for this will be discussed later in the paper.

To more accurately estimate the posterior marginals we suggest an equally simple, but far more effective extension of the MCMC method (E-MCMC) that performs a Monte Carlo averaging of conditional densities drawn from multiple, statistically independent Markov chains. Averaging over the functional forms of the conditional distributions produces more accurate density estimates than averaging over the actual samples themselves [24], [25]. Using multiple Markov chains increases the robustness of the estimates to the presence of multiple modes in the MRF distribution [26]. Incorporating these strategies enhances the fidelity of the estimates, yielding a far greater number of operating points and increasing overall classifier accuracy.

In summary, the contributions of this paper are as follows.

- We generalize the MPM cost function, incorporating class-specific weights, and thus providing a means for varying MRF-based classifier performance. That is, whereas the MPM cost function weights each misclassification equally, the WMPM cost function assigns class-specific penalties.
- To obtain estimates of the posterior marginals that are sufficiently accurate for WMPM estimation we present E-MCMC, an extension of the MCMC algorithm introduced in [18] that performs Monte Carlo averaging of conditional densities drawn from multiple, statistically independent Markov chains.

To illustrate the benefits of WMPM, we integrate it into our MRF-based classification system[1] for detecting cancerous glands on digitized (whole-mount or quarter) histological sections from radical prostatectomies [14]. Over a cohort of 27 images from 10 patient studies, we demonstrate how WMPM can be used to vary classifier performance, enabling the construction of receiver operator characteristic (ROC) curves. Additionally, we compare the abilities of E-MCMC and M-MCMC to estimate the posterior marginal distributions by contrasting the resulting ROC curves produced by WMPM using both techniques. Assuming an equal number of iterations for both methods, the most pertinent results are as follows: 1) E-MCMC yields ROC curves with several orders of magnitude more operating points than those produced with M-MCMC, 2) E-MCMC allows the choice of virtually any true (or false) positive rate (within [0,1]), while M-MCMC restricts these rates to small subintervals of [0,1], and 3) E-MCMC outperforms M-MCMC in terms of classification accuracy.

The remainder of the paper is organized as follows. In Section II we review the Bayesian estimation of MRFs, derive the MPM estimation criteria, and describe the M-MCMC method for estimating the posterior marginals. Section III introduces WMPM estimation and E-MCMC. In Section IV we demonstrate the utility of WMPM and E-MCMC by integrating them into our system for detecting prostate cancer on digitized histological sections. In Section V we discuss our findings and present our concluding remarks.

II. Review of Markov Random Fields and Maximum Posterior Marginal Estimation

A. Markov Random Field Definitions and Notation

Let the set $S = \{1, 2, \ldots, N\}$ reference N sites to be classified. Each site $s \in S$ has two associated random variables: $X_s \in \Lambda \equiv \{\omega_1, \omega_2, \ldots, \omega_L\}$ indicating its state (class) and $Y_s \in \mathbb{R}^D$ representing its D-dimensional feature vector. Particular instances of X_s and Y_s are denoted by the lowercase variables $x_s \in \Lambda$ and $y_s \in \mathbb{R}^D$. Let $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ refer to all random variables X_s and Y_s in aggregate. The state spaces of $\mathbf{X}$ and $\mathbf{Y}$ are the Cartesian products $\Omega = \Lambda^N$ and $\mathbb{R}^{D \times N}$. Instances of $\mathbf{X}$ and $\mathbf{Y}$ are denoted by the lowercase variables $\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \Omega$ and $\mathbf{y} = (y_1, y_2, \ldots, y_N) \in \mathbb{R}^{D \times N}$. See Table I for a list and description of the commonly used notations and symbols in this paper.

Let $G = \{S, E\}$ establish an undirected graph structure on the sites, where S and E are the vertices (sites) and edges, respectively. A neighborhood η_s is the set containing all sites that

[1] Note that in [14] we did not use WMPM. We instead employed maximum *a posteriori* estimation [2], which we implemented using iterated conditional modes [9].

TABLE I
LIST OF NOTATION AND SYMBOLS

Symbol	Description	Symbol	Description
S	Set referencing N sites	$\mathbf{X}\in\Omega$	Collection of all X_s: $\mathbf{X}=(X_1,X_2,\dots,X_N)$
Λ	Range of X_s and x_s: $\Lambda\equiv\{\omega_1,\omega_2,\dots,\omega_L\}$	$\mathbf{x}\in\Omega$	Instance of $\mathbf{X}$: $\mathbf{x}=(x_1,x_2,\dots,x_N)$
$X_s\in\Lambda$	Random variable indicating state at site s	Ω	Range of $\mathbf{X}$ and $\mathbf{x}$: $\Omega=\Lambda^N$
$x_s\in\Lambda$	Instance of X_s	$\mathbf{Y}\in\mathbb{R}^{D\times N}$	Collection of all Y_s: $\mathbf{Y}=(Y_1,Y_2,\dots,Y_N)$
D	Number of features	$\mathbf{y}\in\mathbb{R}^{D\times N}$	Instance of $\mathbf{Y}$: $\mathbf{y}=(y_1,y_2,\dots,y_N)$
$Y_s\in\mathbb{R}^D$	Random variable indicating feature vector at site s	η_s	Set of sites that neighbor $s\in S$
$y_s\in\mathbb{R}^D$	Instance of Y_s	$\mathbf{x}_{\text{-}s}$	$\mathbf{x}_{\text{-}s}=(x_1,\dots,x_{s-1},x_{s+1},\dots,x_N)$
$\widehat{\mathbf{x}}\in\Omega$	Estimate of $\mathbf{X}$	$\mathbf{x}_{\eta_s}$	$\mathbf{x}_{\eta_s}=\left(x_{\eta_s(1)},\dots,x_{\eta_s(\lvert\eta_s\rvert)}\right)$
$C(\mathbf{x},\widehat{\mathbf{x}})$	Cost of choosing $\widehat{\mathbf{x}}$ when the true labels are $\mathbf{x}$	$R(\mathbf{X}\lvert\widehat{\mathbf{x}},\mathbf{y})$	Bayesian risk (expected cost): $\mathrm{E}[C(\mathbf{X},\widehat{\mathbf{x}})\lvert\mathbf{y}]$
b	Burn-in period of Markov chain	$\alpha(\cdot)$	Weighting function $\alpha:\Lambda\rightarrow\mathbb{R}^+$
c	Number of independent Markov chains	m	Number of MCMC iterations after equilibrium

share an edge with s, i.e., $\eta_s=\{r : r\in S, r\neq s, \{r,s\}\in E\}$. If P is a probability measure defined over Ω then the triplet (G,Ω,P) is called a random field. The random field $\mathbf{X}$ is a Markov random field if its local conditional probability density functions satisfy the Markov property: $P(X_s=x_s|\mathbf{X}_{-s}=\mathbf{x}_{-s})=P(X_s=x_s|\mathbf{X}_{\eta_s}=\mathbf{x}_{\eta_s})$, where $\mathbf{x}_{-s}=(x_1,\dots,x_{s-1},x_{s+1},\dots,x_N)$, $\mathbf{x}_{\eta_s}=(x_{\eta_s(1)},\dots,x_{\eta_s(|\eta_s|)})$, and $\eta_s(i)\in S$ is the ith element of the set η_s. Note that in places where it does not create ambiguity we will simplify the probabilistic notations by omitting the random variables, e.g., $P(\mathbf{x})\equiv P(\mathbf{X}=\mathbf{x})$.

```
Gibbs Sampler
Input: Initial labeling $\mathbf{x}^0$
Output: Final labeling $\mathbf{x}^k$ after iteration $k$
$k = 0$
do
    $k = k + 1$
    $\mathbf{x}^k = \mathbf{x}^{k-1}$
    for $\forall\ s \in S$ do
        sample $x_s^k$ from $P(X_s = x_s^k|\mathbf{x}_{\text{-}s}^k, \mathbf{y})$
    end for
while $k < \infty$
```

Fig. 1. Algorithm for the Gibbs sampler.

B. Maximum Posterior Marginals (MPM) Cost Function

Given an observation of the feature vectors $\mathbf{Y}$, we would like to estimate the states $\mathbf{X}$. Bayesian estimation advocates selecting the estimate $\widehat{\mathbf{x}}\in\Omega$ that minimizes the conditional risk (expected cost) [17]

$$R(\mathbf{X}|\widehat{\mathbf{x}},\mathbf{y})=\mathrm{E}\left[C(\mathbf{X},\widehat{\mathbf{x}})|\mathbf{y}\right]=\sum_{\mathbf{x}\in\Omega}C(\mathbf{x},\widehat{\mathbf{x}})P(\mathbf{x}|\mathbf{y}) \tag{1}$$

where E indicates expected value and $C(\mathbf{x},\widehat{\mathbf{x}})$ is the cost of selecting labels $\widehat{\mathbf{x}}$ when the true labels are $\mathbf{x}$.

In [18] Marroquin *et al.* suggested the following cost function:

$$C_{\mathrm{MPM}}(\mathbf{x},\widehat{\mathbf{x}})=\sum_{s\in S}\left[1-\delta(x_s-\widehat{x}_s)\right]. \tag{2}$$

This function counts the number of sites in $\widehat{\mathbf{x}}$ that are labeled incorrectly, with each misclassification accruing an identical cost. Inserting (2) into (1) yields

$$\begin{aligned}R_{\mathrm{MPM}}(\mathbf{X}|\widehat{\mathbf{x}},\mathbf{y})&=\sum_{\mathbf{x}\in\Omega}\left(\sum_{s\in S}\left[1-\delta(x_s-\widehat{x}_s)\right]\right)P(\mathbf{x}|\mathbf{y})\\&=\sum_{s\in S}\sum_{\mathbf{x}\in\Omega}P(\mathbf{x}|\mathbf{y})\\&\quad-\sum_{s\in S}\sum_{\mathbf{x}\in\Omega}\delta(x_s-\widehat{x}_s)P(\mathbf{x}|\mathbf{y})\\&=|S|-\sum_{s\in S}P(\widehat{x}_s|\mathbf{y})\end{aligned} \tag{3}$$

where $|S|$ is the cardinality of the set S. The distributions $P(\widehat{x}_s|\mathbf{y})$ are called the posterior marginals. Minimizing (3) over $\widehat{\mathbf{x}}$ is equivalent to independently maximizing each of these posterior marginals with respect to its corresponding $\widehat{x}_s$. Hence, this estimation criterion is termed maximum posterior marginals (MPM). An exhaustive search for each optimal $\widehat{x}_s\in\Lambda$ is nearly always appropriate since $|\Lambda|$ is usually small.

C. Estimating the Posterior Marginals

As can be seen from (3), MPM requires estimates of the posterior marginals $P(x_s|\mathbf{y})$. Unfortunately, the range Ω of $\mathbf{X}$ is far too large to calculate them by the direct marginalization of $P(\mathbf{x}|\mathbf{y})$. However, it is possible to construct a Markov chain that yields random samples of $P(\mathbf{x}|\mathbf{y})$—and consequently $P(x_s|\mathbf{y})$. Using these random samples, a Monte Carlo procedure can then estimate the posterior marginals. Such a Markov chain Monte Carlo (MCMC) approach was first proposed by Marroquin *et al.* in [18]. Marroquin's MCMC method, which we will abbreviate as M-MCMC, is now explained in greater detail.

Given each site's conditional probability density function $P(x_s|\mathbf{x}_{-s},\mathbf{y})$, the Gibbs sampler [2] (see Fig. 1) generates a Markov chain $(\mathbf{X}^0,\mathbf{X}^1,\mathbf{X}^2,\dots)$ with equilibrium distribution $P(\mathbf{x}|\mathbf{y})$, where $\mathbf{X}^k$ is a random variable indicating the state of the chain at iteration k. (See [27] for an excellent discussion of the Gibbs sampler.) The proportion of time the chain spends—after reaching equilibrium—in any state $\mathbf{x}$ is given by $P(\mathbf{x}|\mathbf{y})$, i.e., each $\mathbf{X}^k$ represents a sample from the distribution $P(\mathbf{x}|\mathbf{y})$. The convergence to $P(\mathbf{x}|\mathbf{y})$ is independent of the starting conditions [28]; and consequently, $\mathbf{x}^0$ can be selected at random from Ω. Determining the number of iterations b needed for the Markov chain to reach equilibrium—called the burn-in period—is difficult, and depends upon the particular distribution $P(\mathbf{x}|\mathbf{y})$ and the initial conditions $\mathbf{x}^0$. Usually the burn-in period b is selected empirically.

Note that in practice the computation of $P(x_s|\mathbf{x}_{-s},\mathbf{y})$ in Step 6 (Fig. 1) is straight-forward. Consider that $P(x_s|\mathbf{x}_{-s},\mathbf{y})$ reduces to $P(x_s|\mathbf{x}_{\eta_s},y_s)$ by consequence of the Markov property and the typical assumption that the observations $\mathbf{Y}$ are conditionally independent given their associated states, i.e., $P(\mathbf{y}|\mathbf{x}) = \prod_{s\in S} P(y_s|x_s)$. Furthermore, $P(x_s|\mathbf{x}_{\eta_s},y_s) \propto P(y_s|x_s)P(x_s|\mathbf{x}_{\eta_s})$ by Bayes law.

As previously stated, if a Markov chain has an equilibrium distribution $P(\mathbf{x}|\mathbf{y})$, then the proportion of time the chain spends in any state $\mathbf{x}$ is given by $P(\mathbf{x}|\mathbf{y})$. This implies that the fraction of time the specific site s spends in state x_s is given by $P(x_s|\mathbf{y})$. Therefore, one method for estimating $P(x_s|\mathbf{y})$ is as follows:

$$\widehat{P}(X_s=\omega|\mathbf{y}) \approx \frac{1}{m}\sum_{k=b+1}^{b+m} \delta\left(x_s^k-\omega\right) \tag{4}$$

where δ is the discrete (Kronecker) delta function, $\omega \in \Lambda$, and m is the number of iterations past equilibrium needed to accurately estimate $P(\mathbf{x}|\mathbf{y})$. Thus, (4) simply records the number of times x_s^k assumes each of the $|\Lambda|$ possible classes, and then normalizes by the total number of recorded samples. That is, (4) performs density estimation by histogramming. The value for m, like b, is typically chosen empirically.[2]

III. Weighted Maximum Posterior Marginals

The MPM cost function weights each misclassification equally. We now generalize this cost function, allowing misclassifications of certain classes to be penalized more heavily than others. This creates a natural preference for specific classes and, as previously mentioned, a means for adjusting classifier performance.

A. Weighted Maximum Posterior Marginals (WMPM) Cost Function

With MPM estimation each incorrect label $\widehat{x}_s$ accrues an identical cost of one, regardless of the true state x_s. To penalize incorrect estimates differently for different classes we propose the following cost function:

$$C_{\text{WMPM}}(\mathbf{x},\widehat{\mathbf{x}}) = \sum_{s\in S}\alpha(x_s)\left[1-\delta(x_s-\widehat{x}_s)\right] \tag{5}$$

where the positive weighting function $\alpha : \Lambda \rightarrow \mathbb{R}^+$ indicates the cost of mislabeling site s when its true label is x_s. The expected cost (risk) is found by inserting (5) into (1), yielding

$$\begin{aligned} &R_{\text{WMPM}}(\mathbf{X}|\widehat{\mathbf{x}},\mathbf{y}) && (6)\\ &\quad=\sum_{\mathbf{x}\in\Omega}\left(\sum_{s\in S}\alpha(x_s)\left[1-\delta(x_s-\widehat{x}_s)\right]\right)P(\mathbf{x}|\mathbf{y})\\ &\quad=\sum_{s\in S}\sum_{\mathbf{x}\in\Omega}\alpha(x_s)P(\mathbf{x}|\mathbf{y})-\sum_{s\in S}\alpha(\widehat{x}_s)P(\widehat{x}_s|\mathbf{y}). && (7)\end{aligned}$$

Since the first term is not a function of $\widehat{\mathbf{x}}$, minimizing (6) over $\widehat{\mathbf{x}}$ is equivalent to independently maximizing each of the weighted posterior marginals $\alpha(\widehat{x}_s)P(\widehat{x}_s|\mathbf{y})$ with respect to its corresponding $\widehat{x}_s$. Thus, weighted maximum posterior marginal (WMPM) estimation advocates choosing the $\widehat{\mathbf{x}} = (\widehat{x}_1,\widehat{x}_2,\ldots,\widehat{x}_N)$ that individually maximizes $\alpha(\widehat{x}_s)P(\widehat{x}_s|\mathbf{y})$ for all $s \in S$. As with MPM, an exhaustive search for each $\widehat{x}_s$ is appropriate since $|\Lambda|$ is typically small.

B. Estimating the Posterior Marginals Using an Extended MCMC (E-MCMC) Method

WMPM, like MPM, requires estimates of the posterior marginals $P(x_s|\mathbf{y})$. Unfortunately, certain limitations (discussed below) inherent in M-MCMC render it inappropriate for WMPM estimation. Accordingly, we present an extension of M-MCMC (E-MCMC) that mitigates these limitations.

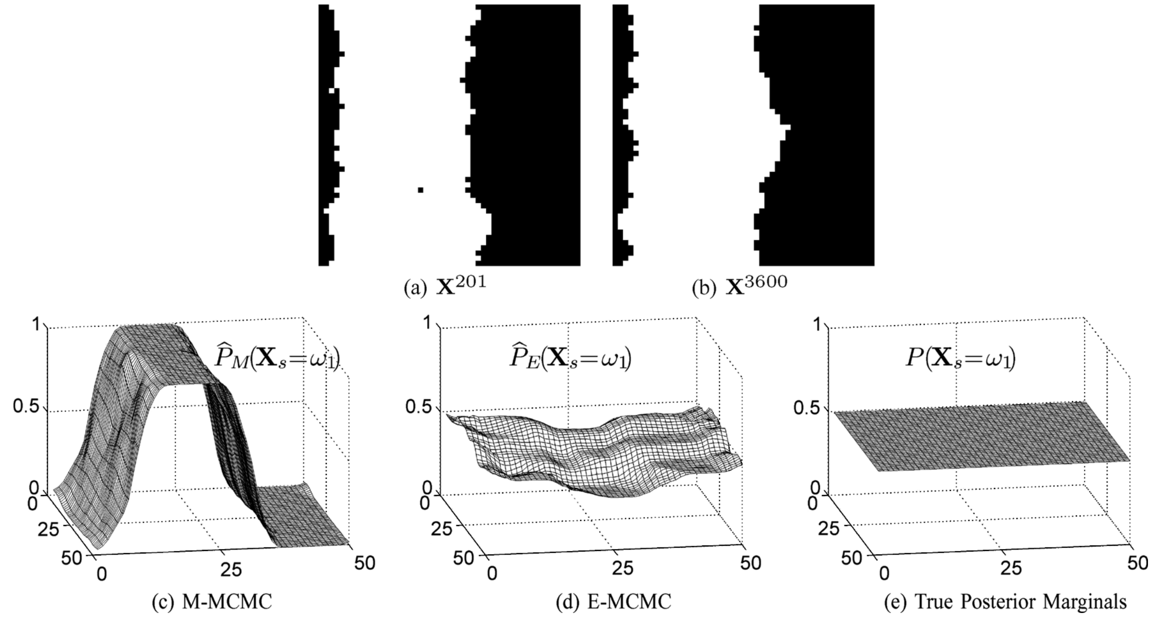
M-MCMC suffers from two major deficiencies. First, it is not robust to poorly-mixing chains. Consecutive samples in a simulated Markov chain generally exhibit high autocorrelations. This is not surprising as the Gibbs sampler generates sample $\mathbf{X}^k$ from $\mathbf{X}^{k-1}$. The autocorrelation can become problematic if $P(\mathbf{x})$ is multimodal. In this event the Gibbs sampler can become trapped in a single mode—typically the mode closest to the initial starting point $\mathbf{x}^0$—for a large number of iterations, and the resulting samples will not accurately reflect $P(\mathbf{x}|\mathbf{y})$. Such a chain is said to be poorly-mixing.

For example, consider a 50 × 50 grid of pixels whose distribution $P(\mathbf{x})$ is defined by the Potts [30] model (see Appendix B) with the two possible classes ω_1 and ω_2. Note that in this example we have no observations $\mathbf{Y}$. The Gibbs sampler can be used to generate samples from $P(\mathbf{x})$ by constructing a Markov chain $(\mathbf{X}^0,\mathbf{X}^1,\mathbf{X}^2,\ldots)$ with equilibrium distribution $P(\mathbf{x})$. Fig. 2(a) illustrates sample $\mathbf{X}^{201}$ (ω_1 is white). Fig. 2(b) illustrates sample $\mathbf{X}^{3600}$. Notice the similarity between the samples despite the large difference $(3600-201=3399)$ in their iteration indices. This suggests that the chain is poorly-mixing, and that the Markov chain has likely become stuck in a mode of $P(\mathbf{x}|\mathbf{y})$. Assuming $m=3400$ samples and a burn-in period of $b=200$, Fig. 2(c) illustrates the estimates of the posterior marginals $\widehat{P}_M(\mathbf{X}_s=\omega_1)$ using M-MCMC. The resulting estimates, which clearly reflect the samples shown in Fig. 2(a) and (b), are quite different from the true posterior marginals: $P(X_s=\omega_1)=1/2$ for all $s \in S$ [see Fig. 2(e)]. Note that the value 1/2 follows from the symmetry or exchangeability argument [31]. That is, $P(\mathbf{x})$ is symmetric with respect to ω_1 and ω_2, and exchanging their roles would not alter their respective probabilities. Therefore, their probabilities must be equal. (See Appendix B for further insight into the symmetry of the Potts prior $P(\mathbf{x})$.)

The second deficiency associated with M-MCMC is the resolution of the probability estimates, which is limited to $1/m$. That is, M-MCMC yields probability estimates that must be elements of the set $\{0,1/m,2/m,\ldots,1\}$. This is easily seen from (4). Thus, two probabilities separated by less than $1/m$ may not be distinguishable. Furthermore, the total number of unique probabilities is bounded by $m+1$. As we will see in Section IV, the number of unique probabilities determines the number of possible operating points.

We can rectify both deficiencies by extending M-MCMC. First, to improve the robustness with respect to multimodal distributions, (4) can be extended to sample over multiple chains,

[2] Dubes and Jain [29] refer to b and m as "magic" numbers.


(a) $\mathbf{X}^{201}$ (b) $\mathbf{X}^{3600}$

(c) M-MCMC (d) E-MCMC (e) True Posterior Marginals

Fig. 2. (a) Sample of Potts distribution ($\beta = 1$ and 8-connected neighborhood) with binary classes ω_1 (white) and ω_2 (black) generated by Gibbs sampler at iteration $k = 201$. (b) Sample of binary Potts distribution generated by Gibbs sampler at iteration $k = 3600$. (c) Estimate of posterior marginals using M-MCMC with $b = 200$ and $m = 3400$. (d) Estimate of posterior marginals using E-MCMC with $b = 200$, $m = 25$, and $c = 16$. (e) The true posterior marginals are $P(X_s = \omega_1) = 1/2$ for all $s \in S$. Note that E-MCMC (d) yielded more accurate estimates of the true marginals (e) than did M-MCMC (c).

with each chain initialized to a random state in Ω to ensure statistical independence (between chains) [26]. Second, to improve the resolution, instead of estimating $P(x_s|\mathbf{y})$ directly from samples of $\mathbf{X}^k$ as in (4), we can leverage the following result [24], [25]:

$$\begin{aligned} P(x_s|\mathbf{y}) &= \int P(x_s|\mathbf{x}_{-s},\mathbf{y})P(\mathbf{x}_{-s}|\mathbf{y})d\mathbf{x}_{-s} \\ &= \mathrm{E}\left[P(x_s|\mathbf{X}_{-s},\mathbf{y})\right]. \end{aligned} \quad (8)$$

Equation (8) states that the marginal distribution $P(x_s|\mathbf{y})$ is the expected value of the random function $P(x_s|\mathbf{X}_{-s},\mathbf{y})$. Since the expectation is with respect to $\mathbf{X}_{-s}$, the result is a distribution. Summarizing, the marginal distribution $P(x_s|\mathbf{y})$ of X_s can be found by averaging the function $P(x_s|\mathbf{x}_{-s},\mathbf{y})$ over all the possible states of $\mathbf{x}_{-s}$. Note that under the typical assumptions of the Markovity of $\mathbf{X}$ and the conditional independence of the observations $\mathbf{Y}$, the distribution $P(x_s|\mathbf{x}_{-s},\mathbf{y})$ reduces to $P(x_s|\mathbf{x}_{\eta s},y_s)$ and $\mathrm{E}[P(x_s|\mathbf{X}_{-s},\mathbf{y})]$ simplifies to $\mathrm{E}[P(x_s|\mathbf{X}_{\eta s},y_s)]$.

Incorporating multiple Markov chains along with the result from (8) into (4) yields our extended (and Rao–Blackwellized [32]) MCMC method (E-MCMC)

$$\widehat{P}(X_s = \omega|\mathbf{y}) \approx \frac{1}{c\cdot m}\sum_{j=1}^{c}\sum_{k=b+1}^{b+m} P\left(X_s = \omega|\mathbf{x}_{-s}^{j,k},\mathbf{y}\right) \quad (9)$$

where c is the total number of Markov chains and $\mathbf{x}_{-s}^{j,k}$ are the states of all sites except s in Markov chain j at iteration k. By averaging over the functional forms $P(x_s|\mathbf{x}_{-s},\mathbf{y})$—instead of the samples themselves as in (4)—E-MCMC eliminates the previous issue of resolution. To see this consider the following: 1) a single functional "sample" $P(x_s|\mathbf{x}_{-s},\mathbf{y})$ in (9) updates our estimate of $\widehat{P}(x_s|\mathbf{y})$ for all possible states of $x_s \in \Lambda$, and not just for the current state x_s^k as with (4) and 2) the degree of contribution to $\widehat{P}(x_s|\mathbf{y})$ is not fixed at $1/m$ as in (4), but instead varies according to the value of $P(x_s|\mathbf{x}_{-s}^{j,k},\mathbf{y})$. From a practical perspective, averaging over the functional forms of the conditional probability density functions $P(x_s|\mathbf{x}_{-s}^{j,k},\mathbf{y})$ significantly increases the number of unique probabilities as compared to averaging over the samples themselves; this is important since, as mentioned previously, the number of operating points is determined by the number of unique probabilities.

In general, the distributions $P(x_s|\mathbf{x}_{-s},\mathbf{y})$ contain more information about $P(x_s|\mathbf{y})$ than the individual samples X_s^k, and consequently yield more accurate estimates [24], [25]. Note that for poorly-mixing chains, extracting more than a single sample per chain (i.e., $m > 1$) provides little benefit (since the samples are highly dependent). Nonetheless, for certain chains and values of m it can be advantageous [27]. Also, some researchers have suggested that, depending on the burn-in period and the autocorrelation between samples in the chain, a single long chain may be more effective per sample than several shorter chains in certain circumstances [33], [34]. However, with the ubiquity of multiprocessor machines—each chain can be executed on a separate processor—such circumstances are becoming exceedingly rare.

It is insightful to return to the previous example in Fig. 2 and apply E-MCMC. The resulting estimates of the posterior marginals $\widehat{P}_E(\mathbf{X}_s = \omega_1)$ (with $b = 200$, $m = 25$, and $c = 16$) are illustrated in Fig. 2(d). Note that though M-MCMC and E-MCMC used the identical number of iterations ($16 \times [200 + 25] = 3600$), E-MCMC generated much more accurate estimates of the true marginals [Fig. 2(e)] than did M-MCMC [Fig. 2(c)].

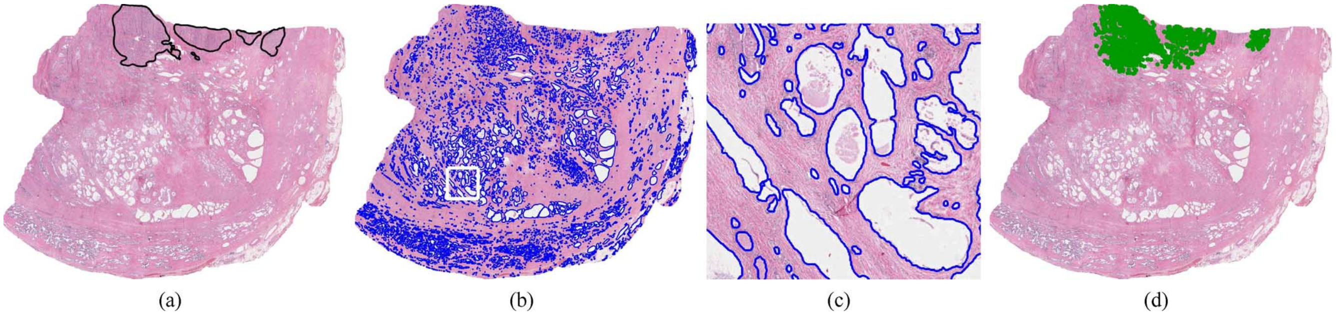


Fig. 3. (a) H&E stained whole-mount prostate histology section; black ink mark indicates "ground-truth" of CaP extent as delineated by a pathologist. (b) Result of automated gland segmentation [14]. (c) Magnified view of white box in (b). (d) Green dots indicate the centroids of those glands labeled as malignant.

IV. Experimental Results: Detection of Prostate Cancer on Histological Sections

In this section we incorporate WMPM into our MRF-based classification system for detecting carcinoma of the prostate (CaP) in (whole-mount or quarter) histological sections (HSs) from radical prostatectomies (RPs). Specifically, we show that by varying the class-specific weights inherent in the WMPM estimation criteria, we can arbitrarily adjust the detection sensitivity/specificity and generate receiver operator characteristic (ROC) curves. Additionally, we illustrate the advantages of the using E-MCMC instead of M-MCMC to estimate the requisite posterior marginals.

A. *System Description*

The analysis of histological sections from RPs plays a significant role in the diagnosis and treatment of prostate cancer. The most salient information in these HSs is derived from the morphology and architecture of the glandular structures. Since complex tasks such as Gleason grading [35]–[37] consider only the cancerous glands, an initial process capable of rapidly identifying these glands is highly desirable. In [14] we introduced an automated system for detecting CaP glands in Hematoxlyn and Eosin (H&E) stained tissue sections. The primary goal of this system is to eliminate regions that are not likely to be cancerous, thereby reducing the computational load of further, more sophisticated analyses. Consequently, in a clinical setting the algorithm should operate at a high detection sensitivity, ensuring that very little CaP is discarded.

It is important to mention that in [14] our CaP detection system did not use WMPM. It instead employed MAP estimation [2], implemented using iterated conditional modes (ICM) [9], a deterministic analogue of the Gibbs sampler. To adjust classifier performance we modified the single element clique potentials of the Markov prior, extending an idea first proposed by Comer *et al.* [19] (see Monaco *et al.* [13]). However, this approach requires rerunning ICM with every adjustment of the potentials; and consequently, generating a dense ROC curve in a reasonable amount of time is difficult. Thus, we were motivated to develop WMPM, since adjusting classifier performance via WMPM only requires comparing the appropriately weighted posterior marginal probabilities—a very rapid operation.

Fig. 3(a) illustrates a prostate HS from a RP specimen. The black lines indicate the spatial extent of CaP as delineated by a pathologist (and verified by a second pathologist). The numerous white regions are the glands—cavities in the tissue through which fluid flows—which our system automatically identifies and segments. Fig. 3(b) illustrates the segmented gland boundaries in blue. Fig. 3(c) provides a magnified view of the white box in Fig. 3(b). Following gland segmentation, the algorithm measures the area of each gland. Since malignant glands tend to be smaller than benign glands [38], this is a discriminative feature. Furthermore, malignant (benign) glands tend to proximate other malignant (benign) glands; this is modeled using a Markov prior. The WMPM classifier leverages these properties to label each gland as either malignant or benign. Fig. 3(d) illustrates the centroids of those glands labeled as malignant.

We now formally express this CaP detection problem using the MRF nomenclature established in Section II-A. Let the set $S = \{1, 2, \ldots, N\}$ reference the N segmented glands in a HS. Each site has an associated state $X_s \in \Lambda \equiv \{\omega_1, \omega_2\}$, where ω_1 and ω_2 indicate malignancy and benignity, respectively. The random variable $Y_s \in \mathbb{R}$ indicates the area of gland s. All feature vectors Y_s are assumed conditionally independent and identically distributed (i.i.d.) given their corresponding states, i.e., $P(\mathbf{y}|\mathbf{x}) = \prod_{s\in S} P(y_s|x_s)$. Each conditional distribution $P(y_s|x_s)$ is modeled parametrically using a mixture of Gamma distributions [14]; these distributions are fit from training samples using maximum likelihood estimation. The tendency for neighboring glands to share the same label is incorporated with a Markov prior $P(\mathbf{x})$ modeled using a probabilistic pairwise Markov model (PPMM) [14]; the PPMM is trained using maximum pseudo-likelihood estimation (MPLE) [9]. See Appendix C for a discussion of PPMMs. Two glands are considered neighbors if the distance between their centroids is less than 0.9 mm.

B. *Preliminaries*

The dataset consists of 27 digitized H&E stained histological sections from RPs obtained from 10 patients. The HSs primarily contain CaP with Gleason scores ranging from 6 to 8. All specimens were digitized at 40× magnification (0.25 μm per pixel) using an Aperio whole-slide digital scanner. A single pathologist then annotated the spatial extent of CaP on each digitized specimen, thus establishing the "ground truth" for classifier evaluation. (Note, all annotations were reviewed by a second pathologist.) An example annotation is shown in Fig. 3(a); the black line delineating CaP extent is overlaid on the image, and

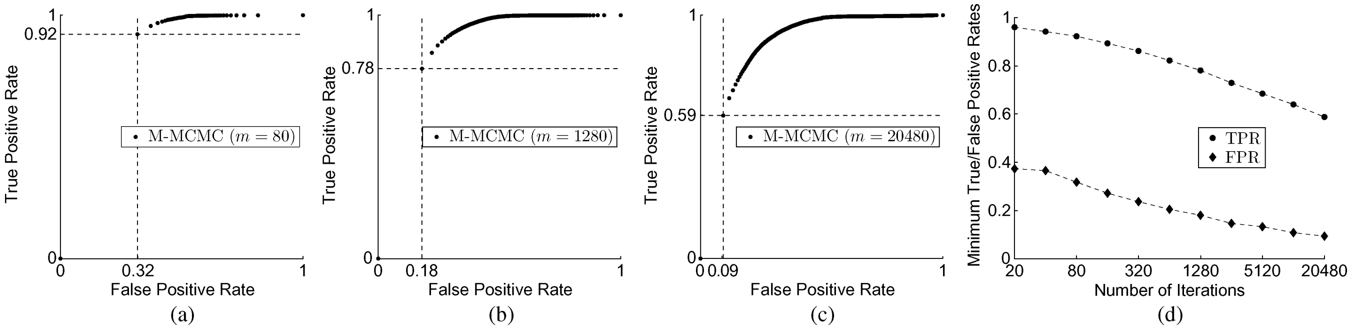


Fig. 4. (a)–(c) ROC curves using M-MCMC with $m \in \{80, 1280, 20480\}$ iterations. The dashed lines in (a)–(c) indicate the minimum TPR and FPR. (d) Plot of the minimum TPRs and FPRs using M-MCMC as a function of the number of iterations m.

was not present during processing. The Aperio scanner creates a multiresolution image pyramid for each digitized HS. The detection system processes the single image in this pyramid whose pixel width is 8 μm. This resolution is 1/32 of that available to the pathologist during annotation.

To assess performance we define the following: true positives (TP) are those segmented glands identified as cancerous by both the expert-provided “ground-truth”[3] and the automated system, true negatives (TN) are those segmented glands identified as benign by both the truth and the automated system, false positives (FP) are those segmented glands identified as benign by the truth and malignant by the automated system, and false negatives (FN) are those segmented glands identified as malignant by the truth and benign by the automated system. The true positive rate (TPR) and false positive rate (FPR) are given by TP/(TP+FN) and FP/(TN+FP), respectively. Note that the TPR and FPR are synonymous with the sensitivity and one minus the specificity, respectively.

WMPM classifies glands as follows: gland s belongs to ω_1 if $\alpha(\omega_1)\widehat{P}(\mathbf{X}_s = \omega_1|\mathbf{y}) > \alpha(\omega_2)\widehat{P}(\mathbf{X}_s = \omega_2|\mathbf{y})$, where $\widehat{P}(x_s|\mathbf{y})$ is the posterior marginal of gland s estimated using either M-MCMC or E-MCMC. More intuitively, pixel s is classified as ω_1 if $\widehat{P}(\mathbf{X}_s = \omega_1|\mathbf{y}) > T$, where the threshold $T \in [0, 1]$ is

$$T = \frac{\alpha(\omega_2)}{\alpha(\omega_1) + \alpha(\omega_2)}. \tag{10}$$

We can generate a ROC curve by varying T from 0 to 1, measuring the aggregate TP, FP, FN, and TN across all images, and then computing the TPR and FPR.

The following statements hold for all subsequent experiments: 1) the initial labeling $\mathbf{x}^0$ for each Markov chain is drawn randomly from a uniform distribution over Ω, 2) the training/testing datasets are determined by leave-one-out cross-validation, and 3) each Markov chain employs a burn-in period of $m = 10$ iterations (which was chosen empirically).

C. Experiment I: Quantitative Comparison of ROC Curves Using M-MCMC and E-MCMC

In the first experiment we compare the ROC curves generated by WMPM using M-MCMC and E-MCMC. Fig. 4(a)–(c) illustrate the ROC curves for M-MCMC with $m \in \{80, 1280, 20480\}$ iterations. As discussed previously, M-MCMC yields probability estimates whose resolution is limited to $1/m$. This manifests as a restriction upon the ranges of achievable TPRs and FPRs, and an upper bound of $m + 1$ on the total number of operating points. The dashed lines in Fig. 4(a)–(c) indicate the minimum possible TPR and FPR in each ROC curve. Fig. 4(d) plots these minimum rates as a function of the number of iterations m. The relationship is approximately log-linear. For example, reducing the true positive rate by 0.05 requires doubling the number of iterations. If this trend continues, then generating an ROC curve that extends to the origin would require over 83 million samples, and approximately half a year of processing time (estimates assume running our program on a 2.66 GHz Intel Xeon processor).

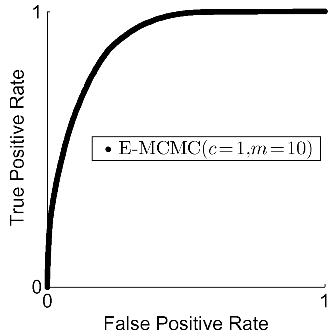


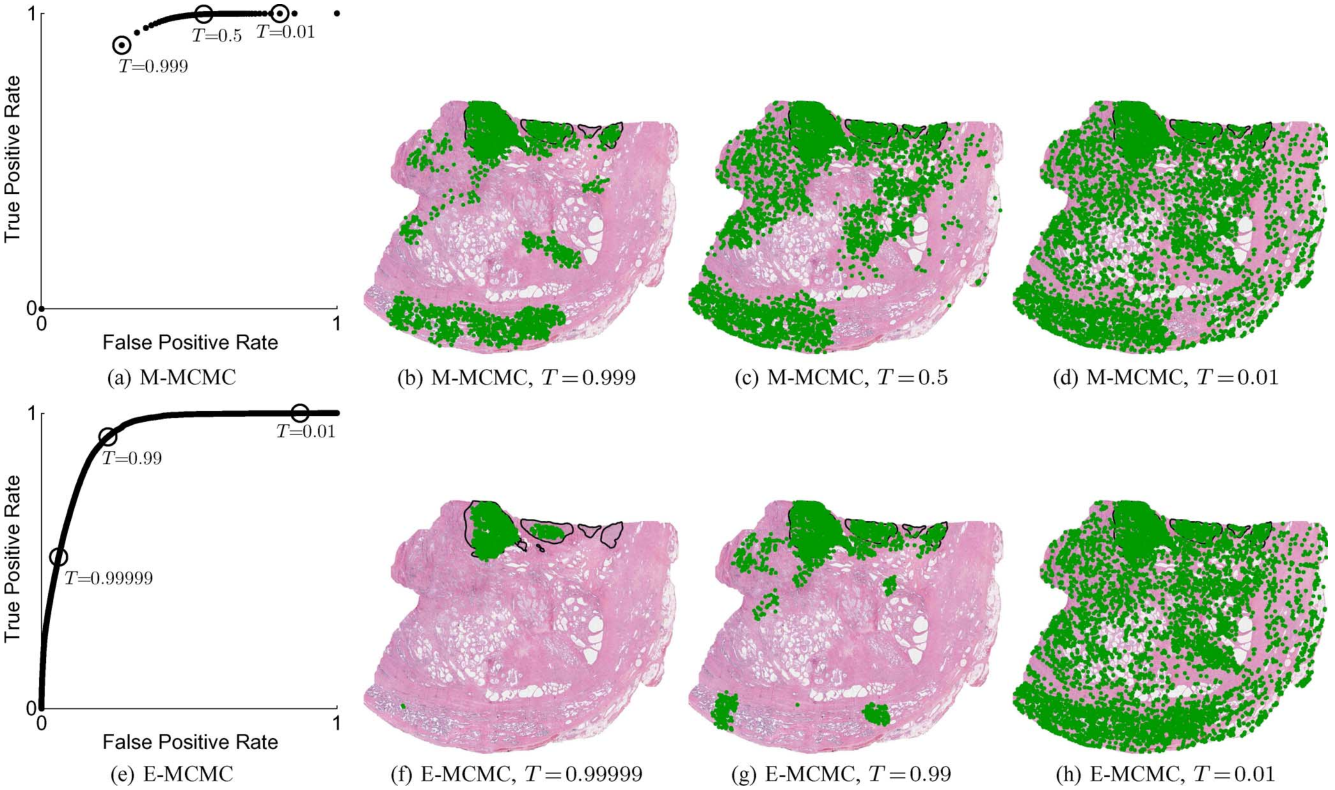
Fig. 5. ROC curve using E-MCMC with $m = 10$ iterations and a single Markov chain (i.e., $c = 1$).

Fig. 5 illustrates the ROC curve using E-MCMC with $m = 10$ iterations and $c = 1$ chains. Thus, even when using a single chain with only ten iterations, the resultant ROC curve is so densely populated that it appears continuous. Specifically, we have the following statistics: 1) the minimum and maximum TPRs (besides 0 and 1) are 4.2×10^{-5} and $1 - 4.2 \times 10^{-5}$, 2) the minimum and maximum FPRs are 1.4×10^{-5} and $1 - 1.4 \times 10^{-5}$, 3) 92 070 of the 94 999 total posterior marginal probabilities (i.e., there are 94 999 segmented glands across all 27 images) are unique, and 4) the maximum difference between TPRs (FPRs) measured at consecutive operating points is 2.5×10^{-5} (7.0×10^{-5}).

D. Experiment II: Qualitative Results of CaP Detection on Histological Sections Using M-MCMC and E-MCMC

The previous experiment demonstrated that using M-MCMC to estimate the posterior marginals limited the range of pos-

[3] Note that a segmented gland is considered cancerous with respect to the “ground-truth” if its centroid lies within a expert-provided CaP delineation.



(a) M-MCMC (b) M-MCMC, $T = 0.999$ (c) M-MCMC, $T = 0.5$ (d) M-MCMC, $T = 0.01$

(e) E-MCMC (f) E-MCMC, $T = 0.99999$ (g) E-MCMC, $T = 0.99$ (h) E-MCMC, $T = 0.01$

Fig. 6. (a), (e) ROC curves of CaP detection system on HSs using M-MCMC ($m = 150$ and $b = 10$) and E-MCMC ($m = 10, b = 10$, and $c = 8$). (b)–(d) Centroids of the glands (green dots) labeled as malignant using M-MCMC for $T \in \{0.999, 0.5, 0.01\}$. Corresponding system performances at these T values are indicated by the hollow black circles in (a). (f)–(h) Centroids of the glands labeled as malignant using E-MCMC for $T \in \{0.99999, 0.99, 0.01\}$. Corresponding system performances at these T values are indicated by the hollow black circles in (e). Note that even at the lowest (nonzero) FPR, WMPM using M-MCMC still misclassifies a considerable number of benign glands. By contrast, WMPM using E-MCMC demonstrates the existence of an operating point at which the system detects the majority of the cancerous glands while incurring almost no false positives.

sible sensitivities/specificities of our MRF-based CaP detection system. The current experiment qualitatively depicts the practical impact of this limitation by illustrating detection results at different operating points. For comparison, we provide these results alternately using M-MCMC and E-MCMC to estimate the posterior marginals.

To ensure a fair evaluation, we confine both M-MCMC ($b = 10$, $m = 150$) and E-MCMC ($b = 10$, $m = 10$, and $c = 8$) to an identical number of MCMC iterations $(8 \times [10 + 10] = 10 + 150 = 160)$. The choice of 160 is reasonable since processing this number of iterations requires approximately one minute (leaving sufficient time for the remainder of the detection process). The selection of eight chains for E-MCMC reflects the fact that our computer has eight processors, and each chain can be processed in parallel. Thus, E-MCMC can actually estimate the marginals eight times faster than M-MCMC.

Fig. 6(a) and (e) provides plots of the ROC curves when employing M-MCMC and E-MCMC, respectively. The remaining subfigures in Fig. 6 provide qualitative examples of the final classification results for these MCMC methods at three different thresholds T. The green dots indicate the centroids of those glands labeled as malignant. The system performances associated with the thresholds T are indicated with black circles on the corresponding ROC curves.

Again, M-MCMC restricts the number of possible TPRs and FPRs, diminishing the benefit of using WMPM. Specifically, Fig. 6(a) illustrates that even at the lowest (nonzero) FPR, WMPM still misclassifies a considerable number of benign glands as malignant. By contrast, Fig. 6(e) (using E-MCMC) demonstrates the existence of an operating point at which the system detects the majority of the cancerous glands while incurring almost no false positives. This operating point is not available with M-MCMC.

E. Experiment III: Comparison of Classifier Performance Using M-MCMC and E-MCMC

In the final experiment we demonstrate that using WMPM with E-MCMC as compared to M-MCMC yields superior classifier performance for any reasonable number of iterations m. We also show that when employing E-MCMC, increasing the number of chains c is a more effective means for enhancing performance than increasing the number of iterations m.

For number of chains $c \in \{1, 2, 4, 8, 16, 32, 64\}$ and iterations $m \in \{10, 20, 40, 80, 160\}$ we use E-MCMC to generate ROC curves over 10 leave-one-out cross-validation trials. (Since E-MCMC is a stochastic algorithm, each execution produces different results.) The area under the ROC curve (AUC) is calculated for each curve. Linear interpolation is used to make the curves continuous for integration. From each set of 10 AUCs we measure the mean $\overline{\mathrm{AUC}}_m^c$ and the standard deviation. Fig. 7(a) plots $\overline{\mathrm{AUC}}_m^8$ for $m \in \{10, 20, 40, 80, 160\}$ numbers of iterations, using a constant number of $c = 8$ chains. Fig. 7(b) plots

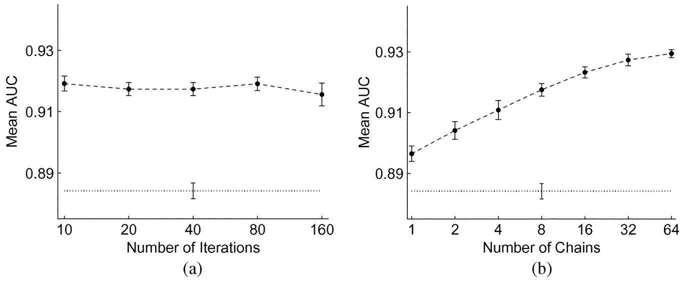


Fig. 7. (a) Mean AUCs (dashed line with dots) for gland detection system using E-MCMC with (a) $c = 8$ and $m \in \{10, 20, 40, 80, 160\}$ and (b) $m = 10$ and $c \in \{1, 2, 4, 8, 16, 32, 64\}$ from 10 leave-one-out cross-validation trials over 27 images from 10 patient studies. The error bars in both figures indicate the standard deviations of the measurements. Increasing the number of iterations provides no statistically significant improvement in mean AUC. Each increase in the number of chains from 1 to 32 does result in a statistically significant improvement in mean AUC. The dotted lines in both (a) and (b) illustrate the mean AUC for M-MCMC with $m = 20480$. Each mean AUC resulting from E-MCMC is significantly greater than the mean AUC using M-MCMC.

$\overline{\text{AUC}}_{10}^{c}$ for $c \in \{1, 2, 4, 8, 16, 32, 64\}$ numbers of chains, but with a constant number of $m = 10$ iterations. The error bars in Fig. 7(a) and (b) indicates the standard deviations.

Measuring statistical significance using a paired t-test with a significance level of 0.01, we can conclude the following: 1) for any number of iterations $m_1, m_2 \in \{10, 20, 40, 80, 160\}$ the difference $\Phi^{m_1,m_2} = \overline{\text{AUC}}_{m_1}^{8} - \overline{\text{AUC}}_{m_2}^{8}$ is not statistically significant under the null hypothesis that $\Phi^{m_1,m_2} = 0$, 2) if the number of Markov chains $c_1, c_2 \in \{1, 2, 4, 8, 16, 32\}$ and $c_1 > c_2$, then the difference $\Phi^{c_1,c_2} = \overline{\text{AUC}}_{10}^{c_1} - \overline{\text{AUC}}_{10}^{c_2}$ is statistically significant under the null hypothesis that $\Phi^{c_1,c_2} = 0$, and 3) the difference $\Phi^{64,32} = \overline{\text{AUC}}_{10}^{64} - \overline{\text{AUC}}_{10}^{32}$ is not statistically significant under the null hypothesis that $\Phi^{64,32} = 0$. It is worth recapitulating these statements regarding E-MCMC less formally: 1) increasing the number of samples results in no statistically significant difference in the mean AUC, 2) each increase in the number of chains from 1 to 32 results in a statistically significant increase in mean AUC, and 3) increasing the number of chains beyond 32 offers no improvement in mean AUC.

Similarly, we use M-MCMC with $m = 20480$ to generate ROC curves over 10 leave-one-out cross-validation trials. We then calculate the mean and standard deviation of the resulting AUCs. The dotted lines in Fig. 7(a) and (b) illustrate the mean AUC for M-MCMC with $m = 20480$. (Note that these lines are not functions of c.) The error bars indicate the standard deviations. Note that each mean AUC resulting from E-MCMC is greater than the mean AUC using M-MCMC (by margins that are statistically significant). This is a remarkable result since M-MCMC leverages far more samples m, and requires many more total MCMC iterations $(m + b)$.

With E-MCMC, increasing the number of chains is a more effective means of improving classifier performance than increasing the number of iterations. This is not surprising since each additional chain introduces a truly independent sample. Samples within the same chain are only independent when there is a sufficient number of iterations separating them. Since increasing the number of samples in a single chain [Fig. 7(a)] did not improve classification performance, it appears in this instance that the necessary separation is extremely large.

On a final note, since M-MCMC can yield sparsely sampled ROC curves over certain FPRs, one might inquire as to the validity of using AUCs to compare the classification performances of E-MCMC and M-MCMC. Let us explore this question in greater detail. The construction of ROC curves using finite datasets necessarily results in discrete operating points (paired sensitivities/specificities), and not continuous curves. The "underlying" continuous ROC curves—which are required for calculating the AUCs—must be interpolated. Therefore, any evaluation of performance using AUCs presupposes this interpolation is sufficiently accurate. The dense curves produced by E-MCMC [see Fig. 5 and Fig. 6(e)] seem more than adequate for such interpolation. However, even with $m = 20480$, M-MCMC [see Fig. 4(c)] yields operating points that are relatively sparse near the origin. Thus, linear interpolation might underestimate the true curve, and thus unfairly disadvantage the AUC (during comparison). Though we believe that any error in AUC estimation is negligible, we provide further evidence substantiating the superior performance of E-MCMC over M-MCMC. Instead of considering the entire ROC curve, we examine the TPRs of both M-MCMC ($m = 20480$) and E-MCMC ($c \in \{1, 2, 4, 8, 16, 32, 64\}$, $m = 10$) at three FPRs (0.25, 0.3, and 0.35) near which M-MCMC yields a dense sampling. Fig. 8 plots the mean and standard deviation of the resulting TPRs measured over the 10 leave-one-out cross-validation trials. As expected, the mean TPRs of E-MCMC—regardless of the number of chains c—exceed those of M-MCMC by margins that are statistically significant (using a paired t-test with a significance level of 0.02).

V. Concluding Remarks

The ability to adjust classifier performance (sensitivity/specificity) with respect to each class is essential for a variety of applications. Unfortunately, most MRF-based classifiers use estimation criteria such as maximum posterior marginals (MPM) and maximum *a posteriori* (MAP) [2] estimation that restrict their performance to a single, static operating point. To address this problem we introduced WMPM, an extension of MPM that allows for adjusting classifier performance by incorporating class-specific weighting into the MPM cost function. That is, whereas the MPM cost function weights each misclassification equally, WMPM provides class-specific penalties.

Ultimately, WMPM estimation reduces to the selection (at each site) of the class that maximizes that site's weighted posterior marginal distribution. Interestingly, this solution is analogous to the familiar means for varying the performance of univariate Bayesian systems—or multivariate systems with statistically independent variables: at each site the *a posteriori* probability associated with each class is appropriately weighted, and then the class with the greatest weighted probability is chosen. In the two-class case this reduces to the familiar thresholding. The analogous methodology for MRFs (i.e., WMPM) replaces the *a posteriori* probabilities with the posterior marginal probabilities.

The most complex aspect of both WMPM and MPM results from the need to estimate the posterior marginals. However, WMPM requires more accurate estimates of the posterior

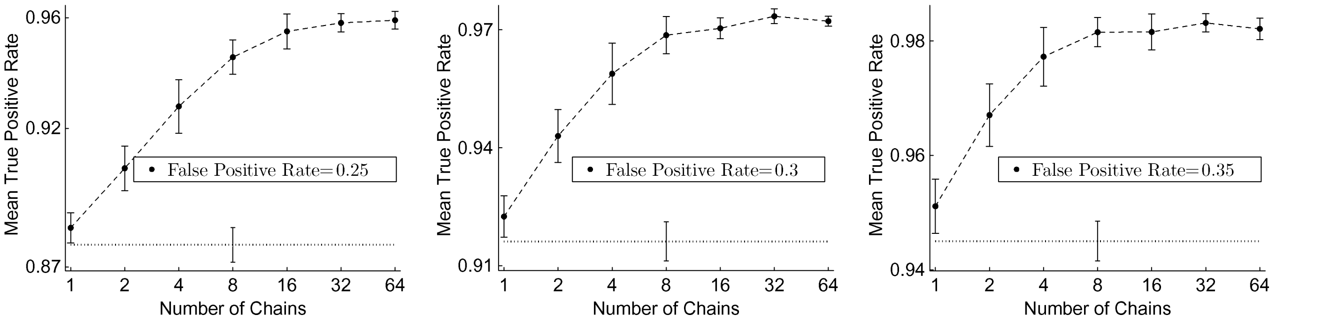


Fig. 8. Mean TPRs for gland detection system using E-MCMC (dashed line with dots) and M-MCMC (dotted line) at FPRs of 0.25, 0.3, and 0.35 from 10 leave-one-out cross-validation trials over 27 images from 10 patient studies. The error bars indicate the standard deviations of the measurements. For a given FPR, each mean TPR resulting from E-MCMC, regardless of the number of chains c, is significantly greater than the mean TPR using M-MCMC. Note that E-MCMC and M-MCMC use $m = 10$ and $m = 20480$ iterations, respectively.

marginals than MPM, and unfortunately the prevalent Markov chain Monte Carlo estimation method proposed by Marroquin *et al.* for use with MPM is inadequate for WMPM. Consequently, we presented E-MCMC, an extension of Marroquin's MCMC method that 1) samples over multiple chains instead of a single chain and 2) uses an ensemble average of conditional probability density functions instead of averaging over the Monte Carlo samples themselves.

We validated the efficacy of WMPM and E-MCMC by incorporating them into our automated system for detecting cancerous glands on digitized histological sections from radical prostatectomies [14]. Assuming a similar number of iterations for both M-MCMC and E-MCMC we observed the following: 1) E-MCMC yielded ROC curves with several orders of magnitude more operating points than those of M-MCMC, 2) E-MCMC allowed (virtually) any choice of true and false positive rates in [0,1], while M-MCMC confined these rates to much smaller subintervals, and 3) E-MCMC produced superior classification accuracy than M-MCMC as measured by area under the ROC curve.

To our knowledge the only previously reported means for adjusting the performance of MRF-based classifiers was to modify the single element clique potentials of the Markov prior [19]. In fact, our cancer detection system reported in [14] applied an extension of this methodology [13] to adjust the operating point of the MAP estimate. Unfortunately, this approach necessitates reperforming a complex MAP estimation procedure (e.g., relaxation procedures [2], [9], loopy belief propagation [39], or graph cuts [40]) with every change in the clique potentials. By contrast, modifying classifier performance with WMPM only requires adjusting the weights, and then comparing the weighted posterior marginals; the time-consuming step of estimating the marginals need only be performed once.

Before concluding, it is worthwhile to briefly consider two other techniques that could be used to vary the performance of MRF-based classifiers. The first leverages a unique property of iterated conditional modes (ICM), and was tangentially suggested in a seminal paper by Besag [9]. ICM is an iterative, deterministic procedure that converges to a local maximum of the MAP probability of a MRF. ICM requires the initial state of the MRF from which to begin the iteration; the choice of this state determines the local maximum to which ICM converges. Thus, varying the initial conditions can vary the classification results. However, the different modes of the MAP probability (to which ICM converges) do not necessarily correspond to meaningful classifications in a Bayesian sense. That is, this method, though perhaps intuitively appealing, seems to lack mathematical justification. The second possibility is to employ fuzzy MRFs [41], [42]. In theory, thresholds could be applied to each site's fuzzy membership values, yielding different classifications. However, fuzzy membership was intended to indicate the degree to which a single site belongs to each of the possible classes (e.g., to account for partial volume effects), and not to reflect the probability of belonging to a specific class. Thus, constructing ROC curves in this manner appears heuristic.

Finally, note that we elected to demonstrate WMPM using an application with two classes (cancer and benign) because it simplifies presentation and comprehension while allowing the construction of ROC curves. Additionally, any multiclass problem can be decomposed into a series of binary class problems. Nonetheless, all the derivations and conclusions in this paper are applicable to any number of classes.

Appendix

A. Gibbs Formulation

The connection between the Markov property and the joint probability density function of $\mathbf{X}$ is revealed by the Hammersley–Clifford (Gibbs–Markov equivalence) theorem [43]. This theorem states that a random field (G, Ω, P) with $P(\mathbf{x}) > 0$ for all $\mathbf{x} \in \Omega$ satisfies the Markov property if, and only if, it can be expressed as a Gibbs distribution

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left\{\sum_{c \in \mathcal{C}} V_c(\mathbf{x})\right\} \tag{11}$$

where $Z = \sum_{\mathbf{x} \in \Omega} \exp\{\sum_{c \in \mathcal{C}} V_c(\mathbf{x})\}$ is the normalizing constant and V_c are positive functions, called clique potentials, that depend only on those x_s such that $s \in c$. A clique c is any subset of S which constitutes a fully connected subgraph of G; the set $\mathcal{C}$ contains all possible cliques. Note that typically $|\Omega| = |\Lambda|^N$ is too large to deterministically evaluate Z. The following reveals the forms of the local conditional probability density functions:

$$P\left(x_s | \mathbf{x}_{\eta_s}\right) = \frac{1}{Z_s} \exp\left\{\sum_{c \in \mathcal{C}_s} V_c(\mathbf{x})\right\} \tag{12}$$

where $\mathcal{C}_s$ represents $\{c \in \mathcal{C} : s \in c\}$ and $Z_s = \sum_{x_s \in \Lambda} \exp\{\sum_{c \in \mathcal{C}_s} V_c(\mathbf{x})\}$. For proofs of Markov formulations and theorems, see Geman [44].

B. Potts Model

Gaussian MRFs notwithstanding, the Potts [30] Markov prior $P(\mathbf{x})$, a multiclass generalization of the Ising automodel [45], is the most prevalent MRF formulation. The potential functions of the Potts model are pairwise. That is, only two-element cliques yield values that are not identically one. The local conditional probability density functions (and implicitly the potential functions) are defined as follows:

$$P\left(x_s|\mathbf{x}_{\eta_s}\right) = \frac{1}{Z_s} \exp\left\{\beta \sum_{r \in \eta_s} \delta(x_s - x_r)\right\} \tag{13}$$

where $\beta \in \mathbb{R}$ and Z_s is the normalizing constant that ensures $\sum_{x_s \in \Lambda} P(x_s|\mathbf{x}_{\eta_s}) = 1$. Note that greater values of β produce "smoother" solutions.

C. Probabilistic Pairwise Markov Models

Before discussing probabilistic pairwise Markov models (PPMMs), we must first introduce additional notation. As discussed previously, $P(\cdot)$ indicates the probability of event $\{\cdot\}$. For instance, $P(X_s = x_s)$ and $P(\mathbf{X} = \mathbf{x})$ signify the probabilities of the events $\{X_s = x_s\}$ and $\{\mathbf{X} = \mathbf{x}\}$. Note that we simplified such notations in the paper—when it did not cause ambiguity—by omitting the random variable, e.g., $P(\mathbf{x}) \equiv P(\mathbf{X} = \mathbf{x})$. We now introduce $p(\cdot)$, which indicates a generic (discrete) probability function; for example, p_u might be a uniform distribution. The notations $P(\cdot)$ and $p(\cdot)$ are useful in differentiating $P(x_s)$ which indicates the probability that $\{X_s = x_s\}$ from $p_u(x_s)$ which refers to the probability that a uniform random variable assumes the value x_s.

Continuing, in place of potential functions (i.e., a Gibbs formulation), PPMMs [14] formulate the local conditional probability density functions (LCPDFs) $P(x_s|\mathbf{x}_{\eta_s})$ of an MRF in terms of pairwise density functions, each of which models the interaction between two neighboring sites. This formulation facilitates the creation of relatively sophisticated LCPDFs (and hence priors), increasing our ability to model complex processes. Within the context of our CaP detection system, we previously demonstrated the superiority of PPMMs over the prevalent Potts model [14]. The PPMM formulation of the LCPDFs is as follows:

$$P\left(x_s|\mathbf{x}_{\eta_s}\right) = \frac{1}{Z_s} p_0(x_s) \prod_{r \in \eta_s} p_{1|0}(x_r|x_s) \tag{14}$$

where the normalizing constant Z_s ensures summation to one, p_0 is the probability density function (PDF) describing the stationary site s, and $p_{1|0}$ represents the conditional PDF describing the pairwise relationship between site s and its neighboring site r. The numbers 0 and 1 replace the letters s and r to indicate that the probabilities are identical across all sites, i.e., the MRF is stationary. Furthermore, p_0 and $p_{1|0}$ are related in the sense that they are a marginal and conditional distribution of the joint distribution $p_{0,1}$, i.e., $p_{0,1}(x_s, x_r) = p_0(x_s)p_{1|0}(x_r, x_s)$. We are free to choose any forms for p_0 and $p_{1|0}$, under the caveat that $p_{0,1}$ be symmetric to ensure stationarity.

References

[1] S. C. Agner, S. Soman, E. Libfeld, M. McDonald, K. Thomas, S. Englander, M. Rosen, D. Chin, J. Nosher, and A. Madabhushi, "Textural kinetics: A novel dynamic contrast-enhanced (DCE)-MRI feature for breast lesion classification," *J. Digital Imag.*, pp. 1–18, May 2010.

[2] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Pattern Recog. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[3] T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 901–914, Apr. 1992.

[4] A. Farag, A. El-Baz, and G. Gimel'farb, "Precise segmentation of multimodal images," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 952–968, Apr. 2006.

[5] S. Awate, T. Tasdizen, and R. Whitaker, "Unsupervised texture segmentation with nonparametric neighborhood statistics," in *Comput. Vis. ECCV*, 2006, pp. 494–507.

[6] X. Liu, D. L. Langer, M. A. Haider, Y. Yang, M. N. Wernick, and I. S. Yetik, "Prostate cancer segmentation with simultaneous estimation of Markov random field parameters and class," *IEEE Trans. Med. Imag.*, vol. 28, no. 6, pp. 906–915, Jun. 2009.

[7] B. Scherrer, F. Forbes, C. Garbay, and M. Dojat, "Distributed local MRF models for tissue and structure brain segmentation," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1278–1295, Aug. 2009.

[8] J. Tohka, I. Dinov, D. Shattuck, and A. Toga, "Brain MRI tissue classification based on local Markov random fields," *Magn. Reson. Imag.*, vol. 28, no. 4, pp. 557–573, May 2010.

[9] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Stat. Soc. Series B (Methodological)*, vol. 48, no. 3, pp. 259–302, 1986.

[10] M. A. T. Figueiredo and J. M. N. Leitao, "Unsupervised image restoration and edge location using compound Gauss-Markov random fields and the MDL principle," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1089–1102, Aug. 1997.

[11] R. Paget and I. D. Longstaff, "Texture synthesis via a noncausal nonparametric multiscale Markov random field," *IEEE Trans. Image Process.*, vol. 7, no. 6, pp. 925–931, Jun. 1998.

[12] A. Zalesny and L. Van Gool, "A compact model for viewpoint dependent texture synthesis," in *Smile Workshop*, London, U.K., 2001, pp. 124–143.

[13] J. Monaco, S. Viswanath, and A. Madabhushi, "Weighted iterated conditional modes for random fields: Application to prostate cancer detection," in *Workshop on Probabilistic Models for Medical Image Analysis (in Conjunction With MICCAI)*, London, U.K., pp. 209–217.

[14] J. P. Monaco, J. Tomaszewski, M. Feldman, I. Hagemann, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi, "High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models," *Med. Image Anal.*, vol. 14, no. 4, pp. 617–629, 2010.

[15] R. Brem, J. Rapelyea, G. Zisman, J. Hoffmeister, and M. DeSimio, "Evaluation of breast cancer with a computer-aided detection system by mammographic appearance and histopathology," *Cancer*, vol. 104, no. 5, pp. 931–935, 2005.

[16] J. Baker, E. Rosen, J. Lo, E. Gimenez, R. Walsh, and M. Soo, "Computer-aided detection (CAD) in screening mammography: Sensitivity of commercial cad systems for detecting architectural distortion," *Am. J. Roentgenol.*, vol. 181, no. 4, pp. 1083–1088, 2003.

[17] R. Duda, P. Hart, and D. Stork*, Pattern Classification.* New York: Wiley, 2001.

[18] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Am. Stat. Assoc.*, vol. 82, no. 397, pp. 76–89, 1987.

[19] M. L. Comer and E. J. Delp, "Segmentation of textured images using a multiresolution Gaussian autoregressive model," *IEEE Trans. Image Process.*, vol. 8, no. 3, pp. 408–420, Mar. 1999.

[20] J. Simmons, P. Chuang, M. Comer, J. Spowart, M. Uchic, and M. De Graef, "Application and further development of advanced image processing algorithms for automated analysis of serial section image data," *Modell. Simulat. Mater. Sci. Eng.*, vol. 17, no. 2, p. 025002, 2009.

[21] B. Tso and R. Olsen, "A contextual classification scheme based on MRF model with improved parameter estimation and multiscale fuzzy line process," *Remote Sensing Environ.*, vol. 97, no. 1, pp. 127–136, 2005.

[22] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, pp. 1087–1092, Jun. 1953.

[23] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[24] J. Liu, W. Wong, and A. Kong, "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes," *Biometrika*, vol. 81, no. 1, pp. 27–40, 1994.

[25] A. Gelfand and A. Smith, "Sampling-based approaches to calculating marginal densities," *J. Am. Stat. Assoc.*, vol. 85, no. 410, pp. 398–409, 1990.

[26] A. Gelman and D. Rubin, "Inference from iterative simulation using multiple sequences," *Stat. Sci.*, vol. 7, no. 4, pp. 457–472, 1992.

[27] G. Casella and E. George, "Explaining the Gibbs sampler," *Am. Stat.*, vol. 46, no. 3, pp. 167–174, 1992.

[28] L. Tierney, "Markov chains for exploring posterior distributions," *Ann. Stat.*, vol. 22, no. 4, pp. 1701–1728, 1994.

[29] R. C. Dubes, A. K. Jain, S. G. Nadabar, and C. C. Chen, "MRF model-based algorithms for image segmentation," in *Proc. 10th Int. Conf. Pattern Recognit.*, 1990, vol. 1, pp. 808–814.

[30] R. B. Potts, "Some generalised order-disorder transformations," in *Proc. Cambridge Philos. Soc.*, 1952, vol. 48, pp. 106–109.

[31] A. Gelman, J. Carlin, H. Stern, and D. Rubin*, Bayesian Data Analysis*. Boca Raton, FL: CRC, 2004.

[32] G. Casella and C. Robert, "Rao-blackwellisation of sampling schemes," *Biometrika*, vol. 83, no. 1, pp. 81–94, 1996.

[33] C. Geyer, "Practical Markov chain Monte Carlo," *Stat. Sci.*, vol. 7, no. 4, pp. 473–483, 1992.

[34] A. Raftery and S. Lewis, "[Practical Markov chain Monte Carlo]: Comment: One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo," *Stat. Sci.*, vol. 7, no. 4, pp. 493–497, 1992.

[35] D. Gleason, "Classification of prostatic carcinomas," *Cancer Chemotherapy Rep.*, vol. 50, pp. 125–128, 1966.

[36] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted Bayesian multi-resolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans. Biomed. Eng.*, to be published.

[37] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and Gleason grading of histological images," *IEEE Trans. Med. Imag.*, vol. 26, no. 10, pp. 1366–1378, Oct. 2007.

[38] V. Kumar, A. Abbas, and N. Fausto*, Robbins and Cotran Pathologic Basis of Disease*. Philadelphia, PA: Saunders, 2004.

[39] J. Yedidia, W. Freeman, and Y. Weiss, "Generalized belief propagation," *Adv. Neural Inf. Process. Syst.*, pp. 689–695, 2000.

[40] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[41] S. Ruan, B. Moretti, J. Fadili, and D. Bloyet, "Fuzzy Markovian segmentation in application of magnetic resonance images," *Computer Vis. Image Understand.*, vol. 85, no. 1, pp. 54–69, 2002.

[42] F. Salzenstein and C. Collet, "Fuzzy Markov random fields versus chains for multispectral image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1753–1767, Nov. 2006.

[43] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. R. Stat. Soc. Series B (Methodological)*, vol. 36, no. 2, pp. 192–236, 1974.

[44] D. Geman*, Random Fields and Inverse Problems in Imaging*. New York: Springer, 1991, vol. 1427, Lecture Notes in Mathematics, pp. 113–193.

[45] E. Ising, "Beitrag zur theorie des ferromagnetismus," *Zeitschrift fr Physik*, vol. 31, pp. 253–258, 1925.