

1 **Short Note on Two Output-dependent Hidden**
2 **Markov Models**

3 Jing-Hao Xue^{a,*}, D. Michael Titterington^a

4 ^a*Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK*

5 **Abstract**

6 The purpose of this note is to study the assumption of “mutual information inde-
7 pendence”, which is used by Zhou (2005) for deriving an output-dependent hidden
8 Markov model, the so-called discriminative HMM (D-HMM), in the context of deter-
9 mining a stochastic optimal sequence of hidden states. The assumption is extended
10 to derive its generative counterpart, the G-HMM. In addition, state-dependent rep-
11 resentations for two output-dependent HMMs, namely HMMSDO (Li, 2005) and
12 D-HMM, are presented.

13 *Key words:* Discriminative models; Generative models; Mutual information
14 independence; Output-dependent hidden Markov model

* Corresponding author. Tel.: +44 141 330 2474; fax: +44 141 330 4814.

Email addresses: jinghao@stats.gla.ac.uk (Jing-Hao Xue),

mike@stats.gla.ac.uk (D. Michael Titterington).

1 Introduction

2 Generative models like normal-based discriminant analysis and discriminative
3 models like logistic regression are comprehensively investigated and compared
4 in the machine learning literature (Rubinstein and Hastie, 1997; Ng and Jor-
5 dan, 2001). Amongst the latent (hidden) variable models for structured data
6 such as time series, hidden Markov models (HMMs) for discrete-valued hidden
7 states and state-space models (SSMs) for continuous-valued hidden states are
8 widely used.

9 Traditionally, an HMM is generative because it models a distribution $P(O_1^n|S_1^n)$,
10 the data generation process (DGP) of the observed output sequence, $O_1^n =$
11 o_1, \dots, o_n , given the hidden state sequence, $S_1^n = s_1, \dots, s_n$, and thus $P(O_1^n|S_1^n)$,
12 a state-dependent term, is included in the criterion for determining a sto-
13 chastic optimal sequence of hidden states. Recently, Zhou (2005) proposes
14 a discriminative hidden Markov model (D-HMM), which includes output-
15 dependent terms $P(s_t|O_1^n), t = 1, \dots, n$, in the criterion, based on an assump-
16 tion of “mutual information independence”. Meanwhile, Li (2005) presents
17 the so-called “hidden Markov models with states depending on observations”
18 (HMMSDO), which assumes that the current state s_t depends not only on
19 the last state s_{t-1} but also on the last output o_{t-1} , so that output-dependent
20 terms $P(s_t|s_{t-1}, o_{t-1})$ are included in the criterion.

21 Both the D-HMM and HMMSDO show superior performance in determining
22 the optimal state sequence for certain applications. Zhou (2005) shows that the
23 D-HMM outperforms the corresponding generative hidden Markov model (G-
24 HMM) for part-of-speech tagging and phrase chunking; Li (2005) shows that

1 HMMSDO outperforms the standard HMM for prediction of protein secondary
 2 structures when the training set is large enough.

3 In this note, we shall study the assumption of “mutual information indepen-
 4 dence” that is used for deriving the D-HMM (Zhou, 2005) in the context of
 5 determining an optimal state sequence, and then extend it to derive its gener-
 6 ative counterpart, the G-HMM. In addition, state-dependent representations
 7 for these two output-dependent HMMs will be presented.

8 **2 Generative HMM**

9 Following the notation used by Zhou (2005), the definition of the optimal
 10 hidden state sequence S_1^n based on the observed output sequence O_1^n is that
 11 of the maximum a posteriori (MAP) estimator S^* of S_1^n :

$$S^* = \operatorname{argmax}_{S_1^n} \{ \log P(S_1^n | O_1^n) \} . \quad (1)$$

12 The G-HMM rewrites the criterion (1) through applying Bayes’ theorem and
 13 ignoring the item determined purely by O_1^n as

$$S^* = \operatorname{argmax}_{S_1^n} \{ \log P(S_1^n) + \log P(O_1^n | S_1^n) \} ,$$

14 which is further factorised as

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \log \left(P(o_1 | S_1^n) \prod_{k=2}^n P(o_k | O_1^{k-1}, S_1^n) \right) \right\} .$$

15 In order to make this formulation tractable, an assumption that O_1^n is condi-
 16 tionally independent given S_1^n is in general introduced as, for all $k \in \{2, \dots, n\}$,

$$P(o_k | O_1^{k-1}, S_1^n) = P(o_k | S_1^n) , \quad (2)$$

1 and thus based on such a conditional independence assumption, the MAP
 2 estimator for the G-HMM is simplified to

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \sum_{i=1}^n \log P(o_i | S_1^n) \right\} . \quad (3)$$

3 The G-HMM is regarded as being generative because it directly models the
 4 DGP $P(o_i | S_1^n)$ of the observed o_i from the hidden S_1^n .

5 In practice, as for the standard HMM, the assumption (2) is further simplified
 6 to

$$P(o_k | O_1^{k-1}, S_1^n) = P(o_k | S_1^n) = P(o_k | s_k) , \quad (4)$$

7 and thus the MAP estimator of the standard HMM is

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \sum_{i=1}^n \log P(o_i | s_i) \right\} . \quad (5)$$

8 3 Discriminative HMM from Mutual Information Independence

9 The D-HMM rewrites the criterion (1) through applying Bayes' theorem, but
 10 not ignoring the item determined purely by O_1^n , as

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} \right\} .$$

11 To make this formulation tractable, an assumption that the mutual informa-
 12 tion ($MI(S_1^n, O_1^n) = \log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)}$) between S_1^n and O_1^n is independent with
 13 respect to each hidden s_i was introduced by Zhou (2005) as

$$MI(S_1^n, O_1^n) = \sum_{i=1}^n MI(s_i, O_1^n) , \quad (6)$$

14 or, in more detail,

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^n \log \frac{P(s_i, O_1^n)}{P(s_i)P(O_1^n)} = \sum_{i=1}^n \log \frac{P(s_i | O_1^n)}{P(s_i)} . \quad (7)$$

1 Based on such a representation, the MAP estimator for the D-HMM is sim-
 2 plified as (Zhou, 2005)

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \sum_{i=1}^n \log P(s_i|O_1^n) - \sum_{i=1}^n \log P(s_i) \right\}. \quad (8)$$

3 The D-HMM is regarded as being discriminative because the criterion (8)
 4 includes directly the discriminative process $P(s_i|O_1^n)$, representing an output-
 5 dependence of a hidden state s_i on all the observed outputs O_1^n .

6 We shall make four observations about the D-HMM.

7 First, it is noted that the criterion (8) is simultaneously to maximise the max-
 8 imum posterior marginal (MPM) estimator $\sum_{i=1}^n \log P(s_i|O_1^n)$ of $\log P(S_1^n|O_1^n)$
 9 and to maximise the distance between the state transition model $\log P(S_1^n)$
 10 and its independent-based counterpart $\sum_{i=1}^n \log P(s_i)$.

11 Second, in order to satisfy the assumption (7) underlying the D-HMM, it is
 12 required that

$$\prod_{k=2}^n \frac{P(s_k|S_1^{k-1}, O_1^n)}{P(s_k|S_1^{k-1})} = \prod_{k=2}^n \frac{P(s_k|O_1^n)}{P(s_k)}.$$

13 Since this is valid for any value of s_k , it follows that, for all $k \in \{2, \dots, n\}$,

$$\frac{P(s_k|S_1^{k-1}, O_1^n)}{P(s_k|S_1^{k-1})} = \frac{P(s_k|O_1^n)}{P(s_k)}. \quad (9)$$

14 Third, the assumption (7) can be rewritten as

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^n \log \frac{P(s_i, O_1^n)}{P(s_i)P(O_1^n)} = \sum_{i=1}^n \log \frac{P(O_1^n|s_i)}{P(O_1^n)}. \quad (10)$$

15 Based on such a representation, the MAP estimator (8) for the D-HMM can
 16 be rewritten, with the term $\sum_{i=1}^n \log P(O_1^n)$ determined purely by O_1^n being
 17 ignored, as

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \sum_{i=1}^n \log P(O_1^n|s_i) \right\}. \quad (11)$$

1 Therefore, the D-HMM can also be represented as being generative because
 2 the criterion (11) includes a generative-like process $P(O_1^n|s_i)$, representing a
 3 state-dependence of all the observed outputs O_1^n on a hidden state s_i .

4 Fourth, it can be seen that, when the assumption (6) of mutual information
 5 independence develops from independence between pairs (s_i, O_1^n) into that be-
 6 tween local pairs (s_i, o_i) such that $MI(S_1^n, O_1^n) = \sum_{i=1}^n MI(s_i, o_i)$, the criteria
 7 (11) and (8) degenerate into the criterion (5), indicating that the D-HMM
 8 degenerates into the standard HMM.

9 4 Generative HMM from Mutual Information Independence

10 Furthermore, similarly to the assumption (6) proposed by Zhou (2005), an
 11 assumption that mutual information between S_1^n and O_1^n is independent with
 12 respect to each observed o_i can be introduced here as

$$MI(S_1^n, O_1^n) = \sum_{i=1}^n MI(S_1^n, o_i) , \quad (12)$$

13 or, in more detail,

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^n \log \frac{P(S_1^n, o_i)}{P(S_1^n)P(o_i)} = \sum_{i=1}^n \log \frac{P(o_i|S_1^n)}{P(o_i)} . \quad (13)$$

14 Based on such a representation, we can obtain another generative model and
 15 its MAP estimator, with the term $\sum_{i=1}^n \log P(o_i)$ determined purely by O_1^n
 16 being ignored, as

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \sum_{i=1}^n \log P(o_i|S_1^n) \right\} . \quad (14)$$

17 This estimator is in fact the estimator (3) of the G-HMM, *i.e.*, the G-HMM
 18 can be derived under the assumption (12), a type of mutual information in-

1 dependence.

2 Similarly, we shall make three observations about this G-HMM, which is de-
3 rived from mutual information independence.

4 First, in order to satisfy the assumption (13) of the G-HMM, it is required
5 that, for all $k \in \{2, \dots, n\}$,

$$\frac{P(o_k|O_1^{k-1}, S_1^n)}{P(o_k|O_1^{k-1})} = \frac{P(o_k|S_1^n)}{P(o_k)}. \quad (15)$$

6 Therefore, under the MAP criterion (1), the conditions (15) and (2) have the
7 same effect on determining the optimal hidden S_1^n .

8 Second, the assumption (13) can be rewritten as

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^n \log \frac{P(S_1^n, o_i)}{P(S_1^n)P(o_i)} = \sum_{i=1}^n \log \frac{P(S_1^n|o_i)}{P(S_1^n)}. \quad (16)$$

9 Based on such a representation, the MAP estimator (14) for the G-HMM can
10 be rewritten, with the terms related to $\log P(S_1^n)$ being combined, as

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ (1-n) \log P(S_1^n) + \sum_{i=1}^n \log P(S_1^n|o_i) \right\}. \quad (17)$$

11 Therefore, in this sense, the G-HMM can also be represented as being dis-
12 criminative because the criterion (17) includes a discriminative-like process
13 $P(S_1^n|o_i)$, representing an output-dependence of all the hidden states S_1^n on
14 an observed output o_i .

15 Third, it can be seen that, when the assumption (12) of mutual information
16 independence develops from independence between pairs (S_1^n, o_i) into that be-
17 tween local pairs (s_i, o_i) such that $MI(S_1^n, O_1^n) = \sum_{i=1}^n MI(s_i, o_i)$, the criteria
18 (17) and (14) degenerate into the criterion (5), indicating that the G-HMM
19 degenerates into the standard HMM.

1 **5 Equivalence between G-HMM and D-HMM**

2 Once we assume a fully independent mutual information between any state-
 3 output combination (s_i, o_j) as

$$MI(S_1^n, O_1^n) = \sum_{i=1}^n \sum_{j=1}^n MI(s_i, o_j) , \quad (18)$$

4 or, in more detail,

$$\begin{aligned} \log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} &= \sum_{i=1}^n \sum_{j=1}^n \log \frac{P(s_i, o_j)}{P(s_i)P(o_j)} \\ &= \sum_{i=1}^n \sum_{j=1}^n \log \frac{P(o_j|s_i)}{P(o_j)} = \sum_{i=1}^n \sum_{j=1}^n \log \frac{P(s_i|o_j)}{P(s_i)} , \end{aligned} \quad (19)$$

5 this assumption results in two criteria, one generative and the other discrimi-
 6 native, with the MAP estimators as

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \sum_{i=1}^n \sum_{j=1}^n \log P(o_j|s_i) \right\} , \quad (20)$$

7

$$S^* = \operatorname{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \sum_{i=1}^n \sum_{j=1}^n \log P(s_i|o_j) - \sum_{i=1}^n \{n \log P(s_i)\} \right\} , \quad (21)$$

8 respectively. These two criteria are equivalent.

9 In the context of determining an optimal sequence of hidden states, apart
 10 from the equivalence above, up to now, we find two occurrences of equivalence
 11 between a discriminative representation of the MAP criterion and its genera-
 12 tive counterpart: one is for the D-HMM between the criteria (8) and (11), the
 13 other is for the G-HMM between the criteria (17) and (14).

14 We shall further illustrate such equivalence with two simple but related HMMs:
 15 one is a generative-like state-dependent model, which assumes that the current
 16 output o_t depends not only on the current state s_t but also on the last state
 17 s_{t-1} ; the other is a discriminative-like output-dependent model, the so-called

1 HMMSDO (Li, 2005), which assumes that the current state s_t depends not
 2 only on the last state s_{t-1} but also on the last output o_{t-1} .

3 The joint distribution of the first generative-like state-dependent model is

$$P(S_1^n, O_1^n) = P(s_1)P(o_1|s_1) \prod_{i=2}^n P(s_i|s_{i-1})P(o_i|s_i, s_{i-1}) . \quad (22)$$

4 This distribution can be rewritten as

$$\begin{aligned} P(S_1^n, O_1^n) &= P(o_1, s_1) \prod_{i=2}^n P(s_i, o_i|s_{i-1}) \\ &= P(o_1)P(s_1|o_1) \prod_{i=2}^n P(o_i|s_{i-1})P(s_i|s_{i-1}, o_i) , \end{aligned} \quad (23)$$

5 which leads to a discriminative-like output-dependent part $P(s_i|s_{i-1}, o_i)$ in the
 6 distribution. In fact, the difference between the probabilistic directed acyclic
 7 graphs (DAGs) corresponding to the joint distributions (22) and (23) is only
 8 in that directions of edges from s_i to o_i are reversed.

9 Similarly, the joint distribution of the discriminative-like output-dependent
 10 HMMSDO, with $P(s_i|s_{i-1}, o_{i-1})$ included, is (Li, 2005)

$$P(S_1^n, O_1^n) = P(s_1)P(o_1|s_1) \prod_{i=2}^n P(s_i|s_{i-1}, o_{i-1})P(o_i|s_i) . \quad (24)$$

11 This distribution can be rewritten as

$$\begin{aligned} P(S_1^n, O_1^n) &= P(s_1)P(o_n|s_n) \prod_{i=2}^n P(s_i, o_{i-1}|s_{i-1}) \\ &= P(s_1)P(o_n|s_n) \prod_{i=2}^n P(s_i|s_{i-1})P(o_{i-1}|s_i, s_{i-1}) , \end{aligned} \quad (25)$$

12 which leads to a no longer discriminative-like output-dependence in the dis-
 13 tribution. In fact, the difference between the DAGs corresponding to the joint
 14 distributions (24) and (25) is only in that directions of edges from s_i to o_{i-1}
 15 are reversed. In practice, whether or not $P(o_{i-1}|s_i, s_{i-1})$ is reasonable needs
 16 to be justified, because it means that the current output depends on the next

1 state.

2 **6 Summary**

3 This note has suggested that the mutual information assumption (12) resulted
4 in the G-HMM, while another mutual information assumption (6) resulted in
5 the D-HMM. However, in practice, whether or not the assumptions are reason-
6 able and how the corresponding HMMs perform can be data-dependent; re-
7 search efforts to explore an adaptive switching between or combination of these
8 two models may be worthwhile. Meanwhile, this note has suggested that the
9 so-called output-dependent HMMs could be represented in a state-dependent
10 manner, and vice versa, essentially by application of Bayes' theorem.

11 **References**

- 12 Li, Y., 2005. Hidden Markov models with states depending on observations.
13 *Pattern Recognition Letters* 26 (7), 977–984.
- 14 Ng, A. Y., Jordan, M. I., 2001. On discriminative vs. generative classifiers: a
15 comparison of logistic regression and naïve bayes. In: NIPS. pp. 841–848.
- 16 Rubinstein, Y. D., Hastie, T., 1997. Discriminative vs. informative learning.
17 In: KDD. pp. 49–53.
- 18 Zhou, G. D., 2005. Direct modelling of output context dependence in discrim-
19 inative hidden Markov model. *Pattern Recognition Letters* 26 (5), 545–553.