CHAPTER
TWENTY-EIGHT

Introduction to Mortgage-Backed Securities

Anyone stupid enough to promise to be responsible for a stranger's debts deserves to have his own property held to guarantee payment.
—Proverbs 27:13

A **mortgage-backed security** (**MBS**) is a bond backed by an undivided interest in a pool of mortgages. MBSs traditionally enjoy high returns, wide ranges of products, high credit quality, and liquidity [432]. The mortgage market has witnessed tremendous innovations in product design [54]. The complexity of the products and the prepayment option mandate the deployment of advanced models and software techniques. In fact, the mortgage market probably could not have operated efficiently without them [659]. Although our focus will be mainly on residential mortgages, the underlying principles are applicable to other types of assets as well.

28.1 Introduction

A mortgage is a loan secured by the collateral of real estate property. The lender – the **mortgagee** – can foreclose the loan by seizing the property if the borrower – the **mortgagor** – defaults, that is, fails to make the contractual payments. An MBS is issued with pools of mortgage loans as the collateral. The cash flows of the mortgages making up the pool naturally reflect upon those of the MBS. There are three basic types of MBSs: **mortgage pass-through security** (**MPTS**), **collateralized mortgage obligation** (**CMO**), and **stripped mortgage-backed security** (**SMBS**).

The mortgage sector is by far the largest in the debt market (see Fig. 28.1). The mortgage market conceptually is divided between a primary market, also called the **origination market**, and a secondary market in which mortgages trade. The secondary market includes the market for loans that are not securitized, called **whole loans**, and the market for MBSs.

Individual mortgages are unattractive for many investors. To start with, often at hundreds of thousands of U.S. dollars or more, they demand too much investment. Most investors also lack the resources and knowledge to assess the credit risk involved. Furthermore, a traditional mortgage is fixed rate, level payment, and fully amortized with the percentage of **principal and interest** (**P&I**) varying from month to month, creating accounting headaches. Finally, prepayment levels fluctuate with a host of factors, making the size and the timing of the cash flows unpredictable.

Mortgage debt outstanding (U.S.$ millions)

	1994	*1995*	*1996*	*1997*	*1998*	*1999*
Total outstanding	4,392,794	4,603,981	4,877,536	5,211,286	5,736,638	6,387,651
By holder:						
Commercial banks	1,012,711	1,090,189	1,145,389	1,245,315	1,337,217	1,495,717
Savings institutions	596,191	596,763	628,335	631,826	643,957	668,634
Life insurance cos	210,904	213,137	208,162	206,840	213,640	229,333
Federal/agency	315,580	308,757	295,192	286,167	292,636	320,105
Mortgage pools/trusts	1,730,004	1,863,210	2,040,848	2,239,350	2,589,764	2,954,836
Individuals/others	527,404	531,926	559,609	601,788	659,425	719,026

Figure 28.1: Mortgage debt outstanding 1994–1999. Source: Federal Reserve Bulletin.

A liquid market for individual mortgages did not appear until the mortgage institutions started securitizing their mortgage holdings in 1970. Individual, illiquid mortgages were then turned into marketable securities that were easier to analyze and trade. Today, financial intermediaries buy mortgages and place them in a pool. Interests in the pools are then sold to investors. These undivided ownership interests in the loans that collateralize the security are called **participation certificates** (**PC**s). The intermediary receives the mortgage payments from homeowners or servicing organizations and passes them to investors. The intermediary also guarantees that it will pay investors all the P&I that are due in case of default. Several of the above-mentioned problems are solved or alleviated by this arrangement. For instance, the minimum investment is reduced. The credit risk of the homeowners is virtually eliminated because of the intermediary's guarantees. As a result, the credit strength of the PC as seen by the investor is shifted from the homeowner and the property to the intermediary.

28.2 Mortgage Banking

The original lender is called the **mortgage originator**. It can be thrifts, commercial banks, mortgage bankers, life insurance companies, or pension funds. There are three revenues for the mortgage originator with regard to a new mortgage. It can hold the mortgage for investment or sell the mortgage to an investor or conduit. **Conduits** are either federally sponsored credit agencies or private companies that pool mortgages. Finally, it can use the mortgage as collateral for the issuance of a security. In this way, the mortgage becomes part of a pool of mortgages that are the collateral for a security – it is securitized.

Mortgage insurance is often required for guarding against default. Besides private mortgage insurers, three U.S. government agencies guarantee mortgages for qualified borrowers: the Federal Housing Administration (FHA), the Department of Veterans Affairs (VA), and the Rural Housing Service (RHS) [327]. Loans not guaranteed or insured by the FHA, VA, or RHS are called **conventional loans**. On the other hand, loans that comply with the underwriting standards for sale or conversion to MBSs issued and guaranteed by two federally sponsored credit agencies are called **conforming mortgages**. The two agencies are the Federal National Mortgage Association (FNMA or "Fannie Mae") and the Federal Home Loan Mortgage

Loan Information:	
Balances:	
Principal Balance on 10/03/97	$155,520.31
Escrow Balance on 10/03/97	$3,015.82
Payment Factors:	
Interest Rate	7.12500%
Principal & Interest	$1,702.96
Escrow Payment	$700.32
Total Payment:	$2,403.28
Year-to-Date:	
Interest	$8,514.63
Taxes	$5,665.60
Principal	$6,817.07

Figure 28.2: Typical monthly mortgage statement.

Corporation (FHLMC or "Freddie Mac"). Both are now public companies. Mortgage bankers also originate FHA-insured and VA-guaranteed mortgage loans for sale in the form of Ginnie Mae pass-throughs. Ginnie Mae stands for Government National Mortgage Association (GNMA). MBSs issued by Fannie Mae or Freddie Mac are primarily sold by mortgage banking firms directly to securities dealers. FHA/VA/RHS mortgage loans are packaged for sale as pass-through securities guaranteed by Ginnie Mae and sold also primarily to securities dealers. Conventional loans exceeding the maximum amounts required for conformance are called **jumbo loans**.

A mortgage needs to be serviced. Principal, interest, and escrow funds for taxes and insurance are collected from the borrowers. Taxes and premiums are paid, and P&I are distributed to the investors of the loans. The issuer often has to advance P&I payments due if uncollected, which is referred to as **MBS servicing** [298]. Accounting and monthly reporting are also part of servicing. The **servicing fee** is a percentage of the remaining principal of the loan at the beginning of each month. It is part of the interest portion of the mortgage payment as far as the borrower is concerned. The monthly cash flow from the mortgage hence consists of three parts: servicing fee, interest payment net of the servicing fee, and the scheduled principal repayment. There is a secondary market for servicing rights. The cash flow of servicing right is uncertain because of the prepayment uncertainty. Figure 28.2 shows a typical monthly mortgage statement.

28.3 Agencies and Securitization

The existence of a secondary market is key to the liquidity of mortgages. Government agencies were created by Congress to foster the growth of this market. The means of providing such liquidity was the creation of securities backed by a pool of mortgages and guaranteed by these agencies. With the increase in liquidity and the reduction in credit risk comes the creation of products offering varieties of risk/return patterns. These products in turn attract investors to participate in the mortgage market (see Fig. 28.3).

Mortgage securitization commenced in February 1970 with the issuance of Ginnie Mae Pool #1, a mortgage pass-through. Explosive growth of the market came

Outstanding volume of agency MBSs (U.S.$ billions)

	GNMA	*FNMA*	*FHLMC*	*Total*		*GNMA*	*FNMA*	*FHLMC*	*Total*
1980	93.9	—	17.0	110.9	**1990**	403.6	299.8	321.0	1,024.4
1981	105.8	0.7	19.9	126.4	**1991**	425.3	372.0	363.2	1,160.5
1982	118.9	14.4	43.0	176.3	**1992**	419.3	445.0	409.2	1,273.5
1983	159.8	25.1	59.4	244.3	**1993**	414.0	495.5	440.1	1,349.6
1984	180.0	36.2	73.2	289.4	**1994**	450.9	530.3	460.7	1,441.9
1985	212.1	55.0	105.0	372.1	**1995**	472.3	583.0	515.1	1,570.4
1986	262.7	97.2	174.5	534.4	**1996**	506.2	650.7	554.3	1,711.2
1987	315.8	140.0	216.3	672.1	**1997**	536.8	709.6	579.4	1,825.8
1988	340.5	178.3	231.1	749.9	**1998**	537.4	834.5	646.5	2,018.4
1989	369.9	228.2	278.2	876.3	**1999**	582.0	960.9	749.1	2,292.0

Issuance of agency MBSs (U.S.$ billions)

	GNMA	*FNMA*	*FHLMC*	*Total*		*GNMA*	*FNMA*	*FHLMC*	*Total*
1980	20.6	—	2.5	23.1	**1990**	64.4	96.7	73.8	234.9
1981	14.3	0.7	3.5	18.5	**1991**	62.6	112.9	92.5	268.0
1982	16.0	14.0	24.2	54.2	**1992**	81.9	194.0	179.2	455.2
1983	50.7	13.3	21.4	85.4	**1993**	138.0	221.4	208.7	568.1
1984	28.1	13.5	20.5	62.1	**1994**	111.2	130.6	117.1	359.0
1985	46.0	23.6	41.5	111.1	**1995**	72.9	110.5	85.9	269.2
1986	101.4	60.6	102.4	264.4	**1996**	100.9	149.9	119.7	370.5
1987	94.9	63.2	75.0	233.1	**1997**	104.3	149.4	114.3	368.0
1988	55.2	54.9	39.8	149.9	**1998**	150.2	326.1	250.6	726.9
1989	57.1	69.8	73.5	200.4	**1999**	152.8	300.7	233.0	686.5

Figure 28.3: Agency MBSs 1980–1999. Source: Public Securities Association.

later in late 1981 when Fannie Mae and Freddie Mac started their mortgage swap programs. These developments allow mortgage holders – primarily thrifts – to sell their mortgages to agencies in return for agency-guaranteed pass-through securities backed by the same mortgages. Developments such as these have profound social implications. For example, they lower the cost of financing home ownership.

Among the three housing-related federal agencies, Ginnie Mae, Freddie Mac, and Fannie Mae, only Ginnie Mae is a government corporation within the Department of Housing and Urban Development (HUD).[1] Its guarantee hence carries the full faith and credit of the U.S. Treasury. MBSs with such a guarantee are perceived to have zero default risk. Ginnie Mae guarantees only government-insured or government-guaranteed loans in its programs, whereas Freddie Mac and Fannie Mae are government-sponsored enterprises that mainly use conventional mortgages in their programs. Securities offered by Ginnie Mae, Freddie Mac, and Fannie Mae are commonly referred to as "Ginnie Maes," "Freddie Macs," and "Fannie Maes," respectively.

Agency guarantees come in two forms. One type guarantees the *timely* payment of P&I. Under this guarantee, the P&I will be paid when due even if some of the mortgagors do not pay the monthly mortgage on time, if at all. Pass-throughs carrying this form of guarantee are called **fully modified pass-throughs**. For instance, Ginnie Mae either uses excess cash or borrows from the Treasury if the homeowner

payments are late. All Ginnie Mae MBSs are fully modified pass-throughs. The second type guarantees the timely payment of interest and the ultimate payment of principal, say within a year. Pass-throughs carrying this form of guarantee are referred to as **modified pass-throughs**. Guarantees turn defaults into prepayments from the investor's point of view.

Although Fannie Mae and Freddie Mac buy only conforming mortgages, private conduits buy both conforming and nonconforming mortgages. Being **nonconforming** does not imply greater credit risk. Without explicit or implicit government guarantees on the underlying loans, the so-called **private-label** or **conventional pass-throughs**, which made their debut in 1977, receive high credit ratings through **credit enhancements**.

Traditional mortgages are fixed rate. Record-high fixed mortgage rates in the early 1980s led to the development of adjustable-rate mortgages (ARMs), which were first marketed in late 1983. ARMs are attractive for many reasons. First, the initial rate is typically several percentage points below that of fixed-rate mortgages. It is hence called the "**teaser**" **rate**. Because the home buyer qualifies for the mortgage at the initial loan rate, ARMs allow more people to qualify for a mortgage loan. By the same token, the home buyer can qualify for a larger loan with ARM financing. Second, the index used to adjust the rate is usually tied to a widely recognized and available index. This makes pricing and hedging practical. Third, the interest rate adjustments permitted by ARMs are capped, which insulates the mortgagor from loan payment shock during prolonged periods of rising interest rates. Fourth, ARMs represent an attractive investment for institutional investors such as thrifts and savings and loans because ARMs match their variable-rate liabilities better (review Subsection 4.2.3 for this point). Naturally, ARMs are less competitive against fixed-rate mortgages during the periods when the fixed mortgage rates are relatively low.

ARMs financing reduces the housing industry's sensitivity to interest rate fluctuations because borrowers can choose between fixed- and adjustable-rate mortgages based on the prevailing interest rate levels. MPTSs backed by ARMs were created by Fannie Mae in 1984.

28.4 Mortgage-Backed Securities

In the simplest kind of MBS, the MPTs, payments from the underlying mortgages are passed from the mortgage holders through the servicing agency, after a fee is subtracted, and distributed to the security holder on a pro rata basis (see Fig. 28.4). This means that the holder of a $25,000 certificate from a $1 million pool is entitled to $2_{1/2}$% of the cash flow paid by the mortgagors. Because of higher marketability, a pass-through is easier to sell than its individual loans.

A pass-through still exposes the investor to the total prepayment risk associated with the underlying mortgages. Such risk is undesirable from an asset/liability perspective. To deal with prepayment uncertainty, CMOs were created in June 1983 by Freddie Mac with the help of the then First Boston. Unlike mortgage pass-throughs, which have a single maturity and are backed by individual mortgages, CMOs are *multiple*-maturity, *multi*class debt instruments collateralized by pass-throughs, SMBSs, and whole loans. The process of using pass-throughs and SMBSs to create CMOs is called **resecuritization**. The total prepayment risk is now divided among classes of bonds called **classes** or **tranches**.[2] The principal, scheduled and

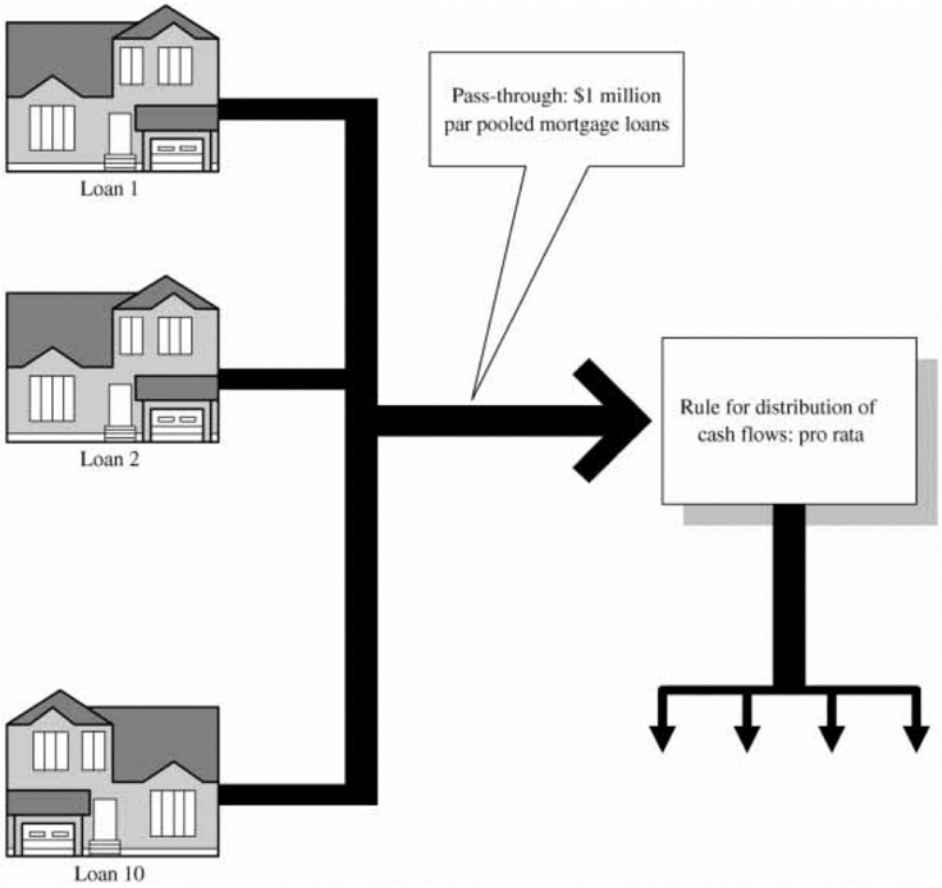


Figure 28.4: Mortgage pass-throughs.

prepaid, is allocated on a prioritized basis so as to redistribute the prepayment risk among the tranches in an unequal way.

In the **sequential tranche paydown structure**, for example, Class A receives principal paydown and prepayments before Class B, which in turn does it before Class C, and so on. Each tranche thus has a different effective maturity. Each tranche may even have a different coupon rate. CMOs were the first successful attempt to alter mortgage cash flows in a security form that attracts a wide range of investors (see Fig. 28.5).

EXAMPLE 28.4.1 Consider a two-tranche sequential pay CMO backed by $1,000,000 of mortgages with a 12% coupon and 6 months to maturity. The cash flow pattern for each tranche with zero prepayment and zero servicing fee is shown in Fig. 28.6. The calculation can be carried out first for the Total columns, which make up the amortization schedule, before the cash flow is allocated. Note that tranche A is retired after 4 months, and tranche B starts principal paydown at the end of month four.

EXAMPLE 28.4.2 (*Continued*) When prepayments are present the calculation is slightly more complex. Suppose the **single monthly mortality** (**SMM**) per month is 5%, which means that the prepayment amount is 5% of the remaining principal. The remaining principal at month i *after* prepayment then equals the scheduled remaining

Outstanding volume of agency collateralized mortgage obligations (U.S.$ billions)

	GNMA	FNMA	FHLMC	Total		GNMA	FNMA	FHLMC	Total
1987	–	0.9	–	0.9	**1994**	–	315.0	263.7	578.7
1988	–	11.6	10.9	22.5	**1995**	–	294.0	247.0	540.9
1989	–	47.6	47.6	95.2	**1996**	–	283.4	237.6	521.0
1990	–	104.3	83.4	187.7	**1997**	17.5	328.6	233.6	579.7
1991	–	193.3	43.0	336.3	**1998**	29.0	311.4	260.3	600.8
1992	–	276.9	217.0	494.0	**1999**	52.5	293.6	316.1	662.1
1993	–	323.4	264.1	587.6					

Issuance of agency collateralized mortgage obligations (U.S.$ billions)

	GNMA	FNMA	FHLMC	Total		GNMA	FNMA	FHLMC	Total
1987	–	0.9	–	0.9	**1994**	3.1	56.3	73.1	132.6
1988	–	11.2	13.0	24.2	**1995**	1.9	8.2	15.4	25.4
1989	–	37.6	39.8	77.3	**1996**	9.5	26.6	34.1	70.2
1990	–	60.9	40.5	101.4	**1997**	7.9	74.8	84.4	167.0
1991	–	101.8	72.0	173.8	**1998**	13.6	76.3	135.2	225.1
1992	–	154.8	131.3	286.1	**1999**	29.6	50.6	119.6	199.7
1993	–	168.0	143.3	311.3					

Figure 28.5: Agency CMOs 1987–1999. Source: Public Securities Association.

principal as computed by Eq. (3.8) times $(0.95)^i$. This done for all the months, the total interest payment at any month is the remaining principal of the previous month times 1%. And the prepayment amount equals the remaining principal times 0.05/0.95 (the division by 0.95 yields the remaining principal *before* prepayment). Figure 28.7 tabulates the cash flows of the same two-tranche CMO under 5% SMM. For instance, the total principal payment at month one, $204,421, can be verified as follows. The scheduled remaining principal is $837,452 from Fig. 28.6. The remaining principal is hence $837452 \times 0.95 = 795579$, which makes the total principal payment $1000000 - 795579 = 204421$. Because tranche A's remaining principal is $500,000, all 204,421 dollars go to tranche A. Incidentally, the prepayment is $837452 \times 5\% = 41873$ (alternatively, $795579 \times 0.05/0.09$). Note that tranche A is retired after 3 months, and tranche B starts principal paydown at the end of month three.

	Interest			Principal			Remaining principal		
Month	A	B	Total	A	B	Total	A	B	Total
							500,000	500,000	1,000,000
1	5,000	5,000	10,000	162,548	0	162,548	337,452	500,000	837,452
2	3,375	5,000	8,375	164,173	0	164,173	173,279	500,000	673,279
3	1,733	5,000	6,733	165,815	0	165,815	7,464	500,000	507,464
4	75	5,000	5,075	7,464	160,009	167,473	0	339,991	339,991
5	0	3,400	3,400	0	169,148	169,148	0	170,843	170,843
6	0	1,708	1,708	0	170,843	170,843	0	0	0
Total	10,183	25,108	35,291	500,000	500,000	1,000,000			

Figure 28.6: CMO cash flows without prepayments. The total monthly payment is $172,548. Month-$i$ numbers reflect the ith monthly payment.

	Interest			*Principal*			*Remaining principal*		
Month	*A*	*B*	*Total*	*A*	*B*	*Total*	*A*	*B*	*Total*
							500,000	500,000	1,000,000
1	5,000	5,000	10,000	204,421	0	204,421	295,579	500,000	795,579
2	2,956	5,000	7,956	187,946	0	187,946	107,633	500,000	607,633
3	1,076	5,000	6,076	107,633	64,915	172,548	435,085	435,085	
4	0	4,351	4,351	0	158,163	158,163	0	276,922	276,922
5	0	2,769	2,769	0	144,730	144,730	0	132,192	132,192
6	0	1,322	1,322	0	132,192	132,192	0	0	0
Total	9,032	23,442	32,474	500,000	500,000	1,000,000			

Figure 28.7: CMO cash flows with prepayments. Month-i numbers reflect the ith monthly payment.

SMBSs were created in February 1987 when Fannie Mae issued its Trust 1 SMBS. For SMBSs, the P&I are divided between the PO strip and the IO strip. In the scenarios of Examples 28.4.1 and 28.4.2, the IO strip receives all the interest payments under the Interest/Total column, whereas the PO strip receives all the principal payments under the Principal/Total column. These new instruments allow investors to better exploit anticipated changes in interest rates. Because the collateral for an SMBS is a pass-through, this is yet another example of resecuritization. CMOs and SMBSs are usually called **derivative MBS**s

➤ **Exercise 28.4.1** Repeat the calculations in Example 28.4.2 under 3% SMM.

28.5 Federal Agency Mortgage-Backed Securities Programs

28.5.1 Government National Mortgage Association ("Ginnie Mae")

Security guaranteed by Ginnie Mae is called an MBS. Ginnie Mae issues its MBSs under one of two programs, GNMA I (established in 1970) and GNMA II (established in 1983). The two programs differ in terms of the collateral underlying the pass-throughs. For example, GNMA I MBSs require all loans in a pool to be approximately homogeneous [297]. A GNMA I MBS is issued with an annual coupon rate that is 0.50% lower than the coupon rate on the underlying mortgages because of guarantee and servicing fees. MBSs backed by **adjustable-payment mortgages** (**APM**s) are issued under the GNMA II program.

The issuer of a Ginnie Mae security passes through the scheduled P&I payments on the underlying mortgages to security holders each month even if the issuer does not collect payments from some mortgagors. It also passes through any additional principal prepayments because of foreclosure settlements. If the issuer defaults on the monthly payments, Ginnie Mae assumes responsibility for the timely payment of P&I.

➤ **Exercise 28.5.1** Even without prepayments, the scheduled monthly payment to MBS holders increases slightly over time. Why?

28.5.2 Federal Home Loan Mortgage Corporation ("Freddie Mac")

Freddie Mac was created on July 24, 1970, as a government-charted corporation. It became a public corporation like Fannie Mae in 1989. Freddie Mac seeks to

increase liquidity and available credit for the conventional mortgage market by establishing and maintaining a secondary market for such mortgages. It started issuing pass-through securities in 1971, which was the first time conventional mortgages were securitized with a federal agency guarantee. Its mortgage pass-throughs are referred to as PCs. Unlike the Ginnie Mae pass-throughs, the Freddie Mac pass-throughs guarantee only eventual repayment of principal. In the fall of 1990, Freddie Mac introduced its Gold PC, which has stronger guarantees: All Gold PCs are fully modified pass-throughs. Freddie Mac securities are not backed by the full faith and credit of the U.S. government. The credit of its securities is perceived to be equivalent to that of securities issued by U.S. government agencies ("U.S. agency" status).

Freddie Mac issues CMOs and SMBSs besides PCs. All Freddie Mac CMOs have semiannual payments much like bonds. They also use only fixed-rate mortgages as collateral and a guaranteed sinking fund to establish minimum principal prepayments.

28.5.3 Federal National Mortgage Association ("Fannie Mae")

Established in 1938, Fannie Mae is the oldest of the three agencies and one of the largest corporations in the United States in terms of assets (U.S.$575 billion as of the end of 1999). It introduced the mortgage pass-through program in 1981. Pass-throughs issued by Fannie Mae are called MBSs. Fannie Mae guarantees the timely payment of both principal and interest on its MBS whether or not the payments have been collected from the borrower. The guarantee encompasses principal payments resulting from foreclosure or prepayment; the securities are fully modified pass-throughs, in other words. Although Fannie Mae obligations are not backed by the full faith and credit of the U.S. government, they carry "U.S. agency" status in the credit markets.

28.6 Prepayments

The prepayment option sets MBSs apart from other fixed-income securities. The exercise of options on most securities is expected to be "rational" in the sense that it will be executed only when it is profitable to do so. This kind of "rationality" is weakened when it comes to the homeowner's decision to prepay. For example, even when the prevailing mortgage rate, called the **current coupon**, exceeds the mortgage's loan rate, some loans remain prepaid.

Prepayment risk refers to the uncertainty in the amount and timing of the principal prepayments in the pool of mortgages that collateralize the security. This risk can be divided into **contraction risk** and **extension risk**. Contraction risk refers to the risk of having to reinvest the prepayments at a rate lower than the coupon rate when interest rates decline. Extension risk is due to the slowdown of prepayments when interest rates climb, making the investor earn the security's lower coupon rate rather than the market's higher rate. Prepayments can be in whole or in part; the former is called **liquidation**, and the latter **curtailment**. Prepayments, however, need not always result in losses (see Exercise 28.6.1). The holder of a pass-through security is exposed to the total prepayment risk associated with the underlying pool of mortgage loans, whereas the CMO is designed to alter the distribution of that risk among investors.

Besides prepayment risk, investors in mortgages are exposed to at least three other risks: interest rate risk, credit risk, and liquidity risk. Interest rate risk is inherent in any fixed-income security. Credit risk is the risk of loss from default. It is almost nonexisting for FHA-insured and VA-guaranteed mortgages. As for privately insured mortgage, the risk is related to the credit rating of the company that insures the mortgage. Liquidity risk is the risk of loss if the investment must be sold quickly.

➤ **Exercise 28.6.1** There are reasons prepayments arising from lower interest rates increase the return of a pass-through if it was purchased at a discount. What are they?

28.6.1 Causes and Characteristics

Prepayments have at least five components [4, 433].

Home sale ("housing turnover"). The sale of a home generally leads to the prepayment of mortgage because of the full payment of the remaining principal. This **due-on-sale** clause applies to most conventional loans. Exceptions are FHA/VA mortgages, which are **assumable**, meaning the buyer can assume the existing loan.

Refinancing. Mortgagors can refinance their home mortgage at a lower mortgage rate. This is the most volatile component of prepayment and constitutes the bulk of it when prepayments are extremely high.

Default. This type of prepayment is caused by foreclosure and subsequent liquidation of a mortgage. It is relatively minor in most cases.

Curtailment. As the extra payment above the scheduled payment, curtailment applies to the principal and shortens the maturity of fixed-rate loans. Its contribution to prepayments is minor.

Full payoff (liquidation). There is evidence that many mortgagors pay off their mortgage completely when it is very **seasoned** and the remaining balance is small. Full payoff can also be due to natural disasters. It is important for only very seasoned loans.

Prepayments exhibit certain characteristics [504]. They usually increase as the mortgage ages – first at an increasing rate and then at a decreasing rate. They are higher in the spring and summer and lower in the fall and winter. They vary by the geographic locations of the underlying properties. Prepayments increase when interest rates drop but with a time lag. If prepayments were higher for some time because of high refinancing rates, they tend to slow down. Perhaps homeowners who do not prepay when rates have been low for a prolonged time tend never to prepay.

Figure 28.8 illustrates the typical price/yield curves of the Treasury and pass-through. As yields fall and the pass-through's price moves above a certain price, it flattens and then follows a downward slope. This phenomenon is called the price compression of premium-priced MBSs. It demonstrates the negative convexity of such securities.

➤ **Exercise 28.6.2** Given that refinancing involves certain fixed costs, which will tend to prepay faster, mortgage securities backed by 15-year mortgages or 30-year mortgages?

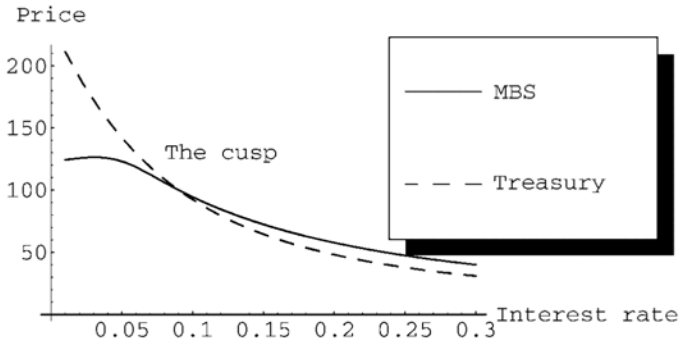


Figure 28.8: MBS vs. Treasury. Both are 15-year securities paying a 9% coupon rate in mortgage-equivalent yield. The segment above 100 means the security is premium-priced, whereas the segment below 100 signifies discount securities. Price compression occurs as yields fall through a threshold. The **cusp** represents that point.

28.6.2 An Analysis of the Incentive to Refinance

Consider a loan with a mortgage rate r_o for a term of n months. Let the scheduled monthly payment of the original loan be C. At the time of refinancing, the mortgage rate for a new n-month loan is r_n, and a monthly payments have been remitted. Both r_o and r_n are monthly rates.

From Eq. (3.8), the remaining principal at the time of refinancing is

$$C\,\frac{1-(1+r_o)^{-n+a}}{r_o}. \tag{28.1}$$

At the current rate r_n, the future cash flow of the original loan has a PV of

$$\sum_{i=1}^{n-a} C(1+r_n)^{-i} = C\,\frac{1-(1+r_n)^{-n+a}}{r_n}.$$

Therefore the net monetary savings are

$$C\,\frac{1-(1+r_n)^{-n+a}}{r_n} - C\,\frac{1-(1+r_o)^{-n+a}}{r_o}. \tag{28.2}$$

Divide the preceding expression by expression (28.1) to obtain the savings per dollar of the remaining principal as

$$\frac{r_o}{r_n}\,\frac{1-(1+r_n)^{-n+a}}{1-(1+r_o)^{-n+a}} - 1.$$

For loans that have not seasoned sufficiently, the preceding expression is roughly

$$\frac{r_o}{r_n} - 1. \tag{28.3}$$

This heuristic argument points to using the *ratio* of loan rates rather than the *difference* to measure the incentive to refinancing [433].

➤ **Exercise 28.6.3** Does it make economic sense to refinance a mortgage if rates have not changed?

➤ **Exercise 28.6.4** Consider a mortgagor who refinances every a months with an n-month loan every time. Show that the monthly payment after the ith refinancing is

$$\text{original balance} \times \left[\frac{(1+r)^n - (1+r)^a}{(1+r)^n - 1} \right]^i \frac{r(1+r)^n}{(1+r)^n - 1},$$

where r is the unchanging monthly mortgage rate.

➤ **Exercise 28.6.5** Which represents a better deal, refinancing from an 8% loan to a 6% loan or from an 11.5% loan to a 9.5% loan?

Additional Reading

This chapter reviewed the mortgage markets, the institutions, the securitization of mortgages, and various mortgage products. Consult [54, 323, 325, 330, 331, 432, 469, 698, 799] for more background information and particularly [54] for a history of the MBS market. References [320, 324, 328] are also rich sources of information. See [54, Table 3.1] and [432, Exhibit 24-3] for other differences between Freddie Mac and Ginnie Mae pass-throughs. That securitization lowers the mortgage rates is not without its dissents [404].

NOTES

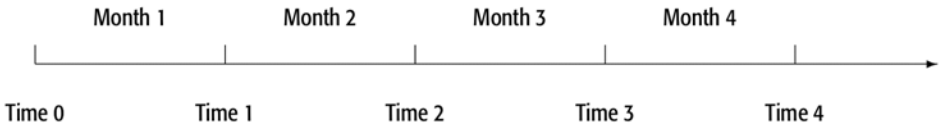
1. Fannie Mae used to be a government agency before being sold to the public in 1968.
2. *Tranche* is a French word for "slice."

CHAPTER
TWENTY-NINE

Analysis of Mortgage-Backed Securities

Oh, well, if you cannot measure, measure anyhow.
Frank H. Knight (1885–1972)

Compared with other fixed-income securities, the MBS is unique in two respects. First, its cash flow consists of PRINCIPAL AND INTEREST (P&I). Second, the cash flow may vary because of prepayments in the underlying mortgages. This chapter covers the MBS's cash flow and valuation. We adopt the following time line when discussing cash flows:



Because mortgage payments are paid in arrears, a payment for month i occurs at time i, that is, end of month i. The end of a month is identified with the beginning of the coming month.

29.1 Cash Flow Analysis

A traditional mortgage has a fixed term, a fixed interest rate, and a fixed monthly payment. Figure 29.1 illustrates the scheduled P&I for a 30-year, 6% mortgage with an initial balance of $100,000. Figure 29.2 shows how the remaining principal balance decreases over time. In the early years, the P&I consists mostly of interest. Then it gradually shifts toward principal payment with the passage of time. However, the total P&I payment remains the same each month, hence the term *level* pay. Identical characteristics hold for the pool's P&I payments in the absence of prepayments and servicing fees.

From the discussions in Section 3.3, we know that the remaining principal balance after the kth payment is

$$C\,\frac{1-(1+r/m)^{-n+k}}{r/m}, \tag{29.1}$$

where C is the scheduled P&I payment of an n-month mortgage making m payments per year and r is the annual mortgage rate. For mortgages, $m = 12$. The remaining

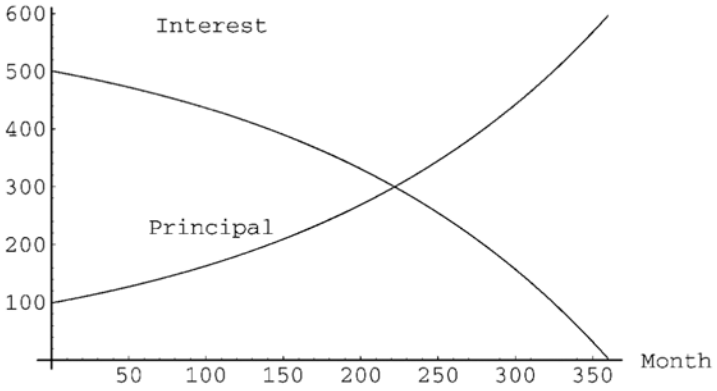


Figure 29.1: Scheduled P&I payments. The schedule is for a 30-year 6% mortgage with an original loan amount of $100,000.

principal balance after k payments can be expressed as a portion of the original principal balance; thus

$$\text{Bal}_k \equiv 1 - \frac{(1+r/m)^k - 1}{(1+r/m)^n - 1} = \frac{(1+r/m)^n - (1+r/m)^k}{(1+r/m)^n - 1}. \tag{29.2}$$

We can verify this equation by dividing balance (29.1) by Bal_0. The remaining principal balance after k payments is simply

$$\text{RB}_k \equiv \mathcal{O} \times \text{Bal}_k,$$

where $\mathcal{O}$ is the original principal balance.

The term **factor** denotes the portion of the remaining principal balance to its original principal balance expressed as a decimal [729]. So Bal_k is the monthly factor when there are no prepayments. It is also known as the **amortization factor**. When the idea of factor is applied to a mortgage pool, it is called the **paydown factor on the pool** or simply the **pool factor** [298].

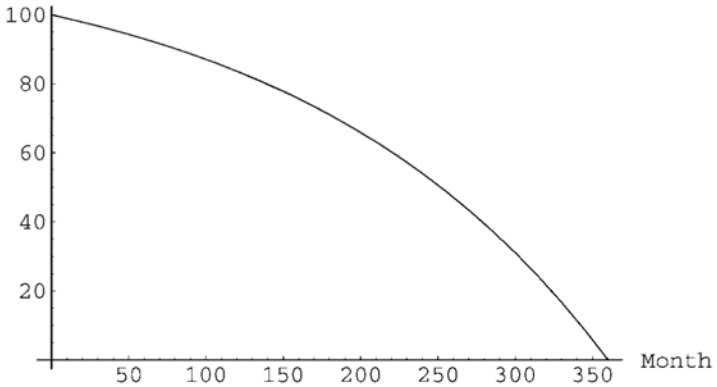


Figure 29.2: Scheduled remaining principal balances. Plotted are the remaining principal balances as percentages of par after each scheduled payment is made.

EXAMPLE 29.1.1 The remaining balance of a 15-year mortgage with a 9% mortgage rate after 54 months is

$$\mathcal{O} \times \frac{[1+(0.09/12)]^{180} - [1+(0.09/12)]^{54}}{[1+(0.09/12)]^{180} - 1} = \mathcal{O} \times 0.824866.$$

In other words, roughly 82.49% of the original loan amount remains after 54 months.

By the amortization principle, the tth interest payment is

$$I_t \equiv \mathrm{RB}_{t-1} \times \frac{r}{m} = \mathcal{O} \times \frac{r}{m} \times \frac{(1+r/m)^n - (1+r/m)^{t-1}}{(1+r/m)^n - 1}.$$

The principal part of the tth monthly payment is

$$P_t \equiv \mathrm{RB}_{t-1} - \mathrm{RB}_t = \mathcal{O} \times \frac{(r/m)(1+r/m)^{t-1}}{(1+r/m)^n - 1}. \tag{29.3}$$

The scheduled P&I payment at month t, or $P_t + I_t$, is therefore

$$(\mathrm{RB}_{t-1} - \mathrm{RB}_t) + \mathrm{RB}_{t-1} \times \frac{r}{m} = \mathcal{O} \times \left[\frac{(r/m)(1+r/m)^n}{(1+r/m)^n - 1} \right], \tag{29.4}$$

indeed a level pay independent of t. The term within the brackets, called the **payment factor** or **annuity factor**, represents the monthly payment for each dollar of mortgage.

EXAMPLE 29.1.2 The mortgage in Example 3.3.1 has a monthly payment of

$$250{,}000 \times \frac{(0.08/12) \times [1+(0.08/12)]^{180}}{[1+(0.08/12)]^{180} - 1} = 2{,}389.13$$

by Eq. (29.4), in total agreement with the number derived there.

➤ **Exercise 29.1.1** Derive Eq. (29.4) from Eq. (3.6).

➤ **Exercise 29.1.2** Consider two mortgages with identical remaining principals but different mortgage rates. Show that their remaining principal balances after the next monthly payment will be different; in fact, the mortgage with a lower mortgage rate amortizes faster.

29.1.1 Pricing Adjustable-Rate Mortgages

We turn to ARM pricing as an interesting application of derivatives pricing and the analysis above. Consider a 3-year ARM with an interest rate that is 1% above the 1-year T-bill rate at the beginning of the year. This 1% is called the **margin**. For simplicity, assume that this ARM carries annual, not monthly, payments. The T-bill rates follow the binomial process, in boldface, in Fig. 29.3, and the risk-neutral probability is 0.5. How much is the ARM worth to the issuer?

Each new coupon rate at the reset date determines the level mortgage payment for the months until the next reset date as if the ARM were a fixed-rate loan with the new coupon rate and a maturity equal to that of the ARM. This implies, for example, that in the interest rate tree of Fig. 29.3 the scenario $\mathrm{A} \to \mathrm{B} \to \mathrm{E}$ will leave our 3-year ARM with a remaining principal at the end of the second year different from

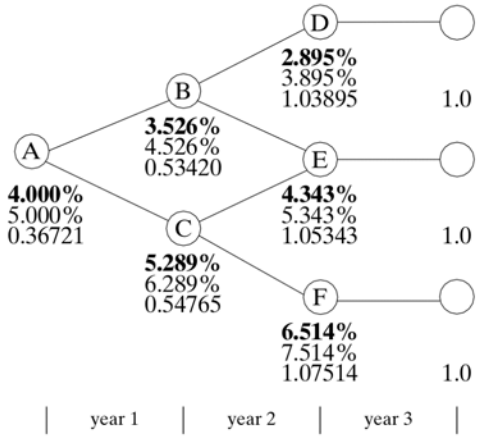


Figure 29.3: ARM's payment factors under stochastic interest rates. Stacked at each node are the T-bill rate, the mortgage rate (which is 1% above the T-bill rate), and the payment factor for a mortgage initiated at that node and ending at the end of year three (based on the mortgage rate at the same node, of course). The short rates are from Fig. 23.8.

that under the scenario $A \to C \to E$ (see Exercise 29.1.2). This path dependency calls for care in algorithmic design to avoid exponential complexity.

The idea is to attach to each node on the binomial tree the annual payment per \$1 of principal for a mortgage initiated at that node and ending at the end of year three – in other words, the payment factor [546]. At node B, for example, the annual payment factor can be calculated by Eq. (29.4) with $r = 0.04526, m = 1$, and $n = 2$ as

$$\frac{0.04526 \times (1.04526)^2}{(1.04526)^2 - 1} = 0.53420.$$

The payment factors for other nodes in Fig. 29.3 are calculated in the same manner.

We now apply backward induction to price the ARM (see Fig. 29.4). At each node on the tree, the net value of an ARM of value \$1 initiated at that node and ending at the end of the third year is calculated. For example, the value is zero at terminal nodes because the ARM is immediately repaid. At node D, the value is

$$\frac{1.03895}{1.02895} - 1 = 0.0097186,$$

which is simply the NPV of the payment 1.03895 next year (note that the issuer makes a loan of \$1 at D). The values at nodes E and F can be computed similarly. At

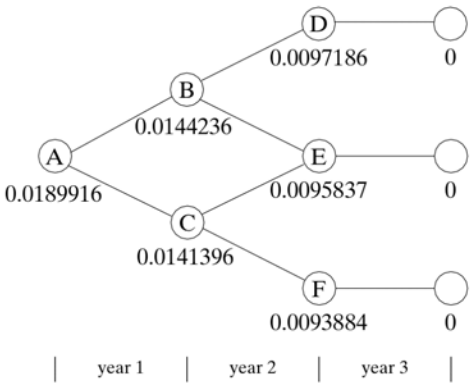


Figure 29.4: Backward induction for ARMs.

node B, we first figure out the remaining principal balance after the payment 1 year hence as

$$1 - (0.53420 - 0.04526) = 0.51106,$$

because \$0.04526 of the payment of \$0.53426 constitutes interest. The issuer will receive \$0.01 above the T-bill rate next year, and the value of the ARM is either \$0.0097186 or \$0.0095837 per \$1, each with probability 0.5. The ARM's value at node B thus is

$$\frac{0.51106 \times (0.0097186 + 0.0095837)/2 + 0.01}{1.03526} = 0.0144236.$$

The values at nodes C and A can be calculated similarly as

$$\frac{[\,1 - (0.54765 - 0.06289)\,] \times (0.0095837 + 0.0093884)/2 + 0.01}{1.05289} = 0.0141396,$$

$$\frac{[\,1 - (0.36721 - 0.05)\,] \times (0.0144236 + 0.0141396)/2 + 0.01}{1.04} = 0.0189916,$$

respectively. The value of the ARM to the issuer is hence \$0.0189916 per \$1 of loan amount. The complete algorithm appears in Fig. 29.5. The above idea of **scaling** has wide applicability for pricing certain classes of path-dependent securities [449, 546].

ARMs are indexed to publicly available indices such as LIBOR, the constant-maturity Treasury (CMT) rate, and the Cost of Funds Index (COFI). The CMT rates are based on the daily CMT yield curve constructed by the Federal Reserve Bank

Algorithm for pricing ARMs:

```
input:   n, r[n][n], s;
real     P[n], f, p;
integer  i, j;
for (j = 0 to n-1) { // Nodes at time n-1.
         f := 1 + r[n-1][j] + s;   //(29.4) with n = 1.
         P[j] := f/(1 + r[n-1][j]) - 1;
}
for (i = n-2 down to 0) // Nodes at time i.
         for (j = 0 to i)  {
            f := (r[i][j] + s)(1 + r[i][j] + s)^(n-i) x
                 ((1 + r[i][j] + s)^(n-i) - 1)^(-1);  //See (29.4).
            p := 1 - (f - r[i][j] - s);
            P[j] := (p x (P[j] + P[j+1]) x 0.5 + s) x
                 (1 + r[i][j])^(-1);
         }
return P[0];
```

Figure 29.5: Algorithm for pricing ARMs. $r[\,i\,][\,j\,]$ is the $(j+1)$th T-bill rate for period $i+1$, the ARM has n periods to maturity, s is the margin, f stores the payment factors, and p stores the remaining principal amounts. All rates are measured by the period. In general, the floating rate may be based on the k-period Treasury spot rate plus a spread. Then Programming Assignment 29.1.3 can be used to generate the k-period spot rate at each node.

of New York and published weekly in the Federal Reserve's *Statistical Release* H.15 [525]. Cost of funds for thrifts indices are calculated based on the monthly weighted average interest cost for thrifts. The most popular cost of funds index is the 11th Federal Home Loan Bank Board District COFI [325, 330, 820].

If the ARM coupon reflects fully and instantaneously current market rates, then the ARM security will be priced close to par and refinancings rarely occur. In reality, adjustments are imperfect in many ways. At the reset date, a margin is added to the benchmark index to determine the new coupon. ARMs also often have **periodic rate caps** that limit the amount by which the coupon rate may increase or decrease at the reset date. They also have **lifetime caps** and **floors**. To attract borrowers, mortgage lenders usually offer a below-market initial rate (the "teaser" rate). The **reset interval**, the time period between adjustments in the ARM coupon rate, is often annual, which is not frequent enough. Note that these terms are easy to incorporate into the pricing algorithm in Fig. 29.5.

➢ **Programming Assignment 29.1.3** Given an n-period binomial short rate tree, design an $O(kn^2)$-time algorithm for generating k-period spot rates on the nodes of the tree. This tree documents the dynamics of the k-period spot rate.

➢ **Programming Assignment 29.1.4** Implement the algorithm in Fig. 29.5. The binomial T-bill rate tree and the mortgage rate as a spread over the T-bill rate are parts of the input.

➢ **Programming Assignment 29.1.5** Consider an IAS with an amortizing schedule that depends solely on the prevailing k-period spot interest rate. This swap's cash flow depends on only the prevailing principal amount and the prevailing k-period spot interest rate. Design an efficient algorithm to price this swap on a binomial short rate tree.

29.1.2 Expressing Prepayment Speeds

The cash flow of a mortgage derivative is determined from that of the mortgage pool. The single most important factor complicating this endeavor is the unpredictability of prepayments. Recall that prepayment represents the principal payment made in excess of the scheduled principal amortization. We need only compare the amortization factor Bal_t of the pool with the reported factor to determine if prepayments have occurred. The amount by which the reported factor exceeds the amortization factor is the prepayment amount.

Single Monthly Mortality

An SMM of ω means that $\omega\%$ of the scheduled remaining balance at the end of the month will prepay. In other words, the SMM is the percentage of the remaining balance that prepays for the month. Suppose the remaining principal balance of an MBS at the beginning of a month is \$50,000, the SMM is 0.5%, and the scheduled principal payment is \$70. Then the prepayment for the month is $0.005 \times (50{,}000 - 70) \approx 250$ dollars. If the same monthly prepayment speed s is maintained since the issuance of the pool, the remaining principal balance at month i will be $\text{RB}_i \times (1 - s/100)^i$. It goes without saying that prepayment speeds must lie between 0% and 100%.

EXAMPLE 29.1.3 Take the mortgage in Example 29.1.1. Its amortization factor at the 54th month is 0.824866. If the actual factor is 0.8, then the SMM for the initial period

of 54 months is

$$100 \times \left[1 - \left(\frac{0.8}{0.824866} \right)^{1/54} \right] = 0.0566677.$$

In other words, roughly 0.057% of the remaining principal is prepaid per month.

Conditional Prepayment Rate

The **conditional prepayment rate** (**CPR**) is the annualized equivalent of an SMM:

$$\text{CPR} = 100 \times \left[1 - \left(1 - \frac{\text{SMM}}{100} \right)^{12} \right].$$

Conversely,

$$\text{SMM} = 100 \times \left[1 - \left(1 - \frac{\text{CPR}}{100} \right)^{1/12} \right].$$

For example, the SMM of 0.0566677 in Example 29.1.3 is equivalent to a CPR of

$$100 \times \left\{ 1 - \left[1 - \left(\frac{0.0566677}{100} \right)^{12} \right] \right\} = 0.677897.$$

Roughly 0.68% of the remaining principal is prepaid annually. Figure 29.6 plots the P&I cash flows under various prepayment speeds. Observe that with accelerated prepayments, the principal cash flow is shifted forward in time.

PSA

In 1985 the Public Securities Association (PSA) standardized a prepayment model. The PSA standard is expressed as a monthly series of CPRs and reflects the increase in CPR that occurs as the pool seasons [619]. The PSA standard postulates the following prepayment speeds: The CPR is 0.2% for the first month, increases thereafter by 0.2% per month until it reaches 6% per year for the 30th month, and then stays at 6% for the remaining years. (At the time the PSA proposed its standard, a seasoned 30-year GNMA's typical prepayment speed was ~6% CPR [260].) The PSA benchmark is also referred to as **100 PSA**. Other speeds are expressed as some percentage of PSA. For example, 50 PSA means one-half the PSA CPRs, 150 PSA means one-and-a-half

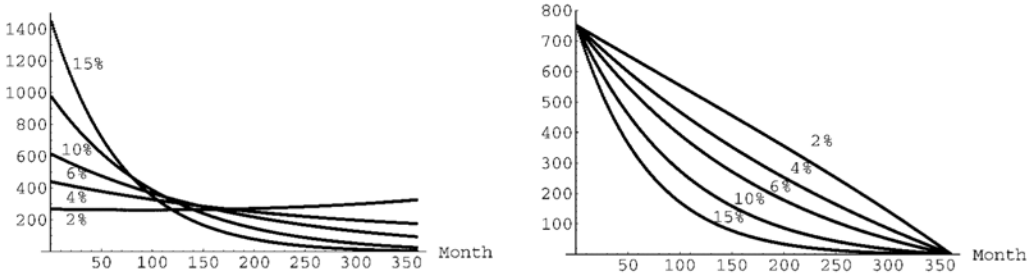


Figure 29.6: Principal (left) and interest (right) cash flows at various CPRs. The 6% mortgage has 30 years to maturity and an original loan amount of $100,000.

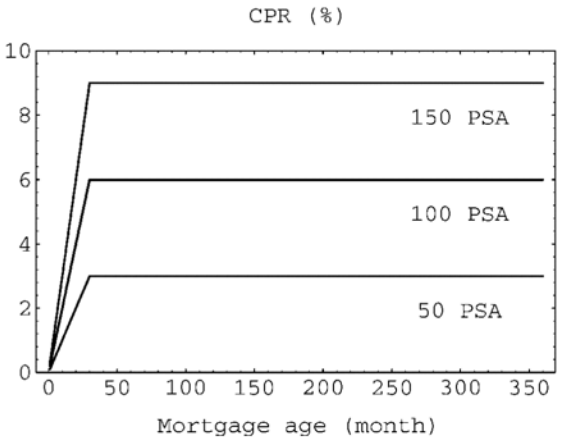


Figure 29.7: The PSA prepayment assumption.

the PSA CPRs, and so on. Mathematically,

$$\text{CPR} = \begin{cases} 6\% \times \frac{\text{PSA}}{100}, & \text{if the pool age exceeds 30 months} \\ 0.2\% \times m \times \frac{\text{PSA}}{100}, & \text{if the pool age } m \le 30 \text{ months} \end{cases}. \tag{29.5}$$

See Fig. 29.7 for an illustration and Fig. 29.8 for the cash flows at 50 and 100 PSAs. Conversely,

$$\text{PSA} = \begin{cases} 100 \times \frac{\text{CPR}}{6}, & \text{if the pool age exceeds 30 months} \\ 100 \times \frac{\text{CPR}}{0.2 \times m}, & \text{if the pool age } m \le 30 \text{ months} \end{cases}.$$

See Fig. 29.9 for the conversion algorithm.

Conversion between PSA and CPR/SMM requires knowing the age of the pool. A prepayment speed of 150 PSA implies a CPR of $0.2\% \times 2 \times (150/100) = 0.6\%$ if the pool is 2 months old, but a CPR of $6\% \times 1.5 = 9\%$ if the pool age exceeds 30 months.

➤ **Exercise 29.1.6** Consider the following PSA numbers:

Month	6	12	18	24	30	36
PSA	100	130	154	230	135	125

Compute their equivalent CPRs.

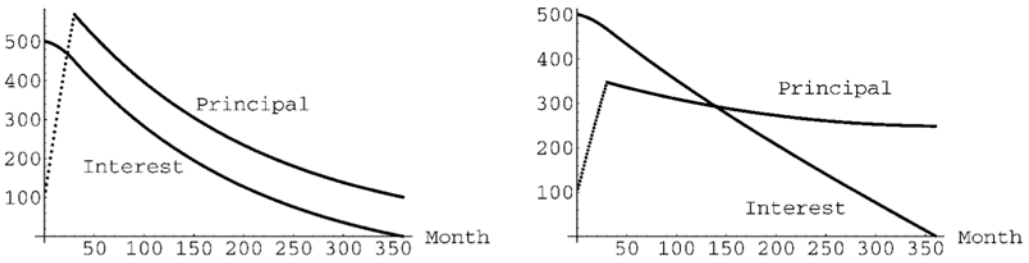


Figure 29.8: P&I payments at 100 PSA (left) and 50 PSA (right). The 6% mortgage has 30 years to maturity and an original loan amount of $100,000.

PSA-to-SMM algorithm:

```
input:    n, PSA, age;
real      SMM[1..n], cpr;
integer   i;
PSA := PSA/100;
for (i = 1 to n) {
        if [i + age ≤ 30]
              cpr := 0.2 × (i + age) × PSA/100;
        else  cpr := 6.0 × PSA/100;
        SMM[i] := 1 − (1 − PSA × cpr)^(1/12);
}
return SMM[ ];
```

Figure 29.9: PSA-to-SMM conversion. The pool has n more monthly cash flows, PSA is the prepayment speed, and age is the number of months since the pool's inception. SMM[ ] stores the prepayment vector in decimal, the ith of which denotes the SMM during month i as seen from now.

➤ **Exercise 29.1.7** Is the SMM assuming 200 PSA twice the SMM assuming 100 PSA?

29.1.3 Prepayment Vector and Cash Flow Analysis

Although it tries to capture, if crudely, how prepayments vary with age, the PSA should be viewed as a market convention rather than as a model. Instead of a single PSA number, a vector of PSAs generated by a prepayment model should be used to describe the monthly prepayment speed through time. The monthly cash flows can be derived thereof.

Similarly, the CPR should be seen purely as a measure of speed rather than a model. When we treat a single CPR number as the true prepayment speed, that number will be called the **constant prepayment rate** for obvious reasons. This simple model fails to address the empirical fact that pools with new production loans typically prepay at a slower rate than seasoned pools. As in the PSA case, a vector of CPRs should be preferred. In practice, a vector of CPRs or SMMs is easier to work with than a vector of PSAs because of the lack of dependence on the pool age. In any case, a CPR vector can always be converted into an equivalent PSA vector and vice versa.

To price an MBS, we start with its cash flow, that is, the periodic P&I under a static prepayment assumption as given by a prepayment vector. The invoice price is now $\sum_{i=1}^{n} C_i/(1+r)^{\omega-1+i}$, where C_i is the cash flow at time i, n is the **weighted average maturity** (**WAM**),[1] r is the discount rate, and ω is the fraction of period from settlement until the first P&I payment date. The WAM is the weighted average remaining term of the mortgages in the pool, where the weight for each mortgage is the remaining balance. The r that equates the above with the market price is called the (**static**) **cash flow yield**. The **implied PSA** is the single PSA speed producing the same cash flow yield.

MBSs are quoted in the same manner as U.S. Treasury notes and bonds. For example, a price of 94-05 means $94_{5/32}$% of par value. Sixty-fourth of a percent is expressed by appending "+" to the price. Hence, the price 94-05+ represents $94_{11/64}$% of par value.

Cash Flow

Each cash flow is composed of the principal payment, the interest payment, and the principal prepayment. Let B_k denote the actual remaining principal balance at month k. Given the pool's actual remaining principal balance at time $i-1$ (i.e., B_{i-1}), the P&I payments at time i are

$$\overline{P_i} \equiv B_{i-1}\left(\frac{\text{Bal}_{i-1}-\text{Bal}_i}{\text{Bal}_{i-1}}\right) = B_{i-1}\,\frac{r/m}{(1+r/m)^{n-i+1}-1}, \tag{29.6}$$

$$\overline{I_i} \equiv B_{i-1}\,\frac{r-\alpha}{m}, \tag{29.7}$$

where α is the **servicing spread** (or servicing fee rate), which consists of the servicing fee for the servicer as well as the guarantee fee. The prepayment at time i is

$$\text{PP}_i = B_{i-1}\,\frac{\text{Bal}_i}{\text{Bal}_{i-1}} \times \text{SMM}_i,$$

where SMM_i is the prepayment speed for month i. If the total principal payment from the pool is $\overline{P_i}+\text{PP}_i$, the remaining principal balance is

$$\begin{aligned} B_i &= B_{i-1} - \overline{P_i} - \text{PP}_i \\ &= B_{i-1}\left[1-\left(\frac{\text{Bal}_{i-1}-\text{Bal}_i}{\text{Bal}_{i-1}}\right)-\frac{\text{Bal}_i}{\text{Bal}_{i-1}}\times\text{SMM}_i\right] \\ &= \frac{B_{i-1}\times\text{Bal}_i\times(1-\text{SMM}_i)}{\text{Bal}_{i-1}}. \end{aligned} \tag{29.8}$$

Equation (29.8) can be applied iteratively to obtain

$$B_i = \text{RB}_i \times \prod_{j=1}^{i}(1-\text{SMM}_j). \tag{29.9}$$

Define $b_i \equiv \prod_{j=1}^{i}(1-\text{SMM}_j)$. Then the scheduled P&I is

$$\overline{P_i} = b_{i-1}P_i, \quad \overline{I_i} = b_{i-1}I_i' \tag{29.10}$$

where $I_i' \equiv \text{RB}_{i-1}\times(r-\alpha)/m$ is the scheduled interest payment. The scheduled cash flow and the b_is determined from the prepayment vector are therefore all that are needed to calculate the projected actual cash flows. Note that if the servicing fees do not exist (that is, $\alpha=0$), the projected monthly payment *before* prepayment at month i becomes

$$\overline{P_i} + \overline{I_i} = b_{i-1}(P_i+I_i) = b_{i-1}C, \tag{29.11}$$

where C is the scheduled monthly payment on the original principal. See Fig. 29.10 for a linear-time algorithm for generating the mortgage pool's cash flow.

Servicing and guarantee fees are deducted from the gross **weighted average coupon** (**WAC**) of the aggregate mortgage P&I to obtain the **pass-through rate**. The WAC is the weighted average of all the mortgage rates in the pool, in which the weight used for each mortgage is the remaining balance. The servicing spread

Mortgage pool cash flow under prepayments:

```
input:   n, r (r > 0), SMM[ 1..n ];
real     B[ n+1 ], P[ 1..n ], Ī[ 1..n ], PP[ 1..n ], b;
integer  i;
b := 1;
B[ 0 ] := 1;
for (i = 1 to n)  {
         b := b × (1 − SMM[ i ]);   //See (29.9).
         B[ i ] := b × ((1+r)^n − (1+r)^i)/((1+r)^n − 1);  //  See (29.2).
         P[ i ] := B[ i − 1 ] − B[ i ];
         Ī[ i ] := B[ i − 1 ] × r;   //See (29.7).
         PP [ i ] := B[ i ] × SMM[ i ]/(1 − SMM[ i ]);
}
return B[ ], P[ ], Ī[ ], PP[ ];
```

Figure 29.10: Mortgage pool cash flow under prepayments. SMM is the prepayment vector, and the mortgage rate r is a monthly rate. The pool has n monthly cash flows, and its principal balance is \$1. B stores the remaining principals, P are the principal payments (prepayments included), $\overline{I}$ are the interest payments, and PP are the prepayments. The prepayments are calculated based on Exercise 29.1.9, part (1).

for an MBS represents both the guarantee fee and the actual servicing fee itself. For example, a Ginnie Mae MBS with a 10.5% pass-through rate has a total servicing of 0.50%, of which 0.44% is retained by the servicer and 0.06% is remitted to Ginnie Mae. The figure most visible to the investor is the pass-through rate, but the amortization of P&I is a function of the gross mortgage rate of the individual loans making up the pool.

➤ **Exercise 29.1.8** Show that the scheduled monthly mortgage payment at month i is

$$B_{i-1}\frac{(r/m)(1+r/m)^{n-i+1}}{(1+r/m)^{n-i+1}-1}.$$

➤ **Exercise 29.1.9** Verify that (1) $\mathrm{PP}_i = B_i[\,\mathrm{SMM}_i/(1-\mathrm{SMM}_i)\,]$ and (2) the actual principal payment $\overline{P_i} + \mathrm{PP}_i$ is $b_{i-1}(P_i + \mathrm{RB}_i \times \mathrm{SMM}_i)$ (not $b_i P_i$).

➤ **Exercise 29.1.10** Verify Eqs. (29.9) and (29.10).

➤ **Exercise 29.1.11** Derive Eq. (29.11) by using Eqs. (29.2) and (29.4).

➤ **Exercise 29.1.12** Derive the PVs of the PO and IO strips based on current-coupon mortgages under constant SMM and zero servicing spread.

➤ **Exercise 29.1.13** Show that a pass-through backed by traditional mortgages with a mortgage rate equal to the market yield is priced at par regardless of prepayments. Assume either zero servicing spread or a pass-through rate equal to the market yield. (Prices of par-priced pass-throughs are hence little affected by variations in the prepayment speed.)

➢ **Programming Assignment 29.1.14** Implement the algorithm in Fig. 29.10.

29.1.4 Pricing Sequential-Pay CMOs

Consider a three-tranche sequential-pay CMO backed by $3,000,000 of mortgages with a 12% coupon and 6 months to maturity. The three tranches are called A, B, and Z. All three tranches carry the same coupon rate of 12%. The Z tranche consists of **Z bonds**. A Z bond receives no payments until all previous tranches are retired. Although a Z bond carries an explicit coupon rate, the owed interest is accrued and added to the principal balance of that tranche. For that reason, Z bonds are also called **accrual bonds** or **accretion bonds**. When a Z bond starts receiving cash payments, it becomes a pass-through instrument.

Assume that the ensuing monthly interest rates are 1%, 0.9%, 1.1%, 1.2%, 1.1%, and 1.0%. Assume further that the SMMs are 5%, 6%, 5%, 4%, 5%, and 6%. We want to calculate the cash flow and the fair price of each tranche.

We can compute the pool's cash flow by invoking the algorithm in Fig. 29.10 with $n = 6$, $r = 0.01$, and $\text{SMM} = [\,0.05, 0.06, 0.05, 0.04, 0.05, 0.06\,]$. We can derive individual tranches' cash flows and remaining principals thereof by allocating the pool's P&I cash flows based on the CMO structure. See Fig. 29.11 for the breakdown. Note that the Z tranche's principal is growing at 1% per month until all previous tranches are retired. Before that time, the interest due the Z tranche is used to retire A's and B's principals. For example, the $10,000 interest due tranche Z at month one is directed to tranche A instead, reducing A's remaining principal from $386,737 to $376,737 while increasing Z's from $1,000,000 to $1,010,000. At month four, the interest amount that goes into tranche Z, $10,303, is exactly what is required of Z's remaining principal of $1,030,301. The tranches can be priced

Month		*1*	*2*	*3*	*4*	*5*	*6*
Interest rate		1.0%	0.9%	1.1%	1.2%	1.1%	1.0%
SMM		5.0%	6.0%	5.0%	4.0%	5.0%	6.0%
Remaining principal (B_i)							
	3,000,000	2,386,737	1,803,711	1,291,516	830,675	396,533	0
A	1,000,000	376,737	0	0	0	0	0
B	1,000,000	1,000,000	783,611	261,215	0	0	0
Z	1,000,000	1,010,000	1,020,100	1,030,301	830,675	396,533	0
Interest ($\overline{I_i}$)		30,000	23,867	18,037	12,915	8,307	3,965
A		20,000	3,767	0	0	0	0
B		10,000	20,100	18,037	2,612	0	0
Z		0	0	0	10,303	8,307	3,965
Principal		613,263	583,026	512,195	460,841	434,142	396,534
A		613,263	376,737	0	0	0	0
B		0	206,289	512,195	261,215	0	0
Z		0	0	0	199,626	434,142	396,534

Figure 29.11: CMO cash flows. Month-i numbers reflect the ith monthly payment. "Interest" and "Principal" denote the pool's P&I and distributions to individual tranches. Interest payments may be used to make principal payments to tranches A, B, and C. The Z bond thus protects earlier tranches from extension risk.

as follows:

$$\text{tranche A} = \frac{20000+613263}{1.01} + \frac{3767+376737}{1.01\times 1.009} = 1000369,$$

$$\text{tranche B} = \frac{10000+0}{1.01} + \frac{20100+206289}{1.01\times 1.009} + \frac{18037+512195}{1.01\times 1.009\times 1.011} + \frac{2612+261215}{1.01\times 1.009\times 1.011\times 1.012} = 999719,$$

$$\text{tranche Z} = \frac{10303+199626}{1.01\times 1.009\times 1.011\times 1.012} + \frac{8307+434142}{1.01\times 1.009\times 1.011\times 1.012\times 1.011} + \frac{3965+396534}{1.01\times 1.009\times 1.011\times 1.012\times 1.011\times 1.01} = 997238.$$

This CMO has a total theoretical value of \$2,997,326, slightly less than its par value of \$3,000,000. See the algorithm in Fig. 29.12.

We have seen that once the interest rate path and the prepayment vector for that interest rate path are available, a CMO's cash flow can be calculated and the CMO priced. Unfortunately, the remaining principal of a CMO under prepayments is, like an ARM, path dependent. For example, a period of high rates before dropping to the current level is not likely to result in the same remaining principal as a period of low rates before rising to the current level. This means that if we try to price a 30-year CMO on a binomial interest rate model, there will be $2^{360} \approx 2.35 \times 10^{108}$ paths to consider! As a result, Monte Carlo simulation is the computational method of choice. It works as follows. First, one interest rate path is generated. Based on that path, the prepayment model is applied to generate the pool's principal, prepayment, and interest cash flows. Now the cash flows of individual tranches can be generated and their PVs derived. The above procedure is repeated over many interest rate scenarios. Finally, the average of the PVs is taken.

➤ **Exercise 29.1.15** Calculate the monthly prepayment amounts for Fig. 29.11.

➢ **Programming Assignment 29.1.16** Implement the algorithm in Fig. 29.12 for the cash flows of a four-tranche sequential CMO with a Z tranche. Assume that each tranche carries the same coupon rate as the underlying pool's mortgage rate. Figures 29.13 and 29.14 plot the cash flows and remaining principal balances of one such CMO.

29.1.5 Weighted Average Life

The **weighted average life** (**WAL**) of an MBS is the average number of years that each dollar of unpaid *principal* due on the mortgages remains outstanding. It is computed by

$$\text{WAL} \equiv \frac{\sum_{i=1}^{m} i\,P_i}{12\times P},$$

where m is the remaining term to maturity in months, P_i is the principal repayment i months from now, and P is the current remaining principal balance.[2] See Fig. 29.15 for an illustration. Usually, the greater the anticipated prepayment rate, the shorter

Sequential CMO cash flow generator:

```
input:   n, r (r > 0), SMM[1..n], O[1..4];
real     B[n+1], P[1..n], Ī[1..n]; // Pool cash flows.
real     B[1..4][n+1], P[1..4][1..n], Ī[1..4][1..n];
real     P, I;
integer  i, j;
Call the algorithm in Fig. 29.10 for B[n+1], P[1..n], Ī[1..n];
for (j = 1 to 4) { B[j][0] := O[j]; } // Original balances.
for (i = 1 to n) { // Month i.
        P := P[i]; I := Ī[i]; // Pool P&I for month i.
        for (j = 1 to 3)   { // Tranches A, B, C.
            Ī[j][i] := B[j][i-1] × r;   // Interest due tranche j.
            I := I - Ī[j][i];
            if [ B[j][i-1] ≤ P ]   { // Retire it.
                P := P - B[j][i-1]; P[j][i] := B[j][i-1];
                B[j][i] := 0;
            } else {
                B[j][i] := B[j][i-1] - P; P[j][i] := P; P := 0;
            }
        }
        for (j = 1 to 3)  { // Interest as prepayment for A, B, C.
            if [ B[j][i] ≤ I ]  { // Retire it.
                Ī[j][i] := Ī[j][i] + B[j][i]; I := I - B[j][i];
                B[j][i] := 0;
            } else {
                B[j][i] := B[j][i] - I; Ī[j][i] := Ī[j][i] + I;
                I := 0;
            }
        }
        // Tranche Z.
        Ī[4][i] := I; P[4][i] := P;
        B[4][i] := B[4][i-1] × (1+r) - P - I;
}
return B[ ][ ], P[ ][ ] Ī[ ][ ];
```

Figure 29.12: Sequential CMO cash flow generator. SMM is the prepayment vector, and the mortgage rate r is a monthly rate. The pool has n monthly cash flows, and its principal balance is assumed to be \$1. B stores the remaining principals, P are the principal payments (prepayments included), and $\bar{I}$ are the interest payments. Tranche 1 is the A tranche, tranche 2 is the B tranche, and so on. $\mathcal{O}$ stores the original balances of individual tranches as fractions of \$1.

the average life. Given a static prepayment vector, the WAL increases with coupon rates because a larger proportion of the payment in early years is then interest, delaying the repayment of principal. The implied PSA is sometimes defined as the single PSA speed that gives the same WAL as the static prepayment vector.

29.2 Collateral Prepayment Modeling

The interest rate level is the most important factor in influencing prepayment speeds. The MBS typically experiences accelerating prepayments after a lag when the prevailing mortgage rate becomes 200 basis points below the WAC. This event is known

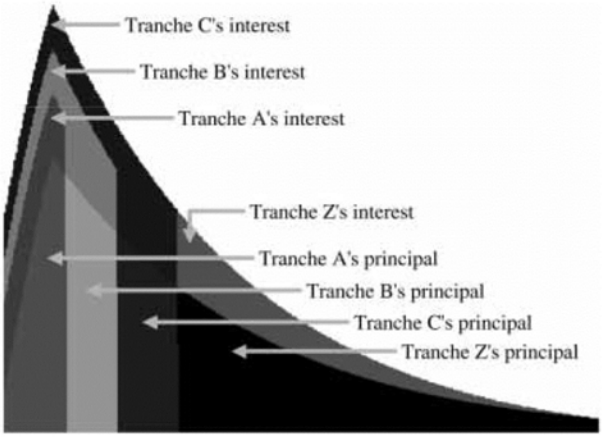


Figure 29.13: Cash flows of a four-tranche sequential CMO. The mortgage rate is 6%, the actual prepayment speed is 150 PSA, and each tranche has an identical original principal amount.

as the **threshold for refinancing**. The prepayment speed accelerates rapidly and then tends to "**burn out**" and settle at a lower speed. The subsequent times when rates fall through the refinancing threshold will not produce the same response. Over time, the pool is left mostly with mortgagors who do not refinance under any circumstances, and the pool's interest rate sensitivity falls. Next to refinancing incentive, loan size is also critical as the monetary savings are proportional to it [4].

The age of a pool has a general impact on prepayments. Refinancing rates are generally lower for new loans than seasoned ones. Interest rate changes and other human factors have little impact on prepayment speeds for the early years of the pool's life. Afterwards, the pool begins to experience such factors that can lead to higher prepayment speeds, such as the sale of the house. This increase in prepayment speeds will stabilize to a steady state with age. We must add that given sufficient refinancing incentives, prepayment speeds can rise sharply even for new loans.

Refinancing is not the only reason prepayments accelerate when interest rates decline. Lower interest rates make housing more affordable and may trigger the trade-up to a bigger house. However, by and large, very high prepayment speeds are primarily due to refinancings, not housing turnover.

In prepayment modeling, the WAC instead of the pass-through rate is the governing factor. To start with, MBSs with identical pass-through rate may have different WACs, which almost surely result in different prepayment characteristics. The

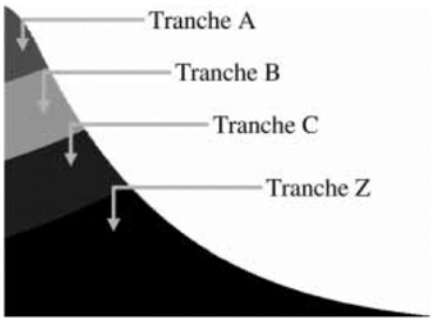


Figure 29.14: Remaining principal balances of a four-tranche sequential CMO. The CMO structure is identical to the one in Fig. 29.13. Tranche Zs principal balance grows until it becomes the current-pay tranche.

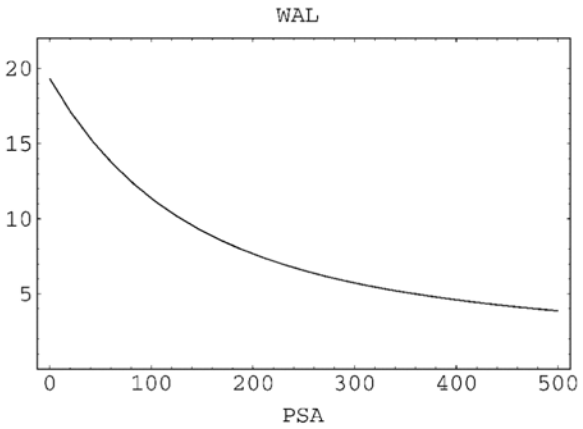


Figure 29.15: WAL under various PSAs. The underlying mortgages have 30 years to maturity and a 6% coupon rate.

WAC may also change over time because, absent prepayments, mortgages with lower coupons amortize faster than those with higher coupons (see Exercise 29.1.2). This makes the WAC increase over time. With prepayments, however, mortgages with higher coupons prepay faster, making the pool's WAC decline over time.

Each mortgage type (government-insured, conventional, and so forth) has a different prepayment behavior. For example, Freddie Mac and Fannie Mae pass-throughs seem to take longer to season than Ginnie Maes, and prepayment rates for 15-year mortgage pass-throughs usually exceed those of comparable-coupon 30-year pass-throughs [54, 325].

From the analysis above, a prepayment model needs at least the following factors: current and past interest rates, state of the economy (especially the housing market), WAC, current coupon rate, loan age, loan size, agency and pool type, month of the year, and burnout. Although we have discussed prepayment speeds at the pool level, a model may go into individual loans to generate the pool's cash flow if such information is available and the benefits outweigh the costs [4, 259]. A long-term average of the projected speeds is typically reported as the model's projected prepayment vector. This projection can be a weighted average of the projected speeds, the single speed that gives the same weighted average life as the vector, or the single speed that gives the same yield as the vector [433].

A PO is purchased at a discount. Because its cash flow is returned at par, a PO's dollar return is simply the difference between the par value and the purchase price. The faster that dollar return is realized, the higher the yield. Prepayments are therefore beneficial to POs. In declining mortgage rates, not only do prepayments accelerate, the cash flow is also discounted at a lower rate; consequently, POs appreciate in value. The opposite happens when mortgage rates rise (see Fig. 29.16). In summary, POs have positive duration and do well in bull markets.

An IO, in contrast, has no par value. Any prepayments reduce the pool principal and thus the interest as well. When mortgage rates decline and prepayments accelerate, an IO's price usually declines even though the cash flow will be discounted at a lower rate. If mortgage rates rise, the cash flow improves. However, beyond a certain point, the price of an IO will decline because of higher discount rates (see Fig. 29.17). An IO's price therefore moves in the same direction as the change in mortgage rates

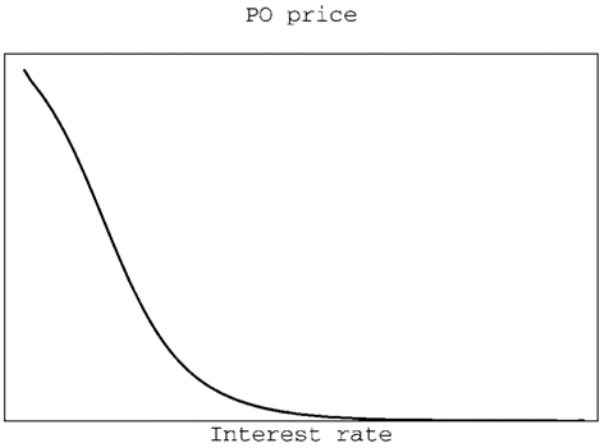


Figure 29.16: Price of PO.

over certain ranges (negative duration, in other words). Unlike most fixed-income securities, IOs do best in bear markets.

SMBSs are extremely sensitive to changes in prepayment speeds (see Exercise 29.2.2). These securities are often combined with other types of securities to alter the return characteristics. For example, because the PO thrives on the acceleration of prepayment speeds, it serves as an excellent hedge against MBSs whose price flattens or declines if prepayments accelerate, whereas IOs can hedge the interest rate risk of securities with positive duration.

➤ **Exercise 29.2.1** Divide the borrowers into slow and fast refinancers. (More refined classification is possible.) The slow refinancers are assumed to respond to refinancing incentive at a higher rate than fast refinancers. Describe how this setup models burnout.

➤ **Exercise 29.2.2** From Exercise 29.1.12, show that the prices of PO and IO strips are extremely sensitive to prepayment speeds.

Figure 29.17: Price of IO. IOs and POs do not have symmetric exposures to rate changes.

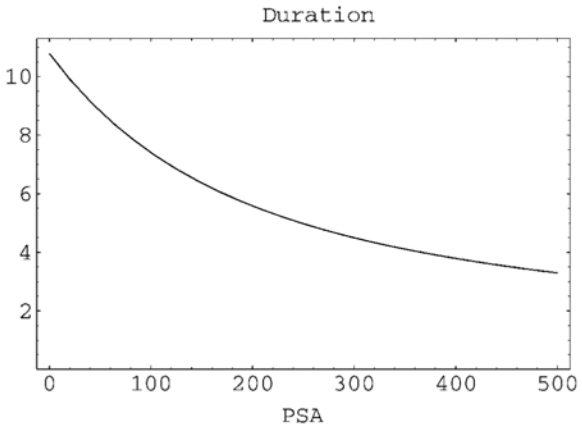


Figure 29.18: MD under various PSAs. The coupon rate and the market yield are assumed to be 6%. The underlying mortgages have 30 years to maturity.

➤ **Exercise 29.2.3** Firms that derive income from servicing mortgages can be viewed as taking a long position in IOs. Why?

29.3 Duration and Convexity

Duration is more important for the evaluation of pass-throughs than the WAL, which measures the time to the receipt of the principal cash flows [247, 619]. Figure 29.18 illustrates the Macaulay duration (MD) of a pass-through under various prepayment assumptions. The MD derived under a static prepayment vector, which does not change as yields change, is also called **static duration** or **cash flow duration**.

Duration is supposed to reveal how a change in yields affects the price, that is,

$$\text{percentage price change} \approx -\text{effective duration} \times \text{yield change}. \tag{29.12}$$

Relation (29.12) has obvious applications in hedging. However, static duration is inadequate for that purpose because the cash flow of an MBS depends on the prevailing yield. The most relevant measure of price volatility is the effective duration,

$$\frac{\partial P}{\partial y} \approx \frac{P_- - P_+}{P_0(y_+ - y_-)},$$

where P_0 is the current price, P_- is the price if yield is decreased by Δy, P_+ is the price if yield is increased by Δy, y is the initial yield, $y_+ \equiv y + \Delta y$, and $y_- \equiv y - \Delta y$. Figure 29.19 plots the effective duration of an MBS. For example, it says that the effective duration is approximately six at 9%; a 1% change in yields will thus move the price by roughly 6%. The prices P_+ and P_- are often themselves expected values calculated by simulation. To save computation time, either $(P_- - P_0)/(P_0 \Delta t)$ or $(P_0 - P_+)/(P_0 \Delta t)$ may be used instead, as only one of P_- and P_+ needs to be calculated then.

Similarly, convexity $\partial^2 P/\partial y^2$ can be approximated by the effective convexity:

$$\frac{P_+ + P_- - 2 \times P_0}{P_0[\,0.5 \times (y_+ - y_-)\,]^2}.$$

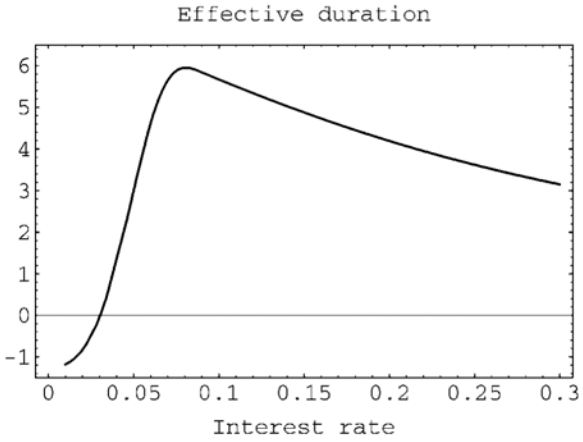


Figure 29.19: Effective duration. The MBS is from Fig. 28.8.

See Fig. 29.20 for an illustration. Convexity can improve first-order formula (29.12) by adding second-order terms,

$$\text{percentage price change} \approx -\text{effective duration} \times \text{yield change} + 0.5 \times \text{convexity} \times (\text{yield change})^2.$$

We saw in Fig. 28.8 that an MBS's price increases at a decreasing rate as the yield falls below the cusp because of accelerating prepayments, at which point it starts to decrease. This negative convexity is evident in Fig. 29.20. Therefore, even if the MD, which is always positive, is acceptable for current-coupon and moderately discount MBSs, it will not work for premium-priced MBSs.

➤ **Exercise 29.3.1** Suppose that MBSs are priced based on the premise that there are no prepayments until the 12th year, at which time the pool is repaid completely. This is called the **FHA 12-year prepaid-life concept**. Argue that premium-priced MBSs are overvalued and discount MBSs are undervalued if prepayments occur before the 12th year. (Studies have shown that the average life is much shorter than 12 years [577].)

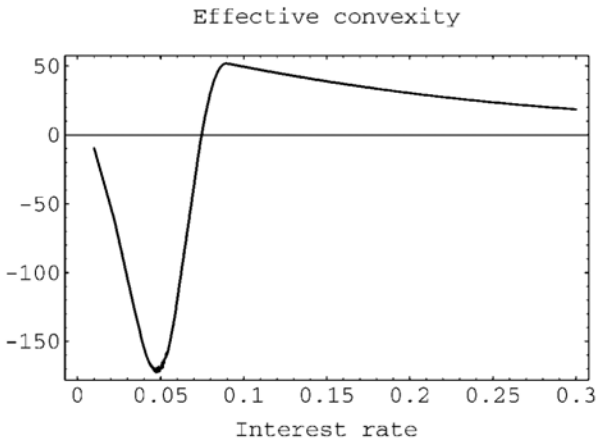


Figure 29.20: Effective convexity. The MBS is from Fig. 28.8.

➤ **Exercise 29.3.2** Modified duration $(1/P)\sum_{i=1}^{n} iC_i(1+y)^{-(i+1)}$ cannot be negative for pass-throughs. On the other hand, effective duration, which approximates modified duration, can be negative, as shown in Fig. 29.19. Why?

➤ **Exercise 29.3.3** A hedger takes a long position in MBSs and hedges it by shorting T-bonds. Assess this strategy.

➤ **Exercise 29.3.4** Consider options on mortgage pass-through forwards. Argue that Black's model tends to overstate the call value and to underestimate the put value.

29.4 Valuation Methodologies

Mortgage valuation involves modeling the uncertain cash flow and computing its PV. As in Section 27.4, the three basic approaches to valuing MBSs are static cash flow yield, option modeling, and OAS. Because their valuation is more technical and relies more on judgment than do other fixed-income securities, not to mention such issues as prepayment risk, credit quality, and liquidity, MBSs are priced to a considerable yield spread over the Treasuries and corporate bonds.

29.4.1 The Static Cash Flow Yield Methodology

When an internal rate of return is calculated with the static prepayment assumption over the life of the security, the result is the (static) cash flow yield, we recall. The static cash flow yield methodology compares the cash flow yield on an MBS with that on "comparable" bonds. For this purpose, it is inappropriate to use the stated maturity of the MBS because of prepayments. Instead, either the MD or the WAL under the same prepayment assumption can be used.

Although simple to use, this methodology sheds little light on the relative value of an MBS. Its problems, besides being static, are that (1) the projected cash flow may not be reinvested at the cash flow yield, (2) the MBS may not be held until the final payout date, and (3) the actual prepayment behavior is likely to deviate from the assumptions.

The static spread methodology goes beyond the cash flow yield by incorporating the Treasury yield curve. The static spread to the Treasuries is the spread that makes the PV of the projected cash flow from the MBS when discounted at the spot rate plus the spread equal its market price (review Section 5.4).

29.4.2 The Option Pricing Methodology

Virtually all mortgage loans give the homeowner the right to prepay the mortgage at any time. The homeowner in effect holds an option to call the mortgage. The totality of these rights to prepay constitutes the embedded call option of the pass-through. Because the homeowner has the right to call a pro rata portion of the pool, the MBS investor is short the embedded call; therefore,

$$\text{pass-through price} = \text{noncallable pass-through price} - \text{call option price}.$$

The option pricing methodology prices the call option by an option pricing model. It then estimates the market price of the noncallable pass-through by

$$\text{noncallable pass-through price} = \text{pass-through price} + \text{call option price}.$$

The preceding price is finally used to compute the yield on this theoretical bond that does not prepay. This yield is called the option-adjusted yield.

The option pricing methodology was criticized in Subsection 27.4.2. It has additional difficulties here. Prepayment options are often "irrationally" exercised. Furthermore, a partial exercise is possible as the homeowner can prepay a portion of the loan; there is not one option but many, one per homeowner. Finally, valuation of the call option becomes very complicated for CMO bonds.

29.4.3 The Option-Adjusted-Spread Methodology

The OAS methodology has four major parts [382]. The interest rate model is the first component. Then there is the prepayment model, which is the single most important component. Although the prepayment model may be deterministic or stochastic, there is evidence showing that deterministic models that are accurate on average are good enough for pass-throughs, IOs, and POs [428, 433]. The **cash flow generator** is the third component. It calculates the current coupon rates for the interest rate paths given by the interest rate model. It then generates the P&I cash flows for the pool as well as allocating them for individual securities based on the prepayment model and security information such as CMO rules. Note that the same pool cash flow drives many securities. Finally, the equation solver calculates the OAS. Because several paths of interest rates are used, many statistics are often computed as well. See Fig. 29.21 for the overall structure.

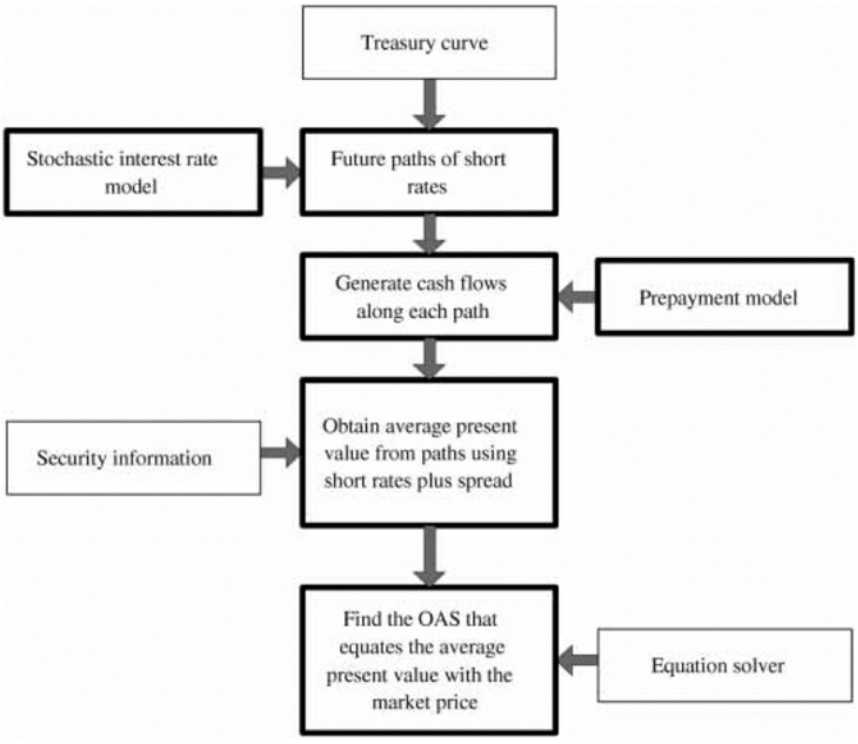


Figure 29.21: OAS computation framework for MBSs. Components boxed by thinner borders are supplied externally.

The general valuation formula for uncertain cash flows can be written as

$$\text{PV} = \lim_{N\to\infty} \frac{1}{N} \sum_{N \text{ paths } r^*} \frac{C_n^*}{(1+r_1^*)(1+r_2^*)\cdots(1+r_n^*)}, \tag{29.13}$$

where r^* denotes a risk-neutral interest rate path for which r_i^* is the ith one-period rate and C_n^* is the cash flow at time n under this scenario. The summation averages over a large number of scenarios whose distribution matches the interest rate dynamics. The average over scenarios must also match the current spot rate curve, i.e.,

$$\frac{1}{(1+f_1)(1+f_2)\cdots(1+f_n)} = \lim_{N\to\infty} \frac{1}{N} \sum_{N \text{ paths } r^*} \frac{1}{(1+r_1^*)(1+r_2^*)\cdots(1+r_n^*)},$$

$n = 1, 2, \ldots$, where f_i are the implied forward rates.

The Monte Carlo valuation of MBSs is closely related to Eq. (29.13). The interest rate model randomly produces a set of risk-neutral rate paths. The cash flow is then generated for each path. Finally, we solve for the spread s that makes the average discounted cash flow equal the market price:

$$P = \lim_{N\to\infty} \frac{1}{N} \sum_{N \text{ paths } r^*} \frac{C_n^*}{(1+r_1^*+s)(1+r_2^*+s)\cdots(1+r_n^*+s)}.$$

This spread s is the OAS. The implied cost of the embedded option is then calculated as

$$\text{option cost} = \text{static spread} - \text{OAS}.$$

A common alternative averages the cash flows first and then calculates the OAS as the spread that equates this average cash flow with the market price. Although this approach is more efficient, it will generally give a different spread.

OAS calculation is very time consuming. The majority of the cost lies in generating the cash flows. This is because CMOs can become arbitrarily complex in their rules for allocating the cash flows. Such complexity requires special **data structures** in software design. The computational costs are then multiplied by the many runs of the Monte Carlo simulation.

OAS can be seen to measure the risk premium for bearing systematic risks in the mortgage market. Under this interpretation, the OAS methodology identifies investments with the best potential for excess returns. Being statistically derived, the prepayment model will always be out of date and provide only a crude forecast for future conditions. Therefore an alternative interpretation is that no such risk premium exists: A nonzero OAS simply implies that the market is trading off a different set of prepayment assumptions [34]. This view suggests that one investigate the implied prepayment assumptions [188].

➤ **Exercise 29.4.1** Argue that the OAS with zero interest rate volatility, called the **zero-volatility OAS**, corresponds to the static spread.

➢ **Programming Assignment 29.4.2** Implement the OAS computation for the four-tranche sequential CMO under the BDT model. Assume a constant SMM.

Duration and Convexity

Effective duration and convexity can be computed if the OAS is held constant. The results are called the OAS duration and the **OAS convexity**, respectively [323, 325]. Key rate durations, introduced in Section 27.5 and calculated like the OAS duration, are most useful in identifying the segments of the yield curve that most affect the MBS value [260]. Note that the OAS duration is at least twice as expensive as the OAS in terms of computation time because at least one of P_+ and P_- has to be computed by simulation. The OAS convexity is three times as expensive because both P_+ and P_- have to be computed.

Prepayment risk can represent the risk that the market price reflects prepayment assumptions that are different from the model. An interesting measure of prepayment risk is the **prepayment duration**. It is the percentage change in price, with the OAS held constant, for a given percentage deviation in speeds from some base level projection (see Exercise 29.2.2) [198, 328, 433, 815].

Holding Period Returns

The HPR assesses the MBS over a holding period. The FV at the horizon consists of the projected P&I cash flows, the interest on the reinvestment thereof, and the projected horizon price. The monthly total return is

$$\left(\frac{\text{total future amount}}{\text{price of the MBS}}\right)^{1/\text{number of months}} - 1.$$

To calculate the preceding return, prepayment assumptions, reinvestment rates, and interest rate dynamics are all needed. These assumptions are not independent.

The OAS can be combined with the HPR analysis. First we create a few static interest rate and prepayment scenarios for the holding period. The prepayment assumptions are in the form of prepayment vectors. We then calculate the HPR for each scenario by assuming that the OAS remains unchanged at the horizon.

Additional Reading

See [54, 55, 260, 330, 829] for more information on MBSs, [54, 55, 124, 259, 260, 276, 297, 323, 325, 330, 595, 619, 649, 788, 789, 818, 896] for the valuation of MBSs, [134, 188, 197, 198, 438, 454] for OAS analysis, and [142, 715] for the Monte Carlo valuation of MBSs. Monte Carlo simulation typically provides an unbiased estimate [478]. Application of the variance-reduction techniques and quasi-Monte Carlo methods in Chap. 18 can result in less work [197, 354]. Parallel processing for much faster performance has been convincingly demonstrated [601, 794, 892, 893]. Additional information on duration measures can be found in [33, 258, 272, 394, 429, 504, 889]. Many yield concepts are discussed in [406]. See [118, 220, 268, 361, 411, 430, 431, 433, 540] for prepayment models, [260] for a historical account, and [296] for early models. Factors used in prepayment modeling are considered in [54, 330, 430, 433]. The FHA 12-year prepaid-life concept is discussed in [54, 363]. Valuation of MBSs may profit from two-factor models because prepayments tend to depend more on the long-term rate [456]. See [316] for the prepayments of multifamily MBSs and [203, 742, 864] for the empirical analysis of prepayments. Burnout modeling is discussed in [199]. The refinancing waves of 1991–1993 cast some doubts on the burnout concept, however

[433]. Consult [524] for hedging MBSs and [460] for options on MBSs. The 11th District COFI is analyzed in [347, 684], and the CMT rates are compared with the on-the-run yields in [525].

NOTES

1. Also known as the **weighted average remaining maturity** (**WARM**).
2. **Payment delays** should be incorporated in the WAL calculation: 14 (actual) or 45 days (stated) for GNMA Is, 19 (actual) or 50 days (stated) for GNMA IIs, 24 (actual) or 55 (stated) for Fannie Mae MBSs, 44 (actual) or 75 (stated) for Freddie Mac non-Gold PCs, and 14 (actual) or 45 (stated) for Freddie Mac Gold PCs. The **stated payment delay** denotes the number of days between the first day of the month and the date the servicer actually remits the P&I to the investor [54, 330].

CHAPTER THIRTY

Collateralized Mortgage Obligations

Capital can be understood only as motion, not as a thing at rest.
Karl Marx (1818–1883), *Das Kapital*

Mutual funds combine diverse financial assets into a portfolio and issue a single class of securities against it. CMOs reverse that process by issuing a diverse set of securities against a relatively homogeneous portfolio of assets [660]. This chapter surveys CMOs. The tax treatment of CMOs is generally covered under the provisions of the Real Estate Mortgage Investment Conduit (REMIC) rules of 1986. As a result, CMOs are often referred to as **REMIC**s [162, 469].

30.1 Introduction

The complexity of a CMO arises from layering different types of payment rules on a prioritized basis. In the first-generation CMOs, the sequential-pay CMOs, each class of bond would be retired sequentially. A sequential-pay CMO with a large number of tranches will have very narrow cash flow windows for the tranches. To further reduce prepayment risk, tranches with a principal repayment schedule were introduced. They are called **scheduled bonds**. For example, bonds that guarantee the repayment schedule when the actual prepayment speed lies within a specified range are known as **planned amortization class** bonds (**PAC**s). PACs were introduced in August 1986 [141]. Whereas PACs offer protection against both contraction and extension risks, some investors may desire protection from only one of these risks. For them, a bond class known as the **targeted amortization class** (**TAC**) was created.

Scheduled bonds expose certain CMO classes to less prepayment risk. However, this can occur only if the redirection in the prepayment risk is absorbed as much as possible by other classes referred to as the **support bonds** or **companion bonds**. Support bonds are a necessary by-product of the creation of scheduled tranches.

Pro rata bonds provide another means of layering. Principal cash flows to these bonds are divided proportionally, but the bonds can have different interest payment rules. Suppose the WAC of the collateral is 10%, tranche B1 receives 40% of the principal, and tranche B2 receives 60% of the principal. Given this pro rata structure, many choices of interest payment rules are possible for B1 and B2 as long as the interest payments are nonnegative and the WAC does not exceed 10%. The coupon rates can even be floating. One possibility is for B1 to have a coupon of 5% and B2 to have a coupon of 13.33%. Bonds with pass-through coupons that are higher

and lower than the collateral coupon have thus been created. Bonds like B1 are called **synthetic discount securities** and bonds like B2 are called **synthetic premium securities**. An extreme case is for B1 to receive 99% of the principal and have a 5% coupon and B2 to receive only 1% of the principal and have a 505% coupon. In fact, first-generation IOs took the form of B2 in July 1986 [55].

IOs have either a **nominal principal** or a notional principal. A nominal principal represents actual principal that will be paid. It is called "nominal" because it is extremely small, resulting in an extremely high coupon rate. A case in point is the B2 class with a 505% coupon above. A notional principal, in contrast, is the amount on which interest is calculated. An IO holder owns none of the notional principal. Once the notional principal amount declines to zero, no further payments are made on the IO.

30.2 Floating-Rate Tranches

A form of pro rata bonds are floaters and inverse floaters whose combined coupon does not exceed the collateral coupon. A floater is a class whose coupon rate varies directly with the change in the reference rate, and an inverse floater is a class whose coupon rate changes in the direction opposite to the change in the reference rate. When the coupon on the inverse floater changes by x times the amount of the change in the reference rate, this multiple x is called its **slope**. Because the interest comes from fixed-rate mortgages, floaters must have a coupon cap. Similarly, inverse floaters must have a coupon floor. Floating-rate classes were created in September 1986.

Suppose the floater has a principal of P_f and the inverse floater has a principal of P_i. Define $\omega_\mathrm{f} \equiv P_\mathrm{f}/(P_\mathrm{f}+P_\mathrm{i})$ and $\omega_\mathrm{i} \equiv P_\mathrm{i}/(P_\mathrm{f}+P_\mathrm{i})$. To make the structure self-supporting, the coupon rates of the floater, c_f, and the inverse floater, c_i, must satisfy $\omega_\mathrm{f} \times c_\mathrm{f} + \omega_\mathrm{i} \times c_\mathrm{i} = \mathrm{WAC}$, or

$$c_\mathrm{i} = \frac{\mathrm{WAC} - \omega_\mathrm{f} \times c_\mathrm{f}}{\omega_\mathrm{i}}.$$

The slope is clearly $\omega_\mathrm{f}/\omega_\mathrm{i}$. To make sure that the inverse floater will not encounter a negative coupon, the cap on the floater must be less than $\mathrm{WAC}/\omega_\mathrm{f}$. In fact, caps and floors are related by

$$\mathrm{floor} = \frac{\mathrm{WAC} - \omega_\mathrm{f} \times \mathrm{cap}}{\omega_\mathrm{i}}.$$

EXAMPLE 30.2.1 Consider a CMO deal that includes a floater with a principal of \$64 million and an inverse floater with a principal of \$16 million. The coupon rate for the floating-rate class is LIBOR+ 0.65 and that for the inverse floater is $42.4 - 4 \times \mathrm{LIBOR}$ The slope is thus four. The WAC of the two classes is

$$\frac{64}{80} \times \text{floater coupon rate} + \frac{16}{80} \times \text{inverse floater coupon rate} = 9\%,$$

regardless of the level of LIBOR. Consequently the coupon rate on the underlying collateral, 9%, can support the aggregate interest payments that must be made to these two classes. If we set a floor of 0% for the inverse floater, the cap on the floater is 11.25%.

A variant of the floating-rate CMO is the **superfloater** introduced in 1987. In a conventional floating-rate class, the coupon rate moves up or down on a one-to-one basis with the reference rate subject to caps and floors. A superfloater's coupon rate, in comparison, changes by some multiple of the change in the reference rate, thus magnifying any changes in the value of the reference rate. Superfloater tranches are bearish because their value generally appreciates with rising interest rates.

Suppose that the initial LIBOR is 7% and the coupon rate for a superfloater is based on this formula:

$$(\text{initial LIBOR} - 40 \text{ basis points}) + 2 \times (\text{change in LIBOR}).$$

The following table shows how the superfloater changes its coupon rate as LIBOR changes. The coupon rates for a conventional floater of LIBOR plus 50 basis points are also listed for comparison.

LIBOR Change (Basis Points)	*−300*	*−200*	*−100*	*0*	*+100*	*+200*	*+300*
Superfloater	0.6	2.6	4.6	6.6	8.6	10.6	12.6
Conventional floater	4.5	5.5	6.5	7.5	8.5	9.5	10.5

A superfloater provides a much higher yield than a conventional floater when interest rates rise and a much lower yield when interest rates fall or remain stable. We verify this by looking at the above table by means of spreads in basis points to LIBOR in the next table:

LIBOR Change (Basis Points)	*−300*	*−200*	*−100*	*0*	*+100*	*+200*	*+300*
Superfloater	−340	−240	−140	−40	60	160	260
Conventional floater	50	50	50	50	50	50	50

➤ **Exercise 30.2.1** Repeat the calculations in the text by using the following formula:

$$(\text{initial LIBOR} - 50 \text{ basis points}) + 1.5 \times (\text{change in LIBOR}).$$

➤ **Exercise 30.2.2** Argue that the maximum coupon rate that could be paid to a floater is higher than would be possible without the inclusion of an inverse floater.

30.3 PAC Bonds

PAC bonds may be the most important innovation in the CMO market [141]. They are created by calculation of the cash flows from the collateral by use of two prepayment speeds, a fast one and a slow one. Consider a **PAC band** of 100 PSA (the **lower collar**) to 300 PSA (the **upper collar**). Figure 30.1 shows the principal payments at the two collars. Note that the principal payments under the higher-speed scenario are higher in the earlier years but lower in later years. The shaded area represents the principal payment schedule that is guaranteed for every possible prepayment speed between 100% and 300% PSAs. It is calculated by taking the minimum of the principal paydowns at the lower collar and those at the upper collar. This schedule is called the **PAC schedule**. See Fig. 30.2 for a linear-time cash flow generator for a simple CMO containing a PAC bond and a support bond.

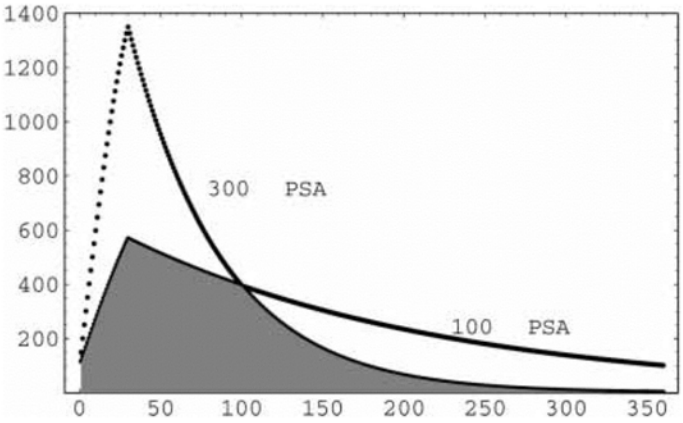


Figure 30.1: PAC schedule. The underlying mortgages are 30-year ones with a total original loan amount of $100,000,000 (the numbers on the *y* axis are in thousands) and a coupon rate of 6%. The PAC schedule is determined by the principal payments at 100 PSA and 300 PSA.

Adherence to the amortization schedule of the PAC takes priority over those of all other bonds. The cash flow of a PAC bond is therefore known as long as its support bonds are not fully paid off. Whether this happens depends to a large extent on the CMO structure, such as priority and the relative sizes of PAC and non-PAC classes. For example, a relatively small PAC is harder to break than a larger PAC, other things being equal.

If the actual prepayment speed is 150 PSA, the principal payment pattern of the PAC bond adheres to the PAC schedule. The cash flows of the support bond "flow around" the PAC bond (see Fig. 30.3). The cash flows are neither sequential nor pro rata; in fact, the support bond pays down *simultaneously* with the PAC bond. Because more than one class of bonds may be receiving principal payments at the same time, structures with PAC bonds are called **simultaneous-pay CMO**s. At the lower prepayment speed of 100 PSA, far less principal cash flow is available in the early years of the CMO. As all the principal cash flows go to the PAC bond in the early years, the principal payments on the support bond are deferred and the support bond extends. The support bond does, however, receive more interest payments.

If prepayments move outside the PAC band, the PAC schedule may not be met. At 400 PSA, for example, the cash flows to the support bond are accelerated. After the support bond is fully paid off, all remaining principal payments go to the PAC bond, shortening its life. See Fig. 30.4 for an illustration. The support bond thus absorbs part of the contraction risk. Similarly, should the actual prepayment speed fall below the lower collar, then in subsequent periods the PAC bond has priority on the principal payments. This reduces the extension risk, which is again absorbed by the support bond.

The PAC band guarantees that if prepayments occur at any single constant speed within the band *and* stay there, the PAC schedule will be met. However, the PAC schedule may not be met even if prepayments on the collateral always vary within the band over time. This is because the band that guarantees the original PAC schedule can expand and contract, depending on actual prepayments. This phenomenon is known as **PAC drift**.

PAC cash flow generator:

```
input:   n, r (r > 0), SMM[ 1..n ], PSA_u, PSA_ℓ, O[ 2 ];
real     P[ 1..n ], Ī[ 1..n ]; // Pool cash flows.
real     smm[ 1..n ], B[ 2 ][ n+1 ], P[ 2 ][ 1..n ], Ī[ 2 ][ 1..n ], P, I;
integer  i;
Call the algorithm in Fig. 29.10 with SMM[ 1..n ] for
         P[ 1..n ] and Ī[ 1..n ]; // Pool cash flows.
Call the algorithm in Fig. 29.9 for smm[ 1..n ] based on PSA_u;
Call the algorithm in Fig. 29.10 with smm[ 1..n ] and
         store the principal cash flow in P[ 0 ][ 1..n ];
Call the algorithm in Fig. 29.9 for smm[ 1..n ] based on PSA_ℓ;
Call the algorithm in Fig. 29.10 with smm[ 1..n ] and
         store the principal cash flow in P[ 1 ][ 1..n ];
for (i = 1 to n) {P[ 0 ][ i ] := min(P[ 0 ][ i ], P[ 1 ][ i ]); }
// PAC schedule per one dollar of original principal:
Normalize P[ 0 ][ 1..n ] so that the n elements sum to one;
B[ 0 ][ 0 ] := O[ 0 ]; B[ 1 ][ 0 ] := O[ 1 ]; // Original balances.
for (i = 1 to n) { // Month i.
         P := P[ i ]; I := Ī[ i ]; // Pool P&I for month i.
         P[ 1 ][ i ] := min(0, P − O[ 0 ] × P[ 0 ][ i ], B[ 1 ][ i − 1 ]);
         B[ 1 ][ i ] := B[ 1 ][ i − 1 ] − P[ 1 ][ i ];
         Ī[ 1 ][ i ] := B[ 1 ][ i − 1 ] × r; // Support bond done.
         P := P − P[ 1 ][ i ];
         P[ 0 ][ i ] := P;
         B[ 0 ][ i ] := B[ 0 ][ i − 1 ] − P;
         Ī[ 0 ][ i ] := I − Ī[ 1 ][ i ]; // PAC bond done.
}
return B[ ][ ], P[ ][ ], Ī[ ][ ];
```

Figure 30.2: PAC cash flow generator. SMM[ ] stores the actual prepayment speeds, PSA_u and PSA_ℓ form the PAC band, and the mortgage rate r is a monthly rate. The pool has n monthly cash flows, and its principal balance is assumed to be \$1. $\mathcal{O}$ stores the original balances of individual bonds as fractions of \$1; in particular, bond 0 is the PAC bond, bond 1 is the support bond, and $\mathcal{O}[0]+\mathcal{O}[1]=1$. B stores the remaining principals, P are the principal payments (prepayments included), and $\bar{I}$ are the interest payments. The CMO deal contains one PAC tranche and one support tranche.

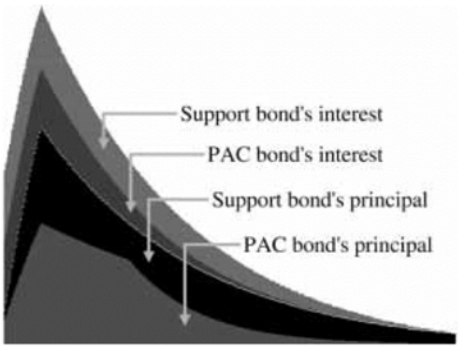


Figure 30.3: Cash flows of a PAC bond at 150 PSA. The mortgage rate is 6%, the PAC band is 100 PSA to 300 PSA, and the actual prepayment speed is 150 PSA.

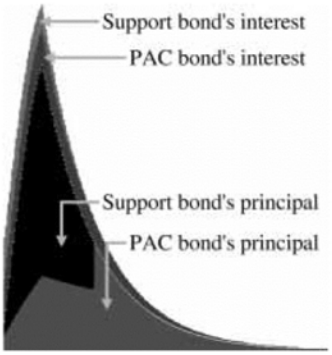


Figure 30.4: Cash flows of a PAC bond at 400 PSA. The mortgage rate is 6%, the PAC band is 100 PSA to 300 PSA, and the actual prepayment speed is 400 PSA.

PACs can be divided sequentially to provide narrower paydown structures. These **sequential PAC**s narrow the range of years over which principal payments occur. See Fig. 30.5 for an illustration. Although these bonds are all structured with the same band, the actual range of speeds over which their schedules will be met may differ. We can take a CMO bond and further structure it. For example, the sequential PACs could be split by use of a pro rata structure to create high and low coupon PACs. We can also replace the second tranche in a four-tranche ABCZ sequential CMO with a PAC class that amortizes starting in year four, say. But note that tranche C may start to receive prepayments that are in excess of the schedule of the PAC bond. It may even be retired earlier than tranche B.

Support bonds themselves can have cash flows prioritized so as to reduce prepayment risk. Support bonds with schedules, also referred to as **PAC II bonds**, are supported by other support bonds without schedules. PACs in a structure in which there are PAC II level bonds are called **PAC I bonds**.

➢ **Programming Assignment 30.3.1** Implement the cash flow generator in Fig. 30.2.

➢ **Programming Assignment 30.3.2** Implement the cash flow generator for sequential PAC bonds.

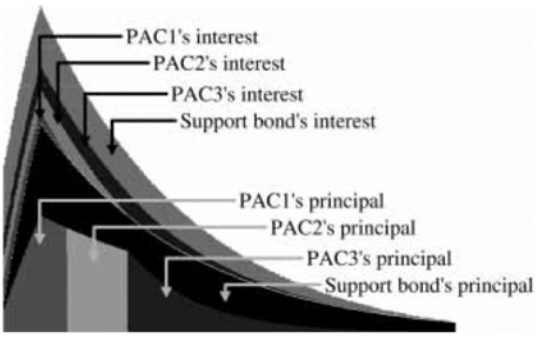


Figure 30.5: Cash flows of sequential PAC bonds. The mortgage rate is 6%, the PAC band is 100 PSA to 300 PSA, and the actual prepayment speed is 150 PSA. The three PAC bonds have identical original principal amounts.

30.4 TAC Bonds

Created in 1986, TAC bonds, just as PAC bonds, have priority over other bond classes that do not have a schedule for principal repayment. PACs have a higher priority over TACs, however. TAC bonds have a single PSA prepayment speed over which the principal repayment schedule is guaranteed. When prepayments exceed the speed, the excess principal is paid to the support bonds first. However, when prepayments fall short of the speed, TAC bonds will extend. TACs are therefore designed to provide protection against contraction risk but not extension risk.

30.5 CMO Strips

A class in a CMO structure can be a **CMO strip**. A CMO strip that is created when an IO is stripped from a CMO bond is called a **bond IO**. For example, this stripping mechanism creates an **inverse IO** from an inverse floater and a **PAC IO** from a PAC bond. Bond IOs lower the coupon of the CMO tranche. Some people call bond IOs **IOettes** to distinguish them from IO strips created off the entire collateral. A PO class that is neither a PAC nor a TAC is called a **super PO**. Like a PO strip, such bonds are purchased at a substantial discount from par and are returned at par. When prepayments accelerate as interest rates decline, "super" performance follows, hence the name.

30.6 Residuals

All CMOs contain a residual interest composed of the excess of collateral cash flows plus any reinvestment income over the payments for principal, interest, and expenses. This excess cash flow is called the **CMO residual**. The residual arises in part because credit rating agencies require CMOs to be overcollateralized in order to receive AAA credit rating: The cash flows must be sufficient to meet all the obligations under any prepayment scenario.

Another source of residual cash flow is reinvestment income. There is usually a delay between the time the payments from the collateral are received and the time they are remitted to the CMO bondholders. For example, whereas the mortgages in the collateral pay monthly, most CMOs pay quarterly or semiannually. The CMO trustee is therefore able to reinvest the pool cash flows before distribution dates. To be conservative in calculating the funds needed to meet future obligations, the rating agencies require that the trustee assume a relatively low reinvestment rate. CMO trustees have been able to reinvest at higher rates, and the excess is retained as a residual.

Additional Reading

See [14, 54, 161, 260, 325, 439, 758] for in-depth analyses of CMOs.

CHAPTER THIRTY-ONE

Modern Portfolio Theory

> Truly important and significant hypotheses will be found to have "assumptions" that are wildly inaccurate descriptive representations of reality.
>
> Milton Friedman, *"The Methodology of Positive Economics"*

This chapter starts with the **mean-variance theory** of portfolio selection. This theory provides a tractable framework for quantifying the risk–return trade-off of assets. We then investigate the equilibrium structure of asset prices. The result is the celebrated **Capital Asset Pricing Model** (**CAPM**, pronounced cap-m). The CAPM is the foundational quantitative model for measuring the risk of a security. Alternative asset pricing models based on factor analysis are also presented. The practically important concept of value at risk (VaR) for risk management concludes the chapter.

31.1 Mean–Variance Analysis of Risk and Return

Risk is the chance that expected returns will not be realized. We adopt standard deviation of the rate of return as the measure of risk.[1] This choice, although not without its critics, is standard in portfolio analysis and has nice statistical properties. Investors are presumed to prefer higher expected returns and lower variances.

Assume that there are n assets with random rates of return, $r_1, r_2, \ldots, r_n$. The expected values of these returns are $\bar{r}_i \equiv E[r_i]$. If we form a portfolio of these n assets by using (capitalization) weights $\omega_1, \omega_2, \ldots, \omega_n$, the portfolio's rate of return is

$$\boldsymbol{r} = \omega_1 r_1 + \omega_2 r_2 + \cdots + \omega_n r_n$$

with mean $\bar{\boldsymbol{r}} = \sum_{i=1}^{n} \omega_i \bar{r}_i$ and variance

$$\boldsymbol{\sigma}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j \sigma_{ij} = \sum_{i \neq j} \omega_i \omega_j \sigma_{ij} + \sum_{i=1}^{n} \omega_i^2 \sigma_i^2,$$

where σ_i^2 represents the variance of r_i and σ_{ij} represents the covariance between r_i and r_j. Note that $\sigma_{ii} = \sigma_i^2$.

The portfolio's total risk as measured by its variance consists of (1) $\sum_{i \neq j} \omega_i \omega_j \sigma_{ij}$, the **systematic risk** associated with the correlations between the returns on the assets in the portfolio, and (2) $\sum_{i=1}^{n} \omega_i^2 \sigma_i^2$, the **specific** or **unsystematic risk** associated

with the individual variances alone. Every possible weighting scheme $\omega_1, \omega_2, \ldots, \omega_n$ with $\sum_{i=1}^{n} \omega_i = 1$ corresponds to a portfolio, with negative weights meaning short sales. The constraints $\omega_i \geq 0$ can be added to exclude short sales. A portfolio $\boldsymbol{\omega} \equiv [\, \omega_1, \omega_2, \ldots, \omega_n \,]^{\mathrm{T}}$ that satisfies all the specified constraints is said to be a **feasible portfolio**.

Interestingly, if the returns[2] of the assets are uncorrelated, i.e., $\sigma_{ij} = 0$ for $i \neq j$, the variance of the portfolio's return decreases toward zero as n increases, provided that the portfolio is well diversified. For example, with $\omega_i = 1/n$,

$$\boldsymbol{\sigma}^2 = \sum_{i=1}^{n} \omega_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{n} \sigma_i^2}{n^2} \leq \frac{\sigma_{\max}^2}{n},$$

where $\sigma_{\max} \equiv \max_i \sigma_i$. This shows the power of diversification. Diversification, however, has its limits when asset returns are correlated. To see this point, assume that (1) all the returns have the same variance s^2, (2) the return correlation is a constant z, hence $\sigma_{ij} = zs^2$ for $i \neq j$, and (3) $\omega_i = 1/n$. The variance of $\boldsymbol{r}$ then is

$$\boldsymbol{\sigma}^2 = \sum_{i \neq j} \frac{zs^2}{n^2} + \sum_{i=1}^{n} \frac{s^2}{n^2} = n(n-1)\frac{zs^2}{n^2} + \frac{s^2}{n} = zs^2 + (1-z)\frac{s^2}{n},$$

which cannot be reduced below the average covariance zs^2.

These two examples demonstrate that specific risk and systematic risk behave very differently as the number of assets included in the portfolio grows. In general, as the portfolio gets larger and is well diversified, the specific risk tends to zero, whereas the systematic risk converges to the average of all the covariances for all pairs of assets in the portfolio. Markowitz called this phenomenon the **law of the average covariance** [644]. Systematic risk therefore does *not* disappear with diversification.

Consider a two-dimensional diagram with the horizontal axis denoting standard deviation and the vertical axis denoting mean. This is called the **mean–standard deviation diagram**. Every feasible portfolio with mean return rate $\bar{\boldsymbol{r}}$ and standard deviation $\boldsymbol{\sigma}$ can be represented as a point at $(\boldsymbol{\sigma}, \bar{\boldsymbol{r}})$ on the diagram; it is an **obtainable mean–standard deviation combination**. The set of feasible points form the **feasible set**. In general, the feasible set is a solid two-dimensional region and convex to the left. Thus the straight line segment connecting any two points in the set does not cross the left boundary of the set. For a given expected rate of return, the feasible point with the smallest variance is the corresponding left boundary point. The left boundary of the feasible set is hence called the **minimum-variance set**, and the point on this set having the minimum variance is the **minimum-variance point** (**MVP**). Most investors will choose the portfolio with the smallest variance for a given mean. Such investors are risk averse because they seek to minimize risk as measured by the standard deviation. Similarly, most investors will choose the portfolio with the highest mean for a given level of standard deviation (i.e., the highest point on a given vertical line). Therefore only the subset of the minimum-variance set above the MVP will be of interest. An obtainable mean–standard deviation combination is **efficient** if no other obtainable combinations have either higher mean and no higher variance or less variance and no less mean. The set of efficient combinations is termed the **efficient frontier**, and the corresponding portfolios are termed the **efficient portfolios**. See Fig. 31.1.

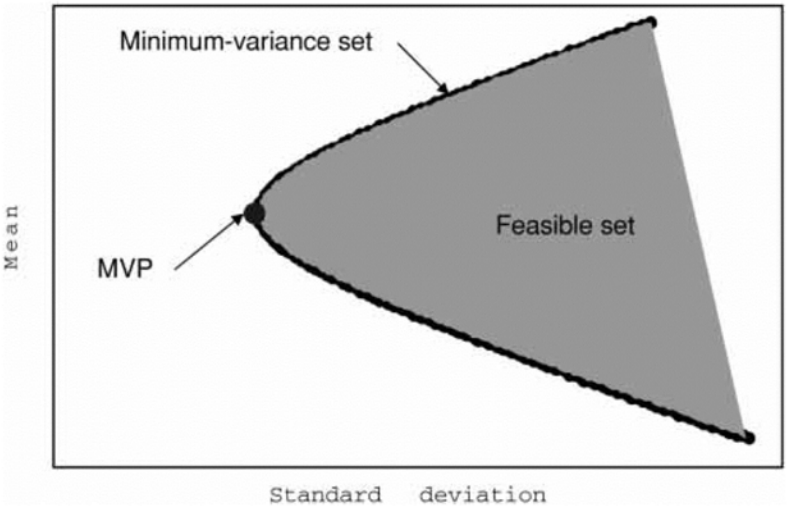


Figure 31.1: Feasible, minimum-variance, and efficient sets. The points in the minimum-variance set (that are above the MVP) form the efficient frontier, which is also called the **efficient set**. When short sales are not allowed, the feasible set is bounded because its mean lies within $[\min_i \bar{r}_i, \max_i \bar{r}_i]$ and its standard deviation lies within $[0, \max_i \sigma_i]$ (see Exercise 31.1.3).

Here is the mathematical formulation for the minimum-variance portfolio with a given mean value $\bar{\boldsymbol{r}}$ that is due to Markowitz in 1952 [641]:

$$\begin{aligned} \text{minimize} \quad & (1/2)\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_i\omega_j\sigma_{ij}, \\ \text{subject to} \quad & \sum_{i=1}^{n}\omega_i\bar{r}_i = \bar{\boldsymbol{r}}, \\ & \sum_{i=1}^{n}\omega_i = 1. \end{aligned}$$

Short selling can be prohibited if $\omega_i \geq 0$ for $i = 1, 2, \ldots, n,$ is added to the constraints. (The factor $1/2$ in front of the variance will simplify the analysis later.) The preceding **Markowitz problem** is a quadratic programming problem. It is a single-period investment theory that specifies the trade-off between the mean and the variance of a portfolio's rate of return.[3]

The Markowitz problem can be solved as follows. The weights ω_i and the two **Lagrange multipliers** λ and μ for an efficient portfolio satisfy

$$\begin{aligned} & \sum_{j=1}^{n}\sigma_{ij}\omega_j - \lambda\bar{r}_i - \mu = 0 \quad \text{for } i = 1, 2, \ldots, n, \\ & \sum_{i=1}^{n}\omega_i\bar{r}_i = \bar{\boldsymbol{r}}, \\ & \sum_{i=1}^{n}\omega_i = 1. \end{aligned}$$

There are $n+2$ equations with $n+2$ unknowns: $\omega_1, \omega_2, \ldots, \omega_n, \lambda, \mu$. Because the equations are linear, they can be easily solved (see Fig. 19.2). If the goal is to obtain the highest return for a given level of variance σ_p^2, then the problem becomes

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{n} \omega_i \bar{r}_i, \\ \text{subject to} \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j \sigma_{ij} = \sigma_p^2, \\ & \sum_{i=1}^{n} \omega_i = 1. \end{aligned}$$

Sophisticated quadratic programming techniques are needed to solve it.

Striking conclusions can be drawn from the mean–variance framework. Suppose that two solutions are available: (1) $(\omega_1, \lambda_1, \mu_1)$ with expected return rate $\bar{r}_1$ and (2) $(\omega_2, \lambda_2, \mu_2)$ with expected return rate $\bar{r}_2$. Direct substitution shows that $(\alpha\omega_1 + (1-\alpha)\omega_2, \alpha\lambda_1 + (1-\alpha)\lambda_2, \alpha\mu_1 + (1-\alpha)\mu_2)$ is also a solution to the $n+2$ equations and corresponds to the expected return rate $\alpha\bar{r}_1 + (1-\alpha)\bar{r}_2$. Thus the combined portfolio $\alpha\omega_1 + (1-\alpha)\omega_2$ also represents a point in the minimum-variance set. To use this result, suppose that ω_1 and ω_2 are two different portfolios in the minimum-variance set. Then as α varies over $-\infty < \alpha < \infty$, the portfolios defined by $\alpha\omega_1 + (1-\alpha)\omega_2$ sweep out the entire minimum-variance set. In particular, if ω_1 and ω_2 are efficient, they will generate all other efficient points. This is the **two-fund theorem**. Hence all investors seeking efficient portfolios need consider investing in combinations of only these two funds instead of individual stocks. This conclusion rests on the assumptions, among others, that everyone cares about only means and variances, that everyone has the same assessment of the parameters (means, variances, and covariances), that short selling is allowed, and that a single-period framework is appropriate.

➤ **Exercise 31.1.1** Express the efficient portfolio in matrix form.

➤ **Exercise 31.1.2** Construct a portfolio with zero risk from two perfectly negatively correlated assets without short sales.

➤ **Exercise 31.1.3** Let $C \equiv [\sigma_{ij}]$ be a positive definite matrix. (1) Prove that $\max_i \sigma_{ii}$ is the maximum value of $\sum_i \sum_j \omega_i \omega_j \sigma_{ij}$ under the constraints $\sum_i \omega_i = 1$ and $\omega_i \geq 0$. (2) How about the minimum value under the same constraints? (You may assume that the row sums of C^{-1} are all nonnegative.)

➤ **Exercise 31.1.4** Let $P(t)$ denote the asset price at time t. Define $r(T) \equiv [P(T)/P(0)] - 1$ as the holding period rate of return for a period of length T and $r_c(T) \equiv \ln(P(T)/P(0))$ as the continuous holding period rate of return for the same period. Under the assumption that asset prices are lognormally distributed, derive the relations between the mean and the variance of $r(T)$ and those of $r_c(T)$.

➤ **Exercise 31.1.5** Consider a portfolio P of n assets each following an independent geometric Brownian motion process with identical mean and variance, $dS_i/S_i = \mu dt + \sigma\, dW_i$. Each asset has the same weight of $1/n$ in the portfolio. Show that this portfolio's expected rate of return, $E[\ln(P(t)/P(0))]/t$, exceeds each

individual asset's expected rate of return, $E[\ln(S_i(t)/S_i(0))]/t$, by $(1-1/n)\sigma^2/2$. (Volatility is thus not synonymous with risk.)

31.1.1 Adding the Riskless Asset

The riskless asset by definition has a return that is certain; its return has zero volatility. The riskless return's covariance with any risky asset's return is thus zero. The presence of the riskless asset in a portfolio implies lending or borrowing cash at the riskless rate: Lending means a long position in the asset, whereas borrowing means a short position. Clearly the riskless asset has to be a zero-coupon bond whose maturity matches the investment horizon.

The shape of the feasible set changes dramatically when the riskless asset is available. Let r_f denote the riskless rate of return. Start with the feasible set defined by *risky* assets. Now for each portfolio in this set, say portfolio A, form combinations with the riskless asset. These new combinations trace out the infinite straight line originating at the riskless point, passing through the risky portfolio, and continuing indefinitely: the r_f–A ray in Fig. 31.2. There is a ray of this type for every portfolio in the feasible set. The totality of these rays forms a triangularly shaped feasible set. If borrowing of the riskless asset is not allowed, we can adjoin only the line segment between the riskless asset and points in the original feasible set but cannot extend the line further. The inclusion of these line segments leads to a feasible set with a straight-line front edge but a rounded top: the r_f–P–Q curve in Fig. 31.2. Note that investors who hold some riskless assets invest the remaining funds in portfolio P, as they are on the r_f–P segment.

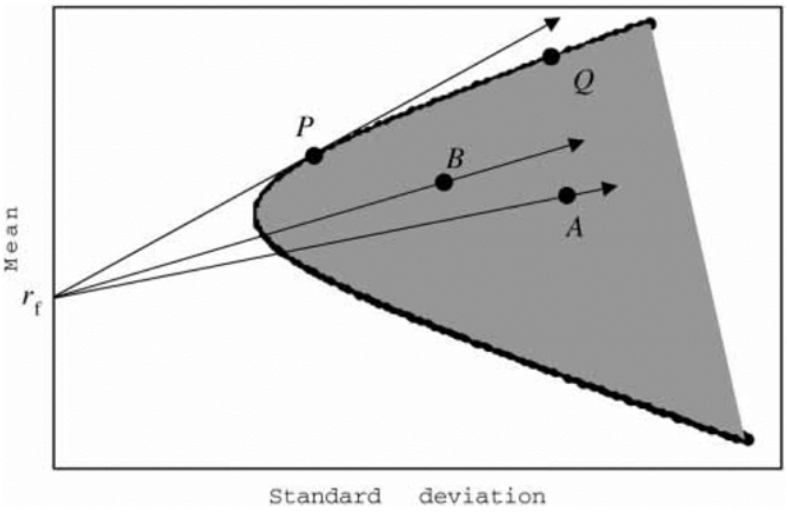


Figure 31.2: The efficient frontier with riskless lending and borrowing. The shaded area is the feasible set defined by risky assets. The line segment between r_f and A consists of combinations of portfolio A and lending, whereas the line segment beyond A consists of combinations of portfolio A and borrowing. The equation for the line is $y = r_f + x(\bar{r}_A - r_f)/\sigma_A$. The same observation can be made of any risky portfolio such as B, P, and Q. The ray through the tangent portfolio P defines the efficient frontier.

A special portfolio, denoted by P in Fig. 31.2, lies on the tangent point between the feasible set and a ray passing through r_f. When both borrowing and lending of the riskless asset are available, the efficient frontier is precisely this ray. Any efficient portfolio therefore can be expressed as a combination of P and the riskless asset. We have thus proved Tobin's **one-fund theorem**, which says there is a single fund of risky assets such that every efficient portfolio can be constructed as a combination of the fund and the riskless asset.

Identifying the tangent point P is computationally easy. For any point $(\sigma, \bar{r})$ in the feasible set defined by risky assets, we can draw a line between the riskless asset and that point as in Fig. 31.2. The slope is equal to $\theta \equiv (\bar{r} - r_f)/\sigma$, which has the interesting interpretation of the excess return per unit of risk. The tangent portfolio is the feasible point that maximizes θ. Assign weights $\omega_1, \omega_2, \ldots, \omega_n$ to the n risky assets such that $\sum_{i=1}^n \omega_i = 1$. The weight on the riskless asset in the tangent fund is zero. As a result, $\bar{r} - r_f = \sum_{i=1}^n \omega_i(\bar{r}_i - r_f)$, and

$$\theta = \frac{\sum_{i=1}^n \omega_i(\bar{r}_i - r_f)}{\sqrt{\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}\omega_i\omega_j}}.$$

Setting the derivative of θ with respect to each ω_j equal to zero leads to the equations

$$\lambda \sum_{i=1}^n \sigma_{ij}\omega_i = \bar{r}_j - r_f, \quad j = 1, 2, \ldots, n,$$

where $\lambda \equiv \sum_{i=1}^n \omega_i(\bar{r}_i - r_f)/(\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}\omega_i\omega_j) = (\bar{r} - r_f)/\sigma^2$. Making the substitution $v_i = \lambda\omega_i$ for each i simplifies the preceding equations to

$$\sum_{i=1}^n \sigma_{ij}v_i = \bar{r}_j - r_f, \quad j = 1, 2, \ldots, n. \tag{31.1}$$

We solve these linear equations for the v_is (see Fig. 19.2) and determine ω_i by setting $\omega_i = v_i/(\sum_{j=1}^n v_j)$. A negative ω_i means that asset i needs to be sold short.

If riskless lending and borrowing are disallowed, the whole efficient frontier can be traced out by solving Eq. (31.1) for all possible riskless rates, because an efficient portfolio is a tangent portfolio to a ray extending from *some* riskless rate (consult Fig. 31.2 again). However, there is a better way. Observe that v_i are linear in r_f; in other words, $v_i = c_i + d_i r_f$ for some constants c_i and d_i. We can find c_i and d_i by first solving Eq. (31.1) for v_i under two different r_fs, say r_f' and r_f''. The solutions v_i' and v_i'' correspond to two efficient portfolios. Now we solve

$$v_i' = c_i + d_i r_f',$$
$$v_i'' = c_i + d_i r_f''$$

for the unknown c_i and d_i for each i. By treating r_f as a variable and varying it, we can trace out the entire frontier. Just as the two-fund theorem says, two efficient portfolios suffice to determine the frontier.

➤ **Exercise 31.1.6** What would the one-fund theorem imply about trading volumes?

31.1.2 Alternative Efficient Portfolio Selection Models

In the **Black model**, portfolios are chosen subject only to $\sum_{i=1}^{n} \omega_i = 1$. In the **standard portfolio selection model**, short sales are disallowed, and the constraints are

$$\begin{aligned} &\sum_{i=1}^{n} \omega_i = 1, \\ &\omega_i \geq 0, \quad i = 1, 2, \ldots, n. \end{aligned}$$

By law or by policy, there may be restrictions on the amounts that can be invested in any one security. To handle them, one may augment the standard model with upper bounds:

$$\begin{aligned} &\sum_{i=1}^{n} \omega_i = 1, \\ &\omega_i \geq 0, \quad i = 1, 2, \ldots, n, \\ &\omega_i \leq u_i, \quad i = 1, 2, \ldots, n. \end{aligned}$$

In the **Tobin–Sharpe–Lintner model**, the portfolios are chosen subject to

$$\begin{aligned} &\sum_{i=1}^{n+1} \omega_i = 1, \\ &\omega_i \geq 0, \quad i = 1, 2, \ldots, n. \end{aligned}$$

The variable ω_{n+1} represents the amount lent (or borrowed if ω_{n+1} is negative). The covariances $\sigma_{n+1,i}$ are of course zero for $i = 1, 2, \ldots, n+1$. Limited borrowing can be modeled by the addition of the constraint $\omega_{n+1} \leq u_{n+1}$. In the **general portfolio selection model**, a portfolio is feasible if it satisfies

$$\begin{aligned} &A\boldsymbol{\omega} = \boldsymbol{b}, \\ &\boldsymbol{\omega} \geq 0, \end{aligned}$$

where A is any $m \times n$ matrix and $\boldsymbol{b}$ is an m-dimensional vector [642].

➤ **Exercise 31.1.7** Two portfolio selection models are **strictly equivalent** if they have the same set of obtainable mean–standard deviation combinations. Prove that any model that does not impose the nonnegative constraint on $\boldsymbol{\omega}$ is strictly equivalent to some general portfolio selection model, which does.

31.2 The Capital Asset Pricing Model

Imagine a world in which all investors are mean–variance portfolio optimizers and they share the same expectation as to expected returns, variances, and covariances. Also assume zero transactions cost. By the one-fund theorem, every investor will hold some amounts of the riskless asset and the same portfolio of risky assets. As all risky assets must be held by somebody, an immediate implication is that every investor holds the **market portfolio** in equilibrium regardless of one's degree of risk aversion. The market portfolio, which consists of all *risky* assets, is furthermore efficient (see Exercise 31.2.1).[4]

Given that the single efficient fund of risky assets is the market portfolio, the efficient frontier consists of a single straight line emanating from the riskless point

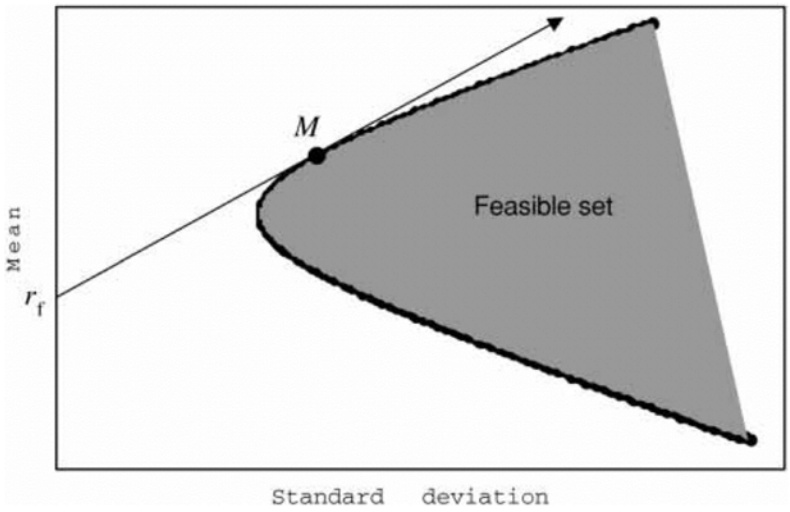


Figure 31.3: The capital market line. The point M stands for the market portfolio. The shaded set is the feasible set defined by risky assets. Investors can adjust the risk level by changing their holdings of riskless asset; for example, risk can be increased by holding negative amounts of the riskless asset.

and passing through the market portfolio. No complex computation is needed to determine the efficient frontier. This line is called the **capital market line**, which shows the relation between the expected rate of return and the risk for efficient portfolios (see Fig. 31.3). Prices should adjust so that efficient assets and portfolios fall on the line. Individual risky securities and inefficient portfolios, in contrast, will plot below the line. This is Sharpe's famous CAPM of 1964 [796],[5] which was independently arrived at by Lintner [605] and Mossin [680]. This model is fundamental to the equilibrium pricing of risky assets.

The capital market line states that

$$\bar{r} = r_f + \frac{\bar{r}_M - r_f}{\sigma_M}\sigma,$$

where $\bar{r}_M$ and σ_M are the expected value and the standard deviation of the market rate of return and $\bar{r}$ and σ are the expected value and the standard deviation of the rate of return of any efficient asset. Observe that as risk increases, the expected rate of return must also increase. The slope of the capital market line is $(\bar{r}_M - r_f)/\sigma_M$, which is called the **market price of risk**. It tells by how much the expected rate of return of an efficient portfolio must increase if the standard deviation of that rate increases by one unit. The market price of risk is also known as the **Sharpe ratio** [798].

The capital market line relates the expected rate of return of an efficient portfolio to its standard deviation, but it does not show how the expected rate of return of an individual asset relates to its individual risk. That relation is stated in the following theorem.

THEOREM 31.2.1 *If the market portfolio M is efficient, the expected return $\bar{r}_j$ of any asset j satisfies $\bar{r}_j - r_f = \beta_j(\bar{r}_M - r_f)$, where $\beta_j \equiv \sigma_{j,M}/\sigma_M^2$ and $\sigma_{j,M} \equiv \mathrm{Cov}[r_j, r_M]$.*

The value β_i is referred to as the **beta** of an asset. An asset's beta is all that needs to be known about its risk characteristics. The value $\bar{r}_i - r_f$ is the expected excess

Company	*Beta*	*Company*	*Beta*
America Online	2.43	Intel	1.03
AT&T	0.82	Merck	0.87
Citigroup	1.68	Microsoft	1.49
General Motors	1.01	Sun Micro.	1.19
IBM	1.07	Wal-Mart	1.20

Figure 31.4: Betas of some U.S. corporations. America Online merged with Time-Warner in 2001. Source: Standard & Poor's, May 8, 2000.

rate of return of asset i. It is the amount by which the rate of return is expected to exceed the riskless rate. Likewise, $\overline{r}_M - r_{\rm f}$ is the expected excess rate of return of the market portfolio. The CAPM says that the expected excess rate of return of an asset is proportional to the expected excess rate of return of the market portfolio, and the proportionality factor is beta. Beta, not volatility, is the measure of a security's risk, and the method of beating the market is to assume greater risk, i.e., beta. Figure 31.4 shows the betas of some U.S. corporations. We can estimate beta by regressing the excess return on the asset against the excess return on the market.

The CAPM formula in Theorem 31.2.1 shows a linear relation between beta and the expected rate of return for all assets whether they are efficient or not. This relationship, when plotted on a beta expected-return diagram, is termed the **security market line**. All assets fall on the security market line; in particular, the market is the point at $\beta = 1$.

Essentially the same arguments go through even if there is no riskless asset (see Exercise 31.2.12). The role of the riskless rate of return is then played by the mean rate of return in which the line in Fig. 31.3 intercepts the axis of mean rate of return.

➤ **Exercise 31.2.1** Verify that the market portfolio is efficient.

➤ **Exercise 31.2.2** Prove the security market line formula in Theorem 31.2.1.

31.2.1 More on the CAPM

The portfolio beta is the weighted average of the betas of the individual assets in the portfolio. Specifically, suppose a portfolio contains n assets with the weights $\omega_1, \omega_2, \ldots, \omega_n$. The rate of return of the portfolio is $r \equiv \sum_i \omega_i r_i$. Hence $\mathrm{Cov}[\, r, r_M \,] = \sum_i \omega_i \sigma_{i,M}$. It follows immediately that the portfolio beta equals $\sum_i \omega_i \beta_i$.

Write asset i's rate of return as

$$r_i = r_{\rm f} + \beta_i (r_M - r_{\rm f}) + \epsilon_i, \tag{31.2}$$

where $E[\, \epsilon_i \,] = 0$ by the CAPM. Now take the covariance of r_i with r_M in Eq. (31.2) to yield

$$\sigma_{i,M} = \beta_i^2 \sigma_M^2 + \mathrm{Cov}[\, \epsilon_i, r_M \,] = \sigma_{i,M} + \mathrm{Cov}[\, \epsilon_i, r_M \,].$$

Therefore $\mathrm{Cov}[\, \epsilon_i, r_M \,] = 0$ and

$$\sigma_i^2 = \beta_i^2 \sigma_M^2 + \mathrm{Var}[\, \epsilon_i \,]. \tag{31.3}$$

It is important to note that the total risk σ_i^2 is a sum of two parts. The first part, $\beta_i^2\sigma_M^2$, is the systematic risk. This is the risk associated with the market as a whole, also called the **market risk**. It cannot be reduced by diversification because every asset with nonzero beta contains this risk. The second part $\mathrm{Var}[\,\epsilon_i\,]$ is the specific risk. This risk is uncorrelated with the market and can be reduced by diversification. Only the systematic risk has any bearing on returns.

Consider an asset on the capital market line with a beta of β and an expected rate of return equal to $\overline{r} = r_{\mathrm{f}} + \beta(\overline{r}_M - r_{\mathrm{f}})$. This asset, which is efficient, must be equivalent to a combination of the market portfolio and the riskless asset. Its standard deviation is therefore $\beta\sigma_M$, which implies it has only systematic risk but no specific risk by Eq. (31.3). Now consider another asset with the same beta β. According to the CAPM, its expected rate of return must be $\overline{r}$. However, if it carries specific risk, it will not fall on the capital market line. The specific risk is thus the distance by which the portfolio lies below the capital market line.

Although stated in terms of expected returns, the CAPM is also a pricing model. Suppose an asset is purchased at a known price P and later sold at price Q. The rate of return is $r \equiv (Q - P)/P$. By the CAPM,

$$\frac{\overline{Q} - P}{P} = r_{\mathrm{f}} + \beta(\overline{r}_M - r_{\mathrm{f}}),$$

where β is the beta of the asset. Solve for P to obtain

$$P = \frac{\overline{Q}}{1 + r_{\mathrm{f}} + \beta(\overline{r}_M - r_{\mathrm{f}})}. \tag{31.4}$$

Hence the CAPM can be used to decide whether the price for a stock is "right." Note that the risk-adjusted interest rate is $r_{\mathrm{f}} + \beta(\overline{r}_M - r_{\mathrm{f}})$, not r_{f}.

Equation (31.4) can take another convenient form. The value of beta is

$$\beta = \frac{\mathrm{Cov}[\,r, r_M\,]}{\sigma_M^2} = \frac{\mathrm{Cov}[\,(Q/P) - 1, r_M\,]}{\sigma_M^2} = \frac{\mathrm{Cov}[\,Q, r_M\,]}{P\sigma_M^2}.$$

Substituting this into pricing formula (31.4) and dividing by P yields

$$1 = \frac{\overline{Q}}{P(1 + r_{\mathrm{f}}) + \mathrm{Cov}[\,Q, r_M\,](\overline{r}_M - r_{\mathrm{f}})/\sigma_M^2}.$$

Solve for P again to obtain

$$P = \frac{1}{1 + r_{\mathrm{f}}}\left\{\overline{Q} - \frac{\mathrm{Cov}[\,Q, r_M\,](\overline{r}_M - r_{\mathrm{f}})}{\sigma_M^2}\right\}. \tag{31.5}$$

This demonstrates it is the asset's covariance with the market that is relevant for pricing.

➤ **Exercise 31.2.3** For an asset uncorrelated with the market (that is, with zero beta), the CAPM says its expected rate of return is the riskless rate even if this asset is very risky with a large standard deviation. Why?

➤ **Exercise 31.2.4** If an asset has a negative beta, the CAPM says its expected rate of return should be less than the riskless rate even if this asset is very risky with a large standard deviation. Why? (For example, we saw in Chap. 24 that IO strips earn less than the riskless rate despite their high riskiness.)

➤ **Exercise 31.2.5** Why must all portfolios with the same expected rate of return but different total risks fall on the same point on the security market line?

➤ **Exercise 31.2.6** (1) Verify that pricing formula (31.4) is linear (the price of the sum of two assets is the sum of their prices, and the price of a multiple of an asset is the same multiple of the price). (2) Derive the same results from the no-arbitrage principle.

31.2.2 Portfolio Insurance

Portfolio insurance is a trading strategy that protects a portfolio from market declines but without losing the opportunity to participate in market rallies – in a word, a protective put [772]. Using puts to protect a portfolio from falling below a specified floor is a simple example of *static* portfolio insurance. Alternatives to static schemes are dynamic strategies that create synthetic options with stocks and bonds. Dynamic strategies, however, generate high transactions costs. This problem was mitigated by the introduction of stock index futures. Compared with the underlying assets, futures can be traded at much lower transactions costs in achieving the desired mixture of risky and riskless assets.[6]

Let the value of the index be S and each put be on \$100 times the index. Consider a diversified portfolio with a beta of β. If for each $100 \times S$ dollars in the portfolio, one put contract is purchased with strike price X, the value of the portfolio is protected against the possibility of the index's falling below the floor of X. Our goal is to implement this protective put. Specifically, to protect each dollar of the portfolio against falling below W at time T, we buy β put contracts for each $100 \times S$ dollars in the portfolio. Note that the total number of puts bought is $\beta V/(S \times 100)$, where V is the current value of the portfolio. The strike price X is the index value when the portfolio value reaches W.

Let r be the interest rate and q the dividend yield. Suppose that the index reaches S_T at time T. The excess return of the index over the riskless interest rate is $(S_T - S)/S + q - r$, and the excess return of the portfolio over the riskless interest rate is $\beta((S_T - S)/S + q - r)$. The return from the portfolio is therefore $\beta[\,(S_T - S)/S + q - r\,] + r$, and the increase in the portfolio value net of the dividends is $\beta[\,(S_T - S)/S + q - r\,] + r - q$. Therefore the portfolio value per dollar of the original value is

$$1 + \beta\left(\frac{S_T - S}{S} + q - r\right) + r - q = \beta\,\frac{S_T}{S} + (\beta - 1)(q - r - 1). \tag{31.6}$$

Choose X to be the S_T that makes Eq. (31.6) equal W; in other words,

$$X = [\,W + (q - r - 1)(1 - \beta)\,]\,\frac{S}{\beta}.$$

From Eq. (31.6), the portfolio value is less than W by $\beta(\Delta S/S)$ if and only if the index value is less than X by $\beta(\Delta S/S)(S/\beta) = \Delta S$. Exercising the options therefore induces a matching cash inflow of

$$\beta\,\frac{\Delta S}{100 \times S} \times 100 = \beta\,\frac{\Delta S}{S}.$$

The strategy's cost is $P\beta V/(S \times 100)$, where P is the put premium with strike price X.

Clearly a higher strike price provides a higher floor of WV dollars at a greater cost. This trade-off between the cost of insurance and the level of protection is typical of any insurance. The total wealth of course has a floor of

$$WV - \frac{P\beta V}{S \times 100}.$$

EXAMPLE 31.2.2 Start with $S = 1000$, $\beta = 1.5$, $q = 0.02$, and $r = 0.07$ for a period of 1 year. We have the following relations between the index value and the portfolio value per dollar of the original value.

Index value in a year	*1200*	*1100*	*1000*	*900*	*800*
Portfolio value in a year	1.275	1.125	0.975	0.825	0.675

For example, if the portfolio starts at \$1 million and the insured value is \$0.825 million, then $(1.5 \times 1{,}000{,}000)/(100 \times 1{,}000) = 15$ put contracts with a strike price of 900 should be purchased.

➤ **Exercise 31.2.7** Redo Example 31.2.2 with $S = 1000, \beta = 2, q = 0.01$, and $r = 0.05$.

➤ **Exercise 31.2.8** Consider a portfolio worth \$1,000 times the S&P 500 Index and with a beta of 1.0 against the index. Argue that buying 10 put index options with a strike price of 1,000 insures against the portfolio value's dropping below \$1,000,000.

➤ **Exercise 31.2.9** A mutual fund manager believes that the market is going to be relatively calm in the near future and writes a covered index call. Analyze it by following the same logic as that of the protective put.

➤ **Exercise 31.2.10** A bank offers the following financial product to a mutual fund manager planning to buy a certain stock in the near future. If the stock price is over \$50, the manager buys it at \$50. If the stock price is below \$40, the manager buys it at \$40. If the stock price is between the two, the manager buys at the spot price. Analyze the underlying options.

31.2.3 Critical Remarks

Fire those CAPM-peddling consultants.
—Louis Lowenstein [620]

Although the CAPM is widely used by practitioners [592], many of its assumptions have been controversial. It assumes either normally distributed asset returns or quadratic utility functions. It furthermore assumes that investors care about only the mean and the variance of returns, which implies that they view upside and downside risks with equal distaste. In reality, portfolio returns are not, strictly speaking, normally distributed, and investors seem to distinguish between upside and downside risks. The theory posits, unrealistically, that everyone has identical information about the returns of all assets and their covariances. Even if this assumption were valid, it would not be easy to obtain accurate data. Usually, the variances and covariances can be accurately estimated, but not the expected returns (see Example 20.1.1). Unfortunately, errors in means are more critical than errors in variances, and

errors in variances are more critical than errors in covariances [204]. The assumption that all investors share a common investment horizon is rarely the case in practice.

The CAPM assumes that all assets can be bought and sold on the market. The assets include not just securities, but also real estate, cash, and even human capital. Because the market portfolio is difficult to define, in reality proxies for the market portfolio are used [799]. The trouble is that different proxies result in different beta estimates for the same security (see Exercise 31.2.12). Finally, a single risk factor does not seem adequate for describing the cross section of expected returns [145, 336, 424, 635, 636, 666].

➤ **Exercise 31.2.11** Why are security analysts' 1-year forecasts worse than 5-year ones?

➤ **Exercise 31.2.12** Prove that using any efficient portfolio for the risky assets as the proxy for the market portfolio results in linear relations between the expected rates of return and the betas, just as in the CAPM.

31.3 Factor Models

The mean–variance theory requires that many parameters be estimated: n for the expected returns of the assets and $n(n+1)/2$ for their covariances. Luckily, asset returns can often be explained by a much smaller number of underlying sources of randomness called factors. A factor model represents the connection between factors and individual returns. In this section a factor model of the return process for asset pricing is presented, the **Arbitrage Pricing Theory** (**APT**).

31.3.1 Single-Factor Models

We start with single-factor models. Suppose there are n assets with rates of return, $r_1, r_2, \ldots, r_n$. There is a single factor f, which is a random quantity such as the return on a stock index for the holding period. The rates of return and the factor are related by

$$r_i = a_i + b_i f + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where a_i and b_i are constants. The b_is are the factor loadings or **factor betas**, and they measure the sensitivity of the return to the factor. Without loss of generality, let $E[\,\epsilon_i\,] = 0$. Assume that ϵ_i are uncorrelated with f: $\mathrm{Cov}[\,f, \epsilon_i\,] = 0$. Furthermore, assume that they are uncorrelated with each other, i.e., $E[\,\epsilon_i \epsilon_j\,] = 0$ for $i \neq j$.[7] Any correlation between asset returns thus arises from a common response to the factor. The variance of ϵ_i is denoted by $\sigma_{\epsilon_i}^2$ and that of f by σ_f^2. There is a total of $3n+2$ parameters: $a_i, b_i, \sigma_{\epsilon_i}^2, \overline{f}$, and σ_f^2. The following results are straightforward:

$$\begin{aligned}
\overline{r}_i &= a_i + b_i \overline{f}, \\
\sigma_i^2 &= b_i^2 \sigma_f^2 + \sigma_{\epsilon_i}^2, \\
\mathrm{Cov}[\,r_i, r_j\,] &= b_i b_j \sigma_f^2, \quad i \neq j, \\
b_i &= \mathrm{Cov}[\,r_i, f\,] / \sigma_f^2.
\end{aligned}$$

The preceding simple covariance matrix leads to very efficient algorithms for the portfolio selection problems in Subsection 31.1.1 [317]. The single-factor model is due to Sharpe [795].

The return on a portfolio can be analyzed similarly. Consider a portfolio constructed with weights ω_i. Its rate of return is just

$$r = a + bf + \epsilon,$$

where $a \equiv \sum_{i=1}^{n} \omega_i a_i$, $b \equiv \sum_{i=1}^{n} \omega_i b_i$, and $\epsilon \equiv \sum_{i=1}^{n} \omega_i \epsilon_i$. The portfolio's beta b is hence the average of the underlying assets' betas b_i (recall the law of the average covariance). It is easy to verify that $E[\,\epsilon\,] = 0$, $\mathrm{Cov}[\,f, \epsilon\,] = 0$, and $\mathrm{Var}[\,\epsilon\,] = \sum_{i=1}^{n} \omega_i^2 \sigma_{\epsilon_i}^2$. Finally, the variance of r is

$$\sigma^2 = b^2 \sigma_f^2 + \mathrm{Var}[\,\epsilon\,],$$

similar to Eq. (31.3). Among the total risk above, the systematic part is $b^2\sigma_f^2$. The systematic risk, which is due to the $b_i f$ terms, results from the factor f that influences every asset and is therefore present even in a diversified portfolio. The $\mathrm{Var}[\,\epsilon\,]$ term represents the specific risk. The specific risk, which is due to the ϵ_i terms, can be made to go to zero through diversification. It is also called the **diversifiable risk**.

➤ **Exercise 31.3.1** Assume that the single factor f is the market rate of return, r_M. Write the return processes as $r_i - r_f = \alpha_i + b_i(r_M - r_f) + \epsilon_i$. As usual, $E[\,\epsilon_i\,] = 0$ and ϵ_i is uncorrelated with the market return. Show that $b_i = \mathrm{Cov}[\,r_i, r_M\,]/\mathrm{Var}[\,r_M\,]$, as in the CAPM.

31.3.2 Multifactor Models

Now there are two factors f_1 and f_2, and the rate of return of asset i takes the form

$$r_i = a_i + b_{i1} f_1 + b_{i2} f_2 + \epsilon_i.$$

As in Subsection 31.3.1, assume that $E[\,\epsilon_i\,] = 0$ and that ϵ_i is uncorrelated with the factors and ϵ_j for $j \neq i$. The formulas for the expected rates of return and covariances are

$$\begin{aligned}
\overline{r}_i &= a_i + b_{i1}\overline{f}_1 + b_{i2}\overline{f}_2, \\
\sigma_i^2 &= b_{i1}^2\sigma_{f_1}^2 + b_{i2}^2\sigma_{f_2}^2 + 2b_{i1}b_{i2}\,\mathrm{Cov}[\,f_1,\, f_2\,] + \sigma_{\epsilon_i}^2, \\
\mathrm{Cov}[\,r_i, r_j\,] &= b_{i1}b_{j1}\sigma_{f_1}^2 + b_{i2}b_{j2}\sigma_{f_2}^2 + (b_{i1}b_{i2} + b_{j1}b_{j2})\,\mathrm{Cov}[\,f_1,\, f_2\,], \quad i \neq j.
\end{aligned}$$

From the preceding equations,

$$\begin{aligned}
\mathrm{Cov}[\,r_i,\, f_1\,] &= b_{i1}\sigma_{f_1}^2 + b_{i2}\,\mathrm{Cov}[\,f_1,\, f_2\,], \\
\mathrm{Cov}[\,r_i,\, f_2\,] &= b_{i2}\sigma_{f_2}^2 + b_{i1}\,\mathrm{Cov}[\,f_1,\, f_2\,].
\end{aligned}$$

These give two equations that can be solved for b_{i1} and b_{i2}. Factor models with more than two factors are easy generalizations. For U.S. stocks, between 3 and 15 factors may be needed [623].

➤ **Exercise 31.3.2** Describe a procedure to convert a set of correlated factors into a set of uncorrelated factors, which are easier to handle.

31.3.3 The Arbitrage Pricing Theory (APT)

The factor-model framework leads to an alternative theory of asset pricing, Ross's APT, which is a theory about equilibrium under factor models [765]. The APT does not require that investors evaluate portfolios on the basis of means and variances. Neither is a quadratic utility function required. Instead, (1) the mean–variance framework is replaced with a factor model for returns, (2) investors are assumed to prefer a greater return to a lesser return when returns are certain, and (3) the universe of assets is assumed to be large.

Consider first a special case in which the rates of return observe the one-factor model:

$$r_i = a_i + b_i f.$$

This factor model has no residual errors, and the uncertainty associated with a return is due solely to the uncertainty in the factor f. Interestingly, the values of a_i and b_i must be related if arbitrage opportunities are to be excluded. Here is the argument. Consider two assets i and j with $b_i \neq b_j$. Now form a portfolio with weights $\omega_i \equiv \omega$ for asset i and $\omega_j \equiv 1-\omega$ for asset j. Its rate of return is

$$r = \omega a_i + (1-\omega)\, a_j + (\omega b_i + (1-\omega)\, b_j)\, f.$$

If we select $\omega = b_j/(b_j - b_i)$ to make the coefficient of f zero, the rate of return r becomes

$$\lambda_0 \equiv \frac{a_i b_j - a_j b_i}{b_j - b_i}.$$

This portfolio is riskless because the equation for r contains no random elements. If there happens to be a riskless asset, then $\lambda_0 = r_{\mathrm{f}}$. Even if riskless assets do not exist, all portfolios constructed without dependence on f must have the same rate of return, λ_0. Now $\lambda_0(b_j - b_i) = a_i b_j - a_j b_i$, which can be rearranged as

$$\frac{a_j - \lambda_0}{b_j} = \frac{a_i - \lambda_0}{b_i}.$$

As this relation holds for all i and j, there is a constant c such that

$$\frac{a_i - \lambda_0}{b_i} = c$$

for all i.[8] The values of a_i and b_i are thus related by $a_i = \lambda_0 + b_i c$. The expected rate of return of asset i is now

$$\overline{r}_i = a_i + b_i \overline{f} = \lambda_0 + b_i c + b_i \overline{f} = \lambda_0 + b_i \lambda_1, \tag{31.7}$$

where $\lambda_1 \equiv c + \overline{f}$. We see that once the constants λ_0 and λ_1 are known, the expected return of an asset is determined entirely by the factor betas b_i. The above analysis can be generalized (see Exercise 31.3.4).

THEOREM 31.3.1 *Let there be n assets whose rates of return are governed by $m < n$ factors according to the equations $r_i = a_i + \sum_{j=1}^{m} b_{ij} f_j$, $i = 1, 2, \ldots, n$. Then there exist constants $\lambda_0, \lambda_1, \ldots, \lambda_m$ such that $\overline{r}_i = \lambda_0 + \sum_{j=1}^{m} b_{ij}\lambda_j$, $i = 1, 2, \ldots, n$. The value λ_i is called the market price of risk associated with factor f_i or simply the* **factor price**.

Next we consider general multifactor models with residual errors,

$$r_i = a_i + \sum_{j=1}^{m} b_{ij} f_j + \epsilon_i,$$

where $E[\epsilon_i] = 0$ and $\sigma_{\epsilon_i}^2 \equiv E[\epsilon_i^2]$. As before, ϵ_i is assumed to be uncorrelated with the factors and with the residual errors of other assets. Let us form a portfolio by using the weights $\omega_1, \omega_2, \ldots, \omega_n$ with $\sum_{i=1}^{n} \omega_i = 1$. The rate of return of the portfolio is

$$r = a + \sum_{j=1}^{m} b_j f_j + \epsilon,$$

where $a \equiv \sum_{i=1}^{n} \omega_i a_i$, $b_j \equiv \sum_{i=1}^{n} \omega_i b_{ij}$, and $\epsilon \equiv \sum_{i=1}^{n} \omega_i \epsilon_i$. Let $\sigma_{\epsilon_i} \leq S$ for some constant S for all i. Assume that the portfolio is *well diversified* in the sense that $\omega_i \leq W/n$ for some constant W for all i – no single asset dominates the portfolio. Then

$$\text{Var}[\epsilon] = \sum_{i=1}^{n} \omega_i^2 \sigma_{\epsilon_i}^2 \leq \frac{1}{n^2} \sum_{i=1}^{n} W^2 S^2 = \frac{W^2 S^2}{n} \to 0$$

as $n \to \infty$. Combined with the fact $E[\epsilon] = 0$, the residual error ϵ of a well-diversified portfolio selected from a very large number of assets is approximately zero.[9]

A riskless portfolio in terms of zero sensitivity to all factors was used in the proof of Theorem 31.3.1. We just showed that the portfolio remains riskless under the more general models as long as it is well diversified. The existence of a riskless well-diversified portfolio suffices to extend Theorem 31.3.1 to the more general models (see Exercise 31.3.5).

The APT and the CAPM are not directly comparable [289]. Neither of the two models' assumptions imply the other's. The CAPM makes strong assumptions about the probability distribution of assets' rates of return, agents' utility functions, or both. The APT on the other hand, makes strong assumptions about assets' equilibrium rates of return. However, because the APT does not identify the factors, the CAPM can be made consistent with the APT and vice versa. For example, consider a two-factor model $r_i = a_i + b_{i1} f_1 + b_{i2} f_2 + \epsilon_i$. Under the APT model, $\overline{r}_i = r_f + b_{i1}\lambda_1 + b_{i2}\lambda_2$. Let portfolio j $(j = 1, 2)$ have an expected return rate of $\lambda_j + r_f$ and a beta value of β_{f_j}. Clearly portfolio j's only source of risk is factor f_j. Because the CAPM says that $\lambda_j = \beta_{f_j}(\overline{r}_M - r_f)$,

$$\overline{r}_i = r_f + b_{i1}\beta_{f_1}(\overline{r}_M - r_f) + b_{i2}\beta_{f_2}(\overline{r}_M - r_f) = r_f + (b_{i1}\beta_{f_1} + b_{i2}\beta_{f_2})(\overline{r}_M - r_f).$$

The beta is thus a weighted sum of the underlying factors' betas with the factor betas as the weights. Different factor betas are the reason different assets have different betas.

➤ **Exercise 31.3.3** For the one-factor APT, what will become of λ_1 if the CAPM holds?

➤ **Exercise 31.3.4** Prove Theorem 31.3.1.

➤ **Exercise 31.3.5** Complete the proof of the APT under the general factor models.

31.4 Value at Risk

> Anyone that relied on so-called
> value-at-risk models has been crucified.
> —The Economist, 1999 [309]

Introduced in 1983, the VaR is an attempt to provide a single number for senior management that summarizes the total risk in a portfolio of financial assets. The VaR calculation is aimed at making a statement of the form: "We are c percent certain not to lose more than V dollars in the next m days." The variable V is the VaR of the portfolio. The VaR is therefore an estimate, with a given degree of confidence, of how much one can lose from one's portfolio over a given time horizon, or

$$\text{Prob}[\text{ change in portfolio value} \le -\text{VaR}\,] = 1 - c,$$

where c is the confidence level (see Fig. 31.5). The VaR is usually calculated assuming "normal" market circumstances, meaning that extreme market conditions such as market crashes are not considered. It has become widely used by corporate treasurers and fund managers as well as financial institutions.

For the purposes of measuring the adequacy of bank capital, the Bank for International Settlements (BIS) sets the confidence level $c = 0.99$ and the time horizon $m = 10$ (days) [293]. Another interesting application is in investment evaluation. Here risk is viewed in terms of the impact of the prospective change on the overall value at risk, i.e., *incremental* VaR, and we go ahead with the investment if the incremental VaR is low enough relative to the expected return [283].

Suppose returns are normally distributed and independent on successive days.[10] We consider a single asset first. We assume that the stock price is S, whose daily volatility for the return rate $\Delta S/S$ is σ. Because the time horizon m is usually small, we assume that the expected price change is zero. The standard deviation of the stock price over this time horizon is $S\sigma\sqrt{m}$. The VaR of holding one unit of the stock is $2.326 \times S\sigma\sqrt{m}$ if the confidence level is 99% and $1.645 \times S\sigma\sqrt{m}$ if

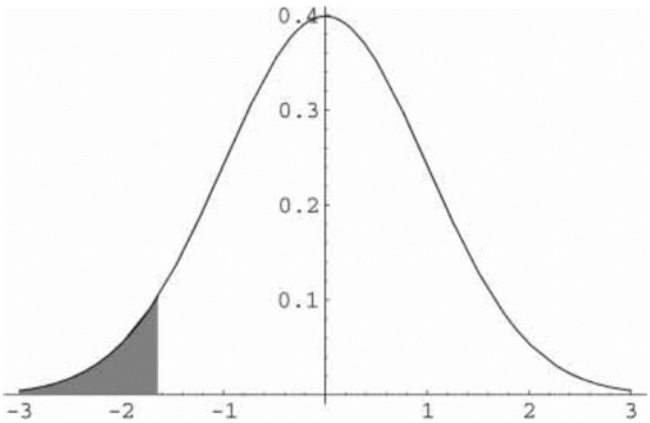


Figure 31.5: Confidence level and VaR. The diagram shows a confidence level of 95% under the standard normal distribution; the shaded area is 5% of the total area under the density function. It corresponds to 1.64485 standard deviations from the mean.

Confidence level (c)	*Number of standard deviations*
95%	1.64485
96%	1.75069
97%	1.88079
98%	2.05375
99%	2.32635

Figure 31.6: Confidence levels and standard deviations from the mean. The table samples confidence levels and their corresponding numbers of standard deviations from the mean when the random variable is normally distributed.

the confidence level is 95%. In general, the VaR is $-N^{-1}(1-c)$ times the standard deviation, or

$$-N^{-1}(1-c)\, S\sigma\sqrt{m},$$

where $N(\cdot)$ is the distribution function of the standard normal distribution (see Fig. 31.6). The preceding equation makes it easy to convert one horizon or confidence level to another. For example, the relation between 99% VaR and 95% VaR is

$$\text{VaR}\,(95\%) = \text{VaR}\,(99\%) \times (1.645/2.326).$$

Similarly, the variance of an m-day return should be m times the variance of a 1-day return. The m-day VaR thus equals $\sqrt{m}$ times the 1-day VaR, which is also called the **daily earnings at risk**. When m is not small, the expected annual rate of return μ needs to be considered. In that case, the drift $S\mu m/T$ is subtracted from the VaR when there are T trading days per annum.

Now consider a portfolio of assets. Assume that the changes in the values of asset prices have a multivariate normal distribution. Let the daily volatility of asset i be σ_i, let the correlation between the returns on assets i and j be ρ_{ij}, and let S_i be the market value of the positions in asset i. The VaR for the whole portfolio then is

$$-N^{-1}(1-c)\sqrt{m}\sqrt{\sum_i \sum_j S_i S_j \sigma_i \sigma_j \rho_{ij}}\,.$$

This way of computing the VaR is called the **variance–covariance approach** [518]. It was popularized by J.P. Morgan's RiskMetrics™ (1994).

The variance–covariance methodology may break down if there are derivatives in the portfolio because the returns of derivatives may not be normally distributed even if the underlying asset is. Nevertheless, if movements in the underlying asset are expected to be very small because, say, the time horizon is short, we may approximate the sensitivity of the derivative to changes in the underlying asset by the derivative's delta as follows. Consider a portfolio P of derivatives with a single underlying asset S. Recall that the delta of the portfolio, δ, measures the price sensitivity to S, or approximately $\Delta P/\Delta S$. The standard deviation of the distribution of the portfolio is $\delta S\sigma\sqrt{m}$, and its VaR is $-N^{-1}(1-c)\,\delta S\sigma\sqrt{m}$. In general when there are many underlying assets, the VaR of a portfolio containing options becomes

$$-N^{-1}(1-c)\sqrt{m}\sqrt{\sum_i \sum_j \delta_i \delta_j S_i S_j \sigma_i \sigma_j \rho_{ij}},$$

where δ_i denotes the delta of the portfolio with respect to asset i and S_i is the value of asset i. This is called the **delta approach** [527, 878]. The delta approach

essentially treats a derivative as delta units of its underlying asset for the purpose of VaR calculation. This is not entirely unreasonable because such equivalence does hold instantaneously. It becomes questionable, however, as m increases.

Rather than using asset prices, VaR usually relies on a limited number of basic market variables that account for most of the changes in portfolio value [603]. As mentioned in Subsection 31.3.1, this greatly reduces the complexity related to the covariance matrix because only the covariances between the market variables are needed now. Typical market variables are yields or bond prices, exchange rates, and market returns. A basic instrument is then associated with each market variable. A security is now approximated by a portfolio of these basic instruments. Finally, its VaR is reduced to those of the basic instruments.

➤ **Exercise 31.4.1** What is the VaR of a futures contract on a stock?

➤ **Exercise 31.4.2** If the stock price follows $dS = S\mu\,dt + S\sigma\,dW$, what is its VaR τ years from now at c confidence?

31.4.1 Simulation

The Monte Carlo simulation is a general method to estimate the VaR, particularly for derivatives [571]. It works by computing the values of the portfolio over many sample paths, and the VaR is based on the distribution of the values. Figure 31.7 contains an algorithm for n asset prices following geometric Brownian motion:

$$\frac{dS_j}{S_j} = \mu_j\,dt + \sigma_j\,dW_j, \quad j = 1, 2, \ldots, n,$$

where the n factors, dW_j, are correlated. As always, *actual* returns, not risk-neutral returns, should be used. For short time horizons, this distinction is not critical for most cases. In practice, to save computation time, a stock with a beta of β is mapped to a position in β times the index. Of course, this approach ignores the stock's specific risk. A related simulation method, called **historical simulation**, utilizes historical data [518]. It is identical to the Monte Carlo simulation except that the sample paths are generated by sampling the historical data as if they are to be repeated in the future.

Brute-force Monte Carlo simulation is inefficient when the number of factors is large. Fortunately, factor analysis and principal components analysis can often reduce the number of factors needed in the simulation [2, 509]. Let C denote the covariance matrix of the n factors $dW_1, dW_2, \ldots, dW_n$. Let $u_i \equiv [\,u_{1i}, u_{2i}, \ldots, u_{ni}\,]^{\mathrm{T}}$ be the eigenvectors of C and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the corresponding positive eigenvalues. Hence $\lambda_i u_i = C u_i$ for $i = 1, 2, \ldots, n$. Recall that each eigenvalue indicates how much of the variation in the data its corresponding eigenvector explains. By the Schur decomposition theorem, the eigenvectors can be assumed to be orthogonal to each other. Normalize the eigenvectors such that $|u_i|^2 = \lambda_i$ and define

$$dZ_j \equiv \lambda_j^{-1} \sum_{k=1}^{n} u_{kj}\,dW_k.$$

It follows that

$$dW_i = \sum_{k=1}^{n} u_{ik}\,dZ_k,$$

VaR with Monte Carlo simulation:

```
input:   p̄, c, n, C[n][n], S[n], μ[n], Δt, m, N;
real     S[m+1][n], y[n], dW[n], P[n][n], p[N];
real     ξ(); // ξ() ~ N(0,1).
integer  i, j, k;
Let P be such that C = PP^T;   // See p. 248.
for (j = 0 to n−1) { S[0][j] := S[j]; }
for (k = 0 to N−1) {
        for (i = 1 to m) {
            for (j = 0 to n−1)
                y[j] := ξ() × √Δt;
            dW := Py;
            for (j = 0 to n−1) {
                S[i][j] := S[i−1][j] × ((1 + μ[j]) × Δt + √C[j][j] × dW[j]);
            }
        }
        Calculate the horizon portfolio value p[k];
}
Sort p[0], p[1], ..., p[N−1] in non-decreasing order;
return p̄ − p[⌊(1−c)N−1⌋];
```

Figure 31.7: VaR with Monte Carlo simulation. The expected rates of return $\mu[\,]$ and the covariances are annualized. There are n assets, the portfolio's initial value is $\overline{p}$, c is the confidence level, C is the covariance matrix for the annualized asset returns, m is the number of days until the horizon, the number of replications is N, and $S[\,]$ stores the initial asset prices. Recall that $C = P\,P^{\intercal}$ is the Cholesky decomposition of C. The portfolio's values at the horizon date are calculated and stored in $p[\,]$. Here we need pricing models and assume that early exercise is not possible during the period. The appropriate percentile is returned after sorting.

where $dZ_k\,dZ_j = 0$ for $j \neq k$ and $dZ_k\,dZ_k = dt$ (see Exercise 31.4.3, part (3)). If the empirical analysis shows that all but the first m principal components are small, then

$$dW_i \approx \sum_{k=1}^{m} u_{ik}\,dZ_k.$$

As a result, the asset price processes can be approximated by

$$\frac{dS_j}{S_j} \approx \mu_j\,dt + \sigma_j \sum_{k=1}^{m} u_{ik}\,dZ_k, \quad j = 1, 2, \ldots, n.$$

Only m orthogonal factors $dZ_1, dZ_2, \ldots, dZ_m$ remain.

➤ **Exercise 31.4.3** Prove that (1) $C = PP^{\intercal}$, where P's ith column is the eigenvector u_i, (2) $P^{-1} = \mathrm{diag}[\,\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_n^{-1}\,]\,P^{\intercal}$, (3) $P\,[\,dZ_1, dZ_2, \ldots, dZ_n\,]^{\intercal} = [\,dW_1, dW_2, \ldots, dW_n\,]^{\intercal}$, and (4) $P^{\intercal}P = \mathrm{diag}[\,\lambda_1, \lambda_2, \ldots, \lambda_n\,]$.

31.4.2 Critical Remarks

VaR relies on certain assumptions that are inconsistent with empirical evidence. Many implementations assume that asset returns are normally distributed. This simplifies the computation considerably but is inconsistent with the empirical evidence,

which finds that many returns have fat tails, both left and right, at both daily and monthly time horizons. Extreme events are hence much more likely to occur in practice than would be predicted based on the assumption of normality [857]. A standard measure of tail fatness is kurtosis. Price jumps and stochastic volatility can be used to generate fat tails (see, e.g., Exercise 20.2.1) [293]. Although daily market returns are not normal [743], for longer periods, say 3 months, returns are quite close to being normally distributed [592].

The method of calculating the VaR depends on the horizon. A method yielding good results over a short horizon may not work well over longer horizons. The method of calculating the VaR also depends on asset types. If the portfolio contains derivatives, methods different from these used to analyze portfolios of stocks may be needed [464].

The ability to quantify risk exposure into a number represents the single most powerful advantage of the VaR. However, the VaR is extremely dependent on parameters, data, assumptions, and methodology. Although it should be part of an effective risk management program, the VaR is not sufficient to control risk [61, 464]. On occasion, it becomes necessary to quantify the magnitude of the losses that might accrue under events less likely than those analyzed in a standard VaR calculation. The procedures used to quantify potential loss exposures under such special circumstances are called **stress tests** [571]. A stress test measures the loss that could be experienced if a set of factors are exogenously specified.

31.4.3 VaR for Fixed-Income Securities

In contrast to stock prices, bond prices tend to move together because much of the movement is systematic, the common factor being the interest rate. For this reason, bond portfolio management does not require that the portfolios be well diversified. Instead, a few bonds of differing maturities can usually hedge the price fluctuations in any single bond or portfolio of bonds [91].

Duration (see Section 4.2) and key rate duration (see Section 27.5) were used to quantify the interest rate exposure of fixed-income portfolios and securities. A VaR methodology can also be based on duration. If S refers to the initial yield of a fixed-income instrument with duration D, the VaR for a long position in the instrument is $1.645 \times \sigma SD$ for a 95% confidence level. As before, VaR analysis requires parameters for the *actual* term structure dynamics. Simulation-based VaR usually conducts factor analysis before the actual simulation [804]. Three orthogonal factors seem to be sufficient (see Subsection 19.2.5).

The variance–covariance approach to the VaR is more complicated [470, 720]. First, an "equivalent" portfolio of standard zero-coupon bonds is obtained for each bond (this is called **cash flow mapping**). Then the historical volatility of spot rates and the correlations between them are used to construct a 95% confidence interval for the dollar return. It is difficult to apply this approach to securities with embedded options, however.

Additional Reading

In 1952, Markowitz and Roy independently published their papers that mark the era of modern portfolio theory [642, 644]. Our presentation of modern portfolio theory

was drawn from [317, 623]. General treatment of mean–variance can be found in [643]. See [82, 174, 403, 760, 862] for additional information on beta and [81, 332, 399] for the issue of expected return and risk. The framework of modern portfolio theory can be applied to real estate [389, 407]. See [28] for modern portfolio theory's applicability in Japan. One of the reasons cited for the choice of standard deviation as the measure of risk is that it is easier to work with than the alternatives [799]. An interesting theory from experimental psychology, the **prospect theory**, says that an investor is much more sensitive to reductions in wealth than to increases, which is called **loss aversion** [533, 534]. Spreads can be used to profit from such behavioral "biases" [180]. Consult [592] for an approach beyond mean–variance analysis.

See [826] for development of the utility function. Optimization theory is discussed in [278, 687]. See [346, 470, 567, 646, 693] for additional information on portfolio insurance. Consult [317, 623, 673] for investment performance evaluation; there seems to be no consistent performance for mutual funds [634]. See [96, 360, 799] for security analysis, such as technical analysis and fundamental analysis, and [132] on market timing.

Consult [314] for the VaR of derivatives, [484, 691, 720, 771] for VaR when returns are not normally distributed, [8] for managing VaR using puts, [293, 390, 464, 484, 720, 804, 857] for the VaR of fixed-income securities. The **Cornish–Fisher expansion** is useful for correcting the skewness in distribution during VaR calculations [470, 522].

NOTES

1. Sometimes we use variance of return as the measure of risk for convenience.
2. "Rate of return" and "return" are often used interchangeably as only single-period analysis is involved.
3. Markowitz's Ph.D. dissertation was initially voted down by Friedman on the grounds that "It's not math, it's not economics, it's not even business administration." See [64, p. 60].
4. The S&P 500 Index often serves as the proxy for the market portfolio. **Index funds** are mutual funds that attempt to duplicate a stock market index. Offered in 1975 under the name of Vanguard Index Trust, the Vanguard 500 Index Fund is the first index mutual fund. It tracks the S&P 500 and became the largest mutual fund in April of 2000.
5. Sharpe sent the paper in 1962 to the *Journal of Finance*, but it was quickly rejected [64, p. 194].
6. Dynamic strategies rely on the market to supply the needed liquidity. The Crash of 1987 and the Russian and the LTCM crises of 1998 demonstrate that such liquidity may not be available at times of extreme market movements [308, 654]. As prices began to fall during the Crash of 1987, portfolio insurers sold stock index futures. This activity in the futures market led to more selling in the cash market as program traders attempted to arbitrage the spreads between the cash and futures markets. Further price declines led to more selling by portfolio insurers, and so on [567, 647].
7. The CAPM does not require that the residuals ϵ_i be uncorrelated; see Eq. (31.2).
8. See Eqs. (15.12) and (24.10) for similar arguments.
9. Chebyshev's inequality in Exercise 13.3.10, part (2), supplies the intuition.
10. "Return" means price change ΔS or simple rate of return $\Delta S/S$. This is consistent with the stochastic differential equation $\Delta S = S\mu\,\Delta t + S\sigma\sqrt{\Delta t}\,\xi$ when Δt is small.

CHAPTER THIRTY-TWO

Software

Test everything. Hold on to the good.
I Thessalonians 5:21

32.1 Web Programming

The software for the book is Web-centric in that a reasonably updated Web browser is all that is needed to run it. The software is transferred to the user when clicked; no installation is necessary. As the collection of software expands, *The Capitals* Web page will reflect that. This new medium of software distribution excels the traditional way of bundling software with each book in a floppy disk or a CD-ROM [705].

The Web promises to be a platform that is independent of the computer's operating system and hardware. That means a program or document written in HTML (Hypertext Markup Language [167]) can be run everywhere, and the author is relieved from worrying about the potentially infinite number of computer systems that may access the code. As of now, this promise is not yet fully realized. To start with, the same document or program often elicits different behaviors from browsers of various companies or even browsers of the same family but with different versions. Browsers may implement only a subset of the standard plus a few nonstandard features. Additional complications are the possible versions of Java shipped with the browser and window systems running on top of the operating system. Fortunately, in most cases these problems are either inessential or can be avoided by upgrading the browser and not using nonstandard features.

32.2 Use of *The Capitals* Software

Open *The Capitals* at

www.csie.ntu.edu.tw/~lyuu/capitals.html.

See Fig. 32.1 for a typical look. Now click on any program to run it. For example, click on "mortgage" to generate the calculator in Fig. 32.2. Many financial problems can be solved by using two or more programs simultaneously. For example, one can run "spot & forward rates from coupon bonds" to calculate the spot rates. Then copy these rates into "seq. CMO pricer (vector)" to price CMOs. As another example, consider pricing CMO tranches at 10 years before maturity. We can run

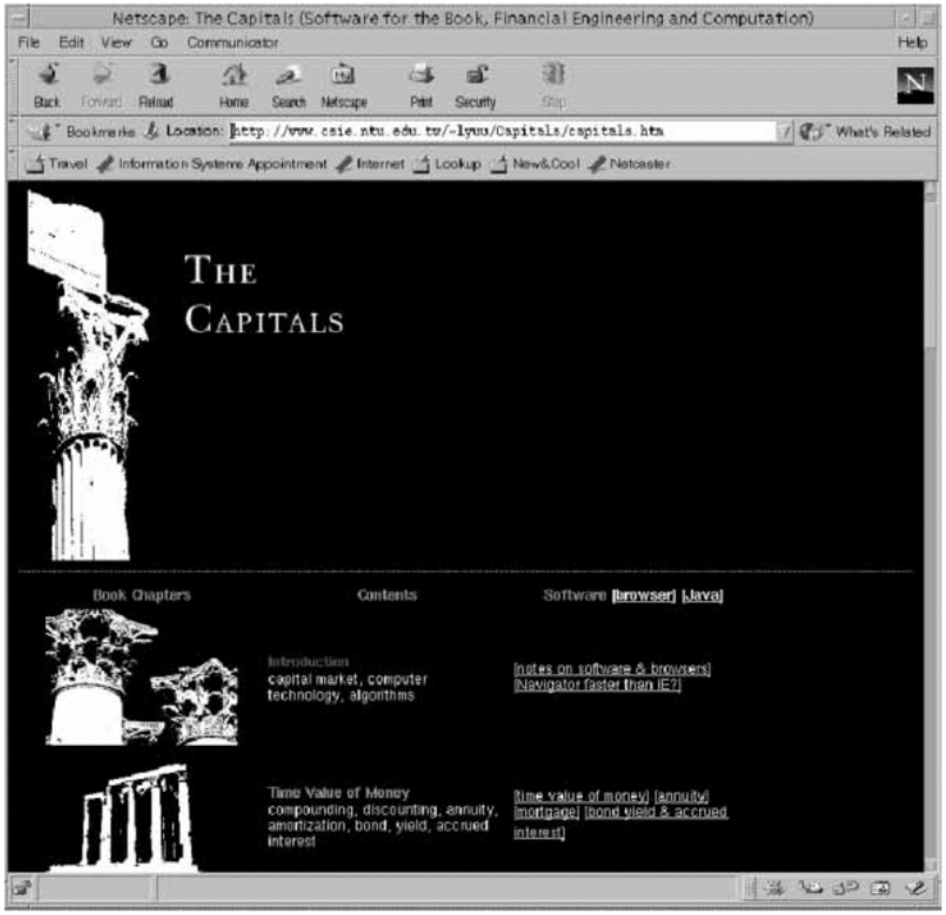


Figure 32.1: *The Capitals* page. This page is displayed by the Netscape browser in a Unix environment. The looks may differ from browser to browser and can be altered by changing the browser's settings.

"seq. CMO (vector)" to derive the tranches' remaining balances at that time. Then plug those numbers as original principals into "seq. CMO pricer (vector)" with 10 years remaining.

Some programs can be applied to situations not originally intended. For instance, CMO programs can be used for individual mortgages by allocating the entire principal to the first tranche; the cash flow of an SMBS can be tabulated by "pool P&I tabulator (vector)," and so on.

The following guidelines are recommended to run *The Capitals* software smoothly.

- Netscape Navigator 4.0 or higher, or Microsoft's Internet Explorer 4.0 or higher.
- Enable Java.
- Use Java 1.1.4 or higher.

Check "notes on software & browsers" for additional information.

The programs at *The Capitals* are written in JavaScript [357] and Java [356, 467].[1] Because it is the Java-enabled Web browser that interprets and executes the code, user interaction and processing are offloaded to the user's computer. This client/server

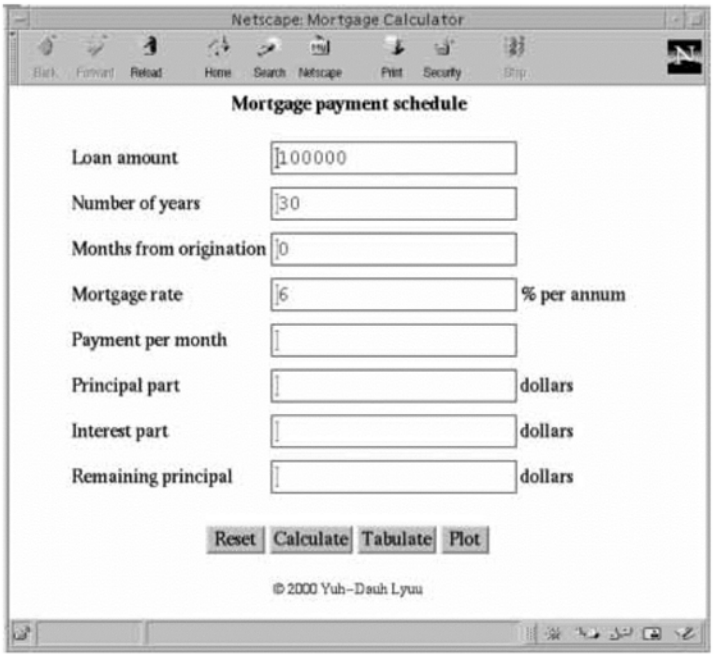


Figure 32.2: Mortgage calculator.

architecture is more efficient than having many clients' computing and interaction tasks running on the server, slowing it down for everybody. The unstated assumptions have been that the user's computer is reasonably powerful and the network is reasonably fast [719]. Java programs that run in a Web browser are called **Java applets** (see Fig. 32.3) [264].

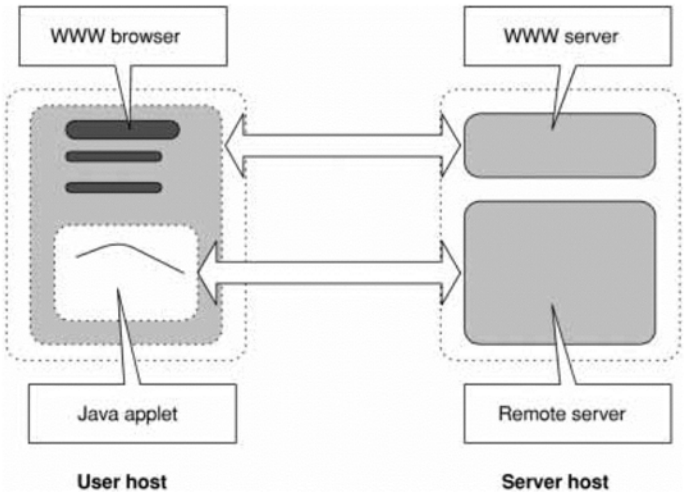


Figure 32.3: Java client/server architecture on the Web. The Java programming language was released by Sun in May of 1995. It promised platform-independent client/server software systems [356, 416].

32.3 Further Topics

Some computation-intensive tasks can take advantage of parallel processing for much faster performance. A good example is Monte Carlo MBS pricing. It starts by breaking the job into several tasks, each of which, on a different computer, simulates a fraction of the interest rate scenarios and calculates the average price. The averages are then collected to obtain the overall average price. Note that once the work has been divided, no communication among the tasks is needed before the collection stage. Good speed-ups have been obtained [528, 601, 794, 892, 893]. In contrast, a task that cannot be structured in such a way as to limit the amount of communication, hence dependency, among the tasks will not result in good performance [588]. Only computation-intensive problems are worthwhile to parallelize.

The Web technology is young and evolving quickly. Users and developers have been willing to tolerate many annoyances because they are witnessing something that promises to change the way society works. If the history of the auto industry is any guide, it will take decades for the technology to mature. Fortunately, thanks to the efforts and dedication of many corporations and computer professionals, the Web has become the most important and easy-to-use platform for software.

NOTE

1. JavaScript is not Java, but it has a similar syntax.

> There is nothing new to be discovered in physics now [1900].
>
> William Thomson (aka Lord Kelvin) (1824–1907)

CHAPTER THIRTY-THREE

Answers to Selected Exercises

More questions may be easier to answer than just one question.
Imre Lakatos (1922–1974), *Proofs and Refutations*

CHAPTER 2

Exercise 2.2.2: (1) Recall that $\sum_{i=1}^{n} i = n(n+1)/2$. (2) Use $\sum_{i=1}^{n} i^2 = (2n^3+3n^2+n)/6$. (3), (4) Use $\sum_{i=0}^{k} 2^i = 2^{k+1}-1$. (5) Use Euler's summation formula [461, p. 18],

$$\int_a^{b+1} g(x)\,dx \le \sum_{i=a}^{b} g(i) \le \int_{a-1}^{b} g(x)\,dx.$$

CHAPTER 3

Exercise 3.1.1: It is sufficient to show that $g(m) \equiv (1+\frac{1}{m})^m$ is an increasing function of m. Note that

$$g'(m) = g(m)\left[\ln\left(1+\frac{1}{m}\right) - \frac{1}{m+1}\right].$$

We can show the expression within the brackets to be positive by differentiating it with respect to m.

An alternative approach is to expand $g(m)$ and $g(m+1)$ as polynomials of $x \equiv 1/m$ using the binomial expansion. It is not hard to see that every term in $g(m+1)$, except the one of degree $m+1$ (which $g(m)$ does not have), is at least as large as the term of the same degree in $g(m)$. This approach does not require calculus.

Exercise 3.1.2: Monthly compounding, i.e., 12 times per annum. This can be verified by noting that $18.70/12 = 1.5583$.

Exercise 3.1.3: (1) The computing power's growth function is $(1.54)^n$, where n is the number of years since 1987. The equivalent continuous compounding rate is 43.18%. Because the memory capacity has quadrupled every 3 years since 1977, the function is $4^{n/3} = (1.5874)^n$, where n is the number of years since 1977. The equivalent continuous compounding rate is 46.21%. (2) It is $(500{,}000/300{,}000)^{1/4} - 1 \approx 13.6\%$. Data are from [574].

Exercise 3.2.1:

$$\text{PV} = \sum_{i=0}^{nm-1} C\left(1+\frac{r}{m}\right)^{-i} = C\,\frac{1-\left(1+\frac{r}{m}\right)^{-nm}}{(r/m)}\left(1+\frac{r}{m}\right).$$

Exercise 3.3.1: We derived Eq. (3.8) by looking forward into the future. We can also derive the same relation by looking back into the past: Right after the kth payment, the remaining principal is the value of the original principal minus the value of all the payments made to date, which is exactly