CHAPTER 6

Model selection and checking

In the basic HMM with m states, increasing m always improves the fit of the model (as judged by the likelihood). But along with the improvement comes a quadratic increase in the number of parameters, and the improvement in fit has to be traded off against this increase. A criterion for model selection is therefore needed.

In some cases, it is sensible to reduce the number of parameters by making assumptions on the state-dependent distributions or on the t.p.m. of the Markov chain. For an example of the former, see p. 175, where, in order to model a series of categorical observations with 16 circular categories, von Mises distributions (with two parameters) are used as the state-dependent distributions. For an example of the latter, see Section 13.3, where we describe discrete state-space stochastic volatility models which are m-state HMMs with only three or four parameters. Notice that in this case the number of parameters does not increase at all with increasing m.

In this chapter we give a brief account of model selection in HMMs (Section 6.1), and then describe the use of pseudo-residuals in order to check for deficiencies in the selected model (Section 6.2).

6.1 Model selection by AIC and BIC

A problem which arises naturally when one uses hidden Markov (or other) models is that of selecting an appropriate model, e.g. of choosing the appropriate number of states m, sometimes described as the 'order' of the HMM, or of choosing between competing state-dependent distributions such as Poisson and negative binomial. Although the question of order estimation for an HMM is neither trivial nor settled (see Cappé *et al.*, 2005, Chapter 15), we need some criterion for model comparison. The material outlined below is based on Zucchini (2000), which gives an introductory account of model selection.

Assume that the observations $x_1, \ldots, x_T$ were generated by the unknown 'true' or 'operating' model f and that one fits models from two different approximating families, $\{g_1 \in G_1\}$ and $\{g_2 \in G_2\}$. The goal of model selection is to identify the model which is in some sense the best.

There exist at least two approaches to model selection. In the fre-

quentist approach one selects the family estimated to be closest to the operating model. For that purpose one defines a discrepancy (a measure of 'lack of fit') between the operating and the fitted models, $\Delta(f, \hat{g}_1)$ and $\Delta(f, \hat{g}_2)$. These discrepancies depend on the operating model f, which is unknown, and so it is not possible to determine which of the two discrepancies is smaller, i.e. which model should be selected. Instead one bases selection on estimators of the expected discrepancies, namely $\widehat{\mathrm{E}}_f(\Delta(f, \hat{g}_1))$ and $\widehat{\mathrm{E}}_f(\Delta(f, \hat{g}_2))$, which are referred to as model selection criteria. By choosing the Kullback–Leibler discrepancy, and under the conditions listed in Appendix A of Linhart and Zucchini (1986), the model selection criterion simplifies to the Akaike information criterion (AIC):

$$\mathrm{AIC} = -2 \log L + 2p\,,$$

where $\log L$ is the log-likelihood of the fitted model and p denotes the number of parameters of the model. The first term is a measure of fit, and decreases with increasing number of states m. The second term is a penalty term, and increases with increasing m.

The Bayesian approach to model selection is to select the family which is estimated to be most likely to be true. In a first step, before considering the observations, one specifies the priors, i.e. the probabilities $\Pr(f \in G_1)$ and $\Pr(f \in G_2)$ that f stems from the approximating family. In a second step one computes and compares the posteriors, i.e. the probabilities that f belongs to the approximating family, given the observations, $\Pr(f \in G_1 \mid \mathbf{x}^{(T)})$ and $\Pr(f \in G_2 \mid \mathbf{x}^{(T)})$. Under certain conditions (see e.g. Wasserman (2000)), this approach results in the Bayesian information criterion (BIC) which differs from AIC in the penalty term:

$$\mathrm{BIC} = -2 \log L + p \log T\,,$$

where $\log L$ and p are as for AIC, and T is the number of observations. Compared to AIC, the penalty term of BIC has more weight for $T > e^2$, which holds in most applications. Thus the BIC often favours models with fewer parameters than does the AIC.

For the earthquakes series, AIC and BIC both select three states: see Figure 6.1, which plots AIC and BIC against the number of states m of the HMM. The values of the two criteria are provided in Table 6.1. These values are also compared to those of the independent mixture models of Section 1.2.4. Although the HMMs demand more parameters than the comparable independent mixtures, the resulting values of AIC and BIC are lower than those obtained for the independent mixtures.

Several comments arise from Table 6.1. Firstly, given the serial dependence manifested in Figure 2.1, it is not surprising that the independent mixture models do not perform well relative to the HMMs. Secondly, although it is perhaps obvious *a priori* that one should not even try to

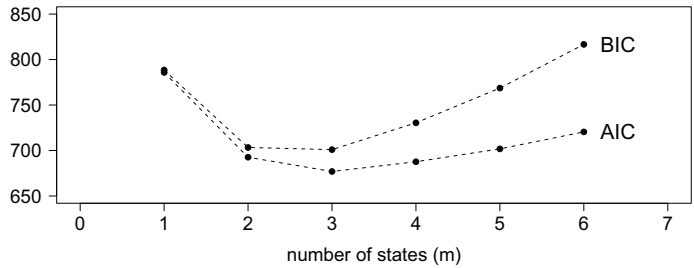

Figure 6.1 *Earthquakes series: model selection criteria AIC and BIC.*

Table 6.1 *Earthquakes data: comparison of (stationary) hidden Markov and independent mixture models by AIC and BIC.*

model	k	$-\log L$	AIC	BIC
'1-state HM'	1	391.9189	785.8	788.5
2-state HM	4	342.3183	692.6	703.3
3-state HM	9	329.4603	**676.9**	**701.0**
4-state HM	16	327.8316	687.7	730.4
5-state HM	25	325.9000	701.8	768.6
6-state HM	36	324.2270	720.5	816.7
indep. mixture (2)	3	360.3690	726.7	734.8
indep. mixture (3)	5	356.8489	723.7	737.1
indep. mixture (4)	7	356.7337	727.5	746.2

fit a model with as many as 25 or 36 parameters to 107 observations, and dependent observations at that, it is interesting to explore the likelihood functions in the case of HMMs with five and six states. The likelihood appears to be highly multimodal in these cases, and it is easy to find several local maxima by using different starting values. A strategy that seems to succeed in these cases is to start all the off-diagonal transition probabilities at small values (such as 0.01) and to space out the state-dependent means over a range somewhat less than the range of the observations.

According to both AIC and BIC, the model with three states is the most appropriate. But more generally the model selected may depend on the selection criterion adopted. The selected model is displayed on p. 51, and the state-dependent distributions, together with the resulting marginal, are displayed in Figure 3.1 on p. 51.

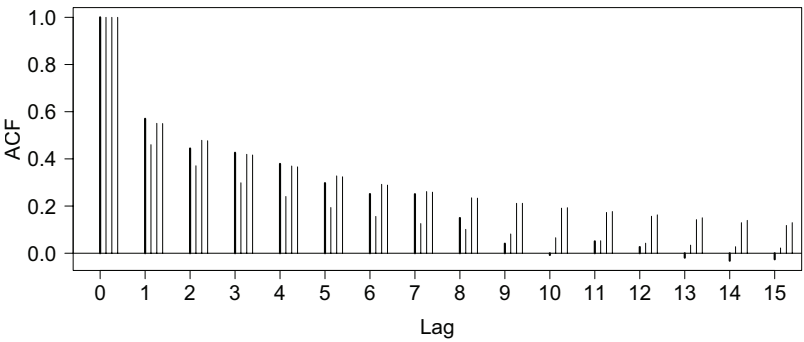

Figure 6.2 *Earthquakes data: sample ACF and ACF of three models. The bold bars on the left represent the sample ACF, and the other bars those of the HMMs with (from left to right) two, three and four states.*

It is also useful to compare the autocorrelation functions of the HMMs with two to four states with the sample ACF. The ACFs of the models can be found by using the results of Exercise 4 of Chapter 2, and appeared on p. 52. In tabular form the ACFs are:

k:	1	2	3	4	5	6	7	8
observations	0.570	0.444	0.426	0.379	0.297	0.251	0.251	0.149
2-state model	0.460	0.371	0.299	0.241	0.194	0.156	0.126	0.101
3-state model	0.551	0.479	0.419	0.370	0.328	0.292	0.261	0.235
4-state model	0.550	0.477	0.416	0.366	0.324	0.289	0.259	0.234

In Figure 6.2 the sample ACF is juxtaposed to those of the models with two, three and four states. It is clear that the autocorrelation functions of the models with three and four states correspond well to the sample ACF up to about lag 7. However, one can apply more systematic diagnostics, as will now be shown.

6.2 Model checking with pseudo-residuals

Even when one has selected what is by some criterion the 'best' model, there still remains the problem of deciding whether the model is indeed adequate; one needs tools to assess the general goodness of fit of the model, and to identify outliers relative to the model. In the simpler context of normal-theory regression models (for instance), the role of residuals as a tool for model checking is very well established. In this section

we describe quantities we call pseudo-residuals which are intended to fulfil this role much more generally, and which are useful in the context of HMMs. We consider two versions of these pseudo-residuals (in Sections 6.2.2 and 6.2.3); both rely on being able to perform likelihood computations routinely, which is certainly the case for HMMs. A detailed account, in German, of the construction and application of pseudo-residuals is provided by Stadie (2002). See also Zucchini and MacDonald (1999).

6.2.1 Introducing pseudo-residuals

To motivate pseudo-residuals we need the following simple result. Let X be a random variable with continuous distribution function F. Then $U \equiv F(X)$ is uniformly distributed on the unit interval, which we write

$$U \sim \mathrm{U}(0,1).$$

The proof is left to the reader as Exercise 1.

The **uniform pseudo-residual** of an observation x_t from a continuous random variable X_t is defined as the probability, under the fitted model, of obtaining an observation less than or equal to x_t:

$$u_t = \Pr(X_t \leq x_t) = F_{X_t}(x_t).$$

That is, u_t is the observation x_t transformed by its distribution function under the model. If the model is correct, this type of pseudo-residual is distributed U(0,1), with residuals for extreme observations close to 0 or 1. With the help of these uniform pseudo-residuals, observations from different distributions can be compared. If we have observations $x_1, \ldots, x_T$ and a model $X_t \sim F_t$, for $t = 1, \ldots, T$ (i.e. each x_t has its own distribution function F_t), then the x_t-values cannot be compared directly. However, the pseudo-residuals u_t are identically U(0,1) (if the model is true), and can sensibly be compared. If a histogram or quantile-quantile plot ('qq-plot') of the uniform pseudo-residuals u_t casts doubt on the conclusion that they are U(0,1), one can deduce that the model is not valid.

Although the uniform pseudo-residual is useful in this way, it has a drawback if used for outlier identification. For example, if one considers the values lying close to 0 or 1 on an index plot, it is hard to see whether a value is very unlikely or not. A value of 0.999, for instance, is difficult to distinguish from a value of 0.97, and an index plot is almost useless for a quick visual analysis.

This deficiency of uniform pseudo-residuals can, however, easily be remedied by using the following result. Let Φ be the distribution function of the standard normal distribution and X a random variable with distribution function F. Then $Z \equiv \Phi^{-1}(F(X))$ is distributed standard

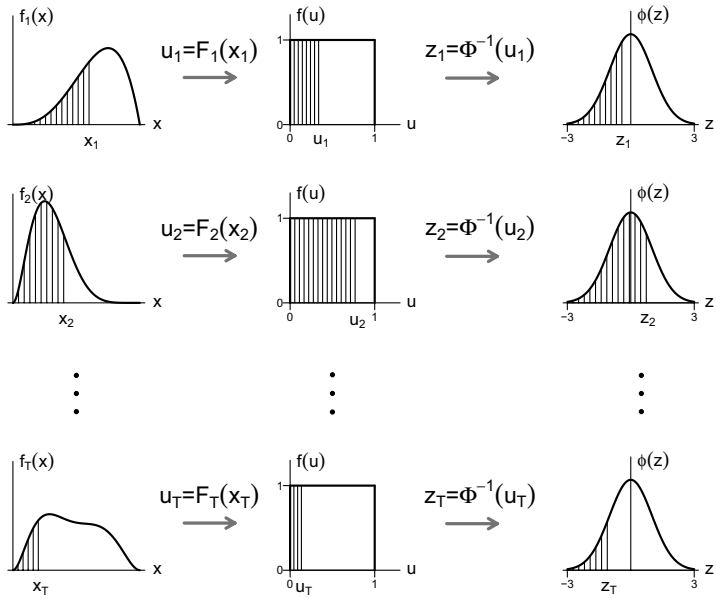

Figure 6.3 *Construction of normal pseudo-residuals in the continuous case.*

normal. (For the proof we refer again to Exercise 1.) We now define the **normal pseudo-residual** as

$$z_t = \Phi^{-1}(u_t) = \Phi^{-1}(F_{X_t}(x_t)).$$

If the fitted model is valid, these normal pseudo-residuals are distributed standard normal, with the value of the residual equal to 0 when the observation coincides with the median. Note that, by their definition, normal pseudo-residuals measure the deviation from the median, and not from the expectation. The construction of normal pseudo-residuals is illustrated in Figure 6.3. If the observations $x_1, \ldots, x_T$ were indeed generated by the model $X_t \sim F_t$, the normal pseudo-residuals z_t would follow a standard normal distribution. One can therefore check the model either by visually analysing the histogram or qq-plot of the normal pseudo-residuals, or by performing tests for normality.

This normal version of pseudo-residuals has the advantage that the absolute value of the residual increases with increasing deviation from the median and that extreme observations can be identified more easily on a normal scale. This becomes obvious if one compares index plots of uniform and normal pseudo-residuals.

Note that the theory of pseudo-residuals as outlined so far can be

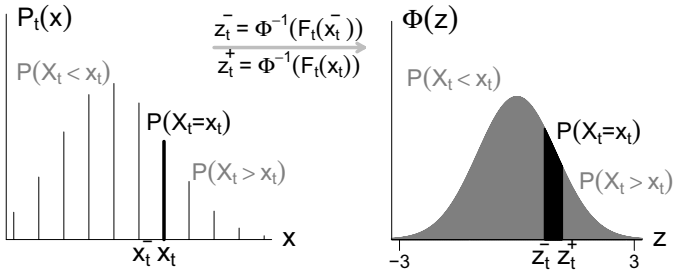

Figure 6.4 *Construction of normal pseudo-residuals in the discrete case.*

applied to continuous distributions only. In the case of discrete observations, the pseudo-residuals can, however, be modified to allow for the discreteness. The pseudo-residuals are no longer defined as points, but as intervals. Thus, for a discrete random variable X_t with distribution function F_{X_t} we define the uniform pseudo-residual segments as

$$[u_t^-; u_t^+] = [F_{X_t}(x_t^-); F_{X_t}(x_t)], \tag{6.1}$$

with x_t^- denoting the greatest realization possible that is strictly less than x_t, and we define the normal pseudo-residual segments as

$$[z_t^-; z_t^+] = [\Phi^{-1}(u_t^-); \Phi^{-1}(u_t^+)] = [\Phi^{-1}(F_{X_t}(x_t^-)); \Phi^{-1}(F_{X_t}(x_t))]. \tag{6.2}$$

The construction of the normal pseudo-residual segment of a discrete random variable is illustrated in Figure 6.4.

Both versions of pseudo-residual segments (uniform and normal) contain information on how extreme and how rare the observations are, although the uniform version represents the rarity or otherwise more directly, as the length of the segment is the corresponding probability. For example, the lower limit u_t^- of the uniform pseudo-residual interval specifies the probability of observing a value strictly less than x_t, $1 - u_t^+$ gives the probability of a value strictly greater than x_t, and the difference $u_t^+ - u_t^-$ is equal to the probability of the observation x_t under the fitted model. The pseudo-residual segments can be interpreted as interval-censored realizations of a uniform (or standard normal) distribution, if the fitted model is valid. Though this is correct only if the parameters of the fitted model are known, it is still approximately correct if the number of estimated parameters is small compared to the size of the sample (Stadie, 2002). Diagnostic plots of pseudo-residual segments of discrete random variables necessarily look rather different from those of continuous random variables.

It is easy to construct an index plot of pseudo-residual segments or to plot these against any independent or dependent variable. However, in order to construct a qq-plot of the pseudo-residual segments one has to specify an ordering of the pseudo-residual segments. One possibility is to sort on the so-called 'mid-pseudo-residuals' which are defined as

$$z_t^m = \Phi^{-1}\left(\frac{u_t^- + u_t^+}{2}\right). \tag{6.3}$$

Furthermore, the mid-pseudo-residuals can themselves be used for checking for normality, for example via a histogram of mid-pseudo-residuals.

Now, having outlined the properties of pseudo-residuals, we can consider the use of pseudo-residuals in the context of HMMs. The analysis of the pseudo-residuals of an HMM serves two purposes: the assessment of the general fit of a selected model, and the detection of outliers. Depending on the aspects of the model that are to be analysed, one can distinguish two kinds of pseudo-residual that are useful for an HMM: those that are based on the conditional distribution given all other observations, which we call **ordinary pseudo-residuals**, and those based on the conditional distribution given all preceding observations, which we call **forecast pseudo-residuals**.

That the pseudo-residuals of a set of observations are identically distributed (either U(0,1) or standard normal) is their crucial property. But for our purposes it is not important whether such pseudo-residuals are independent of each other; indeed we shall see in Section 6.3.2 that it would be wrong to assume of ordinary pseudo-residuals that they are independent.

Note that Dunn and Smyth (1996) discuss (under the name 'quantile residual') what we have called normal pseudo-residuals, and point out that they are a case of Cox–Snell residuals (Cox and Snell, 1968).

6.2.2 Ordinary pseudo-residuals

The first technique considers the observations one at a time and seeks those which, relative to the model and *all* other observations in the series, are sufficiently extreme to suggest that they differ in nature or origin from the others. This means that one computes a pseudo-residual z_t from the conditional distribution of X_t, given $\mathbf{X}^{(-t)}$; a 'full conditional distribution', in the terminology used in MCMC (Markov chain Monte Carlo). For continuous observations the normal pseudo-residual is

$$z_t = \Phi^{-1}\left(\Pr(X_t \leq x_t \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)})\right).$$

If the model is correct, z_t is a realization of a standard normal random variable. For discrete observations the normal pseudo-residual segment

is $[z_t^-; z_t^+]$, where

$$z_t^- = \Phi^{-1}\left(\Pr(X_t < x_t \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)})\right)$$

and

$$z_t^+ = \Phi^{-1}\left(\Pr(X_t \leq x_t \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)})\right).$$

In the discrete case the conditional probabilities $\Pr(X_t = x \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)})$ are given by Equations (5.1) and (5.2) in Section 5.1; the continuous case is similar, with probabilities replaced by densities.

Section 6.3.1 applies ordinary pseudo-residuals to HMMs with one to four states fitted to the earthquakes data, and a further example of their use appears in Figure 9.3 and the corresponding text.

6.2.3 Forecast pseudo-residuals

The second technique for outlier detection seeks observations that are extreme relative to the model and all *preceding* observations (as opposed to all other observations). In this case the relevant conditional distribution is that of X_t given $\mathbf{X}^{(t-1)}$. The corresponding (normal) pseudo-residuals are

$$z_t = \Phi^{-1}\left(\Pr(X_t \leq x_t \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)})\right)$$

for continuous observations; and $[z_t^-; z_t^+]$ for discrete, where

$$z_t^- = \Phi^{-1}\left(\Pr(X_t < x_t \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)})\right)$$

and

$$z_t^+ = \Phi^{-1}\left(\Pr(X_t \leq x_t \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)})\right).$$

In the discrete case the required conditional probability $\Pr(X_t = x_t \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)})$ is given by the ratio of the likelihood of the first t observations to that of the first $t-1$:

$$\Pr(X_t = x \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}) = \frac{\boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x)\mathbf{1}'}{\boldsymbol{\alpha}_{t-1}\mathbf{1}'}.$$

The pseudo-residuals of this second type are described as forecast pseudo-residuals because they measure the deviation of an observation from the median of the corresponding one-step-ahead forecast. If a forecast pseudo-residual is extreme, this indicates that the observation concerned is an outlier, or that the model no longer provides an acceptable description of the series. This provides a method for the continuous monitoring of the behaviour of a time series. An example of such monitoring is given at the end of Section 15.4: see Figure 15.5.

The idea of forecast pseudo-residual appears — as 'conditional quantile residual' — in Dunn and Smyth (1996); in the last paragraph on

p. 243 they point out that the quantile residuals they describe can be extended to serially dependent data. The basic idea of (uniform) forecast pseudo-residuals goes back to Rosenblatt (1952), however. Both Brockwell (2007) and Dunn and Smyth describe a way of extending what we call forecast pseudo-residuals to distributions other than continuous. Instead of using a segment of positive length to represent the residual if the observations are not continuous, they choose a point distributed uniformly on that segment. The use of a segment of positive length has the advantage, we believe, of explicitly displaying the discreteness of the observation, and indicating both its extremeness and its rarity.

Another example of the use of forecast pseudo-residuals appears in Figure 9.4 and the corresponding text.

6.3 Examples

6.3.1 Ordinary pseudo-residuals for the earthquakes

In Figure 6.5 we show several types of residual plot for the fitted models of the earthquakes series, using the first definition of pseudo-residual, that based on the conditional distribution relative to all other observations, $\Pr(X_t = x \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)})$. The relevant code appears in A.2.10. It is interesting to compare the pseudo-residuals of the selected three-state model to those of the models with one, two and four states.

As regards the residual plots provided in Figure 6.5, it is clear that the selected three-state model provides an acceptable fit while, for example, the normal pseudo-residuals of the one-state model (a single Poisson distribution) deviate strikingly from the standard normal distribution. If, however, we consider only the residual plots (other than perhaps the qq-plot), and not the model selection criteria, we might even accept the two-state model as an adequate alternative.

Looking at the last row of Figure 6.5, one might be tempted to conjecture that, if a model is 'true', the ordinary pseudo-residuals will be independent, or at least uncorrelated. From the example in the next section we shall see that that would be an incorrect conclusion.

6.3.2 Dependent ordinary pseudo-residuals

Consider the stationary Gaussian AR(1) process $X_t = \phi X_{t-1} + \varepsilon_t$, with the innovations ε_t independent standard normal. It follows that $|\phi| < 1$ and $\mathrm{Var}(X_t) = 1/(1-\phi^2)$.

Let t lie strictly between 1 and T. The conditional distribution of X_t given $\mathbf{X}^{(-t)}$ is that of X_t given only X_{t-1} and X_{t+1}. This latter conditional distribution can be found by noting that the joint distribution of X_t, X_{t-1} and X_{t+1} (in that order) is normal with mean vector $\mathbf{0}$ and

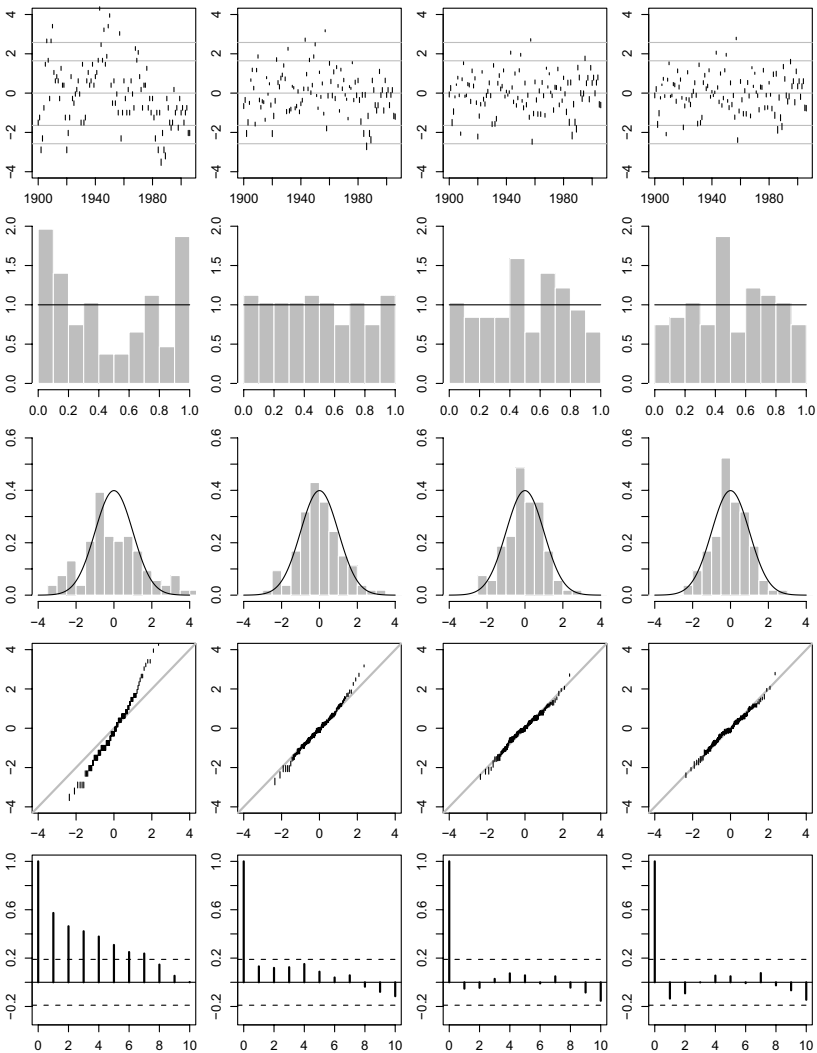

Figure 6.5 *Earthquakes: ordinary pseudo-residuals. Columns 1–4 relate to HMMs with (respectively) 1, 2, 3, 4 states. The top row shows index plots of the normal pseudo-residuals, with horizontal lines at 0,* ± 1.96, ± 2.58. *The second and third rows show histograms of the uniform and the normal pseudo-residuals. The fourth row shows quantile-quantile plots of the normal pseudo-residuals, with the theoretical quantiles on the horizontal axis. The last row shows the autocorrelation functions of the normal pseudo-residuals.*

covariance matrix

$$\mathbf{\Sigma} = \frac{1}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi \\ \phi & 1 & \phi^2 \\ \phi & \phi^2 & 1 \end{pmatrix}.$$

The required conditional distribution then turns out to be normal with mean $\phi(X_{t-1}+X_{t+1})/(1+\phi^2)$ and variance $(1+\phi^2)^{-1}$. Hence the corresponding uniform pseudo-residual is

$$\Phi\left(\left(X_t - \frac{\phi}{1+\phi^2}(X_{t-1}+X_{t+1})\right)\sqrt{1+\phi^2}\right),$$

and the normal pseudo-residual is

$$z_t = \left(X_t - \frac{\phi}{1+\phi^2}(X_{t-1}+X_{t+1})\right)\sqrt{1+\phi^2}.$$

By the properties of ordinary normal pseudo-residuals (see Section 6.2.2), z_t is unconditionally standard normal; this can also be verified directly.

Whether (e.g.) the pseudo-residuals z_t and z_{t+1} are independent is not immediately obvious. The answer is that they are not independent; the correlation of z_t and z_{t+1} can (for $t > 1$ and $t < T-1$) be obtained by routine manipulation, and turns out to be

$$-\phi(1-\phi^2)/(1+\phi^2);$$

see Exercise 3. This is opposite in sign to ϕ and smaller in modulus. For instance, if $\phi = 1/\sqrt{2}$, the correlation of z_t and z_{t+1} is $-\phi/3$. (In contrast, the corresponding *forecast* pseudo-residuals do have zero correlation at lag 1.)

One can, also by routine manipulation, show that $\mathrm{Cov}(z_t, z_{t+2}) = 0$ and that, for all integers $k \geq 3$,

$$\mathrm{Cov}(z_t, z_{t+k}) = \phi\,\mathrm{Cov}(z_t, z_{t+k-1}).$$

Consequently, for all integers $k \geq 2$,

$$\mathrm{Cov}(z_t, z_{t+k}) = 0.$$

□

6.4 Discussion

This may be an appropriate point at which to stress the dangers of over-interpretation. Although our earthquakes model seems adequate in important respects, this does not imply that it can be interpreted substantively. Nor indeed are we aware of any convincing seismological interpretation of the three states we propose. But models need not have a substantive interpretation to be useful; many useful statistical models

are merely empirical models, in the sense in which Cox (1990) uses that term.

Latent-variable models of all kinds, including independent mixtures and HMMs, seem to be particularly prone to over-interpretation, and we would caution against the error of reification: the tendency to regard as physically real anything that has been given a name. In this spirit we can do no better than to follow Gould (1997, p. 350) in quoting John Stuart Mill:

> The tendency has always been strong to believe that whatever received a name must be an entity or being, having an independent existence of its own. And if no real entity answering to the name could be found, men did not for that reason suppose that none existed, but imagined that it was something peculiarly abstruse and mysterious.

Exercises

1.(a) Let X be a continuous random variable with distribution function F. Show that the random variable $U = F(X)$ is uniformly distributed on the interval $[0, 1]$, i.e. $U \sim \mathrm{U}(0, 1)$.

(b) Suppose that $U \sim \mathrm{U}(0, 1)$ and let F be the distribution function of a continuous random variable. Show that the random variable $X = F^{-1}(U)$ has the distribution function F.

(c) i. Give the explicit expression for F^{-1} for the exponential distribution, i.e. the distribution with density function $f(x) = \lambda\, e^{-\lambda x},\ x \geq 0$.

ii. Verify your result by generating 1000 uniformly distributed random numbers, transforming these by applying F^{-1}, and then examining the histogram of the resulting values.

(d) Show that for a continuous random variable X with distribution function F, the random variable $Z = \Phi^{-1}(F(X))$ is distributed standard normal.

2. Consider the AR(1) process in the example of Section 6.3.2. That the conditional distribution of X_t given $\mathbf{X}^{(-t)}$ depends only on X_{t-1} and X_{t+1} is fairly obvious because X_t depends on $X_1, \ldots, X_{t-2}$ only through X_{t-1}, and on $X_{t+2}, \ldots, X_T$ only through X_{t+1}.

   Establish this more formally by writing the densities of $\mathbf{X}^{(T)}$ and $\mathbf{X}^{(-t)}$ in terms of conditional densities $p(x_u \mid x_{u-1})$ and noting that, in their ratio, many of the factors cancel.

3. Verify, in the AR(1) example of Section 6.3.2, that for appropriate ranges of t-values:

(a) given X_{t-1} and X_{t+1},

$$X_t \sim \mathrm{N}\Big(\phi(X_{t-1}+X_{t+1})/(1+\phi^2), (1+\phi^2)^{-1}\Big);$$

(b) $\mathrm{Var}(z_t) = 1$;

(c) $\mathrm{Corr}(z_t, z_{t+1}) = -\phi(1-\phi^2)/(1+\phi^2)$;

(d) $\mathrm{Cov}(z_t, z_{t+2}) = 0$; and

(e) for all integers $k \geq 3$, $\mathrm{Cov}(z_t, z_{t+k}) = \phi\,\mathrm{Cov}(z_t, z_{t+k-1})$.

4. Generate a two-state stationary Poisson–HMM $\{X_t\}$ with t.p.m. $\mathbf{\Gamma} = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}$, and with state-dependent means $\lambda_1 = 2$ and $\lambda_2 = 5$ for $t = 1$ to 100, and $\lambda_1 = 2$ and $\lambda_2 = 7$ for $t = 101$ to 120. Fit a model to the first 80 observations, and use forecast pseudo-residuals to monitor the next 40 observations for evidence of a change.

5. Consider again the soap sales series introduced in Exercise 5 of Chapter 1.

   (a) Use AIC and BIC to decide how many states are needed in a Poisson–HMM for these data.

   (b) Compute the pseudo-residuals relative to Poisson–HMMs with 1–4 states, and use plots similar to those in Figure 6.5 to decide how many states are needed.