CHAPTER 18

GIBBS SAMPLING

Gibbs sampling is a simulative Markov chain approach to sample from a multidimensional probability distribution. Below we explain the concept, give an example from physics, and then give a financial application.

§1. The Concept

Suppose we wish to sample pairs (X, Y) drawn from the two dimensional joint mass function $p_{(X,Y)}(x, y)$. Let $p_Y(y)$ denote Y's marginal mass function and $p_{(X \mid Y=y)}(x)$ denote X's conditional mass function given $Y = y$, so $p_{(X,Y)}(x, y) = p_{(X \mid Y=y)}(x)\, p_Y(y)$.

Suppose $(X_0, Y_0) \sim p_{(X,Y)}$ and, if $Y_0 = y$, choose $X_1 \sim p_{(X \mid Y=y)}$. Then

$$\begin{aligned} P[X_1 = x, Y_0 = y] &= P[X_1 = x \mid Y_0 = y]\, P[Y_0 = y] \\ &= p_{(X \mid Y=y)}(x)\, p_Y(y) \;=\; p_{(X,Y)}(x, y). \end{aligned}$$

That is, $(X_0, Y_0) \sim p_{(X,Y)} \implies (X_1, Y_0) \sim p_{(X,Y)}$, where X_1 is drawn from the conditional distribution of X given the realized value of Y_0. Now, if $X_1 = x$, choose $Y_1 \sim p_{(Y \mid X=x)}$. The above reasoning will also show that $(X_1, Y_1) \sim p_{(X,Y)}$. These dynamics for $(X_0, Y_0) \to (X_1, Y_1)$ (first choose X_1 according to X's conditional distribution given Y_0, then draw Y_1 according to Y's conditional distribution given X_1) may be repeated indefinitely producing the sequence

$$(X_0, Y_0),\ (X_1, Y_1),\ (X_2, Y_2),\ \ldots$$

Since the dynamics are the same at each step and the distribution of (X_{n+1}, Y_{n+1}) depends only on the realized value of (X_n, Y_n) we have a homogeneous Markov chain on the state space

$$S \;=\; \{(x, y) : x \in \text{Range}\,(X), y \in \text{Range}\,(Y)\}.$$

We have argued above that $p_{(X,Y)}$, which is a probability distribution on S, is invariant under the dynamics of this MC. If the MC is ergodic (which it will be in our applications) it will follow that no matter how (X_0, Y_0) is selected the distribution of (X_n, Y_n) will converge to $p_{(X,Y)}$ as n gets large. This is the essence of Gibbs sampling.

We have illustrated this for a two dimensional distribution, but the approach works in higher dimensions as well. Our first example illustrates this for a 40,401 dimensional distribution!

§2. First Application: Ising Models

Ising models (pronounced "ezing") are mathematical models of ferromagnetism. In ferromagnetic material electrons possess what physicists call a *spin*, "up" (+1) or "down" (−1). In magnetized material the spins of electrons are highly aligned. The question is how does this happen spontaneously in some materials? In the words of Wikipedia on Ising models:

> "Once the electron's spin was discovered, it was clear that magnetism should be due to a large number of electrons spinning in the same direction. It was natural to ask how the electrons all know which direction to spin, because the electrons on one side of a magnet don't directly interact with the electrons on the other side. They can only influence their neighbors. The Ising model was designed to investigate whether a large fraction of the electrons could be made to spin in the same direction using **only local forces**." (Bold added.)

By the way, we see the same phenomenon in schools of fish. Fish are not like tourists, following a leader with a flag. Each fish responds to the actions of only nearby fish, yet the school exhibits extreme coherence of motion. How??

In a two dimensional Ising model, the action takes place on a lattice — infinite in the mathematical model but necessarily finite for simulation purposes. Our lattice will be the 201 × 201 points with the integer coordinates

$$L \;=\; \{(i,j) : -100 \le i \le 100,\ -100 \le j \le 100\}.$$

A *configuration* x will be an allocation of the spins, +1 or −1, to the 40,401 lattice points and x_{ij} will denote the spin at the lattice site (i,j). Let S denote all such configurations — there are $2^{40,401}$ of them. Formally

$$S \;=\; \big\{\text{all functions } x \text{ from } L \text{ to } \{-1,+1\}\big\}.$$

Each (i,j) on the interior of the lattice has four *neighbors*: north $(i,j+1)$; south $(i,j-1)$; east $(i+1,j)$; and west $(i-1,j)$. For points on the boundary, we use the convention $100+1=-100$ and $-100-1=100$ so, for example, the southeast corner point $(100,-100)$ has the four neighbors: north $(100,-99)$; south $(100,100)$; east $(-100,-100)$; and west $(99,-100)$. Roughly speaking this is equivalent to pasting the southern boundary to the northern boundary to form a cylinder and then pasting the western boundary (now a circle) to the eastern boundary (also a circle). Essentially the action is now taking place on the surface of an inner tube! The point of this is that, with this convention, every lattice site (including those on the boundary) has four neighbors. We will write $(i,j)\leftrightarrow(i',j')$ if (i',j') is a neighbor of (i,j), and we note that this is a symmetric relation: $(i,j)\leftrightarrow(i',j') \iff (i',j')\leftrightarrow(i,j)$.

We now put a probability measure on the $2^{40,401}$ elements of S. For configuration $x \in S$, let $D(x)$ denote the number of pairs of neighbors in L with spins in the configuration x that disagree. Formally

$$D(x) = \\ \#\big\{\text{pairs } (i,j) \text{ and } (i',j') \in L \ \text{ with } (i,j)\leftrightarrow(i',j') \text{ and } x_{ij} \neq x_{i'j'}\big\},$$

where the # means "number of elements of". [Here we count a pair $(i,j)\leftrightarrow(i',j')$ and $(i',j')\leftrightarrow(i,j)$ only once — not twice.] For some $\beta \geq 0$ (called the inverse temperature) put $g(x) = e^{-\beta D(x)}$ and then $Z = \sum_{x \in S} g(x)$. Finally we define $P(x) = \frac{g(x)}{Z}$. We see that $P(x) > 0$ for each configuration x, and $\sum_{x \in S} P(x) = 1$ due to the normalizing constant Z. Note that large values of β discourage neighboring disagreements (by making $g(x)$ small if $D(x)$ is large) and therefore encourage the alignment of neighboring spins. At the other extreme, if $\beta = 0$ then each $g(x) = 1$ and the distribution P is uniform on all configurations x. At this extreme the alignment of spins is completely irrelevant.

We seek to sample configurations $x \sim P$. We do this via Gibbs sampling. We start with a random configuration x where each x_{ij} is independently ± 1 with equal likelihood. We simulate 50 thousand steps of the Markov chain, where one step of our MC involves a complete raster scan of the lattice sites. This is implemented below via pseudo code:

```
for (n = 1; n <= 50000; n++) {
  for (i = -100; i <= 100; i++) {
    for (j = -100; j <= 100; j++) {
      (∗) choose new $x_{ij}$ according to its conditional distribution
            given the other spins in the $x$ configuration;
    }
  }
}.
```

To implement $(*)$, let A denote a specification of all the spins in L *except* for the spin at site (i,j). We seek to compute $P[\{x_{ij} = +1\} \mid A]$ and then choose x_{ij}'s spin randomly according to this conditional probability. Let D_A denote the number of pairs of neighbors in L *other than those four pairs involving site* (i,j) with spins in the A configuration that disagree. Let b denote the number of site (i,j)'s neighbors that have spin $+1$ — so $0 \leq b \leq 4$ and $4-b$ is the number of (i,j)'s neighbors that have spin -1. If we choose $x_{ij} = +1$ (respectively -1) the total number of disagreements in L will be $D_A + 4 - b$ (respectively $D_A + b$).

Hence

$$
\begin{aligned}
&P[\{x_{ij} = +\,1\} \mid A] \\
&\qquad = P[A \cap \{x_{ij} = +1\}] \,/\, PA \\
&\qquad = P[A \cap \{x_{ij} = +1\}] \,/\, \big(P[A \cap \{x_{ij} = +1\}] + P[A \cap \{x_{ij} = -1\}]\big) \\
&\qquad = \frac{e^{-\beta(D_A+4-b)}}{Z} \bigg/ \left(\frac{e^{-\beta(D_A+4-b)}}{Z} + \frac{e^{-\beta(D_A+b)}}{Z}\right) \\
&\qquad = \frac{1}{1 + e^{-\beta(2b-4)}}.
\end{aligned}
$$

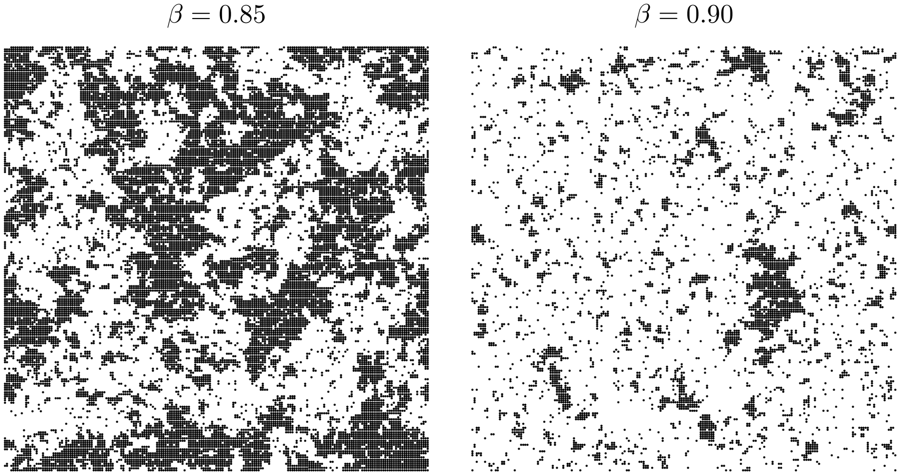
We note that $P[\{x_{ij} = +1\} \mid A]$ does not depend on the entire configuration A but rather only on the spins at (i,j)'s four neighbors, they determine b and are the "local forces" referred to in the Wikipedia quote. This is convenient and makes $P[\{x_{ij} = +1\} \mid A]$ easy to compute. The table below shows $P[\{x_{ij} = +1\} \mid A]$ for $\beta = 0$, 0.5, and 1.0 and the five possible values of b. When $\beta = 0$ the choice of spin for (i,j) does not depend on A and is ± 1 with equal likelihood. As β increases we see that these probabilities increasingly encourage (i,j)'s spin to agree with the majority of its neighbors' spins.

Figure 1. $P[\{x_{ij} = +1\} \mid A]$ for various values of b and β.

	$b = 0$	$b = 1$	$b = 2$	$b = 3$	$b = 4$
$\beta = 0$	0.500	0.500	0.500	0.500	0.500
$\beta = .5$	0.119	0.269	0.500	0.731	0.881
$\beta = 1$	0.018	0.119	0.500	0.881	0.982

Figure 2 illustrates how the distribution P responds to the inverse temperature β. The figure shows the $n = 50{,}000$ configuration for two different βs. On the left $\beta = 0.85$ and there remain competing large clusters of $+1$ spins and -1 spins ($+1$ spins are shown with a darkened pixel, -1 spins with a white pixel). On the right, where $\beta = 0.90$, one particular spin has largely taken over the lattice. (In the fish analogy, the fish are mostly swimming in the same direction!) These two images are visually typical for these two βs.

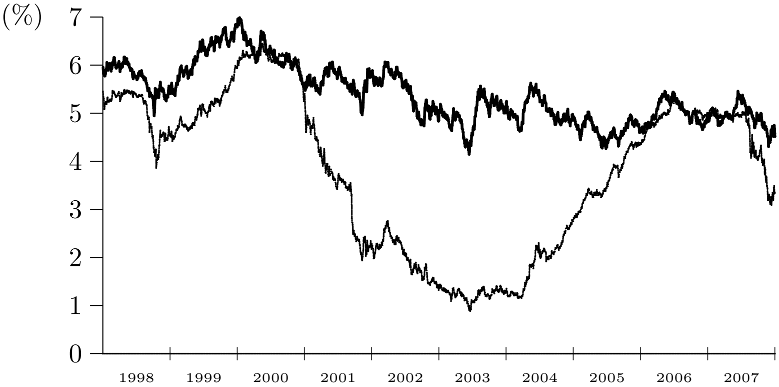
Figure 2. Representative configurations for two different betas.



§3. A Financial Application

In this example we seek to model the behavior of a short term Treasury yield as a function of a long term Treasury yield. Figure 3 below shows daily data for the 1 year (thin line) and 20 year (bold line) Treasury yields over the 10 year period from January 1998 to December 2007 — there are 2,501 items of data.

Figure 3. 1-year and 20-year (bold) Treasury yields.



Let $(x_t : 1 \le t \le 2{,}501)$ denote the time series for the 20 year yield and (y_t) the corresponding time series for the 1 year yield. We take the 20 year (x_t) data as given (i.e., non stochastic) and view the 1 year (y_t) as a realization of a stochastic process (Y_t) governed by the dynamics:

$$\begin{aligned} Y_t &= \beta x_t + Z_t \\ Z_t &= \alpha Z_{t-1} + \epsilon_t, \quad \text{and} \\ \epsilon_t &\sim \text{independent Normal}\,(0, \sigma^2)\text{s}. \end{aligned}$$

Here β, α, and σ^2 are three parameters that we seek to estimate using the data. This model basically says that the short rate Y is some fraction β of the long rate x. Since the yield curve is generally positively sloped (look at the data), we expect $0 < \beta < 1$. We do not expect this relationship to be precise, so we overlay the spread Z. Our data is daily, so we don't expect this spread to change rapidly. Hence the autocorrelated model, where we expect α to be close to 1 and σ to be fairly small (since the data is daily).

Our strategy is to select a prior distribution f for $(\beta, \alpha, \sigma^2)$ and use Gibbs sampling to sample from f_D, the posterior distribution given the data $D = \{Y_1 = y_1, \ldots, Y_{2501} = y_{2501}\}$. First we choose initial values β_0, α_0, and σ_0^2. It will not matter how we do this because our Markov chain will converge to the invariant distribution f_D regardless of the initial state. We choose to draw the initial values from the prior distribution, as specified below. Then, for $1 \leq n \leq 11{,}000$ we exercise the following steps: (1) given α_{n-1} and σ_{n-1}^2, draw a new $\beta_n \sim f_D(\beta \,|\, \alpha_{n-1}, \sigma_{n-1}^2)$; (2) given σ_{n-1}^2 and the new β_n, draw a new $\alpha_n \sim f_D(\alpha \,|\, \beta_n, \sigma_{n-1}^2)$; and (3) given the new β_n and α_n, draw a new $\sigma_n^2 \sim f_D(\sigma^2 \,|\, \beta_n, \alpha_n)$. After n is large, each new triple $(\beta_n, \alpha_n, \sigma_n^2)$ will be drawn from the joint $f_D(\beta, \alpha, \sigma^2)$ distribution. We ignore the first 1,000 triples to allow convergence to f_D and keep the remaining 10,000 observations.

Sampling repeatedly from f_D in this manner will generate conditional marginal density functions $f_D(\beta)$, $f_D(\alpha)$, and $f_D(\sigma^2)$, from which MAP estimates for those parameters may be calculated. See [39], page 624, for more on this approach.

The Prior Distribution $f(\beta, \alpha, \sigma^2)$. We will take both β and α to be normally distributed with mean and variance 1, and σ^2 to have a scaled inverted χ^2 distribution, specifically $\sigma^2 \sim \lambda \cdot \text{Inv-}\chi_m^2$, where $\lambda = 0.0025$ and $m = 6$. These choices are rather arbitrary, but fortunately they do not influence the results greatly due to the abundance of data. They were selected to give ample probability mass to the region where we expect their respective values to lie, as discussed above. With these marginal distributions, we take these three parameters to be independent.

Step 1. Choosing β, given α and σ^2. Suppose we know the values of α and σ^2. From our model we see that, for $2 \leq t \leq 2{,}501$ we have,

$$\begin{aligned} Y_t - \beta x_t &= Z_t, \quad \text{and} \\ Y_{t-1} - \beta x_{t-1} &= Z_{t-1}, \quad \text{so} \\ Y_t - \alpha Y_{t-1} - \beta(x_t - \alpha x_{t-1}) &= Z_t - \alpha Z_{t-1} \;=\; \epsilon_t, \end{aligned}$$

where *(third line)* above is *(first line)* $-\alpha\cdot$ *(second line)*. Putting $\widetilde{Y}_t = Y_t - \alpha Y_{t-1}$ and $\widetilde{x}_t = x_t - \alpha x_{t-1}$ we get that

$$\widetilde{Y}_t \;=\; \beta\widetilde{x}_t + \epsilon_t, \quad 2 \le t \le 2{,}501, \text{ where } \epsilon_t \sim \text{Normal}\,(0,\sigma^2).$$

Since we know α, we may calculate the $\widetilde{x}_t$ and the realized values of $\widetilde{Y}_t$, namely $\tilde{y} = y_t - \alpha y_{t-1}$. We can then use standard OLS regression (refer to §2, Chapter 17) to estimate β:

$$\widehat{\beta} \;=\; \frac{\sum_t \widetilde{x}_t\widetilde{y}_t}{\sum_t \widetilde{x}_t^2} \;\sim\; \text{Normal}\left(\beta, \sigma^2_{\widehat{\beta}}\right),$$

where, since we know σ^2, we may calculate $\sigma^2_{\widehat{\beta}} = \sigma^2 \Big/ \sum_t \widetilde{x}_t^2$. To summarize, we have $\beta \sim$ the prior Normal$\,(1,1^2)$ and then $\widehat{\beta} \sim \text{Normal}\,(\beta,\sigma^2_{\widehat{\beta}})$. Reflecting the data $\widehat{\beta}$, we conclude that the posterior distribution of β is

$$\beta \;\sim\; \text{Normal}\left(\frac{\sigma^2_{\widehat{\beta}} + \widehat{\beta}}{\sigma^2_{\widehat{\beta}} + 1}, \frac{\sigma^2_{\widehat{\beta}}}{\sigma^2_{\widehat{\beta}} + 1}\right),$$

using equation (4) from Chapter 15 (with $1 \to \mu_0$, $1 \to \sigma_0^2$, and $\sigma^2_{\widehat{\beta}} \to \sigma^2$).

Remark. We have *assumed* here without justification that the summary statistic $\widehat{\beta}$ is *sufficient* in the sense that $f_D(\beta\,|\,\alpha,\sigma^2) = f_{\widehat{\beta}}(\beta\,|\,\alpha,\sigma^2)$. Such convenient assumptions are common when working with data. Indeed, the convenient assumption of normality is frequently made without justification.

Following the program of Gibbs, then, we draw a number β from this posterior conditional distribution for β given α and σ^2.

Step 2. Choosing α, given β and σ^2. Proceeding with Gibbs, our next step is to draw $\alpha \sim f_D(\alpha\,|\,\beta,\sigma^2)$. Our model says that $Z_t = Y_t - \beta x_t$. Since we know β at this step we may calculate the realized values of Z_t: $z_t = y_t - \beta x_t$. At time t, our model says that

$$Z_t \;=\; \alpha z_{t-1} + \epsilon_t, \quad 2 \le t \le 2{,}501,$$

and we may use the realized values z_t to estimate α via OLS regression:

$$\widehat{\alpha} \;=\; \frac{\sum_t z_t z_{t-1}}{\sum_t z_{t-1}^2} \;\sim\; \text{Normal}\,(\alpha,\sigma^2_{\widehat{\alpha}}).$$

Importantly, here again since we know σ^2 we may calculate $\sigma^2_{\widehat{\alpha}} = \sigma^2 \Big/ \sum_t z_{t-1}^2$. Summarizing, $\alpha \sim$ the prior Normal$\,(1,1^2)$ and then $\widehat{\alpha} \sim \text{Normal}\,(\alpha,\sigma^2_{\widehat{\alpha}})$. Reflecting the data $\widehat{\alpha}$, we conclude that the posterior distribution of α is

$$\alpha \;\sim\; \text{Normal}\left(\frac{\sigma^2_{\hat{\alpha}} + \widehat{\alpha}}{\sigma^2_{\hat{\alpha}} + 1}, \frac{\sigma^2_{\hat{\alpha}}}{\sigma^2_{\hat{\alpha}} + 1}\right).$$

Here again we are assuming that the summary statistic $\widehat{\alpha}$ is sufficient, i.e., that $f_D(\alpha \mid \beta, \sigma^2) = f_{\hat{\alpha}}(\alpha \mid \beta, \sigma^2)$.

We now draw a number α from this posterior conditional distribution given β and σ^2.

Step 3. Choosing σ^2, given β and α. Here we draw $\sigma^2 \sim f_D(\sigma^2 \mid \beta, \alpha)$. Our model says that (ϵ_t) are iid Normal$\,(0, \sigma^2)$. Knowing β, we can compute the realized Z_t sequence: $z_t = y_t - \beta x_t$. Now, knowing α, we can compute the realized ϵ_t sequence for $2 \le t \le 2{,}501$: call it $e_t = z_t - \alpha z_{t-1}$ and form the statistic

$$r^2 \;=\; \frac{\sum_{t=2}^{2{,}501} e_t^2}{2{,}500}.$$

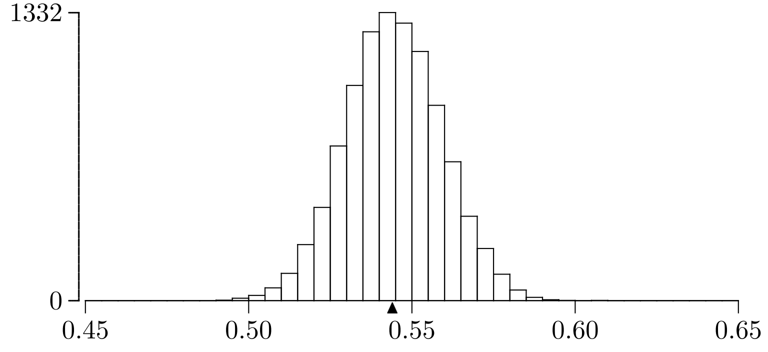
We have the following: $\sigma^2 \sim 0.0025 \cdot \text{Inv-}\chi^2_6$ (the prior distribution); then we observe the data $(\epsilon_t : 2 \le t \le 2{,}501)$ which are iid Normal$\,(0, \sigma^2)$ with realized values (e_t). It follows from Example 2 of Chapter 15 that the posterior distribution of σ^2 is

$$\sigma^2 \sim \lambda \cdot \text{Inv-}\chi^2_m, \text{ where } m = 2{,}506 \text{ and } \lambda = \frac{6 \cdot 0.0025 + 2{,}500 \cdot r^2}{2{,}506}.$$

Yet again we have assumed sufficiency, with the hope that $f_D(\sigma^2 \mid \beta, \alpha) = f_{r^2}(\sigma^2 \mid \beta, \alpha)$. We then draw $\sigma^2 \sim \lambda \cdot \text{Inv-}\chi^2_m$ and go back to step 1.

The Results. In `TreasuryModel.cpp` we execute steps (1), (2), and (3) above 11,000 times. Ignoring the first 1,000, we have 10,000 observations of $(\beta, \alpha, \sigma^2)$ drawn from the desired f_D distribution. Figure 4 below shows a histogram for the resulting β statistic. The mean value for β over the 10,000 samples is $\widehat{\beta} = 0.544$, as highlighted by the '▲' in Figure 4. As this lies in the histogram bar of maximal height (with 1,332 observations) it seems reasonable to take this $\widehat{\beta}$ as the MAP estimate.

Figure 4: Histogram showing the conditional marginal distribution $f_D(\beta)$.



Similar results for α and σ^2 are $\widehat{\alpha} = 0.999$ and $\widehat{\sigma}^2 = 0.038^2$. We note that the Z_t process is highly autocorrelated, as expected. Figure 5 shows the realization of Z_t using these estimated parameter values.

Figure 5: The realized Z_t time series.

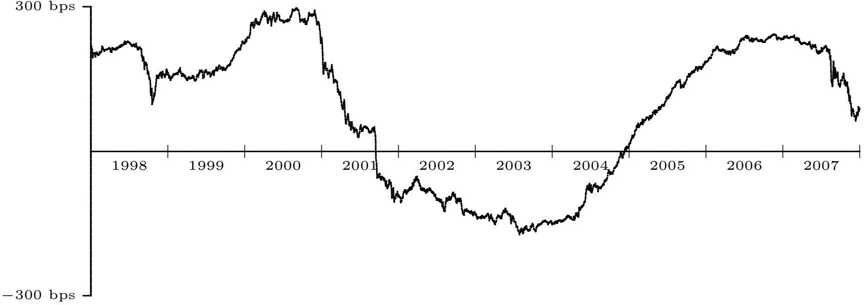
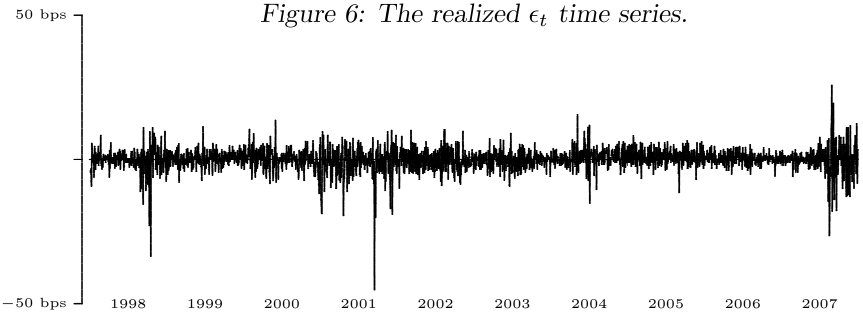


Figure 6 shows the realization of the ϵ_t process. This is supposed to be white noise, i.e., iid Normal $(0, \sigma^2)$ data. Clearly this is not the case. We see obvious spikes in volatility near the August 1998 embassy bombings in Tanzania and Kenya, 9/11/2001, and also in 2007 leading up to the financial crisis.

Figure 6: The realized ϵ_t time series.



Accompanying Code

The program `TreasuryModel.cpp` implements the model described above, generating output files for viewing with TeX. A histogram of the ϵ_ts may be viewed with `Histogram.tex`. Subsequently a histogram of the β statistic may be viewed with the same TeX program. Time series for the Z_t and ϵ_t statistics may be viewed with `Show_Z_t.tex` and `Show_e_t.tex`, respectively.