Modélisation de l'évolution de la présence d'une infection par un modèle bayésien de Markov caché à observations continues. Application à la surveillance de la BVD en France

Aurélien Madouasse[1] & Matthieu Trotreau & Grégoire Kuntz[2] & Jean-Michel Galharret[3]

[1] *Oniris, INRAE, BIOEPAR, Nantes 44300, France, aurelien.madouasse@oniris-nantes.fr*
[2] *Innoval, GDS Bretagne, gregoire.kuntz@innoval.com*
[3] *Oniris, INRAE, StatSC, Nantes 44300, France, jean-michel.galharret@oniris-nantes.fr*

Résumé. Un modèle bayésien de Markov caché a été développé pour estimer les caractéristiques de tests sérologiques et la dynamique d'une infection à partir de données longitudinales de surveillance. L'approche proposée consiste à modéliser les résultats de tests comme un mélange de deux distributions (l'une pour les échantillons séronégatifs et l'autre pour les séropositifs). En l'absence de seuil de détection, une règle de décision basée sur les paramètres du mélange est proposée. Cette approche est comparée par simulations à l'approche classique (binarisation du résultat du test comme observation de la variable latente). Les résultats montrent une bonne estimation des paramètres du mélange et de meilleures performances que le modèle classique dans tous les scénarios épidémiologiques considérés. Le modèle a été appliqué aux données du programme de surveillance du virus de la diarrhée virale bovine (BVD) en Bretagne à partir de résultats trimestriels de deux tests sérologiques sur le lait de tank entre 2014 et 2020.

Mots-clés. Modèle bayésien de Markov caché, mélange de lois.

Abstract. A Bayesian Hidden Markov Model (BHMM) has been developed to estimate test characteristics and infection dynamics from longitudinal surveillance data. The proposed approach involves modeling test results as a mixture of two distributions (one for seronegative samples and the other for seropositive). In the absence of a detection threshold, we propose a decision rule based on the parameters of this mixture. This approach is compared through simulations to the classical approach (binarization of the test result as an observation of the latent variable). The results show an accurate estimation of the mixture parameters and outperforms the classical model across all considered epidemiological scenarios. The model was applied to data from the Bovine Viral Diarrhea (BVD) virus surveillance program in Brittany, using quarterly ELISA test results from bulk tank milk between 2014 and 2020.

Keywords. Hidden Markov Model, Gaussian mixture ...

1 Modèle Gaussien de Markov caché

1.1 Introduction

Le modèle de Markov caché (Hidden Markov Model, HMM), introduit par Baum (1966), est un processus stochastique utilisé pour représenter une variable catégorielle latente qui évolue en temps discret ($t = 1, \ldots, T$) selon une dynamique markovienne : l'état suivant de la chaîne ne dépend que de son état actuel et est indépendant des états précédents. À tout moment, l'état latent Z_t détermine la distribution d'une variable observée Y_t. Dans les programmes de surveillance de maladie, un HMM à deux états latents est considéré et Y_t peut être un résultat de test binaire (par exemple, positif ou négatif) ou bien le résultat d'une mesure de concentration d'anticorps (e.g. tests ELISA). Dans le premier cas, Y_t suit une distribution de Bernoulli avec un paramètre p dépendant de Z_t (see Madouasse et al., 2022). Dans le second cas, Y_t peut être modélisée par une distribution de probabilité.

Un HMM a deux états latents est caractérisé par la prévalence initiale de la maladie $\pi_1 = P(Z_1 = 1)$ et deux probabilités de transition $\tau_1 = P(Z_t = 1|Z_{t-1} = 0)$, $\tau_2 = P(Z_t = 1|Z_{t-1} = 1)$. Ces probabilités sont supposées être indépendantes du temps (i.e., les chaines sont homogènes). On suppose que les résultats de tests sont distribués selon un mélange gaussien de paramètres (μ_0, σ_0) (séronégatifs) et (μ_1, σ_1) (séropositifs).

$$P(Y < y|Z = z) = (1 - z)\Phi_0(y) + z\Phi_1(y),$$

où $z = 0, 1$, $y \in \mathbb{R}$, et Φ_i est la fonction de répartition de la loi normale de paramètres (μ_i, σ_i). La figure 1 résume ce modèle gaussien de Markov caché.

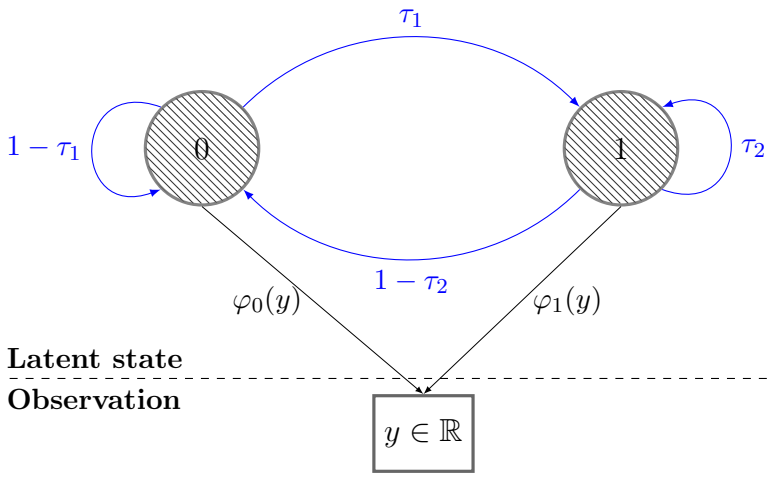

FIGURE 1 – HMM à deux états avec des données observées distribuées selon un mélange gaussien. φ_i correspond à la densité gaussienne de moyenne μ_i et de variance σ_i^2.

1.2 Inférence bayésienne

Soit $\theta = (\pi_1, \tau_1, \tau_2, \mu_0, \sigma_0, \mu_1, \sigma_1)$ l'ensemble des paramètres du modèle. Étant donné une séquence de T observations $Y_{1:T} = Y_1, \ldots, Y_T$, six types d'inférence sont effectués :

Filtering : Calculer pour $t < T$, la probabilité de l'état Z_t conditionnement aux observations $Y_{1:t}$.

Smoothing : Calculer la probabilité de Z_t conditionnellement à toutes les observations $Y_{1:T}$.

Forecasting : Pour $h \in \mathbb{N}$, calculer la probabilité de Z_{t+h} conditionnellement aux observations $Y_{1:t}$.

Evaluation : Calculer la vraisemblance de la séquence $Y_{1:T}$.

Decoding : Déterminer la séquence $Z_{1:T}$ la plus probable étant données les observations $Y_{1:T}$.

Learning : Estimer le paramètre θ sachant $Y_{1:T}$.

L'algorithme forward, l'algorithme backward et l'algorithme de Viterbi sont utilisés pour résoudre ces problèmes (voir par exemple Viterbi (1967); Rabiner (1989)). Parmi ces tâches, l'estimation de θ (*Learning*) peut être réalisée par maximum de vraisemblance, l'algorithme EM ou l'inférence bayésienne (voir par exemple Ghahramani, 2001).

La distribution jointe des observations $Y_{1:T}$ et des états latents $Z_{1:T}$ conditionnement à θ est

$$P(Y_{1:T}, Z_{1:T}|\theta) = P(Z_1)P(Y_1|Z_1, \mu_0, \sigma_0, \mu_1, \sigma_1) \prod_{t=2}^{T} P(Z_t|Z_{t-1})P(Y_t|Z_t, \mu_0, \sigma_0, \mu_1, \sigma_1),$$

où $P(Z_1)$ dépend uniquement de π_1, $P(Z_t|Z_{t-1})$ uniquement de τ_1, τ_2. De plus on a : $P(Y_{1:T}|\theta) = \sum_{Z_{1:T}} P(Z_{1:T}Y_{1:T}|\theta)$ dont on déduit la loi a posteriori de $\theta|Y_{1:T}$.

1.3 Règle de décision basée sur la donnée continue

Soit Z l'état latent correspondant à un résultat de test y. On adopte la règle de décision suivante :

$$Z = \begin{cases} 1 \text{ lorsque } \Phi_0(y) > 1 - \Phi_1(y), \\ 0 \text{ sinon.} \end{cases} \tag{1}$$

Un calcul simple montre que si $s = \frac{\sigma_0\mu_1 + \sigma_1\mu_0}{\sigma_0 + \sigma_1}$, alors on a $\Phi_0(s) = 1 - \Phi_1(s)$. Donc cette règle de décision est équivalente à déclarer un résultat de test y positif lorsque $y > s$. Par ailleurs, la sensibilité (i.e. le taux de vrais positifs) et la spécificité (i.e. le taux de faux négatifs) associées à ce seuil sont égales à $\Phi_0(s) = 1 - \Phi_1(s)$. La loi a posteriori de s peut être déduite de celles des paramètres du mélange et servir de référence en l'absence de seuil de détection.

Scenario	τ_1	τ_2	μ_0	μ_1	Seuil (h)	Se	Sp
1	0.1	0.85	25	75	50	0.95	0.95
2	0.1	0.85	25	50	37.5	0.79	0.79
3	0.1	0.85	30	50	37.5	0.79	0.69
4	0.1	0.85	25	45	37.5	0.69	0.79
5	0.2	0.7	25	75	50	0.95	0.95
6	0.4	0.4	25	75	50	0.95	0.95

TABLE 1 – Valeurs des paramètres utilisés pour les simulations.

2 Résultats

2.1 Données simulées

Pour les simulations les écarts types des deux distributions du mélange ont été fixés à $\sigma_0 = \sigma_1 = 15$. La prévalence initiale de la maladie a été fixée à $\pi_1 = 0.4$. La table 1 fournit les valeurs des autres paramètres utilisés pour les simulations. Le choix de τ_2 a été réalisé tel que $\tau_2 = 1 - \frac{\tau_1(1-\pi_1)}{\pi_1}$ qui est standard en épidémiologie car il correspond à une situation où l'infection est équilibrée (i.e. le nombre d'individus éliminant l'infection est équilibré par le nombre d'individus l'acquérant). La valeur du seuil h permet de transformer les données en résultats binaires (i.e. $Z = 0$ pour $y < h$) et ainsi utiliser le modèle proposé par Madouasse et al. (2022). La sensibilité et la spécificité de ce seuil h sont respectivement égales à $1 - \Phi_1(h)$ et $\Phi_0(h)$. Les scénarios 1 à 4 correspondent à un chevauchement de plus en plus important entre les deux composantes du mélange. Les scénarios 5 et 6 correspondent à une augmentation de la probabilité de devenir infecté (respectivement $\tau_1 = 0.2$ et $\tau_1 = 0.4$). Le graphe 2 fournit les résultats de la comparaison des deux modèles sur $B = 1500$ échantillons de $n = 100$ individus suivis sur $t = 5$ périodes. Les performances des modèles sont comparables en terme de biais sur les quatre premiers scénarios. Par contre le biais des estimations avec le modèle à réponses binaires est très mauvais pour les paramètres τ_1, τ_2 dans les scénarios 5 et 6 (haut du graphe), alors que celui des estimations obtenues avec le modèle à observations continues reste équivalent aux autres scénarios (1 à 4). Par ailleurs, les écarts types des distributions a posteriori des paramètres du modèle à données continues sont dans tous les scénarios inférieurs à ceux du modèle à données discrètes (bas du graphe).

2.2 Données réelles

Les données sont issues d'un plan de surveillance de la BVD (Diarrhée Bovine Virale) réalisé sur des troupeaux entre 2014 et 2020 dans la région Bretagne à l'aide de deux tests de détection d'anticorps appelés BTMab1 et BTMab2. Les troupeaux ont été testés tous les trois mois, les deux tests étant utilisés en alternance. Étant donné que les tests mesurent la présence d'anticorps dans le tank de lait (sample-pooling), l'état de latence considéré dans cette étude a été défini comme *la présence d'au moins un animal séropositif dans le troupeau.* Le seuil utilisé pour les deux tests avait été fixé à 35 pour toute la durée de l'étude. Les paramètres du modèle ont été estimés chaque année (4 temps de mesure), l'année $(t - 1)$

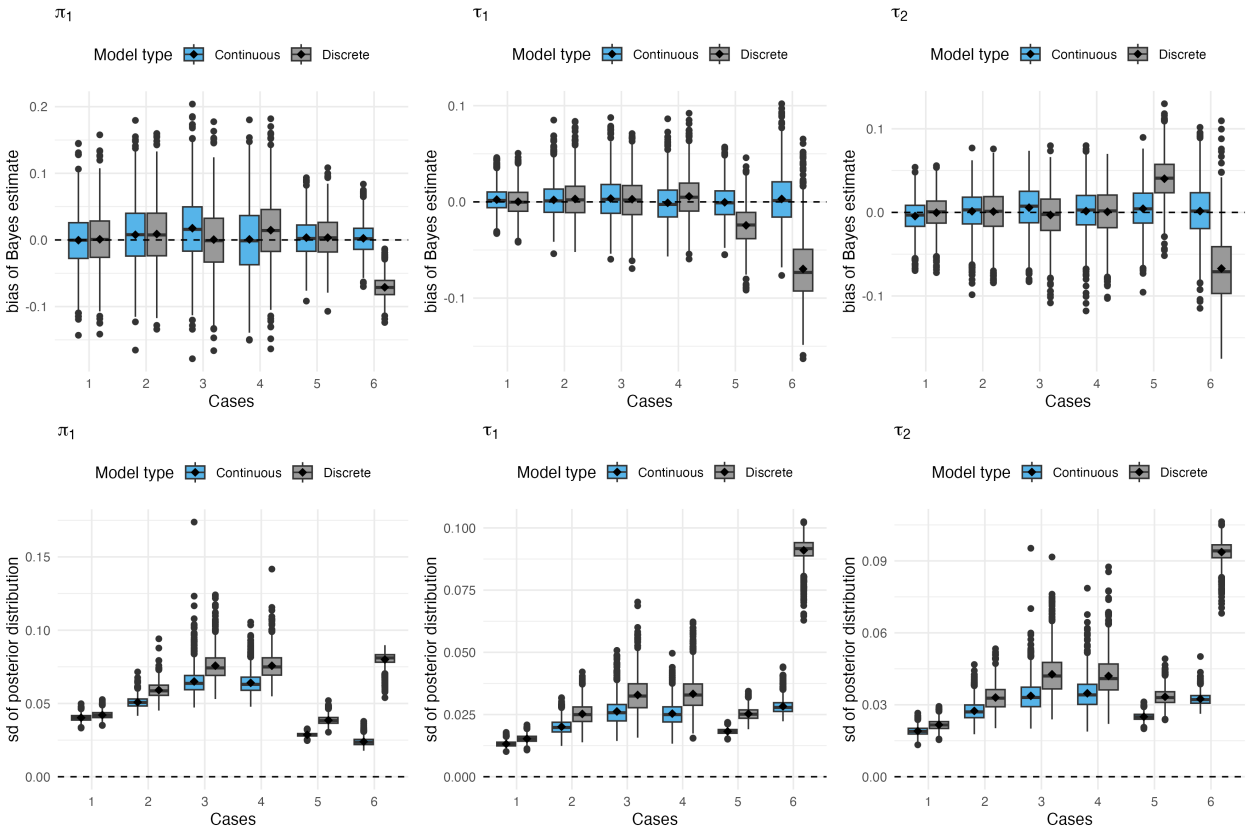

FIGURE 2 – Boxplots du biais de l'estimateur de Bayes des paramètres de dynamique π_1, τ_1, τ_2 (haut) et des écarts types de leur distribution a posteriori (bas) en fonction du scénario considéré et du modèle Continuous : modèle à observations continues) et Discrete : modèle à observations binaires).

servant d'historique pour l'année t. La figure 3 fournit les estimations des caractéristiques des deux tests durant la période d'étude. On constate que la moyenne de la composante des troupeaux infectés du test BTMab1 est proche voir en dessous du seuil de détection dans la plupart des départements. Ainsi, le fait d'utiliser un seuil constant alors qu'il s'agit de données de pooling et qu'on ne tient donc en compte de la prévalence intra-troupeaux de l'infection va détériorer les performances des tests en terme de sensibilité. En revanche, utiliser la règle de décision (1) associée aux données continues aurait permis de garantir une sensibilité minimale donnée par la table 2.

Références

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01) :9–42.

Madouasse, A., Mercat, M., van Roon, A., Graham, D., Guelbenzu, M., Santman Berends, I., van Schaik, G., Nielen, M., Frossling, J., Agren, E., Humphry, R., Eze, J., Gunn, G.,

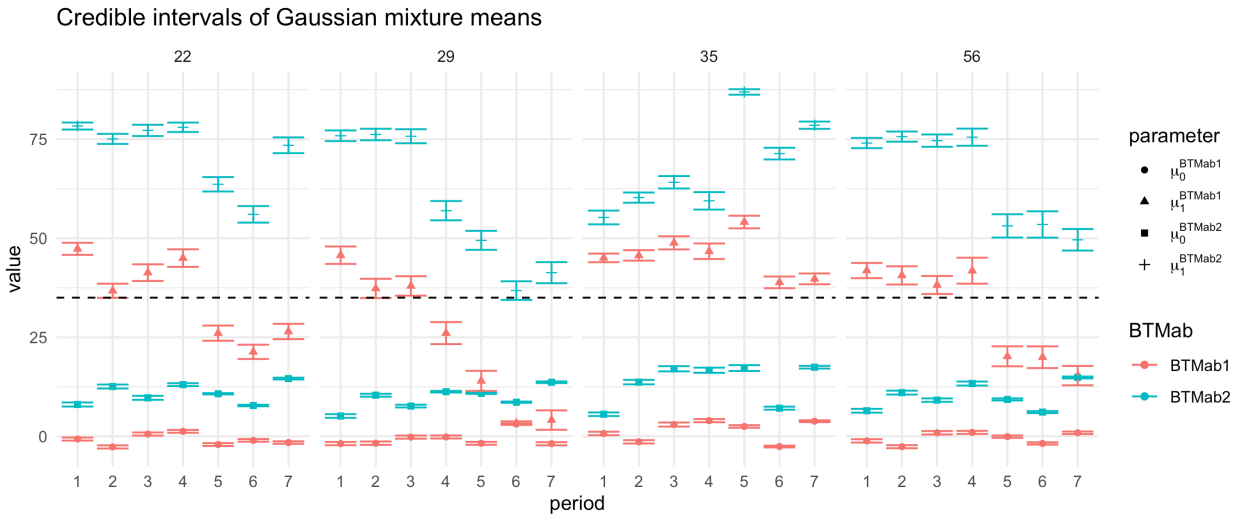

FIGURE 3 – Evolution des paramètres des tests sérologiques au cours de la période d'étude en fonction du département breton.

	période	Seuil	$S_e = S_p$
1	1	17.01	0.90
2	2	10.78	0.85
3	3	13.93	0.85
4	4	12.54	0.85
5	5	4.22	0.75
6	6	3.54	0.71
7	7	4.98	0.76

TABLE 2 – Evolution du seuil de décision de la règle (1) pour le département 22 entre 2014 et 2020. S_e et S_p désignent respectivement la sensibilité et la spécificité du seuil issue de la règle de décision associée.

Henry, M. K., Gethmann, J., More, S. J., Toft, N., and Fourichon, C. (2022). A modelling framework for the prediction of the herd-level probability of infection from longitudinal data. *Peer Community Journal*, 2.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2) :260–269.