Carnegie Mellon University

Generalization Error Bounds for Time Series

A Dissertation Submitted to the Graduate School in Partial Fulfillment of the Requirements

for the degree

Doctor of Philosophy

in

Statistics

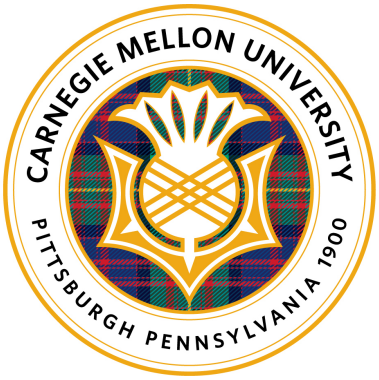
by

Daniel Joseph McDonald

Department of Statistics

Carnegie Mellon University

Pittsburgh, PA 15213



April 6, 2012

ABSTRACT

In this thesis, I derive generalization error bounds — bounds on the expected inaccuracy of the predictions — for time series forecasting models. These bounds allow forecasters to select among competing models, and to declare that, with high probability, their chosen model will perform well — without making strong assumptions about the data generating process or appealing to asymptotic theory. Expanding upon results from statistical learning theory, I demonstrate how these techniques can help time series forecasters to choose models which behave well under uncertainty. I also show how to estimate the β-mixing coefficients for dependent data so that my results can be used empirically. I use the bound explicitly to evaluate different predictive models for the volatility of IBM stock and for a standard set of macroeconomic variables. Taken together my results show how to control the generalization error of time series models with fixed or growing memory.

ACKNOWLEDGMENTS

I would especially like to thank my advisors Cosma Shalizi and Mark Schervish. Without their insight, encouragement, probing questions, excitement, and suggestions, this thesis would have turned out much differently. I am very proud of this work, and I owe both of them a tremendous amount for helping it to become something worthwhile. I would like to thank Dave DeJong for helping to come up with a topic, encouraging the work, and helping me to understand the goals and motivations of macroeconomic researchers. I also thank Larry Wasserman and Alessandro Rinaldo for providing lots of useful insight, keeping me focused, and serving on my committee.

Throughout my graduate career, I have received lots of beneficial encouragement and advice from other faculty members during collaborations, courses, office hours, and informal meetings concerning specific statistical issues I have encountered, life as a PhD student, teaching issues, job market decisions, and life in general. I want to thank Matt Harrison, Surya Tokdar, Chris Genovese, and Larry Wasserman for getting me through my first-semester courses, my first real statistics classes, and inadvertently keeping me enrolled in the program when I thought I was in over my head. I want to thank Jay Kadane for working with me on ADA, challenging my assumptions, and teaching me to thoroughly question everything. I also want to thank Howard Seltman, Joel Greenhouse, Rebecca Nugent, Chad Schafer, Kathryn Roeder, Peter Freeman, Steve Fienberg, Bill Eddy, Valerie Ventura, Oded Meyer, and John Lehoczky for their advice in the classroom and out over the last five years.

This entire process would have been less fun and less worthwhile without the friends that I have made here, especially the first year lunch group — Stacey

Ackerman-Alexeeff, Tracy Sweet, Jionglin Wu, Chris Neff, Dancsi Percival, Anne-Sophie Charest, James Sharpnack, and Darren Homrighausen — as well as Bethany Schwing, Laura Jekel, and George Loewenstein. I also want to thank Michael Sato and Ana Isabel Zorro for being there when I needed someone to talk to.

Finally, I want to thank my parents for their encouragement and confidence.

NOTATION

This thesis uses many different probability measures and σ-fields in different contexts. I list many of the symbols I will use in this document to avoid confusion.

$\mathbb{P}$ — The probability distribution of a single random variable Z or the pair (X, Y); Used only in the context of independence

$\mathbb{P}^n$ — The joint distribution of n independent random variables; The n-fold product measure $\prod_{i=1}^{n} \mathbb{P} = \mathbb{P}^n$

$\mathbb{P}_1$ — The probability distribution of a single random variable Y_1 generated by a dependent process

$\mathbb{P}_{\mathcal{C}}$ — The restriction of a probability measure to a specific σ-field $\mathcal{C}$; also appearing as $\mathbb{P}_t$ if it is the restriction to the σ-field generated by the dependent random variable at time t

$\mathbf{Y}_{i:j}$ — The sequence of dependent random variables $Y_i, \ldots, Y_j$

$\sigma_{i:j}$ — The σ-field generated by the sequence $\mathbf{Y}_{i:j}$

$\mathbb{P}_{i:j}$ — The joint distribution of the sequence $\mathbf{Y}_{i:j}$; A measure on $\sigma_{i:j}$

$\mathbb{P}_{i:j \otimes k:l}$ — The joint distribution of the sequences $\mathbf{Y}_{i:j}$ and $\mathbf{Y}_{k:l}$

$\mathbb{P}_{i:j} \otimes \mathbb{P}_{k:l}$ — The product measure on two sequences of dependent random variables; Under this distribution $\mathbf{Y}_{i:j} \perp\!\!\!\perp \mathbf{Y}_{k:l}$

$\mathbf{Y}_{\infty}$ — An infinite sequence of dependent random variables; Equivalent to $\mathbf{Y}_{-\infty:\infty}$

σ_{∞} — The σ-field generated by $\mathbf{Y}_{\infty}$

$\mathbb{P}_\infty$ — The infinite dimensional distribution on σ_∞

$\mathbb{E}_\mathbb{P}$ — The expected value with respect to the probability distribution $\mathbb{P}$; i.e. $\mathbb{E}_\mathbb{P}[g] := \int g d\mathbb{P}$; When obvious, this may be written as $\mathbb{E}_X$ for the expected value taken with respect to the distribution of the random variable X or simply as $\mathbb{E}$

ACRONYMS

AIC — Akaike Information Criterion (see Akaike [1])

AR — Autoregressive

ARMA — Autoregressive Moving Average

BIC — Bayesian Information Criterion

DSGE — Dynamic Stochastic General Equilibrium

ERM — Empirical Risk Minimization (or Minimizer)

FRB — Federal Reserve Board (of Governors)

GARCH — Generalized Autoregressive Conditional Heteroscedasticity

GDP — Gross Domestic Product

IID — Independent and Identically Distributed

MCM — Multi-Country Model

MCMC — Markov Chain Monte Carlo

MPS — MPS comes from the three collaborative centers where the model was developed by Franco Modigliani, Albert Ando, and Frank de Leeuw of MIT, the University of Pennsylvania, and the Social Science Research Council respectively.

RBC — Real Business Cycle

SRM — Structural Risk Minimization

SV — Stochastic Volatility

VAR — Vector Autoregressive

VARMA — Vector Autoregressive Moving Average

VC — Vapnik-Chervonenkis

RELATED PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011a), "Estimated VC dimension for risk bounds," submitted for publication, arXiv:1111.3404.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011b), "Estimating β-mixing coefficients," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, eds. G. Gordon, D. Dunson, and M. Dudík, vol. 15, JMLR W&CP, arXiv:1103.0941.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011c), "Estimating β-mixing coefficients via histograms," submitted for publication, arXiv:1109.5998.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011d), "Generalization error bounds for stationary autoregressive models,"
arXiv:1103.0942.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011e), "Risk bounds for time series without strong mixing," submitted for publication, arXiv:1106.0730.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011f), "Time series forecasting: model evaluation and selection using nonparametric risk bounds," in preparation.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

THESIS OVERVIEW AND MOTIVATION

1

INTRODUCTION

Researchers in statistics and machine learning have spent countless hours over the past century on a quest to find estimators for huge varieties of applied problems. Sometimes the goal is to be able to describe the unknown distribution from which the data arose so as to inform scientists, government officials, or the general public about phenomena of interest — the age of the universe, the costs and benefits of universal health care, or the effect of coffee or soda on colon cancer [106]. Other times, the goal is more ambitious: to predict the future. Huge numbers of smart people devote time and energy to anticipating stock market fluctuations, marketing experts recommend products consumers are unable to live without, and geneticists wish to learn if different sequences of DNA can predict an individual's susceptibility to a particular disease. When making predictions from data, forecasters are concerned with two important questions: (1) given a new data point, what is the mapping from predictors to responses; and (2) are the predictions any good. I will briefly sketch the manner in which this analysis typically procedes with more details to come in Chapter 3.

To address the first question, suppose that predictors live in some space $\mathcal{X}$ and responses live in another space $\mathcal{Y}$. Many methods of finding a mapping $f : \mathcal{X} \to \mathcal{Y}$ amount to choosing a class of candidate functions $\mathcal{F}$ and then picking the best one by minimizing a loss function $\ell(Y, f(X))$ which measures the performance of f. If $\mathcal{F}$ contains linear functions and $\ell(Y, f(X)) = (Y - f(X))^2$, then this procedure

amounts to ordinary least squares. Using the negative log likelihood as the loss function yields maximum likelihood estimation.

One possible answer to the second question requires the choice of functions $f \in \mathcal{F}$ which minimize the loss in expectation. This quantity,

$$R(f) = \mathbb{E}_{\mathbb{P}}[\ell(Y, f(X))], \tag{1.1}$$

is the generalization error, or risk, of the prediction algorithm. Unfortunately, while it is natural to want this to be small, one usually cannot hope to minimize it. The expectation is taken with respect to the joint distribution of the predictors and the response which also affects the learning algorithm's choice of the optimal f. While assumptions can be made about the true data generating process in order to calculate the risk, this tactic negates the most useful quality of prediction through risk minimization: the risk measures the cost of mistakes with respect to the *unknown* data generating process. Researchers' inability to calculate the risk exactly has engendered work deriving upper bounds for the generalization error.

Besides providing guarantees regarding how bad the expected cost of misprediction can be, generalization error bounds are useful for other reasons. Good bounds allow for straightforward model comparisons without making assumptions on the data generating process in contrast to likelihood based methods. Bounds can also be used to demonstrate the optimality of particular prediction algorithms, bounding the best-case performance with respect to the least favorable data generating process, i.e. minimaxity. Sometimes they can be used to naturally construct well behaved learning algorithms through regularization. These possibilities motivate the calculation of generalization error bounds not only as a theoretical and philosophical indulgence but also for improved applied research.

Prediction problems in statistics and machine learning often assume that training data are independent and identically distributed, but most interesting data are dependent and heterogeneous. Consequently, many existing risk bounds are

useless for some types of problems, especially those involving time series data such as economic forecasting.

Some generalization error bounds are known for time series, but they are not useful for the learning algorithms which often arise in the economic forecasting literature for two reasons. First, most generalization error bounds require that the loss function be bounded, which is inconvenient in a regression setting. Second, existing generalization error bounds for time series rely on quantifying the decay of dependence in the data generating process. While positing known rates for the decay of dependence leads to clean theoretical results, this knowledge is sadly unavailable in reality. Thus it is necessary to be able to estimate these rates from the data. In this thesis, I will (a) derive generalization error bounds for state space models, (b) develop methods for estimating the dependence behavior from the data so that the bound is useful, and (c) use the bounds to evaluate and compare existing economic forecasting methods.

The motivation for this thesis comes mainly from time series forecasting particularly for macroeconomics. In Chapter 2, I discuss the history and current methodology of macroeconomic forecasting, its relationship to standard time series models, and the benefits of generalization error bounds for risk analysis and model selection relative to current practice. Chapter 3 discusses methods for controlling generalization when the data are independent and identically distributed, while Chapter 4 describes how to introduce dependence. The remainder of the thesis presents theoretical results necessary to justify calculating generalization error bounds for macroeconomic time series models as well as a few examples of the use of these bounds in practice.

2

ECONOMIC FORECASTING

2.1 MOTIVATION AND LITERATURE REVIEW

Generalization error bounds are provably reliable, probabilistically valid, non-asymptotic tools for characterizing the predictive ability of forecasting models. The theory underlying these methods is fundamentally concerned with choosing particular functions out of some class of plausible functions so that the resulting predictions will be accurate with high probability. While many of these results are useful only in the context of classification problems (i.e., predicting binary variables) and for independent and identically distributed (IID) data, this thesis shows how to adapt and extend these methods to time series models so that economic and financial forecasting techniques can be evaluated rigorously. In particular, these methods control the expected accuracy of future predictions based on finite quantities of data. This allows for immediate model comparisons without appealing to asymptotic results or making strong assumptions about the data generating process in stark contrast to AIC and similar model selection criteria frequently employed in the literature.

2.2 HISTORY

Between 1975 and 1982, the art of macroeconomic forecasting underwent fairly dramatic changes. Until 1976, macroeconomic forecasting concentrated mainly on the use of "reduced-form" statistical characterizations of the economy. Forecasters ran regressions of data on other data and lags of the data and postulated that certain time-series should be related to others in different ways. The first large scale macroeconomic model of this type arose in 1966 with the implementation of the MPS model.[1] The MPS model consisted of around 60 estimating equations and identities used to forecast economic time series on a quarterly basis (think GDP, unemployment, productivity, inflation, etc.). The MPS model and its counterpart the Multi-Country Model (MCM) which contained some 200 equations developed into the FRB/US and its counterpart FRB/WORLD used since 1996 as the main economic forecasting tools at the Federal Reserve Board of Governors (see Brayton et al. [11] for an overview of this history and Brayton and Tinsley [10] for a discussion of the current version). The two models implemented today each use over 300 equations to forecast both the US economy and that of our trade partners.

These large scale macro models stand in stark contrast to the methods of forecasting used by most academic economists. In 1976, Lucas [61] issued a critique of reduced-form models which became very famous. His basic argument was that the sorts of statistical relationships exploited by the large scale macroeconomic models are useless for evaluating the impact of policy decisions, because without any behavioral theory underlying the construction of the models, only observed associations, the policies are bound to change the estimated parameters. In other words, the policy actions that modelers were attempting to evaluate were endogenous to the model, not exogenous.

1 MPS comes from the three collaborative centers where the model was developed by Franco Modigliani, Albert Ando, and Frank de Leeuw of MIT, the University of Pennsylvania, and the Social Science Research Council respectively.

Kydland and Prescott [55] marked the beginning of the use of dynamic stochastic general equilibrium (DSGE) models to combat this critique. Rather than focusing on statistical relationships, economists aimed to build models for the entire economy that are driven by individuals making decisions based on their preferences. In these models, consumers make decisions based on behavioral "deep" parameters like risk tolerance, the labor-leisure tradeoff, and the depreciation rate that are viewed as independent of things like government spending or monetary policy. The result is a heavily theoretical class of models for forecasting macroeconomic time series and the effects of policy interventions that tries to rely on some notion of behavior — it incorporates individuals making optimal choices under uncertainty based on their preferences. Unlike MPS, the FRB/US model tries to incorporate some of these ideas, but its behavioral equations do not arise from optimization the way a DSGE model's do. The remainder of this section discusses dynamic stochastic general equilibrium models and a simpler, more widely used, structural model as well as the state space representations used to estimate them.

2.3 DYNAMIC STOCHASTIC GENERAL EQUILIBRIUM MODELS

Kydland and Prescott [55] model the aggregate economy by considering a single household, intended to be an infinitely long-lived agent representative of all households and firms. The model which I discuss here, the canonical DSGE model, is called the Real Business Cycle (RBC) model. It takes the form of the following optimization problem.

1. The household seeks to maximize U, the expected discounted flow of utility derived from consumption and leisure

$$\max_{c_t, l_t} U = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t, l_t). \tag{2.1}$$

Here the $\mathbb{E}_0$ is the expectation conditional on information available at time $t = 0$, β is the discount factor on future utility, and $u(\cdot)$ is an instantaneous utility function. Future consumption and leisure are both functions of a random variable.

2. The household can produce "goods" y_t using the production function $g(\cdot)$

$$y_t = z_t g(k_t, n_t), \tag{2.2}$$

where k_t and n_t are capital and labor and z_t is a random process referred to as a technology shock or Solow residual in honor of Solow [91].

3. The remaining equations are as follows:

$$1 = n_t + l_t \tag{2.3}$$

$$y_t = c_t + i_t \tag{2.4}$$

$$k_{t+1} = i_t + (1 - \delta)k_t \tag{2.5}$$

$$\ln z_t = (1 - \rho)\ln \bar{z} + \rho \ln z_{t-1} + \epsilon_t \tag{2.6}$$

$$\epsilon_t \overset{iid}{\sim} N(0, \sigma_\epsilon^2). \tag{2.7}$$

Together, these say that the time spent between labor and leisure in each period must sum to 1, all output (income) is spent on consumption c_t or saved (invested) i_t, capital tomorrow is equal to investment today plus the depreciated capital stock, and the log of the technology shock z_t follows an AR(1) process.

The only uncertainty in the model stems from random innovations to technology. Thus, it is clear that this model has various implications: fiscal policy does nothing, monetary policy does nothing, asset prices do nothing, etc. More elaborate models generally account for most of these things. A published model at the Federal Reserve Board of Governors uses differentiated goods, differentiated firms, sticky prices (they do not adjust immediately), and monetary policy (see

Edge et al. [32]). The current version also adds in trade with 20 countries and uses nearly 100 different time-series. Whether any of this additional flexibility is useful for forecasting is unknown.

Estimation of these models is non-trivial and currently an area of active research. All methods involve solving the constrained optimization problem and then turning the result into a state space model through either linear or non-linear approximation. The parameters are estimated through method of moments techniques called calibration after Kydland and Prescott [55] or likelihood analysis as in Sargent [82]. In either case, the resulting estimated model can be used for forecasting. By nature, a DSGE is a nonlinear system of expectational difference equations, and so estimating the parameters is nontrivial. Likelihood methods typically procede by finding a linear approximation using Taylor expansions and the Kalman filter, though increasingly complex nonlinear methods are now an object of intense interest. See for instance Fernández-Villaverde [34], DeJong and Dave [19] or Dejong et al. [23]

2.4 OTHER METHODS

The DSGE framework relies on specifying and solving a dynamic stochastic optimization problem, using approximation techniques so that it may be mapped into state space form, and then estimating the parameters. This is typically a long and complicated process involving differential equations, linear algebra, and nonlinear maximization. A much simpler, reduced form, tool for forecasting is the vector autoregression or VAR. In its most straightforward version, a VAR(p) is specified as

$$\mathbf{x}_t = \mathbf{B}_1\mathbf{x}_{t-1} + \mathbf{B}_2\mathbf{x}_{t-2} + \cdots + \mathbf{B}_p\mathbf{x}_{t-p} + \mathbf{e}_t \tag{2.8}$$

where $\mathbf{x}_t$ is a $k \times 1$ observation vector, $\mathbf{B}_i$ is a $k \times k$ matrix, and $\mathbf{e}_t$ is a $k \times 1$ mean zero noise term. The model is simple to fit using multiple least squares and gives straightforward forecasts for the time series of interest. However, the number

of parameters grows rapidly: ignoring the covariance structure, the VAR(p) has pk^2 parameters. Since n is necessarily small in economic forecasting problems (usually consisting only of quarterly data since 1950), researchers frequently put a default prior called the Minnesota prior on the $\mathbf{B}_i$ to avoid overfitting. While this regularization results in better out of sample forecasting performance when compared to unrestricted models [26], generalization error bounds may lead to improved learning algorithms.

Many less common economic forecasting methods can be reexpressed in state space form. Dynamic factor models like those in Kim and Nelson [48] are trivially state space models. The turning point forecasting models such as DeJong et al. [20] or Wildi [101] also have state space representations.

Economic forecasting is just one application for time series analysis by state space models. Missile tracking applications as well as other linear dynamical systems motivated the path breaking work of Kalman [47]. More recently, state space models have been used for robot soccer by Ruiz-del Solar and Vallejos [81], to study the effects of a seat belt law on traffic accidents in Great Britain by Harvey and Durbin [42], and for neural decoding applications as in Koyama et al. [54].

2.5 STATE SPACE MODELS

The most general form of a state space model is characterized by the observation equation, the state transition equation, and an initial distribution for the state:

$$y_t = \varphi_O(x_t, \epsilon_t) \tag{2.9}$$

$$x_{t+1} = \varphi_S(x_t, \eta_t) \tag{2.10}$$

$$x_1 \sim \mathbb{P}, \tag{2.11}$$

where ϵ_t are η_t are marginally independent and identically distributed (IID) as well as mutually independent. The vector $\{y_t\}_{t=1}^T$ is observed, and the goal is

to make inferences for the unobserved states $\{x_t\}_{t=1}^T$ as well as any parameters characterizing φ_O, φ_S, and the distributions of ϵ_t and η_t.

In the case where φ_O and φ_S are linear with ϵ_t and η_t normally distributed, the Kalman filter can be used along with maximum likelihood or Bayesian methods to derive closed form solutions for the conditional distributions of the states as well as the parameters of interest given data. However, in many applications, researchers are not so lucky. For nonlinear or non-Gaussian models, approximate solutions exist using the particle filter and its derivatives (see for example Kitagawa [49, 50] and Doucet et al. [27] for an exposition of the particle filter and Koyama et al. [54] and DeJong et al. [22] for improvements).

2.6 MODEL EVALUATION METHODS IN TIME SERIES AND ECONOMICS

There are many ways to estimate the generalization error. Traditionally, time series analysts have performed model selection by a combination of empirical risk minimization, more-or-less quantitative inspection of the residuals — e.g., the Box-Ljung test; see [87] — and penalties like AIC. In many applications, however, what really matters is prediction, and none of these techniques, including AIC, really work to control generalization error, especially for mis-specified models. Empirical cross-validation is a partial exception, but it is tricky for time series; see Racine [77] and references therein.

In economics, forecasters have long recognized the difficulties with these methods of risk estimation, preferring to use a pseudo-cross validation approach instead. This technique chooses a prediction function using the initial portion of a data set and evaluates its performance on the remainder. Athanasopoulos and Vahid [2] compare the predictive accuracy of VAR models with vector autoregressive moving average (VARMA) models using a training sample spanning the 1960s and 1970s and a test set spanning the 1980s and 1990s. Faust and Wright [33] compare forecasts produced by the Federal Reserve called "Greenbook forecasts" with

the predictions of various other atheoretical methods, however they ignore periods of high volatility such as 1979–1983. Christoffel et al. [14] compare the New Area Wide Model for Europe with a Bayesian VAR, a random walk, and sample means. The forecasts are evaluated during the relatively stable period of the late 1990s and early 2000s, and the models are updated yearly, giving pseudo-out-of-sample monthly forecasts. Similarly, Del Negro et al. [24] reestimate DSGE-VARs recursively based on rolling 30 year samples before forecasting two year periods between 1985 and 2000. Smets and Wouters [90] compare DSGE models with Bayesian VARs over a similar period. Edge and Gurkaynak [31] argue that DSGEs (as well as statistical or judgmental methods) perform poorly at predicting GDP or inflation. Numerous other examples of model selection and evaluation through pseudo-out-of-sample forecast comparions can be found throughout the literature.

Procedures such as these provide approximate solutions to the problem of estimating the generalization error, but they can be heavily biased toward overfitting — giving too much credence to the observed data — and hence underestimating the true risk for at least three reasons. First, the held out data, or test set, is used to evaluate the performance of competing models despite the fact that it was already partially used to build those models. For instance, the structures of both exogenous and endogenous variables in DSGEs are partially constructed so as to lead to predictive models which fit closely to the most recent macroeconomic phenomena. The recent housing and financial crises have precipitated numerous attempts to enrich existing DSGEs with mechanisms designed to enhance their ability to predict just such a crisis (see for example Goodhart et al. [40], Gerali et al. [38] and Gertler and Karadi [39]). Testing the resulting models on recent data therefore leads to overconfident declarations about a particular model's forecasting abilities. Second, the distributions of the test set and the data used to estimate the model may be different, i.e., it may be that the observed phenomena reflect only a small sampling of possible phenomena which could occur. Models which forecast well during the early 2000s were typically fit and evaluated using numerous occurrences of stable economic conditions, but few were built to also perform well

during periods of crisis. Finally, large departures from the normal course of events such as the recessions in 1980–82 and periods before 1960 are often ignored as in [33]. While these periods are considered rare and perhaps unpredictable, models which are robust to these sorts of tail events will lead to more accurate predictions in future times of turmoil.

2.7 RISK BOUNDS FOR ECONOMICS AND TIME SERIES

In contrast to the model evaluation techniques typically employed in the literature, generalization error bounds provide rigorous control over the predictive risk as well as reliable methods of model selection. They are robust to wide classes of data generating processes and are finite-sample rather than asymptotic in nature. In a broad sense, these methods give confidence bounds which are constructed based on concentration of measure results rather than appeals to asymptotic normality. The results are easy to understand and can be reported to policy makers interested in the quality of the forecasts. Finally, the results are agnostic about the model's specification: it does not matter if the model is wrong, the parameters have interpretable economic meaning, or whether the estimation of the parameters is performed only approximately (linearized DSGEs or MCMC), one can still make strong claims about the ability of the model to predict the future.

The meaning of such results for forecasters, or for those whose scientific aims center around prediction of empirical phenomena, is plain: they provide objective ways of assessing how good their models really are. There are, of course, other uses for scientific models: for explanation, for the evaluation of counterfactuals (especially, in economics, comparing the consequences of different policies), and for welfare calculations. Even in those cases, however, one must ask *why this model rather than another?*, and the usual answer is that the favored model gets the structure at least approximately right. Empirical evidence for structural correctness, in turn, usually takes the form of an argument from empirical success: *it would be*

very surprising if this model fit the data so well when it got the structure wrong. My results, which directly address the inference from past data-matching to future performance, are thus relevant even to those who do not aim at prediction as such.

Part II

EXISTING THEORY

3

STATISTICAL LEARNING THEORY

The goal of this thesis is to control the risk of predictive models, i.e., their expected inaccuracy on new data from the same source as that used to fit the model. In this chapter, I summarize the basic forms of these results in the literature, filling in what was only lightly sketched in Chapter 1.

3.1 THE TRADITIONAL SETUP

Consider predictors $X \in \mathcal{X}$ and responses $Y \in \mathcal{Y}$. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathcal{Y}$ which take predictors as inputs.

Define a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ which measures the cost of making poor predictions. Throughout this chapter I make the following assumption on the loss function.

Assumption A. $\forall f \in \mathcal{F}$

$$0 \leqslant \ell(y, y') \leqslant M < \infty. \tag{3.1}$$

Then, as in (1.1), I can define the risk of any predictor $f \in \mathcal{F}$.

Definition 3.1 (Risk or generalization error)**.**

$$R(f) := \int \ell(f(X), Y) d\mathbb{P} = \mathbb{E}_{\mathbb{P}}\big[\ell(f(X), Y)\big], \tag{3.2}$$

where $(X, Y) \sim \mathbb{P}$.

The risk or generalization error measures the expected cost of using f to predict Y from X given a new observation. Just to emphasize, the expectation is taken with respect to the distribution $\mathbb{P}$ of the test point (X, Y) which is independent of f; the risk is a deterministic function of f with all the randomness in the data averaged away.

Since the true distribution $\mathbb{P}$ is unknown, so is $R(f)$, but one can attempt to estimate it based on only the observed data. Suppose that I observe a random sample $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ so that $(X_i, Y_i) \overset{iid}{\sim} \mathbb{P}$, i.e. $D_n \sim \mathbb{P}^n$. Define the *training error* or *empirical risk* of f as follows.

Definition 3.2 (Training error or empirical risk)**.**

$$\widehat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i). \tag{3.3}$$

In other words, the in-sample training error, $\widehat{R}_n(f)$, is the average loss over the actual training points. It is easy to see that, because the training data D_n and the test point (X, Y) are IID, then given some fixed function f (chosen independently of the sample D_n),

$$\widehat{R}_n(f) = R(f) + \gamma_n(f), \tag{3.4}$$

where $\gamma_n(f)$ is a mean-zero noise variable that reflects how far the training sample departs from being perfectly representative of the data-generating distribution. Here I should emphasize that $\widehat{R}_n(f)$ is random through the training sample D_n. By the law of large numbers, for such fixed f, $\gamma_n(f) \to 0$ as $n \to \infty$, so, with enough data, one has a good idea of how well any given function will generalize to new data.

However, one is rarely interested in the performance of a single function f without adjustable parameters fixed for them in advance by theory. Rather, researchers are interested in a class of plausible functions $\mathcal{F}$, possibly indexed by some possibly infinite dimensional parameter $\theta \in \Theta$, which I refer to as a model. One function (one particular parameter point) is chosen from the model class by mini-

mizing some criterion function. Maximum likelihood, Bayesian maximum *a posteriori*, least squares, regularized methods, and empirical risk minimization (ERM) all have this flavor as do many other estimation methods. In these cases, one can define the empirical risk minimizer for an appropriate loss function ℓ.

Definition 3.3 (Empirical risk minimizer[1])**.**

$$\widehat{f} := \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{R}_n(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \left(R(f) + \gamma_n(f) \right). \tag{3.5}$$

It is important to note that $\widehat{f}$ is random and measurable with respect to the empirical risk process $\widehat{R}_n(f)$ for $f \in \mathcal{F}$. Choosing a predictor $\widehat{f}$ by empirical risk minimization (tuning the adjustable parameters so that $\widehat{f}$ fits the training data well) conflates predicting future data well (low $R(\widehat{f})$, the true risk) with exploiting the accidents and noise of the training data (large negative $\gamma_n(\hat{f})$, finite-sample noise). The true risk of $\widehat{f}$ will generally be bigger than its in-sample risk precisely because I picked it to match the data well. In doing so, $\widehat{f}$ ends up reproducing some of the noise in the data and therefore will not generalize well. The difference between the true and apparent risk depends on the magnitude of the sampling fluctuations:

$$R(\widehat{f}) - \widehat{R}_n(\widehat{f}) \leqslant \sup_{f \in \mathcal{F}} |\gamma_n(f)| = \Gamma_n(\mathcal{F}) \ . \tag{3.6}$$

In (3.6), $R(\widehat{f})$ is random and measurable with respect to $\widehat{f}$.

The main goal of statistical learning theory is to control $\Gamma_n(\mathcal{F})$ while making minimal assumptions about the data generating process — i.e. to provide bounds on over-fitting. Using more flexible models (allowing more general functional forms or distributions, adding parameters, etc.) has two contrasting effects. On the one hand, it improves the best possible accuracy, lowering the minimum of the true risk. On the other hand, it increases the ability to, as it were, memorize noise for any fixed sample size n. This qualitative observation — a generalization of the bias-variance trade-off from basic estimation theory — can be made use-

1 I will sometimes use the more complete notation $\widehat{f}_{erm}$

fully precise by quantifying the complexity of model classes. A typical result is a confidence bound on Γ_n (and hence on the over-fitting), which says that with probability at least $1-\eta$,

$$\Gamma_n(\mathcal{F}) \leqslant \Phi(\Lambda(\mathcal{F}), n, \eta) \,, \tag{3.7}$$

where $\Lambda(\cdot)$ is some suitable measure of the complexity of the model $\mathcal{F}$. To give specific forms of $\Phi(\cdot)$, I need to show that, for a particular f, $R(f)$ and $\widehat{R}_n(f)$ will be close to each other for any fixed n without knowledge of the distribution of the data. Furthermore, I need the complexity, $\Lambda(\mathcal{F})$, to claim that $R(f)$ and $\widehat{R}_n(f)$ will be close, not only for a particular f, but uniformly over all $f \in \mathcal{F}$. Together these two results will allow me to show, despite little knowledge of the data generating process, how bad the $\widehat{f}$ which I choose will be at forecasting future observations.

3.2 CONCENTRATION

The first step to controlling the difference between the empirical and expected risk is to develop concentration results for fixed functions. These finite sample laws of large numbers control the difference between random variables and their expectations. To illustrate what this means, consider a random variable Z with probability distribution $\mathbb{P}$ such that $\mathbb{P}(a \leqslant Z \leqslant b) = 0$. First I state the following Lemma without proof which bounds the moment generating function of Z.

Lemma 3.4 (Equation 4.16 in [45])**.**

$$\mathbb{E}[\exp\{s(Z - \mathbb{E}[Z])\}] \leqslant \exp\left\{\frac{s^2(b-a)}{8}\right\}. \tag{3.8}$$

Then, I can combine the bound on the moment generating function with Markov's inequality to obtain Hoeffding's inequality [45].

Theorem 3.5 (Hoeffding's inequality). *Let* $Z_1, \ldots, Z_n$ *be IID random variables each with distribution* $\mathbb{P}$ *such that,* $\mathbb{P}(a \leqslant Z \leqslant b) = 0$ *and product measure* $\mathbb{P}^n = \prod_{i=1}^n \mathbb{P}$. *Then,*

$$\mathbb{P}^n(|\overline{Z} - \mathbb{E}[\overline{Z}]| \geqslant \epsilon) \leqslant 2\exp\left\{-\frac{2n\epsilon^2}{(b-a)^2}\right\}. \tag{3.9}$$

To provide some intuition for the general topic of concentration bounds, I provide the following proof.

Proof. First, I use Lemma 3.4 to bound the moment generating function of $\overline{Z} - \mathbb{E}[\overline{Z}]$:

$$\mathbb{E}[\exp\{s(\overline{Z} - \mathbb{E}[\overline{Z}])\}] = \prod_{i=1}^n \mathbb{E}\left[\exp\left\{\frac{s}{n}(Z - \mathbb{E}[Z])\right\}\right] \tag{3.10}$$

$$\leqslant \prod_{i=1}^n \exp\left\{\frac{s^2(b-a)^2}{8n^2}\right\} \tag{3.11}$$

$$= \exp\left\{\frac{s^2(b-a)^2}{8n}\right\}. \tag{3.12}$$

Therefore I can use Markov's inequality and the moment generating function bound:

$$\mathbb{P}^n(\overline{Z} - \mathbb{E}[\overline{Z}] > \epsilon) = \mathbb{P}^n\left(\exp\{s(\overline{Z} - \mathbb{E}[\overline{Z}])\} \geqslant \exp\{s\epsilon\}\right) \tag{3.13}$$

$$\leqslant \frac{\mathbb{E}\left[\exp\{s(\overline{Z} - \mathbb{E}[\overline{Z}])\}\right]}{\exp\{s\epsilon\}} \tag{3.14}$$

$$\leqslant \exp\{-s\epsilon\}\exp\left\{\frac{s^2(b-a)^2}{8n}\right\}. \tag{3.15}$$

This holds for all $s > 0$, so I can minimize the right hand side in s. This occurs for $s = 4n\epsilon/(b-a)^2$. Plugging in gives

$$\mathbb{P}^n(\overline{Z} - \mathbb{E}[\overline{Z}] > \epsilon) \leqslant \exp\left\{-\frac{2n\epsilon^2}{(b-a)^2}\right\}. \tag{3.16}$$

Exactly the same argument holds for $\mathbb{P}^n(\overline{Z} - \mathbb{E}[\overline{Z}] < -\epsilon)$, so by a union bound, I have the result. ∎

Of course, this bound holds for the average of independent bounded random variables, which is not necessarily that interesting. Often, one wants concentration for some well-behaved function of independent random variables. One route to concentration for functions is via McDiarmid's inequality.

Theorem 3.6 (McDiarmid Inequality [63]). *Let $Z_1, \ldots, Z_n$ be IID random variables taking values in a set A. Suppose that the function $f : A^n \to \mathbb{R}$ is $\mathbb{P}^n$-measurable and satisfies*

$$|f(\mathbf{z}) - f(\mathbf{z}')| \leqslant c_i \tag{3.17}$$

whenever the vectors $\mathbf{z}$ and $\mathbf{z}'$ differ only in the i^{th} coordinate. Then for any $\epsilon > 0$,

$$\mathbb{P}^n(f - \mathbb{E}[f] > \epsilon) \leqslant \exp\left\{-\frac{2\epsilon^2}{\sum c_i^2}\right\}. \tag{3.18}$$

In later chapters, I will need both of these results. In the remainder of this section, I show how to obtain concentration for the training error around the risk for two different choices of the random variables Z_i. This will lead to two different ways of controlling Γ_n and hence the generalization error of prediction functions.

3.3 CONTROL BY COUNTING

Suppose I let Z_i be the loss of the i^{th} training point for some fixed function f. Then by Hoeffding's inequality, Theorem 3.5,

$$\mathbb{P}^n(|R(f) - \widehat{R}_n(f)| \geqslant \epsilon) \leqslant 2\exp\left\{-\frac{2n\epsilon^2}{M^2}\right\}. \tag{3.19}$$

This result is quite powerful, it says that the probability of observing data which will result in a training error much different from the expected risk goes to zero exponentially with the size of training set. The only assumption necessary was that $0 \leqslant \ell(y, y') \leqslant M$. In fact, even this assumption can be removed and replaced with some moment conditions.

Of course (3.19) holds for the single function f chosen independently of the data. Instead, I want a similar result to hold simultaneously over all functions $f \in \mathcal{F}$ and in particular, the $\widehat{f}$ choosen using the training data, i.e., I wish to bound $\mathbb{P}^n\left(\sup_{f\in\mathcal{F}} |R(f) - \widehat{R}_n(f)| > \epsilon\right)$.

For "small" models, one can simply count the number of functions in the class and apply the union bound. Suppose that $f_1, \ldots, f_N \in \mathcal{F}$. Then

$$\mathbb{P}^n\left(\sup_{1\leqslant i\leqslant N} |R(f_i) - \widehat{R}_n(f_i)| > \epsilon\right) \leqslant \sum_{i=1}^{N} \mathbb{P}^n\left(|R(f_i) - \widehat{R}_n(f_i)| > \epsilon\right) \tag{3.20}$$

$$\leqslant N \exp\left\{-\frac{2n\epsilon^2}{M^2}\right\}, \tag{3.21}$$

by Theorem 3.5. Most interesting models are not small in this sense, but using an appropriate way of "counting", similar results can be derived.

There are many ways of "counting" the number of effectively distinct functions. A direct, functional analysis, approach leads to covering numbers [76, 75] which partitions functions $f \in \mathcal{F}$ into equivalence classes under some metric. Instead, I focus on a measure which is both intuitive and powerful: Vapnik-Chervonenkis (VC) dimension [96, 97].

VC dimension starts as a notion about a collection of sets.

Definition 3.7 (Shattering)**.** *Let $\mathbb{U}$ be some (infinite) set and S a subset of $\mathbb{U}$ with finite cardinality. Let $\mathcal{C}$ be a family of subsets of $\mathbb{U}$. One says that $\mathcal{C}$* shatters *S if for every $S' \subseteq S$, $\exists C \in \mathcal{C}$ such that $S' = S \cap C$.*

Essentially, $\mathcal{C}$ can shatter a set of points if it can pick out every subset of points in S. This says somehow that $\mathcal{C}$ is very complicated or flexible. The cardinality of the largest set S that can be shattered by $\mathcal{C}$ is the known as its VC dimension.

Definition 3.8 (VC dimension)**.** *The* Vapnik-Chervonenkis (VC) dimension *of a collection $\mathcal{C}$ of subsets of $\mathbb{U}$ is*

$$\mathrm{VCD}(\mathcal{C}) := \sup\{|S| : S \subseteq \mathbb{U} \text{ and } S \text{ is shattered by } \mathcal{C}\}. \tag{3.22}$$

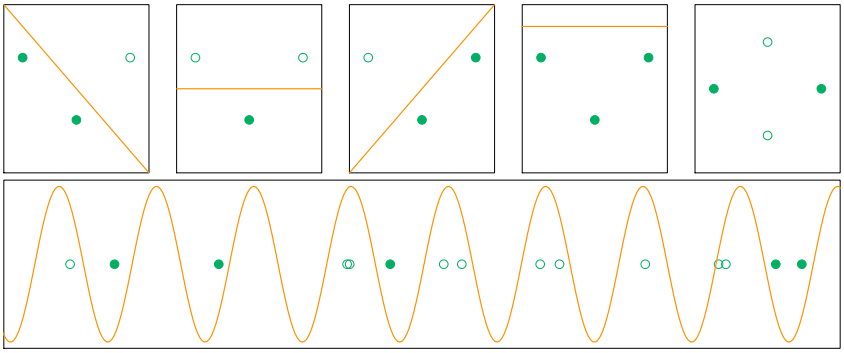


Figure 1: The top panel demonstrates shattering sets of points with linear functions. Here, the points are contained in $\mathbb{R}^2$ so it is possible to shatter three point sets but not four point sets. The bottom panel shows how to shatter points using $\mathcal{F} = \{x \mapsto \sin(\omega x) : \omega \in \mathbb{R}\}$.

Using VC dimension to measure the capacity of function classes is straightforward. Define the indicator function $\mathbb{1}_A(x)$ to take the value 1 if $x \in A$ and 0 otherwise. Suppose that $f \in \mathcal{F}$, $f : \mathbb{U} \to \mathbb{R}$. Then to each f associate the set

$$C_f = \{(u, b) : \mathbb{1}_{(0,\infty)}(f(u) - b) = 1, \quad u \in \mathbb{U}, \quad b \in \mathbb{R}\} \tag{3.23}$$

and associate to $\mathcal{F}$ the class $\mathcal{C}_{\mathcal{F}} := \{C_f : f \in \mathcal{F}\}$.

VC dimension is well understood for some function classes. For instance, if $\mathcal{F} = \{\mathbf{u} \mapsto \boldsymbol{\gamma} \cdot \mathbf{u} : \mathbf{u}, \boldsymbol{\gamma} \in \mathbb{R}^p\}$ then $\text{VCD}(\mathcal{F}) = p + 1$, i.e. it is the number of free parameters in a linear regression plus 1. It does not always have such a nice correspondence with the number of free parameters however. The classic example of such an incongruity is the model $\mathcal{F} = \{u \mapsto \sin(\omega u) : u, \omega \in \mathbb{R}\}$, which has only one free parameter, but $\text{VCD}(\mathcal{F}) = \infty$. This result follows if one can show that for every positive integer J and every binary sequence $(r_1, \ldots, r_J)$, there exists a vector $(u_1, \ldots, u_J)$ such that $\mathbb{1}_{[0,1]}(\sin(\omega u_i)) = r_i$. If I choose $u_i = 2\pi 10^{-i}$, then one can show that taking $\omega = \frac{1}{2}\left(\sum_{i=1}^{J}(1 - r_i)10^i + 1\right)$ solves the system of equations. Both of these examples are shown in Figure 1.

Given a model $\mathcal{F}$ such that $\text{VCD}(\mathcal{F}) = h$, I can control the risk over the entire model. This is one of the milestones of statistical learning theory

Theorem 3.9 (Vapnik and Chervonenkis [98]). *Suppose that* $\mathrm{VCD}(\mathcal{F}) = h$ *and that Assumption A holds. Then,*

$$\mathbb{P}^n\left(\sup_{f\in\mathcal{F}}|R(f)-\widehat{R}_n(f)|>\epsilon\right)\leqslant 4\mathrm{GF}(2n,h)\exp\left\{-\frac{n\epsilon^2}{M^2}\right\}, \tag{3.24}$$

where $\mathrm{GF}(n,h)\leqslant\exp\{h(\log n/h+1)\}$.

The proof of this theorem has a similar flavor to the union bound argument given in (3.20)–(3.21). Essentially, $\mathrm{GF}(n,h)$ counts the effective number of functions in $\mathcal{F}$, i.e., how many can be told apart using only n observations.

This theorem has two corollaries. The first is to give a bound on the expected difference between the training error and the risk for any $f\in\mathcal{F}$. The second is a high probability bound for the expected risk.

Corollary 3.10.

$$\mathbb{E}_{\mathbb{P}^n}\left[\sup_{f\in\mathcal{F}}|R(f)-\widehat{R}_n(f)|\right]=O\left(\sqrt{\frac{h\log n/h}{n}}\right). \tag{3.25}$$

Proof. Define $Z=\sup_{f\in\mathcal{F}}|R(f)-\widehat{R}_n(f)|$, $k_1=4\mathrm{GF}(2n,h)$, and $k_2=1/M^2$. Then,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^n}[Z^2]&=\int_0^\infty\mathbb{P}^n(Z^2>\epsilon)d\epsilon=\int_0^s\mathbb{P}^n(Z^2>\epsilon)d\epsilon+\int_s^\infty\mathbb{P}(Z^2>\epsilon)d\epsilon && (3.26)\\
&\leqslant s+\int_s^\infty\mathbb{P}^n(Z^2>\epsilon)d\epsilon && (3.27)\\
&=s+\int_s^\infty\mathbb{P}^n(Z>\sqrt{\epsilon})d\epsilon && (3.28)\\
&\leqslant s+k_1\int_s^\infty e^{-k_2n\epsilon}d\epsilon && (3.29)\\
&=s+\frac{k_1e^{-k_2ns}}{k_2n}. && (3.30)
\end{aligned}$$

Set $s = \frac{\log k_1}{n k_2}$. Then,

$$\mathbb{E}_{\mathbb{P}^n}[Z] \leqslant \sqrt{\mathbb{E}_{\mathbb{P}^n}[Z^2]} \leqslant \sqrt{\frac{\log k_1}{n k_2} + \frac{1}{n k_2}} \tag{3.31}$$

$$= M\sqrt{\frac{1 + \log 4 GF(2n, h)}{n}} \tag{3.32}$$

which gives the result. ■

Corollary 3.11. *Let $\eta > 0$. Then simultaneously for all $f \in \mathcal{F}$, with probability at least $1 - \eta$,*

$$R(f) \leqslant \widehat{R}_n(f) + M\sqrt{\frac{\log GF(2n, h) + \log 4/\eta}{n}}. \tag{3.33}$$

Proof. Set

$$\eta = 4 GF(2n, h) \exp\left\{-\frac{n\epsilon^2}{M^2}\right\}, \tag{3.34}$$

and solve for ϵ in (3.34) to get the result. ■

The probability statement in Corollary 3.11 is with respect to the joint distribution generating the training data, $\mathbb{P}^n$.

The right side of (3.33) is very similar to standard model selection criteria like AIC or BIC. If one assumes a normal likelihood, then the training error behaves like the negative loglikelihood term while the remainder is the penalty. Here however, the bound holds with high probability despite lack of knowledge of $\mathbb{P}$, and it has nothing to do with asymptotics: it holds for any n. Just like AIC, the penalty term $M\sqrt{\frac{1+\log 4 GF(2n,h)}{n}}$ goes to 0 as $n \to \infty$, and, since Corollary 3.11 holse for all $f \in \mathcal{F}$, it holds in particular for $\widehat{f}$.

3.4 CONTROL BY SYMMETRIZATION

Rather than looking at the losses at each training point and trying to count all the functions in $\mathcal{F}$, one can instead investigate the random variable

$$\Psi_n := \sup_{f \in \mathcal{F}} \left(R(f) - \widehat{R}_n(f) \right). \tag{3.35}$$

Concentrating Ψ_n about its mean follows directly via Theorem 3.6.

Lemma 3.12. *Let Assumption A hold. Then,*

$$\mathbb{P}^n(|\Psi_n - \mathbb{E}[\Psi_n]| > \epsilon) \leqslant 2 \exp\left\{ -\frac{2n\epsilon^2}{M^2} \right\}. \tag{3.36}$$

Proof. Changing one pair (x_i, y_i) can change Ψ_n by no more than $|\ell(y_i, f(x_i))|/n \leqslant M/n$. So by McDiarmid's inequality,

$$\mathbb{P}^n(\Psi_n - \mathbb{E}[\Psi_n] > \epsilon) \leqslant \exp\left\{ -\frac{2n\epsilon^2}{M^2} \right\}. \tag{3.37}$$

Using the same logic

$$\mathbb{P}^n(-\Psi_n + \mathbb{E}[\Psi_n] < -\epsilon) \leqslant \exp\left\{ -\frac{2n\epsilon^2}{M^2} \right\}. \tag{3.38}$$

Taking a union bound gives the result. ∎

One way to handle $\mathbb{E}_{\mathbb{P}^n}[\Psi_n]$ is to use Corollary 3.10. But this is not the only way, and in fact is generally suboptimal. An alternative is to use *Rademacher Complexity* [52, 60, 80, 107, 5].

Definition 3.13 (Rademacher Complexity)**.** *The* empirical Rademacher complexity *of a function class $\mathcal{G}$ composed of functions $g : \mathcal{Z} \to \mathbb{R}$ for some set $\mathcal{Z}$ is*

$$\widehat{\mathfrak{R}}_n(\mathcal{G}) := 2\mathbb{E}_w \left[\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} w_i g(z_i) \right| \middle| (Z_1, \ldots, Z_n) \right], \tag{3.39}$$

where $\mathbf{w} = \{w_i\}_{i=1}^n$ *is a sequence of random variables, independent of each other and everything else, and equal to* $+1$ *or* -1 *with equal probability, and* $Z_1, \ldots, Z_n$ *are IID random variables taking values in the set* $\mathcal{Z}$ *with marginal distributions* $\mathbb{P}$. *The* Rademacher complexity *is*

$$\mathfrak{R}_n(\mathcal{G}) := \mathbb{E}_{\mathbb{P}^n}\left[\widehat{\mathfrak{R}}_n(\mathcal{G})\right]. \tag{3.40}$$

Lemma 3.14.

$$\mathbb{E}_{\mathbb{P}^n}[\Psi_n] \leqslant \mathfrak{R}_n(\ell \circ \mathcal{F}), \tag{3.41}$$

where $\ell \circ \mathcal{F}$ *denotes the function class generated by composing the loss function* ℓ *with functions* $f \in \mathcal{F}$.

Proof.

$$\mathbb{E}_{\mathbb{P}^n}[\Psi_n] = \mathbb{E}_{\mathbb{P}^n}\left[\sup_{f\in\mathcal{F}}(R(f) - \widehat{R}_n(f))\right] \tag{3.42}$$

$$= \mathbb{E}_{\mathbb{P}^n}\left[\sup_{f\in\mathcal{F}}(\mathbb{E}_{\mathbb{P}^n}[\widehat{R}'_n(f)] - \widehat{R}_n(f))\right] \tag{3.43}$$

$$\leqslant \mathbb{E}_{\mathbb{P}^n\otimes\mathbb{P}^n}\left[\sup_{f\in\mathcal{F}}\widehat{R}'_n(f) - \widehat{R}_n(f)\right]. \tag{3.44}$$

where $\widehat{R}'_n(f)$ is based on a "ghost sample" $\{(X'_1, Y'_1), \ldots, (X'_n, Y'_n)\}$ — an imaginary sample from the same distribution, $\mathbb{P}^n$, as the original — which is independent of the original. Now by definition of R,

$$\mathbb{E}_{\mathbb{P}^n}[\Psi_n] \leqslant \mathbb{E}_{\mathbb{P}^n \otimes \mathbb{P}^n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (\ell(f(X'_i), Y'_i) - \ell(f(X_i), Y_i))\right] \tag{3.45}$$

$$= \mathbb{E}_{\mathbb{P}^n \otimes \mathbb{P}^n \otimes \mathbf{w}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} w_i(\ell(f(X'_i), Y'_i) - \ell(f(X_i), Y_i))\right] \tag{3.46}$$

$$\leqslant \mathbb{E}_{\mathbb{P}^n \otimes \mathbb{P}^n \otimes \mathbf{w}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} w_i \ell(f(X'_i), Y'_i)\right]$$

$$+ \mathbb{E}_{\mathbb{P}^n \otimes \mathbb{P}^n \otimes \mathbf{w}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} w_i \ell(f(X_i), Y_i)\right] \tag{3.47}$$

$$= 2\mathbb{E}_{\mathbb{P}^n \otimes \mathbf{w}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} w_i \ell(f(X_i), Y_i)\right]. \tag{3.48}$$

■

Using Rademacher complexity along with Lemma 3.12 gives the following generalization error bound.

Theorem 3.15. *For any $\eta > 0$ and any $f \in \mathcal{F}$, with probability at least $1 - \eta$,*

$$R(f) \leqslant \widehat{R}_n(f) + \mathfrak{R}_n(\ell \circ \mathcal{F}) + M\sqrt{\frac{\log 2/\eta}{2n}}. \tag{3.49}$$

Another benefit of Rademacher complexity is that it can be calculated empirically. One can use the empirical version in place of the expected Rademacher complexity with slight modifications to the risk bound.

Theorem 3.16. *For any $\eta > 0$ and any $f \in \mathcal{F}$, with probability at least $1 - \eta$,*

$$R(f) \leqslant \widehat{R}_n(f) + \widehat{\mathfrak{R}}_n(\ell \circ \mathcal{F}) + 3M\sqrt{\frac{\log 4/\eta}{2n}}. \tag{3.50}$$

Proof. Since changing one point of the sample changes $\widehat{\mathfrak{R}}_n(\ell \circ \mathcal{F})$ by at most $2M/n$, by McDiarmid's inequality

$$\mathbb{P}^n \left(\mathfrak{R}_n(\ell \circ \mathcal{F}) - \widehat{\mathfrak{R}}_n(\ell \circ \mathcal{F}) > \epsilon \right) \leqslant \exp\left\{ -\frac{n\epsilon^2}{2M^2} \right\}. \tag{3.51}$$

Therefore with probability $1 - \eta/2$,

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) \leqslant \widehat{\mathfrak{R}}_n(\ell \circ \mathcal{F}) + M\sqrt{\frac{2\log 1/\eta}{n}}. \tag{3.52}$$

Combining this result with Theorem 3.15 for a confidence parameter $\eta/2$ gives the result since

$$M\sqrt{\frac{2\log 1/\eta}{n}} + M\sqrt{\frac{\log 4/\eta}{2n}} \leqslant 3M\sqrt{\frac{\log 4/\eta}{2n}}. \tag{3.53}$$

■

Good control of $\mathbb{E}[\Psi_n]$ through the Rademacher complexity therefore implies good control of the generalization error. Rademacher complexity is easy to handle for wide ranges of learning algorithms using results in [5] and elsewhere. Support vector machines, kernel methods, and neural networks all have known Rademacher complexities. Furthermore, by applying Lipschitz composition arguments in [57], I need to deal only with the Rademacher complexity of the function class $\mathcal{F}$ rather that of the composition class $\ell \circ \mathcal{F}$. For loss functions ℓ which are ϑ-Lipschitz in their second argument with $\ell(0,0) = 0$, $\mathfrak{R}(\ell \circ \mathcal{F}) \leqslant 2\vartheta\mathfrak{R}(\mathcal{F})$.

3.5 CONCENTRATION FOR UNBOUNDED FUNCTIONS

The main issue with all the results in the previous two sections is that they require bounded loss functions. While in classification, as well as many other settings, this is an intuitively reasonable requirement, this fails for regression. The Rademacher complexity results cannot be extended to unbounded losses, as far as I know, because of the supremum over the function class. The result is that the Rademacher

complexity will always be infinite. The VC method however can be extended to unbounded losses. It simply requires bounding the relative difference between the expected and empirical risks rather than the absolute difference.[2] Similarly, it requires control of the moments of the loss rather than the loss itself.

Assumption B. *Assume that for all* $f \in \mathcal{F}$ *and some* $q > 2$,

$$1 \leqslant \frac{\left(\mathbb{E}_{\mathbb{P}}\left[(\ell(f(X), Y))^q\right]\right)^{1/q}}{R_n(f)} < M. \tag{3.54}$$

Assumption B is still quite general, allowing even some heavy tailed distributions while being more general than the bounded loss requirement. Furthermore, with slight adjustments (see [96, p. 198]), one can allow $1 < q \leqslant 2$. It should be noted that the lower bound is trivially true for any loss distribution.

Theorem 3.17 (Theorem 5.4 in Vapnik [96])**.** *Under Assumption B,*

$$\mathbb{P}^n\left(\sup_{f\in\mathcal{F}} \frac{R(f) - \widehat{R}_n(f)}{R(f)} > \epsilon\right) \leqslant 4 G_{\mathcal{F}}(2n, h) \exp\left\{-\frac{n\epsilon^2}{4\tau^2(q)M^2}\right\}, \tag{3.55}$$

where $\tau(q) = \sqrt[q]{\frac{1}{2}\left(\frac{q-1}{q-2}\right)^{q-1}}$.

This concentration result can also be turned into a risk bound, but the penalty is now multiplicative rather than additive.

Corollary 3.18. *For any* $\eta > 0$ *and any* $f \in \mathcal{F}$*, with probability at least* $1 - \eta$,

$$R(f) \leqslant \frac{\widehat{R}_n(f)}{(1 - \mathcal{E})_+}, \tag{3.56}$$

where

$$\mathcal{E} = 2M\tau(q)\sqrt{\frac{\log G_{\mathcal{F}}(2n, h) + \log 4/\eta}{n}} \tag{3.57}$$

and $(u)_+ = \max(u, 0)$.

2 It is possible that a similar method could be used to generalize the Rademacher complexity to unbounded loss functions. However, I am not aware of any such results in the literature.

3.6 SUMMARY

The concentration results in this chapter work well for independent data. To develop them, I first showed how fast averages concentrate around their expectations: exponentially fast in the size of the data. The second set of results generalizes from a single function to entire function classes. All of these results depend critically on the independence of the random variables, however for time series, I need to be able to handle dependent data.

4

INTRODUCING DEPENDENCE

In this chapter, I show how to move from IID data to dependent data. I will assume conditions of weak dependence. This step draws mainly on the notion of "mixing". Processes are said to be mixing if, as the separation between past and future grows, the events in the past and future approach independence. This idea is illustrated in Figure 2. As a increases, events in the past and future are more widely separated. If, as this separation increases, these events approach independence in some appropriate metric, then the process is said to be mixing.

Because time series data are dependent, the number of data points n in a sample exaggerates how much information the sample contains. Knowing the past allows forecasters to predict future data (at least to some degree), so actually observing those future data points gives less information about the underlying process than in the IID case. Thus, while in Theorem 3.5 the probability of large discrepancies between empirical means and their expectations decreases exponentially in the sample size, in the dependent case, the effective sample size may be much less than n, resulting in looser bounds. Knowing the distance from independence for some particular separation a of a mixing process allows me to determine the effective sample size $\mu < n$.

Figure 2: This figure illustrates "mixing". As a increases, events in the past and future are more widely separated. If, as this separation increases, these events approach independence in some appropriate metric, then the process is said to be mixing.

4.1 DEFINITIONS

Mixing essentially describes the asymptotic dependence behavior of a stochastic process. There are many different versions of mixing which require stronger or weaker conditions on the behavior of the process. For an overview of the strong mixing conditions, see Bradley [9]. These and many weaker versions are discussed in Dedecker et al. [18]. I will be mainly concerned with β-mixing.

Mixing starts fundamentally as a measure of dependence between σ-fields. Consider a standard probability space $(\Omega, \mathcal{S}, \mathbb{P})$ and any two sub-σ-fields $\mathcal{A}$ and $\mathcal{B} \subset \mathcal{S}$.

Definition 4.1 (β-dependence)**.**

$$\beta(\mathcal{A}, \mathcal{B}) := \|\mathbb{P}_{\mathcal{A}\cup\mathcal{B}} - \mathbb{P}_{\mathcal{A}} \otimes \mathbb{P}_{\mathcal{B}}\|_{TV}\,, \tag{4.1}$$

where $\mathcal{A} \cup \mathcal{B} := \{A \cup B : A \in \mathcal{A}, B \in \mathcal{B}\}$ *and* $\mathbb{P}_{\mathcal{C}}$ *denotes the restriction of* $\mathbb{P}$ *to the* σ*-field* $\mathcal{C}$*.*

This definition makes clear that β-dependence is essentially measuring the distance between the joint distribution and the product of the marginal distributions in total variation, i.e. the distance from independence.

While Definition 4.1 provides intuition, it is not the standard definition in the literature. The following Lemma shows the equivalence between Definition 4.1 and that in [9].

Proposition 4.2.

$$\beta(\mathcal{A},\mathcal{B}) = \sup \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|, \tag{4.2}$$

where the supremum is taken over all pairs of finite partitions $\{A_1,\ldots,A_I\}$ *and* $\{B_1,\ldots,B_J\}$ *of* Ω *such that* $A_i \in \mathcal{A}$ *and* $B_j \in \mathcal{B}$ *for each* i *and* j.

In the time series setting, one is interested mainly in the dependence between past and future. This leads to specific choices for the σ-fields. To fix notation, let $\mathbf{Y}_\infty := \{Y_t\}_{t=-\infty}^{\infty}$ be a sequence of random variables where each Y_t is a measurable function from a probability space $(\Omega_t, \mathcal{S}_t, \mathbb{P}_t)$ into a measurable space $\mathcal{Y}$. A block of this random sequence will be written $\mathbf{Y}_{i:j} \equiv \{Y_t\}_{t=i}^{j}$ where i and j are integers, and may be infinite. I use similar notation for the sigma fields generated by these blocks and their joint distributions. In particular, $\sigma_{i:j}$ will denote the sigma field generated by $\mathbf{Y}_{i:j}$, and the joint distribution of $\mathbf{Y}_{i:j}$ will be denoted $\mathbb{P}_{i:j}$.

There are many equivalent definitions of β-mixing (see for instance Doukhan [28], or Bradley [9] as well as Meir [65] or Yu [105]), however the most intuitive is that given in Doukhan [28] which has the framework of Definition 4.1.

Definition 4.3 (β-mixing)**.** *For each* $a \in \mathbb{N}$ *and any* $t \in \mathbb{Z}$*, the* β-mixing coefficient*, or* coefficient of absolute regularity*,* β_a*, is*

$$\beta_a := \sup_t \|\mathbb{P}_{-\infty:t} \otimes \mathbb{P}_{t+a:\infty} - \mathbb{P}_{-\infty:t \otimes t+a:\infty}\|_{TV}\,, \tag{4.3}$$

where $\|\cdot\|_{TV}$ *is the total variation norm. A stochastic process is said to be* absolutely regular*, or* β-mixing*, if* $\beta_a \to 0$ *as* $a \to \infty$.

Loosely speaking, Definition 4.3 says that the coefficient β_a measures the total variation distance between the joint distribution of random variables separated by a time units and a distribution under which random variables separated by a time units are independent. This definition makes clear that a process is β-mixing if the joint probability of events approaches the product of their marginal probabil-

ities as those events become more separated in time, i.e., that $\mathbf{Y}$ is asymptotically independent.

Another characterization, which is occasionally useful, comes from Meir [65].

Proposition 4.4. *The* β*-mixing coefficient,* β_a*, is given by*

$$\beta_a = \sup_t \mathbb{E}_{\mathbb{P}_{-\infty:t}} \sup_{B \in \sigma_{t+a}^{\infty}} |\mathbb{P}_{t+a:\infty}(B \mid \sigma_{-\infty}^{t}) - \mathbb{P}_{t+a:\infty}(B)|. \tag{4.4}$$

The inclusion of the supremum over t in front of the total variation operator gives the greatest generality, however, I will consider only stationary processes.

Definition 4.5 (Stationarity)**.** *A sequence of random variables* $\mathbf{Y}_\infty$ *is* stationary *when all its finite-dimensional distributions are invariant over time: for all* t *and all non-negative integers* i *and* j*, the random vectors* $\mathbf{Y}_{t:(t+i)}$ *and* $\mathbf{Y}_{(t+j):(t+i+j)}$ *have the same distribution.*

Stationarity does not imply that the random variables Y_t are independent across time, rather that the unconditional distribution of Y_t is constant in time. For completeness, I present here a lemma giving the form of the β-mixing under stationarity.

Lemma 4.6. *For stationary professes, the* β-mixing coefficient,

$$\beta_a = \|\mathbb{P}_{-\infty:0} \otimes \mathbb{P}_{a:\infty} - \mathbb{P}_{-\infty:0 \otimes a:\infty}\|_{TV}. \tag{4.5}$$

4.2 MIXING IN THE LITERATURE

Numerous results in the statistics literature rely on knowledge of mixing coefficients. While much of the theoretical groundwork for the analysis of mixing processes was laid years ago (cf. [102, 8, 30, 73, 3, 93, 104, 105]), recent work has continued to use mixing to prove interesting results about the analysis of time-series data. Non-parametric inference under mixing conditions is treated extensively in Bosq [7]. Baraud et al. [4] study the finite sample risk performance of

penalized least squares regression estimators under β-mixing. Kontorovich and Ramanan [53] prove concentration of measure results based on a notion of mixing defined therein which is related to the more common ϕ-mixing coefficients. Ould-Saïd et al. [72] investigate kernel conditional quantile estimation under α-mixing. Steinwart and Anghel [92] show that support vector machines are consistent for time series forecasting under a weak dependence condition implied by α-mixing. Asymptotic properties of nonparametric inference for time series under various mixing conditions are described in Liu and Wu [59]. Finally, Lerasle [58] proposes a block-resampling penalty for density estimation. He shows that the selected estimator satisfies oracle inequalities under both β- and τ-mixing.

Many common time series models are known to be β-mixing, and the rates of decay are known up to constant factors which involve the true parameters of the process. Among the processes for which such knowledge is available are ARMA models [68], GARCH models [12], and certain Markov processes — see Doukhan [28] for an overview of such results. Fryzlewicz and Subba Rao [37] derive upper bounds for the α- and β-mixing rates of non-stationary ARCH processes. To my knowledge, only Nobel [70] approaches a solution to the problem of actually estimating mixing rates (rather than the coefficients themselves) by giving a method to distinguish between different polynomial mixing rate regimes through hypothesis testing.

In addition to the processes known to be mixing, functions of these processes are β-mixing, as I show below. So if $\mathbb{P}_\infty$ could be specified by a dynamic factor model or DSGE or VAR, the observed data would be mixing since these processes are functions of mixing processes.

Lemma 4.7. *Let* $\mathbf{Y}_\infty$ *be stationary and* β*-mixing with coefficients* β_a, $a \in \mathbb{N}$. *Then, for a measurable function* h, $h(\mathbf{Y}_\infty) := (\ldots, h(\mathbf{Y}_{0:d}), h(\mathbf{Y}_{1:d+1}), \ldots)$ *is* β*-mixing with coefficients bounded by* β_{a-d}.

Proof. By Equation 12 in Meir [65, §5], the sequence $(\ldots, \mathbf{Y}_{0:d}, \mathbf{Y}_{1:d+1}, \ldots)$ is β-mixing with coefficients bounded by β_{a-d}. Since h is measurable, then $\sigma(h(\mathbf{Y}_{i:j}))$ is a sub-σ-field of $\sigma_{i:j}$. The result follows from the Definition 4.1. ■

Knowledge of β_a allows me to determine the effective sample size of a given dependent data set $\mathbf{Y}_{1:n}$. In effect, having n dependent-but-mixing data points is like having $\mu < n$ independent ones. Once I determine the correct μ, I can use concentration results for IID data like those in Theorem 3.5 and Theorem 3.9 with small corrections. One possible way of determining μ is to use the technique of blocking described in the next section.

4.3 THE BLOCKING TECHNIQUE

To determine the effective sample size of a given data set, I use the method of blocking outlined by Yu [104, 105].[1] The purpose is to approximate a sequence of dependent variables by an IID sequence. Consider a sample $\mathbf{Y}_{1:n}$ from a stationary β-mixing sequence. Let m_n and μ_n be non-negative integers such that $2m_n\mu_n = n$. Now divide $\mathbf{Y}_{1:n}$ into $2\mu_n$ blocks, each of length m_n. Identify the blocks as follows:

$$U_j = \{Y_i : 2(j-1)m_n + 1 \leqslant i \leqslant (2j-1)m_n\}, \tag{4.6}$$

$$V_j = \{Y_i : (2j-1)m_n + 1 \leqslant i \leqslant 2jm_n\}. \tag{4.7}$$

As in Figure 3, let $\mathbf{U}$ be the entire sequence of odd blocks U_j (the first, third, fifth, etc. blocks), and let $\mathbf{V}$ be the sequence of even blocks V_j. Finally, let $\mathbf{U}'$ be a sequence of blocks which are independent of $\mathbf{Y}_{1:n}$ but such that each block has

1 This technique is actually much older and is often attributed to Bernstein from 1924

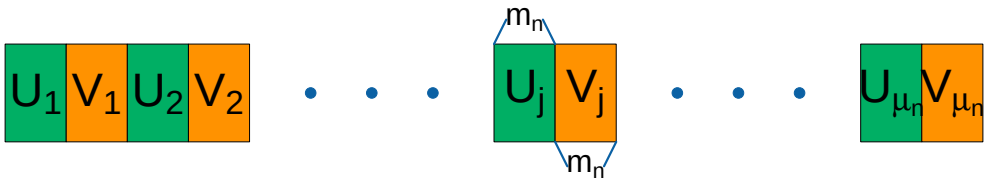


Figure 3: This figure shows how the blocks sequences **U** and **V** are constructed. There are μ "even" blocks U_j and μ "odd" blocks V_j. Each block is of length m_n.

the same distribution as a block from the original sequence. That is construct U'_j such that

$$\mathcal{L}(U'_j) = \mathcal{L}(U_j) = \mathcal{L}(U_1), \tag{4.8}$$

where $\mathcal{L}(\cdot)$ means the probability law of the argument. The blocks $\mathbf{U}'$ are now an IID block sequence, in that for integers $i, j \leqslant 2\mu_n$, $i \neq j$, $U'_i \perp\!\!\!\perp U'_j$, so standard results about IID random variables can be applied to these blocks. See [105] for a more rigorous analysis of blocking. Because the IID $\mathbf{U}'$ blocks are closely related to the dependent **U** blocks, I can use the former to approximate the latter using the following result.

Lemma 4.8 (Lemma 4.1 in [105])**.** *Let ϕ be an event in the σ-field generated by the block sequence* **U***. Then,*

$$|\widetilde{\mathbb{P}}(\phi) - \mathbb{P}^{\mu_n}_{1:m_n}(\phi)| \leqslant \beta_{m_n}(\mu_n - 1), \tag{4.9}$$

where $\widetilde{\mathbb{P}}$ is the joint distribution of the dependent block sequence **U***, and $\mathbb{P}^{\mu_n}_{1:m_n}(\phi)$ is the distribution with respect to the independent sequence,* $\mathbf{U}'$*.*

This lemma essentially gives a method for applying IID results to β-mixing data. Because the dependence decays as the separation between blocks increases, widely spaced blocks are nearly independent of each other. In particular, the difference between probabilities with respect to these nearly independent blocks and probabilities with respect to blocks which are actually independent can be controlled by the β-mixing coefficient.

Proof. I will demonstrate how to prove Lemma 4.8 in the simple case where $m_n = 1$ and $\mu_n = n/2$ to ease notation.

$$
\begin{aligned}
|\widetilde{\mathbb{P}}(\Phi) - \mathbb{P}^{n/2}(\Phi)| &\leqslant \left|\left|\widetilde{\mathbb{P}} - \mathbb{P}^{n/2}\right|\right|_{TV} && (4.10)\\
&\leqslant \left|\left|\widetilde{\mathbb{P}} - \mathbb{P} \times \mathbb{P}_{3,5,\ldots,n-1}\right|\right|_{TV} + \left|\left|\mathbb{P} \times \mathbb{P}_{3,5,\ldots,n-1} - \mathbb{P}^{n/2}\right|\right|_{TV} && (4.11)\\
&= \left|\left|\widetilde{\mathbb{P}} - \mathbb{P} \times \mathbb{P}_{3,5,\ldots,n-1}\right|\right|_{TV} + \left|\left|\mathbb{P}_{3,5,\ldots,n-1} - P^{n/2-1}\right|\right|_{TV} && (4.12)\\
&\leqslant \left|\left|\widetilde{\mathbb{P}} - \mathbb{P} \times \mathbb{P}_{3,5,\ldots,n-1}\right|\right|_{TV} + \|\mathbb{P}_{3,\ldots,n-1} - \mathbb{P} \times \mathbb{P}_{5,\ldots,n-1}\|_{TV}\\
&\quad + \left|\left|\mathbb{P} \times \mathbb{P}_{3,\ldots,n-1} - \mathbb{P}^{n/2-1}\right|\right|_{TV} && (4.13)\\
&= \left|\left|\widetilde{\mathbb{P}} - \mathbb{P} \times \mathbb{P}_{3,\ldots,n-1}\right|\right|_{TV} + \|\mathbb{P}_{3,\ldots,n-1} - \mathbb{P} \times \mathbb{P}_{5,\ldots,n-1}\|_{TV}\\
&\quad + \left|\left|\mathbb{P}_{5,\ldots,n-1} - \mathbb{P}^{n/2-2}\right|\right|_{TV} && (4.14)\\
&\leqslant \cdots (\text{induction}) \cdots\\
&\leqslant \left|\left|\widetilde{\mathbb{P}} - \mathbb{P} \times \mathbb{P}_{3,\ldots,n-1}\right|\right|_{TV} + \|\mathbb{P}_{3,\ldots,n-1} - \mathbb{P} \times \mathbb{P}_{5,\ldots,n-1}\|_{TV}\\
&\quad + \cdots + \left|\left|\mathbb{P}_{n-3,n-1} - \mathbb{P}^2\right|\right|_{TV}. && (4.15)
\end{aligned}
$$

By Lemma 4.6, each total variation term is bounded by β_1 and there are $(n/2-1)$ terms giving the result. ∎

In the time series literature, mixing rates (and therefore the coefficients themselves) are assumed to be known. As mentioned in Section 4.2, many particular process have rates which are known up to constant factors which depend on $\mathbb{P}_\infty$. However, in empirical work, one is faced with a particular data set generated by an unknown process. In the next chapter, I construct a method for estimating mixing coefficients from data without knowledge of $\mathbb{P}_\infty$.

Part III

RESULTS AND APPLICATIONS

5

ESTIMATING MIXING

5.1 INTRODUCTION

This chapter presents the first method for estimating the β-mixing coefficients for stationary time series data given a single sample path. The methodology can be applied to real data if one assumes that they were generated by some unknown β-mixing process. Additionally, it can be used on processes known to be mixing to determine exact mixing coefficients via simulation. Section 5.2 describes the estimator I propose. Section 5.3 presents a necessary preliminary result giving the L^1 convergence rates of histogram density estimators under β-mixing. Section 5.4 states and proves the consistency of our estimator as well as its behavior in finite samples. Section 5.5 demonstrates the performance of the estimator in some simulations.

5.2 THE ESTIMATOR

The first step to deriving my estimator depends on recognizing that the distribution of a finite sample depends only on finite-dimensional distributions. This leads to an estimator of a finite-dimensional version of β_a. Allowing the finite-dimension to increase to infinity with the size of the observed sample gives a consistent estimator of the infinite-dimensional coefficients.

For positive integers d, and a, define

$$\beta_a^d := \|\mathbb{P}_{-d:0} \otimes \mathbb{P}_{a:a+d} - \mathbb{P}_{-d:0 \otimes a:a+d}\|_{TV}. \tag{5.1}$$

Let $\widehat{p}^d$ be the d-dimensional histogram estimator of the joint density of d consecutive observations, and let $\widehat{p}_a^{2d}$ be the $2d$-dimensional histogram estimator of the joint density of two sets of d consecutive observations separated by a time points.

I estimate β_a^d from these two histograms. While it is clearly possible to replace histograms with other choices of density estimators (most notably kernel density estimators), histograms in this case are more convenient theoretically and computationally as explained more fully in Section 5.6. Briefly, the major benefit of histograms is that the total variation distance in Lemma 4.6 is computationally simple regardless of the dimension of the target densities (which will be allowed to approach infinity). If kernels are used instead, this integral will become increasingly difficult to calculate. Define

$$\widehat{\beta}_a^d := \frac{1}{2} \int \left| \widehat{p}_a^{2d} - \widehat{p}^d \otimes \widehat{p}^d \right| \tag{5.2}$$

I show in Theorem 5.5 that, by allowing $d = d_n$ to grow with n, this estimator will converge on β_a. This can be seen most clearly by bounding the ℓ^1-risk of the estimator with its estimation and approximation errors:

$$|\widehat{\beta}_a^{d_n} - \beta_a| \leqslant |\widehat{\beta}_a^{d_n} - \beta_a^{d_n}| + |\beta_a^{d_n} - \beta_a|. \tag{5.3}$$

The first term is the error of estimating β_a^d with a random sample of data. The second term is the non-stochastic error induced by approximating the infinite dimensional coefficient, β_a, with its d-dimensional counterpart, β_a^d. I thus begin by proving the doubly asymptotic convergence of histogram density estimators in Section 5.3, allowing both $d \to \infty$ and $n \to \infty$. Section 5.4 provides rates of convergence for Markov processes and proves consistency for generally β-mixing processes.

5.3 L^1 CONVERGENCE OF HISTOGRAMS

While convergence of density estimators is thoroughly studied in the statistics and machine learning literatures, I am not aware of any results on the L^1 convergence of histograms under β-mixing, which is what this estimator needs.[1] Therefore, I now prove this convergence.

Additionally, the dimensionality of the target density is analogous to the order of the Markov approximation. Therefore, the convergence rates I give are asymptotic in the bandwidth h_n which shrinks as n increases, but also in the dimension d_n which increases with n. Even under these asymptotics, histogram estimation in this sense is not a high dimensional problem. The dimension of the target density considered here is on the order of $\exp\{W(\log n)\}$, where $W(\cdot)$ is the Lambert W function,[2] a rate somewhere between $\log n$ and $\log\log n$.

Theorem 5.1. *If $\widehat{p}$ is the histogram estimator based on a (possibly vector valued) sequence $\mathbf{Y}_{1:n}$ from a β-mixing distribution with stationary density p, then for all $\epsilon > \mathbb{E}\left[\int |\widehat{p} - p|\right]$,*

$$\mathbb{P}_{1:n}\left(\int |\widehat{p} - p| > \epsilon\right) \leqslant 2\exp\left\{-\frac{\mu_n \epsilon_1^2}{2}\right\} + 2(\mu_n - 1)\beta_{m_n} \tag{5.4}$$

where $\epsilon_1 = \epsilon - \mathbb{E}\left[\int |\widehat{p} - p|\right]$.

To prove this result, I use the blocking method of Section 4.3 to transform the dependent β-mixing sequence into a sequence of nearly independent blocks. I then apply McDiarmid's inequality to the blocks to derive asymptotics in the bandwidth of the histogram as well as the dimension of the target density. Combining

1 Early papers on the L^∞ convergence of kernel density estimators (KDEs) include [103, 6, 88]; Freedman and Diaconis [36] look specifically at histogram estimators, and Yu [104] considered the L^∞ convergence of KDEs for β-mixing data and shows that the optimal IID rates can be attained. Tran [94] proves L^2 convergence for histograms under α- and β-mixing. Devroye and Györfi [25] argue that L^1 is a more appropriate metric for studying density estimation, and Tran [93] proves L^1 consistency of KDEs under α- and β-mixing.

2 The Lambert W function is defined as the (multivalued) inverse of $f(w) = w\exp\{w\}$. Thus, $O(\exp\{W(\log n)\})$ is bigger than $O(\log\log n)$ but smaller than $O(\log n)$. See for example Corless et al. [16].

these lemmas allows me to derive rates of convergence for histograms based on β-mixing inputs.

The following lemma provides the doubly asymptotic convergence of the histogram estimator for IID data. It differs from standard histogram convergence results in the bias calculation. In this case I need to be more careful about the interaction between d and h_n.

Lemma 5.2. *For an IID sample $Z_1, \ldots, Z_n$ from some density f on $\mathbb{R}^d$,*

$$\mathbb{E} \int dz |\widehat{p}(z) - \mathbb{E}[\widehat{p}(z)]| = O\left(1/\sqrt{nh_n^d}\right) \tag{5.5}$$

$$\int dz |\mathbb{E}[\widehat{p}(z)] - p(z)| = O(dh_n) + O(d^2 h_n^2), \tag{5.6}$$

where $\widehat{p}$ is the histogram estimate using a grid with sides of length h_n.

Proof of Lemma 5.2. Let α_j be the probability of falling into the j^{th} bin B_j. Then,

$$\mathbb{E} \int |\widehat{p} - \mathbb{E}[\widehat{p}]| = h_n^d \sum_{j=1}^{J} \mathbb{E}\left[\left|\frac{1}{nh_n^d} \sum_{i=1}^{n} \mathbb{1}_{B_j}(Z_i) - \frac{\alpha_j}{h^d}\right|\right] \tag{5.7}$$

$$\leqslant h_n^d \sum_{j=1}^{J} \frac{1}{nh_n^d} \sqrt{\mathbb{V}\left[\sum_{i=1}^{n} \mathbb{1}_{B_j}(Z_i)\right]} \tag{5.8}$$

$$= h_n^d \sum_{j=1}^{J} \frac{1}{nh_n^d} \sqrt{n\alpha_j(1-\alpha_j)} \tag{5.9}$$

$$= \frac{1}{\sqrt{n}} \sum_{j=1}^{J} \sqrt{\alpha_j(1-\alpha_j)} \tag{5.10}$$

$$= O(n^{-1/2}) O(h_n^{-d/2}) = O\left(1/\sqrt{nh_n^d}\right). \tag{5.11}$$

For the second claim, consider the bin B_j centered at $\mathbf{c}$. Let $\mathbb{B}$ be the union of all bins B_j. Assume the following regularity conditions as in [35]:

1. $p \in L^2$ and p is absolutely continuous on $\mathbb{B}$, with a.e. partial derivatives $p_i = \frac{\partial}{\partial z_i} p(\mathbf{z})$

2. $p_i \in L^2$ and p_i is absolutely continuous on $\mathbb{B}$, with a.e. partial derivatives $p_{ik} = \frac{\partial}{\partial z_k} p_i(\mathbf{z})$

3. $p_{ik} \in L^2$ for all i, k.

Using a Taylor expansion

$$p(\mathbf{z}) = p(\mathbf{c}) + \sum_{i=1}^{d} (z_i - c_i) p_i(\mathbf{c}) + O(d^2 h_n^2). \tag{5.12}$$

Therefore, α_j is given by

$$\alpha_j = \int_{B_j} p(z)dz = h_n^d p(c) + O(d^2 h_n^{d+2}) \tag{5.13}$$

since the integral of the second term over the bin is zero. This means that for the j^{th} bin,

$$\mathbb{E}\left[\widehat{p}_n(z)\right] - p(z) = \frac{\alpha_j}{h_n^d} - p(z) \tag{5.14}$$

$$= -\sum_{i=1}^{d} (z_i - c_i) p_i(\mathbf{c}) + O(d^2 h_n^2). \tag{5.15}$$

Therefore,

$$\int_{B_j} \left|\mathbb{E}\left[\widehat{p}_n(z)\right] - p(z)\right| = \int_{B_j} \left| -\sum_{i=1}^{d} (z_i - c_i) p_i(\mathbf{c}) + O(d^2 h_n^2) \right| \tag{5.16}$$

$$\leqslant \int_{B_j} \left| -\sum_{i=1}^{d} (z_i - c_i) p_i(\mathbf{c}) \right| + \int_{B_j} O(d^2 h^2) \tag{5.17}$$

$$= \int_{B_j} \left| \sum_{i=1}^{d} (z_i - c_i) p_i(\mathbf{c}) \right| + O(d^2 h_n^{2+d}) \tag{5.18}$$

$$= O(d h_n^{d+1}) + O(d^2 h_n^{2+d}) \tag{5.19}$$

Since each bin is bounded, I can sum over all J bins. The number of bins is $J = h_n^{-d}$ by definition, so

$$\int dz\, |\mathbb{E}[\widehat{p}_n(z)] - p(z)| = O(h_n^{-d}) \left(O(dh_n^{d+1}) + O(d^2 h_n^{2+d}) \right) \tag{5.20}$$

$$= O(dh_n) + O(d^2 h_n^2). \tag{5.21}$$

■

I can now prove the main result of this section.

Proof of Theorem 5.1. Let g be the L^1 loss of the histogram estimator, $g = \int |p - \widehat{p}_n|$. Here $\widehat{p}_n(y) = \frac{1}{n h_n^d} \sum_{i=1}^n \mathbb{1}_{B_j(y)}(Y_i)$ where $B_j(y)$ is the bin containing z. Let $\widehat{p}_{\mathbf{U}}$, $\widehat{p}_{\mathbf{V}}$, and $\widehat{p}_{\mathbf{U}'}$ be histograms based on the block sequences $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{U}'$ respectively. Clearly $\widehat{p}_n = \frac{1}{2}(\widehat{p}_{\mathbf{U}} + \widehat{p}_{\mathbf{V}})$. Now,

$$\mathbb{P}_{1:n}(g > \epsilon) = \mathbb{P}_{1:n}\left(\int |p - \widehat{p}_n| > \epsilon \right) \tag{5.22}$$

$$= \mathbb{P}_{1:n}\left(\int \left| \frac{p - \widehat{p}_{\mathbf{U}}}{2} + \frac{p - \widehat{p}_{\mathbf{V}}}{2} \right| > \epsilon \right) \tag{5.23}$$

$$\leqslant \mathbb{P}_{1:n}\left(\frac{1}{2} \int |p - \widehat{p}_{\mathbf{U}}| + \frac{1}{2} \int |p - \widehat{p}_{\mathbf{V}}| > \epsilon \right) \tag{5.24}$$

$$= \mathbb{P}_{1:n}(g_{\mathbf{U}} + g_{\mathbf{V}} > 2\epsilon) \tag{5.25}$$

$$\leqslant \mathbb{P}_{\mathbf{U}}(g_{\mathbf{U}} > \epsilon) + \mathbb{P}_{\mathbf{V}}(g_{\mathbf{V}} > \epsilon) \tag{5.26}$$

$$= 2\mathbb{P}_{\mathbf{U}}(g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}}] > \epsilon - \mathbb{E}[g_{\mathbf{U}}]) \tag{5.27}$$

$$= 2\mathbb{P}_{\mathbf{U}}(g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon - \mathbb{E}[g_{\mathbf{U}'}]) \tag{5.28}$$

$$= 2\mathbb{P}_{\mathbf{U}}(g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1), \tag{5.29}$$

where $\epsilon_1 = \epsilon - \mathbb{E}[g_{\mathbf{U}'}]$. Here,

$$\mathbb{E}[g_{\mathbf{U}'}] \leqslant \mathbb{E} \int dz |\widehat{p}_{\mathbf{U}'} - \mathbb{E}[\widehat{p}_{\mathbf{U}'}]| + \int dz |\mathbb{E}[\widehat{p}_{\mathbf{U}'}] - p|, \tag{5.30}$$

so by Lemma 5.2, as long as $\mu_n \to \infty$, $h_n \downarrow 0$ and $\mu_n h_n^d \to \infty$, then for all ϵ there exists $n_0(\epsilon)$ such that for all $n > n_0(\epsilon)$, $\epsilon > \mathbb{E}[g] = \mathbb{E}[g_{\mathbf{U}'}]$. Now applying Lemma 4.8 to the the event $\{g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1\}$ gives

$$2\mathbb{P}_{\mathbf{U}}(g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1) \leqslant 2\mathbb{P}_{\mathbf{U}'}(g_{\mathbf{U}'} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1) + 2(\mu_n - 1)\beta_{m_n} \tag{5.31}$$

where the probability on the right is for the σ-field generated by the independent block sequence $\mathbf{U}'$. Since these blocks are independent, showing that $g_{\mathbf{U}'}$ satisfies the bounded differences requirement allows for the application of McDiarmid's inequality, Theorem 3.6, to the blocks. For any two block sequences $\mathbf{u}'_{1:\mu_n}$ and $\overline{\mathbf{u}}'_{1:\mu_n}$ with $u'_\ell = \overline{u}'_\ell$ for all $\ell \neq j$, then

$$\left|g_{\mathbf{U}'}(\mathbf{u}'_{1:\mu_n}) - g_{\mathbf{U}'}(\overline{\mathbf{u}}'_{1:\mu_n})\right| = \left|\int |\widehat{p}(y;\mathbf{u}'_{1:\mu_n}) - p(y)|dy - \int |\widehat{p}(y;\overline{\mathbf{u}}'_{1:\mu_n}) - p(y)|dy\right| \tag{5.32}$$

$$\leqslant \int |\widehat{p}(y;\mathbf{u}'_{1:\mu_n}) - \widehat{p}(y;\overline{\mathbf{u}}'_{1:\mu_n})|dy \tag{5.33}$$

$$= \frac{2}{\mu_n h_n^d} h_n^d = \frac{2}{\mu_n}. \tag{5.34}$$

Therefore,

$$\mathbb{P}_{1:n}(g > \epsilon) \leqslant 2\mathbb{P}_{\mathbf{U}}(g_{\mathbf{U}'} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1) + 2(\mu_n - 1)\beta_{m_n} \tag{5.35}$$

$$\leqslant 2\exp\left\{-\frac{\mu_n \epsilon_1^2}{2}\right\} + 2(\mu_n - 1)\beta_{m_n}. \tag{5.36}$$

■

5.4 PROPERTIES OF THIS ESTIMATOR

In this section, I derive some properties of my proposed estimator. A finite sample bound for the estimation error is the first step to establishing consistency for $\widehat{\beta}_a^{d_n}$. This result gives convergence rates for estimation of the finite dimensional mixing

coefficient β_a^d and also for Markov processes of known order d, since in this case, $\beta_a^d = \beta_a$. A second result shows convergence of the approximation error. Taken together, I can show that under some conditions $\widehat{\beta}_a^{d_n}$ is consistent.

Theorem 5.3. *Consider a sample* $\mathbf{Y}_{1:n}$ *from a stationary* β*-mixing process* $\mathbb{P}_\infty$*. Let* μ_n *and* m_n *be positive integers such that* $2\mu_n m_n \leqslant n$ *and* $\mu_n \geqslant d > 0$*. Then*

$$\begin{aligned}\mathbb{P}_{1:n}(|\widehat{\beta}_a^d - \beta_a^d| > \epsilon) \leqslant 2\exp\left\{-\frac{\mu_n \epsilon_1^2}{2}\right\} + 2\exp\left\{-\frac{\mu_n \epsilon_2^2}{2}\right\} \\ + 4(\mu_n - 1)\beta_{m_n},\end{aligned} \tag{5.37}$$

where $\epsilon_1 = \epsilon/2 - \mathbb{E}\left[\int |\widehat{p}^d - p^d|\right]$ *and* $\epsilon_2 = \epsilon - \mathbb{E}\left[\int |\widehat{p}_a^{2d} - p_a^{2d}|\right]$.

The proof of Theorem 5.3 relies on the triangle inequality and the relationship between total variation distance and the L^1 distance between densities.

Proof of Theorem 5.3. For any probability measures ν and λ defined on the same probability space with associated densities p_ν and p_λ with respect to some dominating measure π,

$$\|\nu - \lambda\|_{TV} = \frac{1}{2}\int |p_\nu - p_\lambda| d(\pi). \tag{5.38}$$

Note that by stationarity, $\forall t \in \mathbb{N}$, $\mathbb{P}_{0:d} = \mathbb{P}_{t:t+d}$ in the notation of Lemma 4.6. Let $\mathbb{P}_{-d:0\otimes a:a+d}$ be the joint distribution of the bivariate random process created by the initial process and itself separated by a time steps. By the triangle inequal-

ity, one can upper bound β_a^d for any $d = d_n$. Let $\widehat{\mathbb{P}}_{0:d}$ and $\widehat{\mathbb{P}}_{-d:0\otimes a:a+d}$ be the distributions associated with histogram estimators $\widehat{p}^d$ and $\widehat{p}_a^{2d}$ respectively. Then,

$$\begin{aligned}
\beta_a^d &= \|\mathbb{P}_{0:d} \otimes \mathbb{P}_{0:d} - \mathbb{P}_{-d:0\otimes a:a+d}\|_{TV} && (5.39)\\
&= \Big|\Big|\mathbb{P}_{0:d} \otimes P_{0:d} - \widehat{\mathbb{P}}_{0:d} \otimes \widehat{\mathbb{P}}_{0:d} + \widehat{\mathbb{P}}_{0:d} \otimes \widehat{\mathbb{P}}_{0:d} \\
&\quad - \widehat{\mathbb{P}}_{-d:0\otimes a:a+d} + \widehat{\mathbb{P}}_{-d:0\otimes a:a+d} - \mathbb{P}_{-d:0\otimes a:a+d}\Big|\Big|_{TV} && (5.40)\\
&\leqslant \Big|\Big|\mathbb{P}_{0:d} \otimes \mathbb{P}_{0:d} - \widehat{\mathbb{P}}_{0:d} \otimes \widehat{\mathbb{P}}_{0:d}\Big|\Big|_{TV} + \Big|\Big|\widehat{\mathbb{P}}_{0:d} \otimes \widehat{\mathbb{P}}_{0:d} - \widehat{\mathbb{P}}_{-d:0\otimes a:a+d}\Big|\Big|_{TV} \\
&\quad + \Big|\Big|\widehat{\mathbb{P}}_{-d:0\otimes a:a+d} - \mathbb{P}_{-d:0\otimes a:a+d}\Big|\Big|_{TV} && (5.41)\\
&\leqslant 2\Big|\Big|\mathbb{P}_{0:d} - \widehat{\mathbb{P}}_{0:d}\Big|\Big|_{TV} + \Big|\Big|\widehat{\mathbb{P}}_{0:d} \otimes \widehat{\mathbb{P}}_{0:d} - \widehat{\mathbb{P}}_{-d:0\otimes a:a+d}\Big|\Big|_{TV} \\
&\quad + \Big|\Big|\widehat{\mathbb{P}}_{-d:0\otimes a:a+d} - \mathbb{P}_{-d:0\otimes a:a+d}\Big|\Big|_{TV} && (5.42)\\
&= \int |p^d - \widehat{p}^d| + \frac{1}{2}\int |\widehat{p}^d \otimes \widehat{p}^d - \widehat{p}_a^{2d}| + \frac{1}{2}\int |p_a^{2d} - \widehat{p}_a^{2d}| && (5.43)
\end{aligned}$$

where $\frac{1}{2}\int |\widehat{p}^d \otimes \widehat{p}^d - \widehat{p}_a^{2d}|$ is our estimator $\widehat{\beta}_a^d$ and the remaining terms are the L^1 distance between a density estimator and the target density. Thus,

$$\beta_a^d - \widehat{\beta}_a^d \leqslant \int |f^d - \widehat{p}^d| + \frac{1}{2}\int |f_a^{2d} - \widehat{p}_a^{2d}|. \tag{5.44}$$

A similar argument starting from $\beta_a^d = \|\mathbb{P}_{0:d} \otimes \mathbb{P}_{0:d} - \mathbb{P}_{-d:0\otimes a:a+d}\|_{TV}$ shows that

$$\beta_a^d - \widehat{\beta}_a^d \geqslant -\int |p^d - \widehat{p}^d| - \frac{1}{2}\int |p_a^{2d} - \widehat{p}_a^{2d}|, \tag{5.45}$$

so

$$\left|\beta_a^d - \widehat{\beta}_a^d\right| \leqslant \int |p^d - \widehat{p}^d| + \frac{1}{2}\int |p_a^{2d} - \widehat{p}_a^{2d}|. \tag{5.46}$$

Therefore,

$$\mathbb{P}\left(\left|\beta_a^d - \widehat{\beta}_a^d\right| > \epsilon\right) \leqslant \mathbb{P}\left(\int |p^d - \widehat{p}^d| + \frac{1}{2}\int |p_a^{2d} - \widehat{p}_a^{2d}| > \epsilon\right) \tag{5.47}$$

$$\leqslant \mathbb{P}\left(\int |p^d - \widehat{p}^d| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\frac{1}{2}\int |p_a^{2d} - \widehat{p}_a^{2d}| > \frac{\epsilon}{2}\right) \tag{5.48}$$

$$\leqslant 2\exp\left\{-\frac{\mu_n \epsilon_1^2}{2}\right\} + 2\exp\left\{-\frac{\mu_n \epsilon_2^2}{2}\right\} \tag{5.49}$$

$$+ 4(\mu_n - 1)\beta_{m_n}, \tag{5.50}$$

where $\epsilon_1 = \epsilon/2 - \mathbb{E}\left[\int |\widehat{p}^d - p^d|\right]$ and $\epsilon_2 = \epsilon - \mathbb{E}\left[\int |\widehat{p}_a^{2d} - p_a^{2d}|\right]$. ■

Consistency of the estimator $\widehat{\beta}_a^d$ is guaranteed only for certain choices of m_n and μ_n. Clearly $\mu_n \to \infty$ and $\mu_n \beta_{m_n} \to 0$ as $n \to \infty$ are necessary conditions. Consistency also requires convergence of the histogram estimators to the target densities as given in the previous section. As an example to show that this bound can go to zero with proper choices of m_n and μ_n, the following corollary proves consistency for first order Markov processes. Consistency of the estimator for higher order Markov processes can be proven similarly. These processes are geometrically β-mixing as shown in e.g. Nummelin and Tuominen [71].

Corollary 5.4. *Let $\mathbf{Y}_{1:n}$ be a sample from a first order Markov process with $\beta_a = \beta_a^1 = O(\rho^{-a})$ for some $\rho > 1$. Then under the conditions of Theorem 5.3,*

$$\mathbb{E}\left[|\widehat{\beta}_a^1 - \beta_a|\right] = O\left(\sqrt{\frac{W\left(\frac{2}{3}n\log\rho\right)}{n}}\right) \tag{5.51}$$

where $W(\cdot)$ is the Lambert W function.

Proof. Here, c are various constants.

$$\mathbb{E}\left[|\widehat{\beta}_a^1 - \beta_a|\right] = \int_0^\infty d\epsilon \mathbb{P}_{1:n}\left(|\widehat{\beta}_a^1 - \beta_a| > \epsilon\right) \tag{5.52}$$

$$= \int_0^1 d\epsilon \mathbb{P}_{1:n}\left(|\widehat{\beta}_a^1 - \beta_a| > \epsilon\right) \tag{5.53}$$

$$\leqslant c\int_0^1 d\epsilon \exp(-c\mu_n\epsilon^2) + \int_0^1 d\epsilon c\mu_n\beta_{m_n} \tag{5.54}$$

$$\leqslant \sqrt{\frac{c}{\mu_n}} + c\mu_n\rho^{-a}. \tag{5.55}$$

These two terms are balanced by taking

$$\mu_n = O\left(\frac{n}{W\left(\frac{2}{3}n\log\rho\right)}\right) \tag{5.56}$$

giving the result. ∎

My main result in this section establishes consistency of $\widehat{\beta}_a^{d_n}$ as an estimator of β_a for all β-mixing processes provided d_n increases at an appropriate rate. Theorem 5.3 gives finite sample bounds on the estimation error while some measure theoretic arguments show that the approximation error must go to zero as $d_n \to \infty$.

Theorem 5.5. *Let* $\mathbf{Y}_{1:n}$ *be a sample from an arbitrary* β*-mixing process.Let* $d_n = O(\exp\{W(\log n)\})$ *where* W *is the Lambert* W *function. Then* $\widehat{\beta}_a^{d_n} \xrightarrow{P} \beta_a$ *as* $n \to \infty$.

The proof of Theorem 5.5 requires two steps which are given in the following Lemmas. The first specifies the histogram bandwidth h_n and the rate at which d_n (the dimensionality of the target density) goes to infinity. If the dimensionality of the target density were fixed, one could achieve rates of convergence similar to those for histograms based on IID inputs. However, I wish to allow the dimensionality to grow with n, so the rates are much slower as shown in the following lemma.

Lemma 5.6. *For the histogram estimator in Lemma 5.2, let*

$$d_n \sim \exp\{W(\log n)\}, \tag{5.57}$$

$$h_n \sim n^{-k_n}, \tag{5.58}$$

with

$$k_n = \frac{W(\log n) + \frac{1}{2}\log n}{\log n\left(\frac{1}{2}\exp\{W(\log n)\} + 1\right)}. \tag{5.59}$$

These choices lead to the optimal rate of convergence.

Proof. Let $h_n = n^{-k_n}$ for some k_n to be determined. Then I need

$$n^{-1/2}h_n^{-d_n/2} = n^{(k_n d_n - 1)/2} \to 0, \tag{5.60}$$

$$d_n h_n = d_n n^{-k} \to 0, \tag{5.61}$$

and

$$d_n^2 h_n^2 = d_n^2 n^{-2k} \to 0 \tag{5.62}$$

as well as $n \to \infty$. Taking (5.60) and (5.61) first gives

$$n^{(k_n d_n - 1)/2} \sim d_n n^{-k_n} \tag{5.63}$$

$$\Rightarrow \frac{1}{2}(k_n d_n - 1)\log n \sim \log d_n - k_n \log n \tag{5.64}$$

$$\Rightarrow k_n \log n\left(\frac{1}{2}d_n + 1\right) \sim \log d_n + \frac{1}{2}\log n \tag{5.65}$$

$$\Rightarrow k_n \sim \frac{\log d_n + \frac{1}{2}\log n}{\log n\left(\frac{1}{2}d_n + 1\right)}. \tag{5.66}$$

Similarly, combining (5.60) and (5.62) gives

$$k_n \sim \frac{2\log d_n + \frac{1}{2}\log n}{\log n\left(\frac{1}{2}d_n + 2\right)}. \tag{5.67}$$

Equating (5.66) and (5.67) and solving for d_n gives

$$\Rightarrow d_n \sim \exp\{W(\log n)\} \tag{5.68}$$

where $W(\cdot)$ is the Lambert W function. Substituting back into (5.66) gives that

$$h_n = n^{-k_n} \tag{5.69}$$

where

$$k_n = \frac{W(\log n) + \frac{1}{2}\log n}{\log n\left(\frac{1}{2}\exp\{W(\log n)\} + 1\right)}. \tag{5.70}$$

■

It is also necessary to show that as d grows, $\beta_a^d \to \beta_a$. I now state this result. For the proof, see Appendix A.

Lemma 5.7. *β_a^d converges to β_a as $d \to \infty$.*

The basic idea of the proof is to show that β_a^d is a monotone increasing sequence in d which is bounded above by β_a. Therefore it must be that $\lim_{d\to\infty} \beta_a^d \leqslant \beta_a$. Showing that the limit is equal to β_a uses the Hahn decomposition theorem and some measure theoretic results.

I can now prove my main result in Theorem 5.5: that $\widehat{\beta}_a^d$ is a consistent estimator of β_a.

Proof of Theorem 5.5. By the triangle inequality,

$$|\widehat{\beta}^{d_n}(a) - \beta_a| \leqslant |\widehat{\beta}^{d_n}(a) - \beta^{d_n}(a)| + |\beta^{d_n}(a) - \beta_a|.$$

The first term on the right is bounded by the result in Theorem 5.3, where I have shown that $d_n = O(\exp\{W(\log n)\})$ is slow enough for the histogram estimator to remain consistent. That $\beta^{d_n}(a) \xrightarrow{d_n\to\infty} \beta_a$ follows from Lemma 5.7. ■

5.5 PERFORMANCE IN SIMULATIONS

To demonstrate the performance of our proposed estimator, I examine its performance in three simulated examples. The first example is a simple two state Markov

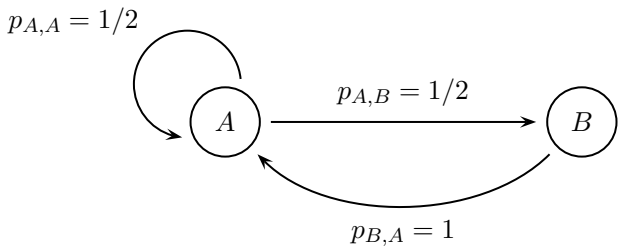


Figure 4: This figure shows the two-state Markov chain S_t used for simulation results

chain. The second example takes this Markov chain as an unobserved input and outputs a non-Markovian binary sequence which remains β-mixing. Finally, I examine an autoregressive model.

As shown in [17], homogeneous recurrent Markov chains are geometrically β-mixing, i.e. $\beta_a = O(\rho^{-a})$ for some $\rho > 1$. In particular, if the Markov chain has stationary distribution $\mathbb{P}_1$ and a-step transition distribution T^a, then

$$\beta_a = \int \mathbb{P}_1(dy) \|T^a(y) - \mathbb{P}_1\|_{TV}. \tag{5.71}$$

Consider first the two-state Markov chain S_t pictured in Figure 4. By direct calculation using (5.71), the mixing coefficients for this process are $\beta_a = \frac{4}{9}\left(\frac{1}{2}\right)^a$. I simulated chains of length $n = 1000$ from this Markov model. Based on 1000 replications, the performance of the estimator is depicted in Figure 5. Here, I have used two bins in all cases, but I allow the Markov approximation to vary as $d \in \{1, 2, 3\}$, even though $d = 1$ is exact. The estimator performs well for $a \leqslant 5$, but begins to exhibit a positive bias as a increases. This is because the estimator is nonnegative, whereas the true mixing coefficients are quickly approaching zero. The upward bias is exaggerated for larger d. This bias will go to 0 as $n \to \infty$.

As an example of a long memory process, I construct, following Weiss [100], a partially observable Markov process which is referred to as the "even process".

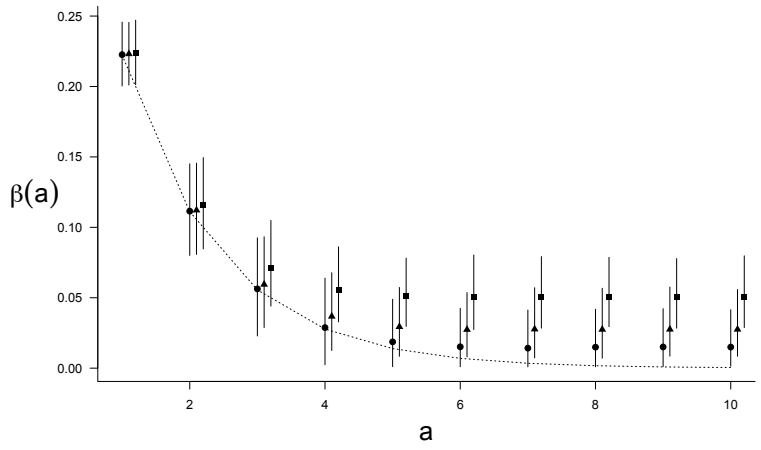


Figure 5: This figure illustrates the performance of our proposed estimator for the two-state Markov chain depicted in Figure 4. I simulated length $n = 1000$ chains and calculated $\widehat{\beta}^d(a)$ for $d = 1$ (circles), $d = 2$ (triangles), and $d = 3$ (squares). The dashed line indicates the true mixing coefficients. I show means and 95% confidence intervals based on 1000 replications.

Let X_t be the observed sequence which takes as input the Markov process S_t constructed above. One observes

$$X_t = \begin{cases} 1 & (S_t, S_{t-1}) = (A, B) \text{ or } (B, A) \\ 0 & \text{else.} \end{cases} \tag{5.72}$$

Since S_t is Markovian, the joint process (S_t, S_{t-1}) is as well, so I can calculate its mixing rate $\beta_a = \frac{8}{9}\left(\frac{1}{2}\right)^a$. The even process must also be β-mixing, and at least as fast as the joint process, since it is a measurable function of a mixing process. However, X_t itself is non-Markovian: sequences of one's must have even lengths, so I need to know how many one's have been observed to know whether the next observation can be zero or must be a one. Thus, the true mixing coefficients are bounded above by $\frac{8}{9}\left(\frac{1}{2}\right)^a$ due to Lemma 4.7, but the coefficients of the observed process areunknown. Using the same procedure as above, Figure 6 shows the estimated mixing coefficients. Again one observes a bias for a large due to the nonnegativity of the estimator.

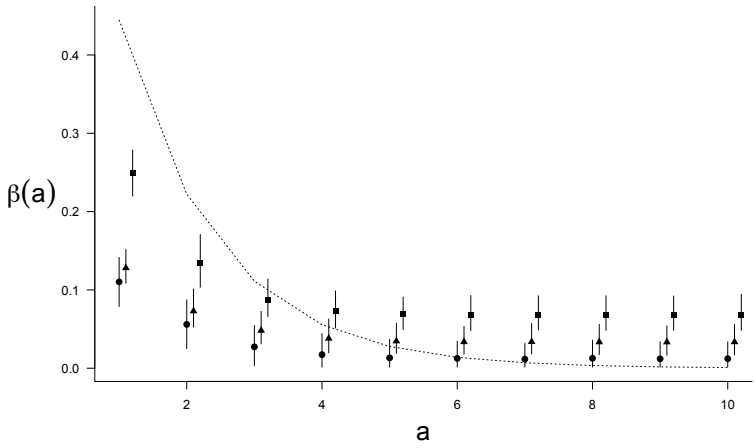


Figure 6: This figure illustrates the performance of our proposed estimator for the even process. Again, I simulated length $n = 1000$ chains and calculated $\widehat{\beta}^{d}(a)$ for $d = 1$ (circles), $d = 2$ (triangles), and $d = 3$ (squares). The dashed line indicates an upper bound on the true mixing coefficients. I show means and 95% confidence intervals based on 1000 replications.

Finally, I estimate the β-mixing coefficients for an AR(1) model

$$Z_t = 0.5Z_{t-1} + \eta_t \qquad \eta_t \overset{iid}{\sim} N(0,1). \tag{5.73}$$

While, this process is Markovian, there is no closed form solution to (5.71), so I calculate it via numerical integration. Figure 7 shows the performance of the estimator for $d = 1$. I select the bandwidth h for each a by minimizing

$$\mathbb{E}[|\widehat{\beta}(a) - \beta_a|] \tag{5.74}$$

where I calculate the expectation based on independent simulations from the process. Figure 7 shows the performance for $n = 3000$. The optimal number of bins is 33, 11, 7, 5, and 3 for $a = 1, \ldots, 5$ and 1 for $a > 5$. However, since the use of one bin corresponds to an estimate of zero, the figure plots the estimate with two bins. Using two bins, one again sees the positive bias for $a > 5$.

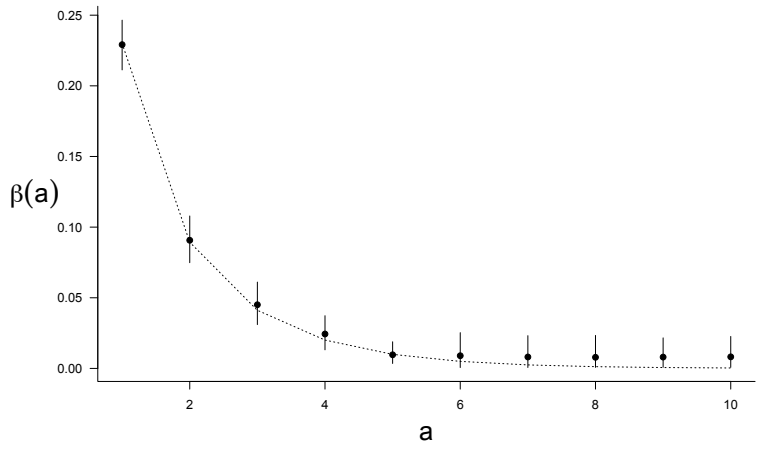


Figure 7: This figure illustrates the performance of our proposed estimator for the AR(1) model. I simulated time series of length $n = 3000$ chains and calculated $\widehat{\beta}(a)$ for $d = 1$. The dashed line indicates the true mixing coefficients calculated via numerical integration. I show sample means and 95% confidence intervals based on 1000 replications.

5.6 DISCUSSION

I have shown that my estimator of the β-mixing coefficients is consistent for the true coefficients β_a under some conditions on the data generating process. There are numerous results in the statistics literature which assume knowledge of the β-mixing coefficients, yet as far as I know, this is the first estimator for them. An ability to estimate these coefficients will allow researchers to apply existing results to dependent data without the need to arbitrarily assume their values. Additionally, it will allow probabilists to recover unknown mixing coefficients for stochastic processes via simulation. Several other mixing and weak-dependence coefficients also have a total-variation flavor, perhaps most notably α-mixing [28, 18, 9]. None of them have estimators, and the same trick might well work for them, too. Despite the obvious utility of this estimator, as a consequence of its novelty, it comes with a number of potential extensions which warrant careful exploration as well as some drawbacks.

Theorem 5.5 does not provide a convergence rate. The rate in Theorem 5.3 applies only to the difference between $\widehat{\beta}^{d}(a)$ and β_a^d. In order to provide a rate in Theorem 5.5, one would need a better understanding of the non-stochastic convergence of β_a^d to β_a. It is not immediately clear that this quantity can converge at any well-defined rate. In particular, it seems likely that the rate of convergence depends on the tail of the sequence $\{\beta_a\}_{a=1}^{\infty}$.

The use of histograms rather than kernel density estimators for the joint and marginal densities is surprising and perhaps not ultimately necessary. As mentioned above, Tran [93] proved that KDEs are consistent for estimating the stationary density of a time series with β-mixing inputs, so perhaps one could replace the histograms in our estimator with KDEs. However, this would need an analogue of the double asymptotic results proven for histograms in Lemma 5.2. In particular, one needs to estimate increasingly higher dimensional densities as $n \to \infty$. This does not cause a problem of small-n-large-d since d is chosen as a function of n, however it will lead to increasingly higher dimensional integration. For histograms, the integral is always trivial, but in the case of KDEs, the numerical accuracy of the integration algorithm becomes increasingly important. This issue could swamp any statistical efficiency gains obtained through the use of kernels. However, this question certainly warrants further investigation.

The main drawback of an estimator based on a density estimate is its complexity. The mixing coefficients are functionals of the joint and marginal distributions derived from the stochastic process $\mathbf{Y}_\infty$, however, it is unsatisfying to estimate densities and solve integrals in order to estimate a single number. Vapnik's main principle for solving problems using a restricted amount of information is "When solving a given problem, try to avoid solving a more general problem as an intermediate step [97, p. 30]." However, despite my estimator's complexity, I am able to obtain nearly parametric rates of convergence to the Markov approximation departing only by logarithmic factors.

6

BOUNDS FOR STATE SPACE MODELS

With the relevant background in Chapters 3–5 in place, I can put the pieces together to present my results. I use β-mixing to find out how much information is in the data and VC dimension to measure the capacity of the state-space model's prediction functions. The result is a bound on the generalization error of the chosen function $\widehat{f}$. In the remainder of this section, I redefine the appropriate concepts in the time series forecasting scenario, I state the necessary assumptions for our results, and I derive risk bounds for wide classes of economic forecasting models. Section 6.1 states and proves risk bounds for the time series forecasting setting, while I demonstrate how to use the results in Section 6.2. Section 6.4 discusses the use of risk bounds for model selection. Finally, Section 6.5 concludes and illustrates the path toward generalizing our methods to more elaborate model classes.

6.1 RISK BOUNDS

6.1.1 *Setup and assumptions*

Consider a finite subsequence of random vectors $\mathbf{Y}_{1:n}$ from a process $\mathbf{Y}_\infty$ defined on a probability space $(\Omega, \sigma_\infty, \mathbb{P}_\infty)$ such that $Y_i \in \mathbb{R}^p$. I make the following assumption on the infinite process.

Assumption C. *Assume that $\mathbb{P}_\infty$ is a stationary, β-mixing distribution with known mixing coefficients β_a, $\forall a > 0$.*

Under stationarity, the marginal distribution of Y_t is the same for all t. I am mainly concerned with the joint distribution of sequences $\mathbf{Y}_{1:n+1}$ wherein one observes the first n observations and attempts to predict time $n+1$. For the remainder of this chapter, I will call this joint distribution $\mathbb{P}$. My results are easily extended to the case of predicting more than one step ahead, but the notation becomes cumbersome.

One defines generalization error and training error in the time series setting slightly differently than in the IID setting. First one needs an appropriate loss function. I will take the loss function ℓ to be some norm $\|\cdot\|$ on $\mathbb{R}^p$, and I will consider prediction functions $f : \mathbb{R}^{n\times p} \to \mathbb{R}^p$

Definition 6.1 (Time series risk)**.**

$$R_n(f) := \mathbb{E}_{\mathbb{P}_{1:n+1}}\Big[\, \|Y_{n+1} - f(\mathbf{Y}_1^n)\| \,\Big]. \tag{6.1}$$

The expectation is taken with respect to the joint distribution $\mathbb{P}$ and therefore depends on n. One may use some or all of the past to generate predictions. A function which takes only the most recent d observations as inputs will be referred to as having *fixed memory* d. Other functions have *growing memory*, i.e., one may use all the previous data to predict the next data point. For this reason, I define two versions of the training error depending on whether or not the memory of the prediction function f is fixed.

Definition 6.2 (Time series training error with memory d)**.**

$$\widehat{R}_n(f) := \frac{1}{n-d-1}\sum_{i=d}^{n-1} \|Y_{i+1} - f(\mathbf{Y}_{i-d+1:i})\| \tag{6.2}$$

Definition 6.3 (Time series training error with growing memory)**.**

$$\widetilde{R}_n(f) := \frac{1}{n-d-1}\sum_{i=d}^{n-1} \|Y_{i+1} - f(\mathbf{Y}_{1:i})\| \tag{6.3}$$

The first case is useful for standard autoregressive forecasting methods, while the second case is applicable to ARMA models, DSGEs, and linear state space models. Additionally, I am writing f as a fixed function, but the dimension of the argument changes with i. This is not an issue for functions which are linear in the data, as is the case with ARMA models, linear state-space models, and linearized DSGEs (see Section 2.3). For nonlinear models, I will consider only the fixed memory version.

6.1.2 *Fixed memory*

I begin with the fixed memory setting before allowing the memory length to grow.

Theorem 6.4. *Given a sample* $\mathbf{Y}_{1:n}$ *such that Assumption B and Assumption C hold, suppose that the model class* $\mathcal{F}$ *has a fixed memory length* $d < n$*. Let* μ *and* a *be integers such that* $2\mu a + d \leqslant n$*. Then, for all* $\epsilon > 0$*,*

$$\begin{aligned}&\mathbb{P}_{1:n}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} > \epsilon\right)\\ &\leqslant 8\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})} + 1\right) - \frac{\mu\epsilon^2}{4\tau^2(q)M^2}\right\} + 2(\mu-1)\beta_{a-d},\end{aligned} \tag{6.4}$$

where $\tau(q) = \sqrt[q]{\frac{1}{2}\left(\frac{q-1}{q-2}\right)^{q-1}}$.

The implications of this theorem are considerable. Given a finite number of observations n, one can say that with high probability, future relative prediction errors will not be much larger than our observed training errors. It makes no difference whether the model is correctly specified. This stands in stark contrast to model selection tools like AIC or BIC which appeal to asymptotic results as in

Claeskens and Hjort [15]. Moreover, given some model class $\mathcal{F}$, one can say exactly how much data is required to have good control of the prediction risk. As the effective data size increases, the righthand side goes to zero given appropriate mixing rates and so the training error is a better and better estimate of the generalization error.

One way to understand this theorem is to visualize the tradeoff between confidence ϵ and effective data μ. Consider the following, drastically simplified version of the result

$$\mathbb{P}_{1:n}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_n(f)}{R_n(f)}>\epsilon\right)\leqslant 8\exp\left\{\ln 2\mu+1-\frac{\mu\epsilon^2}{4}\right\} \tag{6.5}$$

where I have taken the VC dimension to be one and I ignore the extra penalty from the mixing coefficient—i.e. $\beta_a = 0, \forall a > 0$ and therefore $\mu = n$. The goal is to minimize ϵ, thereby ensuring that the relative difference between the expected risk and the training risk is small. At the same time I want to minimize the right side of the bound so that the probability of "bad" outcomes — events such that the relative difference in risks exceeds ϵ — is small. Of course I want to do this with as little data as possible, but the smaller I take ϵ, the larger I must take μ to compensate. I illustrate this tradeoff in Figure 8.

The relative difference between expected and empirical risk is only interesting between zero and one. By construction, it can be no larger than one since $\widehat{R}_n(f) \geqslant 0$, and due to the supremum, events where the training error exceeds the expected risk are irrelevant. Therefore, I am only concerned with $0 \leqslant \widehat{R}(f) \leqslant R_n(f)$, so I need only consider $0 \leqslant \epsilon \leqslant 1$.

The figure is structured so that movement toward the origin is preferable. I have tighter control on the difference in risks with less data. But moving in that direction leads to an increased probability of the bad event — that the difference in risks exceeds ϵ. The bound becomes trivial below the solid black line (the formula says that the bad event occurs with probability larger than one). The desire for the

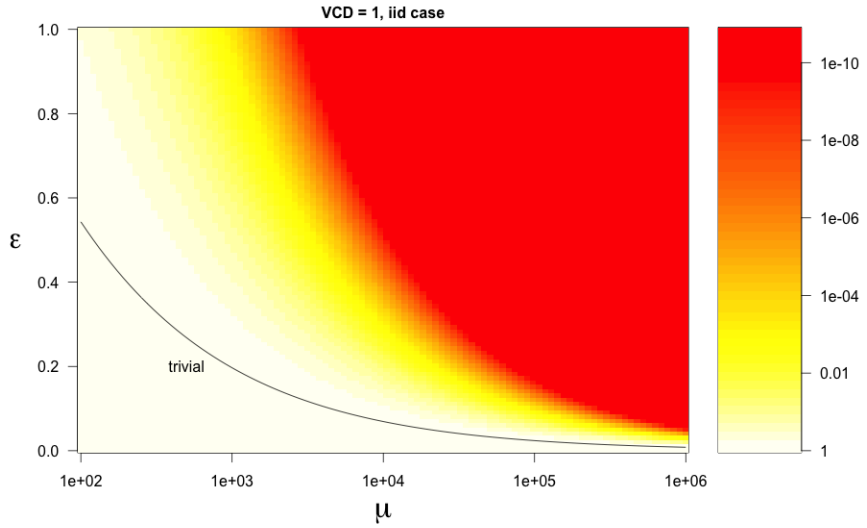


Figure 8: Visualizing the tradeoff between confidence (ϵ, y-axis) and effective data (μ, x-axis). The black curve indicates the region where the bound becomes trivial. Below this line, the probability is bounded by 1. Darker colors indicate lower probability of the "bad" event — that the difference in risks exceeds ϵ. The colors correspond to the natural logarithm of the bound on this probability.

bad event to occur with low probability forces the decision boundary to the upper right.

Another way to interpret the plot is as a set of indifference curves. Anywhere in the same color region is equally desirable in the sense that the probability of bad events is the same. So if I faced a budget constraint trading ϵ and data (i.e. a line with negative slope), I could optimize within the budget set to find the lowest probability allowable.

Before I prove Theorem 6.4 I will state a corollary the form of which is occasionally more convenient.

Corollary 6.5. *Under the conditions of Theorem 6.4, with probability at least* $1-\eta$*, for all* $\eta > 2(\mu-1)\beta_{a-d}$*, the following bound holds simultaneously for all* $f \in \mathcal{F}$ *(including the minimizer of the empirical risk* $\widehat{f}$*):*

$$R_n(f) \leqslant \frac{\widehat{R}_n(f)}{(1-\mathcal{E})_+}. \tag{6.6}$$

Here

$$\mathcal{E} = \frac{2M\tau(q)}{\sqrt{\mu}}\sqrt{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})}+1\right)-\ln(\eta'/8)}, \tag{6.7}$$

$\eta' = \eta - 2(\mu-1)\beta_{a-d}$, $\tau(q) = \sqrt[q]{\frac{1}{2}\left(\frac{q-1}{q-2}\right)^{q-1}}$, *and* $(u)_+ = \max(u,0)$.

Proof of Theorem 6.4. The first step is to move from the actual sample size n to the effective sample size μ which depends on the β-mixing behavior. Now divide $\mathbf{Y}_{1:n}$ into 2μ blocks, each of length a. Identify "odd" blocks $\mathbf{U}$ and "even" blocks $\mathbf{V}$ as in Chapter 4. To repeat,

$$U_j = \{Y_i : 2(j-1)a+1 \leqslant i \leqslant (2j-1)a\}, \tag{6.8}$$

$$V_j = \{Y_i : (2j-1)a+1 \leqslant i \leqslant 2ja\}. \tag{6.9}$$

Let $\mathbf{U} = \{U_j\}_{j=1}^{\mu}$ and let $\mathbf{V} = \{V_j\}_{j=1}^{\mu}$. Finally, let $\mathbf{U}'$ be a sequence of blocks which are independent of $\mathbf{Y}_{1:n}$ but such that each block has the same distribution as a block from the original sequence — i.e.

$$\mathcal{L}(U_j') = \mathcal{L}(U_j) = \mathcal{L}(U_1). \tag{6.10}$$

Let $\widehat{R}_{\mathbf{U}}(f)$, $\widehat{R}_{\mathbf{U}'}(f)$, and $\widehat{R}_{\mathbf{V}}(f)$ be the empirical risk of f based on the block sequences $\mathbf{U}$, $\mathbf{U}'$, and $\mathbf{V}$ respectively. Clearly $\widehat{R}_n(f) = \frac{1}{2}(\widehat{R}_{\mathbf{U}}(f) + \widehat{R}_{\mathbf{V}}(f))$. Define $\tau(q)$ as in the statement of the theorem. Then,

$$
\begin{aligned}
&\mathbb{P}_{1:n}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_n(f)}{R_n(f)} > \epsilon\right) \\
&= \mathbb{P}_{1:n}\left(\sup_{f\in\mathcal{F}}\left[\frac{R_n(f)-\widehat{R}_{\mathbf{U}}(f)}{2R_n(f)} + \frac{R_n(f)-\widehat{R}_{\mathbf{V}}(f)}{2R_n(f)}\right] > \epsilon\right) && (6.11)\\
&\leqslant \mathbb{P}_{1:n}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}}(f)}{R_n(f)} + \sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{V}}(f)}{R_n(f)} > 2\epsilon\right) && (6.12)\\
&\leqslant \mathbb{P}_{\mathbf{U}}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}}(f)}{R_n(f)} > \epsilon\right) + \mathbb{P}_{\mathbf{V}}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{V}}(f)}{R_n(f)} > \epsilon\right) && (6.13)\\
&= 2\mathbb{P}_{\mathbf{U}}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}}(f)}{R_n(f)} > \epsilon\right). && (6.14)
\end{aligned}
$$

Now, apply Lemma 4.8 to the the event $\left\{\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}}(f)}{R_n(f)} > \epsilon\right\}$. This allows one to move from statements about dependent blocks to statements about independent blocks with a slight correction. Therefore,

$$
\begin{aligned}
&2\mathbb{P}_{\mathbf{U}}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}}(f)}{R_n(f)} > \epsilon\right) \\
&\leqslant 2\mathbb{P}_{\mathbf{U}'}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}'}(f)}{R_n(f)} > \epsilon\right) + 2(\mu-1)\beta_{a-d}
\end{aligned}
\tag{6.15}
$$

where the probability on the right is for the σ-field generated by the independent block sequence $\mathbf{U}'$. For convenience, define

$$
R_n^q(f) := \mathbb{E}\left[\|Y_{n+1} - f(\mathbf{Y}_1^n)\|^q\right] \tag{6.16}
$$

despite the obvious abuse of notation. Then,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{U}'}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}'}(f)}{R_n(f)}>\epsilon\right)\\
&=\mathbb{P}_{\mathbf{U}'}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}'}(f)}{R_n(f)}\frac{1}{M}>\frac{\epsilon}{M}\right) &(6.17)\\
&\leqslant\mathbb{P}_{\mathbf{U}'}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}'}(f)}{R_n(f)}\frac{R_n(f)}{\sqrt[q]{R_n^q(f)}}>\frac{\epsilon}{M}\right) &(6.18)\\
&=\mathbb{P}_{\mathbf{U}'}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}'}(f)}{\sqrt[q]{R_n^q(f)}}>\frac{\epsilon}{M}\right) &(6.19)\\
&=\mathbb{P}_{\mathbf{U}'}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_{\mathbf{U}'}(f)}{\sqrt[q]{R_n^q(f)}}>\tau(q)\frac{\epsilon}{M\tau(q)}\right) &(6.20)\\
&\leqslant 8\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})}+1\right)-\frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\}+2(\mu-1)\beta_{a-d}, &(6.21)
\end{aligned}$$

where I have applied Theorem 3.17 to bound the independent blocks. This result is Theorem 6.4. To prove the corollary, set the right hand side equal to η, taking $\eta'=\eta-2(\mu-1)\beta_{a-d}$, and solve for ϵ. Then for all $f\in\mathcal{F}$, with probability at least $1-\eta$,

$$\frac{R_n(f)-\widehat{R}_n(f)}{R_n(f)}\leqslant\epsilon. \tag{6.22}$$

Solving the equation

$$\eta'=8\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})}+1\right)-\frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\} \tag{6.23}$$

implies

$$\epsilon=\frac{2M\tau(q)}{\sqrt{\mu}}\sqrt{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})}+1\right)-\ln(\eta'/8)}=:\mathcal{E}. \tag{6.24}$$

■

The only obstacle to the use of Theorem 6.4 is knowledge of the $\mathrm{VCD}(\mathcal{F})$. For some models, the VC dimension can be calculated explicitly.

Theorem 6.6. *For $\mathcal{F}_{AR}(d)$ the class of AR(d) models*

$$\text{VCD}(\mathcal{F}_{AR}(d)) = d + 1. \tag{6.25}$$

Proof. The VC dimension of a linear classifier $f : \mathbb{R}^d \to \{0, 1\}$ is d (cf. Vapnik [97]). Real valued predictions have an extra degree of freedom. ■

Corollary 6.7. *The class of vector autoregressive models with d lags and k time series has VC dimension $kd + 1$.*

Proof. Here, I am interested in the VC dimension of a multivariate linear classifier. Thus, I must be able to shatter collections of vectors where each vector is a binary sequence of length k. For a VAR, each coordinate is independent, thus, I can shatter a collection of vectors if I can shatter each coordinate projection. The result then follows from Theorem 6.6. ■

Theorem 6.6 applies equally to Bayesian ARs. However, this is likely too conservative as the prior tends to restrict the effective complexity of the function class.[1]

6.1.3 *Growing memory*

Of course, the vast majority of macroeconometric forecasting models have growing memory rather than fixed memory. These model classes include dynamic factor models, ARMA models, and linearized dynamic stochastic general equilibrium models. However, all of these models have the property that forecasts are linear functions of past observations, and in particular, the weight placed on the past decays exponentially under suitable conditions. For this reason, I can recover bounds similar to my previous results even for state-space models.

1 Here I should mention that these risk bounds are frequentist in nature. My meaning is that if I treat Bayesian methods as a regularization technique and predict with the posterior mean or mode, then our results hold. However, from a subjective Bayesian perspective, our results add nothing since all inference can be derived from the posterior. For further discussion of the frequentist risk properties of Bayesian methods under mis-specification, see for example Kleijn and van der Vaart [51], Müller [69] or Shalizi [84]

Linear predictors with growing memory have the following form with $1 \leqslant d < n$:

$$\widehat{\mathbf{Y}}_{d+1}^{n+1} = \mathbf{B}\mathbf{Y}_1^n \tag{6.26}$$

where

$$\mathbf{B} = \begin{bmatrix} b_{d,1} & \cdots & b_{d,d} & & & \\ b_{d+1,1} & \cdots & b_{d+1,d} & b_{d+1,d+1} & & \mathbf{0} \\ \vdots & & \vdots & & \ddots & \\ b_{n,1} & \cdots & b_{n.d} & b_{n,d+1} & \cdots & b_{n,n} \end{bmatrix}. \tag{6.27}$$

With this notation, I can prove the following result about the growing memory linear predictor.

Theorem 6.8. *Given a sample $\mathbf{Y}_1^n$ such that Assumption B and Assumption C hold, suppose that the model class $\mathcal{F}$ is linear in the data and has growing memory. Fix some $1 \leqslant d < n$. Then the following bound holds simultaneously for all $f \in \mathcal{F}$ (including the minimizer of the empirical risk $\widehat{f}$). Let μ and a be integers such that $2\mu a + d \leqslant n$. Then,*

$$\begin{aligned} &\mathbb{P}_{1:n}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widetilde{R}_n(f) - \Delta_d(f)}{R_n(f)} > \tau(q)\epsilon\right) \\ &\leqslant 8\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})} + 1\right) - \frac{\mu\epsilon^2}{4\tau^2(q)M^2}\right\} + 2(\mu-1)\beta_{a-d}, \end{aligned} \tag{6.28}$$

where

$$\Delta_d(f) = \mathbb{E}\left[\|Y_1\|\right]\left\|\sum_{j=1}^{n-d} b_{n,j}\right\| + \frac{1}{n-d-1}\sum_{i=d+1}^{n-1}\left\|\sum_{j=1}^{i-d} b_{i,j}y_j\right\|. \tag{6.29}$$

The $\Delta_d(f)$ term deserves some explanation. It arrises by approximating the growing memory predictor with a finite sample version. The result is an implicit tradeoff: as $d \nearrow n$, $\Delta_d(f) \searrow 0$, but this drives $\mu \searrow 0$, resulting in fewer effective training points whereas larger d has the opposite effect. Also, $\Delta_d(f)$ depends on $\mathbb{E}\left[\|Y_1\|\right]$ which is not necessarily desirable. However, Assumption C has the consequence that there exists L such that $\mathbb{E}\left[\|Y_1\|\right] \leqslant L < \infty$. Finally, I will need $\sum_{j=1}^{n} \left|\left|b_{i,j}\right|\right|$ to be bounded $\forall n$ or $\Delta_d(f) \to \infty$ as $n \to \infty$.

Corollary 6.9. *The following bound holds simultaneously for all $f \in \mathcal{F}$ (including the minimizer of the empirical risk $\widehat{f}$). Let μ and a be integers such that $2\mu a + d \leqslant n$. Then, with probability at least $1-\eta$, for η as in Theorem 6.4,*

$$R_n(f) \leqslant \frac{\widetilde{R}_n(f) + \Delta_d(f)}{(1-\varepsilon)_+} \tag{6.30}$$

where ε and η' are as in Theorem 6.4.

Proof of Theorem 6.8 and Corollary 6.9. Let $\mathcal{F}$ be indexed by the parameters of the growing memory model. Let $\mathcal{F}'$ be the same class of models, but predictions are made based on the truncated memory length d. Then, for any $f \in \mathcal{F}$, and $f' \in \mathcal{F}'$

$$R_n(f) - \widetilde{R}_n(f) \leqslant (R_n(f) - R_n(f')) + (R_n(f') - \widehat{R}_n(f')) + (\widehat{R}_n(f') - \widetilde{R}_n(f)). \tag{6.31}$$

I will need to handle all three terms. The first and third terms are similar. Let $\mathbf{B}$ be as above and define the truncated linear predictor to have the same form but with $\mathbf{B}$ replaced by

$$\mathbf{B}' = \begin{bmatrix} b_{d,1} & b_{d,2} & \cdots & b_{d,d} & & & \\ & b_{d+1,2} & \cdots & b_{d+1,d} & b_{d+1,d+1} & \mathbf{0} & \\ & & & & \ddots & & \\ & & \mathbf{0} & & & & \\ & & & & & b_{n,n-d+1} & \cdots & b_{n,n} \end{bmatrix}. \tag{6.32}$$

Then notice that

$$\widehat{R}_n(f') - \widetilde{R}_n(f) \leqslant |\widehat{R}_n(f') - \widetilde{R}_n(f)| \tag{6.33}$$

$$= \left| \frac{1}{n-d-1} \sum_{i=d}^{n-1} \|Y_{i+1} - \mathbf{b}_i \mathbf{Y}_{i-d+1:i}\| \right.$$

$$\left. - \frac{1}{n-d-1} \sum_{i=d}^{n-1} \left|\left|Y_{i+1} - \mathbf{b}_i' \mathbf{Y}_{i-d+1:i}\right|\right| \right| \tag{6.34}$$

$$\leqslant \frac{1}{n-d-1} \sum_{i=d}^{n-1} \left|\left|(\mathbf{b}_i - \mathbf{b}_i')\mathbf{Y}_{i-d+1:i}\right|\right| \tag{6.35}$$

by the triangle inequality where $\mathbf{b}_i$ is the i^{th} row of $\mathbf{B}$ and analogously for $\mathbf{b}_i'$. Therefore

$$\widehat{R}_n(f') - \widetilde{R}_n(f) \leqslant \frac{1}{n-d-1} \sum_{i=d}^{n-1} \left|\left|(\mathbf{b}_i - \mathbf{b}_i')\mathbf{Y}_{i-d+1}^i\right|\right| \tag{6.36}$$

$$= \frac{1}{n-d-1} \sum_{i=d}^{n-1} \left|\left| \sum_{j=1}^{i-d} b_{i,j} y_j \right|\right| \tag{6.37}$$

For the case of the expected risk, I need only consider the first rows of $\mathbf{B}$ and $\mathbf{B}'$. Using linearity of expectations and stationarity

$$R_n(f) - R_n(f') \leqslant \mathbb{E}\left[\|Y_1\|\right] \left|\left| \sum_{j=1}^{n-d} b_{n,j} \right|\right|. \tag{6.38}$$

Then,

$$R_n(f) - \widetilde{R}_n(f) - \Delta_d(f) \leqslant R_n(f') - \widehat{R}_n(f') \tag{6.39}$$

where

$$\Delta_d(f) = \mathbb{E}\left[\|Y_1\|\right] \left|\left| \sum_{j=1}^{n-d} b_{n,j} \right|\right| + \frac{1}{n-d-1} \sum_{i=d}^{n-1} \left|\left| \sum_{j=1}^{i-d} b_{i,j} y_j \right|\right| \tag{6.40}$$

Divide through by $R_n(f)$ and take the supremum over $\mathcal{F}$ and $\mathcal{F}'$

$$\sup_{f \in \mathcal{F}} \frac{R_n(f) - \widetilde{R}_n(f) - \Delta_d(f)}{R_n(f)} \leqslant \sup_{f' \in \mathcal{F}', f \in \mathcal{F}} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f)}. \tag{6.41}$$

Finally,

$$\sup_{f \in \mathcal{F},\ f' \in \mathcal{F}'} \frac{R_n(f')}{R_n(f)} \leqslant 1 \tag{6.42}$$

since $\mathcal{F}' \subseteq \mathcal{F}$. So,

$$\begin{aligned} \sup_{f' \in \mathcal{F}', f \in \mathcal{F}} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f)} &= \sup_{f' \in \mathcal{F}', f \in \mathcal{F}} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f')} \frac{R_n(f')}{R_n(f)} && (6.43) \\ &\leqslant \sup_{f' \in \mathcal{F}'} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f')}. && (6.44) \end{aligned}$$

Now,

$$\mathbb{P}_{1:n}\left(\sup_{f \in \mathcal{F}} \frac{R_n(f) - \widetilde{R}_n(f) - \Delta_d(f)}{R_n(f)} > \epsilon\right) \leqslant \mathbb{P}_{1:n}\left(\sup_{f' \in \mathcal{F}'} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f')} > \epsilon\right). \tag{6.45}$$

Since $\mathcal{F}'$ is a class with finite memory, I can apply Theorem 6.4 and Corollary 6.5 to get the results. ■

To apply Theorem 6.8, I describe the form of the linear Gaussian state space model. I can then show how to calculate $\Delta_d(f)$ directly from the model and demonstrate that it will behave well as n grows rather than blowing up. Consider the following linear Gaussian state space model, $\mathcal{F}_{SS}$:

$$\begin{aligned} y_t &= A\alpha_t + \epsilon_t, & \epsilon_t &\sim N(0, H), \\ \alpha_{t+1} &= T\alpha_t + \eta_{t+1}, & \eta_t &\sim N(0, Q), \\ & & \alpha_1 &\sim N(a_1, P_1). \end{aligned} \tag{6.46}$$

I make no assumptions about the dimensionality of the parameter matrices A, T, H, Q, a_1, or P_1. The only requirement is stationarity. This amounts to requiring the eigenvalues of T to lie inside the complex unit circle. Stationarity ensures that $\Delta_d(f)$ will be bounded as well as conforming to the assumptions about the data generating process. While $\text{VCD}(\mathcal{F}_{SS})$ is unknown in general, I will actually only

Algorithm 1: Kalman filtering

Recursively generate minimum mean squared error predictions $\widehat{Y}_t$ using the state space model in (6.46).

Set $\widehat{Y}_1 = Aa_1$.
for $1 \leqslant t \leqslant n$ **do**
  Filter

$$\begin{aligned} v_t &= Y_t - \widehat{Y}_t, & F_t &= (AP_tA' + H)^{-1}, \\ K_t &= TP_tA'F_t, & L_t &= T - K_tZ, \\ a_{t+1} &= Ta_t + K_tv_t, & P_{t+1} &= TP_tL_t' + Q. \end{aligned}$$

  Predict

$$\widehat{Y}_{t+1} = Aa_{t+1}.$$

end
return $\widehat{\mathbf{Y}}_{1:n+1}$

need the VC dimension of the finite memory approximation. As I show below, this is linear in the data, so I can simply apply Theorem 6.6.

To forecast using $\mathcal{F}_{SS}$, one uses the Kalman filter [47]. The algorithm proceeds recursively as shown in Algorithm 1. To estimate the unknown parameter matrices, one can proceed in one of two ways: (1) maximize the likelihood returned by the filter; or (2) use the EM algorithm by running the filter and then the Kalman smoother which amounts to the E-step; then maximize the conditional likelihood using ordinary least squares. Bayesian estimation proceeds similarly to the EM approach replacing the M-step with standard Bayesian updates. In either case, one can show (cf. Durbin and Koopman [29]) that given the parameter matrices, the (maximum *a posteriori*) forecast of Y_t is given by

$$\widehat{Y}_{t+1} = A\sum_{j=1}^{t-1}\prod_{i=j+1}^{t} L_iK_jy_j + AK_ty_t + A\prod_{i=1}^{t} L_ia_1 \tag{6.47}$$

This yields the form of $\Delta_d(f)$ for linear state space models. I therefore have the following corollary to Theorem 6.8.

Corollary 6.10. *Let $1 < d < n$. Then the following bound holds simultaneously for all $f \in \mathcal{F}$ where $\mathcal{F}$ is a linear Gaussian state space model. With probability at least $1-\eta$, for η as in Theorem 6.4,*

$$R_n(f) \leqslant \frac{\widetilde{R}_n(f) + \Delta_d(f)}{(1-\varepsilon)_+} \tag{6.48}$$

where ε and η' is as in Theorem 6.4, and

$$\begin{aligned}\Delta_d(f) = \mathbb{E}\left[\|Y_1\|\right] \left\| \sum_{j=1}^{n-d} \prod_{i=j+1}^{n} L_i K_j \right\| \\ + \frac{1}{n-d-1} \sum_{t=d+1}^{n-1} \left\| \sum_{j=1}^{t-d} \prod_{i=j+1}^{t} L_i K_j y_j \right\|.\end{aligned} \tag{6.49}$$

Proof. This follows immediately from Corollary 6.9 and (6.47). ∎

It is simple to compute $\Delta_d(f)$ using Kalman filter output. The corollary allows me to compute risk bounds for wide classes of macroeconomic forecasting models. Dynamic factor models, ARMA models, GARCH models, and even linearized DSGEs have state space representations.

6.2 BOUNDS IN PRACTICE

The theory derived in the previous section is useful both for quantification of the prediction risk and for model selection. In this section, I show how to use some of the results above. I first estimate a simple stochastic volatility model using IBM return data and calculate the bound for the predicted volatility using Theorem 6.8.

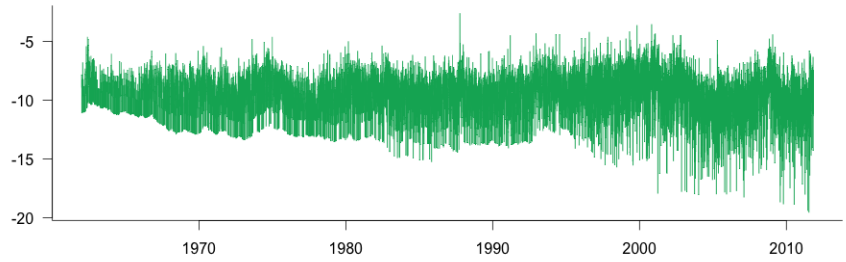


Figure 9: This figure plots daily volatility (squared log returns) for IBM from 1962–2011.

6.2.1 *Stochastic volatility model*

To demonstrate how to use my results, I estimate a standard stochastic volatility model using daily log returns for IBM from January 1962 until October 2011 which gives us $n = 12541$ observations. Figure 9 shows the squared log return series.

The model I investigate is given by

$$y_t = \sigma z_t \exp(\rho_t/2), \qquad z_t \sim N(0,1), \tag{6.50}$$

$$\rho_{t+1} = \phi\rho_t + w_t, \qquad w_t \sim N(0,\sigma_w^2), \tag{6.51}$$

where the disturbances z_t and w_t are mutually and serially independent. This model is nonlinear, but a linear approximation method can be used as in Harvey et al. [43]. I transform the model as follows:

$$\log y_t^2 = \kappa + \frac{1}{2}\rho_t + \xi_t, \tag{6.52}$$

$$\xi_t = \log z_t^2 - \mathbb{E}[\log z_t^2], \tag{6.53}$$

$$\kappa = \log\sigma^2 + \mathbb{E}[\log z_t^2]. \tag{6.54}$$

The noise term ξ_t is no longer normally distributed, but the Kalman filter will still give the minimum mean squared linear estimate of the variance sequence $\boldsymbol{\rho}_{1:n+1}$. Following the transformation, the observation variance is $\pi^2/2$.

To match the data to the model, let y_t be the log returns and remove 688 observations where the return was 0 (i.e., the price did not change from one day to the next). Using the Kalman filter, the negative log likelihood is given by

$$\mathcal{L}(\mathbf{Y}_1^n|\kappa, \phi, \sigma_\rho^2) \propto \sum_{t=1}^{n} \log F_t + v_t^2 F_t^{-1}.$$

Minimizing this gives estimates $\kappa = -9.62$, $\phi = 0.996$, and $\sigma_w^2 = 0.003$. Taking the loss function to be root mean squared error gives a training error of 1.823.

To actually calculate the bound, I need a few assumptions. First, using the methods in Chapter 5, I can estimate $\beta_8 = 0.017$ with 2 bins. For $a > 8$, the optimal number of bins is 1 implying an estimate of 0. While this is likely an underestimate, I will take $\beta_a = 0$ for $a > 8$. Second, take $q = 3$. This choice can be justified by assuming that the distribution of $Y_{n+1} - f(\mathbf{Y}_1^n)$ is standard normal. Then $\left|\left|y_{i+1} - f(\mathbf{Y}_1^i)\right|\right|_2$ has a χ distribution with one degree of freedom, in which case the q^{th} normalized moment M_q, is given in [46] as

$$M_q = \pi^{\frac{q-1}{2q}} \Gamma^{1/q} \left(\frac{q+1}{2}\right). \tag{6.55}$$

Using this formula, $M_3 = 1.46$.

Combining these assumptions with the VC dimension for the stochastic volatility model will allow us to calculate a bound for the prediction risk. For $d = 2$, the VC dimension can be no larger than 3, thus, I may use Corollary 6.10 with η as in Corollary 6.5, i.e., I can take the VC dimension to be 3. Finally, taking $\mu = 538$, $a = 11$, $d = 2$, and $\mathbb{E}||Y_1|| = 1$, I get that $\Delta_2(f) = 0.65 + 1.03 = 1.68$. The result is the bound

$$R_n(f) \leqslant 16.68 \tag{6.56}$$

with probability at least 0.85. In other words, the bound is much larger than the training error, but this is to be expected: the data are highly correlated and so despite the fact that n is large, the effective sample size μ is relatively small.

Model	Training error	AIC-Baseline	Risk bound ($1-\eta > 0.85$)
SV	1.83	-2816	16.68
AR(2)	1.88	-348	6.79
Mean	1.91	0	3.84

Table 1: This table shows the training error and risk bounds for 3 models. AIC is given as the difference from the mean the Mean, the smaller the value, the more support for that model.

For comparison, I also computed the bound for forecasts produced with an AR(2) model (with intercept) and with the mean alone. In the case of the mean, I take $\mu = 658$ and $a = 9$ since in this case, $d = 0$. The results are shown in Table 1. The stochastic volatility model reduces the training error by 5% over predicting with the mean, an increase which is marginal at best. But the resulting risk bound clearly demonstrates that given the small effective sample size, this gain may be spurious: it is likely that the stochastic volatility model is simply over-fitting.

6.2.2 *Real business cycle model*

In this section, I will discuss the methodology for using applying risk bounds to the forecasts generated by the real business cycle model presented in Section 2.3.

To estimate the parameters of this model, I use four data series. In the notation of Section 2.3, these are GDP y_t, consumption c_t, investment i_t, and hours worked n_t. The data are freely available from the Federal Reserve Economic Database (FRED). Appendix B gives the series names and data transformations necessary to replicate the data set that I use. The resulting data set is shown in Figure 10.

The estimation procedure for this model is quite complicated and therefore described more fully in Appendix B. The basic idea is to transform the model of Section 2.3 into a linear state space model with the four observed variables listed above and two unobserved state variables. There is a nonlinear mapping from un-

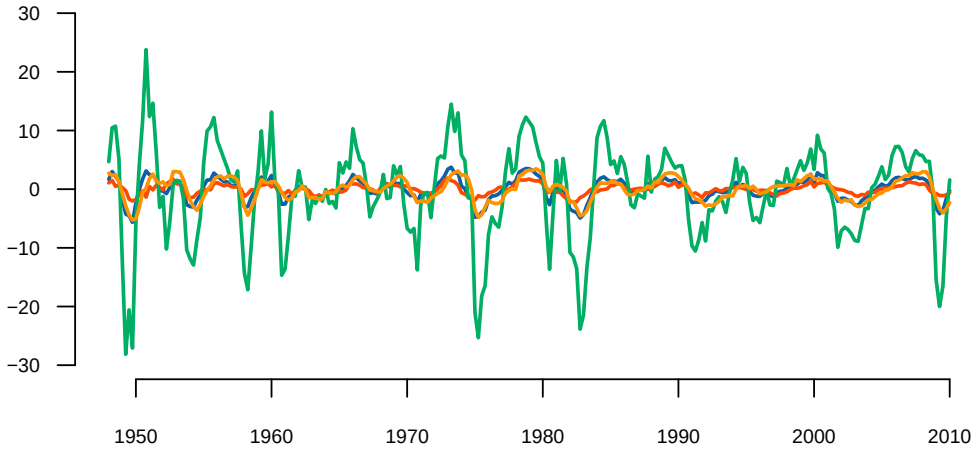


Figure 10: This figure shows the data used to estimate the RBC model. This is quarterly data from 1948:I until 2010:I. The blue line is GDP (output), the red line is consumption, the green line is investment, and the orange line is hours worked. These data are plotted as percentage deviations from trend as discussed in Appendix B.

known parameters to the parameters of the linear state space model, but for each parameter vector, the Kalman filter returns the likelihood, so that likelihood methods are possible. Because the data is uninformative about many of the parameters, I minimize the negative penalized likelihood to estimate them. Then the Kalman filter produces in sample forecasts which are linear in past values of the data so that I could potentially apply the growing memory bound.

For macroeconomic time series, there is not enough data to result in nontrivial bounds regardless of the mixing coefficients or the size of the finite memory approximation. The data shown in Figure 10 has $n = 248$ observations. The minimal possible finite approximation model is therefore a VAR with one lag and four time series, so by Corollary 6.7, it has VC dimension 5. Assuming, as above, that the third normalized moment of the loss function is bounded by $M_3 = 1.46$ and demanding confidence 0.85 ($\eta = 0.15$), then I would need 481 *independent* data vectors to have a non-trivial bound. Under these assumptions, models which have VC dimension 1 or 2 will result in non-trivial bounds for $n = 248$, but nothing

Separation a	# bins	β_a
1	5	0.25
2	4	0.17
3	3	0.03
4	1	0
4	2	0.10

Table 2: Estimated mixing coefficients for the multivariate time series $[y_t,\ c_t,\ i_t,\ n_t]$. I take $d = 1$. The final row shows if I had instead chosen two bins rather than one.

more complicated. Even if I am willing to reduce my confidence, say to $\eta = 0.5$, models with VC dimension larger than 2 are too complicated for this size data set. Allowing the data to be dependent only makes the situation worse.

Using the methods of Chapter 5, I can estimate the β-mixing coefficients of the macroeconomic data set. For the estimator given in (5.2), I take $d = 1$, and I use 5, 4 and 3 bins in the histograms for the lags $a \in \{1, 2, 3\}$. However, after $a = 3$, the estimated mixing coefficients increase, suggesting that the number of bins is too large. This increase also occurs when using either 2 or 4 bins. Together, this suggests, that the positive bias has kicked in, and I should estimate with 1 bin, implying an estimate of $\beta_4 = 0$. Assuming that this is approximately accurate (0 is of course an underestimate), this result suggests that the effective size of the macroeconomic data set is no more than about $\mu = 30$, much smaller then $n = 48$. Assuming $\beta_4 = 0$ and a confidence level of $1 - \eta = 0.85$, I would need around 15,000 quarterly data points to have a nontrivial bound, or about 3700 years of data. The estimated mixing coefficients are shown in Table 2.

In some sense, the empirical results in this section seem slightly unreasonable. Since the results are only upper bounds and may not be tight, it is important to have get an idea as to how tight they may be. I address this issue in the next section.

6.3 HOW LOOSE ARE THE BOUNDS?

In this section, I give some intuition as to how tight (or loose) the bounds presented in Section 6.1 may be. To gain some insight, I will investigate the following quantities

$$T_{erm}(\mathbb{P}_{1:n}) := \int d\mathbb{P}_{1:n}(\mathbf{Y}_{1:n}) R_n(\widehat{f}_{erm}), \tag{6.57}$$

$$T_0(\mathbb{P}_{1:n}) := R_n(f^*) \tag{6.58}$$

$$L(\mathbb{P}_{1:n}) := T_{erm}(\mathbb{P}_{1:n}) - T_0(\mathbb{P}_{1:n}) \tag{6.59}$$

$$L_M := \sup_{\mathbb{P}_{1:n} \in \Pi} L(\mathbb{P}_{1:n}), \tag{6.60}$$

where $\widehat{f}_{erm}$ is the function chosen by minimizing the training error over the class $\mathcal{F}$ and

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} R_n(f). \tag{6.61}$$

Here $\mathbb{P}_{1:n}$ is the joint distribution of a sequence $\mathbf{Y}_{1:n}$. The risk R_n is an expectation taken with respect to the next time point Y_{n+1}, whereas $T_{erm}(\mathbb{P}_{1:n})$ removes the randomness in the procedure to choose $\widehat{f}_{erm}$. I will consider a class of distributions Π is a of which $\mathbb{P}_{1:n}$ is a member.

I will refer to L_M as the "oracle classification loss". It describes how well empirical risk minimization works relative to the best possible predictor $\widehat{f} \in \mathcal{F}$ over the worst distribution π. Vapnik [96] shows that for classification and IID data, for sufficiently large n, there exist constants c and C such that

$$c\sqrt{\frac{h}{n}} \leqslant L_M \leqslant C\sqrt{\frac{h \log n/h}{n}}, \tag{6.62}$$

where $h = \text{VCD}(\mathcal{F})$. In other words, there is a gap of $\log n$ in the rates. Using Rademacher complexity, it is possible to remove this gap, but I will take (6.62) as the baseline and compare results for dependent data to it.

I will derive similar bounds for the β-mixing setting. First I state the following result.

Theorem 6.11. *Given a sample* $\mathbf{Y}_1^n$ *such that Assumption A and Assumption C hold, suppose that the model class* $\mathcal{F}$ *has a fixed memory length* $d < n$. *Let* μ *and* a *be integers such that* $2\mu a + d \leqslant n$. *Then, for all* $\epsilon > 0$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{f\in\mathcal{F}} |R_n(f) - \widehat{R}_n(f)| > \epsilon\right)\\ &\leqslant 8\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})}+1\right) - \frac{\mu\epsilon^2}{M^2}\right\} + 2(\mu-1)\beta_{a-d}.\end{aligned} \tag{6.63}$$

The proof of Theorem 6.11 is exactly like that for Theorem 6.4.

Assumption D. *The time series* $\mathbf{Y}$ *is* exponentially β*-mixing, i. e.*

$$\beta_a = c_1 \exp(-c_2 a^\kappa) \tag{6.64}$$

for some constants c_1 *and* c_2 *and some parameter* κ.

Theorem 6.12. *Under Assumption A and Assumption D, for sufficiently large* n, *there exist constants* c *and* C, *independent of* n *and* h, *such that*

$$c\sqrt{\frac{h}{n}} \leqslant L_M \leqslant C\sqrt{\frac{h\log n^{\kappa/(1+\kappa)}/h}{n^{\kappa/(1+\kappa)}}}. \tag{6.65}$$

Proof. Theorem 6.11 implies that simultaneously

$$\begin{aligned}&\mathbb{P}_{1:n}\left(|R_n(\widehat{f}_{erm}) - \widehat{R}_n(\widehat{f}_{erm})| > \epsilon\right)\\ &\leqslant 8\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})}+1\right) - \frac{\mu\epsilon^2}{M^2}\right\} + 2(\mu-1)\beta_{a-d}\end{aligned} \tag{6.66}$$

and

$$\begin{aligned}&\mathbb{P}_{1:n}\left(|R_n(f^*) - \widehat{R}_n(f^*)| > \epsilon\right)\\&\leqslant 8\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})}+1\right) - \frac{\mu\epsilon^2}{M^2}\right\} + 2(\mu-1)\beta_{a-d}.\end{aligned} \tag{6.67}$$

Since $\widehat{R}_n(\widehat{f}_{erm}) - \widehat{R}_n(f^*) \leqslant 0$, then

$$\begin{aligned}&\mathbb{P}_{1:n}\left(|R_n(\widehat{f}_{erm}) - R_n(f^*)| > 2\epsilon\right)\\&\leqslant 8\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\mathrm{VCD}(\mathcal{F})}+1\right) - \frac{\mu\epsilon^2}{M^2}\right\} + 2(\mu-1)\beta_{a-d}.\end{aligned} \tag{6.68}$$

Then, letting $Z = |R_n(\widehat{f}_{erm}) - R_n(f^*)|$, $k_1 = 8GF(2\mu_n, h)$, and $k_2 = 1/M^2$, proceeding as in the proof of Corollary 3.10, and ignoring constants,

$$\mathbb{E}[Z^2] \leqslant s + k_1'\int_s^M d\epsilon e^{-k_2\mu_n\epsilon} + 4\int_0^M d\epsilon\mu_n\beta_{a_n-d} \tag{6.69}$$

$$L_O \leqslant s + k_1'\int_s^\infty d\epsilon e^{-k_2\mu_n\epsilon} + 4\int_0^M d\epsilon\mu_n\beta_{a_n-d} \tag{6.70}$$

$$= s + \frac{k_1' e^{-k_2\mu_n\epsilon}}{k_2\mu_n} + k_3\mu_n\beta_{a_n-d}. \tag{6.71}$$

Using Assumption D, take $a_n = n^{1/(1+\kappa)}$, $\mu_n = n^{\kappa/(1+\kappa)}$, and $s = \frac{\log k_1'}{n^{\kappa/(1+\kappa)}k_2}$ to balance the exponential terms and linear terms. Then,

$$L_M = O\left(\sqrt{\frac{h\log n^{\kappa/(1+\kappa)}/h}{n^{\kappa/(1+\kappa)}}}\right). \tag{6.72}$$

For the lower bound, apply the IID version. ■

If I instead assume *algebraic mixing*, i.e. $\beta_a = c_1 a^{-r}$, then I can retrieve the same rate where $0 < \kappa < (r-1)/2$ (see Meir [65]). Theorem 6.12 says that in dependent data settings, using the blocking approach developed here, I pay a penalty. In the worst case of exponential mixing where $\kappa = 1$, that penalty is an extra square root

factor. That said, as I will argue in the Chapter 8, the linear term in the risk bound due to the blocking technique may be massive overkill.

6.4 STRUCTURAL RISK MINIMIZATION

My presentation so far has focused on choosing one function $\widehat{f}$ from a model $\mathcal{F}$ and demonstrating that the prediction risk $R_n(\widehat{f})$ is well characterized by the training error inflated by a complexity term. The procedure for actually choosing $\widehat{f}$ has been ignored. Common ways of choosing $\widehat{f}$ are frequently referred to as *empirical risk minimization* or ERM: approximate the expected risk $R_n(f)$ with the empirical risk $\widehat{R}_n(f)$, and choose $\widehat{f}$ to minimize the empirical risk. Many likelihood based methods have exactly this flavor, but more frequently, forecasters have many different models in mind, each with a different empirical risk minimizer.

Regularized model classes (ridge regression, lasso, Bayesian methods) implicitly have this structure — altering the amount of regularization leads to different models $\mathcal{F}$. Methods like these are given by an optimization problem like

$$\widehat{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{R}_n(f) + \lambda \|f\|, \tag{6.73}$$

where $\|\cdot\|$ is some appropriate norm on the function space containing $\mathcal{F}$. This is the Lagrange dual of the constrained optimization problem

$$\begin{aligned} &\min_{f \in \mathcal{F}} \widehat{R}_n(f) \\ &\text{s.t} \quad \|f\| < C, \end{aligned} \tag{6.74}$$

for some constant C. Thus, C further constrains the allowable model space to $\mathcal{F}_C \subseteq \mathcal{F}$. Increasing C (or decreasing λ) leads to larger model classes up to the full class $\mathcal{F}$. More simply, one may just have many different forecasting models from which to choose the best. These scenarios leads to a generalization of ERM called *structural risk minimization* or SRM.

Given a collection of models $\mathcal{F}_1, \mathcal{F}_2, \ldots$ each with associated empirical risk minimizers $\widehat{f}_1, \widehat{f}_2, \ldots$, one wishes to use the function which has the smallest risk. Of course different models have different complexities, and those with larger complexities will tend to have smaller empirical risk. To choose the best function, one therefore penalizes the empirical risk and selects that function which minimizes the penalized version. Model selection tools like AIC or BIC have exactly this form, but they rely on specific knowledge of the data likelihood and use asymptotic approximations to derive an appropriate penalty. In contrast to these methods, I have derived finite sample bounds for the expected risk. This leads to a natural procedure for model selection — choose the predictor which has the smallest bound on the expected risk.

The generalization error bounds in Section 6.1 allow me to perform model selection via the SRM principle without knowledge of the likelihood or appeals to asymptotic results. The penalty accounts for the complexity of the model through the VC dimension. Most useful however is that by using generalization error bounds for model selection, we are minimizing the prediction risk.

If I want to make the prediction risk as small as possible, I can minimize the generalization error bound simultaneously over models $\mathcal{F}$ and functions within those models. This amounts to treating VC dimension as a control variable. Therefore, just like with AIC, I can minimize simultaneously over the empirical risk and the VC dimension. Using the risk bound and following this minimization procedure will lead to choosing the model and function which has the smallest prediction risk, a claim which other model selection procedures cannot make [97, 62].

6.5 CONCLUSION

This chapter demonstrates how to control the generalization error of common macroeconomic forecasting models — ARMA models, vector autoregressions (Bayesian or otherwise), linearized dynamic stochastic general equilibrium models, and

linear state space models. The results I derive give upper bounds on the risk which hold with high probability while requiring only weak assumptions on the true data generating process. These are finite-sample bounds, unlike standard model selection penalties (AIC, BIC, etc.), which only work asymptotically. Furthermore, they do not suffer the biases inherent in other risk estimation techniques such as the pseudo-cross validation approach often used in the economic forecasting literature.

While I have stated these results in terms of standard economic forecasting models, they have very wide applicability. Theorem 6.4 applies to any forecasting procedure with fixed memory length, linear or non-linear. This covers even non-linear DSGEs as long as the forecasts are based on only a fixed amount of past data. The unknown parameters can still be estimated using the entire data set. The results in Theorem 6.8 apply only to methods whose forecasts are linear in the observations, but a similar result could conceivably be derived for nonlinear methods as long as the dependence of the forecast on the past decays in some suitable way.

The bounds I have derived here are the first of their kind for time series forecasting methods typically used in economics, but there are some results for other types of forecasting methods as in Meir [65] and Mohri and Rostamizadeh [66, 67]. Those results require bounded loss functions, as in the IID setting, making them less general than my results, as well as turning on specific forms of regularization which are more rare in economics. For another view on this problem, McDonald et al. [64] shows that using stationarity alone to regularize an AR model leads to bounds which are much worse than those obtained here, despite the stricter assumption of bounded loss.

7

OTHER BOUNDS

In Chapter 6, I used mixing to breed dependent data laws of large numbers from the concentration results for IID random variables in Theorem 3.5 and Theorem 3.6. In this chapter, I take a different approach: I use concentration results for dependent data and show that the corresponding Rademacher complexity is very similar to standard cases.

7.1 CONCENTRATION INEQUALITIES

For IID data, the main tools for developing risk bounds are the inequalities of Hoeffding [45] and McDiarmid [63]. Instead, I will use dependent versions of each which generalize the IID results. These inequalities are derived in van de Geer [95]. They rely on constructing predictable bounds for random variables based on past behavior, rather than assuming *a priori* knowledge of the distribution.

Theorem 7.1 (van de Geer [95] Theorem 2.5)**.** *Consider a random sequence* $\mathbf{Y}_{1:n}$ *where*

$$L_i \leqslant Y_i \leqslant U_i \text{ a.s. for all } i \geqslant 1, \tag{7.1}$$

where $L_i < U_i$ *are* $\sigma_{1:i-1}$*-measurable random variables,* $i \geqslant 1$*. Define*

$$C_n^2 = \sum_{i=1}^{n} (U_i - L_i)^2, \tag{7.2}$$

with the convention $C_0^2 = 0$. *Then for all* $\epsilon > 0$, $c > 0$,

$$\mathbb{P}_{1:n}\left(\sum_{i=1}^{n} Y_i \geqslant \epsilon \text{ and } C_n^2 \leqslant c^2 \text{ for some } n\right) \leqslant \exp\left\{-\frac{2\epsilon^2}{c^2}\right\}. \tag{7.3}$$

Of course if L_i and U_i are non-random, then Theorem 7.1 is the same as the usual Hoeffding inequality. Here however, they must only be forecastable given past values of the random sequence, not forecastable *a priori*.

Theorem 7.2 (van de Geer [95] Theorem 2.6)**.** *Fix* $n \geqslant 1$. *Let* Y_n *be* $\sigma_{1:n}$*-measurable such that*

$$L_i \leqslant \mathbb{E}[Y_n \mid \sigma_{1:i}] \leqslant U_i, \text{ a.s.} \tag{7.4}$$

where $L_i < U_i$ *are* $\sigma_{1:i-1}$*-measurable. Define* C_n^2 *as above. Then for all* $\epsilon > 0$, $c > 0$,

$$\mathbb{P}_{1:n}\left(Y_n - \mathbb{E}[Y_n] \geqslant \epsilon \text{ and } C_n^2 \leqslant c^2\right) \leqslant \exp\left\{-\frac{2\epsilon^2}{c^2}\right\}. \tag{7.5}$$

I will refer to (7.4) as "forecastable boundedness". To see how this generalizes McDiarmid's inequality, I provide the following corollary.

Corollary 7.3. *Let* $f(Y_1, \ldots, Y_n)$ *be some real valued function on* $\mathcal{Y}^n$ *such that*

$$\left|\mathbb{E}[f(Y_1, \ldots, Y_n) \mid \sigma_{1:i}] - \mathbb{E}[f(Y_1, \ldots, Y_n) \mid \sigma_{1:i-1}]\right| \leqslant k_i \tag{7.6}$$

where k_i *is* $\sigma_{1:i-1}$*-measurable. Then,*

$$\begin{aligned}\mathbb{P}_{1:n}\left(f(Y_1, \ldots, Y_n) - \mathbb{E}[f(Y_1, \ldots, Y_n)] > \epsilon \text{ and } \sum_i k_i^2 < c^2\right) \\ < \exp\left\{-\frac{2\epsilon^2}{c^2}\right\}.\end{aligned} \tag{7.7}$$

In particular, this gives a couple of immediate consequences. Suppose that f is bounded. Then,

$$\begin{aligned} k_i &\leqslant \sup_{\mathbf{Y}_i^n} \sup_{\mathbf{Y}_i^{n\prime}} |f(Y_1, \ldots, Y_{i-1}, Y_i, \ldots, Y_n) - f(Y_1, \ldots, Y_{i-1}, Y_i', \ldots, Y_n')| \\ &=: b_i. \end{aligned} \tag{7.8}$$

This contrasts with McDiarmid's inequality in the IID case, wherein one only needs to be concerned with one point that is different. For IID data, starting from (7.6),

$$\begin{aligned} k_i &\leqslant \sup_{Y_i, Y_i'} |f(Y_1, \ldots, Y_{i-1}, Y_i, \ldots, Y_n) - f(Y_1, \ldots, Y_{i-1}, Y_i', \ldots, Y_n)| \\ &=: d_i, \end{aligned} \tag{7.9}$$

if f satisfies bounded differences with constants d_i. In other words, Theorem 7.2 conflates dependence with nice functional behavior.

7.2 RISK BOUNDS

Recall as in Chapter 3 that generalization error bounds can follow from deriving high probability upper bounds on the quantity

$$\Psi_n = \sup_{f \in \mathcal{F}} \left(R(f) - \widehat{R}_n(f) \right), \tag{7.10}$$

which is the worst case difference between the true risk $R(f)$ and the empirical risk $\widehat{R}_n(f)$ over all functions in the class $\mathcal{F}$. In the case of time series, Ψ_n is $\sigma_{1:n}$-measurable, so one can get risk bounds from Theorem 7.2 if one can find suitable L_i and U_i sequences.

Theorem 7.4. *Suppose that Ψ_n satisfies the forecastable boundedness condition (7.4) of Theorem 7.2. Then, for any $0 < \eta \leqslant 1$,*

$$\mathbb{P}\left(R(h) < \widehat{R}_n(h) + \mathbb{E}[\Psi_n] + c\sqrt{\frac{\log 1/\eta}{2}} \text{ or } C_n^2 > c\right) \leqslant 1 - \eta. \tag{7.11}$$

Proof. Applying Theorem 7.2 to the random variable Ψ_n gives

$$\mathbb{P}\left(\Psi_n - \mathbb{E}[\Psi_n] \geqslant \epsilon \text{ and } C_n^2 \leqslant c^2\right) \leqslant \exp\left\{-\frac{2\epsilon^2}{c^2}\right\}. \tag{7.12}$$

Setting the right side of (7.12) equal to η and solving for ϵ gives

$$\epsilon = c\sqrt{\frac{\log 1/\eta}{2}}. \tag{7.13}$$

Substitution and an application of DeMorgan's Law gives the result. ∎

In many cases (as in the examples below), C_n^2 will be deterministic, in which case, the result above is greatly simplified. Essentially, the theorem says that as long as each new Y_i gives additional control on the conditional expectation of Ψ_n, one can ensure that with high probability, forecasts of the future will have only small losses.

Since $\mathbb{E}_{\mathbb{P}_{1:n}}[\Psi_n]$ is often difficult to calculate, I upper bound it with the Rademacher complexity. The standard symmetrization argument for the IID case does not work, but, for time series prediction (as opposed to the more general dependent data case or the online learning case), Rademacher bounds are still available.

Theorem 7.5. *For a time series prediction problem based on a sequence $\mathbf{Y}_{1:n}$,*

$$\mathbb{E}_{\mathbb{P}_{1:n}}[\Psi_n] \leqslant \mathfrak{R}_n(\ell \circ \mathcal{F}). \tag{7.14}$$

The standard way of proving this result in the IID case given in Lemma 3.14 is through introduction of a "ghost sample" $\mathbf{Y}'_{1:n}$ which has the same distribution as $\mathbf{Y}_{1:n}$. Taking empirical expectations over the ghost sample is then the same as

taking expectations with respect to the distribution of $\mathbf{Y}_{1:n}$. Randomly exchanging Y_i with Y_i' by using Rademacher variables allows for control of $\mathbb{E}_{\mathbb{P}_{1:n}}[\Psi_n]$ and leads to the factor of 2 in Definition 3.13. However, in the dependent data setting, this is not quite so easy.

For dependent data, both the ghost sample and the introduction of Rademacher variables arise differently. A similar situation also occurs in the more complex cases of online learning with a (perhaps constrained) adversary choosing the data sequence. It is covered in depth in Rakhlin et al. [78, 79]. With dependent data I will need a different version of the "ghost sample" than that used in the IID case.

Proof of Theorem 7.5. First, rewrite the left side of (7.14):

$$\mathbb{E}_{\mathbb{P}_{1:n}}[\Psi_n] = \mathbb{E}_{\mathbb{P}_{1:n}}\left[\sup_{f\in\mathcal{F}}\left(R_n(f) - \widehat{R}_n(f)\right)\right] \tag{7.15}$$

$$= \mathbb{E}_{\mathbb{P}_{1:n}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbb{P}_{1:n}}[\ell(Y_{n+1}, f(\mathbf{Y}_{1:n}))] - \frac{1}{n}\sum_{i=1}^{n}\ell(Y_{i+1}, f(\mathbf{Y}_{1:n}))\right)\right]. \tag{7.16}$$

At this point, following [78, 79], I introduce a "tangent sequence" $\mathbf{Y}'_{1:n}$. Construct it recursively as follows. Let,

$$\mathcal{L}(Y_1') = \mathcal{L}(Y_1) \tag{7.17}$$

and

$$\mathcal{L}(Y_i'|Y_1, \ldots, Y_{i-1}) = \mathcal{L}(Y_i|Y_1, \ldots, Y_{i-1}). \tag{7.18}$$

Starting from (7.16)

$$\mathbb{E}[\Psi_n] = \mathbb{E}_{\mathbb{P}_{1:n}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbb{P}_{1:n}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(Y_{i+1}, \mathbf{Y}_{1:i})\right] - \frac{1}{n}\sum_{i=1}^{n}\ell(Y_{i+1}, \mathbf{Y}_{1:i})\right)\right] \tag{7.19}$$

$$= \mathbb{E}_{\mathbf{Y}_{1:n}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbf{Y}'_{1:n}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(Y_{i+1}, \mathbf{Y}'_{1:i})\right] - \frac{1}{n}\sum_{i=1}^{n}\ell(Y_{i+1}, \mathbf{Y}_{1:i})\right)\right]. \tag{7.20}$$

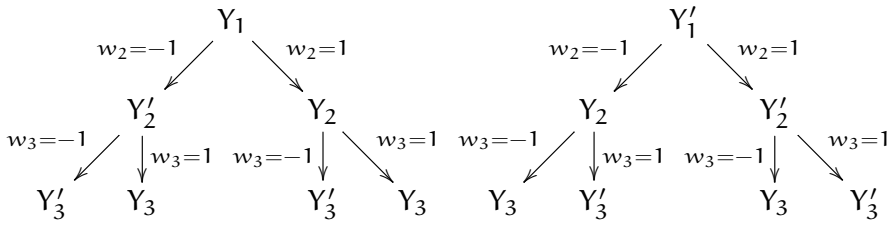


Figure 11: This figure displays the tree structures for $\mathbf{Y}(\mathbf{w})$ and $\mathbf{Y}'(\mathbf{w})$. The path along each tree is determined by one $\mathbf{w}$ sequence, interleaving the "past" between paths.

Then,

$$(7.20) \leqslant \mathbb{E}_{\mathbf{Y}_{1:n},\mathbf{Y}'_{1:n}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\ell(Y_{i+1},\mathbf{Y}'_{1:i})-\ell(Y_{i+1},\mathbf{Y}_{1:i})\right] \tag{7.21}$$

$$=\mathop{\mathbb{E}_{Y_1}}_{Y_1'}\mathop{\mathbb{E}_{Y_2|Y_1}}_{Y_2'|Y_1'}\cdots\mathop{\mathbb{E}_{Y_n|\mathbf{Y}_{1:n-1}}}_{Y_n'|\mathbf{Y}'_{1:n-1}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\ell(Y_{i+1},\mathbf{Y}'_{1:i})-\ell(Y_{i+1},\mathbf{Y}_{1:i})\right], \tag{7.22}$$

where (7.22) is by Jensen's inequality. Now, due to dependence, Rademacher variables must be introduced, carefully as in the adversarial case [78, 79]. Rademacher variables create two tree structures, one associated to the $\mathbf{Y}_{1:n}$ sequence, and one associated to the $\mathbf{Y}'_{1:n}$ sequence. I write these trees as $\mathbf{Y}(\mathbf{w})$ and $\mathbf{Y}'(\mathbf{w})$, where $\mathbf{w}$ is a particular sequence of Rademacher variables (e.g. $(1,-1,-1,1,\ldots,1)$) which creates a path along each tree. For example, consider $\mathbf{w}=\mathbf{1}$. Then, $\mathbf{Y}(\mathbf{w})=(Y_1,\ldots,Y_n)$ and $\mathbf{Y}'(\mathbf{w})=(Y_1',\ldots,Y_n')$, the "always-move-right" path of both tree structures. For $\mathbf{w}=-\mathbf{1}$. Then, $\mathbf{Y}(\mathbf{w})=(Y_1',\ldots,Y_n')$ and $\mathbf{Y}'(\mathbf{w})=(Y_1,\ldots,Y_n)$, the "always-move-left" path of both tree structures. Figure 11 shows the root of the two tree structures.

Changing w_i from $+1$ to -1 exchanges Y_i for Y_i' in both trees and chooses the left child of Y_{i-1} and Y_{i-1}' rather than the right child. In order to talk about the

probability of Y_i conditional on the "past" in the tree, one needs to know the path taken so far. For this, define a selector function

$$\chi(w) := \chi(w, y, y') = \begin{cases} y' & w = 1 \\ y & w = -1. \end{cases} \tag{7.23}$$

Distributions over the trees given by the selector functions then become the objects of interest.

In the time series case, as opposed to the online learning scenario considered in [78], the dependence between future and past means the adversary is *not* free to change predictors and responses separately. Once a branch of the tree is chosen, the distribution of future data points is fixed, and depends only on the preceding sequence. Because of this, the joint distribution of any path along the tree is the same as any other path, i.e. for any two paths $\mathbf{w}, \mathbf{w}'$

$$\mathcal{L}(\mathbf{Y}(\mathbf{w})) = \mathcal{L}(\mathbf{Y}(\mathbf{w}')) \qquad \text{and} \qquad \mathcal{L}(\mathbf{Y}'(\mathbf{w})) = \mathcal{L}(\mathbf{Y}'(\mathbf{w}')). \tag{7.24}$$

Similarly, due to the construction of the tangent sequence, $\mathcal{L}(\mathbf{Y}(\mathbf{w})) = \mathcal{L}(\mathbf{Y}'(\mathbf{w}))$. This equivalence between paths allows us to introduce Rademacher variables swapping Y_i for Y_i' as well as the ability to combine terms.

$$
\begin{aligned}
(7.22) &= \mathbb{E}_{\substack{Y_1 \\ Y_1'}} \mathbb{E}_{w_1} \mathbb{E}_{\substack{Y_2 | \chi(w_1, Y_1, Y_1') \\ Y_2' | \chi(w_1, Y_1', Y_1)}} \mathbb{E}_{w_2} \cdots \mathbb{E}_{\substack{Y_n | \chi(w_{n-1}), \ldots, \chi(w_1) \\ Y_n' | \chi(w_{n-1}), \ldots, \chi(w_1)}} \cdots \\
&\quad \cdots \mathbb{E}_{w_n} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} w_i (h(Y_i') - h(Y_i)) \right] && (7.25) \\
&= \mathbb{E}_{\mathbf{Y}, \mathbf{Y}', \mathbf{w}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} w_i (\ell(Y_{i+1}, \mathbf{Y}'_{1:i}) - \ell(Y_{i+1}, \mathbf{Y}_{1:i})) \right] && (7.26) \\
&\leqslant \mathbb{E}_{\mathbf{Y}', \mathbf{w}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} w_i \ell(Y_{i+1}, \mathbf{Y}'_{1:i}) \right] + \mathbb{E}_{\mathbf{Y}, \mathbf{w}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} w_i \ell(Y_{i+1}, \mathbf{Y}_{1:i}) \right] && (7.27) \\
&= 2\mathbb{E}_{\mathbf{Y}, w} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} w_i \ell(Y_{i+1}, \mathbf{Y}_{1:i}) \right] && (7.28) \\
&= \mathfrak{R}_n(\ell \circ \mathcal{F}). && (7.29)
\end{aligned}
$$

■

While Theorem 7.5 allows me to recover something that looks like the standard Rademacher complexity, it is not quite so simple. Here the expectation is with respect to a dependent sequence rendering it slightly less intuitive. However, another application of Theorem 7.2 yields an empirical version which concentrates around its mean with high probability exactly as in Bartlett and Mendelson [5].

The main issue then in the application of Theorem 7.4 is the determination of the forecastable bounds L_i and U_i from the data generating process. In the next section, I provide a few simple examples to aid intuition.

7.3 EXAMPLES

I consider three different examples which should aid in understanding the nature of the forecastable bounds. Here I present two extreme cases — independence and

complete dependence — as well as an intermediate case. It is important to note that C_n^2 is deterministic in all three cases, though this need not be the case.

7.3.1 *Independence*

For IID data, one simply recovers IID concentration results. As noted in Corollary 7.3, for IID data, the method of bounded differences yields good control. Similarly, Theorem 7.1 gives the same results as Hoeffding's inequality for IID data. Dependence is more interesting.

7.3.2 *Complete dependence*

Let $\mathbf{Y}_{1:n}$ be generated as follows:

$$Y_1 \sim U(a, b), \quad b > a \qquad\qquad Y_i = Y_{i-1}, \quad i \geqslant 2. \tag{7.30}$$

Consider trying to predict the mean $\frac{1}{n}\sum_{i=1}^{n} Y_i$. Then, given no observations, the almost sure upper bound $U_1 = b$ while the lower bound $L_1 = a$. So $(U_1 - L_1)^2 = (b-a)^2$. For $i > 1$, conditional on $\sigma_{1:i}$ (and therefore σ_1), $U_i = L_i$. Thus, $C_n^2 = (b-a)^2$ giving the entirely useless result:

$$\mathbb{P}_{1:n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - (b+a)/2 \geqslant \epsilon\right) < \exp\left\{-\frac{2\epsilon^2}{(b-a)^2}\right\}. \tag{7.31}$$

The right side is independent of n implying that one has essentially observed one data point regardless of n.

7.3.3 *Partial dependence*

Let $\mathbf{Y}_{1:n}$ be generated as follows:

$$Y_0 = 0, \qquad Y_i = \theta Y_{i-1} + \eta_i, \quad i \geqslant 2, \tag{7.32}$$

where $\theta \in (0,1)$ and $\eta_i \overset{iid}{\sim} U(a,b)$ with $b > a$. Again, consider trying to predict the mean $\frac{1}{n}\sum_{i=1}^{n} Y_i$. Define L_i and U_i as follows:

$$L_i = \frac{a}{n}\frac{1-\theta^{n-i}}{1-\theta} + \frac{1}{n}\sum_{k=1}^{i-1} Y_k + \theta Y_{i-1}, \tag{7.33}$$

$$U_i = \frac{b}{n}\frac{1-\theta^{n-i}}{1-\theta} + \frac{1}{n}\sum_{k=1}^{i-1} Y_k + \theta Y_{i-1}. \tag{7.34}$$

From this,

$$C_n^2 = \sum_{i=1}^{n} \frac{(b-a)^2}{n^2(1-\theta)^2}\left(1-\theta^{n-i}\right)^2 \tag{7.35}$$

$$= \frac{(b-a)^2}{n^2(1-\theta)^2(\theta^2-1)}\left(\theta^{2n} - 2\theta^{n+1} - 2\theta^n + n\theta^2 + 2\theta - n + 1\right) \tag{7.36}$$

$$< \frac{(b-a)^2}{n(1-\theta)^2}. \tag{7.37}$$

Therefore, by Theorem 7.2,

$$\mathbb{P}_{1:n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - (b+a)/2 > \epsilon\right) < \exp\left\{-\frac{2n\epsilon^2(1-\theta)^2}{(b-a)^2}\right\}. \tag{7.38}$$

For comparison, if everything was IID, Hoeffding's inequality gives

$$\mathbb{P}_{1:n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - (b+a)/2 > \epsilon\right) < \exp\left\{-\frac{2n\epsilon^2}{(b-a)^2}\right\}. \tag{7.39}$$

Therefore, the dependence in $\mathbf{Y}_1^n$ reduces the effective sample size by $(1-\theta)^2$. If $\theta = 1/2$, then each additional datapoint decreases the probability of a bad event by only a $1/4$ relative to the IID scenario.

7.4 DISCUSSION

In this chapter, I have demonstrated how to control the generalization of time series prediction algorithms. These methods use some or all of the observed past to predict future values of the same series. In order to handle the complicated Rademacher complexity bound for the expectation, I have followed the approach used in the online learning case pioneered by Rakhlin et al. [78, 79], but I show that in this particular case, much of the structure needed to deal with the adversary is unnecessary. This results in clean risk bounds which have a form similar to the IID case.

The main issue with risk bounds for dependent data is that they rely on knowledge of the dependence for application. This is certainly true in this case in that I need to *know* how to choose U_i and L_i such that I almost surely control $\mathbb{E}[\Psi_n]$. For the standard case of bounded loss, there are trivial bounds, but these will not give the necessary dependence on n which would imply learnability of good predictors. More knowledge of the dependence structure of the process is required, though this is in some sense undesirable. Results in the previous chapter also have this requirement.[1] They rely on precise knowledge of the mixing behavior of the data which is unavailable. At the same time, mixing characterizations are often unintuitive conditions based on infinite dimensional joint distributions. The version here depends only on the ability to forecastably bound expectations given increasing amounts of data which is perhaps more natural in applied settings.

1 IID results have an even more onerous requirement: one must be able to rule out any dependence at all.

Part IV

CONCLUSION

8

ADVANCING FURTHER

There are a number of directions future work along the lines pursued herein. In the framework of Chapter 6, it is necessary to know the VC dimension of the model class $\mathcal{F}$ in order to use my bounds. However, this knowledge may be unavailable in practice. Second, the bounds presented in Chapter 6 are often quite loose for a number of theoretical reasons, and it should be possible to tighten them. A third potential extension would be to derive more data-driven methods of establishing risk bounds. In the next few sections, I address each of these issues and give some thoughts as to how future analysis might proceed.

8.1 MEASURING VC DIMENSION

Previous work in Vapnik et al. [99] and Shao et al. [86] proposed methods for measuring the VC dimension of a model class $\mathcal{F}$ by simulating data and estimating the model via empirical risk minimization. In particular [99] shows that the expected maximum deviation between the empirical risks of a classifier on two datasets can be bounded by a function which depends only on the VC dimension of the classifier. In other words, given a collection of classifiers $\mathcal{F}$, and two

data sets $D_n = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ and $D'_n = \{(y'_1, x'_1), \ldots, (y'_n, x'_n)\}$ where $(y_1, x_1), (y'_1, x'_1) \overset{iid}{\sim} \mathbb{P}$, we have the bound

$$\xi(n) := \mathbb{E}_{\mathbb{P}^n}\left[\sup_{f \in \mathcal{F}}(\widehat{R}_n(f, D_n) - \widehat{R}_n(f, D'_n))\right] \leqslant \begin{cases} 1 & n/h^* \leqslant \frac{1}{2} \\ C_1 \frac{\log(2n/h^*)+1}{n/h^*} & \text{if } n/h^* \text{ is small} \\ C_2\sqrt{\frac{\log(2n/h^*)+1}{n/h^*}} & \text{if } n/h^* \text{ is large,} \end{cases} \tag{8.1}$$

where $\mathrm{VCD}(\mathcal{F}) = h^*$. If this bound is tight for all distributions $\mathbb{P}$, then it may be possible to simulate data sets and calculate empirical versions of $\xi(n)$ for different values of n. Then, given constants C_1 and C_2, the right hand side depends only on the unknown VC dimension, so I could solve for it.

Vapnik et al. [99] suggest bounding (8.1) by $\Phi_{h^*}(n)$, viewed as a function of n and parametrized by h:

$$\Phi_h(n) = \begin{cases} 1 & n < h/2 \\ a\frac{\log\frac{2n}{h}+1}{\frac{n}{h}-a''}\left(\sqrt{1 + \frac{a'\left(\frac{n}{h}-a''\right)}{\log\frac{2n}{h}+1}} + 1\right) & \text{else.} \end{cases} \tag{8.2}$$

Here the constants $a = 0.16$, $a' = 1.2$ were determined numerically in [99] to adjust the trade-off between "small" and "large" in (8.1), and $a'' = 0.15$ was chosen so that $\Phi(0.5) = 1$ (this choice depends only on a and a''). If the bound is tight, then since (8.2) is known up to h, one can estimate it given knowledge of the maximum deviation on the left side of (8.1). I do not have such knowledge, but I can generate observations

$$\widehat{\xi}(n) = \Phi_{h^*}(n) + \epsilon(n)$$

at design points n. Here ϵ is mean zero noise (since the bound is tight) having an unknown distribution with support on $[0, 1]$. Given enough such observations at different design points n_ℓ, I can then estimate the true VC dimension h^* using nonlinear least square, but generating $\widehat{\xi}(n_\ell)$ is nontrivial. Vapnik et al. [99] give an

Algorithm 2: Generate $\widehat{\xi}(n_\ell)$

Given a collection of possible classifiers $\mathcal{F}$ and a grid of design points $n_1, \ldots, n_k$, generate $\widehat{\xi}(n_\ell)$. Repeat the procedure at each design point, n_ℓ, m times.

Set $k = 1$
while $k \leqslant m$ **do**
  Generate a data set from the same sample space $\mathcal{Y} \times \mathcal{X}$ as the training sample that is independent of the training sample. The generated set should be of size $2n_\ell$: $\{(y_1, x_1), \ldots, (y_{2n_\ell}, x_{2n_\ell})\}$.
  Split the data set into two equal sets, W and W'.
  Flip the labels (y values) of W'.
  Merge the two sets and train the classifier simultaneously on the entire set: W with the "correct" labels and W' with the "wrong" labels.
  Calculate the training error of the estimated classifier $\widehat{f}$ on W with the 'correct' labels and on W' using the "correct" labels.
  Set $\widehat{\xi}_i(n_\ell) = \left|\widehat{R}_{n_\ell}(\hat{f}, W) - \widehat{R}_{n_\ell}(\hat{f}, W')\right|$.
  $k \leftarrow k + 1$
end
Set $\widehat{\xi}(n_\ell) = \frac{1}{m}\sum_{i=1}^{m} \widehat{\xi}_i(n_\ell)$.

algorithm for generating the appropriate observations. Essentially, at each (fixed) design point $n_\ell : \ell \in \{1, \ldots, k\}$, one simulates m data points $(\widehat{\xi}_i(n_\ell), \Phi_h(n_\ell))$, for $i = 1, \ldots, m$, so as to approximate $\xi(n_\ell)$ as defined in (8.1). This procedure is shown in Algorithm 2.

The problem with this method is that the bound in (8.2) does not actually hold for the constants proposed in [99]. In particular, the tradeoff between "small" and "large" depends on $\mathbb{P}$. For example, construct $\mathbb{P}$ as follows:

$$\begin{aligned} p(x) &= \frac{1}{7} I(x \in \{-3, -2, -1, 0, 1, 2, 3\}) \\ p(y|x) &= \begin{cases} 1 & x < 0 \\ 0 & \text{else.} \end{cases} \end{aligned} \tag{8.3}$$

Take $\mathcal{F} = \{(a, \infty) : a \in \mathbb{R}\}$ which has VC dimension 1. Then I can calculate $\xi(n)$ exactly. Table 3 shows the exact values as well as the "bound". It is clear that in

n	$\xi(n)$	$\Phi_h(n)$
1	0.67	0.72
2	0.50	0.49
3	0.42	0.39

Table 3: This table shows the exact value of $\xi(n)$ for ν as defined in (8.3) as well as $\Phi(n)$. Clearly for $n > 1$, $\xi(n)$ exceeds the bound.

fact, $\xi(n) > \Phi_1(n)$ for some values of n. A complete upper bound which holds uniformly over all possible $\mathbb{P}$ is given in [99] as

$$\begin{aligned}
&\mathbb{E}_\nu \left[\sup_{f \in \mathcal{F}} (\widehat{R}_n(f, W) - \widehat{R}_n(f, W')) \right] \\
&< \min \left\{ 1, \sqrt{\frac{\log 2n/h^* + 1}{n/h^*}} + \frac{3}{\sqrt{nh^*(\log 2n/h^* + 1)}} \right\} \qquad (8.4) \\
&< \min \left\{ 1, 3\sqrt{\frac{\log 2n/h^* + 1}{n/h^*}} \right\}. \qquad (8.5)
\end{aligned}$$

However, this bound is so loose, that the methods of Vapnik et al. [99] and Shao et al. [86] will lead to severe underestimates of the true VC dimension.

A better strategy would be to find a lower bound on $\xi(n)$ which holds over some plausible class of distributions (0 is a trivial lower bound if the ν is a point mass). Given some way of measuring VC dimension, one can derive generalization error bounds which use the measured version rather than the truth. These would have a form much like the following.

Theorem 8.1. *Choose appropriate values of δ and φ based on the lower bound for $\xi(n)$. Let $\epsilon > 0$. Then, for any classifier $f \in \mathcal{F}$ where $\mathcal{F}$ has measured VC dimension $\widehat{h}$,*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| R_n(f) - \widehat{R}_n(f) \right| > \rho \right) \leqslant 4G\Gamma(\widehat{h} + \delta, 2n) \exp\{-n\epsilon^2\}(1 - \varphi) + \varphi. \qquad (8.6)$$

8.2 BETTER BLOCKING

In Section 6.3, I demonstrated that the upper bound may not be tight. In particular, under exponential or algebraic mixing, we gain a logarithmic factor as well as a power of $\kappa/(1+\kappa$ relative to the IID setting. The looseness of these upper bounds is attributable, at least in part, to the way it uses the β-mixing coefficients to bound the difference between IID measures and dependent measures. Recall from the proof of Lemma 4.8 that for an event ϕ in the σ-field generated by the block sequence $\mathbf{U}$,

$$\begin{aligned}|\widetilde{\mathbb{P}}(\phi) - \mathbb{P}^{n/2}(\phi)| \leqslant \left|\left|\widetilde{\mathbb{P}} - \mathbb{P} \times \mathbb{P}_{3,\ldots,n-1}\right|\right|_{TV} + \|\mathbb{P}_{3,\ldots,n-1} - \mathbb{P} \times \mathbb{P}_{5,\ldots,n-1}\|_{TV} \\ + \cdots + \left|\left|\mathbb{P}_{n-3,n-1} - \mathbb{P}^2\right|\right|_{TV} \end{aligned} \tag{8.7}$$

where I have $n/2$ blocks each of length 1, $\widetilde{\mathbb{P}}$ is the joint distribution of these blocks and $\mathbb{P}$ is the marginal distribution of a single block. The final step in the proof was to bound each total variation term with the mixing coefficient β_1.

Of course in the notation of (5.1), one could just as easily state the following result.

Theorem 8.2. *Let ϕ be an event in the σ-field generated by the block sequence $\mathbf{U}$. Then,*

$$|\widetilde{\mathbb{P}}(\phi) - \mathbb{P}^{n/2}(\phi)| \leqslant \sum_{i=1}^{\mu-1} \beta_a^{(2i-1)a}. \tag{8.8}$$

Clearly,

$$\sum_{i=1}^{\mu-1} \beta_a^{(2i-1)a} \leqslant (\mu-1)\beta_a^{(2\mu-3)a} \leqslant (\mu-1)\beta_a \tag{8.9}$$

with equality only when the process is Markovian of order a, so that $\beta_a = \beta_a^a$.

In fact, even the bound in Theorem 8.2 may be too loose. In simulations of the "even process" in Section 5.5, the bound $(\mu-1)\beta_a^a$ holds. If this could be shown to

Algorithm 3: Bootstrapping Risk Bounds

Use the data to resample lengthy time series from the empirical distribution to derive data dependent risk bounds

Take the time series $\mathbf{Y}_1^n$. Fit a model $\widehat{f} \in \mathcal{F}$, and calculate the in-sample risk, $\widehat{R}_n(\widehat{f})$.
Set $b = 1$
while $b \leqslant B$ **do**
    Bootstrap a new series $\mathbf{X}_1^{n+N}$ from $\mathbf{Y}_1^n$, which is several times longer than $\mathbf{Y}_1^n$
    Fit a model to $\mathbf{X}_1^n$, $\widehat{f}_{boot}$, and calculate its in-sample risk, $\widehat{R}_n(\widehat{f}_{boot})$.
    Calculate the test error of $\widehat{f}_{boot}$ on $\mathbf{X}_n^{n+1+N}$ and call it $\widehat{R}_N(\widehat{f}_{boot})$. Because the process is stationary and N is much larger than n, this should be a reasonable estimate of the generalization error of $\widehat{f}_{boot}$.
    Store the difference between the in-sample and generalization risks $\widehat{R}_N(\widehat{f}_{boot}) - \widehat{R}_n(\widehat{f}_{boot})$.
    $b \leftarrow b + 1$
end
Find the $1-\eta$ percentile of the distribution of over-fits. Add this to $\widehat{R}_n(\widehat{f})$.

be true, then the mixing estimation results would be useful in even non-Markovian settings, and it may be possible to remove the $\kappa/(1+\kappa)$ factor in Theorem 6.12.

8.3 BOOTSTRAPPING

An alternative to calculating bounds on forecasting error in the style of statistical learning theory is to use a carefully constructed bootstrap to learn about the generalization error. A fully nonparametric bootstrap for time series data uses the circular bootstrap reviewed in Lahiri [56]. The idea is to wrap the data of length n around a circle and randomly sample blocks of length q. There are n possible blocks, each starting with one of the data points 1 to n. Politis and White [74] give a method for choosing q. Algorithm 3 proposes a bootstrap for bounding the generalization error of a forecasting method.

While intuitively plausible, there is no theory, yet, which says that the results of this bootstrap will actually control the generalization error.

8.4 REGRET LEARNING

Another possible avenue is to target not the *ex ante* risk of the forecast, but the *ex post* regret: how much better might our forecasts have been, in retrospect and on the actually-realized data, had one used a different prediction function from the model $\mathcal{F}$ [13, 78]? In this thesis, I have generally focused on evaluating the performance of predictors $\widehat{f}$ through the risk or perhaps through the oracle risk

$$\mathbb{E}_{\mathbb{P}_1}[\ell(Y_{n+1}, \widehat{f}(\mathbf{Y}_{1:n}))] - \inf_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P}_1}[\ell(Y_{n+1}, f(\mathbf{Y}_{1:n}))]. \tag{8.10}$$

This quantity only makes sense under stationarity, and analysis like that pursued herein used a few other assumptions. If the distribution changes with time, then the above evaluation criterion will not work. Instead, one can consider an extreme case: let an adversary choose the next data point Y_t arbitrarily. In this case, I may choose a different forecasting function at each time point $f_1, \ldots, f_n$ rather than a fixed forecasting function $\widehat{f}$. Now performance is judged through the regret

$$\frac{1}{n}\sum_{i=1}^{n} \ell(Y_{i+1}, f_i(\mathbf{Y}_{1:i})) - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_{i+1}, f_i(\mathbf{Y}_{1:i})). \tag{8.11}$$

Essentially, this amounts to having a pool of experts $\mathcal{F}$ and choosing one expert f_i to make the forecast at each time i. We measure performance of our sequence of forecasts in comparison to the best expert in the pool. If one targets regret rather than risk, one can actually ignore mixing, and even stationarity [85].

9

CONCLUSION

In this thesis, I have demonstrated how to generalize risk bounds from the standard results for independent and identically distributed random variables analyzed with computer science style models to the case of dependent data with time series models. The basic procedure is to start with IID laws of large numbers and use the assumption of mixing combined with the blocking argument to derive laws of large numbers for dependent data. I can then use VC dimension to measure the complexity of real valued function classes. This results in bounds which hold for finite sample sizes, mis-specified models, and broad classes of data generating distributions. All that remains is to actually calculate the bound.

Statistical learning theory has proven itself in many practical applications, but most of its techniques have been developed in ways which have rendered it impossible to apply it immediately to time series forecasting problems.

Most results in statistical learning theory presume that successive data points are independent of one another. This is mathematically convenient, but clearly unsuitable for time series. Recent work has adapted key results to situations where widely-separated data points are asymptotically independent ("weakly dependent" or "mixing" time series). Basically, knowing the rate at which dependence decays lets one calculate how many effectively-independent observations the time series has and apply bounds with this reduced, effective sample size. In Chap-

ter 5, I showed how to estimate the mixing coefficients given one sample from the process.

To develop my results I need to know the complexity of the model classes to which I wish to apply the theory. In Chapter 6, I presented results which apply to linear forecasting models. These work for the vast majority of standard economic forecasting methods — vector autoregressions, linear state space models, and, in particular, linearized DSGEs — but they can not yet work for general nonlinear models with unknown VC dimension. In Chapter 8, I discuss some possible ways to modify my results to deal with this case.

Taken together, these results can provide probabilistic guarantees on a proposed forecasting model's performance. Such guarantees can give policy makers reliable empirical measures which intuitively explain the accuracy of a forecast. They can also be used to pick among competing forecasting methods.

Part V

APPENDIX

A

PROOFS OF SELECTED RESULTS

Proof of Lemma 5.7. Let $\mathbb{P}_{-\infty:0}$ be the distribution on $\sigma_{-\infty:0} = \sigma(\ldots, Y_{-1}, Y_0)$, and let $\mathbb{P}_{a:\infty}$ be the distribution on $\sigma_{a:\infty} = \sigma(Y_a, Y_{a+1}, Y_{a+2}, \ldots)$ for $a > 0$. Let $\mathbb{P}_{-\infty:0 \otimes a:\infty}$ be the distribution on $\sigma_{-\infty:0} \otimes \sigma_{a:\infty}$ (the product sigma-field). Then I can rewrite Definition 4.3 using this notation as

$$\beta_a = \sup_{C \in \sigma_\infty} |\mathbb{P}_{-\infty:0 \otimes a:\infty}(C) - [\mathbb{P}_{-\infty:0} \otimes \mathbb{P}_{a:\infty}](C)|. \tag{A.1}$$

Let $\sigma_{-d:0}$ and $\sigma_{a:a+d}$ be the sub-σ-fields of $\sigma_{-\infty:0}$ and $\sigma_{a:\infty}$ consisting of the d-dimensional cylinder sets for the d dimensions closest together. Let $\sigma_{-d:0} \otimes \sigma_{a:a+d}$ be the product σ-field of these two. For ease of notation define $\sigma^d := \sigma_{-d:0} \otimes \sigma_{a:a+d}$. Then I can rewrite β_a^d as

$$\beta_a^d = \sup_{C \in \sigma^d} |\mathbb{P}_{-\infty:0 \otimes a:\infty}(C) - [\mathbb{P}_{-\infty:0} \otimes \mathbb{P}_{a:\infty}](C)| \tag{A.2}$$

As such $\beta_a^d \leqslant \beta_a$ for all a and d. I can rewrite (A.2) in terms of finite-dimensional marginals:

$$\beta_a^d = \sup_{C \in \sigma^d} |\mathbb{P}_{-d:0 \otimes a:a+d}(C) - [\mathbb{P}_{-d:0} \otimes \mathbb{P}_{a:a+d}](C)|, \tag{A.3}$$

where $\mathbb{P}_{-d:0\otimes a:a+d}$ is the restriction of $\mathbb{P}_\infty$ to $\sigma(Y_{-d+1},\ldots,Y_0,Y_a,\ldots,Y_{a+d-1})$. Because of the nested nature of these sigma-fields,

$$\beta^{d_1}(a) \leqslant \beta^{d_2}(a) \leqslant \beta_a \tag{A.4}$$

for all finite $d_1 \leqslant d_2$. Therefore, for fixed a, $\{\beta_a^d\}_{d=1}^\infty$ is a monotone increasing sequence which is bounded above, and it converges to some limit $L \leqslant \beta_a$. To show that $L = \beta_a$ requires some additional steps.

Let $R = \mathbb{P}_{-\infty:0\otimes a:\infty} - [\mathbb{P}_{-\infty:0} \otimes \mathbb{P}_{a:\infty}]$, which is a signed measure on σ. Let

$$R^d = \mathbb{P}_{-d:0\otimes a:a+d} - [\mathbb{P}_{-d:0} \otimes \mathbb{P}_{a:a+d}],$$

which is a signed measure on σ^d. Decompose R into positive and negative parts as $R = Q^+ - Q^-$ and similarly for $R^d = Q^{+d} - Q^{-d}$. Notice that since R^d is constructed using the marginals of $\mathbb{P}_\infty$, then $R(E) = R^d(E)$ for all $E \in \sigma^d$. Now since R is the difference of probability measures,

$$\begin{aligned} 0 = R(\Omega) &= Q^+(\Omega) - Q^-(\Omega) \\ &= Q^+(D) + Q^+(D^c) - Q^-(D) - Q^-(D^c) \end{aligned} \tag{A.5}$$

for all $D \in \sigma$.

Define $Q = Q^+ + Q^-$. Let $\epsilon > 0$. Let $C \in \sigma$ be such that

$$Q(C) = \beta_a = Q^+(C) = Q^-(C^c). \tag{A.6}$$

Such a set C is guaranteed by the Hahn decomposition theorem (letting C^* be a set which attains the supremum in (A.2), I can throw away any subsets with negative R measure) and (A.5) assuming without loss of generality that $\mathbb{P}_{-\infty:0\otimes a:\infty}(C) > [\mathbb{P}_{-\infty:0} \otimes \mathbb{P}_{a:\infty}](C)$. One can use the field $\sigma' = \bigcup_d \sigma^d$ to approximate σ_∞ in the

sense that, for all ϵ, one can find $A \in \sigma'$ such that $Q(A \Delta C) < \epsilon/2$ (see Theorem D in Halmos [41, §13] or Lemma A.24 in Schervish [83]). Now,

$$\begin{aligned} Q(A \Delta C) &= Q(A \cap C^c) + Q(C \cap A^c) && \text{(A.7)} \\ &= Q^-(A \cap C^c) + Q^+(C \cap A^c) && \text{(A.8)} \end{aligned}$$

by (A.6) since $A \cap C^c \subseteq C^c$ and $C \cap A^c \subseteq C$. Therefore, since $Q(A \Delta C) < \epsilon/2$,

$$\begin{aligned} Q^-(A \cap C^c) &\leqslant \epsilon/2 \\ Q^+(A^c \cap C) &\leqslant \epsilon/2. \end{aligned} \tag{A.9}$$

Also,

$$\begin{aligned} Q(C) &= Q(A \cap C) + Q(A^c \cap C) && \text{(A.10)} \\ &= Q^+(A \cap C) + Q^+(A^c \cap C) && \text{(A.11)} \\ &\leqslant Q^+(A) + \epsilon/2 && \text{(A.12)} \end{aligned}$$

since $A \cap C$ and $A^c \cap C$ are contained in C and $A \cap C \subseteq A$. Therefore

$$Q^+(A) \geqslant Q(C) - \epsilon/2.$$

Similarly,

$$Q^-(A) = Q^-(A \cap C) + Q^-(A \cap C^c) \leqslant 0 + \epsilon/2 = \epsilon/2$$

since $A \cap C \subseteq C$ and $Q^-(C) = 0$ by (A.9). Finally,

$$\begin{aligned} Q^{+d}(A) &\geqslant Q^{+d}(A) - Q^{-d}(A) = R^d(A) && \text{(A.13)} \\ &= R(A) = Q^+(A) - Q^-(A) && \text{(A.14)} \\ &\geqslant Q(C) - \epsilon/2 - \epsilon/2 = Q(C) - \epsilon && \text{(A.15)} \\ &= \beta_a - \epsilon. && \text{(A.16)} \end{aligned}$$

And since $\beta_a^d \geqslant Q^{+d}(A)$, then for all $\epsilon > 0$ there exists d such that for all $d_1 > d$,

$$\beta^{d_1}(a) \geqslant \beta_a^d \geqslant Q^{+d}(A) \geqslant \beta_a - \epsilon. \tag{A.17}$$

Thus, it must be that $L = \beta_a$, so that $\beta_a^d \to \beta_a$ as desired. ■

B

DATA PROCESSING AND ESTIMATION METHODS FOR THE RBC MODEL

B.1 MODEL

Here I give the specific form of the RBC model presented initially in Section 2.3 and estimated in Section 6.2.2. The specific functional forms of the model sketched in Section 2.3 is the following.

$$\max_{c,l} U = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left(\frac{c_t^{\varphi} l_t^{1-\varphi}}{1-\phi} \right)^{1-\phi} \tag{B.1}$$

subject to

$$y_t = z_t k_t^{\alpha} n_t^{1-\alpha}, \tag{B.2}$$

$$1 = n_t + l_t, \tag{B.3}$$

$$y_t = c_t + i_t, \tag{B.4}$$

$$k_{t+1} = i_t + (1-\delta)k_t, \tag{B.5}$$

$$\ln z_t = (1-\rho)\ln \bar{z} + \rho \ln z_{t-1} + \epsilon_t, \tag{B.6}$$

$$\epsilon_t \overset{iid}{\sim} N(0, \sigma_{\epsilon}^2). \tag{B.7}$$

The first step to estimating the model is given by the following system of non-linear stochastic difference equations which are the necessary conditions for the optimization problem.

$$\left(\frac{1-\varphi}{\varphi}\right)\frac{c_t}{l_t} = (1-\alpha)z_t\left(\frac{k_t}{n_t}\right)^{\alpha} \tag{B.8}$$

$$c_t^{\kappa} l_t^{\lambda} = \beta E_t\left\{c_{t+1}^{\kappa} l_{t+1}^{\lambda}\left[\alpha z_{t+1}\left(\frac{n_{t+1}}{k_{t+1}}\right)^{1-\alpha} + (1-\delta)\right]\right\} \tag{B.9}$$

$$c_t + i_t = z_t k_t^{\alpha} n_t^{1-\alpha} \tag{B.10}$$

$$k_{t+1} = i_t + (1-\delta)k_t \tag{B.11}$$

$$1 = n_t + l_t \tag{B.12}$$

$$\ln z_t = (1-\rho)\ln\bar{z} + \rho \ln z_{t+1} + \epsilon_t \tag{B.13}$$

where $\kappa = \varphi(1-\phi) - 1$ and $\lambda = (1-\varphi)(1-\phi)$.

From this system, holding z_t constant, one can calculate the steady state values. These are given by

$$\frac{\widetilde{y}}{\widetilde{n}} = \eta, \tag{B.14}$$

$$\frac{\widetilde{c}}{\widetilde{n}} = \eta - \delta\theta, \tag{B.15}$$

$$\frac{\widetilde{i}}{\widetilde{n}} = \delta\theta, \tag{B.16}$$

$$\widetilde{n} = \left(1 + \left(\frac{1}{1-\alpha}\right)\left(\frac{1-\varphi}{\varphi}\right)[1-\delta\theta^{1-\alpha}]\right)^{-1}, \tag{B.17}$$

$$\widetilde{l} = 1 - \widetilde{n}, \tag{B.18}$$

$$\frac{\widetilde{k}}{\widetilde{n}} = \theta, \tag{B.19}$$

where

$$\theta = \left(\frac{\alpha}{1/\beta - 1 + \delta}\right)^{1/(1-\alpha)}, \tag{B.20}$$

$$\eta = \theta^{\alpha}. \tag{B.21}$$

The next step is to map (B.8)–(B.13) into a linear system of the form

$$Ax_{t+1} = Bx_t + C\nu_{t+1} + D\eta_{t+1}, \tag{B.22}$$

where x_t are the 7 time series in the model $[y_t,\ c_t,\ i_t,\ n_t,\ l_t,\ k_t,\ z_t]$, ν_t are the expectational errors, and η_t are the "exogenous structural shocks" as in DeJong and Dave [19, §5.1]. Taking logs gives the following system:

$$\begin{aligned}
0 &= \log\left(\frac{1-\varphi}{\varphi}\right) + \log c_{t+1} - \log l_{t+1} - \log(1-\alpha) - \log z_{t+1} \\
&\quad - \alpha \log k_t + \alpha \log n_{t+1} && \text{(B.23)} \\
0 &= \kappa \log c_t + \lambda \log l_t - \log \beta - \kappa \log c_{t+1} - \lambda \log l_{t+1} \\
&\quad - \log\left(\alpha \exp(\log z_{t+1}) \frac{\exp[(1-\alpha)\log n_{t+1}]}{\exp[(1-\alpha)\log k_{t+1}]} + 1 - \delta\right) && \text{(B.24)} \\
0 &= \log y_{t+1} - \log z_{t+1} - \alpha \log k_t - (1-\alpha)\log n_{t+1} && \text{(B.25)} \\
0 &= \log y_{t+1} - \log\left(\exp(\log c_{t+1}) + \exp(\log i_{t+1})\right) && \text{(B.26)} \\
0 &= \log k_{t+1} - \log\left(\exp(\log i_{t+1}) + (1-\delta)\exp(\log k_t)\right) && \text{(B.27)} \\
0 &= -\log\left(\exp(\log n_{t+1}) + \exp(\log l_{t+1})\right) && \text{(B.28)} \\
0 &= \log z_{t+1} - \rho \log z_t. && \text{(B.29)}
\end{aligned}$$

Where I have deliberately used k_t rather than k_{t+1} in (B.25). Note that the time dependent terms in the model are now all in log deviations from steady state values. Taking derivatives of the system with respect x_{t+1} and evaluating at the steady state values gives the system matrix A while taking derivatives with respect to x_t gives the system matrix $-B$. The C and D matrices are determined by inspection. In this case, $C = [0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1]$ and $D = [0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0]$.

Given the system in (B.22), I use the method of Sims [89] to transform the model into state space form.[1] The code returns matrices F and G. Finally, to get everything into the form of the linear Gaussian state space model in (6.46),

$$A = F[1:4, 6:7] \qquad H = \texttt{diag}(\epsilon_y, \epsilon_c, \epsilon_i, \epsilon_h) \tag{B.30}$$

$$T = F[6:7, 6:7] \qquad Q = \sigma^2(GG')[6:7, 6:7]. \tag{B.31}$$

Now to return the likelihood, I can run the Kalman filter on (B.30) and (B.31).

B.2 DATA

Once the model is prepared, the data must be prepared. The data to estimate the RBC model is publicly available from the Federal Reserve Economic Database FRED. The necessary series are shown in the Table 4. All of the data is quarterly. The required series are PCESVC96, PCNDGC96, GDPIC1, HOANBS, and CNP16OV. These five series are used to create four series $[y_t', c_t', i_t', h_t']$ as follows:

$$c_t' = 2.5 \times 10^5 \frac{\text{PCESVC96} + \text{PCNDGC96}}{\text{CNP16OV}} \tag{B.32}$$

$$i_t' = 2.5 \times 10^5 \frac{\text{GDPIC1}}{\text{CNP16OV}} \tag{B.33}$$

$$y_t' = c_t + i_t \tag{B.34}$$

$$h_t' = 6000 \frac{\text{HOANBS}}{\text{CNP16OV}}. \tag{B.35}$$

I use the preprocessed data which accompanies DeJong and Dave [19]. This data is available from `http://www.pitt.edu/~dejong/seconded.htm`. I then apply the HP-filter described in Hodrick and Prescott [44] to each series individually to

1 Code for this transformation is available from http://sims.princeton.edu/yftp/gensys/.

Series ID	Description	Unit	Availability
PCESVC96	Real Personal Consumption Expenditures: Services	Billions of Chained 2005 $	1/1/1995
PCNDGC96	Real Personal Consumption Expenditures: Nondurable Goods	Billions of Chained 2005 $	1/1/1995
GDPIC1	Real Gross Domestic Investment	Billions of Chained 2005 $	1/1/1947
HOANBS	Nonfarm Business Sector: Hours of All Persons	Index: 2005=100	1/1/1947
CNP16OV	Civilian Noninstitutional Population	Thousands of Persons	1/1/1948

Table 4: Data series from FRED

calculate trend components $\left[\widetilde{y}_t,\ \widetilde{c}_t,\ \widetilde{i}_t,\ \widetilde{h}_t\right]$. The HP-filter amounts to fitting the smoothing spline

$$\widetilde{\mathbf{x}}_{1:n} = \underset{\mathbf{z}_{1:n}}{\operatorname{argmin}} \sum_{t=1}^{n} (x'_t - z_t)^2 + \lambda \sum_{t=2}^{n-1} ((z_{t+1} - z_t) - (z_t - z_{t-1}))^2, \tag{B.36}$$

with the convention $\lambda = 1600$. I then calculate the detrended series that will be fed into the RBC model as

$$x_t = \log x'_t - \log \widetilde{x}'_t. \tag{B.37}$$

The result is shown in Figure 10.

B.3 ESTIMATION

To perform the estimation, I maximize the likelihood returned by the Kalman filter, but penalize it with priors on each of the "deep" parameters. This is because the likelihood surface is very rough and there exists some prior information about the parameters. Additionally, each of the parameters is constrained to lie in a plausible

		Prior		Constraint	
Parameter	Estimate	Mean	Variance	Lower	Upper
α	0.24	0.29	2.5×10^{-2}	0.1	0.5
β	0.99	0.99	1.25×10^{-3}	0.90	1
ϕ	4.03	1.5	2.5	1	5
φ	0.13	0.6	0.1	0	1
δ	0.03 2	2.5×10^{-2}	1×10^{-3}	0	0.2
ρ	0.89	0.95	2.5×10^{-2}	0.80	1
σ_ϵ	3.45×10^{-5}	1×10^{-4}	2×10^{-5}	0	0.05
σ_y	1.02×10^{-6}	–	–	0	1
σ_c	2.30×10^{-5}	–	–	0	1
σ_i	6.11×10^{-4}	–	–	0	1
σ_n	1.68×10^{-4}	–	–	0	1

Table 5: Priors, constraints, and parameter estimates for the RBC model.

interval. Each parameter has a normal prior with means and variances similar to those in the literature. I generally follow those in DeJong et al. [21]. The priors, constraints (which are strict), and estimates are shown in Table 5.

BIBLIOGRAPHY

[1] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium of Information Theory*, eds. B. N. Petrov and F. Csaki, pp. 267–281.

[2] Athanasopoulos, G., and Vahid, F. (2008), "VARMA versus VAR for macroeconomic forecasting," *Journal of Business and Economic Statistics*, **26**(2), 237–252.

[3] Athreya, K., and Pantula, S. (1986), "A note on strong mixing of ARMA processes," *Statistics & Probability Letters*, **4**(4), 187–190.

[4] Baraud, Y., Comte, F., and Viennet, G. (2001), "Adaptive estimation in autoregression or β-mixing regression via model selection," *The Annals of Statistics*, **29**(3), 839–875.

[5] Bartlett, P. L., and Mendelson, S. (2002), "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, **3**, 463–482.

[6] Bickel, P., and Rosenblatt, M. (1973), "On some global measures of the deviations of density function estimates," *The Annals of Statistics*, **1**(6), 1071–1095.

[7] Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, Springer Verlag, New York, 2nd edn.

[8] Bradley, R. (1983), "Absolute regularity and functions of markov chains," *Stochastic Processes and their Applications*, **14**(1), 67–77.

[9] Bradley, R. C. (2005), "Basic properties of strong mixing conditions. A survey and some open questions," *Probability Surveys*, **2**, 107–144, arXiv:math/0511078.

[10] Brayton, F., and Tinsley, P. (1996), "A guide to FRB/US: A macroeconomic model of the United States," Tech. Rep. 1996-42, Finance and Economics Discussion Series, Federal Reserve Board, Washington, DC.

[11] Brayton, F., Levin, A., Lyon, R., and Williams, J. (1997), "The evolution of macro models at the Federal Reserve Board," in *Carnegie-Rochester Conference Series on Public Policy*, vol. 47, pp. 43–81, Elsevier.

[12] Carrasco, M., and Chen, X. (2002), "Mixing and moment properties of various GARCH and stochastic volatility models," *Econometric Theory*, **18**(01), 17–39.

[13] Cesa-Bianchi, N., and Lugosi, G. (2006), *Prediction, learning, and games*, Cambridge Univ Press, Cambridge, UK.

[14] Christoffel, K., Coenen, G., and Warne, A. (2008), "The new area-wide model of the Euro area: A micro-founded open-economy model for forecasting and policy analysis," Tech. Rep. 944, European Central Bank Working Paper Series, http://www.ecb.int/pub/pdf/scpwps/ecbwp944.pdf.

[15] Claeskens, G., and Hjort, N. L. (2008), *Model Selection and Model Averaging*, no. 27 in Cambridge series in statistical and probabilistic mathematics, Cambridge University Press.

[16] Corless, R., Gonnet, G., Hare, D., Jeffrey, D., and Knuth, D. (1996), "On the Lambert w function," *Advances in Computational Mathematics*, **5**(1), 329–359.

[17] Davydov, Y. (1973), "Mixing conditions for Markov chains," *Theory of Probability and its Applications*, **18**(2), 312–328.

[18] Dedecker, J., Doukhan, P., Lang, G., Leon R., J. R., Louhichi, S., and Prieur, C. (2007), *Weak Dependence: With Examples and Applications*, Springer Verlag, New York.

[19] DeJong, D., and Dave, C. (2011), *Structural macroeconometrics*, Princeton Univ Press, Princeton, 2 edn.

[20] DeJong, D., Dharmarajan, H., Liesenfeld, R., and Richard, J.-F. (2008), "Exploiting non-linearities in GDP growth for forecasting and anticipating regime changes," Tech. rep., University of Pittsburgh.

[21] DeJong, D. N., Ingram, B. F., and Whiteman, C. H. (2000), "A Bayesian approach to dynamic macroeconomics," *Journal of Econometrics*, **98**(2), 203–223.

[22] DeJong, D. N., Dharmarajan, H., Liesenfeld, R., and Richard, J.-F. (2009), "Efficient filtering in state-space representations," Tech. rep., University of Pittsburgh.

[23] Dejong, D. N., Dharmarajan, H., Liesenfeld, R., Moura, G. V., and Richard, J.-F. (2009), "Efficient likelihood evaluation of state-space representations," Tech. rep., University of Pittsburgh.

[24] Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. (2007), "On the fit and forecasting performance of New Keynesian models," *Journal of Business and Economic Statistics*, **25**(2), 123–162.

[25] Devroye, L., and Györfi, L. (1985), *Nonparametric Density Estimation: The* L_1 *View*, John Wiley & Sons, Inc., New York.

[26] Doan, T., Litterman, R., and Sims, C. (1984), "Forecasting and conditional projection using realistic prior distributions," *Econometric Reviews*, **3**(1), 1–100.

[27] Doucet, A., De Freitas, N., and Gordon, N. (2001), *Sequential Monte Carlo Methods in Practice*, Springer Verlag.

[28] Doukhan, P. (1994), *Mixing: Properties and Examples*, Springer Verlag, New York.

[29] Durbin, J., and Koopman, S. (2001), *Time Series Analysis by State Space Methods*, Oxford Univ Press, Oxford.

[30] Eberlein, E. (1984), "Weak convergence of partial sums of absolutely regular sequences," *Statistics & Probability Letters*, **2**(5), 291–293.

[31] Edge, R. M., and Gurkaynak, R. S. (2011), "How useful are estimated DSGE model forecasts?" Finance and Economics Discussion Series 2011-11, Federal Reserve Board, http://federalreserve.gov/pubs/feds/2011/201111/201111abs.html.

[32] Edge, R. M., Kiley, M. T., and Laforte, J.-P. (2007), "Documentation of the research and statistics division's estimated DSGE model of the U.S. economy: 2006 version," Tech. Rep. 2007-53, Finance and Economics Discussion Series, Federal Reserve Board, Washington, DC.

[33] Faust, J., and Wright, J. H. (2009), "Comparing Greenbook and reduced form forecasts using a large realtime dataset," *Journal of Business and Economic Statistics*, **27**(4), 468–479.

[34] Fernández-Villaverde, J. (2009), "The econometrics of DSGE models," Tech. rep., NBER Working Paper Series.

[35] Freedman, D., and Diaconis, P. (1981a), "On the histogram as a density estimator: l_2 theory," *Probability Theory and Related Fields*, **57**(4), 453–476.

[36] Freedman, D., and Diaconis, P. (1981b), "On the maximum deviation between the histogram and the underlying density," *Probability Theory and Related Fields*, **58**(2), 139–167.

[37] Fryzlewicz, P., and Subba Rao, S. (2011), "Mixing properties of ARCH and time-varying ARCH processes," *Bernoulli*, **17**(1), 320–346.

[38] Gerali, A., Neri, S., Sessa, L., and Signoretti, F. (2010), "Credit and banking in a DSGE model of the Euro area," *Journal of Money, Credit and Banking*, **42**, 107–141.

[39] Gertler, M., and Karadi, P. (2011), "A model of unconventional monetary policy," *Journal of Monetary Economics*, **58**, 17–34.

[40] Goodhart, C., Osorio, C., and Tsomocos, D. (2009), "Analysis of monetary policy and financial stability: A new paradigm," Tech. Rep. 2885, CESifo, http://www.cesifo-group.de/portal/page/portal/ifoHome/b-publ/b3publwp/_wp_by_number?p_number=2885.

[41] Halmos, P. (1974), *Measure Theory*, Graduate Texts in Mathematics, Springer-Verlag, New York.

[42] Harvey, A., and Durbin, J. (1986), "The effects of seat belt legislation on British road casualties: A case study in structural time series modelling," *Journal of the Royal Statistical Society. Series A (General)*, **149**(3), 187–227.

[43] Harvey, A., Ruiz, E., and Shephard, N. (1994), "Multivariate stochastic variance models," *The Review of Economic Studies*, **61**(2), 247–264.

[44] Hodrick, R. J., and Prescott, E. C. (1997), "Postwar U.S. business cycles: An empirical investigation," *Journal of Money, Credit, and Banking*, **29**(1), 1–16.

[45] Hoeffding, W. (1963), "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, **58**(301), 13–30.

[46] Johnson, N., Kotz, S., and Balakrishnan, N. (1994), *Continuous univariate distributions*, vol. 2, John Wiley & Sons.

[47] Kalman, R. E. (1960), "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, **82**(1), 35–45.

[48] Kim, C., and Nelson, C. (1998), "Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching," *Review of Economics and Statistics*, **80**(2), 188–201.

[49] Kitagawa, G. (1987), "Non-Gaussian state-space modeling of nonstationary time series," *Journal of the American Statistical Association*, , 1032–1041.

[50] Kitagawa, G. (1996), "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *Journal of Computational and Graphical Statistics*, , 1–25.

[51] Kleijn, B. J. K., and van der Vaart, A. W. (2006), "Misspecification in infinite-dimensional Bayesian statistics," *Annals of Statistics*, **34**, 837–877, arXiv:math/0607023.

[52] Koltchinskii, V., and Panchenko, D. (2000), "Rademacher processes and bounding the risk of function learning," *High Dimensional Probability II*, **47**, 443–459.

[53] Kontorovich, L., and Ramanan, K. (2008), "Concentration inequalities for dependent random variables via the martingale method," *Annals of Probability*, **36**(6), 2126–2158, arXiv:math/0609835.

[54] Koyama, S., Pérez-Bolde, L. C., Shalizi, C. R., and Kass, R. E. (2010), "Approximate methods for state-space models," *Journal of the American Statistical Association*, **105**(489), 170–180.

[55] Kydland, F. E., and Prescott, E. C. (1982), "Time to build and aggregate fluctuations," *Econometrica*, **50**(6), 1345–1370.

[56] Lahiri, S. (2003), *Resampling methods for dependent data*, Springer Verlag.

[57] Ledoux, M., and Talagrand, M. (1991), *Probability in Banach Spaces: Isoperimetry and Processes*, A Series of Modern Surveys in Mathematics, Springer Verlag, Berlin.

[58] Lerasle, M. (2011), "Optimal model selection for density estimation of stationary data under various mixing conditions," *The Annals of Statistics*, **39**(4), 1852–1877.

[59] Liu, W., and Wu, W. (2010), "Simultaneous nonparametric inference of time series," *The Annals of Statistics*, **38**(4), 2388–2421.

[60] Lozano, F. (2000), "Model selection using Rademacher penalization," in *Proceedings of the Second ICSC Symposia on Neural Computation (NC2000). ICSC Adademic Press*.

[61] Lucas, R. E. (1976), "Econometric policy evaluation: A critique," in *The Phillips Curve and Labor Markets*, eds. K. Brunner and A. Meltzer, vol. 1 of *Carnegie-Rochester Conference Series on Public Policy*, Amsterdam: North-Holland.

[62] Massart, P. (2007), "Concentration inequalities and model selection," in *Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*, Springer.

[63] McDiarmid, C. (1989), "On the method of bounded differences," in *Surveys in Combinatorics*, ed. J. Siemons, pp. 148–188, Cambridge University Press.

[64] McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011), "Generalization error bounds for stationary autoregressive models," arXiv:1103.0942.

[65] Meir, R. (2000), "Nonparametric time series prediction through adaptive model selection," *Machine Learning*, **39**(1), 5–34.

[66] Mohri, M., and Rostamizadeh, A. (2009), "Rademacher complexity bounds for non-iid processes," in *Advances in Neural Information Processing*

Systems, eds. D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, vol. 21, pp. 1097–1104, MIT Press, Cambridge, MA.

[67] Mohri, M., and Rostamizadeh, A. (2010), "Stability bounds for stationary φ-mixing and β-mixing processes," *Journal of Machine Learning Research*, **11**, 789–814.

[68] Mokkadem, A. (1988), "Mixing properties of ARMA processes," *Stochastic Processes and their Applications*, **29**(2), 309–315.

[69] Müller, U. K. (2011), "Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix," Tech. rep., Princeton University, http://www.princeton.edu/~umueller/sandwich.pdf.

[70] Nobel, A. (2006), "Hypothesis testing for families of ergodic processes," *Bernoulli*, **12**(2), 251–269.

[71] Nummelin, E., and Tuominen, P. (1982), "Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory," *Stochastic Processes and Their Applications*, **12**(2), 187–202.

[72] Ould-Saïd, E., Yahia, D., and Necir, A. (2009), "A strong uniform convergence rate of a kernel conditional quantile estimator under random left-truncation and dependent data," *Electronic Journal of Statistics*, **3**, 426–445.

[73] Pham, T. D., and Tran, L. T. (1985), "Some mixing properties of time series models," *Stochastic processes and their applications*, **19**(2), 297–303.

[74] Politis, D., and White, H. (2004), "Automatic block-length selection for the dependent bootstrap," *Econometric Reviews*, **23**(1), 53–70.

[75] Pollard, D. (1984), *Convergence of stochastic processes*, Springer Verlag, New York.

[76] Pollard, D. (1990), *Empirical processes: Theory and applications*, Institute of Mathematical Statistics.

[77] Racine, J. (2000), "Consistent cross-validatory model-selection for dependent data: HV-block cross-validation," *Journal of Econometrics*, **99**(1), 39–61.

[78] Rakhlin, A., Sridharan, K., and Tewari, A. (2010), "Online learning: Random averages, combinatorial parameters, and learnability," arXiv:1006.1138.

[79] Rakhlin, A., Sridharan, K., and Tewari, A. (2011), "Online learning: Stochastic and constrained adversaries," arXiv:1104.5070.

[80] Rosenberg, D., and Bartlett, P. (2007), "The Rademacher complexity of co-regularized kernel classes," in *Proceedings of the Eleventh International Conference of Artificial Intelligence and Statistics*, eds. M. Meila and X. Shen, vol. 2, pp. 396–403, JMLR W&CP.

[81] Ruiz-del Solar, J., and Vallejos, P. (2005), "Motion detection and tracking for an AIBO robot using motion compensation and Kalman filtering," in *Lecture Notes in Computer Science 3276 (RoboCup 2004)*, pp. 619–627, Springer Verlag.

[82] Sargent, T. J. (1989), "Two models of measurements and the investment accelerator," *The Journal of Political Economy*, **97**, 251–287.

[83] Schervish, M. (1995), *Theory of Statistics*, Springer Series in Statistics, Springer Verlag, New York.

[84] Shalizi, C. R. (2009), "Dynamics of Bayesian updating with dependent data and misspecified models," *Electronic Journal of Statistics*, **3**, 1039–1074, arXiv:0901.1342.

[85] Shalizi, C. R., Jacobs, A. Z., Klinkner, K. L., and Clauset, A. (2011), "Adapting to non-stationarity with growing expert ensembles," arXiv:1103.0949.

[86] Shao, X., Cherkassky, V., and Li, W. (2000), "Measuring the VC-dimension using optimized experimental design," *Neural computation*, **12**(8), 1969–1986.

[87] Shumway, R., and Stoffer, D. (2000), *Time Series Analysis and Its Applications*, Springer Verlag, New York.

[88] Silverman, B. (1978), "Weak and strong uniform consistency of the kernel estimate of a density and its derivatives," *The Annals of Statistics*, **6**(1), 177–184.

[89] Sims, C. A. (2002), "Solving linear rational expectations models," *Computational Economics*, **20**(1-2), 1–20.

[90] Smets, F., and Wouters, R. (2007), "Shocks and frictions in US business cycles: A Bayesian DSGE approach," *American Economic Review*, **97**(3), 586–606.

[91] Solow, R. M. (1957), "Technical change and the aggregate production function," *The Review of Economics and Statistics*, **39**, 312–320.

[92] Steinwart, I., and Anghel, M. (2009), "Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise," *The Annals of Statistics*, **37**(2), 841–875.

[93] Tran, L. (1989), "The l_1 convergence of kernel density estimates under dependence," *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, **17**(2), 197–208.

[94] Tran, L. (1994), "Density estimation for time series by histograms," *Journal of statistical planning and inference*, **40**(1), 61–79.

[95] van de Geer, S. (2002), "On Hoeffding's inequality for dependent random variables," in *Empirical Process Techniques for Dependent Data*, eds. H. Dehling, T. Mikosch, and M. Sørensen, pp. 161–169, Birkhäuser, Boston.

[96] Vapnik, V. (1998), *Statistical learning theory*, John Wiley & Sons, Inc., New York.

[97] Vapnik, V. (2000), *The Nature of Statistical Learning Theory*, Springer Verlag, New York, 2nd edn.

[98] Vapnik, V., and Chervonenkis, A. (1971), "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, **16**, 264–280.

[99] Vapnik, V., Levin, E., and Cun, Y. L. (1994), "Measuring the VC-dimension of a learning machine," *Neural Computation*, **6**(5), 851–876.

[100] Weiss, B. (1973), "Subshifts of finite type and sofic systems," *Monatshefte für Mathematik*, **77**(5), 462–474.

[101] Wildi, M. (2009), "Real-time US-recession indicator (USRI) a classical cycle perspective with bounceback," Tech. rep., Institute of Data Analysis and Process Design.

[102] Withers, C. S. (1981), "Conditions for linear processes to be strong-mixing," *Probability Theory and Related Fields*, **57**(4), 477–480.

[103] Woodroofe, M. (1967), "On the maximum deviation of the sample density," *The Annals of Mathematical Statistics*, **38**(2), 475–481.

[104] Yu, B. (1993), "Density estimation in the l_∞ norm for dependent data with applications to the Gibbs sampler," *Annals of Statistics*, **21**(2), 711–735.

[105] Yu, B. (1994), "Rates of convergence for empirical processes of stationary mixing sequences," *The Annals of Probability*, **22**(1), 94–116.

[106] Zhang, X., Albanes, D., Beeson, W. L., van den Brandt, P. A., Buring, J. E., Flood, A., Freudenheim, J. L., Giovannucci, E. L., Goldbohm, R. A., Jaceldo-Siegl, K., Jacobs, E. J., Krogh, V., Larsson, S. C., Marshall, J. R., McCullough, M. L., Miller, A. B., Robien, K., Rohan, T. E., Schatzkin, A., Sieri, S., Spiegelman, D., Virtamo, J., Wolk, A., Willett, W. C., Zhang, S. M., and Smith-Warner, S. A. (2010), "Risk of colon cancer and coffee,

tea, and sugar-sweetened soft drink intake: Pooled analysis of prospective cohort studies," *Journal of the National Cancer Institute*, **102**(11), 771–783.

[107] Zhu, X., Rogers, T., and Gibson, B. (2009), "Human Rademacher complexity," in *Advances in Neural Information Processing Systems*, eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, vol. 22, pp. 2322–2330.