Daniel McDonald
Department of Statistics
Carnegie Mellon University

Thesis Proposal

Generalization Error Bounds for State Space Models with an Application to Economic Forecasting

Version: June 29, 2010

Committee: Cosma Shalizi, Mark Schervish, Alessandro Rinaldo, Larry Wasserman, and David N. DeJong

Abstract

In this thesis, I propose to derive entirely data dependent generalization error bounds for state space models. These results can characterize the out-of-sample accuracy of many types of forecasting methods. The bounds currently available for time series data rely both on a quantity describing the dependence properties of the data generating process known as the mixing rate and on a quantification of the complexity of the model space. I will derive methods for estimating the mixing behavior from data and characterize the complexity of state space models. The resulting risk bounds will be useful for empirical researchers at the forefront of economic forecasting as well as for economic policy makers. The bounds can also be applied in other situations where state space models are employed.

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 [55]. 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 strands 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.

To address the first question, suppose that predictors live in some space $\mathcal{X}$ and responses live in another space $\mathcal{Y}$. Finding a mapping $f : \mathcal{X} \to \mathcal{Y}$ amounts 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.

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

$$R(f) = \mathbb{E}[\ell(Y, f(X))], \tag{1}$$

is the generalization error, or risk, of the prediction algorithm. Unfortunately, while this is the natural target to minimize when searching for predictors, it is intractable in most applications. 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 give straightforward techniques for model comparison without making assumptions on the data generating process in contrast to likelihood based methods. They 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 for 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, however in many areas of interest, this is not the case. Consequently, many of the risk bounds in existence 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 data, 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 data rely on

an ability to quantify the dependence behavior for the data generating process, in particular the rate at which the dependence disappears. While knowledge of these rates leads to clean theoretical results, this knowledge is sadly unavailable for applied work. 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.

In section 2, I review the state-of-the-art methods for economic forecasting. Section 3 surveys the literature on prediction risk and generalization error bounds for independent and identically distributed data. Section 4 discusses the notions of mixing for time series and generalization error bounds for dependent data. Finally, in section 5, I outline the goals and research directions for this thesis.

2 Economic forecasting

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. [8] for an overview of this history and Brayton and Tinsley [7] 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 [38] 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.

Kydland and Prescott [36] 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

[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.

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.

2.1 Dynamic stochastic general equilibrium models

Kydland and Prescott [36] 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 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
$$\max_{c_t, l_t} U = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t, l_t). \tag{2}$$
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 stuff y_t using the production function $g(\cdot)$
$$y_t = z_t g(k_t, n_t), \tag{3}$$
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 [50].

3. The remaining equations are as follows:
$$1 = n_t + l_t \tag{4}$$
$$y_t = c_t + i_t \tag{5}$$
$$k_{t+1} = i_t + (1 - \delta)k_t \tag{6}$$
$$\ln z_t = (1 - \rho) \ln \overline{z} + \rho \ln z_{t-1} + \epsilon_t \tag{7}$$
$$\epsilon_t \overset{iid}{\sim} \mathrm{N}(0, \sigma^2). \tag{8}$$
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 Board of Governors uses differentiated goods, differentiated firms, sticky prices (they do not adjust immediately), and monetary policy (see Edge et al. [20]). 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 [36] or likelihood analysis as in Sargent [47]. In either case, the resulting estimated model can be used for forecasting.

2.2 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

$$\mathbf{x}_t = \mathbf{B}_1\mathbf{x}_{t-1} + \mathbf{B}_2\mathbf{x}_{t-2} + \cdots + \mathbf{B}_p\mathbf{x}_{t-p} + \mathbf{e}_t \tag{9}$$

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 [17], 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 that in Kim and Nelson [28] are trivially state space models. The turning point forecasting models such as DeJong et al. [14] or Wildi [53] 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 [26]. More recently, state space models have been used for robot soccer by Ruiz-del Solar and Vallejos [46], to study the effects of a seat belt law on traffic accidents in Great Britain by Harvey and Durbin [23], and for neural decoding applications as in Koyama et al. [34].

2.3 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:

$$y_t = f(x_t, \epsilon_t) \tag{10}$$

$$x_{t+1} = g(x_t, \eta_t) \tag{11}$$

$$x_1 \sim F, \tag{12}$$

where ϵ_t are η_t are i.i.d. and 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 f, g, and the distributions of ϵ_t and η_t.

In the case where f and g 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 [29, 30] and Doucet et al. [18] for an exposition of the particle filter and Koyama et al. [34] and Dejong et al. [15] for improvements).

3 Prediction risk

The goal in constructing any predictive model is to learn some function f which maps available data into predictions. To evaluate these forecasts, one writes down a loss function $\ell(Y, f(X))$ which represents the cost of making forecast errors. Here Y denotes the future data for which predictions are desired. The ideal target to control when making predictions is the risk:

$$R(f) = \mathbb{E}_\mu[\ell(Y, f(X))], \tag{13}$$

where $(X, Y) \sim \mu$.

Since μ is unknown, $R(f)$ is unknown, but researchers often estimate it by the training error

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)). \tag{14}$$

While it is true that $R_n(f)$ converges to $R(f)$ for many algorithms, one can show that when f is chosen by minimizing 14, $\mathbb{E}_\mu[R_n(f)] \leq R(f)$. This means that choosing models based on the training error will give suboptimal results: these models will tend to overfit the data and result in poor out-of-sample predictions.[2] In the statistics and machine learning literature, there are two strategies to obviate this issue. The first is to restrict the class of functions allowed by the algorithm. The second, which encompasses the direction taken in this thesis, is to modify the minimization criterion so as to penalize increased complexity. This section provides some intuition as to why complexity penalization is useful for controlling prediction risk, it then describes some generalization error bounds for i.i.d. data.

3.1 Complexity control

Numerous results exist for penalized model selection in the form of structural risk minimization or regularization. In structural risk minimization, analysts choose a sequence of models of increasing complexity $\{\mathcal{F}_d : d = 1, 2, \ldots\}$ and choose candidate predictors $\widehat{f}_d$ as the solution to the minimization problem

$$\widehat{f}_d = \arg\min_{f \in \mathcal{F}_d} R_n(f). \tag{15}$$

One then chooses the final predictor, $\tilde{f}$, by solving a minimization problem of the form

$$\min_{\widehat{f}_d} R_n(\widehat{f}_d) + \text{pen}(d, n) \tag{16}$$

or

$$\min_{\widehat{f}_d} R_n(\widehat{f}_d) \times \text{pen}(d, n) \tag{17}$$

where the structures of the loss function and the model class $\mathcal{F}_d$ are used to choose the form of the penalty, $\text{pen}(d, n)$. The usefulness of complexity control is best illustrated through simple examples. Here I present explicit risk results for the normal means model and provide generalization error bounds for regression problems.

[2] Many articles in the economic literature compare the forecast performance of models in exactly this way. See for example Athanasopoulos and Vahid [2], Faust and Wright [22], Christoffel et al. [11], Del Negro et al. [16] and Smets and Wouters [49]. Some of these use a cross validation type analysis, fitting the model on the training set and calculating the error on a test set, but this procedure can also be heavily biased: the held out data is used to choose the model class under consideration, the distributions of the test set and the training set may be different, and large deviations from the normal course of events (the recessions in 1980-82) may be ignored.

3.1.1 The normal means model

Suppose that $X_i \sim \mathrm{N}(\theta_i, \sigma^2)$ for $i = 1, \ldots, n$. The goal is to estimate $\theta = (\theta_1, \ldots, \theta_n)'$. Take $\ell(\theta, f(X)) = \sum_{i=1}^n (\theta_i - f(X_i))^2$. Consider the following estimator

$$\widehat{\theta}_i^S = \begin{cases} X_i & i \in S \\ 0 & i \notin S. \end{cases} \tag{18}$$

where $S \subseteq \{1, \ldots, n\}$. Then the risk decomposes into two components

$$R(\widehat{\theta}_i^S) = \sum_{i=1}^n \mathbb{E}[(\widehat{\theta}_i^S - \theta_i)^2] = \sum_{i \notin S} \theta_i^2 + |S|\sigma^2. \tag{19}$$

The first term is the square of the bias while the second term is the variance of the estimator.

Choosing larger S decreases the bias but increases the variance. The maximum likelihood estimator $\widehat{\theta}_i^{MLE}$ is unbiased but has variance $n\sigma^2$, and hence its risk is $n\sigma^2$. It may be possible to give up some bias to decrease the variance and hence make the risk smaller. Ideally one would choose S so as to make $R(\widehat{\theta}_i^S)$ as small as possible, but using $R_n(\widehat{\theta}_i^S)$ is a poor approximation because it is biased. In particular, for the MLE, $R_n(\widehat{\theta}_i^{MLE}) = 0$.

One would like to have an unbiased estimate of the risk to be used for the choice of S. In this case, a little algebra shows that

$$R(\widehat{\theta}_i^S) = \mathbb{E}[R_n(\widehat{\theta}_i^S)] + 2|S|\sigma^2 - n\sigma^2, \tag{20}$$

so it is possible to choose S by minimizing $R_n(\widehat{\theta}_i^S) + 2|S|\sigma^2$. Of course this only works because the distribution of X is known up to a finite dimensional θ.

3.1.2 Regression problems

If the distribution that generated the data is unknown, equality results for risk estimation are no longer possible for finite amounts of data. Penalties like AIC [1], Schwarz criterion [48], and generalized cross-validation [12] come from asymptotic results for linear models where a strict accounting of the number of parameters provides an adequate notion of the complexity of the model space. But in the finite sample case, it is more useful to provide bounds on the size of the generalization error.

Consider the one dimensional regression model

$$y = f(x) + \epsilon \tag{21}$$

where $f(\cdot)$ is some unknown, possibly nonlinear function and ϵ is mean zero random noise. Estimation is based on n i.i.d. training samples from the unknown joint distribution μ. The goal is to choose some estimate $\widehat{f}$ of f from a large class of possible models $\mathcal{F}_d$ which is indexed by some parameters β and a measure of complexity d. For example, take $\mathcal{F}_d$ to be the family of algebraic polynomials

$$\widehat{f}_d(x) = \sum_{i=0}^d \beta_i x^i. \tag{22}$$

Choosing d large will result in interpolating the training points, i.e. a training error of zero, but potentially large predictive risk. Results from Vapnik-Chervonenkis (VC)-theory give an explicit bound on the predictive risk. Calculation of this bound for each model gives a straightforward

method for model selection. Here, the VC dimension is $d+1$. Cherkassky et al. [10] provide a bound for the predictive risk under arbitrary loss: with probability at least $1-\eta$

$$R(\widehat{f}_d) \leq R_n(\widehat{f}_d) \times \left(1 - \sqrt{c - c\ln c - \frac{\ln \eta}{n}}\right)_+^{-1} \tag{23}$$

where $R_n(f)$ is the training error and $c = (d+1)/n$.

3.2 Bounds for I.I.D. data

The statistics and machine learning literature contains many generalization error bounds for learning algorithms based on i.i.d. data for both classification and regression problems. They all depend on some notion of the complexity of the model class: VC-dimension, covering numbers, Rademacher complexity, or algorithmic stability. The discussion here focuses on Rademacher complexities and algorithmic stability since many of the existing bounds for dependent data rely on these complexity notions.

Rademacher complexity is a method of characterizing the capacity of a function class $\mathcal{F}$ by measuring its ability to find functions which correlate well with noise.

Definition 3.1 *For a set of real-valued functions $\mathcal{F}$ with domain $\mathcal{X}$, a distribution P_X on $\mathcal{X}$ and samples of size n from P_X, the Rademacher complexity is*

$$\mathfrak{R}(\mathcal{F}) = \mathbb{E}_X \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right| \right], \tag{24}$$

where $\sigma_1, \cdots, \sigma_n$ are i.i.d. and take values ± 1 with equal probability. Its empirical counterpart is

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right| \right]. \tag{25}$$

Using this notion of complexity leads to generalization error bounds for linear predictors [25], decision trees and support vector machines [3], and combined classifiers such as neural networks and those derived through boosting or bagging [31] among others. For an interesting treatment measuring the Rademacher complexity of human learning capacity see Zhu et al. [56]. Here, the authors ask humans to perform classification exercises while measuring the ability of the average human to learn noise. In one experiment, the true classification of words was according to its length: the word was labeled 1 if it had fewer than 6 letters and 0 otherwise. The size of the function space for human learning turns out to be quite large. Prediction functions used by the participants included for example whether the word "tastes good" or whether the word "relates to motel service". It turns out that human learning satisfies the generalization error bounds, i.e. the risk of the human learning "machine" can be upper bounded in the same way as the sorts of learning algorithms typically deployed by machine learners.

An alternative measure of capacity is algorithmic stability. Algorithmic stability measures the capacity of a learning algorithm by measuring the change in the selected function if the data are perturbed by a small amount. The following definition comes from Mohri and Rostamizadeh [41].

Definition 3.2 *Let $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} = \{Z_1, \ldots, Z_n\}$ be a training sample of size n from the joint distribution of (X, Y). An algorithm is (uniformly) λ-stable if the predictor it returns, $\widehat{f}_{D_n}$, for any two training samples D_n and D_n' that differ by a single point satisfy*

$$\left| \ell(\widehat{f}_{D_n}, Z) - \ell(\widehat{f}_{D_n'}, Z) \right| \leq \lambda \qquad \forall Z \in X \times Y, \tag{26}$$

where $\ell(\cdot, \cdot)$ is the loss function.

Bousquet and Elisseeff [5] use algorithmic stability to derive generalization bounds for a wide class of learning algorithms. Their general result is stated in the theorem below.

Theorem 3.3 *Consider a learning algorithm with uniform stability λ and a non-negative loss function ℓ bounded above by M, for all $Z \in \mathcal{X} \times \mathcal{Y}$ and all training sets D_n. Then for any $n \geq 1$, with probability at least $1 - \eta$,*

$$R(f) \leq R_n(f) + 2\lambda + (4n\lambda + M)\sqrt{\frac{\ln 1/\eta}{2n}}. \tag{27}$$

In particular, Bousquet and Elisseeff [5] give explicit generalization error bounds for k-nearest neighbor classifiers, support vector machines, L_p regularized least squares regression, and minimum relative entropy classification.

4 Time series

Generalization error bounds for i.i.d. data usually arise from an application of Hoeffding's inequality which requires the independence of data points. In order to derive similar results for time series data, a characterization of the dependence structure is necessary. If the dependence is not too strong, then i.i.d. results can be applied to dependent data. This section discusses the notions of mixing for time series data which characterize the dependence of the data and allow for the application of i.i.d. type results. Some resulting generalization bounds for time series data follow.

4.1 Mixing

Mixing for time series data makes the dependence between the future and the past explicit. In particular, mixing rates quantify the decay in the dependence as the future moves farther from the past. There are many definitions of mixing of varying strength (see for example Doukhan [19], Dedecker et al. [13], or Bradley [6]), but for this thesis, the most important notion of mixing is β-mixing. The following definition comes from Doukhan [19].

Definition 4.1 *Let $\{X_i\}_{i=1}^{\infty}$ be random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and denote $X_1^t = \{X_i\}_{i=1}^t$ and $X_{t+m}^{\infty} = \{X_i\}_{i=t+m}^{\infty}$. Let $\sigma_1^t = \sigma(X_1^t)$ and $\sigma_{t+m}^{\infty} = \sigma(X_{t+m}^{\infty})$, be the sigma fields of events generated by the appropriate collections of random variables. Let $\mathbb{P}_t$ be the restriction of $\mathbb{P}$ to σ_1^t, $\mathbb{P}_{t+m}$ be the restriction of $\mathbb{P}$ to σ_{t+m}^{∞} and $\mathbb{P}_{t \otimes t+m}$ be the restriction of $\mathbb{P}$ to $\sigma_{t,t+m} = \sigma(\{X\}_{i=1}^t, \{X\}_{i=t+m}^{\infty})$. Then the coefficient of absolute regularity, or β-mixing coefficient, $\beta(m)$, is given by*

$$\beta(m) = \sup_t ||\mathbb{P}_t \otimes \mathbb{P}_{t+m} - \mathbb{P}_{t \otimes t+m}||_{TV}, \tag{28}$$

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

Definition 4.1 shows that the β-mixing coefficient measures the total variation distance between the joint distribution and the distribution if the time series were independent.

While β-mixing coefficients, or at least the rates of decay, are known for a number of stochastic processes such as ARMA models (Mokkadem [42]), GARCH models (Carrasco and Chen [9]), and Markov processes (see Doukhan [19] for an overview of the literature), there do not appear to be satisfactory methods for estimating these coefficients from time series data (see Meir [39, p. 7]).

Another useful definition of mixing is φ-mixing, which is stronger than β-mixing. The following definition comes from [41].

Definition 4.2 *Using the same notation as in 4.1, the φ-mixing coefficient, $\varphi(m)$, is given by*

$$\varphi(m) = \sup_t \sup_{A \in \sigma_1^t} \sup_{B \in \sigma_{t+m}^{\infty}} \mathbb{E}|P(B \mid A) - P(B)|. \tag{29}$$

A stochastic process is said to be φ-mixing, if $\varphi(m) \to 0$ as $m \to \infty$.

Both of these mixing notions are used in the literature to derive generalization error bounds for dependent data.

4.2 Generalization error bounds for time series

Extending the results from section 3.2 to time series prediction is a fairly recent development. The work of Yu [54] contains many of the uniform law of large numbers results for time series that are typically needed to derive generalization error bounds. Vidyasagar [52] mentions the extension of these results to time series data as an open problem in the literature.

Meir [39] is one of the first papers to construct generalization error bounds for time series data. The general approach is to consider an infinite memory, stationary stochastic process, and decompose the training error of a predictor with finite memory, chosen through empirical risk minimization into three parts:

$$R(\widehat{f}_{p,n,d}) = (R(\widehat{f}_{p,n,d}) - R(f^*_{p,n})) + (R(f^*_{p,n}) - R(f^*_p)) + R(f^*_p) \tag{30}$$

where $\widehat{f}_{p,n,d}$ is an empirical estimate based on finite data of length n, finite memory of length p, and complexity indexed by d; $f^*_{p,d}$ is the oracle with finite memory and given complexity, and f^*_p is the oracle with finite memory over all possible complexities. The three terms amount to an estimation error incurred from the use of noisy data, an approximation error due to the selection of a predictor from a class of limited complexity, and a loss from approximating an infinite memory process with a finite memory process.

There are two main theorems in [39]. The first bounds the estimation error. It requires the stochastic process to be bounded and β-mixing. The second theorem provides a bound for the expected loss of a predictor chosen through structural risk minimization. This bound depends both on the ability to bound the covering numbers of the classes of predictors and on the assumption that for each n, $d = o(a_n)$, where a_n comes from the process of removing the dependence of the stochastic process which is a rather extensive argument analogous to the method of symmetrization and randomization for i.i.d. random variables.

Mohri and Rostamizadeh [40] present Rademacher complexity-based error bounds for non-i.i.d. settings, a generalization of similar existing bounds derived for the i.i.d. case. Their bounds hold in the scenario of dependent samples generated by a stationary β-mixing process. The results are data-dependent and measure the complexity of a class of hypotheses based on the training sample. The empirical Rademacher complexity can be estimated from finite samples and leads to tighter generalization bounds. Their main theorem uses these empirical Rademacher complexities $\mathfrak{R}_{D_\mu}(f)$ where D_μ is a subsample of size a from the original sample D_n.

Theorem 4.3 *Let $\mathcal{F}$ be a space of candidate predictors and $\mathcal{H}$ by the space of induced losses $\ell(Y, f(X))$ for $f \in \mathcal{F}$ such that $\mathcal{H}$ is bounded above by M. Then for any sample D_n drawn from a stationary β-mixing distribution, and for any $\mu, m > 0$ with $2\mu m = n$ and $\eta > 4(\mu - 1)\beta(m)$ where $\beta(m)$ is the mixing coefficient, with probability at least $1 - \eta$,*

$$R(f) \leq R_n(f) + \mathfrak{R}_{D_\mu}(\mathcal{H}) + 3M\sqrt{\frac{\ln 4/\eta'}{2\mu}}, \tag{31}$$

where $\eta' = \eta - 4(\mu - 1)\beta(m)$.

Steinwart and Christmann [51] prove an oracle inequality for generic regularized empirical risk minimization algorithms learning from α-mixing processes, a slightly weaker notion of mixing. To illustrate the inequality, they derive learning rates for least squares SVMs. Since the proof of the oracle inequality uses localization ideas developed for i.i.d. processes, it turns out that these learning rates are close to the optimal rates known in the i.i.d. case.

Mohri and Rostamizadeh [41] study the scenario where the observations are drawn from a stationary φ-mixing or β-mixing sequence. They prove stability-based generalization bounds for both situations. These bounds strictly generalize the bounds given in the i.i.d. case and apply to all stable learning algorithms extending the use of stability-bounds to non-i.i.d. scenarios. The main theorem for φ-mixing sequences follows.

Theorem 4.4 *Let f be the predictor returned by a λ-stable learning algorithm trained on a sample D_n from a φ-mixing stationary distribution with $\varphi(k) = \varphi_0 k^{-r}$ for $r > 1$. Let ℓ be a loss function bounded by $M > 0$. Then for any $\epsilon > 0$,*

$$\mathbb{P}\Big[\,|R(f) - R_n(f)| > \epsilon + \lambda + (r+1)6M\varphi(b)\Big] \leq 2\exp\left\{\frac{-2\epsilon^2(1 + 2\varphi_0 r/(r-1))^{-2}}{n(2\lambda + (r+1)2M\varphi(b) + M/n)^2}\right\}, \quad (32)$$

where $\varphi(b) = \varphi_0\left(\frac{\lambda}{r\varphi_0 M}\right)^{r/(r+1)}$.

They give some examples of applications including support vector regression, kernel ridge regression, and support vector machines. The authors note that their method can be applied when the training and test sets are not independent in contrast to the result in Meir [39], which require the test data to be independent of the training data. This bound generalizes the bound in Kontorovich and Ramanan [33] and Kontorovich [32]. It also matches the i.i.d. stability bound in Bousquet and Elisseeff [4]. Extending these results for the more general β-mixing sequences gives a similar exponential type inequality plus an additive term that depends on the β-mixing rate.

Karandikar and Vidyasagar [27] show that if an algorithm is 'sub-additive' and yields a predictor whose risk can be upper bounded when the data are i.i.d., then the same algorithm will result in predictors whose risk can be bounded if the data is β-mixing. They use this result to derive generalization error bounds in terms of the learning rates for i.i.d. data and the β-mixing coefficients of the data generating process.

All of the results presented in this section suffer from three main drawbacks: they require a priori knowledge of the mixing rate, they require some knowledge of the complexity of the model space $\mathcal{F}$ or loss space $\mathcal{H}$, and they assume bounded loss.

5 Proposed work

The goal of this thesis is to develop theoretical methods to control the risk of state space models, especially those used for time series economic forecasting which rely on little data. Generalization error bounds for state space models would allow forecasters to control the expected cost of predictions and to choose among competing models.

Existing results in the literature are inapplicable to this task for three reasons. First, all the results rely on a characterization of the mixing rates for the data generating process. These rates are assumed known for all of the generalization error bounds in section 4.2. Second, all of the existing bounds require an ability to characterize the complexity of the model space $\mathcal{F}$ using either Rademacher complexities or covering numbers. Third, all of the existing generalization error

bounds for time series data require bounded loss functions which are rarely sensical in the regression setting. In this section, I lay out potential paths for overcoming each of these hurdles and give some preliminary results.

5.1 Estimation of mixing rates

Estimating the mixing rates of time series data is a problem that has not been well studied in the literature. According to Ron Meir, "as far as we are aware, there is no efficient practical approach known at this stage for estimation of mixing parameters [39, p. 7]." The form of the β-mixing rate

$$\beta(m) = \sup_t ||\mathbb{P}_t \otimes \mathbb{P}_{t+m} - \mathbb{P}_{t \otimes t+m}||_{TV}, \tag{33}$$

suggests at least one straightforward, though perhaps naïve approach to estimation which could prove fruitful. One could use nonparametric density estimation for the two marginal distributions as well as the joint distribution, and then calculate the total variation distance using numerical integration. While somewhat simplistic, this method could give good results. However, one would need to show not only that the estimator is unbiased and consistent, but also learn enough about it that the generalization error bound could be properly adjusted to account for the additional uncertainty introduced by using an estimate rather than the true quantity.

Another approach is to bound the β-mixing coefficient with a potentially easier to estimate quantity. Information mixing bounds β-mixing as

$$\beta(m) \leq \sqrt{I(m)}, \tag{34}$$

where $I(m)$ is the information mixing coefficient defined in Bradley [6]. This is just the supremum of the mutual information taken over all t for the densities associated with $\mathbb{P}_t$ and $\mathbb{P}_{t+m}$. The estimation of mutual information has been studied extensively as in Pál et al. [43], Kraskov et al. [35], and Paninski [44].

Most methods for estimating information mixing proceed as in the naïve scenario above by using density estimates to calculate the information mixing rate. Of course, the densities themselves are nuisance parameters and once the densities are estimated, one may as well go for the β-mixing rate instead. An alternate approach in Kraskov et al. [35] uses the distance to the kth nearest neighbor of each point in the joint space to derive an estimator for the mutual information. The resulting formula is very straightforward to apply.

Let $\{X_i\}_{i=1}^N$ be a sample from one marginal distribution, $\{Y_i\}_{i=1}^N$ be a sample from the other marginal distribution, and $Z_i = (X_i, Y_i)$. Denote by $\epsilon(i)$ the distance from Z_i to its kth nearest neighbor using the metric

$$||z - z'|| = \max\left(||x - x'||, ||y - y'||\right), \tag{35}$$

where the norms in the $\mathcal{X}$ and $\mathcal{Y}$ space need not be the same. Let $n_x(i)$ be the number of points x_j whose distance from x_i is less than $\epsilon(i)$ and the same for $n_y(i)$. The estimator for mutual information $I^{(1)}(X, Y)$ is given by

$$I^{(1)}(X, Y) = \psi(k) - \sum_{i=1}^{N} \left[\psi(n_x(i) + 1) + \psi(n_y(i) + 1)\right] + \psi(N) \tag{36}$$

where $\psi(\cdot)$ is the digamma function. The results from applying this estimator to GDP data from 1947 to 2010 to estimate $I(m)$ for $1 \leq m \leq 30$ are shown in Figure 1 for various choices of k.

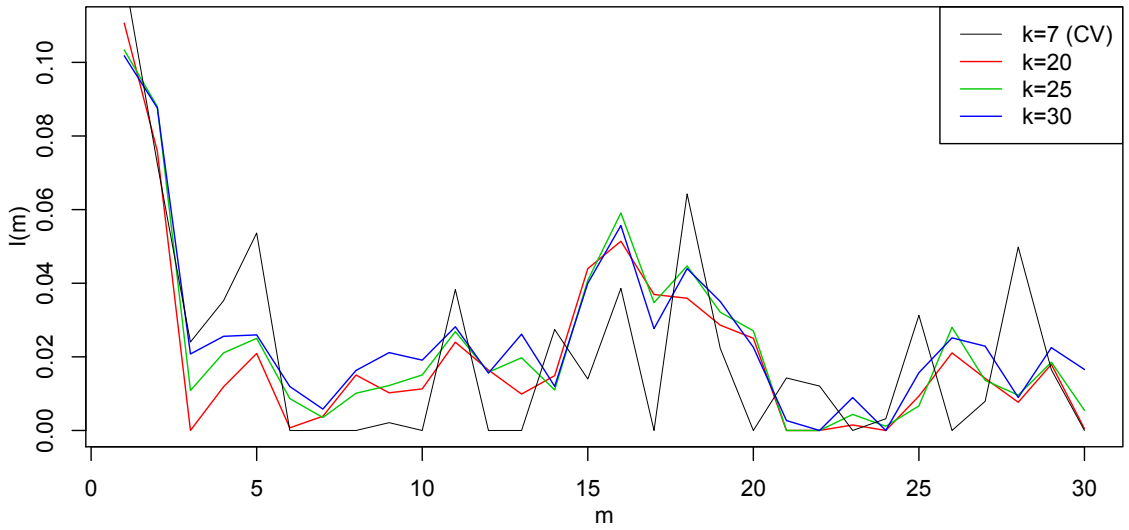

Figure 1: Estimated information mixing rate for GDP data.

The estimator suffers from a number of drawbacks. It is not guaranteed to be nonnegative even though information mixing is positive with $I(m) = 0$ implying independence. Also, as $m \to \infty$, $I(m)$ should decay smoothly to zero, which is clearly not the case in the figure. Deriving an estimator for some measure of mixing which can be used for constructing generalization error bounds is one of the goals of this thesis.

5.2 Model space complexities

Each of the generalization error bounds in section 4.2 has a term which measures the complexity of the model space. This means performing explicit calculations of the covering numbers, Rademacher complexities, or λ-stability for state space models. Alternatively, we could try to use the empirical Rademacher complexity. However, fitting large economic forecasting models once requires significant computing resources. Fitting those same models hundreds or thousands of times in order to calculate the expected risk numerically is almost certainly untenable. Asking forecasters to do this with new theoretical models under consideration is certainly out of the question.

I have been able to derive bounds for the Rademacher complexities for AR(p) models. Extending these results to VARMA and ARMA models, should give some intuition into the necessary path for linear, and eventually nonlinear, state space models.

In order to apply the generalization error bound in Mohri and Rostamizadeh [40] to stationary AR(p) processes, one must calculate or bound the Rademacher complexity of the class of models

$$\mathcal{F}_p = \left\{ \varphi_1, \ldots, \varphi_p : x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} \text{ and } x_t \text{ is stationary} \right\}. \tag{37}$$

The stationarity condition is usually written as the roots of the polynomial

$$p(z) = z^p + \varphi_1 z^{p-1} + \cdots + \varphi_p \tag{38}$$

must lie within the unit circle. Call the space of such coefficients that satisfy this condition the stability domain. For $p = 1$, this domain is easy to characterize: $|\varphi_1| < 1$. For general p, this space

is more complex. A recursive characterization is given in Fam and Meditch [21]. In particular, they show that the space can be bounded by a convex polygon with vertices at the extremes of the stability domain. The vertex with the largest L_2 distance from the origin has coordinates $\left(\binom{p}{1}, \ldots, \binom{p}{p}\right)$. This means that

$$||\varphi||_2^2 \leq \sum_{i=1}^{p} \binom{p}{i}^2 = \binom{2p}{p} - 1, \tag{39}$$

is a necessary condition for $\varphi_1, \ldots, \varphi_p$ to be in the stability domain.

Ordinary linear regressions can be written as kernel regressions. Let

$$\alpha_i = \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'\mathbf{Y}\right)_i \tag{40}$$

$$k(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{X}_i\mathbf{X}_j', \tag{41}$$

where $\mathbf{X}$ is the $n \times p$ design matrix, $\mathbf{Y}$ are the responses, and $\mathbf{X}_i$ is the i^{th} row of the design matrix. Now, requiring

$$\sum_{i,j} \alpha_i \alpha_j k(\mathbf{X}_i, \mathbf{X}_j) \leq \gamma^2, \tag{42}$$

corresponds to the regularization $||\widehat{\beta}^{OLS}||_2^2 \leq \gamma^2$, or ridge regression assuming that the bound is tight.

Returning to the AR(p) model, this means that

$$\mathcal{F}_p \subseteq \overline{\mathcal{F}_p} = \left\{ \varphi_1, \ldots, \varphi_p : x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} \text{ and } ||\varphi||_2^2 \leq \binom{2p}{p} - 1 \right\}. \tag{43}$$

This characterization of the AR(p) model as a regularized kernel regression allows for the application of Lemma 22 in Bartlett and Mendelson [3] to bound the Rademacher complexity of an AR(p) model using either

$$\mathfrak{R}_n(\mathcal{F}_p) \leq \mathfrak{R}_n(\overline{\mathcal{F}_p}) \leq \frac{2}{\sqrt{n}} \sqrt{\left(\binom{2p}{p} - 1\right) \mathbb{E}\mathbf{X_1}\mathbf{X_1}'} \tag{44}$$

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_p) \leq \widehat{\mathfrak{R}}_n(\overline{\mathcal{F}_p}) \leq \frac{2}{\sqrt{n}} \sqrt{\left(\binom{2p}{p} - 1\right) \frac{1}{n} \sum_{t=i}^{n} \mathbf{X_i}\mathbf{X_i}'} \tag{45}$$

where $\widehat{\mathfrak{R}}_n(\mathcal{F}_p)$ is the empirical Rademacher complexity and $\mathfrak{R}_n(\mathcal{F}_p) = \mathbb{E}_X \widehat{\mathfrak{R}}_n(\mathcal{F}_p)$.

5.3 Unbounded loss

A very useful paper by Wenxin Jiang [24] derives an extension to Hoeffding's inequality which applies in the case of unbounded loss and dependent data. This inequality can then be used to bound the generalization error in cases with known dependence structure which is weaker than absolute regularity.

Theorem 5.1 *Let $\{\mathcal{F}_t\}_{-\infty}^{\infty}$ be an increasing sequence of σ-fields and X_t be a random variable that is $\mathcal{F}_t$ measurable for each t. Then for any $\epsilon, C > 0$ and positive integers n, m,*

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{t=1}^{n}(X_t - \mathbb{E}[X_t])\right| > \epsilon\right) \leq 2m \exp\left\{\frac{-n\epsilon^2}{288m^2C^2}\right\} + \frac{6}{\epsilon n}\sum_{t=1}^{n}\mathbb{E}\left|E\left[X_t \mid \mathcal{F}_{t-m}\right] - \mathbb{E}[X_t]\right| + \frac{15}{\epsilon n}\sum_{t=1}^{n}\mathbb{E}\left[|X_t| I_{\{C,\infty\}}(|X_t|)\right]. \tag{46}$$

Here the first term has the exponential type bound similar to the Hoeffding or McDiarmid inequalities used in the derivation of most generalization error bounds. The second term handles the dependence requiring a very weak notion of mixing. The third term puts a condition on the probability of large "tail" values so as to account for unbounded loss functions.

The requirement of bounded loss is not only displeasing from an aesthetic point of view, but also from a practical one. Since these types of risk inequalities are worst-case results, they assume that the loss bound M is attained with probability 1 by the learning algorithm. This has undesirable consequences which are illustrated in the next section.

5.4 Complete bounds for AR models

Combining the results from sections 5.1 and 5.2 with Theorem 4.3 can give a (somewhat) complete bound for the prediction risk of an AR model. Here I use the estimated information mixing rate as a proxy for the true β-mixing rate. This calculation illustrates the delicacy of the result in Theorem 4.3. Predicting the same GDP data with an AR(2) model, let $m = 9$ and $n = 252$. Then for all $\eta > 4 \times 13 \times 0.002 = 0.11$, the third term in the theorem becomes $3M\sqrt{\ln(4/\eta')/28}$ where $\eta' = \eta - 0.11$ and M is chosen as a cutoff to bound the loss function. The other two terms in the bound are both empirical. The empirical risk $R_n(f)$ is 9.62×10^{-5}, and the empirical Rademacher complexity is upperbounded by

$$\frac{8\sqrt{M}}{\mu}\sqrt{\left(\binom{2p}{p} - 1\right)\sum_{i=1}^{n} X_i X_i'} = 0.07\sqrt{M}. \tag{47}$$

Thus, assuming that the estimated information mixing rate bounds the true β-mixing rate of the data generating process, with probability at least $1 - \eta = 0.85$,

$$R(f) \leq 9.62 \times 10^{-5} + 0.07\sqrt{M} + 1.03M \equiv B(M) \tag{48}$$

for some choice of M. This has the rather unfortunate property that

$$R(f) \leq M < B(M) \tag{49}$$

for all M. This illustrates the main difficulty of the bounded loss requirement in the regression setting: the bound applies to the worst case scenario where mistakes always occur at the bound.

A more useful bound can be computed through a bootstrapping procedure. A fully nonparametric version is possible using the circular bootstrap reviewed in Lahiri [37]. The idea is to wrap the data of length T around a circle and randomly sample blocks of length q. There are T possible blocks, each starting with one of the data points 1 to T. To choose q, I used the method of Politis and White [45]. For this data set, the result was a block length of $q = 7$. I ran the bootstrap for $B = 1000$ samples. The strategy was as follows:

1. Take the time series, call it X. Fit an AR(2) model $g(X)$, and calculate the in-sample risk, $R_n(g(X))$.
2. Repeat B times:
   - Bootstrap a new series Y from X, which is several times longer than X; call the initial segment, which is as long as X, Y_1.
   - Fit a model to this, $g_b(Y_1)$, and calculate its in-sample risk, $R_n(g_b(Y_1))$.
   - Calculate the risk of $g_b(Y_1)$ on the rest of Y. Because the process is stationary and Y is much longer than X, this should be a reasonable estimate of the generalization error of $g_b(Y_1)$.
   - Store the difference between the in-sample and generalization risks.
3. Find the $1-\eta$ percentile of the distribution of over-fits. Add this to $R_n(g(X))$.

I chose the new time series to be the length of the data (252) plus 400, giving an extra hundred years. I got the following results for $\eta = 0.05$:

$$\widehat{R}(m(x)) = 9.62 \times 10^{-5} \tag{50}$$

$$\mathrm{Pen}_{1-\eta} = 5.81 \times 10^{-5} \tag{51}$$

$$R(m(x)) \leq \widehat{R}(m(x)) + \mathrm{Pen}_{1-\eta} \tag{52}$$

$$= 1.54 \times 10^{-4} \tag{53}$$

This bound is intuitively sensible and does not suffer from any of the deficiencies of the other bounds, however, there is no theory that supports the coverage claim, and AR models can be fit quickly, whereas general state-space models cannot. Part of this thesis will be to flesh out the theory of this method.

5.5 Conclusions

For this thesis, I plan to concentrate on four complementary yet distinct avenues for progress. The first is to derive estimators for the mixing behavior of time series data. These estimators will allow for useful data dependent generalization error bounds for time series data, whether the learning algorithm is a state space model or not. The second avenue is to continue developing characterizations of the complexity of the model space of increasingly rich state space models. Autoregressive models are a good starting point and suggest the possibility of proceeding by considering autoregressive moving average models before tackling univariate state space models. At the same time, it may be possible to quickly generalize the univariate results to the multivariate least squares algorithm used for vector auto regressions. The third avenue is to use the concentration inequality of Jiang [24] to develop better generalization error bounds with unbounded loss under weaker mixing conditions. Finally, I plan to investigate the conditions under which the bootstrap method results in generalization error bounds with the correct coverage.

The resulting generalization error bounds derived in this thesis can be used by economic forecasters for model selection and for appropriately characterizing out of sample forecast performance. They can also be used to communicate the forecast quality to policy makers. These results are not limited to economic problems, but can be applied to any area using dependent data and state space models.

References

[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, 237–252.

[3] Bartlett, P. L. and Mendelson, S. (2002), "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results," *Journal of Machine Learning Research*, 3, 463–482.

[4] Bousquet, O. and Elisseeff, A. (2001), "Algorithmic Stability and Generalization Performance," *Advances in Neural Information Processing Systems*, pp. 196–202.

[5] Bousquet, O. and Elisseeff, A. (2002), "Stability and Generalization," *The Journal of Machine Learning Research*, 2, 499–526.

[6] Bradley, R. C. (2005), "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," *Probability Surveys*, 2, 107–144.

[7] 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.

[8] 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.

[9] Carrasco, M. and Chen, X. (2002), "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 18, 17–39.

[10] Cherkassky, V., Shao, X., Mulier, F., and Vapnik, V. (1999), "Model Complexity Control for Regression Using VC Generalization Bounds," *IEEE transactions on Neural Networks*, 10, 1075–1089.

[11] 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.

[12] Craven, P. and Wahba, G. (1978), "Smoothing Noisy Data with Spline Functions," *Numerische Mathematik*, 31, 377–403.

[13] Dedecker, J., Doukhan, P., Lang, G., R., J. R. L., Louhichi, S., and Prieur, C. (2007), *Weak Dependence: With Examples and Applications*, Springer Verlag.

[14] 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.

[15] Dejong, D. N., Dharmarajan, H., Liesenfeld, R., and Richard, J.-F. (2009), "Efficient Filtering in State-Space Representations," Tech. rep., University of Pittsburgh.

[16] 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, 123–162.

[17] Doan, T., Litterman, R., and Sims, C. (1984), "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews*, 3, 1–100.

[18] Doucet, A., De Freitas, N., and Gordon, N. (2001), *Sequential Monte Carlo Methods in Practice*, Springer Verlag.

[19] Doukhan, P. (1994), *Mixing: Properties and Examples (Lecture Notes in Statistics)*, Springer, 1 edn.

[20] 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.

[21] Fam, A. T. and Meditch, J. S. (1978), "A Canonical Parameter Space for Linear Systems Design," *IEEE Transactions on Automatic Control*, 23, 454–458.

[22] 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, 468–479.

[23] 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, 187–227.

[24] Jiang, W. (2009), "On Uniform Deviations of General Empirical Risks with Unboundedness, Dependence, and High Dimensionality," *Journal of Machine Learning Research*, 10, 977–996.

[25] Kakade, S. M., Sridharan, K., and Tewari, A. (2008), "On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization," Tech. rep., NIPS.

[26] Kalman, R. E. (1960), "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, 82, 35–45.

[27] Karandikar, R. L. and Vidyasagar, M. (2009), "Probably Approximately Correct Learning with Beta-Mixing Input Sequences," Tech. rep., Indian Statistical Institute.

[28] 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, 188–201.

[29] Kitagawa, G. (1987), "Non-Gaussian State-Space Modeling of Nonstationary Time Series," *Journal of the American Statistical Association*, pp. 1032–1041.

[30] Kitagawa, G. (1996), "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models," *Journal of Computational and Graphical Statistics*, pp. 1–25.

[31] Koltchinskii, V. and Panchenko, D. (2002), "Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers," *The Annals of Statistics*, 30, 1–50.

[32] Kontorovich, L. (2008), "Metric and Mixing Sufficient Conditions for Concentration of Measure," Tech. Rep. 0610427, arXiv.

[33] Kontorovich, L. and Ramanan, K. (2008), "Concentration Inequalities for Dependent Random Variables via the Martingale Method," *Annals of Probability*, 36, 2126–2158.

[34] 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, 170–180.

[35] Kraskov, A., Stögbauer, H., and Grassberger, P. (2004), "Estimating Mutual Information," *Physical Review E*, 69, 1–16.

[36] Kydland, F. E. and Prescott, E. C. (1982), "Time to Build and Aggregate Fluctuations," *Econometrica*, 50, 1345–1370.

[37] Lahiri, S. N. (1999), "Theoretical Comparisons of Block Bootstrap Methods," *Annals of Statistics*, 27, 386–404.

[38] 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.

[39] Meir, R. (2000), "Nonparametric Time Series Prediction Through Adaptive Model Selection," *Machine Learning*, 39, 5–34.

[40] Mohri, M. and Rostamizadeh, A. (2009), "Rademacher Complexity Bounds for Non-IID Processes," *Advances in Neural Information Processing Systems*, 21, 1097–1104.

[41] Mohri, M. and Rostamizadeh, A. (2010), "Stability Bounds for Stationary φ-mixing and β-mixing Processes," *Journal of Machine Learning Research*, 11, 789–814.

[42] Mokkadem, A. (1988), "Mixing properties of ARMA processes," *Stochastic processes and their applications*, 29, 309–315.

[43] Pál, D., Póczos, B., and Szepesvári, C. (2010), "Estimation of Rényi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs," Tech. Rep. 1003.1954v1, arXiv.

[44] Paninski, L. (2003), "Estimation of Entropy and Mutual Information," *Neural Computation*, 15, 1191–1253.

[45] Politis, D. and White, H. (2004), "Automatic Block-Length Selection for the Dependent Bootstrap," *Econometric Reviews*, 23, 53–70.

[46] 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.

[47] Sargent, T. J. (1989), "Two Models of Measurements and the Investment Accelerator," *The Journal of Political Economy*, 97, 251–287.

[48] Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

[49] Smets, F. and Wouters, R. (2007), "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 97, 586–606.

[50] Solow, R. M. (1957), "Technical Change and the Aggregate Production Function," *The Review of Economics and Statistics*, 39, 312–320.

[51] Steinwart, I. and Christmann, A. (2009), "Fast Learning from Non-I.I.D. Observations," Tech. Rep. 2009-1061, NIPS.

[52] Vidyasagar, M. (1997), *A Theory of Learning and Generalization*, Springer Verlag.

[53] Wildi, M. (2009), "Real-Time US-Recession Indicator (USRI) A Classical Cycle Perspective with Bounceback," Tech. rep., Institute of Data Analysis and Process Design.

[54] Yu, B. (1994), "Rates of Convergence for Empirical Processes of Stationary Mixing Sequences," *The Annals of Probability*, 22, 94–116.

[55] 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, 771–783.

[56] Zhu, X., Rogers, T., and Gibson, B. (2009), "Human Rademacher Complexity," Tech. rep., NIPS.