THE VARIATIONAL EM ALGORITHM FOR ON-LINE IDENTIFICATION OF EXTENDED AR MODELS

Václav Šmídl

UTIA, Academy of Sciences, Czech Republic
smidl@utia.cas.cz

Anthony Quinn

Trinity College Dublin, Ireland
aquinn@tcd.ie

ABSTRACT

The AutoRegressive (AR) model is extended to cope with a wide class of possible transformations and degradations. The Variational Bayes (VB) procedure is used to restore conjugacy. The resulting Bayesian recursive identification procedure has many of the desirable computational properties of the classical RLS procedure. During each time-step, an iterative Variational EM (VEM) procedure is required to obtain the necessary moments. The procedure is used to reconstruct an outlier-corrupted AR process and a noisy speech segment. The VB scheme appears to offer improved performance over the related Quasi-Bayes (QB) scheme in the case of time-variant component weights.

1. INTRODUCTION

The AutoRegressive (AR) model has been important in many contexts in DSP, notably in all-pole modelling for speech [1]. Its attraction is the existence of fast recursive algorithms which allow on-line estimation and prediction. Bayesian on-line identification is preserved for the Extended AR (EAR) model [2], while, recently, the Quasi-Bayes (QB) approximation was used for on-line identification of a much richer class, namely the Mixture-based Extension of the AR model (MEAR) [3].

Recently, the on-line Variational Bayes (VB) method was proposed as a general estimation paradigm [4], and applied to non-regressive mixture models. In this paper, VB leads to a closed-form posterior distribution for the MEAR model, and an associated on-line Variational EM (VEM) identification scheme. A wide application context is suggested.

2. EXTENDING THE AUTOREGRESSIVE (AR) MODEL

Consider an r-dimensional observation process, $\boldsymbol{d}_t \in \Re^{r\times 1}$, $t = 1, 2, \ldots$. The data history will be denoted $D_t = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_t]$. The Extended AR (i.e. EAR) model [2, 3] regresses a *known* transformation, $\boldsymbol{g}_y$, of the current observation, $\boldsymbol{d}_t$, onto a set, $\boldsymbol{g}_x$, of *known* transformations of past observations. Specifically, we define internal variables, $\boldsymbol{x}_t \in \Re^{p\times 1}$ and $\boldsymbol{y}_t \in \Re^{r\times 1}$:

$$\boldsymbol{x}_t = \boldsymbol{g}_x\left(D_{t-1}, W_t\right), \quad (1)$$

$$\boldsymbol{y}_t = \boldsymbol{g}_y\left(D_t, W_t\right), \quad (2)$$

where W_t represents a possible known external (i.e. exogenous) variable, and $\boldsymbol{g}_y$ is required to be an invertible, r-to-r mapping from $\boldsymbol{d}_t$ to $\boldsymbol{y}_t$ with non-zero Jacobian, $J_t = J_t\left(\boldsymbol{d}_t\right)$. Then, the EAR model declares that

$$\boldsymbol{y}_t = -A\boldsymbol{x}_t + \boldsymbol{e}_t, \quad (3)$$

$$f\left(\boldsymbol{e}_t|\Omega\right) = \mathcal{N}\left(\boldsymbol{0}_r, \Omega^{-1}\right), \quad (4)$$

where $A \in \Re^{r\times p}$ is a matrix of unknown linear coefficients, and $\boldsymbol{e}_t \in \Re^{r\times 1}$ is a zero-mean Gaussian white noise process (innovations) with unknown precision matrix, $\Omega \in \Re^{r\times r}$. In the sequel, we denote $\boldsymbol{z}_t = [\boldsymbol{y}_t', \boldsymbol{x}_t']'$ and $\boldsymbol{g} = \left[\boldsymbol{g}_y', \boldsymbol{g}_x'\right]'$, and suppress the possible W_t-dependence (1) in the notation.

A mixture-based extension of the EAR model (i.e. the MEAR model) has been proposed [3] in order to allow uncertainty in respect of $\boldsymbol{g}$, via a finite set, $G = [\boldsymbol{g}_1, \ldots, \ \boldsymbol{g}_c]$, of possible cases:

$$f\left(\boldsymbol{d}_t|A, \Omega, \boldsymbol{\alpha}, D_{t-1}, G\right) = \sum_{i=1}^{c} \alpha_i f\left(\boldsymbol{d}_t|A, \Omega, \boldsymbol{x}_{i,t}\right), \quad (5)$$

where

$$f\left(\boldsymbol{d}_t|A, \Omega, \boldsymbol{x}_{i,t}\right) = \left|J_t\left(\boldsymbol{d}_t\right)\right| \mathcal{N}\left(-A\boldsymbol{x}_{i,t}, \Omega^{-1}\right) \quad (6)$$

is the observation model for the ith EAR component, derived using (3,4,1). $\boldsymbol{\alpha} = \{\alpha_i, i = 1, \ldots, c\}$ are the (for now) stationary component weights and $\boldsymbol{x}_{i,t}$ denotes the regression vector (1) for the ith candidate filter, $\boldsymbol{g}_{i,x}$. (5) may be viewed as the marginal of the following switching model which has been augmented by a hidden, uncorrelated label process, $\boldsymbol{l}_t$, indicating the active component at each time, t:

$$f\left(\boldsymbol{d}_t, \boldsymbol{l}_t|A, \Omega, \boldsymbol{\alpha}, D_{t-1}\right) = \prod_{j=1}^{c} \left[f\left(\boldsymbol{d}_t|A, \Omega, \boldsymbol{x}_{i,t}\right) f\left(\boldsymbol{l}_t|\boldsymbol{\alpha}\right)\right]^{l_{j,t}}. \quad (7)$$

Here, $\boldsymbol{l}_t = [l_{1,t}, \ldots, l_{c,t}]'$, being one of the c-dimensional elementary vectors, $\boldsymbol{\delta}_c(i)$, $i = 1, \ldots, c$. Its distribution is assumed multinomial [2]: $f(\boldsymbol{l}_t|\boldsymbol{\alpha}) = Mu(\boldsymbol{\alpha}) = \prod_{j=1}^{c} \alpha_j^{l_{j,t}}$.

3. BAYESIAN RECURSIVE INFERENCE

The EAR model (3,4) is the broadest class for which RLS-style [5] on-line estimation is feasible. The equivalent Bayesian perspective is to evaluate the posterior inference, $f(A, \Omega|D_t)$, $\forall t$, exploiting the key fact that the *Normal-Wishart* ($\mathcal{NW}$) distribution [2]) is *conjugate* to the EAR model (3). It is updated as follows:

$$\mathcal{NW}(A, \Omega|V_t, \nu_t) \propto \propto f(\boldsymbol{d}_t|A, \Omega, D_{t-1})\, \mathcal{NW}(A, \Omega|V_{t-1}, \nu_{t-1}), \quad (8)$$

with the first term on the right-hand side being given by (3,4). In (8), V_t and ν_t are the *sufficient statistics*, with updates as follows, $t > q$:

$$V_t = V_{t-1} + \boldsymbol{z}_t \boldsymbol{z}_t', \quad \nu_t = \nu_{t-1} + 1. \quad (9)$$

The first posterior moments (means) of (8), $\mathsf{E}(A, \Omega|D_t)$, correspond to the classical solution of the normal equations via the covariance method [1]:

$$\widehat{A}_t = V'_{\boldsymbol{ad},t} V^{-1}_{\boldsymbol{aa},t}, \quad \widehat{\Omega}_t = \frac{1}{\nu_t - p + r + 1} \Lambda_t^{-1}, \quad (10)$$

where

$$V_t = \begin{bmatrix} V_{\boldsymbol{dd},t} & V'_{\boldsymbol{ad},t} \\ V_{\boldsymbol{ad},t} & V_{\boldsymbol{aa},t} \end{bmatrix}, \quad \Lambda_t = V_{\boldsymbol{dd},t} - V'_{\boldsymbol{ad},t} V^{-1}_{\boldsymbol{aa},t} V_{\boldsymbol{ad},t}. \quad (11)$$

Here, $V_{\boldsymbol{dd}}$ is the $r \times r$ upper-left sub-block of matrix V_t. Bayesian conjugacy is lost in the MEAR observation model (5). The Bayesian $\mathcal{NW}$ data-update (8) in terms of the MEAR observation model (7) is:

$$f(A, \Omega, \boldsymbol{\alpha}, \boldsymbol{l}_t|D_t) \propto \propto f(\boldsymbol{d}_t, \boldsymbol{l}_t|A, \Omega, \boldsymbol{\alpha}, D_{t-1})\, f(A, \Omega, \boldsymbol{\alpha}|D_{t-1}). \quad (12)$$

The auxiliary random variable, $\boldsymbol{l}_t$, is generated during the data-updating (12), violating the invariance property upon which conjugacy depends.

4. VB-CONJUGACY FOR THE MEAR MODEL

Consider the approximate factorization of (12) into a product of independent terms:

$$f(A, \Omega, \boldsymbol{\alpha}, \boldsymbol{l}_t|D_t) \approx \tilde{f}(A, \Omega, \boldsymbol{\alpha}|D_t)\, \tilde{f}(\boldsymbol{l}_t|D_t). \quad (13)$$

This ensures that a marginal is available for the next update (12), without the need to marginalize over $\boldsymbol{l}_t$. We employ the *Variational Bayes (VB)* procedure to identify the optimal such approximate factorization, being the one which minimizes the Kullback-Leibler Divergence (KLD) [4] of (13) to the true posterior (12). These VB marginals are found to be:

$$\tilde{f}(A, \Omega|D_t) = \mathcal{NW}(V_t, \nu_t), \quad (14)$$

$$\tilde{f}(\boldsymbol{\alpha}|D_t) = \mathcal{D}i(\boldsymbol{\beta}_t), \quad (15)$$

$$\tilde{f}(\boldsymbol{l}_t|D_t) = Mu(\boldsymbol{w}_t), \quad (16)$$

with parameters

$$V_t = V_{t-1} + \sum_{j=1}^{c} w_{j,t} \boldsymbol{z}_{j,t} \boldsymbol{z}'_{j,t}, \quad (17)$$

$$\nu_t = \nu_{t-1} + 1, \quad (18)$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + w_t, \quad (19)$$

$$w_{j,t} \propto |J_{j,t}| \exp\Big[-\frac{1}{2} \boldsymbol{z}'_{j,t} \left[-I_p, \widehat{A}_t\right]' \widehat{\Omega}_t \left[-I_p, \widehat{A}_t\right] \boldsymbol{z}_{j,t} - \frac{1}{2} p\, \boldsymbol{z}'_{j,t} V^{-1}_{\boldsymbol{aa},t} \boldsymbol{z}_{j,t} + \widehat{\ln \alpha_{j\,t}}\Big], \quad (20)$$

using (10), and $\widehat{\ln \alpha_j} = \psi(\beta_i) - \psi\left(\sum_{j=1}^{c} \beta_j\right)$, where $\psi(\cdot)$ denotes the digamma function. The mean value of (16) is $\widehat{\boldsymbol{l}}_t = [w_{1,t}, \ldots, w_{c,t}]'$.

Crucially, then, the VB-marginal has conflated the exact marginal into a single $\mathcal{NW}$ component (14), restoring the necessary functional invariance for the MEAR parameter distribution at each time-step. This will be known as *VB-conjugacy*.

The VB scheme (17)–(20) involves cross-coupling between the VB-moments, i.e. $\widehat{A}$, $\widehat{\Omega}$, the $\widehat{\ln \alpha_j}$ and $\boldsymbol{w}_t$. Hence, the scheme must be iterated to convergence for *each* time-step. This *Variational EM (VEM)* algorithm [4] yields VB-posteriors, (14)–(16), as opposed to a point estimate (ML) under the classical EM scheme. Hence, uncertainties associated with posterior moments are readily available.

The VEM scheme may not be implementable on-line in certain data contexts, since the number of iterations per time-step required for convergence is not known *a priori*. The MEAR model may be inferred without VEM cycles, if optimization is confined only to the parameter distribution, $\tilde{f}(A, \Omega, \boldsymbol{\alpha}|D_t)$, in (13), fixing the remaining term at an appropriate choice. Two choices are now considered.

4.1. Quasi-Bayes (QB) Identification

The closed-form exact marginal of (12) is used [3]:

$$\tilde{f}(\boldsymbol{l}_t|D_t, \boldsymbol{G}) = f(\boldsymbol{l}_t|D_t, \boldsymbol{G}) \propto \mathcal{I}_{\mathcal{D}i}(\boldsymbol{\beta}_{t-1} + \boldsymbol{\delta}_c(i)) \times \mathcal{I}_{\mathcal{NW}}\left(V_{n-1} + \boldsymbol{z}_{i,t} \boldsymbol{z}'_{i,t}, \nu_{n-1} + 1\right), \quad (21)$$

where $\mathcal{I}_{\mathcal{NW}}$ and $\mathcal{I}_{\mathcal{D}i}$ are the closed-form normalizing constants of the $\mathcal{NW}$ and $\mathcal{D}i$ distributions [2]) respectively.

Table 1. Computational complexity for MEAR Identification (m: number of VEM iterations for VB scheme).

algorithm	complexity of one data-update
VB	$m\,(2c+1) \times O\left((r+p)^2\right)$
QB	$2c \times O\left((r+p)^2\right) + c \times O\,(r+p)$
VL	$(c+1) \times O\left((r+p)^2\right) + c \times O\,(r+p)$

4.2. Viterbi-Like (VL) Identification

In each time-step of VB identification, c dyads, $\boldsymbol{z}_{j,t}\boldsymbol{z}'_{j,t}$, $j = 1, \ldots, c$, are used to update the statistics (17). This is unwarranted in cases where one of the dyadic weights, $w_{j,t}$ (20), is dominant, in which case the following certainty equivalence assignment may be successful:

$$\tilde{f}\left(\boldsymbol{l}_t | D_t, \boldsymbol{G}\right) = \delta\left(\boldsymbol{l}_t - \boldsymbol{l}_t^{\mathrm{MAP}}\right), \quad (22)$$

$$\boldsymbol{l}_t^{\mathrm{MAP}} = \arg\max_{\boldsymbol{l}_t} f\left(\boldsymbol{l}_t | D_t, \boldsymbol{G}\right), \quad (23)$$

using (21). Only the dyad corresponding to the inferred active filter (component), $\boldsymbol{l}_t^{\mathrm{MAP}}$, need be used in update (17) at each time step. The computational complexities of the MEAR identification variants are compared in Table 1.

4.3. Time-Variant Component Weights

The assumption of i.i.d. MEAR component weights, $f\left(\boldsymbol{l}_t | \boldsymbol{\alpha}\right)$, may be relaxed, to yield a more realistic varying-weight model reflecting changing observational regimes. In particular, the hidden label field may be modelled as a first-order homogeneous Markov chain, complementing (12) by $f(\boldsymbol{l}_t | \boldsymbol{l}_{t-1}, Q) = \prod_{i,j=1}^{c} q_{i,j}^{l_{i,t}, l_{j,t-1}}$. Using VB-conjugacy (13) once again:

$$f\left(A, \Omega, Q, \boldsymbol{l}_t, \boldsymbol{l}_{t-1} | D_t\right) \approx \approx \tilde{f}\left(A, \Omega, Q | D_t\right) \tilde{f}\left(\boldsymbol{l}_t | D_t\right) \tilde{f}\left(\boldsymbol{l}_{t-1} | D_t\right). \quad (24)$$

The VB principle requires KLD minimization using all three terms as functional variables. The implied VEM scheme involves the same summed-dyad update (17) as for the i.i.d. label case, but the dyadic weights, $w_{j,t}$ (20), are changed, with the last term in the expression being replaced by $\sum_{j=1}^{c} w_{j,t-1} \widehat{\ln q_{i,j}}$. The QB and VL procedures (Sections 4.1 and 4.2 respectively) now require the fixing, as before, of *two* functional variants, being the latter two terms in (24).

5. ESTIMATION OF AN AR MODEL ROBUST TO OUTLIERS

A scalar ($r = 1$) AR process of order p is considered. The presence of isolated outliers is not modelled by the AR model because the outlier-affected observed value does not take part in the future regression. Instead, the process is

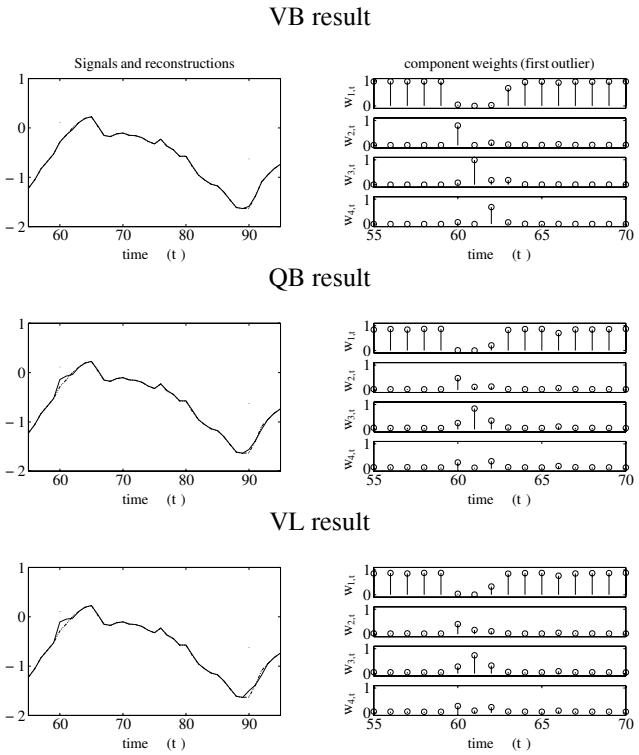

Fig. 1. Reconstruction of an outlier-corrupted AR(2) process.

autoregressive in *unobserved* variable y_t (3), which is observed via

$$d_t = y_t + h_t \omega^{-\frac{1}{2}} e_t. \quad (25)$$

It is convenient to model the outlier as a multiple of the AR innovations, e_t (3), i.e. $h_t = h > 0$ if an outlier occurs, and $h_t = 0$ otherwise. The outlier degrades estimation iff it enters the extended regressor $\boldsymbol{z}_t$. Since $\boldsymbol{z}_t$ is of finite length, and since the outliers are isolated, it is easy to define a finite number of mutually exclusive scenarios [3]. For example, an outlier-corrupted AR(2) process can be approximated by a MEAR model with the following transformations:

$G_1 : \boldsymbol{z}_{1;t} = [d_t, d_{t-1}, d_{t-2}]$	ordinary AR model,
$G_2 : \boldsymbol{z}_{2;t} = \frac{1}{h}[d_t, d_{t-1}, d_{t-2}]$	AR model with higher noise variance, $h^2\omega^{-1}$,
$G_3 : \boldsymbol{z}_{3;t} = [d_t, \hat{y}_{t-1}, d_{t-2}]$	replacing 1-step-delayed output by expected value,
$G_4 : \boldsymbol{z}_{4;t} = [d_t, d_{t-1}, \hat{y}_{t-2}]$	replacing 2-step-delayed output by expected value,

where $\hat{y}_{t-k}$ is an appropriate k-step delayed expectation [3] of the uncorrupted AR process.

Simulation Study A second-order stable AR model with parameters (3) $A = [1.85, -0.95]', \omega = 10$, was simulated with a random outlier on every 30th sample, and $h_t = h = 10$ (25), known *a priori*. In Figure 1, the degraded

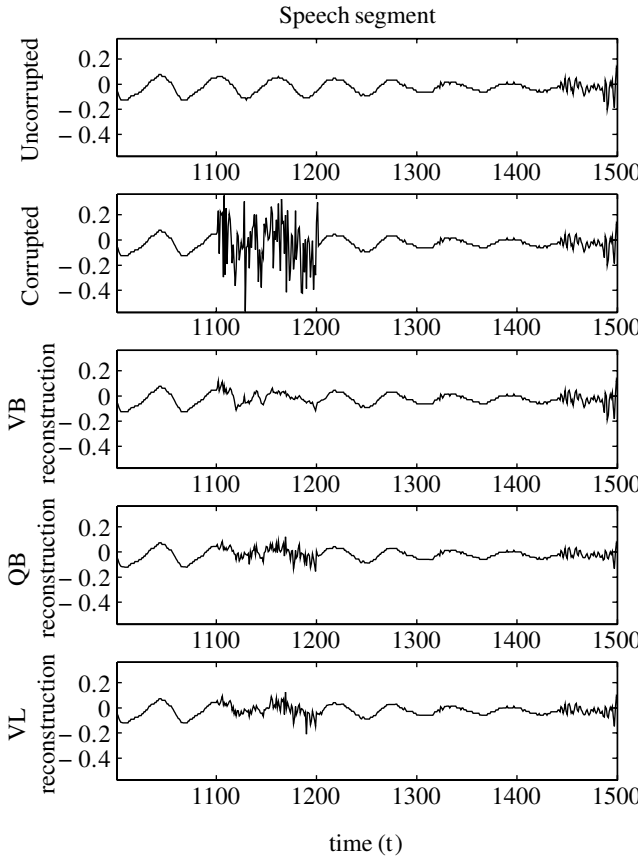

Fig. 2. MEAR-based Speech Reconstruction.

process is shown as the dotted line, along with the reconstruction (solid line). When an outlier occurs, all candidate filters should—by design—be sequentially used to remove the outlier from the estimation formulae (17)-(19). This is achieved almost exactly by the VB algorithm (Figure 1, top right column). The QB and VL algorithms estimate the weights less accurately, as displayed in Figure 1 (right column). A Markov-dependent component sequence is assumed in the MEAR identification procedure (Section 4.3).

6. RECONSTRUCTION OF SPEECH CORRUPTED BY BURST NOISE

A section of the `bbcnews.wav` speech file, sampled at 11 kHz, was corrupted by additive burst noise (Fig. 2). As we do not know the variance of innovations e_n, therefore h_t (25) is assumed to be *unknown* this time. Hence, the MEAR filter-bank was formed by one "direct" filter (modelling the uncorrupted process), and three Kalman Filters, for $h = 3$, $h = 6$, and $h = 10$ respectively. The speech was modelled as non-stationary AR (3) with $p = 8$. A forgetting technique [2] (with forgetting factor $\phi = 0.95$) was used to cope with non-stationarity of the process. Markov filter dependence (Section 4.3) is assumed once again.

Reconstructed values, using the VB, QB and VL methods respectively, are displayed in Figure 2. All three methods successfully suppressed the burst with Mean Squared-Error (MSE) of reconstruction as follows: $\mathrm{MSE}_{\mathrm{VB}} = 20 \times 10^{-4}$, $\mathrm{MSE}_{\mathrm{QB}} = 30 \times 10^{-4}$, and $\mathrm{MSE}_{\mathrm{VL}} = 29 \times 10^{-4}$.

Note that around sample 1450, there is a transition to unvoiced speech, capable of misclassification as burst noise. The VB reconstruction avoided this, but the QB and VL methods tended to suppress the unvoiced speech (Figure 2). The MSE of reconstruction between samples 1450-1500 was: $\mathrm{MSE}_{\mathrm{VB}} = 9 \times 10^{-6}$, $\mathrm{MSE}_{\mathrm{QB}} = 10 \times 10^{-4}$, and $\mathrm{MSE}_{\mathrm{VL}} = 8 \times 10^{-4}$.

7. CONCLUSIONS

The Variational Bayes (VB) procedure presented in this paper enables on-line identification of the MEAR model. The latter significantly extends the modelling potential of the classical AR model. Three VB variants were presented: (i) the full VB procedure, which requires Variational EM (VEM) iterations at each time-step; (ii) the Quasi-Bayes (QB) procedure, which avoids VEM; and (iii) a Viterbi-Like (VL) simplification which further reduces the computational complexity per data-update. The VB procedure is optimal in the KLD-minimization sense, and may have greater flexibility in further model extensions to cope with Markov component weights. It offered moderate performance improvements in the examples presented, while all three variants succeeded in cases where traditional AR modelling cannot cope.

8. REFERENCES

[1] B. Porat, *Digital processing of random signals: theory and methods*. Englewood Cliffs, N.J.: Prentice-Hall, 1994.

[2] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System identification* (P. Eykhoff, ed.), pp. 239–304, Oxford: Pergamon Press, 1981.

[3] V. Šmídl, A. Quinn, M. Kárný, and T. V. Guy, "Robust estimation of autoregressive process using a mixture based filter bank," *Systems & Contol Letters*, 2004. To appear.

[4] M. Sato, "Online model selection based on the variational bayes," *Neural Computation*, vol. 13, pp. 1649–1681, 2001.

[5] L. Ljung and T. Söderström, *Theory and practice of recursive identification*. Cambridge; London: MIT Press, 1983.