STA 214: Probability & Statistical Models

Fall Semester 2004
Mike West

January 19, 2006

Orientation

MATERIAL OMITTED: THIS VERSION DOES NOT HAVE EXERCISES AND SOLUTIONS

1 AR(1) Models

1.1 Introduction

- Time series: Stochastic process in (discrete) time, sometimes (often) equally spaced.
- Scalar time series, continuous measurements - continuous *state space*.
- Simplest non-trivial model: AR(1) - AutoRegressive of order 1.
- Time series is "regressed" on itself - prior value predicts current value.

For $t = 0, 1, \ldots,$ and in principle for theoretical development, $t = -1, -2, \ldots,$

$$x_t = \phi x_{t-1} + \epsilon_t, \qquad \epsilon_t \sim N(0, v). \tag{1}$$

- AR parameter ϕ.
- Innovation (error, evolution error, stochastic input at time t) ϵ_t.
- Innovations are independent, Gaussian (normal), zero-mean and constant variance, and independent of x_{t-1} and all past x_{t-j}. Independence is often written as $e_t \perp\!\!\!\perp e_s$ (for all t, s).
- Innovation is time t random "shock," added to "predicted" value for x_t given by ϕx_{t-1}. The shock is "unpredictable" in the sense that $E(\epsilon_t) = 0$, and the innovations variance v defines the scale of randomness injected into the evolution of the x_t process at each time point.
- Look at models, structure, distribution theory assuming that the *model parameters* θ defined by $\theta = (\phi, v)$ are known. Model analysis in applications involves estimation of parameters and model assessment.
- Notation (not yet, perhaps, standard, but useful): $x_t \leftarrow AR(1|\theta)$, or $AR(1|(\phi, v))$.
- More general models could have different innovations variances at each time, or even non-normal distributions for the innovations.

AR(1) models are of major interest in their own right as simple stochastic process models for many time series applications, but also of very major importance as building blocks of more complex models representing real phenomena. The model class is also a nice setting to introduce core ideas of multivariate distribution theory linked to normal models, to develop initial ideas of simulation - of univariate and structure multivariate distributions, stochastic processes and then posterior distributions for parameters arising in model fitting and Bayesian inference. The AR(1) model class is an example of a class of Markov (Markovian) stochastic processes on a continuous (univariate) state space, which provides - among other things - key examples of Markov chains, and entree to the ideas and theory of Markov chains and of Markov Chain Monte Carlo (MCMC) simulation methods. As components of more complicated probability models, AR(1) models can become "hidden" (latent) processes, so providing examples of hidden Markov models (HMMs).

1.2 Structure and Distribution Theory in Stationary AR(1) Processes

1.2.1 Stationarity

Stationarity of the process means that the $n-$variate joint distribution of $x_{s:s+n-1} = (x_s, x_{s+1}, \ldots, x_{s+n-1})'$ does not depend on s, for any $n \geq 1$. *Weak stationarity* refers to the mean and variance-covariance of the joint distribution, but in the case of linear, normal models those moments characterize the full joint distribution. In particular,

- $n = 1$: each x_t has the same distribution,
- $n = 2$: Each pair of values x_t, x_s has the same bivariate distribution,

and in the current context these are all normal distributions.

1.2.2 Conditional and Marginal Univariate Distributions

Critical to distinguish distributions via the conditioning elements. All distributions are implicitly conditioned on the specified parameter values, for now. Model equation (1) specifies, for all t, the conditional distribution

$$p(x_t|x_{t-1}) = N(\phi x_{t-1}, v). \tag{2}$$

The first-order Markovian property is that, conditional on x_{t-1}, the distribution of x_t does not depend on previous values x_{t-j} for $j > 1$. This is often written as $x_t \perp\!\!\!\perp x_{t-j}, (j > 1)|x_{t-1}$, although a perhaps clearer notation is $(x_t|x_{t-1}) \perp\!\!\!\perp x_{t-j},\ j > 1$.

Assuming weak stationarity alone, write $m = E(x_t)$ and $s = V(x_t)$ so that for all t we know the marginal distribution $x_t \sim N(m, s)$. We can see that

- $m = E(x_t) = E[E(x_t|x_{t-1})] = E(\phi x_{t-1}) = \phi m$, which can only hold if $m = 0$ unless $\phi = 1$, a very special nonstationary *random walk* case.
- Similarly, $s = V(x_t) = E[V(x_t|x_{t-1})] + V[E(x_t|x_{t-1})] = v + \phi^2 s$ so that $s = v/(1 - \phi^2)$. This can only make sense if $|\phi| < 1$, characterising stationary AR(1) models.

1.2.3 Linear Process

The model is a linear time series model, or linear process. Iterate equation (1) to get

$$x_t = \epsilon_t + \phi\epsilon_{t-1} + \ldots + \phi^k\epsilon_{t-k} + \ldots. \tag{3}$$

The process is *linear* - a linear function of current and past innovations, and a sum of independent stochastic elements that are weighted by the AR parameter. If $|\phi| < 1$ the weight at lag k decays as k increases so that the current value of the x process is less and less dependent on the past innovations. An *explosive* (quite nonstationary) process results otherwise.

Linear processes can be non-Gaussian. This framework is Gaussian. Imagine an AR(1) model in which the innovations are independent but from some non-Gaussian distribution, such as a Student T or Cauchy, or other.

1.2.4 Backshift Operator

Backshift operator notation and manipulation: $Bx_t = x_{t-1}$ and $B^k x_t = x_{t-k}$ for all $k > 0$. The model then can be written as $(1 - \phi B)x_t = \epsilon_t$ and this becomes useful for formal manipulations. In algebraic equations the B operator can be treated as if it were a number in (0,1): $x_t = (1 - \phi B)^{-1}\epsilon_t$ and the expansion $(1 - \phi B)^{-1} = 1 + \phi B + \phi^2 B^2 + \ldots$ lead to the (stationary, lagged weights decaying with time) linear representation equation (3).

1.3 Autocorrelations and Full Joint Distributions

- Covariance at lag k: $\gamma(k) = C(x_t, x_{t\pm k})$
- $\gamma(k) = \phi^k s$
- Correlation at lag k: $\rho(k) = \phi^k$
- Use linear representation, or iterated covariances to derive. Autocorrelation at first lag is the defining AR parameter ϕ.
- Not well-defined in nonstationary processes, though some nonstationary processes do have definable conditional autocorrelations.

Any set of x values has a joint normal distribution, a key example is that for (any) n consecutive values, such as $x_{1:n} = (x_1, x_2, \ldots, x_n)'$, a column $n-$vector. We know the mean (0) and the variances and covariances, and the linear representation means that $x_{1:n}$ is a linear function of independent normal innovations so is multivariate normal (see multivariate normal theory reference material too). We write

$$x_{1:n} \sim N(0, \Sigma_n) \tag{4}$$

where 0 is now the $n-$vector of zeros and Σ_n is the variance (or variance-covariance matrix, a symmetric positive definite - SPD - $n \times n$ matrix), which is $\Sigma_n = s\Phi_n$ with correlation matrix

$$\Phi_n = \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \phi & 1 & \phi & \cdots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \cdots & 1 \end{pmatrix}. \tag{5}$$

Some general considerations related to existence of these joint distributions, stationarity and nonstationarity, and also structure of the stationary model:

- A nice exercise in linear algebra: Find Φ_n^{-1}.
- What happens if $\phi = 0$?
- Many applications have $\phi > 0$, but some will involve "oscillatory" behaviour consistent with $\phi < 0$.
- What about the *random walk* limit at $\phi = 1$? The model is well-defined but nonstationary, and this is a very, very important, simple model. Look at $p(x_t|x_0)$ for $t > 0$.

1.4 Bivariate Distributions, Markov Transitions & Reversibility

$$p(x_{t-1}, x_t) = p(x_t|x_{t-1})p(x_{t-1}) \tag{6}$$

- $p(x_t|x_{t-1})$ is normal, linear regression from model equation (1).
- $p(x_{t-1})$ is (stationary) marginal $N(0, s)$.
- Fill in density functions to show the bivariate density is log-quadratic, a normal density.

Key identity for stationary processes: We know that, always in any bivariate distribution,

$$p(x_t) = \int p(x_t|x_{t-1})p(x_{t-1})dx_{t-1}. \tag{7}$$

In this special framework of stationary (linear, normal) AR(1) models, the two univariate margins are the same distribution, $N(0, s)$. The conditional distribution defined by the model is also named the *transition distribution* of the Markov process (or *evolution distribution*). Changing notation a little to highlight this very general, key representation, write $g(u)$ for the density function of $u \sim N(0, s)$, and $t(x|u)$ for the conditional normal density of $(x|u) \sim N(\phi u, v)$, then the key identity above is

$$g(x) = \int t(x|u)g(u)du. \tag{8}$$

The bivariate density is then, in a general setting, $t(x|u)g(u)$. The special stationary linear/Gaussian structure here means this is the same as $t(u|x)g(x)$. See this by looking at the bivariate distribution for $(x, u)'$ with vector mean 0 and 2×2 variance matrix

$$\Sigma_2 = s \begin{pmatrix} 1 & \phi \\ \phi & 1 \end{pmatrix},$$

and then noting that both of the conditional distributions - that for $(x|u)$ and that for $(u|x)$ - are given by $t(\cdot|\cdot)$.

Back to the time series context, this means that the process is *time reversible*, i.e., it looks the same back in time as it does forwards in time. This general result follows directly just by looking at the joint normal distribution, and can be verified by using Bayes theorem: starting with $(x_t|x_{t-1}) \sim N(\phi x_{t-1}, v)$ and $x_{t-1} \sim N(0, s)$, show using Bayes theorem that $(x_{t-1}|x_t) \sim N(\phi x_t, v)$.

1.5 Joint Distributions in Compositional/Markov Form

$$p(x_{1:t}) = p(x_t|x_{t-1})p(x_{t-1}|x_{t-2}) \dots p(x_2|x_1)p(x_1) \tag{9}$$

- Uses Markovian (lag 1) structure.
- Each $p(x_s|x_{s-1})$ is normal, linear regression from model equation (1).
- Initial value x_1 (or any other): $N(0, s)$

1.6 Simulation

What do AR processes look like? Simulate some, for different choices of parameters. How? One way is to just simulate the multivariate normal, but that's not likely most efficient. Best approach is to utilize as much *local structure* in the multivariate normal as possible, and here the *local structure* is Markovian and captured in the compositional decomposition of the density (and corresponding distribution).

Assume we know how to simulate a univariate normal random quantity - we'll discuss that more later. Simulate a *sample path* or *realization* from the AR(1) model of equation (1), generating n consecutive values $x_{1:n}$. This is equivalent to both (a) simulating from the full joint normal distribution - the multivariate normal in $n-$dimensions of equation (5), and (b) simulating through the sequence of conditional (Markovian) densities in the compositional form equation (9). As follows:

- *Initialize:* draw a sample (a simulated value) $x_1 \sim p(x_1) \equiv N(0, s)$, via

  $$x_1 = \sqrt{s} z_1$$

  where $z_1 \sim N(0, 1)$ is a first (synthetic, quasi-random) simulated value in this Monte Carlo exercise.
- For $t = 2 : n$ in sequence, successively generate $(x_t|x_{t-1})$ from the AR(1) model:

  $$x_t = \phi x_{t-1} + \epsilon_t$$

  where now $\epsilon_t = \sqrt{v} z_t$ are realized values of the innovations computed from synthetic Monte Carlo draws $z_t \sim N(0, 1)$, and each evaluated x_{t-1} is"plugged-in" to the equation for the next time step.

2 Likelihoods and Reference Bayesian Inference in AR(1) Models

Key support material:
pages 15-17, 19-23 of the Draft Text Material on time series on the web site are primary.

2.1 Likelihood Functions

Now explicitly recognize dependence on parameters in all densities, so that we have the joint density

$$p(x_{1:n}|\theta) = p(x_n|x_{n-1},\theta)p(x_{n-1}|x_{n-2},\theta)\ldots p(x_2|x_1,\theta)p(x_1|\theta). \tag{10}$$

In θ, this gives the likelihood function:

$$p(x_{1:n}|\theta) \propto (1-\phi^2)^{1/2}v^{-n/2}\exp(-Q^*(\phi)/2v) \tag{11}$$

with

$$Q^*(\phi) = Q(\phi) + (1-\phi^2)x_1^2, \qquad Q(\phi) = \sum_{t=2}^{n}(x_t - \phi x_{t-1})^2. \tag{12}$$

2.2 Reference Bayesian Analysis Conditioning on Initial Values

For n large, the effect of the initial value in the complicating factor $1-\phi^2$ is "small", and the conditional likelihood function is very often used:

$$p(x_{2:n}|x_1,\theta) \propto v^{-(n-1)/2}\exp(-Q(\phi)/2v). \tag{13}$$

- Approximate analysis relative to actual or "full" likelihood;
- Small difference for large n;
- Valid conditional distribution anyhow;
- Can easily compare with actual likelihood;
- Bayesian analysis via simulation methods will easily allow use of full likelihood, as we shall see.

Key results: Conditional likelihood is of standard form, and MLE and reference Bayesian analysis routine. Reference prior is $p(\theta) \propto v^{-1}$ (standard reference analysis for normal random samples and linear regressions, of which this is a special case) so that the posterior is easy:

$$p(\theta|x_{1:n}) \equiv p(\phi, v|x_{1:n}) \propto v^{-(n+1)/2}\exp(-Q(\phi)/2v). \tag{14}$$

Structure of bivariate posterior for $(\phi, v|x_{1:n})$:

- $p(\phi|x_{1:n}, v) \propto \exp(-B(\phi-b)^2/2v)$ with $B = \sum_{t=1}^{n-1} x_t^2$ and $b \equiv \phi_{ML} = B^{-1}\sum_{t=2}^{n} x_t x_{t-1}$. Clearly b relates to the sample correlation at lag 1, and B to the sample variance. Then

  $$(\phi|x_{1:n}, v) \sim N(b, vB^{-1}). \tag{15}$$

  b is the reference posterior mode for ϕ as well as the MLE and LSE.
- Verify the quadratic form can be expressed as

  $$Q(\phi) = Q(b) + B(\phi - b)^2. \tag{16}$$

  This is a special, simple example of the standard decomposition of quadratic forms in linear regression models: $Q(b)$ is the residual sum of squares based on the "fitted" parameter value b.

- Integrating ϕ out of equation (14),

$$p(v|x_{1:n}) \propto v^{-n/2} \exp(-Q(b)/2v). \tag{17}$$

This is the density of an inverse chi-square distribution: $1/v = \kappa/Q(b)$ where $\kappa \sim \chi^2_{n-2}$.

Gammas and chi-square distributions: Write $\psi = 1/v$. By transformation ψ has density proportional to $\psi^{(n-2)/2-1} \exp(\psi Q(b)/2)$, the density of $\psi \sim Ga((n-2)/2, Q(b)/2)$. By transform again, therefore, $\psi Q(b) \sim Ga((n-2)/2, 1/2) \equiv \chi^2_{n-2}$ (by definition of the chi-square distribution). Easy to simulate from $p(v|x_{1:n})$: draw $\kappa \sim \chi^2_{n-2}$ and compute $v = Q(b)/\kappa$.

The two distributions $p(\phi|x_{1:n}, v)$ and $p(v|x_{1:n})$ define the joint posterior. Verify that the implied marginal posterior for $(\phi|x_{1:n})$ is a Student T distribution.

Question and Issue: If we want to ensure that the model represents a stationary process, then how can we have a normal distribution describing likely values and uncertainty about ϕ? It may be concentrated in the stationary region, or we may have to condition; but, conditioning destroys the theory just outlined.

2.3 Posterior Analysis via Parameter Simulation

If we take the view that much parametric inference in based on posterior sampling (we do), then we care less about specific mathematical forms of marginal posteriors than we do about decomposition of posteriors for Monte Carlo analysis. We can make a draw (ϕ, v) from the posterior equation (14) by composition: draw from the margin for v then, conditioning on that value, draw from the conditional for ϕ given v. Same trick that is used in simulation complicated joint distributions of realizations of the x_t process (which we will do again in the next subsection).

One benefit of using simulation is that we'll see if sampled ϕ values live in the stationary region $|\phi| < 1$. Just by looking at how many fall outside gets us into assessing whether or not the fitted model is consistent with stationarity. If we sample thousands of posterior draws, just looking at the fraction within $|\phi| < 1$ is a start (and often the finish) on this issue. Later we'll see how this fits into more formal Monte Carlo analysis using accept-reject methods.

- Generate some synthetic AR(1) data and fit the reference Bayesian analysis. Sample the posterior distribution and explore histograms of posterior samples for each of ϕ and v (or, better, the innovation scale parameter $\sqrt{v}$.)
- Fit the reference analysis to the SOI time series, or any other real data set. Explore posterior histograms, means, etc.

2.4 Predictive Simulation: Forecasting and Model Evaluation

Exploring model fit and implications through simulation for any chosen value of θ is useful, but ignores parameter (estimation) uncertainty. Formally, the relevant distribution for a specified set of unknown future values $x_{(n+1):(n+m)}$ up to $m > 0$ steps ahead is the *predictive* distribution, with density

$$p(x_{(n+1):(n+m)}|x_{1:n}) = \int p(x_{(n+1):(n+m)}|x_{1:n}, \theta) p(\theta|x_{1:n}) d\theta. \tag{18}$$

This can be shown to be a multivariate T distribution, but the use of simulation to generate draws from it is just trivial. Again the key is compositional sampling of a joint distribution:

- Sample a synthetic parameter value from $p(\theta|x_{1:n})$,
- plug-in this generated value for θ in its placeholder in the conditioning of $p(x_{(n+1):(n+m)}|x_{1:n}, \theta)$, and then

- sample $x_{(n+1):(n+m)} \sim p(x_{(n+1):(n+m)}|x_{1:n}, \theta)$.

We have already seen this last step - sampling from the model $AR(1|\theta)$ for a given parameter value, again by composition now sequencing through Monte Carlo draws of $(x_{n+1}|x_n, \theta)$, then $(x_{n+2}|x_{n+1}, \theta)$, and so on. Note that now, however, we must conditional on the pre-initial value x_n to get started, as that as been observed.

Sample realizations - "simulated futures" of the process - generated this way are draws from the formally correct posterior predictive distribution for the evolution of the process into the future. They are based on and incorporate the information relevant to estimation of the parameters, but now also reflect estimation uncertainty: if there is a lot of uncertainty about θ, due to small n for example, then the resulting higher level of uncertainty in the posterior feeds through to the predictions.

- Explore "simulated futures" based on models fitted to synthetic AR(1) data and real data, such as the SOI series.

3 Two Classes of Hidden Markov Models with AR(1) Components

3.1 AR(1) Process Observed with Noise

Observed values of a process are now y_t, and

$$\begin{aligned} y_t &= x_t + \nu_t \\ x_t &\leftarrow AR(1|\theta) \end{aligned}$$

where $\nu_t \sim N(0, w)$ and with $\nu_t \perp\!\!\!\perp \nu_s$ and $\nu_t \perp\!\!\!\perp \epsilon_s$ for all t, s. The ν_t terms are errors of measurement, or of observation, that "corrupt" the signal x_t.

This is a hidden Markov model (HMM), and one of the simplest. It is also one of the more important models in time series as many real processes are not directly observable: measurement error and other forms of technical error, noise obscure the signal (x_t). The first equation is equivalent to $(y_t|x_t) \sim N(x_t, w)$, so that y_t is an unbiased measurement, but not perfect. The relative values of w and v define signal-to-noise characteristics. In a stationary model we know $V(x_t) = s = v/(1 - \phi^2)$ so that the variance of each observation is $s + w$ and the SNR (signal-to-noise ratio) is $s/(s + w)$.

What are the stochastic characteristics of the observed process y_t? The y_t process is stationary and linear, Gaussian so each y_t has a normal marginal distribution, each pair is bivariate normal, and so forth. Is it Markovian? Is it time reversible?

3.2 A Class of Stochastic Volatility Models

A famous class of HMMs are stochastic volatility (SV) models (SVMs) of interest in quantitative finance, the cornerstone of many statistical decision support tools in financial research and investment management. The canonical example:

$$\begin{aligned} y_t &\sim N(0, \sigma_t^2) \\ \sigma_t &= \exp(\mu + x_t) \\ x_t &\leftarrow AR(1|\theta) \end{aligned}$$

- Fix parameters $\theta = (\phi, v)$ and μ and initial state x_0
- Simulate sample paths ("trajectories") of x_t, y_t
- Structure in x_t series
- Structure in y_t series
- Structure in y_t^2 series
- Real data: Returns on international exchange rate markets, where all the action is in the changes in variance of returns, and it is important to model and capture persistence in variances (volatilities).
- Add a dynamic mean: Econometrics and dynamic models - small drifts
- Fit models to data: Assessment, prediction

Some insight into the structure of the model, and also complications in model fitting, arise by noting the implication that, conditional on μ, x_t, the random quantity $y_t^2 = \sigma_t^2 \kappa_t$ where $\kappa_t \sim \chi_1^2$. So if $z_t = \log(y_t^2)/2 = \log(|y_t|)$,

$$\begin{aligned} z_t &= \mu + x_t + \nu_t \\ x_t &\leftarrow AR(1|\theta) \end{aligned}$$

where $\nu_t = \log(\kappa_t)/2$. So we have a model in which the observed quantities, now z_t, are a constant (intercept, level, defining the *baseline* volatility on the log scale) plus a *latent* AR(1) process x_t that defines the time-correlated changes in volatility. This is very similar to the linear-Gaussian AR(1) plus noise model, the simple HMM, but now with two differences: (a) the model has an intercept term, which is only a small detail; (b) the observational noise is non-Gaussian, a bigger issue - the distribution of ν_t is that of 1/2 times the log of a chi-square on 1 degree of freedom, which is not quite Gaussian, somewhat skewed, and a little difficult to work with mathematically. One approach is to approximate this with a normal distribution, and that puts us into the linear, Gaussian hidden AR model framework. More effective approximations exist and can be developed using simulation methods; this model context is one in which we really begin to need Monte Carlo methods for model analysis and prediction.

Multivariate normal distribution theory underlies a good part of the structure and analysis of these models - hidden Markov models with latent AR components, and also the SV models even though the sampling distribution is non-normal. The full joint distribution of any set of log volatilities is, initially, multivariate normal as it is just a linear AR(1) process, and much of the trickery needed in model fitting, especially using MCMC methods, relies very heavily on the implied collections of marginal and conditional distributions. So we do need to know normal distribution theory intimately.

4 Multivariate Normal Theory

See the notes under Supporting Materials on the course web site for much of the theory (and some that may not be so relevant to this course, but still part of the theory and relevant elsewhere). Some elaboration and additions to the theory outlined there, and ties into exploration of normal models and also simulation, are described here.

4.1 Density, Ellipses, Contours

Random quantity x is $p-$vector with a Gaussian/normal distribution, $x \sim N(m, V)$, with pdf

$$p(x) = ((2\pi)^p |V|)^{-1/2} \exp(-Q(x)/2)$$

with

$$Q(x) = (x - m)' V^{-1} (x - m).$$

The family of multivariate normals $x \sim N(m, V)$ is *elliptically symmetric*, being a function only of the quadratic form $Q(x)$ around the centroid m. Points x of constant density lie on elliptical contours (hyper-ellipses in p dimensions - simple ellipses when $p = 2$.) The shape of the density scales as the marginal variances change, but maintain elliptical shapes. The ellipses are oriented along the primary axes when V is diagonal.

Draw some density contours in the bivariate normal for various choices of (m, V).

4.2 Definition and m.g.f.s

The best definition of the multivariate normal is that based on arbitrary linear combinations being univariate normal, and it is best seen, and proven, using either moment generating functions Laplace transforms of p.d.f.s) or characteristic generating functions (Fourier transforms of p.d.f.s). For example, in terms of the moment generating function (m.g.f.):

- Univariate normal: $x \sim N(m, v)$ if and only if $E(\exp(tx)) = \exp(mt + vt^2/2)$ as a function of t, characterizing the normal distribution. This is trivially derived.
- Multivariate normal: in d dimensions, $x \sim N(m, V)$ has m.g.f. $E(\exp(t'x)) = \exp(t'm + t'Vt/2)$ as a function of the $p-$vector-valued argument t, characterizing the (multivariate)normal distribution.
- Any linear combination: scalar random quantity $y = a'x$ has m.g.f. given by

  $$E(\exp(ty)) = E(\exp((ta)'x)) = \exp(t(a'm) + (a'Va)t^2/2)$$

  so that $y \sim N(a'm, a'Va)$.
- Similar direct calculations deliver the results about the normal distributions any affine function $y = Ax + b$ and general linear forms. These are central results in much of statistics.
- The use of Laplace and Fourier transformations in statistical work is limited, but they are very important in theoretical development of probability models when interest lies in weighted averages of random quantities, and characterization of distributions, as this key example shows.

4.3 Partitioned Normal Distributions

The $p-$dimensional normal random quantity $x \sim N(m, V)$ is partitioned as

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

with each subvector x_i of dimension p_i. The mean and variance partition conformably:

$$m = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \qquad V = \begin{pmatrix} V_1 & R \\ R' & V_2 \end{pmatrix},$$

and of course $x_i \sim N(m_i, V_i)$ (by an application of the linear transformation of normals).

Assume that V is non-singular (invertible) so that each of the partitioned variance matrices is too. The *precision matrix* of x is

$$K = V^{-1}.$$

Standard linear algebraic results (or easy derivation) give the inverse of any partitioned matrix, here simplified due to the symmetry of V. The precision matrix K is partitioned conformably with V and given by

$$K = V^{-1} = \begin{pmatrix} K_1 & H \\ H' & K_2 \end{pmatrix},$$

with entries

- $K_1^{-1} = V_1 - RV_2^{-1}R'$,
- $H = -K_1 R V_2^{-1}$,
- $K_2 = V_2^{-1} + H'K_1^{-1}H$.

4.4 Partitioned Normal Distributions: Precision Matrix

In statistical modeling and inference, a central role is played by the regression structure of conditional distributions, and these are trivially derived using standard linear algebraic expressions for the inverses of partitioned matrices. The conditional distributions (regressions) relate intimately to the structure of the precision matrix. Take the zero-mean case: $x \sim N(0, V) = N(0, K^{-1})$ with $p(x) \propto \exp(-Q(x)/2)$ where

$$Q(x) = x'Kx = x_1'K_1x_1 + 2x_1'Hx_2 + x_2'K_2x_2.$$

By inspection,

$$p(x_1|x_2) \propto \exp(-Q(x_1|x_2)/2)$$

where

$$Q(x_1|x_2) = x_1'K_1x_1 + 2x_1'Hx_2.$$

This is quadratic in x_1 (for any conditioning value of x_2) which implies the conditional normality. Center and complete the quadratic to obtain

$$\begin{aligned} Q(x_1|x_2) &= (x_1 + K_1^{-1}Hx_2)'K_1(x_1 + K_1^{-1}Hx_2) \\ &= (x_1 - RV_2^{-1}x_2)'K_1(x_1 - RV_2^{-1}x_2) \end{aligned}$$

based on the formula for H, whereupon $E(x_1|x_2) = A_1x_2$ and $V(x_1|x_2) = K_1^{-1}$ where A_1 has each of the forms $A_1 = RV_2^{-1}$ and $A_1 = -K_1^{-1}H$.

In the general case with a non-zero mean, $x \sim N(m, V)$ and then, similarly,

$$(x_1|x_2) \sim N(m_1 + A_1(x_2 - m_2), K_1^{-1}).$$

The expression $A_1 = RV_2^{-1}$ is the traditional form and shows how the regression of x_1 on x_2 is based on the covariance elements R being rotated and scaled by the variance V_2 of the conditioning variables. The second expression for A_1 relates to the elements H of the precision matrix K, and has critical uses, as follows.

4.5 Precision Matrix and Univariate Complete Conditional Distributions

An extremely important special case is when $p_1 = 1$ so that x_1 is scalar and $x_2 = x_{2:p} = x_{1:p\setminus 1}$. By extension the same theory holds for the *univariate complete conditional distribution* of any element x_i, that is $p(x_i|x_{1:p\setminus i})$, $(i = 1, \ldots, p)$.

Work now in terms of all the *univariate* elements $x = (x_1, \ldots, x_p)'$ and $m = (m_1, \ldots, m_p)'$, as well as the elements $K_{i,j}$ of the full precision matrix K $(i, j = 1, \ldots, p)$.

- The complete conditional normal distribution of x_1 has $V(x_1|x_{1:p\setminus 1}) = 1/K_{1,1}$ and

$$E(x_1|x_{1:p\setminus 1}) = m_1 - \sum_{j\in(1:p\setminus 1)} K_{1,j}(x_j - m_j)/K_{1,1}.$$

  The conditional regression of x_1 on the other variables is such that the regression coefficient on each x_j is $-K_{1,j}/K_{1,1}$.
- By extension, for any $i = 1, \ldots, p$, the complete conditional normal distribution of $(x_i|x_{1:p\setminus i})$ is normal with moments

$$E(x_i|x_{1:p\setminus i}) = m_i + \sum_{j\in(1:p\setminus i)} \gamma_{i,j}(x_j - m_j) \qquad \text{and} \qquad V(x_i|x_{1:p\setminus i}) = 1/K_{i,i}$$

  where

$$\gamma_{i,j} = -K_{i,j}/K_{i,i}.$$

This shows explicitly how the elements of the precision matrix K of the full joint distribution determine all the relevant conditional structure.

- Zeros in the precision matrix K define, and are defined by, *conditional independencies* in $p(x)$. The precision $K_{i,j} = 0$ if any only if the complete conditional distribution of x_i does not depend on x_j, equivalent to $x_i \perp\!\!\!\perp x_j$ conditional on all $x_k, k \in 1 : p \setminus (i, j)$.
- This is a key aspect of analysis of multivariate models and, in particular, underlies basic ideas in *Gaussian graphical models*.

4.6 Example: AR(1) in Noise - A Hidden Markov Model

$$\begin{aligned} y_t &= x_t + \nu_t \\ x_t &\leftarrow AR(1|(\phi, v)) \end{aligned}$$

with $\nu_t \sim N(0, w)$ and with $\nu_t \perp\!\!\!\perp \nu_s$ and $\nu_t \perp\!\!\!\perp \epsilon_s$ for all t, s.

Consider the following: suppose the parameters (ϕ, v, w) are specified and we want to explore what the data y tells us about the hidden (latent, unobservable, unknown) process x. The objective is then to compute and understand, and perhaps simulate, from distributions for sets of x values conditional on some observations from the y series. We will do this for a consecutive n time points, so we want to find and interpret

$$p(x_{1:n}|y_{1:n}).$$

This is the $n-$dimensional conditional distribution from the $2n-$dimensional joint distribution of $x_{1:n}$ and $y_{1:n}$ jointly, so we find that first and then use the above theory to condition.

For any $n > 0$,

$$\begin{aligned} y_{1:n} &= x_{1:n} + \nu_{1:n} \\ x_{1:n} &\sim N(0, \Sigma_n) \end{aligned}$$

with Σ_n as earlier derived, $\Sigma_n = s\Phi_n$ with $s = v/(1-\phi^2)$ and correlation matrix

$$\Phi_n = \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \phi & 1 & \phi & \cdots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \cdots & 1 \end{pmatrix}.$$

Also, $x_{1:n} \perp\!\!\!\perp \nu_{1:n}$ and

$$\nu_{1:n} \sim N(0, wI).$$

Notice that we are working in multivariate normals now, for vectors of length n. Let's now move to $p = 2n$ dimensions and look at the joint distribution of $x_{1:n}$ and $y_{1:n}$. Based on the linearity of y in x and the normal, independence structure, we know this is normal, so we just need the moments. Clearly $E(y_{1:n}) = 0$. Then

- $V(y_{1:n}) = V(x_{1:n} + \nu_{1:n}) = \Sigma_n + wI$, and
- $C(x_{1:n}, y_{1:n}) = E(x_{1:n}y'_{1:n})$ (since the means are zero), so that the covariance is

$$E(x_{1:n}x'_{1:n} + x_{1:n}\nu'_{1:n}) = V(x_{1:n}) + 0 = \Sigma_n.$$

Hence

$$\begin{pmatrix} x_{1:n} \\ y_{1:n} \end{pmatrix} \sim N(0, V_n) \qquad \text{with} \qquad V_n = \begin{pmatrix} \Sigma_n & \Sigma_n \\ \Sigma_n & \Sigma_n + wI \end{pmatrix}.$$

This is a very special example with a highly structured variance matrix. Apply the conditional normal theory to produce the required distribution: in the earlier notation we have $V_1 = \Sigma_n, R = R' = \Sigma_n$ and $V_2 = \Sigma_n + wI$. It follows that $A \equiv A_1 = \Sigma_n(\Sigma_n + wI)^{-1}$ and $K_1^{-1} = wA$, so that

$$(x_{1:n}|y_{1:n}) \sim N(Ay_{1:n}, wA).$$

Explore some examples, including evaluation of the conditional distribution here for some specified AR parameter values and signal-to-noise ratios s/q where $q = s + w$.

Key points to note:

- In the conditional mean $E(x_{1:n}|y_{1:n}) = Ay_{1:n}$ as a set of point estimates of the signal, each x_t is estimated by a weighted linear combination of y_t and other values, with highest weight on y_t and weights decaying away from t. This is *smoothing* - y_t is an unbiased estimate of x_t, but neighbouring y values also provide information since they are correlated with x_t.
- The conditional mean $E(x_{1:n}|y_{1:n}) = Ay_{1:n}$ as a set of point estimates of the signal is "shrunk" towards zero relative to the data point estimates (zero being the "initial" or prior point estimate under the marginal normal distribution for the signal. The *shrinkage effect* is greater for larger values of w (smaller signal-to-noise). That is, the more noise we have in the measurement error process, the smoother the estimated signal will be (the harder it is to "track" the real signal). Play with some simulations and varying parameter values to explore and appreciate this.

Notice that we have now discussed core aspects of theory, analysis and simulation for (a) inference on AR(1) model parameters when we actually observe the x process (the earlier Bayesian reference analysis and simulation of the posterior for $(\phi, v|x_{1:n})$), and now (b) inference on the x process itself when it is latent, through the conditional $p(x_{1:n}|y_{1:n}, (\phi, v, w))$. These two components are almost all - but not quite all - that we need to move to the final stages of full analysis, inference and prediction in the HMM AR(1) model. That will include posterior simulations to generate inference on all the three parameters, (ϕ, v, w), together with the latent $x_{1:n}$. More on that later, when we have the relevant concepts and initial tools of Gibbs sampling, the first - and absolutely central - example of MCMC simulation methods.

4.7 Linear Transforms and Cholesky for Simulation

If $x \sim N(m, V)$ then $x = m + L\epsilon$ where the $p-$vector $\epsilon \sim N(0, I)$ and $LL' = V$. This holds for any *matrix square-root* L of V. The Cholesky decomposition of any (strictly) positive definite symmetric matrix V has L as a lower triangular matrix, and is compute efficiently using only the diagonal and lower (or upper) off-diagonal elements of V. The diagonal elements of L are positive. Samples of x are generated by this transformation, sampling ϵ as a vector of p independent, standard univariate normals. The correlation structure and scaling of $p(x)$ is accounted for through the Cholesky component L.

There are other matrix decompositions, including eigen-decompositions, of use in multivariate normal models and including alternative simulation methods. The Cholesky is, however, standard and numerically most efficient in cases of non-singular normal distributions.

4.8 Singular/Degenerate Normal Distributions

What if V is singular - non-negative definite? Consider the bivariate normal with unit variances and correlation 0.999' as a key example. In general cases, V is non-singular and so rank-deficient.

5 Eigenstructure of Variance Matrices

5.1 Non-singular case

$p \times p$ positive definite and symmetric matrix V.

- V has p positive eigenvalues $d_1 > \ldots > d_p > 0$. Write $D = \text{diag}(d_1, \ldots, d_p)$.
- The p eigenvectors - $p-$vectors $e_1, \ldots e_p$ - defined by $Ve_j = d_j e_j$ for $j = 1, \ldots, p$, are orthogonal: $e_j' e_j = 1$ and $e_i' e_j = 0$ for $i \neq j$. The (eigenvector column) matrix $E = [e_1, \ldots e_p]$ is orthogonal: $E'E = EE' = I$.
- From definition $VE = ED$ we have the key eigen-decomposition (or spectral decomposition):

$$V = EDE'.$$

- Principal components decomposition, also related to singular value decomposition (SVD) as we shall see later.
- $V(x) = V = V(E\epsilon)$ for any $p-$vector random quantity ϵ such that $V(\epsilon) = D$. In the normal case, take elements of ϵ as p independent normals with variances d_j. Otherwise, in non-normal cases the variance matrix analysis is the same, though the elements will generally be dependent (though uncorrelated). The transform

$$x = E\epsilon$$

is key to principal components analysis (PCA) and other statistical computations. For example, simulation of x can be done in the normal case by simulation of p independent normals in ϵ, an alternative to the Cholesky method.

- Reciprocally, $\epsilon = E'x$ has variance matrix D.
- $\epsilon_i = e_i'x$ is the i^{th} principal component transformation of x. Note that if $x \sim N(0, V)$ then $\epsilon_i \sim N(0, d_i)$.
- In normal (and other elliptically symmetric) distributions, this principal component transformation represents a rotation of the axis of the ellipses: the density $p(\epsilon)$ has the same shape as $p(x)$ but the elliptical contours are aligned with the primary axes ϵ_1, ϵ_2, etc - decorrelation is ellipse realignment.
- $Tr(V) = \sum_{i=1}^{p} d_i = d'1$ is *the total variation* under $p(x)$. Note also that $|V| = |D| = \prod_{i=1}^{p} d_i$. The total variation under $p(x)$ is the same as that under $p(\epsilon)$. The i^{th} principal component ϵ_i contributes $100d_i/(d'1)\%$ of this total variation; the first eigenvalue is the largest, so the first component is the "dominant" component, and so on. If, for example, d_1 is really large compared to the rest of the eigenvalues, then $p(x)$ is very heavily concentrated around that one-dimensional subspace; a bivariate normal with very high correlation is a simple and useful example. If a number of eigenvalues are relatively very small, then $p(x)$ is coming close to concentrating in fewer than p dimensions, consistent with V approaching singularity.
- Precision matrix $K = V^{-1} = ED^{-1}E'$, so that K has same eigenvectors as V and the reciprocal eigenvalues.

5.2 Singular case

$p \times p$ symmetric matrix V of rank $k < p$, so V^{-1} is undefined. Now V has just k positive eigenvalues and the remaining $p - k$ are zero.

- Eigenvalues are $d_1 > \ldots > d_k > d_{k+1} = 0, \ldots, d_p = 0$.
- Eigen-decomposition of V is now
$$V = EDE'$$
where E is $p \times k$ of full rank k, and has columns that are the eigenvectors of the positive eigenvalues of the $k \times k$ matrix $D = \mathrm{diag}(d_1, \ldots, d_k)$.
- Now $E'E = I$ $(k \times k)$ as before, but the $p \times p$ matrix EE' is not the identity (it is of rank k so non-singular.)
- Only $k < p$ principal components matter: $p - k$ constraints on the elements of x are implied.
- $x = E\epsilon$ where now the action is in the $k-$ dimensions of ϵ with $V(\epsilon) = D$.
- Generalized inverse: $V^- = ED^{-1}E'$ plays the role of the precision matrix, again of rank $k < p$.
- The p.d.f. of the singular normal $x \sim N(m, V)$ is defined in terms of this generalized inverse and the reduced dimension:
$$p(x) = ((2\pi)^k |D|)^{-1/2} \exp(-Q(x)/2)$$
with
$$Q(x) = (x - m)'V^-(x - m).$$
- Standard software (Matlab, R/Splus) generally delivers full eigen-decompositions, including the zero eigenvalues and (some) eigenvectors they correspond to in the null space of V. We generally need to reduce the output to the dimension of relevance, k. Numerical instabilities creep in (quickly) in higher dimensional problems. Note also that software packages differ in how they choose to order the eigenvalues and eigenvectors - Matlab, for example, orders them in decreasing rather than increasing order.

6 Simulation and Basics of Simple Random Variate Generation

Stochastic simulation methods are central tools in modern science and technology. Simulation methods for random variate generation and numerical approximation via Monte Carlo integration are, at a real practical level, simply approaches to exploring (usually mathematically complicated and multivariate) distributions. We use the term (stochastic) *simulation* interchangeably with *Monte Carlo*.

6.1 Uniform Pseudo-Random Generators

The uniform distribution underlies all practical simulation methods. See Robert & Casella (§2.1) for discussion of algorithms for generating high-quality approximations to uniform U(0,1) variates. These are all deterministic algorithms, including chaotic process models, and generate *pseudo-random numbers* - sequences that are, these days and for most practical purposes, independent and uniformly distributed in the sense that a finite sequence of values so generated cannot be distinguished from a theoretically exact uniform random sample using statistical tests. Random quantities so generated are always in fact based on integer sequences taking values in $(0, 1, \ldots, M)$ where M is a very large integer. Most effective are so-called *congruential* methods have periods of recurrence (return to starting value, or seed, and then repeat) that are very large. Scientific software and packages such as Matlab and R/Splus implement very high-quality and high-period algorithms, and these should be used as standard in most statistical and scientific work. The current Matlab (v6) `rand` generator, for example, claims to be able to generate all the floating point numbers in the closed interval $[2^{-53}, 1 - 2^{-53}]$ and with a period in excess of 10^{400}. For some studies it is of interest to explore the results of a simulation analysis based on repeat evaluation using the same sequence of uniform random variates; in such cases, if a long or complex simulation is involved, the underlying pseudo-random number generator can be reset to begin at the same initial value - the *seed* - for repeat analyses.

6.2 Univariate Distributions via the Inverse CDF Transform

For a c.d.f. $P(\cdot)$, generate $x \sim P(x)$ via

- $x = P^-(u)$ with $u \sim U(0, 1)$ where

$$P^-(u) = \inf\{x : P(x) \geq u\}.$$

If $P(x)$ is continuous, $P^-(u) = P^{-1}(u)$ is the standard inverse function, or *quantile function* of the distribution. The generalized inverse P^- applies also for discrete distributions, and for more interesting cases of mixed discrete and continuous distributions, examples in which $P(x)$ is piecewise continuous but exhibits jumps, for example. P^- defines the generalized quantile function of the distribution.

Some key examples and specific results are of note.

- *Location and/or scale transformations*.
  If $P(x) = P_0((x - m)/s)$ for some location m and scale $s > 0$, and where P_0 is a specified standard distribution, then $x = m + sz$ where $z \sim P_0$. See this by simply noting that $P^-(u) = m + sP_0^-(u)$. Key examples of the normal, Cauchy and other T distributions, but also exponential, gamma and other distributions that arise commonly.
- x is exponential with rate $1/t$, or mean t, if $P(x) = 1 - \exp(-x/t)$ so that $x = -t\log(1 - u)$. Equivalently, $x = -t\log(u)$. This is also an example of scaling from the standard (unit) exponential when $t = 1$.
- $x \sim C(0, 1)$ has $p(x) = 1/\{\pi(1+x^2)\}$ so that $P(x) = 1/2+\arctan(x)/\pi$; thus $x = \tan(\pi(u-1/2))$. Location/scale extensions have $x = m + s\tan(\pi(u - 1/2))$.

6.3 Transformation Methods

In many examples, direct transformations create samples from distributions of interest based on from samples from more standard distributions.

- $x \sim Ga(a, b)$ is a scale transform from the standard gamma with shape a, i.e., $x = y/b$ where $y \sim Ga(a, 1)$.
- x is standard lognormal if $x = \exp(z)$ where $z \sim N(0, 1)$.
- Beta and Dirichlet (generalised, multivariate versions of the beta) are transformations of gamma variables.
- Consider the AR(1) model and suppose we know the innovations variance and assume the conditional posterior $(\phi|x_{1:n}) \sim N(b, v/B)$ for ϕ where $b = \sum_{t=2}^{n} x_t x_{t-1}/B$ and $B = \sum_{t=1}^{n-1} x_t^2$. Ignore for now the stationarity condition and assume that this normal distribution concentrates in $(-1, 1)$ with negligible probability outside. One quantity of interest in some studies is τ, the *correlation length*, or *correlation half-life*, i.e., the time it takes for the absolute value of the correlation between x_t and $x_{t+\tau}$ to decay to 0.5. Since the correlation at lag k is ϕ^k, then $\tau = -\log(2)/\log(\phi)$. Simulate values for ϕ from the normal posterior, and simply transform them to get a sample from $p(\tau|x_{1:n})$.
- The Box-Muller transformation (Robert & Casella, §2.2) is a (venerable) example of a bivariate transformation: two $U(0, 1)$ variates $u_1 \perp\!\!\!\perp u_2$ transform to two $N(0, 1)$ variates $x_1 \perp\!\!\!\perp x_2$ via $x_1 = \sqrt{-2\log(u_1)}\cos(2\pi u_2)$ and $x_2 = \sqrt{-2\log(u_1)}\sin(2\pi u_2)$, easily verified by direct transformation of the bivariate p.d.f. of (u_1, u_2) utilizing the Jacobian. (Most standard software does not use this method nowadays, though it still represents a useful example of transformation methods.)
- Location/scale transformations provide other examples, including the multivariate normal: $x = m + L\epsilon$ samples the $p-$dimensional $N(m, LL')$ distribution by transformation from a set of p independent univariate normal elements of ϵ.

It is easy to underestimate the importance of transformation methods - they are central to simulation-based analysis of complex, high-dimensional models. Suppose we have a large sample from a distribution $P(x)$ for some $p-$vector x, and are interested in a set of q quantities in a $q-$vector y where each element of y may be a complicated function of x. Analytic calculations are impossible, but transforming the Monte Carlo sample is trivial so long as each y_i can be evaluated as a function of x. This generates a sample of vectors y and we can just by inspection look at histograms and other aspects to explore the implied marginal distribution of any element of y.

6.4 Simulation via Convolutions: Mixtures of Distributions

We have already met this in several key example.

- *T distributions*: A scalar quantity x has a standard (Student) T distribution with $\nu > 0$ degrees of freedom if the p.d.f. is proportional to $\{1 + x^2/\nu\}^{-(\nu+1)/2}$. The p.d.f. has the form

$$p(x) \propto \int_0^\infty \lambda^{1/2} \exp(-x^2\lambda/2) p(\lambda) d\lambda$$

  where $p(\lambda)$ is the p.d.f. of the gamma distribution $\lambda \sim Ga(\nu/2, 1/2)$. Location and scale transformations of x generate the full class of univariate T distributions.
- *Multivariate T distributions*: The above is a special case of $p = 1$ in the multivariate T distribution where $p(x) \propto \{1 + (x - m)'V^{-1}(x - m)/\nu\}^{-(\nu+p)/2}$. This is a scale mixture of normals too: with conditional variance matrix V, we have $(x|\lambda) \sim N(m, V/\lambda)$ and $\lambda \sim Ga(\nu/2, 1/2)$. Simulate the

T distribution via the mixture (i.e., via convolution) as $x = m + \lambda^{-1/2} L\epsilon$ where L is the Cholesky lower-triangular component of V, ϵ is a vector of p independent standard univariate normals, and λ is a draw from the gamma distribution.

- Consider a stochastic process - such as the AR(1) model - for a series $x_1, x_t, \ldots$ with a model that defines the joint distribution in terms of compositional form *forward in time*:

$$p(x_{1:n}) = p(x_1) \prod_{t=2}^{n} p(x_t | x_{1:(t-1)}).$$

If we sequentially simulate draws from $p(x_1)$, then from $p(x_2|x_1)$ conditional on the simulated value of x_1, and so forth, we generate a sequence $x_{1:n}$ such that, by construction, each x_t is a draw from the implied marginal distribution $p(x_t)$. Hence, within this compositional analysis we have a whole series of embedded convolutional simulations.

6.5 Compositional Sampling

Sampling mixtures is an example of compositional sampling, though the latter is usually used to denote sampling of a joint distribution of interest in its own right, whereas - often - mixture/convolutional sampling quite often simply uses the mathematical form of a density as a mixture, if such is available, as a technical device.

The joint density of any set of p random quantities can be written in *many* compositional forms, and sometimes one is preferred over others for its technical structure in connection with simulation. Generally, $x \sim p(x)$ can be simulated by sequencing through

- simulate x_1 from the univariate margin for x_1, and given that specific value,
- simulate x_2 from the conditional $p(x_2|x_1)$; given that value,
- simulate $x_3 \sim p(x_3|x_2, x_1)$, and so on.

Any subset of elements of x so sampled represents a draw from the corresponding marginal distribution. We have seen several examples already: the example of the normal/gamma structure of the T distribution - where x and λ are drawn from the joint distribution via composition, and also as in the example of simulation of the AR(1), and other stochastic processes where the x_t are naturally ordered in time.

6.6 Posterior Simulation in Bayesian Analysis

Simulation of posterior distributions provides approaches to evaluation of inferences about complicated functions of model parameters that would, otherwise, be difficult to evaluate or even estimate.

- *AR(1) Example*

  In the reference Bayesian analysis of the AR(1) model we can simulate the posterior for the AR(1) model parameters, (conditioning on the initial value x_1) from the posterior directly, as it has a conditional normal-gamma form. sample the gamma (scaled chi-squared) distribution. In the earlier notation, we simulate $p(\phi, v|x_{1:n})$ via composition, drawing $(v^{-1}|x_{1:n}) \sim Ga((n-2)/2, Q(b)/2)$ and then $(\phi|x_{1:n}, v) \sim N(b, vB^{-1})$.

  One aspect of interest in this example is stationarity: the posterior is not constrained to a stationary model, for which $|\phi| < 1$. If we generate a large random sample as above, then any draw (ϕ^i, v^i) such that $|\phi^i| \geq 1$ indicates a nonstationary model, so that we can assess, by Monte Carlo, just how well supported the stationarity assumption is by looking at the proportion of such draws. This line of thinking is formalized below in connection with Monte Carlo integration.

Very often the posterior distribution is much more complicated, mathematically, and direct simulation via composition or convolutional methods is just not possible. In such cases, accept/reject methods, approximations using analytic approximations and importance (weighted) sampling, and especially MCMC are more relevant.

6.7 Mixtures, Compositional Sampling & Missing Data: Some Key Examples

- Mixtures arise natural in certain models - the Student T as a mixture of normals is a nice example. They also arise from parameter uncertainty. The predictive distribution in a Bayesian analysis is a mixture over the (prior or posterior) distribution representing parameter uncertainty: $p(x) = \int p(x|\theta)p(\theta)d\theta$ is the prior predictive distribution in a model for data x and with parameters θ. In predicting new data y from the corresponding model component $p(y|x,\theta)$ - to forecast, or to generate insights into model fit and adequacy - we often use simulation of the implied *posterior predictive distribution* with p.d.f.

  $$p(y|x) = \int p(y|x,\theta)p(\theta|x)d\theta.$$

  We have already seen an example in the AR(1) model where $x = x_{1:n}$ and $y = x_{(n+1):(n+m)}$ is the future m values of a time series process. Note that random sampling models are very special cases in which $y \perp\!\!\!\perp x|\theta$ so that $p(y|x,\theta) = p(y|\theta)$.
- Discrete mixtures of parametric distributions - such as a discrete mixture (weighted average) of normal distributions - are used in many statistical studies, including kernel density estimation, to represent non-standard distributional forms. For example, a discrete mixture of k normal distributions $N(m_j, v_j)$, $(j = 1, \ldots, k)$, is written as

  $$x \sim \sum_{j=1}^{k} w_j N(m_j, v_j),$$

  where each $w_j > 0$ and $\sum_{j=1}^{k} w_j = 1$. The p.d.f. is simply

  $$p(x) = \sum_{j=1}^{k} w_j (2\pi v_j)^{-1/2} \exp(-(x - m_j)^2/(2v_j)).$$

  As k and the (w_j, m_j, v_j) are varied, this can generate densities that are multimodal and skewed, and in fact provide direct analytic approximation to essentially any practically plausible continuous distribution function.

  Here it is of interest to note that it is easy to sample such discrete mixture via convolution: the mixture can be interpreted in terms of a latent (missing, hidden) underlying variable z that "chooses" a mixture component: z is a discrete random variable taking a value from $\{1, 2, \ldots, k\}$ and with $Pr(z = j) = w_j$, $(j = 1, \ldots, k)$. That is, z is a multinomial random quantity with sample size 1 and cell probabilities $w = (w_1, \ldots, w_n)'$, $Z \sim Mn(1, w)$. Then we have the discrete convolutional representation of $p(x)$, i.e.,

  $$p(x) = \sum_{z=1}^{k} p(x|z)p(z)$$

  where $p(x|z) = N(m_z, v_z)$. Simulate a draw from $p(x)$ via (a) draw a value of $z \sim p(z)$, and then (b) sample the implied normal conditional for $(x|z)$.

  As a complete aside, this mixture framework also provides examples of multivariate distributions that are *not* multivariate normal but whose margins are: a mixture of bivariate normals, for example, may

be such that the marginal normal for, say, x_1 in each mixture component is $N(0,1)$, so that the overall margin $p(x_1)$ is also standard normal, whereas of course the full joint distribution is far from normal.

- Mixtures are often induced in statistical analyses by missing data. The above mixture model can be viewed that way, with the implicit latent variable z interpreted as missing data. A simple example in arising with uncertain outcomes in classification testing provides a nice illustration, as well as contacting a number of additional distributional forms.

  A gene variant created during RNA splicing often underlies much of what matters in terms of the function of a gene. A molecular test assess the presence or absence of one of two types of a variant of a specific gene: variant type A or type B. The test reports an outcome $x = 0$ (no variant - the gene is the common or "wild type"), $x = 1$ (type A variant) or $x = 2$ (type B variant). Write $\theta_i = Pr(x = i)$ so that the parameter is $\theta = (\theta_0, \theta_1, \theta_2)'$ subject to $\sum_{i=0}^{2} \theta_i = 1$. Prior information about base rates for gene variants suggests a prior distribution $p(\theta) = Dir(a)$ where $a = (\alpha a_1, \alpha a_2, \alpha a_3)'$ has positive elements; this Dirichlet prior has p.d.f.

$$p(\theta) = c(a)^{-1} \prod_{i=0}^{2} \theta_i^{\alpha a_i - 1} \qquad \text{on} \qquad \sum_{i=0}^{2} \theta_i = 1,$$

  with normalizing constant

$$c(a) = \{\prod_{i=0}^{2} \Gamma(\alpha a_i)\}/\Gamma(\alpha).$$

  The prior mean is a so that $E(\theta_i) = a_i$, and α represents a concentration parameter. The margins for each θ_i are beta priors. The Dirichlet is the natural conjugate prior for the multinomial sampling distribution. For example, if a measurement is made on a new patient, and the test outcome x observed, then

$$p(\theta|x) \propto p(\theta)p(x|\theta) = p(\theta) \prod_{i=0}^{2} \theta_i^{e_i}$$

  where $e_i = I(x = i)$. Thus $(\theta|x) \sim Dir(a + e)$ with $e = (e_0, e_1, e_2)'$.

  Now for a missing data twist: suppose that, on this specific individual, the test is unable to distinguish gene variant A from B, generating the observation $y = \{x \in (1, 2)\}$; the true nature of the variant is hidden, though we learn that it is certainly not the wild type. In this case the data probability is $p(y|\theta) = \theta_1 + \theta_2$ and so the actual posterior is now

$$p(\theta|y) \propto p(\theta)p(y|\theta) = C\{\prod_{i=0}^{2} \theta_i^{\alpha a_i - 1}\}(\theta_1 + \theta_2)$$

  for some (to be computed) normalization constant C. This can be shown (homework exercise) to be a mixture of two Dirichlet distributions,

$$(\theta|y) = w_1 Dir(a + \epsilon_1) + w_2 Dir(a + \epsilon_2)$$

  where $\epsilon_1 = (0, 1, 0)'$ and $\epsilon_2 = (0, 0, 1)'$, and with $w_j = c(a + \epsilon_j)/(c(a + \epsilon_1) + c(a + \epsilon_2))$ for $j = 1, 2$. The mixing reflects the uncertainty about the type of gene variant on the posterior, and learning process, for the underlying mutation rates. Posterior simulation that generates a mixture indicator according to the probabilities w_j then mirrors this uncertainty in simulating relevant θ values.

  As an aside, simulation of Dirichlet distributions is most easy performed via transformation, using the construction of Dirichlet random variables in terms of initial gamma variates. More on this later.

6.8 Monte Carlo Integration

Much of what underlies the use of simulation is the theory underlying Monte Carlo integration. See Robert & Casella (§3.2). Since a primary use is in simulation-based Bayesian computation, we use θ to denote the random quantity of interest; θ will typically represent a vector of parameters, or latent variables, in a statistical model, and the distribution $P(\theta)$ is the posterior distribution based on a model analysis and fit to observed data. The development of Monte Carlo integration is of course quite general and applies to any distribution $P(\theta)$. We work in terms of density functions $p(\theta)$ throughout.

- Interest lies in expectations, typically for many different functions $h(\cdot)$,

$$H = \int h(\theta)p(\theta)d\theta.$$

  Generally includes vector functions, though scalars suffice for most practical purposes (the mean of a vector is the vector of means).
- Random sample $\theta_{1:m}$ where $\theta_i \perp\!\!\!\perp p(\theta)$ for $i = 1, \ldots, m$.
- Monte Carlo approximation:

$$\bar{h} = m^{-1} \sum_{i=1}^{m} h(\theta_i).$$

- *Convergence theories:* Each (transformed random quantity) $h(\theta_i)$ has mean H and, assuming that $E(h^2(\theta)) < \infty$ so that the common variance $\int (h(\theta) - H)^2 p(\theta)d\theta$ of each $h(\theta_i)$ is finite, we have:
  - Law of Large Numbers: ensures almost sure converge of $\bar{h}$ to H;
  - Central Limiting Theorem: $\bar{h} - H$ is asymptotically normal, with asymptotic variance

$$m^{-1} \int (h(\theta) - H)^2 p(\theta)d\theta$$

    that can be approximated by

$$v_m = m^{-2} \sum_{i=1}^{m} (h(\theta_i) - \bar{h})^2.$$

  Hence, taking large Monte Carlo sample sizes m (in the thousands or tens of thousands) can yield very precise, and cheaply computed, numerical approximations to mathematically difficult integrals. The Central Limit Theorem provides effective rule-of-thumb guidelines about precision: for example, an approximate 95% interval estimate associated with the Monte Carlo estimate $\bar{h}$ is $\pm 1.96\sqrt{v_m}$. We refer to $\sqrt{v_m}$ as the (numerical or) Monte Carlo standard error.

6.8.1 A Key Example: Estimating the c.d.f. at a Point

One example takes $h(\theta) = I(\theta \leq x)$ for any x, so that H is the value of the c.d.f. $P(x)$. Then

- $\bar{h}$ is the proportion of the sample values $\theta_{1:n}$ that are below x.
- Each event $\{\theta_i < x\}$ has probability H, and independence implies that $m\bar{h} \sim Bin(m, H)$.
- In this case, $V(\bar{h}) = H(1 - H)/m \approx \bar{h}(1 - \bar{h})/m$. (It is easily checked that this standard binomial variance approximation coincides with the general sample variance formula for v_m above.) Under the normal central limit, a 95% interval estimate of H is then $\bar{h} \pm 1.96\sqrt{\bar{h}(1 - \bar{h})/m}$.

For example, in the AR(1) model we are interested in $Pr(|\phi| < 1) = P(1) - P(-1)$ where $P(\cdot)$ is the posterior c.d.f. for ϕ. Here, then, we want to use Monte Carlo integration to estimate $P(1)$ and $P(-1)$ so as to approximately evaluate the posterior probability that the AR process is stationary.

6.8.2 Histograms of Monte Carlo Draws

Histograms of $\theta_{1:m}$ represent Monte Carlo approximations to $p(\theta)$. The above theory provides access to uncertainty assessments. It is very common to use complex simulation methods to generate large Monte Carlo samples and to then use those samples in creating histograms as well as computing approximate expectations for various functions of interest - Monte Carlo posterior means $(h(\theta) = \theta)$, various posterior probabilities $(h(\theta)$ is an indicator function on some set or range of values of θ,) and more complex transformations of what may be a high-dimensional parameter θ.

6.9 Importance Sampling

Importance sampling is one of the first steps into Monte Carlo analysis in which simulated variates from one distribution are used to explore another - simulation from the "wrong distribution" can be extraordinarily useful. Rejection sampling is another such method, and Metropolis MCMC methods define the encompassing framework. Currently, importance sampling is still of practical interest in

- fairly small problems, in terms of dimension,
- in which the density of the distribution of interest can be easily evaluated, but when it is difficult to sample from directly, and
- when it is relatively easy to identify and simulate from distributions that approximate the distribution of interest.

The core example is Bayesian inference on a parameter θ and, as earlier, the distribution $P(\theta)$ is the posterior distribution based on a model analysis and fit to observed data. We work in terms of density functions $p(\theta)$ throughout. Here, then, we must be able to evaluate $p(\theta)$ as a function at any point, and we assume that we have available an *importance sampling distribution* with p.d.f. $g(\theta)$. Two key requirements are that (a) $g(\cdot)$ is easy to sample from, and (b) the p.d.f. $g(\theta)$ is easy to evaluate at any point, as for $p(\theta)$. Often, the context is one in which $g(\theta)$ has been derived as an analytic approximation to $p(\theta)$, and the closer the approximation, the more accurate the resulting importance sampling MC analysis will be. However, sometimes $g(\theta)$ is just a convenient density that is particularly easy to simulate and evaluate, and in such cases we may easily generate very large Monte Carlo samples to increase the accuracy even though $g(\theta)$ may be a rather crude approximation to $p(\theta)$.

The underlying idea is simple:

- Interest lies in expectations of the form

$$H = \int h(\theta)p(\theta)d\theta.$$

- We can write

$$H = \int h(\theta)w(\theta)g(\theta)d\theta \qquad \text{with} \qquad w(\theta) = p(\theta)/g(\theta).$$

  This shows that the expectation of $h(\theta)$ under $p(\theta)$ is just that of $h(\theta)w(\theta)$ under $g(\theta)$.
- Using direct Monte Carlo integration,

$$\bar{h} = m^{-1}\sum_{i=1}^{m} w(\theta_i)h(\theta_i),$$

  where we draw the random sample $\theta_i \perp\!\!\!\perp g(\theta)$ for $i = 1, \ldots, m$. We are sampling from the "wrong" distribution.

- The measure "how wrong" at each simulated θ_i value is the *importance weight*

$$w(\theta_i) = p(\theta_i)/g(\theta_i).$$

These ratios "weight" the sample estimates $h(\theta_i)$ to "correct" for the fact that we sampled the wrong distribution.

See Robert and Casella §3.3 for length discussion of optimality questions and convergence (of $\bar{h}$ to H) and the association limit theorems. The focus on computing specific expectations means that choices of g will depend on h to deliver best results. However, in much statistical work, and especially this context of a posterior analysis, we are interested in many possible expectations, and generally view the Monte Carlo analysis as one in which $g(\theta)$ will be chosen to ensure good MC accuracy for various posterior characteristics.

Critical considerations (see Robert and Casella §3.3):

- MC estimate $\bar{h}$ has the expectation H, and is generally almost surely convergent to H under conditions below.
- $Var(\bar{h}) = m^{-1} \int h^2(\theta)p^2(\theta)/g(\theta)d\theta - H^2$. For ranges of functions h, these variances are going to be finite for cases in which, generally, $w(\theta) = p(\theta)/g(\theta)$ is bounded and decays rapidly in the tails of the target $p(\theta)$. Smaller variance, and hence superior MC approximations, are achieved for densities whose tails dominate those of the target.
- The last feature means that importance sampling distributions should be chosen to have tails at least as fat as the target - common then to consider distributions such as T distributions, that decay like reciprocal powers of θ, when the target has exponential decay such as is exhibited in normal and other exponential family models. Theoretical investigation of the tail decay of $p(\theta)$ will help guide thinking about importance sampling distributions.
- Obviously require the support of $g(\theta)$ to be the same as, or contain, that of $p(\theta)$.

Problems in which $w(\theta)$ can be computed are rare in statistics. Quite commonly, we know $p(\theta)$ only up to a constant of normalization - especially true in Bayesian analysis using Bayes' Theorem. Then we renormalise the importance weights:

$$\hat{h} = \sum_{i=1}^{m} w_i h(\theta_i) \qquad \text{where} \qquad w_i = w(\theta_i)/\sum_{j=1}^{m} w(\theta_i).$$

- This is "as if" we are using a discrete distribution to approximate $p(\theta)$: that distribution with point masses (=probabilities) w_i at the points θ_i.
- Good discussion of the convergence of $\hat{h}$, and associated asymptotic (in m) theory, is given in Geweke (1989, *Econometrica*, **97**, pp1317-1339). In particular, $\hat{h} - H$ is asymptotically normal under broad conditions, basically those noted above on the relative tail weight matter. The asymptotic numerical or Monte Carlo standard error for guiding thinking about the accuracy of specific approximations is σ_m where

$$\sigma_m^2 = (mW_m)^{-1} \sum_{i=1}^{m} (h(\theta_i) - \hat{h})^2 w(\theta_i)^2 \qquad \text{with} \qquad W_m = \sum_{i=1}^{m} w(\theta_i)^2.$$

Exploring the empirical distribution of the sampled weights w_i guides understanding of how well, or how poorly, a particular importance sampling approximation may be. Weights that are close to uniform are desirable, and very unevenly distributed weights are not; it is easy to generate examples in which one or a very small number of weights are very large and dominate the rest, due to poor choice of $g(\theta)$.

See Robert and Casella for several examples. Some others:

- Supposes $p(\theta) \propto p_0(\theta) I(\theta \in A)$ where p_0 is the p.d.f. of a known, standard distribution, such as a multivariate T distribution, and where A is some specifed set of values. For example, a univariate T distribution for θ conditional on $\theta > m$ for some specified constant m. In this case an obvious importance sampling distribution is $g(\theta) = p_0(\theta)$. Then $g(\theta)/p(\theta) = 1$ if $\theta \in A$ and zero otherwise. This may be very efficient, and will be when the target distribution concentrates heavily on the set A.
- The above example and that below are instances of the following setup. Suppose that we know the form of $p(\theta) = a(\theta)g(\theta)$ where $g(\theta)$ is a p.d.f. with the desirable characteristics for an importance sampler, and $a(\theta)$ is some easily evaluated function, perhaps a contribution to the likelihood for θ from a statistical model under with $p(\theta)$ is the posterior density - but a factor that complicates the posterior analysis. Then, we may use "part" of the posterior for simulation, and correct based on the weight $w(\theta) \propto a(\theta)$.
- In the simple AR(1) model we have a posterior $p(\phi, v|x_{1:n})$ that has a slightly complicated mathematical form and cannot be directly sampled. We have, however, the conditional normal/inverse gamma posterior for (ϕ, v) arising from the AR model conditioning on x_1 and ignoring the contribution to the posterior from that initial observation; this conditional posterior is easily sampled and evaluated.

  See the worked Matlab example on this interesting application of importance sampling.

6.9.1 Resampling for Inference and Prediction

- Treating the importance weighted samples as a discrete approximation to the posterior $p(\theta)$ allows for MC estimation of integrals that include, for example, aspects of predictive distributions. For predicting a "future" random quantity y based on a model $p(y|\theta)$, we use the predictive density

$$p(y) = \int p(y|\theta)p(\theta)d\theta.$$

  Forecasting the future of a time series is one example. In such contexts, predictive expectations follow from the importance sampler via integration; e.g.; $E(y) = \int h(\theta)p(\theta)d\theta$ where $h(\theta) = E(y|\theta)$.
- Often we want to simulate predictive distributions. This is easily done via composition, treating the importance sampling weights as a discrete probability distribution on the sampled θ_i values:
  - Sample from the set $\{\theta_i\}$ according to the multinomial distribution defined by weights $\{w_i\}$; then
  - Sample $y \sim p(y|\theta_i)$.
- Resampling is a device sometimes used to convert the discrete distribution into a uniform discrete distribution, so generating an approximate random sample from $p(\theta)$, although with duplicate values. This is useful when we can easily generate very large MC importance samples, and then resample to "flatten out" the weights and produce an equally weighted version, then use simple Monte Carlo approximations without worrying about weights. This has the feature that points θ_i with very low weights will be disregarded, and is useful generally when the importance sampler is effective and generates well-balanced weights in the first place.

See the worked Matlab example on resampling in the AR(1) model analysis.

6.10 Simple Accept/Reject

Simple accept/reject sampling also uses an importance sampling function and the implied importance ratio to correct samples from the wrong distribution. This method is very widely used in simple random variate generation, especially univariate distributions. It utilizes the same importance ratio ideas, but leads to *exact* corrections and so exact samples from $p(\theta)$. The use nowadays is restricted to univariate or low dimensional problems where it can be the most efficient approach, but in even modestly complicated distributions it can he difficult to implement, since it involves knowledge of an upper bound on the importance ratio. It is particularly hard to implement when the target distribution of interest is known only up to a constant of normalization, the common situation in posterior analysis, and in higher dimensions. More recently introduced and advanced methods, including so-called slice-sampling and related methods, offer some real practical advantages in, again, lower dimensional and mathematically relatively tractable examples. However, the principles underlying accept/reject are, as with importance sampling, simply critical principles in moving to more comprehensive approaches involving accept/reject methods in MCMC algorithms.

- Assume that $w(\theta) = p(\theta)/g(\theta) < M$ for some constant M. With normalized densities we must have $M > 1$, so that $1/M < 1$. If $g(\theta)$ represents a "good" potential importance sampler, then $w(\theta)$ will decay in the tails rapidly, and if it is a "good" approximation to $p(\theta)$ then M should not be too far from 1.
- Generate a *candidate* value $\theta \sim g(\theta)$. If $w(\theta)$ is large, then the candidate seems to represent $p(\theta)$; otherwise, θ is an unlikely value under $p(\theta)$.

The accept/reject theory formalizes this:

1. Accept θ with probability $w(\theta)/M$: if accepted, it is a draw from $p(\theta)$; otherwise *reject* and try again.
2. Equivalently, and operationally, generate $u \sim U(0,1)$ independently of θ. Then *accept* θ as a draw from $p(\theta)$ if, and only if, $u < w(\theta)/M$.

The theory underlying this is simple. We look at the distribution of a random quantity θ that is generated by this algorithm. We need to show simply that the distribution of θ conditional on it having been accepted, i.e., conditional on $u < w(\theta)/M$, is in fact $P(\theta)$. By Bayes' theorem, the p.d.f. of $(\theta|u < w(\theta)/M)$ is just

$$f(\theta|u < w(\theta)/M) = \frac{g(\theta)Pr(u < w(\theta)/M|\theta)}{Pr(u < w(\theta)/M)}.$$

Now

- $Pr(u < w(\theta)/M|\theta) = w(\theta)/M$ since $u \sim U(0,1)$, and
- $Pr(u < w(\theta)/M) = \int Pr(u < w(\theta)/M|\theta)g(\theta)d\theta = \int w(\theta)g(\theta)d\theta/M = \int p(\theta)d\theta/M = 1/M$.

Thus

$$f(\theta|u < w(\theta)/M) = g(\theta)w(\theta) = p(\theta)$$

as required: accepted values have p.d.f. $p(\theta)$.

Rejected values are ... rejects. One key question is the efficiency, based on the proportion of rejects. If we want a sample of size $10,000$ and have to simulated millions due to a high rejection rate, we will probably try another approach. Implicit in the proof of the accept/reject theory above is the fact that $Pr(u < w(\theta)/M) = 1/M$. That is, a pair (θ, u) generated this way has a probability $1/M$ of acceptance. An importance sampler $g(\theta)$ such that M is close to 1 is, naturally, and efficient accept/reject sampler and leads to a high probability of acceptance.

- Read the key example of generating $Ga(a, 1)$ distributions for any positive value of a based on samples from the $Ga(\lfloor a \rfloor, 1)$ where $\lfloor a \rfloor$ is the integer part of a (Robert & Casella, example 2.3.4).
- *Normal from Cauchy:* One simple and illustrative example is simulating the normal from the Cauchy. Take $p(\theta)$ to be standard normal and $g(\theta) = \{\sqrt{s}\pi(1 + x^2/s)\}^{-1}$. The Cauchy is fatter tailed than the normal so we know the importance ratio will decay fast, and that it should deliver an effective importance sampling approach, so it should be of use for rejection too. In this example, you can easily show that $w(\theta) < M_s$ where $M_s = \sqrt{(2\pi/s)}\exp(s/2 - 1)$. It is then trivial to also show that M is minimized at $s = 1$, with $M_1 = \sqrt{(2\pi/e)}$, and so $1/M_1 \approx 0.66$, so about a 2 in 3 acceptance probability.

  See the worked Matlab example.

7 Elements of Markov Chain Structure and Convergence

7.1 Definition

- Random quantities x in a $p-$dimensional state space χ are generated sequentially according to a conditional or *transition* distribution $P(x|x')$, sometimes referred to as a transition kernel (and with several notations), that is defined for all $x, x' \in \chi$. The sequence of quantities $x^{(1)}, x^{(2)}, \ldots, x^{(t)}, \ldots$, is generated from some initial value $x^{(0)}$ via

$$x^{(t)} \sim P(x|x^{(t-1)}) \quad \text{with} \quad x^{(t)} \perp\!\!\!\perp x^{(t-k)}|x^{(t-1)}, \ (k > 1). \tag{19}$$

  The sequence $\{x^{(t)}\}$ is a first-order Markov process, or Markov chain, on the state space χ.
- Work in terms of density functions: whether the state space is discrete, continuous or mixed (as is often the case in statistical modelling applications), use $p(x|x')$ to denote the density of $P(x|x')$, assuming the p.d.f. to be well-defined and unique. Then the Markov process is generated from the initial value $x^{(0)}$ by convolutional sampling, i.e., sequencing through the conditionals

$$x^{(t)} \sim p(x|x^{(t-1)})$$

  for all $t > 0$.
- In this set-up, the Markov process is homogenous: for all t, transitions from $x^{(t-1)}$ to $x^{(t)}$ are governed by the same conditional distribution $P(\cdot|\cdot)$. More general, non-homogenous Markov processes are also of interest, and quite frequently used in statistical work too: that is, processes defined by transition densities $p_t(x|x')$ at step t. Restrict attention here, however, to homogenous processes as much of the use of simulation methodology in MCMC applications involves homogenous processes.

7.2 Transitions and Ergodocity

Markov processes define sequences of conditional distributions for $k-$step ahead transitions via convolutions. In the homogenous Markov chain, we have transition densities simply induced by the specified $1-$step transition density. Define the $k-$step transition p.d.f. $p^k(x|x')$, for each $k \geq 1$, as the conditional density of the state $x = x^{(t+k)}$ given the state $x' = x^{(t)}$ for any time t. Then $p^k(\cdot|\cdot)$ is just the $k-$fold convolution of the transition p.d.f. $p(\cdot|\cdot)$. In detail,

- $p^1(x^{(t+1)}|x^{(t)}) \equiv p(x^{(t+1)}|x^{(t)})$;
- $p^2(x^{(t+2)}|x^{(t)}) = \int p(x^{(t+2)}|x^{(t+1)})p(x^{(t+1)}|x^{(t)})dx^{(t+1)}$;
- $p^3(x^{(t+3)}|x^{(t)}) = \int p(x^{(t+3)}|x^{(t+2)})p^2(x^{(t+2)}|x^{(t)})dx^{(t+2)}$,

and so on with, in general,

- $p^k(x^{(t+k)}|x^{(t)}) = \int p(x^{(t+k)}|x^{(t+k-1)})p^{k-1}(x^{(t+k-1)}|x^{(t)})dx^{(t+k-1)}$.

These densities describe the probability structure governing the evolution of the process through time steps as a function of the core $1-$step transitions. Notice that the process will be free to "wander" anywhere in the state space - with positive density on *any* values $x^{(t+k)} \in \chi$ based on *any* starting value $x^{(t)}$ - if the $k-$step density function $p^k(x|x')$ is positive for all $x, x' \in \chi$. This will be ensured, as a sufficient condition and a key example, when $p(x|x') > 0$ for all $x, x' \in \chi$, for example. Processes with this property are generally well-behaved in the sense of defining stationary processes, under additional (weak) assumptions.

The Markov process is *irreducible* if, starting at any initial state x', the process can reach any other state $x \in \chi$. Irreducibility is formally defined in terms of the $k-$step transition probabilities on subsets of χ :

for any two subsets $A \subseteq \chi$ and $A' \subseteq \chi$, the chain is irreducible if the probability of reaching any state $x \in A$ starting from any state $x' \in A'$ is positive under the $k-$step transition distribution for some finite k. Loosely speaking, starting at any state we have positive probability of reaching any other state. In terms of densities, it suffices to require that $p(x|x') > 0$ for any $x, x' \in \chi$, and this is often the case in MCMC. This also ensures that the Markov process is *recurrent*, meaning that states $x \in A \subseteq \chi$ will have positive density $p^k(x|x')$ for all k as $k \to \infty$; if we could run the chain for ever, any such set A would be revisited infinitely often, i.e., would recur. The process is *aperiodic* if may visit any state $x \in A$ from any $x' \in A'$ in one step, which is a stronger condition and one that is again implied if $p(x|x') > 0$ The practical relevance of irreducibility and aperiodicity is that chains with these properties are convergent: an irreducible, aperiodic Markov process is called an *ergodic* process and defines the main class of stationary, convergence processes of interest in MCMC.

7.3 Stationary Distributions and Convergence

The Markov process has a *stationary* or *invariant* distribution $\Pi(x)$, with density $\pi(x)$, if

$$\pi(x) = \int p(x|x')\pi(x')dx'.$$

More formally in terms of distribution functions, $P(x) = \int P(x|x')d\Pi(x')$.

In statistical applications of MCMC, we generally design Markov chain generating processes with a specific stationary distribution in mind from the context of the statistical model - in a Bayesian posterior analysis, the stationary distribution will be a target posterior distribution of interest. Note the implication that, if the "current state" x' of the chain is known to be distributed according to $\pi(x')$, then the transition to the next state maintains that marginal distribution, in that x is also distributed according to $\pi(x)$. The joint density is $p(x|x')\pi(x')$.

If we know that the process is ergodic (irreducible and aperiodic), then $\pi(x)$ is the *unique* stationary distribution. In such processes:

- $p^k(x|x') \to \pi(x)$ as $k \to \infty$ from any initial value x', so that the process converges in the sense that, looking ahead from any initial state x', the probability distribution on future states x looks more and more like the stationary distribution $\pi(x)$ as k increases. The initial value is eventually "forgotten", and samples $x^{(k)}$ from such a chain will eventually resemble draws (though dependent draws) from $\pi(x) : x^{(k)} \sim \pi(x)$ as $k \to \infty$.
- We say that the Markov process or chain converges to $\pi(x)$.
- Ergodic averages of process values also converge (almost surely) to their limiting expectations under $\pi(x)$, i.e., for functions $h(x)$ that are integrable with respect to $\pi(x)$,

$$m^{-1}\sum_{t=1}^{m} h(x^{(t)}) \to \int h(x)\pi(x)dx$$

  almost surely as $m \to \infty$. This justifies the use of samples from a long Markov chain simulation as building blocks for histograms and sample estimates of features of the stationary distribution. Chains that are initialized at values not drawn from the stationary distribution will often be "run" for an initial period of time until, assumedly, convergence, and sampled values, their averages and so forth then computed after discarding the initial "burn-in" series.
- If $x^{(0)} \sim \pi(\cdot)$ then, for all $t > 0$, $x^{(t)} \sim \pi(\cdot)$; if the chain starts (or, at any time point, is restarted) with a draw from the stationary distribution, then it is already "converged" and all subsequent states come from $\pi(x)$ too.

7.4 Reversibility and Detailed Balance

Some Markov processes are reversible, in that the transitions "backwards" in time are governed by the same distribution $p(x|x')$ as the "forward" transitions. This is true of many practicable MCMC methods in statistical analysis. Clearly the chain is Markov in reverse time. In an ergodic chain with stationary density $\pi(x)$, the "backward transition" density from a current state $x^{(t)}$ is available directly from Bayes' theorem; denoting the density at time t by $p^r(x'|x)$, (r stands for *reversed*) we have

$$p^r(x^{(t-1)}|x^{(t)}) = p(x^{(t-1)})p(x^{(t)}|x^{(t-1)})/p(x^{(t)})$$

where $p(x^{(s)})$ denotes the density of the state at time s. In general this backward transition depends on t and so the backward chain is non-homogenous. However, if the forward chain is assumed to have converged to its stationary distribution, or has been initialized to ensure that, then $p(x^{(s)}) = \pi(x^{(s)})$ and the 1-step backward transition p.d.f. is

$$p^r(x^{(t-1)}|x^{(t)}) = \pi(x^{(t-1)})p(x^{(t)}|x^{(t-1)})/\pi(x^{(t)}).$$

The "backwards" process is then also homogenous, i.e., is given for all t by the transition p.d.f.

$$p^r(x'|x) = \pi(x')p(x|x')/\pi(x).$$

Sometimes we have $p^r(x'|x) \equiv p(x'|x)$, and the Markov process is said to be *reversible*. In such cases, we see that

$$p(x'|x)\pi(x) = p(x|x')\pi(x'),$$

an equation referred to historically as *detailed balance.* In this case, the above identity defines the symmetric joint density of two consecutive states x, x', in either forward or reverse directions, in the stationary process.

Some useful MCMC processes are not reversible, though some of the focus in development of MCMC methods is to define rapidly convergent, reversible processes with specified "target" stationary distributions.

7.5 Revisiting Examples

- The simple AR(1) process $x_t \leftarrow AR(1|\theta)$ is an example of an ergodic, reversible Markov process with state space the real line, when $|\phi| < 1$. Transitions are governed by $(x|x') \sim N(\phi x', v)$ and the stationary distribution is $x \sim N(0, s)$ with $s = v/(1-\phi^2)$. The convergence rate is geometric: like $|\phi^k|$ as $k \to \infty$.
- Consider the simple Cauchy AR(1) process defined by $x_t = \phi x_{t-1} + \epsilon_t$ where $\epsilon_t \sim C(0, v)$ with $|\phi| < 1$ and $v > 0$. Transitions are defined by $p(x|x') \propto \{1 + (x - x')^2/v\}^{-1}$ and the process is ergodic with some stationary density $\pi(x)$ satisfying

  $$\pi(x) \propto \int_{-\infty}^{\infty} \pi(x')\{1 + (x - x')^2/v\}^{-1} dx'.$$

  This is not an easy equation to solve for $\pi(x)$. However, the model is trivially amenable to solution using characteristic functions (Fourier transforms), as follows.

  The characteristic function (c.f.) of any random variable z is $f(s) = E(\exp(isz))$ and is unique. We know, for example, that the c.f. of $C(0, v)$ is the function $\exp(-\sqrt{v}|s|)$. Now, from $x_t = \phi x_{t-1} + \epsilon_t$ and the independence of x_{t-1} and ϵ_t we have

  $$E(\exp(isx_t)) = E(\exp(is\phi x_{t-1}))E(\exp(is\epsilon_t)),$$

or

$$f_\pi(s) = f_\pi(\phi s)\exp(-\sqrt{v}|s|),$$

where $f_\pi(s)$ is the characteristic function of the stationary distribution. We can now simply observe that, the choice $f_\pi(s) = \exp(-\sqrt{r}|s|)$ is a solution if, and only if, $\sqrt{r} = \sqrt{v}/(1-|\phi|)$. Since the c.f. of a distribution is unique, this must be the only solution, and note that it corresponds to the $C(0,r)$ distribution; that is, we have shown that

$$\pi(x) \propto \{1 + x^2/r\}^{-1}$$

where $r = v/(1-|\phi|)^2$.

The comparison with the normal model is of interest, and we note that there are generalizations in which ϵ_t follows a *stable distribution* - distributions whose tailweights span the range between the normal (light) and Cauchy (very heavy). These are distributions in which the c.f. is $\exp(-\sqrt{v}|s|^\alpha)$ for some index $1 \leq \alpha \leq 2$.

We have a suspicion - from empirical exploration of sample paths generated from such models - that the process is not reversible, and it is now demonstrably true: $p^r(x'|x) = \pi(x')p(x|x')/\pi(x)$ does not reduce to the forward transition p.d.f. $p(x'|x)$. In fact, among all linear processes, of which the AR(1) is the simplest case, only normal innovations lead to reversibility. On the other hand, many reversible non-linear processes arise routinely in MCMC (though many MCMC-derived processes are not reversible).

8 Introduction to MCMC via Gibbs Sampling

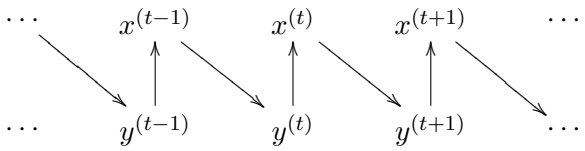
Markov chain Monte Carlo (MCMC) methods simulate dependent sequences of random variables with the goal of generating samples from a stationary Markov process, so that - ultimately - such values represent draws from the stationary distribution of the Markov process. MCMC algorithms are therefore created with a specific, *target* stationary distribution in mind, and the methodology focuses on how to develop specific Markov chains with that stationary distribution. Explicitly, such analysis generate dependent samples, not independent draws, and some attention there rests on the theory and methods of stationary processes and time series analysis to investigate such methods.

8.1 A Simple Example

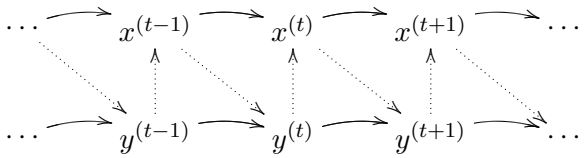
Suppose (x, y) has a bivariate normal distribution, with $x \sim N(0,1)$, $y \sim N(0,1)$ and $C(x,y) = \rho$. The following algorithm generates a sequence of values of x, namely $x^{(1)}, x^{(2)}, \ldots, x^{(t)}, \ldots$: Write $r = 1-\rho^2$ for the conditional variance of each of the conditional distributions, $(x|y) \sim N(\rho y, r)$ and $(y|x) \sim N(\rho x, r)$.

- Choose any value $x = x^{(0)}$.
- For $t = 1, 2, \ldots,$ repeat:
  - Sample a value $y = y^{(t)}$ from the conditional distribution $(y|x)$ at the "current" value of $x = x^{(t-1)}$; that is, generate a draw $y^{(t)}|x^{(t-1)} \sim N(\rho x^{(t-1)}, r)$.
  - Sample $x = x^{(t)}$ from the univariate conditional $p(x|y)$ at the "current" value of $y = y^{(t)}$; that is, draw $x^{(t)}|y^{(t)} \sim N(\rho y^{(t)}, r)$.
  - Save the value of $x^{(t)}$ and continue.

Graphically, the construction is illustrated in this graph, with the conditional independence implicit in the lack of arrows between variables:



The second version of this shows the dependencies induced in the $x^{(t)}$ sequence on marginalization over the $y^{(t)}$ sequence - the dashed arrows now reflect the original conditional dependencies removed by this marginalization. The setup is also obviously symmetric, with the same Markovian dependence structure evident in the $y^{(t)}$ sequence when marginalizing out the $x^{(t)}$ sequence.



This generates a sequence $x^{(1:n)}$ where we stop the process after $t = n$ steps. Some facts:

1. If $x^{(0)}$ is chosen randomly as $x^{(0)} \sim N(0,1)$, then $y^{(1)} \sim N(0,1)$ by composition, and clearly $(x^{(0)}, y^{(0)})$ is a draw from the bivariate normal of interest. Progressively then, for all t,

   - $x^{(t)} \sim N(0,1)$,
   - $y^{(t)} \sim N(0,1)$.
   - $(x^{(t)}, y^{(t)})$ is a draw from the bivariate joint normal $p(x,y)$.
   - For all $t > 1$, $x^{(t)} \perp\!\!\!\perp x^{(t-1)} | y^{(t)}$.
   - For $t > 0$, $x^{(t)} = \rho y^{(t)} + \epsilon^{(t)}$ with $\epsilon^{(t)} \sim N(0, r)$ independently, and $y^{(t)} = \rho x^{(t-1)} + \eta^{(t)}$ with $\eta^{(t)} \sim N(0, r)$ independently, and with $\epsilon^{(t)} \perp\!\!\!\perp \eta^{(s)}$.
   - So $(x^{(t)} | x^{(t-1)}) \sim N(\rho^2 x^{(t-1)}, r(1+\rho^2))$, or $N(\phi x^{(t-1)}, v)$ with $\phi = \rho^2$ and $v = r(1+\rho^2) = 1 - \phi^2$.
   - The $\{x^{(t)}\}$ process is first-order Markov, $x^{(t)} \perp\!\!\!\perp x^{(t-k)} | x^{(t-1)}$ for $k > 1$, and a linear, Gaussian (and time reversible) Markov process.
   - $x^{(t)} \leftarrow AR(1|\theta)$ with $\theta = (\phi, v)$ where $\phi = \rho^2$ and $v = (1 - \phi^2)$.
   - The stationary distribution of the Markov chain is the margin $x^{(t)} \sim N(0,1)$, and the sequence of values $x^{(1:n)}$ generated is such that each $x^{(t)}$ comes from this stationary distribution, but they are not independent: $C(x^{(t)}, x^{(t-k)}) = \phi^k = \rho^{2k}$.
   - Unless ϕ is very close to $\phi = 1$, these correlations decay fast with k, so that samples some time units apart will have weaker and weaker correlations. In a long sequence, selecting out a subsequence at a given spacing of some k units, e.g., $\{x^{(0)}, x^{(k)}, x^{(2k)}, \ldots, x^{(mk)}\}$ will generate a set of m draws that represent an approximately independent $N(0,1)$ sample if k is large enough so that ρ^{2k} is "negligible".

2. If $x^{(0)}$ is chosen as a draw from some other distribution, or just fixed as some arbitrary value, then the sample path $x^{(1:n)}$ starts in a region of the x space that may be far from the stationary $N(0,1)$ distribution. Under the stationary model implied, the sample path will converge to stationarity, "forgetting" the arbitrary initialization at a rate ρ^{2t}. For larger t, $x^{(t)}$ will be closer to sequential draws from the stationary distribution, and convergence will be faster for smaller values of $|\rho|$.

   For example, the time taken for the correlation at lag k to decay to a value c is $\log(c)/(2\log(\rho))$. Some values indicate the implied "convergence" times, rounded up to the nearest integer:

	$\rho = 0.7$	$\rho = 0.9$	$\rho = 0.95$	$\rho = 0.99$
$c = 0.10$	4	11	23	115
$c = 0.05$	5	15	30	150
$c = 0.01$	7	22	45	230

3. Whether exactly (starting from an $x^{(0)} \sim N(0, 1),$) or after the convergence to the stationary distribution (starting at whatever other initial value is chosen) each pair $(x^{(t)}, y^{(t)})$ represents a draw from the joint normal distribution. Again, the draws are serially correlated.

This example serves to introduce the key idea of MCMC: sampling using a Markov chain process. It also introduces a central and most important class of MCMC methods - Gibbs sampling methods, in which the chain is generated by *iterative* resampling from the conditional distributions of a target joint distribution. It is simultaneously a method to simulate (correlated) samples from $p(x, y)$, and - then as a corollary - from the margins $p(x)$ and of course $p(y)$.

8.2 Examples of Bivariate Gibbs Sampling

The above example is a simple example of Gibbs sampling MCMC for a bivariate distribution. Generally, a joint density $p(x, y)$ implies the set of *complete conditional densities* $\{p(x|y), p(y|x)\}$. Gibbs sampling iteratively resamples these conditionals to generate a Markovian sequence of pairs $(x^{(t)}, y^{(t)})$ that, under very general conditions, (eventually) represents a stationary bivariate process that converges to the stationary distribution $p(x, y)$ of the implied Markov chain, in the sense that, as t increases, $(x^{(t)}, y^{(t)}) \sim p(x, y)$. Before going further with the theory and generalizations to more than two dimensions, some additional examples help fix ideas.

Example; Normal-inverse gamma: Suppose $(x|\lambda) \sim N(m, s/\lambda)$ and $\lambda \sim Ga(k/2, kh/2)$. We know that x has a marginal T distribution. The complete conditionals are defined by $p(x|\lambda)$ above and, as is easily checked, $(\lambda|x) \sim Ga((k + 1)/2, (kh + x^2/s)/2)$. This construction is very broadly used in models where, for example, observed data x has an uncertain scale factor λ, as we know arises in normal mixture models such as T and Cauchy distributions.

Example: Beta-binomial with uncertain N: We make a binomial observation $(y|N, \beta) \sim Bin(N, \beta)$ with a uniform prior on β, but that are uncertain about the binomial total and describe that uncertainty via a Poisson distribution, $N \sim Po(m)$, with $N \perp\!\!\!\perp \beta$. We have the joint posterior $p(\beta, N|y)$ under which the complete conditionals are

- the standard beta posterior $(\beta|N, y) \sim Be(1 + y, 1 + N - y)$, and
- $p(N|\beta, y) \propto p(N)p(y|n, \beta) = \text{constant}\{m(1 - \beta)\}^{N-y}/(N - y)!$, for $N \geq y$. This implies that $N = y + n$ where $n \sim Po(m(1 - \beta))$, with full conditional density

$$p(N|\beta, y) = \{m(1 - \beta)\}^{N-y} \exp(-m(1 - \beta))/(N - y)!, \qquad (N \geq y).$$

This is an example of a bivariate posterior density in a Bayesian analysis, and one in which Gibbs sampling applies trivially to simulate sequences $(\beta^{(t)}, N^{(t)})$ that will converge to samples from the joint posterior $p(\beta, N|y)$. It is also an example where the state space of the implied Markov chain is both discrete (N) and continuous (β), whereas all examples so far have been continuous.

8.3 Bivariate Gibbs Sampling in General

In the general bivariate case, Gibbs sampling sequences through the simulations as detailed in the normal example above: for $t = 1, 2, \ldots$, a draw from the $(y|x)$ conditional simulates $(y^{(t)}|x^{(t-1)})$, then a draw from the $(x|y)$ conditional simulates a new value of $(x^{(t)}|y^{(t)})$, and the process continues. We need a slightly modified notation for clarity in the theoretical development. Write the complete conditionals now as

$$\{p_{x|y}(\cdot|\cdot), p_{y|x}(\cdot|\cdot)\}.$$

Then Gibbs sampling generates, for all $t \geq 1$, from

- $y^{(t)} \sim p_{y|x}(\cdot|x^{(t-1)})$, then
- $x^{(t)} \sim p_{x|y}(\cdot|y^{(t)})$, and so on.

This defines a first-order Markov process on the (x, y) space jointly, but also univariate Markov processes on x and y individually. Consider the simulated x process $x^{(1:n)} = \{x^{(1)}, \ldots, x^{(n)}\}$ arising. We can see that the one-step transition p.d.f. is just

$$p(x^{(t)}|x^{(t-1)}) = \int p_{x|y}(x^{(t)}|u)p_{y|x}(u|x^{(t-1)})du.$$

Under very weak conditions, which are basically ensured in cases when the complete conditionals are compatible with a joint and the transition p.d.f. above is positive for all $x^{(t-1)}, x^{(t)}$, the resulting chain will represent a stationary process with limiting stationary distribution given by $p(x) = \int p(x, y)dy$. Convergence to stationarity from an arbitrary initial value $x^{(0)}$ will generally be hastened if the initial value can be chosen or simulated from a distribution close to the marginal $p(x)$, of course, as in the bivariate normal example.

Evidently transitions in the joint space are governed by the bivariate transition p.d.f.

$$p((x^{(t)}, y^{(t)})|(x^{(t-1)}, y^{(t-1)})) = p_{x|y}(x^{(t)}|y^{(t)})p_{y|x}(y^{(t)}|x^{(t-1)}).$$

8.4 Complete Conditional Specification in Bivariate Distributions

- Continue with the simple bivariate context and working in terms of density functions. Gibbs sampling in statistical analysis is commonly used to explore and simulation posterior distributions, so that the generic $p(x, y)$ is then a posterior density for (x, y) from some statistical model, often defined via Bayes's theorem. The complete conditionals $\{p(x|y), p(y|x)\}$ can then be just "read off" by inspection of $p(x, y)$, and all is well so long as this is a proper, joint integrable density function. Note that, as will often be the case, the joint density may be known only up to a constant of normalization.
- Under broadly applicable conditions, a joint density is uniquely defined by the specification of the complete conditionals (Theorem 7.1.19 of Robert and Casella). Given $p(x|y)$ and $p(y|x)$, we can deduce

  $$p(y) \int \frac{p(x|y)}{p(y|x)} dx = 1$$

  so that $p(y)$ is defined directly by the reciprocal of the integral of ratio of conditionals. A similar result exists for $p(x)$. Hence the conditionals define $p(y)$ (and/or $p(x)$) and the joint density is deduced: $p(x, y) = p(x|y)p(y)$. The requirement is simply that the integral here exists and that its reciprocal defines a proper density function for $p(y)$, with a parallel condition for the corresponding integral defining $p(x)$.

- *An example:* take $(x|y) \sim N(\rho y, (1-\rho^2)v)$ and $(y|x) \sim N(\rho x, (1-\rho^2)w)$ where $|\rho| < 1$ and for any variances v, w. It can be easily (if a bit tediously) verified that the integral above leads to the margin $y \sim N(0, w)$, the corresponding margin $x \sim N(0, v)$, and a bivariate normal distribution for (x, y).
- *A second example:* (Robert and Casella, example 7.4.1). If the complete conditionals are exponential with $E(x|y) = y^{-1}$ and $E(y|x) = x^{-1}$, then the integral diverges, and no joint distribution exists that is compatible with these conditionals - they are not *complete conditionals* of any distribution.

9 Gibbs Sampling

9.1 General Framework

Consider a $p-$dimensional distribution with p.d.f. $p(x)$. Consider any partition of x into a set of $q \leq p$ elements $x = \{x_1, \ldots, x_q\}$ where each x_i may be a vector or scalar. The implied set conditional distributions have p.d.f.s

$$p(x_i|x_{-i}), \qquad \text{where } x_{-i} = x \backslash x_i, \qquad (i = 1, \ldots, q). \tag{20}$$

In the extreme case in which x_i is univariate, these define the set of univariate *complete conditional distributions* defined by $p(x)$. In other cases the set of q conditionals represents conditionals implied on vector subsets of x based on the chosen partitioning. To be clear in notation, we now subscript the conditionals by the index of the subvector in the argument, and as needed will explicitly denote the conditioning elements, i.e.,

$$p_i(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_q) = p_i(x_i|x_{-i}) \equiv p(x_i|x_{-i})$$

for each $i = 1, \ldots, q$.

The idea of Gibbs sampling is just the immediate extension from the bivariate case above. With x partitioned in the (essentially arbitrary) order as described, proceed as follows:

1. Choose any value initial value $x = x^{(0)}$.
2. Successively generate values $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, as follows. For each $t = 1, 2, \ldots$, based on the "current" value $x^{(t)}$ sample a new state $x^{(t+1)}$ by this sequence of simulations:

   - draw a new value of x_1, namely

   $$x_1^{(t+1)} \sim p_1(x_1|x_{-1}^{(t)});$$

   - continue through new draws of

   $$x_i^{(t+1)} \sim p_i(x_i|x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, \; x_{i+1}^{(t)}, \ldots, x_q^{(t)})$$

   from $i = 2, \ldots, q-1$, then
   - complete the resampling via

   $$x_q^{(t+1)} \sim p_q(x_q|x_{-q}^{(t+1)}).$$

   This results in a complete update of $x^{(t)}$ to $x^{(t+1)}$, and then passes to the next step. Note how, at each step, the "most recent" sampled values of each element of x in the conditioning is used.

The term *Gibbs sampling* has sometimes been restricted to apply to the case of $p = q$ when all elements x_i are univariate, and the conditional distributions $p_i(\cdot|\cdot)$ are the full set of p complete univariate conditionals. More generally, and far more common in practice, is the use of a partition of x into a set of components

for which the implied q conditionals are easily simulated. It should be intuitively evident that blocking univariate elements together into an element x_i and then sampling $x_i^{(t+1)}$ from the implied conditional leads to - for that block of components of $x^{(t+1)}$ - a simulated value that depends less on the previous value $x^{(t)}$ than it would in the pure Gibbs sampling from univariate conditionals. An extreme example is when we have just one element, so sample directly from $p(x)$. Theory supports this intuition: the more we can block elements of x together to create subvectors x_i of higher dimension, the weaker will be the dependence between successive values $x^{(t)}$ and $x^{(t+1)}$. Hence in examples from here on, and especially practical model examples, the MCMC iterates between parameters and variables of varying dimension. Indeed, this is really key to the broad applicability and power of MCMC generally, and Gibbs sampling in particular.

In the following examples, the state space is a parameter space defined in a statistical model analysis, and Gibbs sampling is used to sample from the posterior distribution in that analysis.

9.2 An Example of Gibbs Sampling Using Completion

"Completion" refers to extending the state, or parameter space to introduce otherwise hidden or latent variables that, in this context, provide access to a Gibbs sampler. The term "data augmentation" is also used, although usually the additional variables introduced are not hypothetical data at all; in some cases, they do, however, have a substantive interpretation, as in problems with truly missing or censored data.

A simple, and very useful example is linear regression modelling under a heavy-tailed error distribution fixes ideas. This is a special case of linear regression with heavy-tailed errors, which underlies "robust regression" estimation - the use of an error distribution that has heavier-tails than normal can provide "protection" against outlying observations (bad or corrupt measurements) that, in the normal model, can have a substantial and undesirable impact on regression parameter estimation. Much empirical experience supports the view that, in many areas of science, natural and experimental processes generate variation that, while leading to (at least approximately) symmetric error distributions, is simply non-normal and usually heavier-tailed.

9.2.1 Reference Normal Linear Regression

Start with the usual normal linear model and its reference Bayesian analysis:

- n observations in $n-$vector response y, known $n \times k$ design matrix H whose columns are the n values of k regressors (predictor variables), and $n-$vector of observational errors ϵ with independent elements $\epsilon_i \sim N(0, \phi^{-1})$ of precision ϕ.
- Elements of y : $y_i = h_i'\beta + \epsilon_i$ where the h_i are the rows of H.
- $y = H\beta + \epsilon$.

In the general notation we have a state $x = (\beta, \phi)$ with the corresponding $p = (k+1)-$dimensional state space. The standard reference posterior $p(x|y) \equiv p(\beta, \phi|y)$ is well-known:

- Reference prior $p(\beta, \phi) \propto \phi^{-1}$ leads to the reference posterior of a normal/inverse gamma form, namely $(\beta|\phi, y) \sim N(b, \phi^{-1}B^{-1})$ and $(\phi|y) \sim Ga((n-k)/2, q/2)$ where $B = H'H$, $b = \hat{\beta} = B^{-1}H'y$ is the LSE, and $q = e'e$ is the sum of squared residuals $e = y - Hb$.

The implied posterior T distribution for β is usually used for inference on regression effects, and this can of course be simulated if desired - most easily via the convolution implicit in the conditional normal/inverse gamma posterior above. That is, the joint posterior $p(\beta, \phi|y)$ is directly simulated by drawing ϕ from its marginal gamma posterior and then, conditional on that ϕ value, drawing β from the conditional normal posterior.

9.2.2 Reference Normal Linear Regression with Known, Unequal Observational Variance Weights

Weighted linear regression (and weighted least square) has $\epsilon_i \sim N(0, \lambda_i^{-1}\phi^{-1})$ independently, allowing differing "weights" λ_i to account for differing precisions in error terms. The reference analysis is trivially modified in this case when the weights $\Lambda = (\lambda_1, \ldots, \lambda_n)'$ are specified. Now,

$$y_i = h_i'\beta + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, \lambda_i^{-1}\phi^{-1}).$$

The changes to the posterior are simply that, now, the regression parameter estimate b, associated precision matrix B and the residual sum of squares are based on the weighted design and response data, via

$$B = H'\Lambda H, \quad b = B^{-1}H'\Lambda y \quad \text{and} \quad q = e'\Lambda e.$$

9.2.3 Heavy-Tailed Errors in Regression

A key example is to change the normal error assumption into a T distribution, say T with $\nu = 5$ degrees of freedom. We can imagine treating ν as an additional parameter to be estimated, but for now consider ν fixed. Evidently, the joint posterior is now complicated:

$$p(\beta, \phi|y) \propto \phi^{n/2-1} \prod_{i=1}^{n} \{1 + \phi(y_i - h_i'\beta)'(y_i - h_i'\beta)/\nu\}^{-(\nu+1)/2},$$

in the $(k+1)-$dimensional parameter space. This is, in general, a horribly complicated (poly-T) function that may be multimodal, skewed differently in different dimensions, and is very hard to explore and summarise.

The completion, or augmentation, method that enables an almost trivial Gibbs sampler just exploits the definition of the T distribution as a scale mixture of normals. Recall that the T distribution for ϵ_i can be written as the marginal distribution from a joint in which

- $\epsilon_i|\lambda_i \sim N(0, \lambda_i^{-1}\phi^{-1})$, and
- $\lambda_i \sim Ga(\nu/2, \nu/2)$.

Write $\Lambda = (\lambda_1, \ldots, \lambda_n)'$ for the new/latent parameters induced here. Then, conditional on Λ, we have the weighted regression model. The augmentation is now clear: extend the state from (β, ϕ) to $x = (x_1, x_2)$ with $x_1 = (\beta, \phi)$ and $x_2 = \Lambda$. The posterior of interest is now $p(x_1, x_2|y)$ - if we can sample this posterior, then the simulated values lead to marginal samples for the primary parameters $x_1 = (\beta, \phi)$, while also adding value in that we will generate posterior draws of the latent "weights" of observations that can be used to explore which observations have large/small weights, based on the fit to the predictors.

The two conditional posteriors of interest are defined:

- For $x_1 = (\beta, \phi)$, we have the weighted regression model conditional on $x_2 = \Lambda$, so we know that we simulate directly from $p(\beta, \phi|\Lambda, y)$ via
  - $(\phi|\Lambda, y) \sim Ga((n-k)/2, q/2)$ and then
  - $(\beta|\phi, \Lambda, y) \sim N(b, \phi^{-1}B^{-1})$

  where $b = H'\Lambda H$, $b = B^{-1}H'\Lambda y$ and $q = e'\Lambda e$ with $e = y - Hb$.
- For $x_2 = \Lambda$, it is trivially seen that

$$p(\Lambda|\beta, \phi, y) = \prod_{i=1}^{n} p(\lambda_i|\beta, \phi, y_i)$$

with independent component margins $p(\lambda_i|\beta,\phi,y_i) \propto p(\lambda_i)p(y_i|\beta,\phi,\lambda_i)$, so that

$$(\lambda_i|\beta,\phi,y_i) \sim Ga((\nu+1)/2, (\nu+\phi\epsilon_i^2)/2)$$

at $\epsilon_i = y_i - h_i'\beta$.

9.3 General Markov Chain Framework of Gibbs Sampling

Gibbs samplers define first-order Markov processes on the full $p-$dimensional state space. Note that the sequence of conditional simulations in §3.1 provides a complete update from the sampled vector $x^{(t-1)}$ to the sampled vector $x^{(t)}$ by successively updating the components. That this is Markov is obvious, and it is clearly also homogenous since the conditional distributions are the same for all t. The implied transition distribution is easily seen to be based on these individual conditionals. For clarity in notation in this section write $F(x|x')$ for the transition distribution function and $f(x|x')$ for the corresponding p.d.f. . Then the Markov process defined by the Gibbs sampler has transition p.d.f. governing transitions from any state x' to new states x, of the form:

$$f(x|x') = p_1(x_1|x'_{-1})\{\prod_{i=2}^{q-1} p_i(x_i|x_1,\ldots,x_{i-1},\ x'_{i+1},\ldots,x'_q)\}p_q(x_q|x_{-q}).$$

It is relatively straightforward to show (see Robert and Casella, §7.1.3) that the joint density $p(x)$ does indeed provide a solution as a stationary distribution of the Markov process, i.e.,

$$p(x) = \int f(x|x')p(x')dx'.$$

Further, under conditions that are every broadly applicable in statistical applications, the process converges to this *unique* stationary distribution: the Markov chain defined by the Gibbs sampler is irreducible and aperiodic, so defining an ergodic process with limiting distribution $p(x)$. Thus, MC samples generated by such Gibbs samplers will *converge* in the sense that, ultimately, $x^{(t)} \sim p(x)$. This is the case in applications with positive conditional densities $p_i(x_i|x_{-i}) > 0$ for all x, for example, and this really highlights the relevance and breadth of applicability in statistical modelling. There are models and examples in which this is not the case, though even then convergent Gibbs samplers may be defined (on a case-by-case basis) or alternative methods that modify the basic Gibbs structure may apply.

In general, the Markov chains generated by Gibbs samplers are not reversible on the full state space, though the induced chains on elements of x are.

9.4 Some Practicalities

9.4.1 General Comments

- The key focuses in setting up and running a Gibbs sampler relate to the choice and specification of the conditional distributions with two considerations: ease and efficiency of the resulting simulations, and resulting dependence between successive states in the resulting Markov chain. Once a sampler is specified, the questions of monitoring it to assess convergence - to decide when to begin to save generated x values following an initial "burn-in" period - are raised.
- A sampler with very low dependence between successive states is generating samples that are close to random samples, and that is the gold-standard. Highly dependent chains may be sub-sampled to save draws some iterations apart, thus weakening the serial dependence. Setting up and running a Gibbs sampler requires specification of sets of conditional distributions as in the examples, and choices about

which specific distributions to use. We have already discussed the idea of *blocking* - using conditional distributions such that the subvectors x_i are of maximal dimension, in an attempt to "weaken" the dependence between successively saved values.

- Initialization is often challenging, and repeat Gibbs runs from different starting values are of relevance. Repeat, differentially initialled Gibbs runs that eventually generate states in the same region suggest that there has been convergence to that region.
- Most Gibbs samplers are non-linear, though linear autocorrelations are the simple, standard tools for examining first-order structure. Plotting sample paths of selected subsets of x over Gibbs iterates, and plotting sample autocorrelation functions for Gibbs output series after an initial burn-in period, are two common exploratory methods.

9.4.2 Sample Autocorrelations

Recall the notation for dependence structure in a stationary univariate series x_t. Assuming second-order moments exist) the autocovariance function is $\gamma(k) = C(x_t, x_{t\pm k})$ with marginal variance $V(x_t) = \gamma(0)$ for each t, and then the corresponding autocorrelation (a.c.f.) function is $\rho(k) = \gamma(k)/\gamma(0)$. Sample autocovariances and autocorrelations based on an observed series $x_{1:n}$ are just the sample analogues, namely

$$\hat{\gamma}(k) = n^{-1} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) \qquad \text{and} \qquad \hat{\rho}(k) = \hat{\gamma}(k)/\hat{\gamma}(0).$$

Simply evaluating and plotting the latter based on a (typically long, perhaps subsampled) Gibbs series gives some insights into the nature of the serial dependence. The concepts and insights into dependence over time we have generated in studying (exhaustively) the linear AR(1) model translate into assessment of serial dependence in Gibbs samples, even though they generally represent non-linear AR processes.

nb. See course web page, under the Examples, Matlab Code and Data link, for simple Matlab functions for computing and plotting the a.c.f.

9.5 Gibbs Sampling in Linear Regression with Multiple Shrinkage Priors

See Supplementary Notes web page link.

9.6 Gibbs Sampling in Hierarchical Models

See Robert & Casella, §7.1.6 for some examples.

10 Gibbs Sampling in a Stochastic Volatility Model: HMMs and Mixtures

10.1 Introduction

Here is an example that both introduces very standard, core manipulation of discrete mixture distributions and that provides a component of a nowadays standard Gibbs sampler for the non-linear stochastic volatility models in financial times series. Recall the canonical SV model. We have an observed price series, such as a stock index or exchange rate, P_t at equally spaced time points t, and model the per-period *returns* $r_t = P_t/P_{t-1} - 1$ as a zero mean but time-varying volatility process

$$\begin{aligned} r_t &\sim N(0, \sigma_t^2), \\ \sigma_t &= \exp(\mu + x_t), \\ x_t &\leftarrow AR(1|\theta) \end{aligned}$$

with $\theta = (\phi, v)$. (In reality, such models are components of more useful models in which the mean of the returns series is non-zero and is modelled via regression on economic and financial predictors). The parameter μ is a baseline log-volatility; the AR(1) parameter ϕ defines persistence in volatility, and the innovations variance v "drives" the levels of activity in the volatility process.

Based on data $r_{1:n}$, the uncertain quantities to infer are $(x_{1:n}, \mu, \theta)$ and MCMC methods will utilize a good deal of what we know about the joint normal distribution - and its various conditionals - of $x_{0:n}$. The major complication arises due to the non-linearity inherent in the observation equation, and one approach to dealing with this uses the log form earlier mentioned: transforming the data to $y_t = \log(r_t^2)/2$, we have

$$\begin{aligned} y_t &= \mu + x_t + \nu_t, \\ x_t &\leftarrow AR(1|\theta), \end{aligned}$$

where $\nu_t = \log(\kappa_t)/2$ and $\kappa_t \sim \chi_1^2$. This is a linear AR(1) HMM but with non-normal observation errors ν_t. We know that we can develop an effective Gibbs sampler for a normal AR(1) HMM, and that suggests using normal distributions somehow.

10.2 Normal Mixture Error Model

We know that we can approximate a defined continuous p.d.f. as accurately as desired using a discrete mixture of normals. Using such a mixture for the distribution of ν_t provides what is nowadays a standard analysis of the SV model, utilizing the idea of completion again - i.e., introducing additional latent variables to provide a *conditionally normal* model for the ν_t, and then including those latent variables in the analysis. Specifically:

- The log-$\chi_1^2/2$ distribution can be very accurately approximated by a discrete mixture of a few normal distributions with known parameters, i.e.,

$$p(\nu_t) = \sum_{j=1}^{J} q_j N(b_j, w_j).$$

  One standard approximation, developed simply by numerical optimization, uses $J = 7$ and conditional moments

q_j :	0.0073	0.0000	0.1056	0.2575	0.3400	0.2457	0.0440
b_j :	−5.7002	−4.9186	−2.6216	−1.1793	−0.3255	0.2624	0.7537
w_j :	1.4490	1.2949	0.6534	0.3157	0.1600	0.0851	0.0418

See Kim, Shephard and Chib, 1998, *Review of Economic Studies* for this. Mixtures with more components can refine the approximation - the Gibbs sampling setup is structurally the same, and just changes in details.

- Introduce latent *indicator variables* $\gamma_t \in \{1:J\}$ with $p(\gamma_t) = q_j$ at $\gamma_t = j$ independently over t and of all other random quantities. Then the mixture of normals can be constructed from the conditionals

$$(\nu_t|\gamma_t = j) \sim N(b_{\gamma_t}, w_{\gamma_t}) \equiv N(b_j, w_j), \quad (j = 1, \ldots, J).$$

The implied marginal distribution of ν_t, averaging with respect to the discrete density $p(\gamma_t)$ is the mixture of normals above. This is the *data augmentation*, or completion, trick once again. We introduce these latent mixture component indicators to induce conditional normality of the ν_t.

10.3 Exploiting Conditional Normality in Normal Mixtures

Some of the key component distributions for Gibbs sampling in the normal mixture SV model are most clearly understood by, first, pulling out of the time series context and considering just one time point - we will drop the t suffix for clarity and for this general discussion. Conditional on $(\gamma = j, x, \mu)$, we have $y \sim N(\mu + x + b_j, w_j)$. Imagine now that, at some point in a Gibbs sampling analysis, we are interested in the conditional posterior for x under a *normal* prior and conditioning on current values of $\gamma = j$ and μ. The likelihood for x in this setup is normal, so the implied conditional posterior for x is normal. Similarly, at any point in a Gibbs sampling analysis when we want to compute the conditional posterior for μ under a *normal* prior and conditioning on current values of $\gamma = j$ and x, the likelihood for μ in this setup is normal, so the implied conditional posterior for μ is normal. Hence, we can immediately see that the *imputation* of the latent mixture component indicators translates the complicated model into a coupled set of normal models within which the theory is simple, and conditional simulations are trivial. This is the key to MCMC in this example, and many others involving mixtures.

To be explicit, if we have a prior $x \sim N(m, M)$ then we have the following component conditional distributions:

- The conditional posterior for $(x|y, \gamma, \mu)$ is normal with moments
  - $E(x|y, \gamma, \mu) = m + A(y - \mu - b_\gamma - m)$ with $A = M/(M + w_\gamma)$ and
  - $V(x|y, \gamma, \mu) = w_\gamma A$.
- The conditional posterior for $(\gamma|y, x, \mu)$ is given by the J probabilities q_j^* over $\gamma = j \in \{1, \ldots, J\}$, via

$$\begin{aligned} q_j^* = \quad & Pr(\gamma = j|y, x, \mu) \propto Pr(\gamma = j)p(y|\gamma = j, x, \mu) \\ & \propto q_j \exp\{-(y - \mu - b_j - x)^2/(2w_j)\}/\sqrt{w_j} \end{aligned}$$

for $j = 1, \ldots, J$. Normalization over j leads to the updated probabilities q_j^*.

These two, coupled conditionals are easy to simulate, and form key elements of the complete Gibbs sampler for the normal mixture SV models.

10.4 An Initial Gibbs Sampler in the SV Model

A full Gibbs sampler can now be defined to simulate from the complete target posterior

$$p(x_{0:n}, \mu, \phi, v|y_{1:n})$$

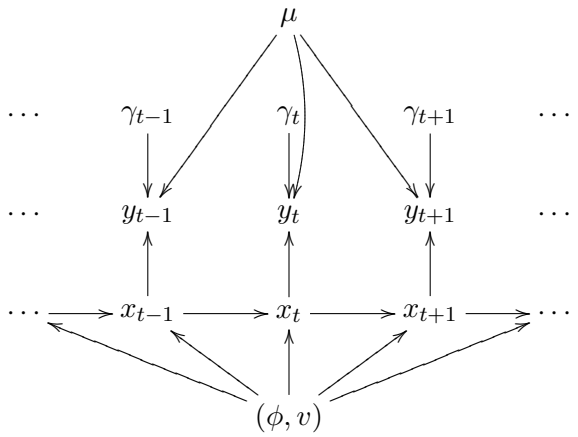
by first extending the state space to include the set of indicators $\gamma_{1:n}$. That is, we sample various conditionals of the full posterior

$$p(x_{0:n}, \gamma_{1:n}, \mu, \phi, v|y_{1:n}).$$

One component of this will be to sample $x_{0:n}$ from a conditionally linear, normal AR(1) HMM, and *all ingredients have been covered: Question 2 of Homework #2 and Question 3 of the Midterm exam*. For this we need to finalize the specification with the prior $p(x_0, \mu, \phi, v)$, and here take these quantities to be independent with prior $p(x_0)p(\mu)p(\phi)p(v)$ where $x_0 \sim N(0, u)$ for some (large?) $u > 0$, $\mu \sim N(g, G)$, $\phi \sim N(c, C)$ and $v^{-1} \sim Ga(a/2, av_0/2)$.

Gibbs sampling can now be defined. One convenient initialization is to set $\mu = \bar{y} = n^{-1}\sum_{t=1}^{n} y_t$ and then, for each $t = 1, \ldots, n$, $x_t = y_t - \bar{y}$ and with $x_0 = 0$. Given these initial values, the sequence of conditionals to sample from to generate each Gibbs step is as follows. Note that each conditional distribution is, formally, explicitly conditioned on values of all other uncertain and observed quantities; however, the highly structured model has various conditional independencies that leads to reduction and simplification in each case. This is characteristic of Gibbs sampling in complex, hierarchical models - we massively exploit this conditional independence structure to simplify the calculations.

The conditional independence structure is exhibited in the graphical model:



1. $p(\gamma_{1:n}|y_{1:n}, x_{0:n}, \mu, \phi, v)$.

   It is immediate that the γ_t are conditionally independent and the above discussion provides the conditional posteriors over $j = 1, \ldots, J$. Namely,

   $$Pr(\gamma_t = j|y_t, x_t, \mu) = q^*_{t,j}$$

   where

   $$q^*_{t,j} \propto q_j \exp\{-(y_t - \mu - b_j - x_t)^2/(2w_j)\}/\sqrt{w_j}$$

   for $j = 1, \ldots, J$, and normalization over j leads to the updated probabilities $q^*_{t,j}$, for each time point t. This defines a discrete posterior for each γ_t that is trivially sampled to generate new mixture component indicators.
2. $p(\phi|y_{1:n}, x_{0:n}, \mu, v)$.

   This is just the normal posterior $p(\phi|x_{0:n}, v)$ for the AR(1) coefficient in the AR(1) model under the $\phi \sim N(c, C)$ prior. This is trivially sampled to generate a new value of ϕ.
3. $p(v|y_{1:n}, x_{0:n}, \mu, \phi)$.

   This is just the posterior $p(v|x_{0:n}, \phi)$ for the AR(1) innovation variance in the AR(1) model under the inverse gamma prior. This is trivially sampled to generate a new value of v.

4. $p(\mu|y_{1:n}, x_{0:n}, \gamma_{1:n}, \phi, v)$.

   This is the posterior for μ under the $\mu \sim N(g, G)$ prior and based on n conditionally normal, independent observations $y_t^* = y_t - x_t - b_{\gamma_t} \sim N(\mu, w_{\gamma_t})$. So the posterior is, easily, normal, say $N(g', G')$ where

$$1/G' = 1/G + 1/W \qquad \text{and} \qquad g' = G'(g/G + \hat{\mu}/W)$$

   where

$$W^{-1} = \sum_{t=1}^{n} w_{\gamma_t}^{-1} \qquad \text{and} \qquad \hat{\mu} = W \sum_{t=1}^{n} w_{\gamma_t}^{-1} y_t^*.$$

5. $p(x_{0:n}|y_{1:n}, \gamma_{1:n}, \mu, \phi, v)$.

   Under the conditioning here we have a (conditionally) linear, normal AR(1) HMM. The one extension to the discussion of this model earlier is that, given the specific values of elements of $\gamma_{1:n}$, the error variances of the normal distributions for the y_t are known and different over time: the earlier constant variance w is replaced by the conditional values w_{γ_t} for each t. We can then run the *forward-filtering, backward sampling (FFBS)* algorithm that uses the Kalman filter-like forward analysis from $t = 1$ up to $t = n$, then moves backwards in time successively simulating the states $x_n, x_{n-1}, \ldots, x_0$ to generate a full sample $x_{0:n}$.

   For clarity here, temporarily drop the quantities $(\gamma_{1:n}, \mu, \phi, v)$ from the conditioning of distributions - they are important but, at this step, fixed at "current" values throughout.

   - *Forward Filtering:*

     Refer back to Question 2 of Homework #2 for full supporting details.

     Beginning at $t = 0$ with $x_0 \sim N(m_0, M_0)$ where $m_0 = 0, M_0 = u$, we have, for all $t = 0, 1, \ldots, n$, the on-line posteriors

$$(x_t|y_{1:t}) \sim N(m_t, M_t)$$

     sequentially computed by the Kalman filtering update equations:

$$m_t = a_t + A_t e_t \qquad \text{and} \qquad M_t = w_{\gamma_t} A_t$$

     where

$$\begin{aligned} e_t &= y_t - \mu - b_{\gamma_t} - a_t, \\ a_t &= \phi m_{t-1}, \\ A_t &= h_t/(h_t + w_{\gamma_t}), \\ h_t &= v + \phi^2 M_{t-1}. \end{aligned}$$

   - *Backward Sampling:*
     - At $t = n$, sample from $(x_n|y_{1:n}) \sim N(m_n, M_n)$.
     - Then, for each $t = n-1, n-2, \ldots, 0$, sample from

$$p(x_t|x_{(t+1):n}, y_{1:n}) \equiv p(x_t|x_{t+1}, y_{1:t})$$

       using the just sampled value of x_{t+1} in the conditioning. Here $(x_t|x_{t+1}, y_{1:t})$ is normal with

$$E(x_t|x_{t+1}, y_{1:t}) = m_t + (\phi M_t/h_{t+1})(x_{t+1} - a_{t+1})$$

       and

$$V(x_t|x_{t+1}, y_{1:t}) = vM_t/h_{t+1}.$$

It is worth commenting on the final step and the relevance of the FFBS component. This is an example of a Hidden Markov Model and this specific approach is used in other such models, including HMMs with discrete state spaces. In the SV model here, an older, alternative approach to simulating the latent x_t values is to use a Gibbs sampler on each of the complete conditionals $p(x_t|x_{-t}, y_{1:n}, \gamma_{1:n}, \mu, \phi, v)$. This is certainly feasible and a simpler technical approach, but will tend to be much less effective in applications, especially in cases - as in finance - where the AR dependence is high. In many such applications ϕ is close to 1, the process being highly persistent; this induces very strong correlations between x_t and x_{t-1}, x_{t+1}, for example, so that the set of complete univariate conditionals will be *very* concentrated. A Gibbs sampler run on these conditionals will then move around the x component of the state space extremely slowly, with very high dependence among successive iterations - it will converge, but very, very slowly indeed. The FFBS approach *blocks* all the x_t variates together and at each Gibbs step generates a completely new sample trajectory $x_{0:n}$ from the relevant joint distribution, and so moves swiftly around the state space, generally rapidly converging.

10.5 Ensuring Stationarity of the SV Model

Some of the uses of the model are to simulate future volatilities: given an MC sample from the above Gibbs analysis, it is trivial to then forward sample the x_t process to generate an MC sample for, say, x_{n+k} for any $k > 0$, and hence deduce posterior samples for $p(r_{n+k}|y_{1:n})$. The spread of these distributions play into considerations of just how wild future returns might be, and currently topical questions of value-at-risk in financial management, for example.

Practical application of SV models require that we modify the set-up to ensure that the SV process is stationary. A model allowing values of $|\phi| > 1$ will generate volatility processes that are explosive, and that is quite unrealistic. The modification is trivial mathematically and - now that the analysis and model fitting uses Gibbs sampling - makes little difference computationally and involves only very modest changes to the above set of conditional distributions. This is an example of how analysis using MCMC allows us to easily overcome technical complications that would otherwise be much more challenging.

The volatility process is stationary if we restrict to $|\phi| < 1$ and assume the stationary distribution for the pre-initial value x_0. That is,

$$(x_0|\phi, v) \sim N(0, v/(1-\phi^2)).$$

Hence, for stationarity we must replace the earlier assumption $x_0 \sim N(0, u)$ by substituting $u = v/(1-\phi^2)$. This depends on the unknown parameters (ϕ, v) but that simply means we need to recognize how this changes some or all of the conditional distributions within the MCMC analysis as a result of this dependence. The impact on the set of five conditional distributions are as follows.

1. The conditional distributions for $\gamma_{1:n}$ are unchanged.
2. The conditional posterior for ϕ is now restricted to the stationary region $|\phi| < 1$. In volatility modelling we expect high, positive values so that we may restrict further to $0 < \phi < 1$ if desired; we shall assume this here for the following development. Further, the posterior for ϕ also now has an additional component contributed by the likelihood term $p(x_0|\phi, v)$ that had earlier been ignored. Thus the conditional is

$$p(\phi|x_{0:n}, v) \propto p(\phi)p(x_{1:n}|x_0, \phi, v)p(x_0|\phi, v), \quad 0 < \phi < 1.$$

We may still use the prior $p(\phi) = N(c, C)$ but now truncated to $0 < \phi < 1$, so that

$$p(\phi|x_{0:n}, v) \propto \exp(-(\phi - c')^2/(2C'))a(\phi), \quad 0 < \phi < 1,$$

where $1/C' = 1/C + B/v$ and $c' = C'(c/C + Bb/v)$ with $B = \sum_{t=1}^{n} x_{t-1}^2$ and where b is the sample autocorrelation at lag-1, $b = \sum_{t=1}^{n} x_t x_{t-1}/B$, as before. Also the term $a(\phi)$ is just

$$a(\phi) \propto p(x_0|\phi, v) = (1-\phi^2)^{1/2} \exp(\phi^2 x_0^2/(2v)).$$

Note that this latter term introduces an additional technical complication, but is a term that apparently contributes little (it is "worth" one observation on the x series, compared to the other $n-1$) and may, at this first step, be ignored. If we ignore $a(\phi)$, the conditional density is a truncated normal, which we can easily simulate. We shall see how to simply modify this using a Metropolis-Hastings MCMC component later on.

3. The conditional for v is similarly modified by the addition of the pre-initial term

$$p(x_0|\phi, v) \propto v^{-1/2} \exp((1-\phi^2)x_0^2/(2v))$$

now viewed as a function of v. This adds one degree-of-freedom to the conditional inverse gamma posterior, and adds the term $(1-\phi^2)x_0^2$ to the sum-of-squares; that is, under the prior $v^{-1} \sim Ga(a/2, av_0/2)$ based on some prior point estimate v_0 and degrees-of-freedom a, we have the conditional posterior

$$(v^{-1}|x_{0:n}, \phi) \sim Ga((a+n+1)/2, (av_0+q)/2)$$

where $q = (1-\phi^2)x_0^2 + \sum_{t=1}^{n} \epsilon_t^2$ with $\epsilon_t = x_t - \phi x_{t-1}$ for $t = 1, \ldots, n$.

4. The conditional distributions for μ is unchanged.
5. The only change to the FFBS analysis is the initialization at $x_0 \sim N(m_0, M_0)$ as described, but now the initial moments are $m_0 = 0$ and $M_0 = v/(1-\phi^2)$.

10.6 Model Re-parametrization: The Centered AR Model

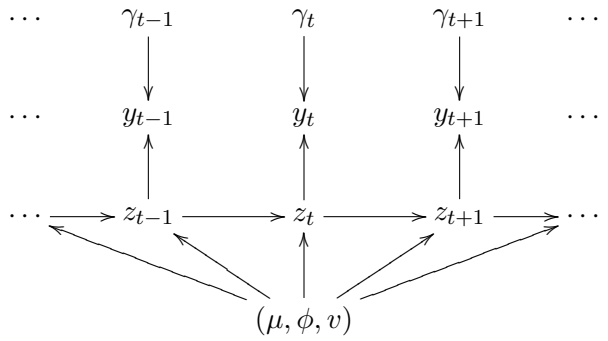
The above Gibbs sampler can be usefully applied to generate exploration of the full posterior, and is provably convergent. However, it is an excellent example of an MCMC analysis that suffers from slow convergence in some components - specifically, the $(x_{0:n}, \mu)$ sub-states, due to choice of parametrization. Some experience in exploring this Gibbs sampler with either simulated or real data quickly reveal that the simulated sample paths for μ will almost always drift far away from regions consistent with expectation under the prior and, in a financial times series analysis, with economic reality. This is coupled with corresponding drifts in the sample paths of the simulated volatility process x_t itself. The problem is one of "soft" identification, and is clarified as follows. The y_t data provide direct observations on the sum $\mu + x_t$. Adding any constant to μ while subtracting that same constant from each of the x_t leaves the model unchanged, so that the posterior will inevitably exhibit very strong *negative* dependence between the values of the x_t process and μ. The Gibbs sampler as structured above couples sampling for μ given $x_{0:n}$ with sampling for $x_{0:n}$ given μ, and so will usually wander around generating samples such that the $\mu + x_t$ are quite appropriate and consistent with the data, but with with values of μ and x_t themselves subject to quite unpredictable and commonly very large shifts up/down respectively. Simply applying the sampler and monitoring convergence will show clear evidence of (a) very strong negative dependence as described, and (b) high positive autocorrelations within the Gibbs iterates indicative of the non-stationary drifting.

This is a common problem in complex, multi-parameter models. Some examples in hierarchical models are discussed by Robert and Casella (§7.1). The practicality is one of appropriate choice of parametrization to define the Gibbs conditional posteriors when more than one such choice is possible, and this example of theoretically implied dependence in the posterior - that is obvious from the structure of the model - is a key feature that, in other models, should suggest re-parametrization. The solution here is to focus on the *centered* AR(1) parametrization, as follows.

Write $z_t = \mu + x_t$ so that $x_t = z_t - \mu$ for each t. Then the model is recast as

$$\begin{aligned} y_t &= z_t + \nu_t, \\ z_t &= \mu + \phi(z_{t-1} - \mu) + \epsilon_t, \end{aligned}$$

and we can denote the second equation - the centered AR(1) model - as $z_t \leftarrow \mu + AR(1|(\phi, v))$. The model has not changed, but the graphical representation now clearly shows the decoupling of μ from the data y_t, and the resulting simplified conditional independence structure:



This parametrization induces a few changes in the conditional distributions for the Gibbs sampler. The basic structure of the conditional distributions is the same, though details change. A few pertinent points are as follows:

- Assuming $|\phi| < 1$ the process z_t is a stationary AR(1) process with a non-zero mean μ, and stationary distribution $z_t \sim N(\mu, v/(1-\phi^2))$. In particular this implies the pre-initial prior now depends on μ as well as (ϕ, v),
$$(z_0|\mu, \phi, v) \sim N(\mu, v/(1-\phi^2)).$$
- The data $y_{1:n}$ provide information directly on $z_{0:n}$ and $\gamma_{1:n}$. Conditional on $z_{1:n}$, the components of posteriors for the parameters (μ, ϕ, v) do not depend on the data.
- Conditioning on $z_{0:n}$ and (ϕ, v) implies a normal likelihood for μ - essentially that from the $n+1$ "observations" $\xi_0 = z_0 \sim N(\mu, v/(1-\phi^2))$ and, for $t = 1, \ldots, n$, $\xi_t \sim N(\mu, v/(1-\phi)^2)$ where $\xi_t = (z_t - \phi z_{t-1})/(1-\phi)$.

10.7 Practicable Gibbs Sampler for the HMM SV Model: Summary Conditionals

The full set of conditionals for Gibbs iterations in this standard, convergent and (at time of writing) most well-behaved Gibbs sampler known, are as follows.

1. For $\gamma_{1:n}$:

   Sample the independent posteriors $p(\gamma_t|y_t, z_t)$ defined by

   $$Pr(\gamma_t = j|y_t, x_t) = q^*_{t,j} \propto q_j \exp\{-(y_t - b_j - z_t)^2/(2w_j)\}/\sqrt{w_j}$$

   for $j = 1, \ldots, J$.
2. For ϕ :

   Sample from

   $$p(\phi|z_{0:n}, \mu, v) \propto g(\phi)a(\phi)$$

where $g(\phi) = N(c', C')$ *truncated* to the stationary region, or to $0 < \phi < 1$ for a positive dependency, and with:

$$1/C' = 1/C + B/v \qquad \text{and} \qquad c' = C'(c/C + Bb/v),$$

where

$$B = \sum_{t=1}^{n} x_{t-1}^2 \qquad \text{and} \qquad b = \sum_{t=1}^{n} x_t x_{t-1}/B$$

with $x_t = z_t - \mu$ for each t. Also, $a(\phi) = (1-\phi^2)^{1/2} \exp(\phi^2 x_0^2/(2v))$.

An approximate strategy ignores the $a(\phi)$ term and just samples the truncated normal $g(\phi)$. A simple Metropolis-Hastings component corrects this, as we shall see.

3. For v :

   Sample from

   $$(v^{-1}|z_{0:n}, \mu, \phi) \sim Ga((a+n+1)/2, (av_0+q)/2)$$

   with $q = (1-\phi^2)x_0^2 + \sum_{t=1}^{n} \epsilon_t^2$ with $\epsilon_t = x_t - \phi x_{t-1}$ and where $x_t = z_t - \mu$ for $t = 0, \ldots, n$.

4. For μ :

   Sample from $(\mu|z_{0:n}, \phi, v) \sim N(g', G')$ where

   $$1/G' = 1/G + (1-\phi^2)/v + n(1-\phi)^2/v$$

   and

   $$g' = G'(g/G + (1-\phi^2)z_0/v + (1-\phi)^2 \sum_{t=1}^{n} \xi_t/v)$$

   where $\xi_t = (z_t - \phi z_{t-1})/(1-\phi)$. This simplifies as

   $$g' = G'(g/G + (1-\phi^2)z_0/v + (1-\phi) \sum_{t=1}^{n} (z_t - \phi z_{t-1})/v).$$

5. For $z_{0:n}$:

   Sample using the FFBS construction to simulate $p(z_{0:n}|y_{1:n}, \gamma_{1:n}, \mu, \phi, v)$.

   - *Forward Filtering:*
     - Set $m_0 = \mu$ and $M_0 = v/(1-\phi^2)$.
     - For $t = 0, 1, \ldots, n$, update the on-line posteriors $(z_t|y_{1:t}) \sim N(m_t, M_t)$ by

       $$m_t = a_t + A_t e_t \qquad \text{and} \qquad M_t = w_{\gamma_t} A_t$$

       where

       $$\begin{aligned} e_t &= y_t - b_{\gamma_t} - a_t, \\ a_t &= \mu + \phi(m_{t-1} - \mu), \\ A_t &= h_t/(h_t + w_{\gamma_t}), \\ h_t &= v + \phi^2 M_{t-1}. \end{aligned}$$

   - *Backward Sampling:*
     - Simulate $(z_n|y_{1:n}) \sim N(m_n, M_n)$.

– For $t = n-1, n-2, \ldots, 0$, sample from the normal conditional for $(z_t|z_{t+1}, y_{1:t})$ using the just sampled value of z_{t+1} in the conditioning. The mean and variance of this distribution are

$$E(z_t|z_{t+1}, y_{1:t}) = m_t + (\phi M_t/h_{t+1})(z_{t+1} - a_{t+1})$$

and

$$V(z_t|z_{t+1}, y_{1:t}) = vM_t/h_{t+1}.$$

This generates (in reverse order) the historical trajectory sample path $z_{0:n}$.

11 MCMC Methods

Gibbs sampling defines a broad class of MCMC methods of key utility in Bayesian statistical analysis as well as other applications of probability models. All Gibbs samplers are special examples of a general approach referred to as MCMC Metropolis-Hastings methods. The original special class of (pure) Metropolis methods is itself another rich and very broadly useful class of MCMC methods, and the MH framework extends it enormously.

11.1 Metropolis Methods

Consider the general framework of random quantities x in a $p-$dimensional state space χ and a target distribution $\Pi(x)$ with p.d.f. $\pi(x)$, discrete continuous or mixed. We cannot sample directly from $\pi(x)$ and so, based on the ideas and intuition from the Gibbs sampling approach, focus on the concept of exploring $\pi(x)$ by randomly wandering around via some form of sequential algorithm that aims to identify regions of higher probability and generate a catalogue of x values representing that probability. Imagine at some time $t-1$ we are at $x = x^{(t-1)} \in \chi$. Looking around in a region near $x^{(t-1)}$ we might see points of higher density, and they would represent "interesting" directions in χ to move towards. We might also see points of lower density that would be of lesser interest though, if the p.d.f. there is not really low, still worth moving towards. Metropolis methods reflect this concept by generating Markov processes on χ that wander around, moving to regions of higher density than at the "current" state, but also allowing for moves to regions of lower density so as to adequately explore the support of $\pi(x)$. The setup is as follows.

- The *target* distribution has p.d.f. $\pi(x)$ on χ.
- A realization of a first-order Markov process is generated, $x^{(1)}, x^{(2)}, \ldots, x^{(t)}, \ldots$, starting from some initial state $x^{(0)}$.
- A *proposal distribution* with p.d.f. $g(x|x')$ is defined for all $x, x' \in \chi$. This is actually a family of proposal distributions, indexed by x'. g must be symmetric in the sense that $g(x|x') = g(x'|x)$.
- At step $t-1$, the current state is $x^{(t-1)}$. A *candidate state* x^* is generated from the current proposal distribution,
$$x^* \sim g(x^*|x^{(t-1)}).$$
- The target density ratio
$$\pi(x^*)/\pi(x^{(t-1)})$$
compares the candidate state to the current state.
- The Metropolis chain moves to the candidate proposed if it has higher density than the current state; otherwise, it moves there with probability defined by the target density ratio. More precisely,

$$x^{(t)} = \begin{cases} x^*, & \text{with probability } \min\{1, \pi(x^*)/\pi(x^{(t-1)})\}, \\ x^{(t-1)}, & \text{otherwise.} \end{cases}$$

Hence the Metropolis chain always moves to the proposed state at time t if the proposed state has higher target density than the current state, and moves to a lower state with lower density in proportion to the density value itself. This leads to a random, Markov process than naturally explores the state space according to the probability defined by $\pi(x)$ and hence generates a sequence that, while dependent, eventually represents draws from $\pi(x)$.

11.1.1 Convergence of Metropolis Methods

Sufficient conditions for ergodicity are that the chain can move anywhere at each step, which is ensured, for example, if $g(x|x') > 0$ everywhere. Many applications satisfy this positivity condition, which we generally assume here. The Markov process generated under this condition is ergodic and has a limiting distribution. We can deduce that $\pi(x)$ is this limiting distribution by simply verifying that it satisfies the detailed balance (reversibility) condition - Metropolis chains are reversible.

Recall that detailed balance is equivalent to symmetry of the joint density of any two consecutive values (x, x') in the chain, hence they have a common marginal density. If we start with $x' \sim \pi(x')$, then the implied margin for x is $\pi(x)$, and the same is true of the chain in reverse. That is,

$$p(x|x')\pi(x') = p(x'|x)\pi(x)$$

where $p(x|x')$ is the transition density induced by the construction of the chain and $p(x'|x)$ that in reverse.

To see that this holds, suppose - with no loss of generality - that we have (x, x') values such that $\pi(x) \geq \pi(x')$. Then:

- Starting at x' the chain moves surely to the candidate $x \sim g(x|x')$. Hence $p(x|x') = g(x|x')$, and the *forward* joint density (left hand side of the detailed balance equation) is $g(x|x')\pi(x')$.
- In a reverse step from x to x' we have $x' \sim g(x'|x)$ with probability $\min\{1, \pi(x')/\pi(x)\} = \pi(x')/\pi(x)$ in this case, and $x' = x$ with probability $1 - \pi(x')/\pi(x)$. So the reverse transition p.d.f. is a mixture of g with the point mass, namely

$$\begin{aligned} p(x'|x) &= \frac{\pi(x')}{\pi(x)} g(x'|x) + \left(1 - \frac{\pi(x')}{\pi(x)}\right) \delta(x' - x) \\ &= \frac{\pi(x')}{\pi(x)} g(x'|x) + 0 \end{aligned}$$

  so that the *reverse* joint density (the right hand side of the detailed balance equation) is $\pi(x')g(x'|x)$. Now the symmetry of g plays a role, since this is the same as $\pi(x')g(x|x')$, and the detailed balance identity holds.

By symmetry the same argument applies for points (x, x') such that $\pi(x) \leq \pi(x')$. It follows that $\pi(x)$ is the stationary p.d.f. satisfying detailed balance, hence the limiting distribution of the Metropolis chain.

11.2 Comments and Examples

1. It is characteristic of Metropolis chains that they remain in the same state for some random number of iterations, based on the "rejected" candidates. Movement about the state space with a reasonably high acceptance rate is desirable, but acceptance rates that are very high are a cause for concern. Note simply that a candidate state x^* that is very close to the current state x' leads to $\pi(x^*) \approx \pi(x')$ and so is highly likely to be accepted. Thus proposal distributions that are very precise have high acceptance rates at the cost of moving very slowly around the state space, making frequent but very small or "local" moves, and so will be very slow to converge to the stationary distribution. On the other hand, very diffuse proposals that generate large moves in state space are much more likely to generate candidates with very low values of $\pi(x^*)$ and hence will suffer from low acceptance rates even though they more freely explore the state space. Thus the choice - or design - of proposal distributions will generally aim to balance acceptance rate with convergence, and is based on a mix of experimentation and customization to the features of $\pi(x)$ to the extent possible.

2. With a $p-$dimensional continuous distribution, any normal distribution with mean x' defines a convergent Metropolis chain, as does any other elliptically symmetric distribution. For example, define $g(x|x')$ as the density of $x = x' + L\epsilon$ where ϵ has a standard normal distribution.
3. One common use, in lower dimensional problems, is to generate an MCMC from an analytic approximation, say a normal approximation to a posterior distribution in a generalized linear model, or other model. For example, suppose $x = \theta$, a multivariate parameter of a statistical model and we have derived an asymptotic normal approximation to the posterior in a data analysis, $\theta \sim N(t, T)$ where, for example, t may be the MLE and T the inverse Hessian matrix in a standard reference analysis of a generalized linear model. Then a Metropolis that proposes states in parameter space that are oriented according to the approximate posterior dependence structure implied by T might start at $\theta^{(0)} = t$ and wander around according to a proposal distribution of the form $(\theta^{(t)}|\theta^{(t-1)}) \sim N(\theta^{(t-1)}, cT)$ for some constant c.

   Some general guidelines can be developed for this specific example, focussing on the choice of c in an attempt to balance the acceptance rate of the chain and speed convergence. Key work of Gelman, Roberts & Gilks (1995, *Bayesian Statistics 5*, eds: Bernardo *et al*, Oxford University Press), discussed in §11.9 of Gelman, Carlin, Stern & Rubin *(Bayesian Data Analysis*, 2nd Edn., 2004, Chapman & Hall), consider such the idealized situation in which the target is in fact normal, $\pi(x) = N(t, T)$, and the proposal $g(x|x') = N(x', cT)$. They derive the rule-of-thumb $c \approx 2.4/p^{1/2}$ with a focus on efficiency of the resulting chain relative to random sampling from the target.

   One very practical strategy is to create an initial normal approximation, run such a Metropolis for a while to produce updated estimates of (t, T) based on the MC draws from that chain, and "tune" the scaling constant c to achieve an acceptance rate of around 25-50%. Theoretical results (see above references) have generated useful heuristics related to this normal set-up that indicate an acceptance rate of about 25% for low dimensions, rising to about 50% in higher dimensions, is consistent with reasonably rapid convergence to the stationary distribution. After a period of such tuning and adapting both c and the estimated moments (t, T), the MH can then be run from that point on to produce a longer MC series for summary inferences (Gelman *et al*, §10.9, 2004).
4. This MCMC method is also referred to as a *random walk Metropolis-Hastings method*. The symmetric proposal distribution generates candidates according to a random walk. The location family example above clearly displays this.

Example. In the simple zero-mean AR(1) model, the approximate conditional normal/inverse gamma posterior, conditioning on the initial value x_1 of the series as fixed and uninformative about the AR parameters $\theta = (\phi, v)$ and also ignoring stationarity constraints, provides a simple initial approximation on which to base the moments (t, T) for the core of a normal proposal. This could be applied directly to the parameters θ, or, generally better, to real-valued transforms of them that make the normal proposals more tenable. In either case, the exact target posterior is easily evaluated. Write $f(\phi, v|x_{1:n})$ for the *conditional* normal/inverse gamma posterior density, so that the exact target posterior for the stationary model is just

$$\pi(\theta) \equiv p(\phi, v|x_{1:n}) \propto \begin{cases} f(\phi, v|x_{1:n})(1-\phi^2)^{1/2} \exp(\phi^2 x_1^2/(2v))/\sqrt{v}, & \text{if } |\phi| < 1 \text{ and } v > 0, \\ 0, & \text{otherwise.} \end{cases}$$

This is easily evaluated to define the accept/reject probabilities in a Metropolis using the random walk proposals $\theta^* \sim N(\theta', cT)$. Note that the constraints $|\phi| < 1$ and $v > 0$ will be applied to candidates, so that any invalid parameter values lead to zero density and are automatically rejected.

See example code on the course web site for this little example.

11.3 Metropolis-Hastings Methods

The more broadly applicable Metropolis-Hastings (MH) methods build on the same basic idea: from a current state, generate a candidate move from a proposal distribution, and accept/reject that candidate according to some measure of its importance under the target density. The setup is very similar to Metropolis methods but simply relaxes the constraint that g be symmetric, with some modification to the accept/reject test as a result.

- The *target* distribution has p.d.f. $\pi(x)$ on χ.
- A realization of a first-order Markov process is generated, $x^{(1)}, x^{(2)}, \ldots, x^{(t)}, \ldots$, starting from some initial state $x^{(0)}$.
- A *proposal distribution* with p.d.f. $g(x|x')$ is defined for all $x, x' \in \chi$. Now, for general MH methods, this can be essentially any conditional distribution.
- At step $t-1$, the current state is $x^{(t-1)}$. A *candidate state* x^* is generated from the current proposal distribution,
$$x^* \sim g(x^*|x^{(t-1)}).$$
- Compare the importance ratio $\pi(x^*)/g(x^*|x^{(t-1)})$ for the candidate state with the corresponding ratio for a "reverse" move, $\pi(x^{(t-1)})/g(x^{(t-1)}|x^*)$, via

$$\begin{aligned} \alpha_t &= \min\left\{1, \frac{\pi(x^*)}{g(x^*|x^{(t-1)})} \Big/ \frac{\pi(x^{(t-1)})}{g(x^{(t-1)}|x^*)}\right\} \\ &= \min\left\{1, \frac{\pi(x^*)}{\pi(x^{(t-1)})} \cdot \frac{g(x^{(t-1)}|x^*)}{g(x^*|x^{(t-1)})}\right\}. \end{aligned}$$

  That is,
$$\alpha_t = \alpha(x^{(t-1)}, x^*)$$
  where
$$\alpha(x', x) = \min\left\{1, \frac{\pi(x)}{\pi(x')} \cdot \frac{g(x'|x)}{g(x|x')}\right\}.$$
- The Metropolis chain moves to the candidate proposed if it has higher importance ratio (importance weight) than the current state; otherwise, it moves there with probability defined by the relative magnitudes of the importance ratios. More precisely,
$$x^{(t)} = \begin{cases} x^*, & \text{with probability } \alpha_t, \\ x^{(t-1)}, & \text{otherwise.} \end{cases}$$

Hence the MH chain evolves in a fashion similar to the Metropolis chain, but with the importance ratio guiding moves rather than the target p.d.f. alone. This allows for much greater freedom in choice of the proposal distribution (no longer required to be symmetric) and the appearance of the proposal p.d.f. values via the importance ratios that now define the accept/reject probabilities simply "corrects" the chain (we are sampling from the "wrong distribution" again) to ensure the resulting process converges to $\pi(x)$. Note the direct tie-in with ideas from both importance sampling and direct accept/reject methods.

11.3.1 Convergence of Metropolis-Hastings Methods

Again, sufficient conditions for ergodicity are that the chain can move anywhere at each step, which is ensured, for example, if $g(x|x') > 0$ everywhere. Convergence is assured and demonstrated precisely as for the Metropolis methods since MH methods generally define reversible processes and hence the stationary distribution appears in the detailed balance identity

$$p(x|x')\pi(x') = p(x'|x)\pi(x)$$

where $p(x|x')$ is the transition density induced by the construction of the chain and $p(x'|x)$ the p.d.f. of the reverse transitions.

Suppose we start the chain at $x' \sim \pi(x')$ and generate x. Suppose first that we have values such that $\alpha(x', x) > 1$. Then:

- Starting at x' the chain moves surely to the candidate $x \sim g(x|x')$. Hence $p(x|x') = g(x|x')$, and the joint density (left hand side of the detailed balance equation) is $g(x|x')\pi(x')$.
- The reverse step from x to x' would be made only with probability

$$\alpha(x, x') = \frac{\pi(x')}{\pi(x)} \frac{g(x|x')}{g(x'|x)} < 1$$

in this case. So that the reverse transition p.d.f. is

$$\begin{aligned} p(x'|x) &= \alpha(x, x')g(x'|x) + (1 - \alpha(x, x'))\delta(x' - x) \\ &= \frac{\pi(x')}{\pi(x)} \frac{g(x|x')}{g(x'|x)} g(x'|x) + 0 = \frac{\pi(x')}{\pi(x)} g(x|x'). \end{aligned}$$

So the joint p.d.f. of the reverse chain (the right hand side of the detailed balance equation) is $p(x'|x)\pi(x) = \pi(x')g(x|x')$. This agrees with the forward result and so detailed balance holds.

By symmetry the same argument applies for points (x', x) such that $\alpha(x', x) < 1$, and it follows that $\pi(x)$ is the stationary p.d.f. satisfying detailed balance, hence the limiting distribution.

11.4 Comments and Examples

1. As with the special case of Metropolis, MH methods generate replicate values of sampled states through rejection of candidates, and some attention is needed in method design to achieve reasonable acceptance and convergence rates.
2. Note that, as in importance sampling, the MH method applies in contexts where the target density (and also the proposal density) is known only up to a constant of normalization, since it is ratios of these densities that are required. This makes the approach useful in Bayesian analysis when dealing with unnormalized posteriors.
3. We have already noted that, if $g(x|x') = g(x'|x)$, then the MH method reduces to the Metropolis method.
4. The class of *independence chain* methods arise as special cases in which $g(x|x') = g(x)$ for all x', and we generate candidates as random samples from g. This obviously ties in intimately with importance sampling, and the conditions that define a useful importance sampler translate to the MH context. Common uses are in models, such as generalized linear regression models, for example, in which we can relatively easily derive analytic approximations, such as normal or T, to a posterior density, and

then - perhaps with some "fattening of the tails" - easily simulate it to generate candidates. Consider again the example of an initial asymptotic normal approximation $\theta \sim N(t, T)$ to a target posterior $\pi(\theta)$ in a statistical analysis. Suitable proposal distributions would be multivariate normals with an inflated variance matrix, say cT for some $c > 1$. Potentially improved domination of the tails of the target $\pi(\theta)$ will generated by proposals that are T distributions with the same mode t and dispersion matrix T, but some fairly low degrees of freedom.

5. Gibbs sampling can be interpreted as a special case of Metropolis-Hastings in which each sub-state is resampled in sequence within each MH iteration. In this case the proposal distribution is defined by the sequence of conditionals for sub-states under the target joint density, and there is no rejection - all candidates are accepted (see Robert and Casella §7.1.4, and also Gelman *et al* 2004, §11.5).
6. *Metropolis-Hastings within Gibbs* refers to the very common use of a Gibbs sampler within which some of the component conditional distributions are sampled via MH. Suppose the usual Gibbs framework in which the state is $x = \{x_1, \ldots, x_q\}$ and, at each Gibbs iterate t, we sequence through aiming to update $x^{(t-1)}$ to $x^{(t)}$ by sampling the component conditionals

$$p_i(x_i | x_1^{(t)}, \ldots, x_{i-1}^{(t)}, \; x_{i+1}^{(t-1)}, \ldots, x_q^{(t-1)}).$$

   Under the target $\pi(x)$ the above conditional is of course simply given from

$$p_i(x_i | x_{-i}) \propto \pi(x)$$

   with each of the elements of $x_{-i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_q\}$ set to their most recently sampled values.

   Suppose now that $p_i(x_i|x_{-i})$ is not easy to directly sample for some i. The MH-within-Gibbs method simply involves generating x_i via some specified MH method. That is, generate and accept/reject a candidate value for x_i given a specified proposal density $g_i(x_i|x_i')$ that may also depend on the current, conditional values of x_{-i}. This generates an overall modified MH method for x, and is a very commonly used approach in complex, structured models.
7. A very common use of MH is in models in which the target is given by

$$\pi(x) \propto a(x)g(x)$$

   and where

   - $g(x)$ is the p.d.f. of a distribution that is easily simulated,
   - $g(x)$ is easy to compute, and
   - $g(x)$ plays a dominant role in determining $\pi(x)$ - i.e., much of the shape and scale of $\pi(x)$ comes through $g(x)$, while $a(x)$ modifies this basic form.

   In such cases, $g(x)$ is a natural choice for a proposal density for an independence chain MH analysis, and the acceptance probability calculation simplifies to $\alpha(x, x') = \min(1, a(x)/a(x'))$. Again this ties in with similar ideas in importance sampling.

Example. In the AR(1) HMM stochastic volatility model, recall that the complete Gibbs sampler for the model in its centered parametrization samples, at each Gibbs iterated, the conditional posterior for the volatility persistence parameter ϕ given the volatility states and other parameters. This has the form $a(\phi)g(\phi)$ where $g(\phi)$ is a truncated normal density for ϕ, truncated to the stationary region or, more practically, to $0 < \phi < 1$, and where $a(\phi) \propto (1-\phi^2)^{1/2} \exp(\phi^2 x_0^2/(2v))$ based on the current Gibbs values of the initial volatility state x_0 and the AR(1) innovation variance v. In this setup it is easy to simulate $g(\phi)$ and so is an obvious, and practically effective, proposal density for an independence chain MH component for ϕ within the Gibbs sampler.

Happy Halloween!

12 Autoregressive Models

12.1 Introduction

AR(p) models for univariate time series are Markov processes with dependence of higher order than lag-1 in the univariate state space. Refer to **** Support notes on Linear Processes, AR models etc **** on course web site, and Chapters 9 and 15 of West & Harrison *Bayesian Forecasting and Dynamic Models*.

12.2 AR(p) Model Form

Data y_t are generated from an AR(p) model

$$y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \epsilon_t \qquad \text{where} \qquad \epsilon_t \sim N(0, v)$$

for $t = 1, 2, \ldots$ (and conceptually $t = 0, -1, \ldots,)$ and with $\epsilon_t \perp\!\!\!\perp \epsilon_s$ for $t \neq s$.

- *Notation:* y_t is used in place of earlier x_t for notation, and the reason will become clear.
- *Notation:* $y_t \leftarrow AR(p|\theta)$ with $\theta = (\phi, v)$ and $\phi = (\phi_1, \ldots, \phi_p)'$.
- y_t is regressed on p past, most recent, consecutive values of the process.
- Linear prediction model.
- Obvious extension of AR(1) models.
- Homogenous Markovian model - the same model applies for all t, since the parameters (ϕ, v) are constant in time.
- y_t is a linear, homogenous Gaussian process, and is time reversible. Other models with non-Gaussian innovations are linear but not reversible.

12.3 Backshift Operators and Characteristic Polynomials

- Operator B such that $By_s = y_{s-1}$ and so $B^k y_s = y_{s-k}$ for all k.
- $y_t = \sum_{j-1}^{p} \phi_j B^j y_t + \epsilon_t$
- $\Phi(B) y_t = \epsilon_t$
- *Characteristic Polynomial* $\Phi(u) = 1 - \phi_1 u - \phi_2 u^2 - \cdots - \phi_p u^p$. This is a polynomial of order p defined on $|u| < 1$.

12.4 Inversion and Moving Averages (MAs)

As in the AR(1) case, iterative substitution of y_{t-1} then y_{t-2} and so on in the right hand side of the model equation represents y_t as a linear function of more distant past y_s and $\epsilon_t, \epsilon_{t-1}, \ldots$. Formally, assuming the inversion can be performed,

$$y_t = \Phi(B)^{-1} \epsilon_t = \Pi(B) \epsilon_t$$

or

$$y_t = \epsilon_t + \pi_1 \epsilon_{t-1} + \pi_2 \epsilon_{t-2} + \cdots$$

where the (likely infinite order) polynomial $\Pi(u) = 1 + \pi_1 u + \pi_2 u^2 + \cdots$ satisfies $\Phi(u)\Pi(u) = 1$.

This gives the explicit representation of y_t as a linear combination of independent innovations, a linear (Gaussian) process. This is also referred to as a *Moving Average (MA)* representation. If the weights π_j cut-off to zero after some finite lag q, we have a finite MA(q) representation.

12.5 Stationarity

A stationary AR(p) process is such that this inversion exists. The moving average weights must decay to zero eventually, otherwise the linear combination of past innovations will explode. If the representation exists, then evidently

- $E(y_t) = 0$ for all t,
- $V(y_t) = v \sum_{j=0}^{\infty} \pi_j^2$ for all t, and this shows that the weights must decay rapidly with j,
- $Cov(y_t, y_{t-k}) = \gamma(k)$ is some function of the π_j weights, but depends only on lag k and not on t.
- These moments and all other properties are determined by the values of p, ϕ, v.

12.6 Characteristic Roots

Write $\Phi(u) = \prod_{j=1}^{p}(1 - \alpha_j u)$ where the α_i are the *reciprocals* of the roots of the characteristic polynomial, i.e., $\Phi(\alpha_j^{-1}) = 0$ for each $j = 1, \ldots, p$.

- $\Pi(u) = \Phi(u)^{-1} = \prod_{i=1}^{p}(1 - \alpha_i u)^{-1}$ exists on $|u| < 1$ if, any only if, $|\alpha_j| < 1$ for each $j = 1, \ldots, p$.
- $\{\alpha_j\}$ are the *characteristic roots* of the AR(p) model.
- Polynomial roots may be real or complex. Complex roots occur in pairs of complex conjugates.
  - A real root is any number in $-1 < r < 1$.
  - A pair of complex conjugate roots has the form $\alpha, \alpha^* = r \exp \pm i\omega$ for some modulus r, such that $0 < r < 1$, and argument (or angle) ω such that $-\pi < \omega < \pi$ (or, in some conventions, $0 < \omega < 2\pi$).
  - Matlab: *abs* and *angle*.
  - R/Splus: *Mod* and *Arg*.
  - $r \exp(\pm i\omega) = r \cos(\omega) \pm ir \sin(\omega)$.
  - $\alpha + \alpha^* = 2r \cos(\omega)$ and $\alpha\alpha^* = r^2$.
  - The period, or wavelength, of the complex roots is $2\pi/\omega$.

12.7 AR(2) Examples

When $p = 2$, $\Phi(u) = 1 - \phi_1 u - \phi_2 u^2 = (1 - \alpha_1 u)(1 - \alpha_2 u)$ may have either two, distinct real roots or one pair of complex conjugate roots. Note that $\phi_1 = \alpha_1 + \alpha_2$ and $\phi_2 = -\alpha_1\alpha_2$. If the roots are real, the autocorrelations in the process are a composition of those of two AR(1) processes with parameters given by the two real roots, and in fact the AR(2) model can be represented as the sum of two such AR processes, as we shall see below in discussion of model decompositions.

If the roots are complex, the AR(2) process is *quasi-periodic*. Sample trajectories have the appearance of "noisy" damped cosine waves of fixed wavelength $2\pi/\omega$. The "noise" is random variation in amplitude and phase of the waveform through time, injected by the innovations sequence. The damping is induced by the modulus r. A low modulus rapidly damps the waveform between time points, prior to the injection of the next innovation. A very persistent waveform, closer to a sinusoidal form, is generated in cases of r close to unity.

Model decomposition theory below shows how all AR(p) models can be decomposed, and hence understood both theoretically and from a quantitative, practical viewpoint, in terms of basic AR(1) and (almost) AR(2) processes.

12.8 Partial Inversion of AR Models

Sometimes we use fit higher-order models and then may interpret part of the structure as arising from a lower-order model with correlated innovations, i.e., ARMA processes.

For example, suppose $p = 3$ and we have one real root ρ and a pair of complex conjugate roots $\alpha, \alpha^* = r\exp(\pm i\omega)$ so that $\Phi(u) = (1-\rho u)(1-\alpha u)(1-\alpha^* u)$. Suppose that ρ is fairly small so that $(1-\rho u)^{-1} \approx 1 + \rho u + \rho^2 u^2$ for $|u| < 1$. Then can rewrite the AR model $\phi(B)y_t = \epsilon_t$ as

$$(1-\alpha B)(1-\alpha^* B)y_t \approx (1 + \rho B + \rho^2 B^2)\epsilon_t$$

or

$$y_t \approx \phi_1' y_{t-1} + \phi_2' y_{t-2} + \epsilon_t + \pi_1 \epsilon_{t-1} + \pi_2 \epsilon_{t-2}$$

where $\phi_1' = 2r\cos(\omega)$, $\phi_2' = -r^2$, $\pi_1 = \rho$ and $\pi_2 = \rho^2$.

- y_t looks like an ARMA(2,2) process - an AR(2) process in which the innovations are correlated and themselves have a second-order structure.
- Obvious extensions to ARMA(p, q) models are of interest.
- In this example with ρ quite small, the second term may be negligible and the process looks approximately like an ARMA(2,1) process.
- Very often, higher-order AR(p) models have underlying, lower-order AR structure that is somewhat obscured by measurement or timing errors that induce correlation between the innovations, and can be recovered through the device of fitting a higher-order model and then using this idea of partial inversion, at least at an exploratory level.

12.9 Reference Conditional Linear Model for Inference and Prediction

Refer back to Homework 3, Question 5, where you fitted AR(p) models to the SOI series and explored inference in the reference analysis

Under this model, *conditioning on the set of initial values* $y_1, \ldots, y_p$ and assuming we observe $n > 2p$ consecutive values, we have a model that has the form of a linear regression, and reference Bayesian inference is standard. With $y_{(p+1):n} = (y_{p+1}, y_{p+2}, \ldots, y_n)'$ and $(n-p) \times p$ autoregressive design matrix

$$H = \begin{pmatrix} y_p & y_{p-1} & \cdots & y_1 \\ y_{p+1} & y_p & \cdots & y_2 \\ \vdots & \vdots & \ddots & \vdots \\ y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{pmatrix},$$

we have the linear model expression

$$y_{(p+1):n} = H\phi + \epsilon_{(p+1):n} \quad \text{with} \quad \epsilon_{p+1:n} \sim N(0, vI).$$

The conditional reference posterior for θ has the compositional form

$$\begin{aligned} (\phi | v, y_{1:n}) &\sim N(b, vB^{-1}), \\ (v^{-1} | y_{1:n}) &\sim Ga((n-2p)/2, Q(b)/2), \end{aligned}$$

with

$$Q(\phi) = (y_{(p+1):n} - H\phi)'(y_{(p+1):n} - H\phi), \quad B = H'H \quad \text{and} \quad b = B^{-1}H' y_{(p+1):n}.$$

Here b is the conditional MLE, LSE and reference posterior mean and mode for ϕ, and $Q(b)$ is the residual sum of squares from the conditional regression analysis.

This is all (trivially) coded in the existing little Matlab and Splus functions. They also provide for posterior simulation - direct sampling from this conditional reference posterior, and simulation of predictive distributions - *"sampling the future"*. Again, refer back to Homework 3, Question 5 and the analysis code and solutions.

12.10 Linear Systems: State Space Representation

Introduce the $p-$dimensional state vector $x_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})'$ for all t. Then the AR(p) model may be re-expressed as

$$\begin{aligned} y_t &= F'x_t \\ x_t &= Gx_{t-1} + F\epsilon_t \end{aligned}$$

where

$$F = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Here G is the *state evolution, or transition matrix* in the extended state space representation. Note that this maps the state from one to p dimensions and so converts the p^{th} order Markovian dependence to a first order dependence.

- This representation is one reason for the notational use of y_t for data, since now x_t is the $p-$vector state variable.
- An easy extension to a latent AR process - HMM with the underlying hidden state being a higher-order model - is given by the "AR(p) in noise" extension in which $y_t = F'x_t + \nu_t$.

12.11 Forecast Function and Eigenstructure

Insight into the dependence structure is generated by inspection of the *forecast function* $f_t(k) = E(y_{t+k}|y_{1:t})$ of the process as a function of the "look-ahead" horizon $k = 1, 2, \ldots$. It is easily seen that $E(x_{t+k}|y_{1:t}) = G^k x_t$ so that $f_t(k) = F'G^k x_t$.

Now G is a square matrix. Assume that G has distinct, non-zero eigenvalues - this will be almost surely the case when a ϕ is derived from a model fit to real data. The G has an eigendecomposition $G = E\Lambda E^{-1}$ where the $p \times p$ eigenvector matrix E has columns that are the eigenvectors of the corresponding elements of the diagonal matrix Λ. That is, $G\lambda_j = e_j\lambda_j$ where $E = [e_1, \ldots, e_p]$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$. The eigenvalues and vectors can be real or complex valued. Since G is real valued, any complex eigenvalues must occur in conjugate pairs. Then $G^k = E\Lambda^k E^{-1}$ and so

$$f_t(k) = \sum_{j=1}^{p} c_{t,j}\lambda_j^k$$

for some (real or complex) numbers $c_{t,j}$ that depend on E and x_t.

- A real valued eigenvalue λ_j contributes a term λ_j^k to the forecast function. This will explode unless $|\lambda_j| < 1$, and will otherwise decay to zero with increasing k.

- A pair of complex conjugate eigenvalues $r\exp(\pm i\omega)$ will contribute terms proportional to

$$r^k \exp(\pm i\omega k) = r^k\{\cos(\omega k) \pm i\sin(\omega k)\}.$$

  Since the forecast function is real the two terms must have canceling imaginary components, so leading to a term proportional to

$$r^k \cos(\omega k + a_t)$$

  for some phase a_t that depends on the current state x_t. Unless $|r| < 1$ this will also be explosive; if $|r| < 1$, then the form is a damped cosine wave of fixed wavelength $2\pi/\omega$ and current state dependent phase.
- y_t is a linear combination of these processes.
- With models of higher order p, many such component processes exist. Some may be of real practical (physical, economic) significance and interest, whereas some - often those of very low moduli, will often represent high frequency or short-term noise.
- Inference on (ϕ, v) (and p) leads to inference on the eigenstructure of the model and hence on the underlying components in the time series. Simulation of the posterior for (ϕ, v) leads easily to simulated values of the eigenvalues and so forth. Plug-in estimates of ϕ, such as the reference posterior mean b, provide a start.

Key result: The eigenvalues λ_j are precisely the characteristic roots: $\lambda_j = \alpha_j$ for $j = 1, \ldots, p$. *(See Homework #8).* Hence a stationary AR(p) is characterized by a G matrix that has all eigenvalues of less than unit modulus, whether real or complex. In such cases, the forecast function is clearly "well-behaved", with the components damping out as k increases, eventually leading to $f_t(k) \to 0 = E(y_{t+k})$ as k increases.

12.12 Autocorrelations

In the state space representation, $E(x_t) = 0$ and $V(x_t) = S$ where S satisfies $S = GSG' + U$ and U is the $p \times p$ matrix of zeros except for $U_{1,1} = v$. We can see the form of the a.c.f. of y_t easily. Since $y_t = x_t'F$ we have

$$\gamma(k) = E(y_{t+k}y_t) = F'E(x_{t+k}x_t')F.$$

Now $x_{t+k} = G^k x_t +$ terms involving $\epsilon_{t+1}, \ldots, \epsilon_{t+k}$. Hence $E(x_{t+k}x_t') = G^k S$ and so

$$\gamma(k) = F'G^k SF = \sum_{j=1}^{p} g_j \lambda_j^k$$

for some constant g_j that depend on (ϕ, v). As a result, the autocorrelations $\rho(k)$ have the same *form* as a function of lag k; that is, precisely the same form as the forecast function. Autocorrelations of AR(p) processes are a mixture of damped AR(1)-like terms that decay exponentially (real eigenvalues) and may oscillate (real negative eigenvalues), and damped AR(2)-like cosine forms (complex conjugate pairs of eigenvalues).

12.13 AR Model and Process Decompositions

The eigentheory is completed with an explicit representation of y_t in terms of underlying (latent, but identifiable) components of AR(1) and/or AR(2) (actually ARMA(2,1)) forms.

- Define a transformed $p-$vector state variable $z_t = E^{-1}x_t$. Also set $F_o = E'F$ and $F_e = E^{-1}F$.
- Then $y_t = F_o'z_t$ and $z_t = \Lambda z_{t-1} + F_e\epsilon_t$ for all t.

- Elements of $z_t = (z_{t,1}, z_{t,2}, \ldots, z_{t,p})'$ are individual AR(1) processes, with $z_{t,j}$ having AR coefficient λ_j. They are correlated since they are driven by the same innovations.
- Real eigenvalues lead to real components: $z_{t,j} \leftarrow AR(1|(\lambda_j, v_j))$.
- Complex eigenvalues $r\exp(\pm i\omega)$ lead to pairs of *complex* and conjugate AR(1) processes. The linear combination of two such processes must be real, and this leads to a real component that has the structure of an ARMA(2,1) component in which the AR(2) part is damped by $r > 0$, and quasi-periodic with fixed frequency ω, i.e., wavelength $2\pi/\omega$, but time-dependent amplitude and phase.

See Chapters 9 - especially - and 15 of West & Harrison Bayesian Forecasting and Dynamic Models *for further details*.

13 Sequential Learning and TVAR models

13.1 Sequential Representations of Learning in Autoregressive Models

Investigation of the details of how posterior distributions for (ϕ, v) are sequentially updated as new data arises is useful for, among other things, defining a framework for extension to time-varying parameter models.

We can recast the model in the following regression form. With the $p-$dimensional state vector $x_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})'$ for all t we have

$$y_t = x_{t-1}'\phi + \epsilon_t$$

and, under the regression analysis as already described, the posterior for $\theta = (\phi, v)$ has the conjugate normal-inverse gamma form (see §12.9). We need a change of notation, and now write the posterior as

$$\begin{aligned}(\phi|v, y_{1:t}) &\sim N(m_t, vM_t),\\ (v^{-1}|y_{1:t}) &\sim Ga(n_t/2, n_t s_t/2).\end{aligned}$$

This conditional posterior is valid for all times t, and so the defining quantities $\{m_t, M_t, n_t, s_t\}$ are naturally related as time t varies. In particular, the "time t update" involves the mapping from their values at $t-1$ - based on data $y_{1:(t-1)}$ - to their values at t, representing the additional information provided by the time t observation y_t.

The following key theory defines the sequential updating, and is quite general.

Suppose $p(\phi, v|y_{1:(t-1)})$ has the above normal-inverse gamma from with defining parameters

$$\{m_{t-1}, M_{t-1}, n_{t-1}, s_{t-1}\}.$$

Then:

- The *one-step ahead forecast distribution* conditional on v is

  $$(y_t|y_{1:(t-1)}, v) \sim N(x_{t-1}'m_{t-1}, q_t v)$$

  with $q_t = 1 + x_{t-1}'M_{t-1}x_{t-1}$.
- The *time t posterior distribution* is $p(\phi, v|y_{1:t}) = p(\phi|v, y_{1:t})p(v|y_{1:t})$ and is normal-inverse gamma with parameters $\{m_t, M_t, n_t, s_t\}$ that are computed as follows:

  - $m_t = m_{t-1} + A_t e_t$,
  - $M_t = M_{t-1} - A_t A_t' q_t$,
  - $n_t = n_{t-1} + 1$,
  - $s_t = (n_{t-1}s_{t-1} + e_t^2/q_t)/(n_{t-1} + 1)$.

  with

  - $e_t = y_t - x_{t-1}'m_{t-1}$, the *one-step ahead forecast error,* and
  - $A_t = M_{t-1}x_{t-1}/q_t$, the *adaptive coefficient $p-$vector,*

These results flow from standard normal theory and Bayes' theorem (see Multivariate Normal Theory notes). Some comments and alternative expressions are of interest:

- The update for m_t is a "predictor/corrector" form: The prior or "predicted" value for ϕ, namely m_{t-1}, is corrected by the weighted forecast error. A large forecast error implies a large correction, and vice-versa.

- $M_t < M_{t-1}$, so that we apparently always gain information.
- The degrees of freedom n_t increases by one per observation.
- The error variance estimate s_t is updated as a weighted average of the predicted estimate s_{t-1} and the forecast error scaled by q_t. A larger forecast error leads to an inflation of the estimate of error variance.
- All conditional normal distributions convert to the corresponding T distributions on marginalization over v. For example,
  - $(y_t|y_{1:(t-1)})$ is univariate T on n_{t-1} degrees of freedom, with mean $x'_{t-1}m_{t-1}$ and scale $q_t s_{t-1}$, whereas
  - the posterior $p(\phi|y_{1:t})$ is $p-$variate T distributed on n_t degrees of freedom with mean m_t and scale matrix $M_t s_t$.
- Very important alternative representations of $\{m_t, M_t\}$ are the forms derived directly from Bayes' theorem, namely

$$m_t = M_t(M_{t-1}^{-1}m_{t-1} + x_{t-1}y_t) \qquad \text{and} \qquad M_t^{-1} = M_{t-1}^{-1} + x_{t-1}x'_{t-1}.$$

These are standard formulæ, though the new, alterative representations above are both computationally more efficient and numerically more stable as no matrix inversions are required. Some additional practical points related to initialization:

- Since the updating results hold true for all t, it is clear that we can now consider analysis based on any initial prior of the normal-inverse gamma form at $t = p$. We need to consider the initial point as $t = p$ since the required regression vector x_t is available only for $t \geq p$. That is, analysis can be initialized at any specified values of $\{m_p, M_p, n_p, s_p\}$ prior to implementing the sequential form of the analysis beginning with the "first" observation at $t = p + 1$.
- An alternative initialization involves fitting the reference posterior distribution based on an initial set of $q > p$ observations, and then beginning the sequential analysis at a "first" time point $t = q+1$ with $\{m_q, M_q, n_q, s_q\}$ defined by the reference posterior.

13.1.1 Useful Linear Algebraic Identities

For reference, the latter equation and the alterative form of M_t are related to a key and broadly useful matrix identity, in this case - dropping the time indices for generality and clarity - simply

$$(M^{-1} + xx')^{-1} = M - Mxx'M/(1 + x'Mx)$$

for any positive definite and symmetric $p \times p$ matrix M and $p-$vector x.

A more general version involves a $p \times q$ matrix X and a $q \times q$ positive definite symmetric matrix V, when the identity is

$$(M^{-1} + XV^{-1}X')^{-1} = M - MX(V + X'MX)^{-1}X'M.$$

13.2 Model Likelihood Function

One most useful side product of the sequential analysis is an easy evaluation of the *model likelihood* - the joint density of observations unconditional on model parameters. Simply note that, at each time t we obtain the one-step ahead forecast, or predictive density $p(y_t|y_{1:(t-1)})$. This is the univariate T p.d.f.

$$p(y_t|y_{1:(t-1)}) = c(n_{t-1})(s_{t-1}q_t)^{-1/2}\left(1 + \frac{e_t^2}{n_{t-1}s_{t-1}q_t}\right)^{-(n_{t-1}+1)/2}$$

where $c(\nu)$ is the constant of normalization of the T density on ν degrees of freedom,

$$c(\nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}}.$$

Then, the joint p.d.f. of any set of observations $y_{k:n}$ is given by composition as

$$p(y_{k:n}|y_{1:k-1}) = \prod_{t=k}^{n} p(y_t|y_{1:(t-1)}).$$

For example, if we start with $t = p+1$ and some initial values of parameters $\{m_p, M_p, n_p, s_p\}$ as discussed above, then $k = p+1$ - the above expression defines the joint density of the data from thereon.

The value of the joint density function is also named the *marginal likelihood* of the model. For example, we may rerun the analysis at different values of the model order p, and then $p(y_{1:n})$ is actually $p(y_{1:n}|p)$, the value of the data density conditional on that value of p. As we change p and rerun the analysis, this maps out the likelihood function over p for assessment of model order. A prior distribution over p might then be used to convert to a posterior distribution for model order, for example.

13.3 TVAR Models

Some non-stationary phenomena can be regarded as *"locally stationary"* in the sense that a standard model, such as an AR(p) model, provides an adequate and useful functional form for the data generating process, but its adequacy relies on permitting the defining parameters to take different values over time. This simple concept is in fact quite powerful in some applications, and in fact underlies very widely used methods of local smoothing in time series in many areas of social and natural sciences, engineering signal processing, and short-term forecasting in business and economics. The comprehensive framework of *dynamic state space models* (West and Harrison, 1997) demonstrates the broad scope of applicability of the concept, as well as various models classes.

A core class of such non-stationary models is the class of *Time-Varying Autoregressive Models (TVAR)*, a rather simple but nevertheless very useful extension of AR models.

13.3.1 Concept: Random Drift in Parameters

The key concept is simply to allow model parameters to vary randomly in time, according to a stochastic process model. We will examine just random walk models in detail, but the concept is more general. The AR(p) model $y_t \leftarrow AR(p|(\phi, v))$ is extended to the TVAR(p) model in which, at any time t,

$$y_t = x'_{t-1}\phi_t + \epsilon_t \equiv \sum_{j=1}^{p} \phi_{t,j} y_{t-j} + \epsilon_t$$

where

$$\phi_t = (\phi_{t,1}, \phi_{t,2}, \ldots, \phi_{t,p})'$$

is the - now - time-varying AR parameter vector. Any model for time variation might be considered. A simple random walk model has several points of recommendation. Such a model has the form

$$\phi_t = \phi_{t-1} + \omega_t$$

where ω_t is a sequence of zero-mean $p-$vector variates with $\omega_t \perp\!\!\!\perp \omega_s$ for $t \neq s$. The ω_t variates constitute a sequence of random "shocks" to the model that define the temporal evolution of the AR parameters.

This random walk parameter evolution is simple and has the attraction that it is neutral with respect to directions of change in parameters, and allows them to "wander" freely over time so permitting substantial change in model form over long periods.

13.3.2 Sequential Learning in TVAR Models

Suppose that, at time $t-1$, information on the *current AR parameter* is summarized via $(\phi_{t-1}|v, y_{1:(t-1)}) \sim N(m_{t-1}, vM_{t-1})$ while $(v^{-1}|y_{1:(t-1)}) \sim Ga(n_{t-1}/2, n_{t-1}s_{t-1}/2)$ just as in the static parameter model. Then:

- Assume that $\phi_t = \phi_{t-1} + \omega_t$ where

  $$(\omega_t|v, y_{1:(t-1)}) \sim N(0, vW_t)$$

  independently of ϕ_{t-k} for $k \geq 1$. Note that we include v as a constant factor in the variance of ω_t for consistency throughout.
- It easily follows that the "evolution" of the parameter to time t results in a (prior) distribution for the changed value that is just

  $$(\phi_t|v, y_{1:(t-1)}) \sim N(m_{t-1}, vM_{t|t-1})$$

  where $M_{t|t-1} = M_{t-1} + W_t$.

- Observing y_t, $p(\phi_t|v, y_{1:t}) \sim N(m_t, vM_t)$ where the update equations are precisely as in the AR model but now the initial distribution has a variance matrix component $M_{t|t-1}$ in place of M_{t-1}.
- The update for $p(v|y_{1:t})$ is also essentially as in the AR model but again with the small change that $M_{t|t-1}$ replaces M_{t-1}.

In summary, the sequential learning proceeds as follows:

1. The *time* $t-1$ *posterior* is parameterized by $\{m_{t-1}, M_{t-1}, n_{t-1}, s_{t-1}\}$;
2. The *one-step ahead prior at* $t-1$ is

$$\begin{aligned}(\phi_t|v, y_{1:(t-1)}) &\sim N(m_{t-1}, vM_{t|t-1}),\\ (v^{-1}|y_{1:(t-1)} &\sim Ga(n_{t-1}, n_{t-1}s_{t-1});\end{aligned}$$

3. The *one-step ahead forecast distribution* conditional on v is

$$(y_t|y_{1:(t-1)}, v) \sim N(x'_{t-1}m_{t-1}, q_t v)$$

   with $q_t = 1 + x'_{t-1}M_{t|t-1}x_{t-1}$.
4. The *time* t *posterior distribution* is $p(\phi, v|y_{1:t}) = p(\phi|v, y_{1:t})p(v|y_{1:t})$ and is normal-inverse gamma with parameters $\{m_t, M_t, n_t, s_t\}$ that are computed as follows:

   - $m_t = m_{t-1} + A_t e_t$,
   - $M_t = M_{t|t-1} - A_t A'_t q_t$,
   - $n_t = n_{t-1} + 1$,
   - $s_t = (n_{t-1}s_{t-1} + e_t^2/q_t)/(n_{t-1}+1)$.

   with

   - $e_t = y_t - x'_{t-1}m_{t-1}$, the *one-step ahead forecast error,* and
   - $A_t = M_{t|t-1}x_{t-1}/q_t$, the *adaptive coefficient* $p-$*vector,*

Again, these results flow from standard normal theory and Bayes' theorem (see Multivariate Normal Theory notes). Also, again exactly as in the static AR case, important alternative representations of $\{m_t, M_t\}$ are the forms derived directly from Bayes' theorem, namely

$$m_t = M_t({M_{t|t-1}}^{-1}m_{t-1} + x_{t-1}y_t) \qquad \text{and} \qquad M_t^{-1} = {M_{t|t-1}}^{-1} + x_{t-1}x'_{t-1}.$$

Note that the AR(p) model is, of course, recovered as the special case in which $W_t = 0$ so that $M_{t|t-1} = M_{t-1}$. Otherwise, the model now allows for parameter change through non-zero W_t matrices. Controlling the degree of change is very often desirable, in order that a model parametrization change smoothly consistent with scientific context. Within a Gaussian framework, this involves care in specifying or controlling the estimation of these variance matrices.

13.3.3 Concept: Random Parameter Change as Loss of Information

Between times points $t-1$ and t, the increased in variance from M_{t-1} to $M_{t|t-1} = M_{t-1} + W_t$ reflects a decreased precision, i.e., loss of information. This concept of information loss is key to a parsimonious and efficient method of structuring parameter change models - the notion of *information discounting*. One specific application of the more general concept of *variance matrix discounting* (West and Harrison, 1997, chapter 6) to structure and specify such random-change models - in fact, the simplest such approach - is to

assume a constant *rate of loss of information* about parameters over time. For a fixed, specified *discount factor* δ with $0 < \delta < 1$, specify

$$W_t = M_{t-1}(\delta^{-1} - 1).$$

That is, the loss of information W_t is a (usually rather small) fraction of the existing information M_{t-1}. For instance, $\delta = 0.99$ implies a per period loss of information of about 1%, whereas $\delta = 0.9$ leads to information attrition at a rate of about 11% per period.

One key implication is that, simply,

$$M_{t|t-1} = M_{t-1}/\delta,$$

the discount factor inducing a (usually small) inflation in the elements of the variance matrix between time points. The discount factor δ can be specified, and analysis repeated with differing values. The theory of sequential/compositional computation of the model likelihood applies directly to provide for assessment of different values along with the model order p.

14 Data Arrays and Decompositions

14.1 Variance Matrices and Eigenstructure

Consider a $p \times p$ positive definite and symmetric matrix V - a model parameter or a sample variance matrix. The eigenstructure is of interest in understanding patterns of association and underlying structure that may be lower dimensional, in the sense that highly correlated - collinear - variables may be "driven" by a common underlying but unobserved factor, or simply redundant measures of the same phenomenon.

- Write
$$V = EDE'$$
where $D = \text{diag}(d_1, \ldots, d_p)$ is the diagonal matrix of eigenvalues of V and the corresponding eigenvectors are the columns of the orthogonal matrix E. Inversely, $E'VE = D$.
- If V is the variance matrix of a generic random $p-$vector x, then E maps x to uncorrelated variates and back; that is, there exists a $p-$vector f such that $V(f) = D$ and $x = Ef$, or $f = E'x$. The representation $x = Ef$ may be referred to as a *factor decomposition* of x; the uncorrelated elements of f are factors that, through the linear combinations defined by the map E, generate the patterns of variation and association in the elements of x. The j^{th} factor in f impacts the i^{th} element of x through the weight $E_{i,j}$, and for this reason E may be referred to as the *factor loadings* matrix.
- The factors with largest variances - the largest eigenvalues - play dominant roles in defining the levels of variation and patterns of association in the elements of x. Factor i contributes $100d_i/\sum_{j=1}^p d_j\%$ of the *total variation* in V, namely $\sum_{j=1}^p d_j = \text{tr}(V)$.
- If V is singular - rank deficient of rank $r < p$ - the same structure exists but $p - r$ of the eigenvalues are zero. Now $D = \text{diag}(d_1, \ldots, d_r)$ represents the non-zero and positive eigenvalues, and E is no longer square but $p \times r$ with $E'E = I$, now the $r \times r$ identity. Further, $x = Ef$ and $f = E'x$ where f is a factor vector with $V(f) = D$. This clearly represents the precise collinearities among the elements of x - there are only r free dimensions of variation. In non-singular cases, very small eigenvalues indicate a context of high collinearities, approaching singularity.
- This decomposition - both the eigendecomposition of V and the resulting representation $x = Ef$ - is also known as the *principal component decomposition*. Principal component analysis (PCA) involves evaluation and exploration of the empirical factors computed based on a sample estimate of the variance matrix of a $p-$dimensional distribution.

14.2 Data Arrays, Sample Variances and Singular Value Decompositions

Consider the data array from n observations on p variables, denoted by the $n \times p$ matrix X whose rows are samples and columns are variables. Observation/case i has values in the $p-$vector x_i, and x_i' is the i^{th} column of X. The $p \times n$ matrix X' has variables as rows, and n samples as columns x_i.

Assume the variables are centered - i.e., have zero mean, or that the sample means have been subtracted - so that sample covariances are represented in the $p \times p$ matrix $V = S/n$ where $S = X'X = \sum_{i=1}^n x_i x_i'$. (The divisor could be taken as $n - 1$, as a matter of detail.)

- V and S have the same eigenvectors and eigenvalues that are that same up to the factor n, i.e., $V = EDE'$ and $S = ED_sE'$ where $D_s = nD$. This holds whether or not S, and so V, is of full rank: E is $p \times r$ of rank r and $D = \text{diag}(d_1, \ldots, d_r)$ with positive values. The rank r of S cannot, of course, exceed that of X, so $r \leq \min(p, n)$. In particular, if $p > n$ then $r \leq n < p$. That is, the rank is at most the sample size when there are more variables than samples.

- The *singular value decomposition* of the data matrix X is

$$X' = EF$$

where the $r \times n$ matrix F is such that FF' is diagonal. In fact, we see that $F = E'X'$ so that $FF' = E'SE = D_* = nD$. The r elements nd_i are also known as the *singular values* of X.

  - A more common form of the SVD is

$$X' = ED_*^{1/2}F_*$$

where the $r \times n$ matrix $F_* = D_*^{-1/2}F$ is such that $F_*F_*' = I$, the $r \times r$ identity.
  - For example, the Matlab and R *svd* functions generate outputs in this form. The rows of F_* simple represent standardized (unit variance) versions of the r factors in F.
  - In cases of $p < n$, both X' and E are $p \times n$ matrices, having more columns than rows - they are "long and skinny" matrices.
  - In cases of $p > n$, r can be no more than the sample size. Then both X' and E are "tall and skinny", with E is $p \times r$ having possibly fewer than n columns in rank reduced cases.
  - Standard SVD routines of software packages generally produce redundant decompositions and the computation is inefficient. For example, in cases with $p > n$, the standard Matlab function returns E of dimension $p \times p$ and $D_*^{1/2}$ as $p \times n$ with the lower $p - n$ rows filled with zeros. The function can be flagged to produce E of dimension $p \times n$ and just the reduced $D_s^{1/2}$ with the n relevant eigenvalues. Check the documentation in Matlab and R; see also the cover Matlab function *svd0* on the course web site.

- Write $F = (f_1, \ldots, f_n)$ so that $x_i = Ef_i$ and $f_i = E'x_i$. The f_i are the n sample values of the *singular factor* $p-$vectors, and E provides the loadings of the data variables on the singular factors.

Finally, consider the precision matrix corresponding to V. We have $K = V^-$ which is the regular inverse if V is non-singular, or the generalized inverse otherwise (recall that the generalized inverse satisfies $VV^-V = V$ and $V^-VV^- = V^-$.) With $V = EDE'$ we have

$$K = ED^-E'$$

where:

- if V is non-singular, then E is $p \times p$ and $D = D^{-1} = \text{diag}(1/d_1, \ldots, 1/d_p)$;
- if V is singular of rank $r < p$, then E is $p \times r$ and $D^- = \text{diag}(1/d_1, \ldots, 1/d_r)$.

Note how the patterns of loadings of variables on factors, defined by the elements of E, also plays major roles in defining the elements of the precision matrix.

See the course data page for exploration of patterns of association in time series exchange rate returns, and some exploratory Matlab code.

15 Wishart Distributions: Variance and Precision Matrices

The Wishart distributions arise as models for random variation and descriptions of uncertainty about variance and precision matrices. They are of particular interest in sampling and inference on covariance and association structure in multivariate normal models, and in ranges of extensions in regression and state space models.

15.1 Definition and Structure

Suppose that Ω is a $p \times p$ symmetric matrix of random quantities

$$\Omega = \begin{pmatrix} \omega_{1,1} & \omega_{1,2} & \omega_{1,3} & \cdots & \omega_{1,p} \\ \omega_{1,2} & \omega_{2,2} & \omega_{2,3} & \cdots & \omega_{2,p} \\ \omega_{1,3} & \omega_{2,3} & \omega_{3,3} & \cdots & \omega_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_{1,p} & \omega_{2,p} & \omega_{3,p} & \cdots & \omega_{p,p} \end{pmatrix}.$$

Suppose that the joint density of the $p(p+1)/2$ univariate elements defining Ω is given by

$$p(\Omega) = c|\Omega|^{(n-p-1)/2} \exp\{-\mathrm{tr}(\Omega A^{-1})/2\}$$

for some constant *degrees of freedom* $n \geq p$ and $p \times p$ positive definite symmetric matrix A, and that this density is defined and non-zero only when Ω is positive definite, and hence non-singular. This is the p.d.f. of a Wishart distribution for Ω. The Wishart is a multivariate extension of the gamma distribution, as the form of the p.d.f. intimates.

Some notation, comments and key properties are noted (see Lauritzen, 1996, *Graphical Models* (O.U.P.), Appendix C, for good and detailed development of many aspects of the theory of normal and Wishart distributions.)

- The standard notation is $\Omega \sim W_p(n, A)$.
- A is the *location matrix* parameter of the distribution.
- $E(\Omega) = nA$ and $E(\Omega^{-1}) = A^{-1}/(n-p-1)$ (the latter only defined when $n > p+1$.)
- The random matrix $\Sigma = \Omega^{-1}$ has an *inverse Wishart distribution*.
- The normalizing constant c is given by

$$c^{-1} = |A|^{n/2} 2^{np/2} \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma((n+1-i)/2).$$

- In the exponent of the p.d.f. , $\mathrm{tr}(\Omega A^{-1}) = \mathrm{tr}(A^{-1}\Omega)$.
- The distribution is proper and defined via the p.d.f. if and only if the degrees of freedom is no less than the dimension, $n \geq p$, but then applies for any value of n, not only integer values.
- The eigendecomposition of Ω is $\Omega = \Phi\Delta\Phi'$ where Φ is the $p \times p$ orthogonal matrix whose columns are eigenvalues of Ω, and $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_p)$ are the positive eigenvalues. If $(a_1, \ldots, a_p)$ are the (also positive) eigenvalues of A, then

$$p(\Omega) \propto \{\prod_{i=1}^{p} \delta_i^{(n-p-1)/2} a_i^{-n/2}\} \exp\{-\mathrm{tr}(\Omega A^{-1})/2\}.$$

The Wishart distribution is a multivariate version of the gamma distribution. Further, marginal distributions of diagonal elements and block diagonal elements of Ω are also Wishart distributed. Specifically:

- If $p = 1$, write $\omega = \Omega$ and $a = A$, both now scalars. The p.d.f. shows that $\omega \sim Ga(n/2, 1/(2a))$ or $\omega = a\kappa$ where $\kappa \sim \chi^2_n$.

- Partition Ω as

$$\Omega = \begin{pmatrix} \Omega_{1,1} & \Omega_{1,2} \\ \Omega'_{1,2} & \Omega_{2,2} \end{pmatrix}$$

where $\Omega_{1,1}$ is $q \times q$ with $q < p$, $\Omega_{2,2}$ is $(p-q) \times (p-q)$ and $\Omega_{1,2}$ is $q \times (p-q)$. Partition A conformably, with elements $A_{1,1}, A_{2,2}$ and $A_{1,2}$. Then

$$\Omega_{1,1} \sim W_q(n, A_{1,1}) \quad \text{and} \quad \Omega_{2,2} \sim W_{p-q}(n, A_{2,2}).$$

- The diagonal elements have gamma marginal distributions, $\omega_{i,i} \sim Ga(n/2, 1/(2a_{i,i})$ where $a_{i,i}$ is the i^{th} diagonal element of A. That is, $w_{i,i} = a_{i,i}k_i$ where $k_i \sim \chi^2_n$.

These are just a few key properties of the Wishart distribution, there being much more theory of relevance in multivariate analysis and also statistical modelling that relates to the joint and conditional distributions of matrix sub-elements of Ω. In particular, Bayesian analysis of Gaussian graphical models relies heavily on such structure for both graphical model development and for specification of prior distributions over graphical models (see Lauritzen, 1996, *Graphical Models* (O.U.P.), Appendix C, for summary of key theoretical results.)

15.2 Wishart Sampling Distributions for Sample Variance Matrices

The Wishart distribution arises naturally as the sampling distribution of (to a constant) sample variance matrices in multivariate normal populations, as follows:

1. Suppose n observations $x_i \sim N(0, \Sigma)$ with $x_i \perp\!\!\!\perp x_j$ for $i \neq j$, and

$$S = \sum_{i=1}^{n} x_i x'_i = X'X$$

where X is the $n \times p$ data matrix whose rows are x'_i. The usual sample variance matrix is then $\hat{\Sigma} = S/n$. This is a sufficient statistic for Σ and the MLE of Σ. We have

$$(S|\Sigma) \sim W_p(n, \Sigma)$$

with $E(S|\Sigma) = n\Sigma$ so that $\hat{\Sigma}$ is an unbiased estimate of Σ.

2. Suppose n observations $x_i \sim N(\mu, \Sigma)$ with $x_i \perp\!\!\!\perp x_j$ for $i \neq j$, and

$$S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' = \tilde{X}'\tilde{X}$$

where $\tilde{X}$ is the $n \times p$ centered data matrix whose rows are $(x_i - \bar{x})'$. The usual sample variance matrix is then $\hat{\Sigma} = S/(n-1)$ and we have $S \perp\!\!\!\perp \bar{x}$ with

$$(S|\Sigma) \sim W_p(n-1, \Sigma),$$

and now $E(S|\Sigma) = (n-1)\Sigma$ so that $\hat{\Sigma}$ is an unbiased estimate of Σ.

15.3 Wishart Priors and Posteriors for Precision Matrices in Multivariate Normal Models

Consider a random sample $x_{1:n}$ from the $p-$dimensional normal distribution with zero mean, $(x_i|\Sigma) \sim N(0, \Sigma)$, and set $\Omega = \Sigma^{-1}$ for the precision matrix, supposing Σ and Ω to be non-singular.

- The likelihood function is
$$p(x_{1:n}|\Omega) \propto |\Omega|^{n/2} \exp\{-\text{tr}(\Omega S)/2\}$$
where
$$S = \sum_{i=1}^{n} x_i x_i' = X'X$$
where X is the $n \times p$ data matrix.
- The standard reference prior is $p(\Omega) \propto |\Omega|^{-(p+1)/2}$ over the space of positive definite symmetric matrices. This leads to the standard reference posterior for a normal precision matrix
$$p(\Omega|x_{1:n}) \propto |\Omega|^{(n-p-1)/2} \exp\{-\text{tr}(\Omega S)/2\}$$
so that $(\Omega|x_{1:n}) \sim W_p(n, S^{-1})$. Also, Σ has an inverse Wishart posterior distribution. Posterior expectations are
$$E(\Omega|x_{1:n}) = nS^{-1} = \hat{\Sigma}^{-1}$$
and
$$E(\Sigma|x_{1:n}) = E(\Omega^{-1}|x_{1:n}) = S/(n-p-1) = (n/(n-p-1))\hat{\Sigma}$$
if $n > p+1$. The sample variance matrix $\hat{\Sigma}$ is the *harmonic posterior mean* of Σ.

In the case when also estimating the normal mean, $(x_i|\mu, \Sigma) \sim N(\mu, \Sigma)$, the reference prior is $p(\mu, \Omega) \propto |\Omega|^{-(p+1)/2}$ and, on integrating over μ, the implied reference posterior is similar to that of the known mean case, viz $(\Omega|x_{1:n}) \sim W_p(n-1, S^{-1})$ where now S is the centered sum of squares with each x_i replaced by $x_i - \bar{x}$.

The Wishart is also the conjugate proper prior for normal precision matrices, and much use of this fact is made in Bayesian analysis of Gaussian graphical models as well as state space modelling for multivariate time series (REFERENCES). In particular, with a prior $\Omega \sim W_p(n_0, A_0)$ where $A_0 = S_0^{-1}$ for some "prior sum of squares" matrix S_0 and "prior sample size" n_0, the posterior based on the above likelihood function is clearly $W_p(n_1, A_1)$ where $n_1 = n_0 + n$ and $A_1 = (S_0 + S)^{-1}$.

15.4 Constructive Properties and Simulating Wishart Distributions

A fundamental and practically critical property of the family of Wishart distributions is standardization. Just as we standardize normal distributions to zero mean and unit scale, we standardize Wishart distributions to identity location matrices. This is one use of a more generally useful property of transformations.

- Suppose $\Omega \sim W_p(n, A)$.
- For any $q \times p$ matrix C with $q \leq p$, we have $C\Omega C' \sim W_q(n, CAC')$. (It turns out that this extends to $q > p$ when the implied distribution is a singular Wishart, as discussed below.)
- If $q = p$ and C is such that $CAC' = I$, we have the standard Wishart, $W_p(n, I)$.
- Conversely, suppose that $\Psi \sim W_p(n, I)$ and $A = PP'$ for any non-singular $p \times p$ matrix P. (i.e., set $C^{-1} = P$ above). Then $\Omega = P\Psi P' \sim W_p(n, A)$.

This shows how to simulate $W_p(n, A)$ for any location matrix A based on samples from the standard Wishart. The matrix P can be any non-singular square root of A, such as the Cholesky factor of A when A is non-singular or, more generally, the factor generated from the singular value decomposition of A. The latter will apply in singular and non-singular cases. That is, if $A = EBE'$ with $p \times p$ eigenvector matrix E and $p \times p$ diagonal matrix of positive eigenvalues B, then we can use $P = EB^{1/2}$. Compared to the Cholesky decomposition this has an advantage of being numerically more stable and also extending to cases in which A is singular, or close to singular.

The *Bartlett decomposition* of the standard Wishart distribution $W_p(n, I)$ provides an efficient direct simulation algorithm, as well as useful theory. If we can efficiently simulate the standard Wishart, then the last point above shows how we can use that to create samples from any Wishart distribution. The Bartlett decomposition, and hence construction, is as follows:

- For fixed dimension p and integer $n \geq p$, generate independent normal and chi-square random quantities to define the upper triangular matrix

$$U = \begin{pmatrix} \gamma_1 & z_{1,2} & z_{1,3} & \cdots & z_{1,p} \\ 0 & \gamma_2 & z_{2,3} & \cdots & z_{2,p} \\ 0 & 0 & \gamma_3 & \cdots & z_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \gamma_p \end{pmatrix}$$

  where the entries are independent random draws with diagonal elements $\gamma_i \sim \chi^2_{n-i+1}$ for $i = 1, \ldots, p$, and the upper off-diagonal elements $z_{i,j} \sim N(0, 1)$ for $i = 1, \ldots, p$ and $j = i + 1, \ldots, p$.
- Then (Odell and Fieveson, JASA 1968), the random matrix $\Psi = U'U \sim W_p(n, I)$.
- Hence, if $A = PP'$ for any non-singular $p \times p$ matrix P, we can sample from $\Omega \sim W_p(n, A)$ by generating U and computing $\Omega = (UP')'UP'$.

Some uses of simulation include the ease with which posterior inference on complicated functions of Ω can be derived. For example, inference may be desired for:

- *Correlations:* the correlation between elements i and j of x are $\sigma_{i,j}/\sqrt{\sigma_{i,i}\sigma_{j,j}}$ where the σ terms are the relevant entries in $\Sigma = \Omega^{-1}$.
- *Complete conditional regression coefficients and covariance selection.* Recall that if $x = (x_1, \ldots, x_p)'$ has zero mean normal distribution with precision matrix Ω, then

$$(x_i | x_{1:p\setminus i}, \Omega) \sim N(m_i(x_{1:p\setminus i}), 1/\omega_{i,i})$$

  where

$$m_i(x_{1:p\setminus i}) = \sum_{j=1:p\setminus i} \gamma_{i,j} x_j \quad \text{and} \quad \gamma_{i,j} = -\omega_{i,j}/\omega_{i,i}.$$

This last example shows that the posterior for Ω in a data analysis therefore immediately provides direct inferences, via simulation of the elements of the implied γ terms, for the *partial regression coefficients* in each of the p implied linear regressions. This assumes, of course, a "full model" in the sense that each x_j has, with probability one, a non-zero coefficient in each regression. The study of *covariance selection* and *Gaussian graphical models* focuses on questions of just what variables are relevant as predictors in each of these p conditional distributions.

15.5 Reduced Rank Cases - Singular Wishart Distributions

Sometimes we are directly interested in non-singular (reduced rank, or rank deficient) variance matrices and cases that arise directly from location matrices A of reduced rank. For example, in the normal sampling model suppose that X is rank deficient due to collinearities among the variables, so that S is non-singular. More often, A may be close to singular, then using the modified method below will be numerically stable. The real utility arises in problems in which $p > n$ in that analysis, so that the rank of S is usually n or may be less than n, and certainly lower than p due to dimensionality.

The general framework of possibly reduced rank distributions also includes the regular Wishart as a special case.

- Suppose that A has rank $r \leq p$ with eigendecomposition $A = EBE'$ where E is $p \times r$, $E'E = I$ and $B = \text{diag}(b_1, \ldots, b_r)$ where each $d_i > 0$. This allows A to be rank deficient.
- The generalized inverse of A is $A^- = EB^{-1}E'$.
- Suppose $\Omega = P\Psi P'$ where $P = EB^{1/2}$ and where $\Psi \sim W_r(n, I)$. Then Ω is rank deficient and so singular when $r < p$. In those cases, Ω has the *singular Wishart distribution*.
- The p.d.f. is $p(\Omega) \propto \prod_{i=1}^{r} \delta_i^{(n-r-1)/2} \exp\{-\text{tr}(\Omega A^-)/2\}$ where $(\delta_1, \ldots, \delta_r)$ are the r positive eigenvalues of Ω.
- Simulation is still direct: simulate a regular, non-singular Wishart $\Psi \sim W_r(n, I)$ and transform to the rank deficient Ω.

For the reference analysis of the normal variance/precision model, a singular sample variance matrix (arising, as indicated by example, in cases of $p > n$,) leads to $A = S^-$. With $S = X'X = E(nD)E'$ as earlier explored, this implies $A = EBE'$ as above, where now $B = (nD)^{-1}$.

16 Elements of Graphical Models

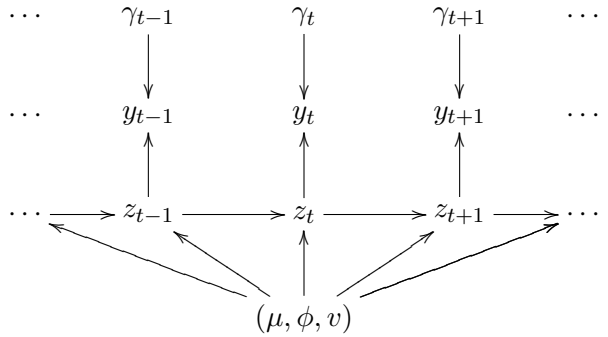
16.1 An Example

The AR(1) SV HMM provides a nice set of examples.

Recall the model in centered form, defined by a set of conditional distributions that imply the full joint density over all states and parameters:

$$p(y_{1:n}, z_{0:n}, \gamma_{1:n}, \mu, \phi, v) = \{\prod_{t=1}^{n} p(y_t|z_t, \gamma_t)p(z_t|z_{t-1}, \mu, \phi, v)p(\gamma_t)\}p(z_0)p(\mu, \phi, v),$$

and, as a detail, $p(\mu, \phi, v) = p(\mu)p(\phi)p(v)$. The strong set of conditional independencies implicit here are encoded in the graph



Variables and parameters are nodes of the graph, and edges imply dependency: a *directed edge*, or arrow, from node a to b is associated with a conditional dependence of b on a in the joint distribution defined by the set of conditional distributions the graph describes. The graph is *directed* since all edges have arrows, and it is *acyclic* since there are no cycles resulting from the defined set of arrows - this latter fact results from the specification of the graph from the set of conditional distributions defined by the full, proper joint distribution over

$$X_n \equiv \{y_{1:n}, z_{0:n}, \gamma_{1:n}, \mu, \phi, v\}.$$

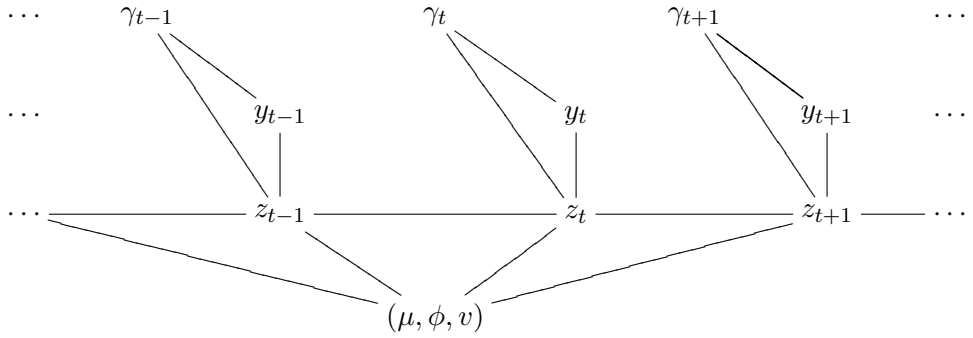
Example. Consider the node for y_t. The existence of arrows from each of z_t and γ_t to y_t, coupled with the lack of arrows to y_t from any other nodes in the graph, imply that the conditional distribution for y_t *conditional on all other variables* depends on, but only on, z_t and γ_t. That is,

$$y_t \perp\!\!\!\perp \{X_n \setminus (z_t, \gamma_t)\} \mid (z_t, \gamma_t).$$

The variables (z_t, γ_t) are *parents* of y_t in the directed graph; y_t is a *child* of each of z_t and γ_t.

Some of the implied dependencies among variables in graphs are defined through common children. For example, γ_t and z_t are parents of y_t, so share a common "bond" that implies association in the overall joint distribution. We have already seen the relevance of this in the Gibbs sampling MCMC analysis: the conditional posterior for the z_t depends on the γ_t, and vice-versa, though the original DAG representation - consistent with the model specification - has no edges between γ_t and z_t. The *undirected graph* that the model implies is shown below, now with the full set of relevant edges related to dependencies in the overall

joint distribution.



Any conditional distribution, and the relevant variables associated with a specific node, can be "read off" this graph. A node has edges linking to its *neighbours*, and that set of neighbours of any node represent the variables on which it depends; conditioning on the neighbours renders the variable at the target node conditionally independent of all other variables.

16.2 General Structure and Terminology for Directed Graphical Models

- A joint distribution for $x = (x_1, \ldots, x_p)'$ has p.d.f. $p(x)$ that may be factorized in, typically, several or many ways. Fixing the order of the variables as above, the usual compositional form of the p.d.f. is

$$p(x) = \prod_{i=1}^{p} p(x_i | x_{(i+1):p}).$$

- In the i^{th} term here, it may be that some of the variables x_j, for $j \in (i+1) : p$, do not in fact play a role. The *parents* of x_i are those variables that do appear and play roles in defining the compositional conditional; that is, knowledge of all x_j for $j \in pa(i) \subseteq \{(i+1) : p\}$ are necessary and sufficient to render x_i conditionally independent of x_k for $k \in \{(i+1) : p\}$ but $k \notin pa(i)$.
- Generally, a joint distribution may factorize as

$$p(x) = \prod_{i=1}^{p} p(x_i | x_{pa(i)})$$

  in a number of ways as the indices are permuted. Each such factorization is consistent with a graphical representation in terms of a *directed, acyclic graph* (DAG) in which the p nodes correspond to the variables x_i, and *directed edges* (arrows) are drawn from node x_j to node x_i if, and only if, $j \in pa(i)$.
- Simulation of $p(x)$ via compositional sampling is easy given any specification in terms of a joint density factored over a DAG.
- Reversing arrows and adding additional directed edges is the root to identifying specific conditional distributions in a DAG. We have already seen in the SV HMM example how to just "read" a DAG representation of a complex, multivariate distribution to identify the variables relevant for a particular conditional of interest. There, for example, the conditional distribution (conditional posterior) for z_t given all other variables depends on the evident parents $\{z_{t-1}, z_{t+1}, (\mu, \phi, v)\}$, but also $\{y_t, \gamma_t\}$. The dependence on y_t is implied since y_t is a child of z_t in the DAG; the additional dependence on γ_t arises since z_t shares parentage of y_t with γ_t. This kind of development is clearly defined through conversion of DAGs to undirected graphs, discussed below. Its relevance in developments such as Gibbs sampling in complicated statistical models is apparent.

16.2.1 Multivariate Normal Example: Exchange Rate Data

Exploratory regression analysis of international exchange rate returns for 12 currencies yields the following set of coupled regressions in "triangular" form. The data are daily returns relative to the US dollar (USD) over a period of about three years to the end of 1996. The returns are centered so that the distribution can be considered zero mean, and, for currencies labeled as below, regressions are specified for currency i conditional on subset of currencies $j > i$:

Index	*Country*	*Currency*	*Parents*	$pa(i)$
1	Canada	CAD (dollar)	-	-
2	New Zealand	NZD (dollar)	AUD	3
3	Australia	AUD (dollar)	GBP, DEM	6,12
4	Japan	JPY (yen)	CHF, DEM	10,12
5	Sweden	SEK (krone)	GBP, FRF	6,9
6	Britain	GBP (pound)	DEM	12
7	Spain	ESP (peseta)	BEF, FRF, DEM	8,9,12
8	Belgium	BEF (franc)	FRF, NLG, DEM	9,11,12
9	France	FRF (franc)	NLG	11
10	Switzerland	CHF (franc)	NLG, DEM	11,12
11	Netherlands	NLG (guilder)	DEM	12
12	Germany	DEM (mark)		

This defines as set of conditional distributions that cohere to give the joint distribution with density defined as the product. The corresponding DAG is below.

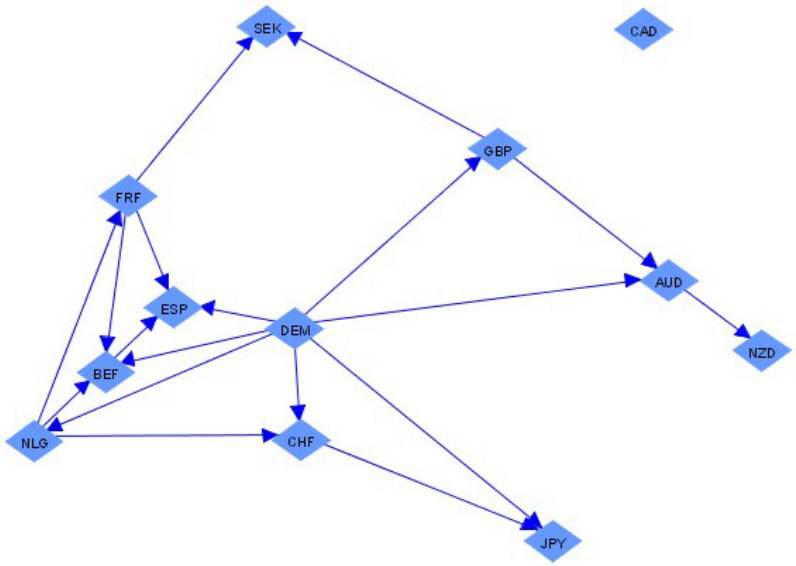


Figure 1: Acyclic directed graph corresponding to the set of conditional regressions defining a joint distributoin for exchange rate returns.

See Matlab code and example on the course web site.

16.3 Joint Distributions and Undirected Graphs

16.3.1 Undirected Graphical Models

Undirected graphs are useful representations of the dependencies in a joint distribution via qualitative display of the full set of complete conditional distributions $p(x_i|x_{-i})$. The set of conditioning variables x_j that in fact play a role in defining $p(x_i|x_{-i})$ are the *neighbours* of x_i, and we write $ne(i)$ for the set of indices of these neighbours. Hence $x_i \perp\!\!\!\perp x_k \mid \{x_j : j \in ne(i)\}$ for all $k \neq i$ such that $k \neq ne(i)$. The complete conditionals are then the set

$$p(x_i|x_{-i}) \equiv p(x_i|x_{ne(i)}), \qquad i = 1, \ldots, p.$$

Some key facts are that:

- Given a specified joint density, it is easy to generate the complete conditionals and neighbor sets simply by inspection, viz.
$$p(x_i|x_{ne(i)}) \propto p(x),$$
and identifying the normalization constant (that depends generally on $x_{ne(i)}$.)
- Neighbourhood membership is symmetric in the sense that $j \in ne(k)$ if, any only if, $k \in ne(j)$.

16.3.2 Graphs from DAGs

In some analyses an initial representation of a joint distribution in terms of a factorization over a DAG is the initial basis for an undirected graphical model. Starting with a DAG, note that:

1. A directed edge from a node x_j to node x_i implies a dependency in the joint distribution.

   This indicates that directed edges - arrows - in a DAG induce undirected edges representing conditional dependencies in the implied graph. That is, if $x_j \in pa(i)$ then $x_j \in ne(i)$.

2. If a pair of nodes x_j and x_k are parents of a node x_i in a DAG then they are, by association through the child node x_i, dependent.

   This indicates that *all* parents of a given node will be associated under the joint distribution - associated through common children if not already directly linked in the DAG. That is, for any j and k such that $\{j, k\} \in pa(i)$ we have $j \in ne(k)$ and $k \in ne(j)$ as well as, of course, $\{j, k\} \in ne(i)$.

These two steps define the process to generate the unique undirected graph from a specified DAG. Note that, generally, an undirected graph can be consistent with more than one DAG, since the DAG relates to one specific form of compositional factorization of the joint density and there may be several or many others. The choice of ordering of variables is key in defining the DAG, whereas the graph represents the full joint distribution without regard to ordering.

Dropping arrow heads is then the first step towards generating a graph from a DAG. The second step - inserting an undirected edge between any nodes that share parentage but are not already connected in the DAG - is referred to as *moralization* of the graph; all parents of any child must be married.

16.3.3 Multivariate Normal Example: Exchange Rate Data

The graph corresponding to the model Figure 1 is below.

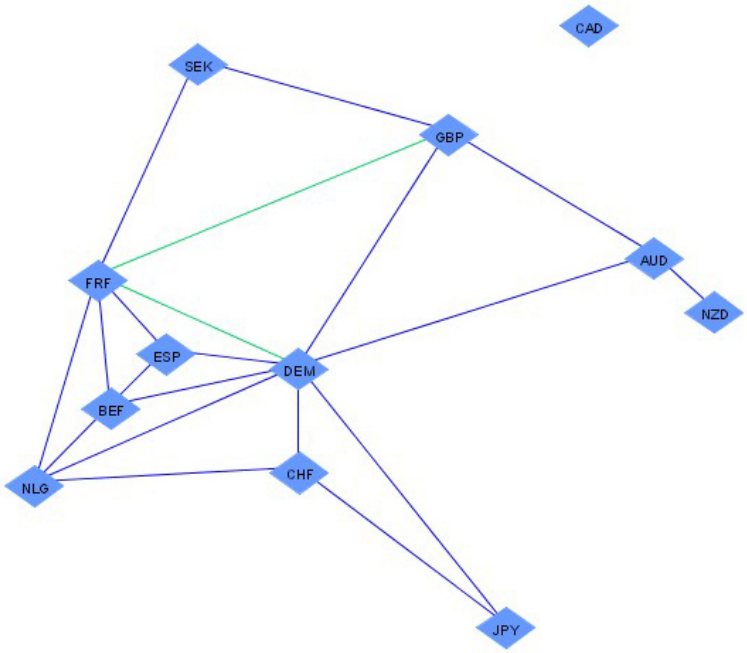


Figure 2: Undirected graph corresponding to the DAG Figure 1 for the fitted distribution of exchange rate returns.

16.4 Factorization of Graphical Models

16.4.1 General Decompositions of Graphs

A multivariate distribution over an undirected graph can usually be factorized in a number of different ways, and breaking down the joint density into components can aid in interpretation and computations with the distribution. Key factorizations relate to the graph theoretic decompositions of arbitrary planar graphs.

Consider an undirected graph $G = (V_G, E_G)$ defined by a set of nodes V_G and a set of edges E_G. If two nodes $v, u \in V_G$ are neighbours, then E_G contains the edge (v, u) and we write $v \sim u$ in G. Hence $E_G = \{(i, j) : i \sim j\}$. The relation $v \sim u$ is of course symmetric and equivalent to each of $j \in ne(i)$ and $i \in ne(j)$. In the context of graphical models for joint distributions of the $p-$vector random variable, $V = \{x_1, \ldots, x_p\}$.

- Consider any subset of nodes $V_A \subseteq V_G$, and write E_A for the corresponding edge set in G. Then (V_A, E_A) defines a subgraph - an undirected graph on nodes in V_A.
- Consider two subgraphs $A = (V_A, E_A)$ and $B = (V_B, E_B)$ of G, and identify the intersection $S = (V_S, E_S)$ where $V_S = V_A \cap V_B$. Then the subgraph S *separates* A and B in G. Simply, there are no nodes $v \in V_A \setminus V_S$ and $u \in V_B \setminus V_S$, such that $v \sim u$. The intersection graph S is called a *separator* of A and B in G.

- Any subgraph A is *complete* if it has all possible edges: $E_A = \{(i,j) : \text{ for all } i,j \in V_A\}$. Every node in a complete graph A is a neighbour of every other such node. The subgraph is *fully connected*.
- A subgraph $S \in G$ is a *clique* of G if it is a *maximally complete subgraph* of G. That is, S is complete and we cannot add a further node that shares an edge with each node of S. Proper subgraphs of S (all subgraphs apart from S itself) are complete but not maximal. (Also, for cliques, we denote the graph by $S \equiv V_S$ since the edge set is, by definition, full and so the notation is redundant.)

Graphs can be decomposed, often in many different ways, into sequences of interconnecting subgraphs separated by complete subgraphs. Such a decomposition is known as a *junction tree* representation of the graph - a tree since it defines an ordered sequence of subgraphs with a tree structure. Based on a specified (usually quite arbitrary) ordering of the nodes, a junction tree decomposition has the form

$$G \;:\longrightarrow\; J_G = [C_1, S_1, C_2, S_2, \cdots, C_{k-1}, S_{k-1}, C_k]$$

where

- the C_i, S_i are proper subgraphs (each with at least 2 nodes);
- each S_i is a complete subgraph of G;
- S_i separates subgraphs C_{i-1} and C_i.

The junction tree is a set of some number k *prime component* subgraphs of G linked by the sequence of $k-1$ separating subgraphs.

Example: A 9 node graph. The figures below provides an example on a $p = 9$ node graph.

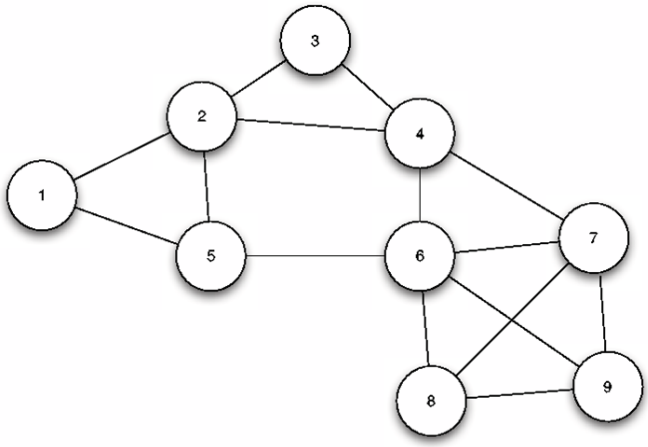


Figure 3: A graph G on $p = 9$ nodes.

The example graph G of Figure 3 can be decomposed sequentially as follows:

- $S_1 = (2,5)$ separates component C_1 with $V_{C_1} = (1,2,5)$ from the rest (Figure 4);
- further separation comes from $S_2 = (4,6)$ (Figure 5);
- then $S_3 = (2,4)$ and $S_4 = (6,7)$, in sequence, provide additional separation.
- This results in a set of $k = 5$ components with these 4 separators (Figure 6) and the corresponding junction tree representation (Figure 7).

16.4.2 Decomposition of distributions over graphs

If the graph G represents the conditional dependency structure in $p(x)$ where $x = (x_1, \ldots, x_p)'$ and the nodes represent the x_i variables, the joint p.d.f. factorizes corresponding to any decomposition of the graph into a junction tree. This is a general and powerful result that is related to the Hammersley-Clifford characterization of joint densities. To avoid unnecessary complications, suppose $p(x) > 0$ everwhere. Then, based on a junction tree representation

$$G \ :\longrightarrow \ J_G = [C_1, S_1, C_2, S_2, \cdots, C_{k-1}, S_{k-1}, C_k]$$

we have the density decomposition

$$p(x) = \frac{\prod_{i=1}^{k} p(x_{C_i})}{\prod_{i=1}^{k-1} p(x_{S_i})}$$

where $x_{C_i} = \{x_j : \ j \in V_{C_i}\}$ is the set of variables in component i, and $x_{S_i} = \{x_j : \ j \in S_i\}$ is the set of variables in separator i.

- That is, the joint density factors as a product of joint densities of variables within each prime component divided by the product of joint densities of variables within each separating complete subgraph. This is quite general.
- One nice, intuitive way to view this decomposition is that the joint density is a product over all component densities, but that implied "double counting" of variables in separators requires a "correction" in terms of "taking out" the contributions from the densities on separators via division.
- That there may be several or many such decompositions simply represents alternative factorizations of the density.

Example: The 9 node graph. In the example note that four of the five prime components are themselves complete, whereas one is not a complete subgraph (an "incomplete" prime component). A joint density $p(x)$ in the graph in Figure 3 then has the representation

$$p(x) = \frac{p(x_1, x_2, x_5)p(x_2, x_3, x_4)p(x_2, x_4, x_5, x_6)p(x_4, x_6, x_7)p(x_6, x_7, x_8, x_9)}{p(x_2, x_5)p(x_2, x_4)p(x_4, x_6)p(x_6, x_7)}.$$

Some insight is gained by noting that this can be written as

$$\begin{aligned} p(x) &= \frac{p(x_1, x_2, x_5)}{p(x_2, x_5)} \frac{p(x_2, x_3, x_4)}{p(x_2, x_4)} \frac{p(x_2, x_4, x_5, x_6)}{p(x_4, x_6)} \frac{p(x_4, x_6, x_7)}{p(x_6, x_7)} p(x_6, x_7, x_8, x_9) \\ &= p(x_1|x_2, x_5)p(x_3|x_2, x_4)p(x_2, x_5|x_4, x_6)p(x_4|x_6, x_7)p(x_6, x_7, x_8, x_9), \end{aligned}$$

i.e., one specific compositional form corresponding to a DAG on the implied elements. This shows how the general component-separator decomposition can naturally yield representations of joint densities that are useful for simulation of the joint distribution as well as for investigating dependency structure. It is an example of how we can construct DAGs from graphs.

16.5 Additional Comments

Marginalization induces complexity in graphical models. If $G = (V, E)$ is the graph for $p(x_{1:p})$, and $A \subset x_{1:p}$, then the graph representing the marginal density $p(x_A)$ will generally have more edges than the subgraph (A, E_A) of G. The variables removed by marginalization generally induce edges representing the now "hidden" dependencies.

A simple example is the "star" graph defined by the joint density $p(x_{1:p}) = p(x_1)\prod_{i=2}^{p} p(x_i|x_1)$. This graph encodes the dependencies $x_i \perp\!\!\!\perp x_j|x_1$ for all $i, j > 1$. Here x_1 is a key variable inducing dependencies among all the others. On marginalisation over x_1 the result is a complete graph on $x_{2:p}$, a much less "sparse" graph in this $p-1$ dimensional margin than we see for the subgraph in the full p dimensions. Node x_1 separates all other nodes in the full graph.

16.6 Special Cases of Decomposable Graphs

Decomposable graphs are special but key examples.

- A *decomposable graph* G is a graph in which there are no edge cycles of path length four or more - the graph is *triangulated.* The 9 node example above is non-decomposable since it has a four cycle; modifying that graph to add one edge - between nodes 4 and 5 or between nodes 2 and 6 - leads to a triangulated and hence decomposable graph.
- One of the relevant features of decomposable graphs is that all prime components are complete - that is, any junction tree representation defines a graph decomposition as a sequence of intersecting complete subgraphs.

16.7 Gaussian Graphical Models

In the Gaussian case, lack of an edge between nodes i and j in a graph implies a conditional independence consistent with the implied zero in the precision matrix of the multivariate normal distribution. Hence

- Decomposition into a junction tree implies that, within each separator S_i, all variables are conditionally dependent since the separator is a complete subgraph.
- If the graph is decomposable, we can view each component subgraph in the junction tree (all prime components and separators) as defining a set of subsets of jointly normal variables, within each of which there is a full set of dependencies.

This latter feature leads to a statistical theory of inference on Gaussian graphical models utilizing what are termed *hyper-Wishart distributions* - or *hyper-Markov Wishart distributions* - for inference. For a given graph G, these models imply that the prior and posterior distributions for the full precision matrix of $p(x)$ - whatever its sparsity structure may be as a result of the missing edges in G - are within a relatively tractable family of hyper-Wishart distributions. The framework extends to non-decomposable graphs, with subsequent computational challenges in understanding these distributions when p is more than small integer value. However, the major, major open questions - questions at the frontiers of research in statistics and computational science and the interfaces with machine learning - are those of exploring spaces of graphs themselves to make inferences on the dependence structure itself.

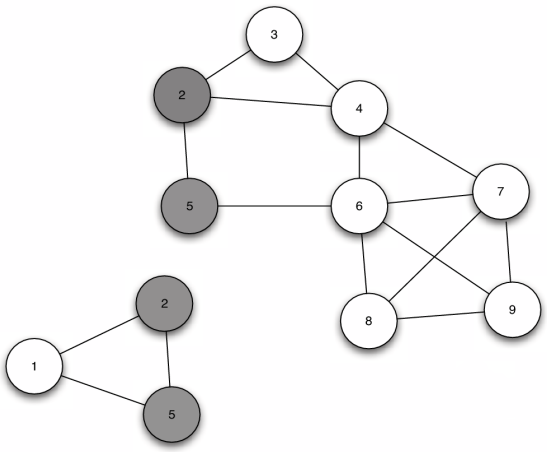


Figure 4: $S_1 = (2, 5)$ separates component C_1 from the graph.

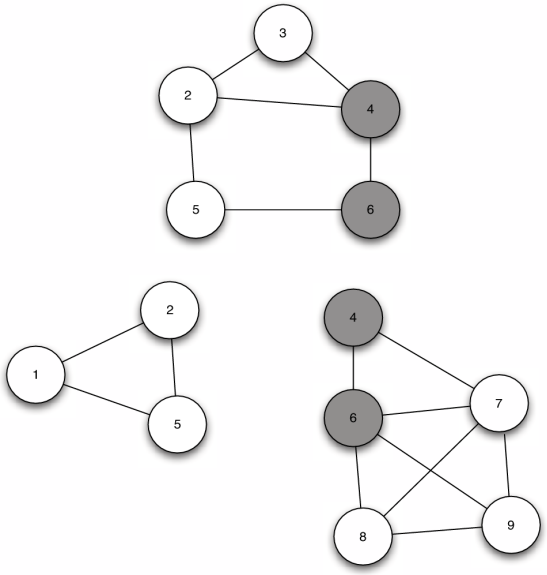


Figure 5: $S_2 = (4, 6)$ defines further separation.

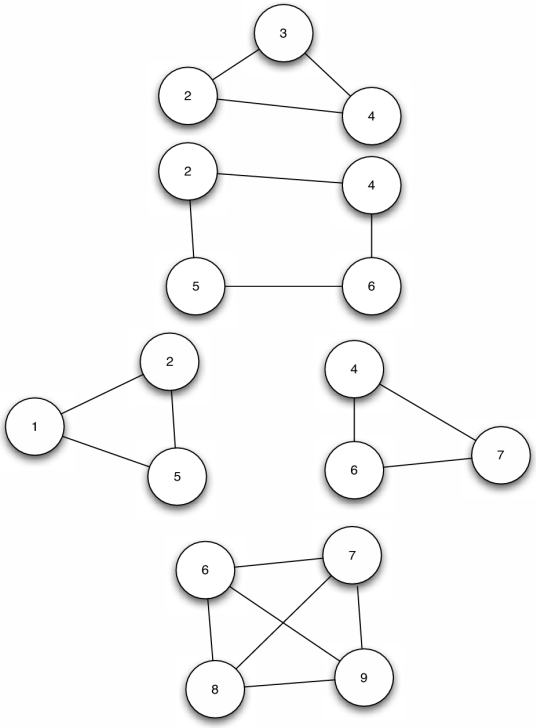


Figure 6: The graph G is separated into $k = 5$ components via the 4 separators.

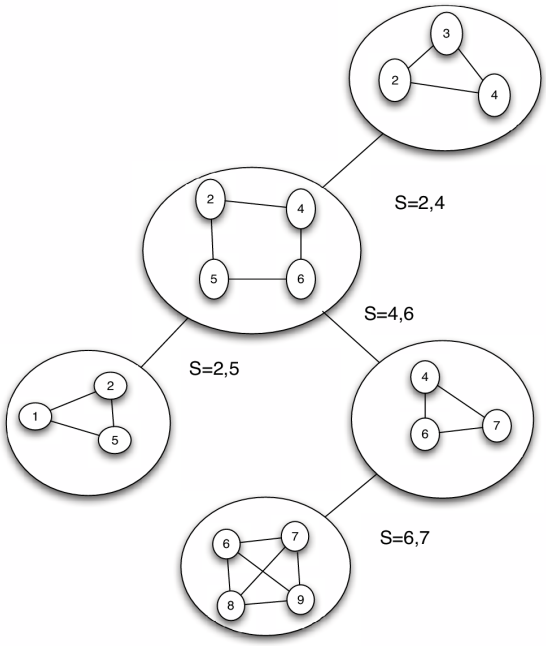


Figure 7: The resulting junction tree representation of G.