Large Multi-scale Spatial Modeling Using Tree Shrinkage Priors

Rajarshi Guhaniyogi and Bruno Sanso

Department of Statistics,

University of California, Santa Cruz, CA 95064,

correspondence email: rguhaniy@ucsc.edu

Abstract: This document contains posterior computation for MDCT model, posterior computation of MDCT with non Gaussian data, two dimensional illustration of MDCT with binary spatial data, theoretical properties of MDCT along with the proofs.

2. Posterior computation for MDCT with Gaussian model

We proceed to do parametric inference with data $(y(\boldsymbol{s}_i),\boldsymbol{x}(\boldsymbol{s}_i))_{i=1}^n$ at locations $\mathcal{S}=\{\boldsymbol{s}_1,...,\boldsymbol{s}_n\}$. Stacking responses and predictors across locations we obtain, $\boldsymbol{y}=(y(\boldsymbol{s}_1),...,y(\boldsymbol{s}_n))'$, $\boldsymbol{X}=[\boldsymbol{x}(\boldsymbol{s}_1):\cdots:\boldsymbol{x}(\boldsymbol{s}_n)]'$. Let $\boldsymbol{K}$ be an $n\times(J(1)+\cdots+J(R))$ matrix whose ith row is given by $(K(\boldsymbol{s}_i,\boldsymbol{s}_1^1,\phi_1),...,K(\boldsymbol{s}_i,\boldsymbol{s}_R^{J(R)},\phi_R))'$. Further assume $\boldsymbol{\beta}^r=(\beta_1^r,...,\beta_{J(r)}^r)'$ and $\boldsymbol{\beta}=(\boldsymbol{\beta}^1,...,\boldsymbol{\beta}^R)'$, $y_{i,r,j}=y_i-\sum_{(k_1,k_2)\neq(j,r)}K(\boldsymbol{s}_i,\boldsymbol{s}_{k_1}^{k_2},\phi_{k_2})\beta_{k_1}^{k_2}$, $\boldsymbol{y}_{r,j}=(y_{1,r,j},...y_{n,r,j})'$, $\boldsymbol{K}_{r,j}=(K(\boldsymbol{s}_1,\boldsymbol{s}_j^r,\phi_r),...,K(\boldsymbol{s}_n,\boldsymbol{s}_j^r,\phi_r))'$. The full conditional distributions of $\boldsymbol{\gamma}$, σ^2, β_j^r and $\delta_{j,r}$ are readily available in closed form and are given by

- $\boldsymbol{\gamma}|-\sim N((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{K}\boldsymbol{\beta}),\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$
- $\sigma^2|-\sim IG\left(\frac{n}{2}+c,d+\frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\gamma}-\boldsymbol{K}\boldsymbol{\beta}||^2\right)$
- $\beta_j^r|-\sim N\left(\frac{\frac{\boldsymbol{K}_{r,j}'\boldsymbol{y}_{r,j}}{\sigma^2}}{\frac{1}{\alpha_j^r}+\frac{\boldsymbol{K}_{r,j}'\boldsymbol{K}_{r,j}}{\sigma^2}},\frac{1}{\frac{1}{\alpha_j^r}+\frac{\boldsymbol{K}_{r,j}'\boldsymbol{K}_{r,j}}{\sigma^2}}\right)$. Parallelized block updating of $\boldsymbol{\beta}$ is described in

the Section 4.1.

- Recall the definition of father node in Section 3.2. Additionally define $father^2(\boldsymbol{s}_j^r)$ as the father node of the father node of $\boldsymbol{s}_j^r$. Similarly, $father^3, ..., father^R$ node are defined. Let $\alpha_{j,r,-1} = \prod_{l=r}^{R} \delta_{father^{r+1-l}(\boldsymbol{s}_j^l),l-1}$ and $\alpha_{j,1,-1} = 1$. Then $\delta_1|- \sim Gamma\left(1 + \frac{J(1)+\cdots+J(R)}{2}, 1 + \frac{1}{2}\sum_{r=1}^{R}\sum_{j=1}^{J(r)}\left[(\beta_j^r)^2/\alpha_{j,r,-1}\right]\right)$

- Let $\alpha_{k,l,-r,-j} = \delta_1 \prod_{h=l}^{r+2} \delta_{father^{l+1-h}(\boldsymbol{s}_k^h),h-1} \prod_{h=r}^{2} \delta_{father^{r+1-h}(\boldsymbol{s}_j^h),h-1}$, $\alpha_{j,r,-r,-j} = 1$. Then $\delta_j^r|- \sim Gamma\left(c + \frac{\#\boldsymbol{\beta}_{j,r}^{Subtree}}{2}, 1 + \frac{1}{2}\sum_{l\geq r, \boldsymbol{s}_k^r \in Subtree(\boldsymbol{s}_j^r)} (\beta_k^l)^2/\alpha_{k,l,-r,-j}\right)$ for $r > 1$.

- Finally at each iteration, joint posterior distribution is maximized over a discrete grid of η values fixing all other parameters at the current iterate, $\eta \in \{1, ..., h_\eta\}$, h_η is an integer. In all simulation studies and in the real data analysis, we never found the maximization of the posterior over η to occur for η values more than 5. Thus, we fix $h_\eta = 5$ for all empirical investigations. We must mention that the posterior is maximized at $\eta = 1, 2$ in most of the iterations.

3. Posterior computation for MDCT with non Gaussian data

With scale mixture of Gaussian representation for the t-distribution, equation (5.10) in the main article can be written as

$$y_i \sim N(\boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\gamma} + \sum_{r=1}^{R}\sum_{j=1}^{J(r)} K(\boldsymbol{s}, \boldsymbol{s}_j^r, \phi_r)\beta_j^r, \frac{\sigma^2}{\lambda_i}),\ \lambda_i \sim Gamma(\nu/2, \nu/2).\ i = 1, .., n.$$

The priors on other parameters are kept the same as in the section with Gaussian data. Let $\boldsymbol{D} = diag(1/\lambda_1, ..., 1/\lambda_n)$. Full conditional posterior distributions of the parameters come in closed form given as following.

- $\gamma|- \sim N((\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}(\boldsymbol{y}-\boldsymbol{K\beta}), \sigma^2(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1})$

- $\sigma^2|- \sim IG\left(\frac{n}{2}+c, d+\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\gamma}-\boldsymbol{K\beta})'\boldsymbol{D}^{-1}(\boldsymbol{y}-\boldsymbol{X\gamma}-\boldsymbol{K\beta})\right)$

- $\lambda_i|- \sim Gamma(\frac{\nu+1}{2}, \frac{\nu}{2}+(y_i-\boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\gamma}-\sum_{r=1}^{R}\sum_{j=1}^{J(r)}K(\boldsymbol{s},\boldsymbol{s}_j^r,\phi_r)\beta_j^r)^2/2)$

- $\beta_j^r|- \sim N\left(\frac{\frac{\boldsymbol{K}'_{r,j}\boldsymbol{D}^{-1}\boldsymbol{y}_{r,j}}{\sigma^2}}{\frac{1}{\alpha_j^r}+\frac{\boldsymbol{K}'_{r,j}\boldsymbol{D}^{-1}\boldsymbol{K}_{r,j}}{\sigma^2}}, \frac{1}{\frac{1}{\alpha_j^r}+\frac{\boldsymbol{K}'_{r,j}\boldsymbol{D}^{-1}\boldsymbol{K}_{r,j}}{\sigma^2}}\right)$. Parallelized block updating of $\boldsymbol{\beta}$ is performed as described in Section 4.1.

- Recall the definition of father node in Section 3.2. Additionally define $father^2(\boldsymbol{s}_j^r)$ as the father node of the father node of $\boldsymbol{s}_j^r$. Similarly, $father^3, ..., father^R$ node are defined. Let $\alpha_{j,r,-1} = \prod_{l=r}^{R}\delta_{father^{r+1-l}(\boldsymbol{s}_j^l),l-1}$ and $\alpha_{j,1,-1}=1$. Then
  $\delta_1|- \sim Gamma\left(1+\frac{J(1)+\cdots+J(R)}{2}, 1+\frac{1}{2}\sum_{r=1}^{R}\sum_{j=1}^{J(r)}\left[(\beta_j^r)^2/\alpha_{j,r,-1}\right]\right)$

- Let $\alpha_{k,l,-r,-j} = \delta_1\prod_{h=l}^{r+2}\delta_{father^{l+1-h}(\boldsymbol{s}_k^h),h-1}\prod_{h=r}^{2}\delta_{father^{r+1-h}(\boldsymbol{s}_j^h),h-1}$, $\alpha_{j,r,-r,-j}=1$. Then
  $\delta_j^r|- \sim Gamma\left(c+\frac{\#\boldsymbol{\beta}_{j,r}^{Subtree}}{2}, 1+\frac{1}{2}\sum_{l\geq r,\boldsymbol{s}_k^r\in Subtree(\boldsymbol{s}_j^r)}(\beta_k^l)^2/\alpha_{k,l,-r,-j}\right)$ for $r>1$.

- Finally at each iteration, joint posterior distribution is maximized over a discrete grid of η values fixing all other parameters at the current iterate, $\eta\in\{1,...,h_\eta\}$, h_η is an integer. In all simulation studies and in the real data analysis, we never found the maximization of the posterior over η to occur for η values more than 5. Thus, we fix $h_\eta = 5$ for all empirical investigations. We must mention that the posterior is maximized at $\eta = 1,2$ in most of the iterations.

4. Two dimensional illustration of MDCT with binary spatial data

To demonstrate the flexibility offered by MDCT as opposed to predictive methods (such as LaGP), performance of MDCT is investigated under non-Gaussian binary spatial data.

For this purpose $10,500$ observations within $[0,1] \times [0,1]$ domain are generated from the probit spatial regression model. More precisely, with $\boldsymbol{x}(\boldsymbol{s}_i)$ as the predictor vector at $\boldsymbol{s}_i$, the response y_i is simulated using

$$y_i \overset{ind}{\sim} Ber(p_i)$$

$$\Phi^{-1}(p_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\gamma} + w_0(\boldsymbol{s}_i).$$

The model includes an intercept γ_0 and a predictor $\boldsymbol{x}(\boldsymbol{s})$ drawn i.i.d from from $N(0,1)$ with the corresponding coefficient γ_1, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$. $\boldsymbol{w}_0 = (w_0(\boldsymbol{s}_1), ..., w_0(\boldsymbol{s}_n))'$ is an n dimensional vector that follows a multivariate normal distribution with mean $\mathbf{0}_n$ and the covariance matrix of the order $n \times n$ specified through the Matérn (see equation (5.9) in the main draft) class of correlation functions. A random subset of 10000 observations are selected for model fitting and the rest is used to judge performance of MDCT as a binary classifier.

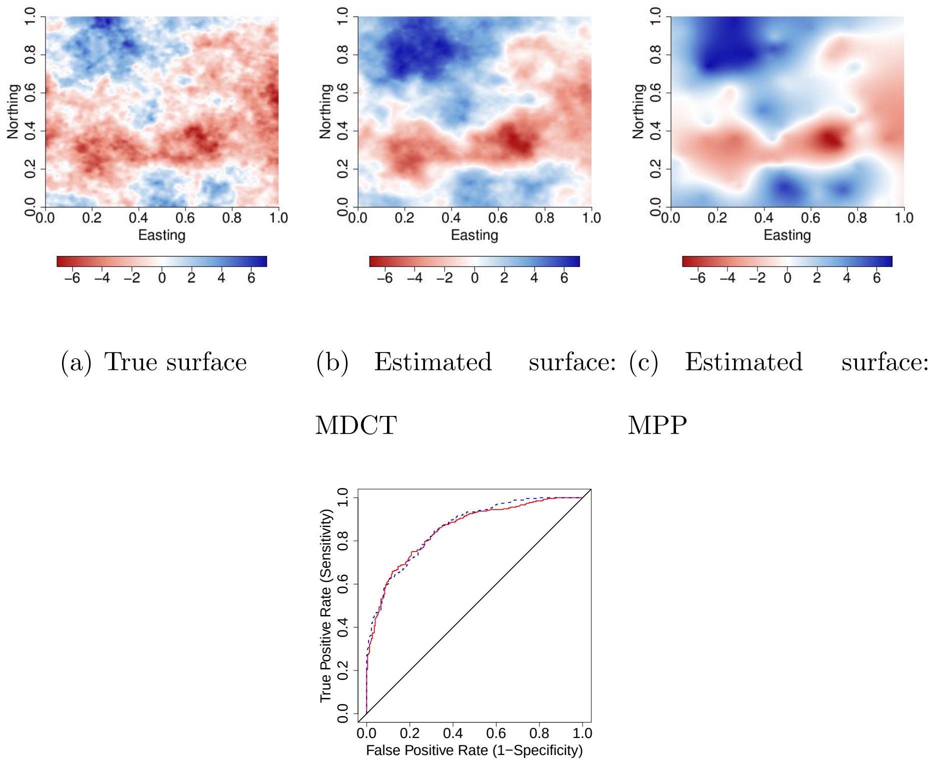
To implement MDCT for binary spatial data, we employ

$$y_i \overset{ind}{\sim} Ber(p_i), \Phi^{-1}(p_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\gamma} + \sum_{r=1}^{R} \sum_{j=1}^{J(r)} K(\boldsymbol{s}, \boldsymbol{s}_j^r, \phi_r)\beta_j^r.$$

The posterior computation of all model parameters proceeds with the standard data augmentation procedure (see Albert and Chib (1993)). Due to space constraint, we omit all the details and plan to elaborate it in a future article. For the sake of our exposition, MDCT is implemented with 3 resolutions having a total of 2100 basis functions.

Note that the binary regression precludes the possibility of employing LaGP as a competitor. On the other hand, `LatticeKrig` package implements LatticeKrig only for continuous response. Thus as a competitor, binary spatial regression with modified predictive process is implemented in `R` package `spBayes`.

Figure 1 shows the true surface and estimated surfaces from MDCT and MPP. Since



(a) True surface

(b) Estimated surface: MDCT

(c) Estimated surface: MPP

(d) ROC out of sample

Figure 1: (a) True surface generating the data. Figures (b) and (c) present the posterior predictive mean of estimated spatial surfaces from MDCT and MPP. (d) shows out of sample ROC curves for MPP and MDCT. Dotted line presents ROC for MDCT, while solied line presents ROC for MPP.

the surface estimation from binary spatial regression is a notoriously challenging problem, it comes with no surprise that the performance of all competitors deteriorate when compared with continuous response case discussed in Section 5.2. However, among the two competitors MDCT outperforms MPP considerably. It becomes clear from Figure 1 that MPP undergoes massive oversmoothing and loses most of the local features in the spatial surface. MDCT also experiences smoothing, though to a much lesser degree than MPP. Referring to Figure 1(d), MDCT appears to be marginally better than MPP in terms of out of sample classification (Area under the ROC curve for MDCT is 0.72, while the same for MPP is 0.69). The binary spatial regression analysis further corroborates the flexibility and accuracy of MDCT.

5. Theoretical properties

We establish convergence results for multiscale DCT regression model under the simplifying assumption that the predictor coefficient $\boldsymbol{\gamma} = (0, \ldots, 0)$.

Define two metrics in the function space given by

$$||w||_\infty = \sup_{\boldsymbol{s}\in\mathcal{D}} |w(\boldsymbol{s})|,$$
$$||w||_\zeta = \max_{k\leq\lfloor\zeta\rfloor} \sup_{\boldsymbol{s}\in\mathcal{D}} |D^k w(\boldsymbol{s})| + \max_{\tilde{k}\leq\lfloor\zeta\rfloor} \sup_{\boldsymbol{s},\boldsymbol{s}'\in\mathcal{D}} \frac{|D^k w(\boldsymbol{s}) - D^k w(\boldsymbol{s}')}{||\boldsymbol{s}-\boldsymbol{s}'||^{\zeta-\lfloor\zeta\rfloor}},$$

where $D^k = \frac{\partial^{k_1+k_2}}{\partial s_1^{k_1}\partial s_2^{k_2}}$, for $k_1, k_2 \in \mathbb{N}$ and $\boldsymbol{s} = (s_1, s_2)'$. Further define the sets

$$\Theta_\zeta = \left\{ w(\boldsymbol{s}) : w(\boldsymbol{s}) = \sum_{r=1}^{R}\sum_{j=1}^{J(r)} K(\boldsymbol{s}, \boldsymbol{s}_j^r, \phi_r)\beta_j^r, R \in \mathbb{N}, \boldsymbol{s}_j^r \in \mathcal{R}^2, \beta_j^r \in \mathcal{R}, ||w||_\zeta < \infty, \right\}$$
$$\Theta_\zeta^n = \{w \in \Theta_\zeta : ||w||_\zeta < n^\alpha, \alpha \in (1/2, 1]\}$$
$$\Theta_{\zeta,c} = \text{Closure under } ||\cdot||_\infty \text{ of } \Theta_\zeta$$
$$\mathcal{B}_{\epsilon,n} = \left\{ w \in \Theta_\zeta^n : \frac{1}{n}\sum_{i=1}^{n} |w(\boldsymbol{s}_i) - w_0(\boldsymbol{s}_i)| < \epsilon, \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon \right\}.$$

Theorem 1. *Let $\mathcal{P}_{w_0,\sigma_0^2}$ denotes the true data generating joint distribution of $\{y_i\}$. Assume*

(a) $\mathcal{D}$ is compact.

(b) $K(\cdot,\cdot,\phi_r)$ is continuous.

Simplifying assumption is merely to ease notation and calculations; all results generalize in a straightforward manner.

(c) $w_0 \in \Theta_{\zeta,c}$, $||w_0||_\zeta < \infty$, *for some* ζ.

Then for any $(w_0, \sigma_0^2) \in \Theta_{\zeta,c} \times \mathcal{R}^+$ *and for any* $\epsilon > 0$,

$$\lim_{n\to\infty} \Pi((w, \sigma^2) \in \mathcal{B}_{\epsilon,n} | y_1, ..., y_n) = 0$$

almost surely under $\mathcal{P}_{w_0,\sigma_0^2}$.

Theorem 1 establishes consistency of estimating the data generating surface w_0 and the true error variance σ_0^2. The proof proceeds along the same line of arguments outlined in Choi and Schervish (2007), Pillai (2008) and is provided in the Appendix.

Proof of Theorem 1:

We begin by stating and proving a lemma that will be useful in the proof of the theorem.

Lemma 1. *Consider a ball of radius* δ *around* (w_0, σ_0^2) *given by*

$$B_\delta(w_0, \sigma_0^2) = \left\{ (w, \sigma^2) : ||w - w_0||_\infty < \delta, \left| \frac{\sigma^2}{\sigma_0^2} - 1 \right| < \delta \right\}.$$

Then $\pi(B_\delta(w_0, \sigma_0^2)) > 0$, *for all* $\delta > 0$.

Proof. Since $w_0 \in \Theta_c$, $\exists w_*(\boldsymbol{s}) = \sum_{r=1}^{R^*} \sum_{j=1}^{J(r)} K(\boldsymbol{s}, \boldsymbol{s}_j^{r*}, \phi_r)\beta_j^{r*}$, s.t. $||w_* - w_0||_\infty < \delta/2$. Note that $K(\cdot, \cdot, \phi_r)$ is a continuous function on a compact set $\mathcal{D}$, implying $K(\cdot, \cdot, \phi_r)$ to be a uniformly continuous function. Thus, $\exists$ M, s.t. $M = \sup_{\boldsymbol{s}\in\mathcal{D}} \max_{r=1,..,R^*; j=1:J(r)} |K(\boldsymbol{s}, \boldsymbol{s}_j^{r*}, \phi_r)|$. Assume further that $\eta = \sum_{r=1}^{R^*} \sum_{j=1}^{J(r)} |\beta_j^{r*}|$. Since K is uniformly continuous, one can choose $\boldsymbol{s}_j^r$'s such that $\sup_{\boldsymbol{s}\in\mathcal{D}} |K(\boldsymbol{s}, \boldsymbol{s}_j^r, \phi_r) - K(\boldsymbol{s}, \boldsymbol{s}_j^{r*}, \phi_r)| < \frac{\delta}{4\eta \sum_{r=1}^{R^*} J(r)}$. Define the set

$$\mathcal{I} = \left\{ \{\beta_j^r\} : |\beta_j^r - \beta_j^{r*}| < \frac{\delta}{4M \sum_{r=1}^{R^*} J(r)} \right\}.$$

Clearly, for the set of all $w(\boldsymbol{s}) = \sum_{r=1}^{R^*} \sum_{j=1}^{J(r)} K(\boldsymbol{s}, \boldsymbol{s}_j^r, \phi_r)\beta_j^r$, with $\boldsymbol{s}_j^r$ is chosen as above and

β_j^r chosen from $\mathcal{I}$, we have

$$\begin{aligned}|w_0(\boldsymbol{s}) - w(\boldsymbol{s})| &\leq |w_0(\boldsymbol{s}) - w_*(\boldsymbol{s})| + |w_*(\boldsymbol{s}) - w(\boldsymbol{s})| \\ &\leq \frac{\delta}{2} + \sum_{r=1}^{R^*}\sum_{j=1}^{J(r)} |K(\boldsymbol{s}, \boldsymbol{s}_j^r, \phi_r)||\beta_j^r - \beta_j^{r*}| + \sum_{r=1}^{R^*}\sum_{j=1}^{J(r)} |\beta_j^{r*}||K(\boldsymbol{s}, \boldsymbol{s}_j^r, \phi_r) - K(\boldsymbol{s}, \boldsymbol{s}_j^{r*}, \phi_r)| \\ &\leq \frac{\delta}{2} + \frac{\delta M \sum_{r=1}^{R^*} J(r)}{4M\sum_{r=1}^{R^*} J(r)} + \frac{\delta \eta \sum_{r=1}^{R^*} J(r)}{4\eta \sum_{r=1}^{R^*} J(r)} = \delta.\end{aligned}$$

Thus $\mathcal{I} \times \left\{\sigma^2 : \left|\frac{\sigma^2}{\sigma_0^2} - 1\right| < \delta\right\} \subseteq B_\delta(w_0, \sigma_0^2)$. Since, the prior on all β_j^r are continuous on the entire real line and the prior on σ^2 is also continuous on $\mathcal{R}^+$, it trivially holds that $\pi(B_\delta(w_0, \sigma_0^2)) \geq \pi\left(\mathcal{I} \times \left\{\sigma^2 : \left|\frac{\sigma^2}{\sigma_0^2} - 1\right| < \delta\right\}\right) > 0$. This concludes the proof of the lemma.

□

We will now proceed with the proof of Theorem 1. Our aim is to check that all conditions of Theorem in Choi and Schervish (2007) are satisfied. Let $H_i = \frac{N\left(y_i|w_0(\boldsymbol{s}_i),\sigma_0^2\right)}{N(y_i|w(\boldsymbol{s}_i),\sigma^2)}$, and $K_i(w, w_0) = \mathrm{E}(H_i)$ and $V_i(w, w_0) = \mathrm{Var}(H_i)$. It is easy to check that (Choi and Schervish (2007))

$$\begin{aligned}K_i(w, w_0) &= \frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2} - \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{1}{2}\frac{(w(\boldsymbol{s}) - w_0(\boldsymbol{s}))^2}{\sigma^2} \\ V_i(w, w_0) &= \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma^2} - 1\right)^2 + \frac{\sigma_0^4}{\sigma^4}(w(\boldsymbol{s}) - w_0(\boldsymbol{s}))^2.\end{aligned}$$

Thus for every $\epsilon > 0$, there exists a $\delta > 0$ such that $(w(\cdot), \sigma^2) \in B_\delta(w_0, \sigma_0^2)$ implies $K_i(w, w_0) < \epsilon, \forall\, i$ and $\sum_{i=1}^\infty \frac{V_i(w,w_0)}{i^2} < \infty$. Thus condition (i) is satisfied. Condition (ii), i.e. the prior positivity has already been proved to be satisfied by Lemma 1.

Finally, the condition of having an exponentially consistent sequence of tests follows along the same line as the proof of Theorem 2 in Choi and Schervish (2007). This concludes the theorem.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, pp. 669–679.

Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis 90*, pp. 1969–1987.

Pillai, N. S. (2008). *Levy random measures: posterior consistency and applications.* Ph.D. Thesis, Duke University.