Time series forecasting: model evaluation and selection using nonparametric risk bounds

Daniel J. McDonald
Carnegie Mellon University
danielmc@cmu.edu

Version: December 21, 2011

Abstract

We derive generalization error bounds — bounds on the expected inaccuracy of the predictions — for traditional time series forecasting models. These bounds allow forecasters to select among competing models and to guarantee that with high probability, their chosen model will perform well without making strong assumptions about the data generating process or appealing to asymptotic theory. Extending results from statistical learning theory, we demonstrate how these techniques can benefit economic and financial forecasters interested in choosing models which behave well under uncertainty and mis-specification. We provide results which apply to many standard economic and financial forecasting tools including VARs, state space models, linearized DSGEs, etc.

1 Introduction

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 paper 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.

To fix ideas, imagine IID data $((Y_1, X_1), \ldots, (Y_n, X_n))$ with $(Y_i, X_i) \in \mathcal{X} \times \mathcal{Y}$, some prediction function $f : \mathcal{X} \to \mathcal{Y}$, and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ which measures the cost of poor predictions. The *generalization error* or *risk* of f is

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

where the expectation is taken with respect to ν, the joint distribution of (Y, X). The generalization error measures the inaccuracy of our predictions when we use f on future data, making it a very natural criterion for model selection or as a way to provide performance guarantees. Of course, to actually calculate it, we need knowledge of the distribution ν and a single fixed function f to use for predictions, neither of which is common. Because explicitly calculating the risk is infeasible,

forecasters attempt to estimate it, necessitating specific assumptions on ν. An alternative which we employ here is to derive upper bounds which hold for large classes of models and distributions.

There are many ways to estimate the generalization error, and a comprehensive review is beyond the scope of this paper. Traditionally, time series analysts have performed model selection by a combination of empirical risk minimization (choosing a function by minimizing the empirical analogue of (1)), more-or-less quantitative inspection of the residuals — e.g., the Box-Ljung test; see [45] — 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 [41] 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 vector autoregressive (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 [17] compare Greenbook forecasts produced by the Federal Reserve with the predictions of various atheoretical methods, however they ignore periods of high volatility such as 1979–1983. Christoffel et al. [9] 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. [13] reestimate DSGE-VARs recursively based on rolling 30 year samples before forecasting two year periods between 1985 and 2000. Smets and Wouters [46] compare dynamic stochastic general equilibrium (DSGE) models with Bayesian VARs over a similar period. Edge and Gurkaynak [16] 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. [21], Gerali et al. [19] and Gertler and Karadi [20]). 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 [17]. 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.

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), we can still make strong claims about the ability of the model to predict the future.

Our main results in Section 4 assert that for wide classes of time series models (including VARs, state-space models, and linearized DSGEs), the expected cost of poor predictions is bounded by the model's in-sample performance inflated by a term which balances the amount of observed data with the complexity of the model. The bound holds with high probability under the unknown distribution ν assuming only mild conditions — existence of some moments, stationarity, and the decay of temporal dependence as data points become widely separated in time. As a preview, the following theorem provides the general form of the result.

Meta-Theorem 1.1 (Essentially)**.** *Given a time series $Y_1, \ldots, Y_n$ satisfying some mild conditions and a prediction function f chosen from a class of functions $\mathcal{F}$ (possibly by using the observed sample), then, with probability at least $1 - \eta$,*

$$R(f) \leq \widehat{R}_n(f) C_{\nu,\mathcal{F}}(\eta, n) \tag{2}$$

where $R(f)$ is the expected cost of making prediction errors on new samples, $\widehat{R}_n(f)$ is the average cost of in-sample prediction errors, $C_{\nu,\mathcal{F}}(\eta, n) \geq 1$ balances the complexity of the model from which f was chosen with the amount of data used to choose it.

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.* Our 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.

The remainder of this paper is structured as follows. Section 2 provides motivation and background for our results, giving intuition in the IID setting by focusing on concentration of measure ideas and characterizations of model complexity. Section 3 gives the explicit assumptions we make and describes how to leverage powerful ideas from time series to generalize the IID methods. Section 4 states and proves risk bounds for the time series forecasting setting, while we demonstrate how to use the results in §5. Finally, Section 6 concludes and illustrates the path toward generalizing our methods to more elaborate model classes.

2 Statistical learning theory

Our goal 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. To orient readers new to this approach, we sketch how

classical results in the IID setting are obtained. For simplicity, let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq [-K/2, K/2]$. Let $f : \mathcal{X} \to \mathcal{Y}$ be some function used for making predictions of Y from X.

We define a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ which measures the cost of making poor predictions. Throughout, we will take

$$\ell(y, y') = \left|\left|y - y'\right|\right|, \tag{3}$$

where $||\cdot||$ is some appropriate norm. Then as before the risk of any predictor $f \in \mathcal{F}$ is given by

$$R(f) = \mathbb{E}_\nu\big[\, ||Y - f(X)|| \,\big], \tag{4}$$

where $(X, Y) \sim \nu$. The risk or generalization error measures the expected cost of using f to predict Y from X given a new observation.

Since the true distribution ν is unknown, so is $R(f)$, but we can attempt to estimate it based on only our observed data. We define the *training error* or *empirical risk* of f as

$$\widehat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} ||Y_i - f(X_i)|| \,. \tag{5}$$

In other words, the in-sample training error, $\widehat{R}_n(f)$, is the average loss over the actual training points. Because the true risk is an expectation value, we can say that

$$\widehat{R}_n(f) = R(f) + \gamma_n(f), \tag{6}$$

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. By the law of large numbers, for each fixed f, $\gamma_n(f) \to 0$ as $n \to \infty$, so, with enough data, we have a good idea of how well any given function will generalize to new data.

However, economists rarely use a single function f without adjustable parameters fixed for them in advance by theory. Rather, there is a class of plausible functions $\mathcal{F}$, possibly indexed by some parameters $\theta \in \Theta$, which we refer to as a model. We pick out one function (choose one particular parameter point) from the model class by minimizing the in-sample loss. This means

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{R}_n(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \left(R(f) + \gamma_n(f)\right). \tag{7}$$

Tuning the parameters so that $\widehat{f}$ fits the training data well thus conflates predicting future data well (low $R(\widehat{f})$, the true risk) with exploiting the accidents and noise of the training data (large negative $\gamma_n(\widehat{f})$, finite-sample noise). The true risk of $\widehat{f}$ will generally be bigger than its in-sample risk precisely because we 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:

$$R(\widehat{f}) - \widehat{R}_n(\widehat{f}) \leq \sup_{f \in \mathcal{F}} |\gamma_n(f)| = \Gamma_n(\mathcal{F}) \,. \tag{8}$$

The main goal of statistical learning theory is to mathematically control $\Gamma_n(\mathcal{F})$ by finding tight bounds on it while making minimal assumptions about the unknown data-generating process; 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 usefully 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 $1-\eta$,

$$\Gamma_n(\mathcal{F}) \leq \Phi(\Psi(\mathcal{F}), n, \eta) \ , \tag{9}$$

where $\Psi(\cdot)$ measures the complexity of the model $\mathcal{F}$. To give specific forms of $\Phi(\cdot)$, we 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, and we need to understand the complexity, $\Psi(\mathcal{F})$, so that we can claim $R(f)$ 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 us to show, despite little knowledge of the data generating process, how bad the $\widehat{f}$ which we choose will be at forecasting future observations.

2.1 Concentration

The first step to controlling the difference between the empirical and expected risk is to show that for a single fixed $f \in \mathcal{F}$, $R(f) - \widehat{R}_n(f)$ is small with high probability. This follows from a standard Chernoff bound coupled with Hoeffding's inequality [23].

Theorem 2.1. *For any $f \in \mathcal{F}$,*

$$\mathbb{P}_\nu(|R(f) - \widehat{R}_n(f)| \geq \epsilon) \leq 2\exp\left\{-\frac{2n\epsilon^2}{K^2}\right\}. \tag{10}$$

Proof. First, we use Hoeffding's inequality to bound the moment generating function of the difference $R(f) - \widehat{R}_n(f)$:

$$\mathbb{E}[\exp\{s(R(f) - \widehat{R}_n(f))\}] = \prod_{i=1}^{n} \mathbb{E}\left[\exp\left\{\frac{s}{n}\left[R(f) - \ell(Y_i, f(X_i))\right]\right\}\right] \tag{11}$$

$$\leq \prod_{i=1}^{n} \exp\left\{\frac{s^2K^2}{8n^2}\right\} = \exp\left\{\frac{s^2K^2}{8n}\right\}. \tag{12}$$

Now, for a fixed f, we have $\mathbb{E}[\widehat{R}_n(f)] = R(f)$. Therefore we can apply Markov's inequality and the moment generating function bound:

$$\mathbb{P}_\nu(R(f) - \widehat{R}_n(f) > \epsilon) = \mathbb{P}_\nu\left(\exp\{s(R(f) - \widehat{R}_n(f))\} \geq \exp\{s\epsilon\}\right) \tag{13}$$

$$\leq \frac{\mathbb{E}\left[\exp\{s(R(f) - \widehat{R}_n(f))\}\right]}{\exp\{s\epsilon\}} \tag{14}$$

$$\leq \exp\{-s\epsilon\}\exp\left\{\frac{s^2K^2}{8n}\right\}. \tag{15}$$

This holds for all $s > 0$, so we can minimize the right hand side in s. This occurs for $s = 4n\epsilon/M^2$. Plugging in gives

$$\mathbb{P}_\nu(R(f) - \widehat{R}_n(f) > \epsilon) \leq \exp\left\{-\frac{2n\epsilon^2}{K^2}\right\}. \tag{16}$$

Exactly the same argument holds for $\mathbb{P}_\nu(R(f) - \widehat{R}_n(f) < -\epsilon)$, so by a union bound, we have the result. □

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 $||y - f(x)|| < K$. In fact, even this assumption can be removed and replaced with some moment assumptions which will be the case for our main results.

Of course this bound holds for the single function f. Instead, we want a similar result to hold simultaneously over all functions $f \in \mathcal{F}$ and in particular, the $\widehat{f}$ we choose using the training data, i.e., we wish to bound $\mathbb{P}_\nu \left(\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}_n(f)| > \epsilon \right)$.

2.2 Capacity

For "small" models, we 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 we have

$$\mathbb{P}_\nu \left(\sup_{1 \le i \le N} |R(f_i) - \widehat{R}_n(f_i)| > \epsilon \right) \le \sum_{i=1}^{N} \mathbb{P}_\nu \left(|R(f_i) - \widehat{R}_n(f_i)| > \epsilon \right) \le N \exp\left\{ -\frac{2n\epsilon^2}{K} \right\}, \tag{17}$$

by Theorem 2.1. Most interesting models are not small in this sense, but given the appropriate way of counting functions, similar results can be derived.

There are a number of measures for the size or capacity of a model. Algorithmic stability [27, 5, 4] measures the sensitivity of the chosen function to small tweaks to the data. Similarly, maximal discrepancy [49] asks how different the predictions could be if two functions are chosen using two separate data sets. A more direct, functional analysis, approach leads to covering numbers [40, 39] which partitions functions $f \in \mathcal{F}$ into equivalence classes under some metric. Rademacher complexity [3] directly describes a model's ability to fit random noise. We focus on a measure which is both intuitive and powerful: Vapnik-Chervonenkis (VC) dimension [48, 49].

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

Definition 2.2. *Let $\mathbb{U}$ be some (infinite) set and S a finite subset of $\mathbb{U}$. Let $\mathcal{C}$ be a family of subsets of $\mathbb{U}$. We say 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 largest set S that can be shattered by $\mathcal{C}$ is the known as its VC dimension.

Definition 2.3 (VC dimension)**.** *The* Vapnik-Chervonenkis (VC) dimension *of a collection $\mathcal{C}$ of subsets of $\mathbb{U}$ is*

$$\text{VCD}(\mathcal{C}) := \sup\{|S| : S \subseteq \mathbb{U} \text{ and } S \text{ is shattered by } \mathcal{C}\}. \tag{18}$$

Using VC dimension to measure the capacity of function classes is straightforward. Define the indicator function $\mathbf{1}_A(x)$ 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

$$C_f = \{u \cup a : \mathbf{1}_{(0,\infty)}(f(u) - b) = 1, \quad u \in \mathbb{U}, \quad b \in \mathbb{R}\} \tag{19}$$

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 $\mathcal{F} = \{\mathbf{x} \mapsto \boldsymbol{\gamma} \cdot \mathbf{x} : \boldsymbol{\gamma} \in \mathbb{R}^p\}$ 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 is the model $\mathcal{F} = \{x \mapsto \sin(\omega x) : \omega \in \mathbb{R}\}$, which has only one free parameter, but $\text{VCD}(\mathcal{F}) = \infty$.[1]

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

Theorem 2.4 (Vapnik and Chervonenkis [50])**.** *Suppose that* $\text{VCD}(\mathcal{F}) = h$ *and* $0 \leq \ell(y, y') \leq K$*. Then,*

$$\mathbb{P}_\nu \left(\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}_n(f)| > \epsilon \right) \leq 4GF(2n, h) \exp\left\{ -\frac{n\epsilon^2}{K^2} \right\}, \tag{20}$$

where $GF(n, h) = \exp\{h(\log x/h + 1)\}$*.*

The proof of this theorem has a similar flavor to the union bound argument given in (17). Essentially, $GF(n, h)$ counts the effective number of functions in $\mathcal{F}$, i.e., how many can be told apart using only n observations. This theorem has as an immediate corollary a bound for the expected risk. Since the probability statement holds for all functions, it holds in particular for that function which minimizes the empirical risk, $\widehat{f}$.

Corollary 2.5. *For any* $\eta > 0$ *and any* $f \in \mathcal{F}$*, with probability at least* $1 - \eta$*,*

$$R(f) \leq \widehat{R}_n(f) + K\sqrt{\frac{\log GF(2n, h) - \log \eta/4}{n}}. \tag{21}$$

The only term that is random is the training error, hence the fact that this statement holds with high probability. The penalty term goes to zero as $n \to \infty$. Also, the right side 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 ν and it has nothing to do with asymptotic normality: it holds for any n.

These concentration results work well for independent data. The first shows exactly how fast averages concentrate around their expectations: exponentially fast in the size of the data. The second result generalizes the first from a single function to entire function classes. Both results depend critically on the independence of the random variables, however in the case of interest, we need to be able to handle dependent data. Because time-series data are dependent, the number of data points n in a sample $\mathbf{Y}_1^n$ 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 2.1 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.

3 Time series

In moving from the IID setting to time series forecasting, we need a number of modifications to our initial setup. To be more explicit, we present the following notation and definitions. Rather than observing in put out pairs (Y_i, X_i), we observe a single sequence of random variables $\mathbf{Y}_1^n :=$

[1] This result follows if we can show that for any positive integer J and any binary sequence $(r_1, \ldots, r_J)$, there exists a vector $(x_1, \ldots, x_J)$ such that $\mathbf{1}_{[0,1]}(\sin(\omega x_i)) = r_i$. If we choose $x_i = 2\pi 10^{-i}$, 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.

$(Y_1, \ldots, Y_n)$ where each Y_i takes values in a measurable space $\mathcal{Y}$.[2] We are interested in using functions which take past observations as inputs and predict future values of the process. Suppose, given data from time 1 to time n, we wish to predict time $n+1$.

Since we no longer have IID data, we will need a few restrictions on the sorts of dependent processes we can consider. We first remind the reader of the notion of (strict or strong) stationarity.

Definition 3.1 (Stationarity)**.** *A sequence of random variables* $\mathbf{Y}$ *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 t, only that the unconditional distribution of Y_t is constant in time. From among all the stationary processes, we will discuss only a subset thereof in which widely-separated observations are asymptotically independent. Without this assumption, convergence of the training error to the expected risk could occur arbitrarily slowly, preventing the derivation of finite sample results.[3] The next definition describes the nature of the serial dependence which we are willing to allow.

Definition 3.2 (β-Mixing)**.** *Consider a stationary random sequence* $\mathbf{Y}_{-\infty}^{\infty}$ *defined on a probability space* $(\Omega, \Sigma, \mathbb{P})$*. Let* $\sigma_i^j = \sigma(\mathbf{Y}_i^j)$ *be the* σ*-field of events generated by the appropriate collection of random variables. Let* $\mathbb{P}_0$ *be the restriction of* $\mathbb{P}$ *to* $\sigma_{-\infty}^0$*,* $\mathbb{P}_a$ *be the restriction of* $\mathbb{P}$ *to* σ_a^{∞}*, and* $\mathbb{P}_{0\otimes a}$ *be the restriction of* $\mathbb{P}$ *to* $\sigma(\mathbf{Y}_{-\infty}^0, \mathbf{Y}_a^{\infty})$*. The* coefficient of absolute regularity*, or* β-mixing coefficient*,* β_a*, is given by*

$$\beta_a := ||\mathbb{P}_0 \times \mathbb{P}_a - \mathbb{P}_{0\otimes a}||_{TV}, \tag{22}$$

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

This is only one of many equivalent characterizations of β-mixing (see Bradley [6] for others). This definition makes clear that a process is β-mixing if the joint probability of events which are widely separated in time increasingly approaches the product of the individual probabilities, i.e., that $\mathbf{Y}$ is asymptotically independent. Many common time series models are known to be β-mixing, and the rates of decay are known up to constant factors given the true parameters of the process. Among the processes for which such knowledge is available are ARMA models [37], GARCH models [7], and certain Markov processes — see Doukhan [14] for an overview of such results. Additionally, functions of these processes are β-mixing, so if $\mathbb{P}$ could be specified by a dynamic factor model or DSGE or VAR, the observed data would satisfy this condition.

Knowledge of β_a allows us 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 we determine the correct μ, we can use concentration results for IID data like those in Theorem 2.1 and 2.4 with small corrections.

4 Risk bounds

With the relevant background in place, we can put the pieces together to present our results. We 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

[2] We will take $\mathcal{Y} = \mathbb{R}^p$ throughout, but this assumption is not necessary.

[3] In fact, Adams and Nobel [1] demonstrate that for ergodic processes, finite VC dimension is enough to give consistency.

error of the chosen function $\widehat{f}$. In the remainder of this section, we redefine the appropriate concepts in the time series forecasting scenario, we state the necessary assumptions for our results, and we derive risk bounds for wide classes of economic forecasting models.

4.1 Setup and assumptions

We observe a finite subsequence of random vectors $\mathbf{Y}_1^n$ from a process $\mathbf{Y}_{-\infty}^{\infty}$ defined on a probability space $(\Omega, \Sigma, \mathbb{P}_\infty)$ such that $Y_i \in \mathbb{R}^p$. We make the following assumption on the infinite process.

Assumption A. *Assume that $\mathbb{P}_\infty$ is a stationary, β-mixing distribution with known mixing coefficients β_a, $\forall a > 0$.*[4]

Under stationarity, the marginal distribution of Y_t is the same for all t. We are mainly concerned with the joint distribution of sequences Y_1^{n+1} wherein we observe the first n observations and attempt to predict time $n+1$. For the remainder of this paper, we will call this joint distribution $\mathbb{P}$. Our results are easily extended to the case of predicting more than one step ahead, but the notation becomes cumbersome.

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

Definition 4.1 (Time series risk)**.**

$$R_n(f) := \mathbb{E}_{\mathbb{P}}\Big[\, ||Y_{n+1} - f(\mathbf{Y_1^n})|| \,\Big]. \tag{23}$$

The expectation is taken with respect to the joint distribution $\mathbb{P}$ and therefore depends on n. We 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, we define two versions of the training error depending on whether or not the memory of the prediction function f is fixed.

Definition 4.2 (Time series training error with memory d)**.**

$$\widehat{R}_n(f) := \frac{1}{n-d-1} \sum_{i=d}^{n-1} \Big|\Big|Y_{i+1} - f(\mathbf{Y}_{i-d+1}^i)\Big|\Big| \tag{24}$$

Definition 4.3 (Time series training error with growing memory (at least d))**.**

$$\widetilde{R}_n(f) := \frac{1}{n-d-1} \sum_{i=d}^{n-1} \Big|\Big|Y_{i+1} - f(\mathbf{Y}_1^i)\Big|\Big| \tag{25}$$

The first case is useful for standard VAR forecasting methods, while the second case as applicable to ARMA models, DSGEs, and linear state space models. Additionally, we are writing f as a fixed function, but the dimension of the argument changes with i. This is not an issue for functions

[4] Of course, in practice we do not know the data generating process, so we do not know β_a. McDonald et al. [32] shows how to estimate the mixing coefficients based on a sample from a mixing process.

which are linear in the data, as is the case with ARMA models, linear state-space models, and linearized DSGEs.[5] For nonlinear models, we will consider only the fixed memory version.

To control the generalization error for time series forecasting, we make one final assumption which is more general than the bounded loss assumption we used in §2, in particular, it allows for unbounded loss as long as we can control some moments of the risk.

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

$$1 \leq \frac{\left(\mathbb{E}_{\mathbb{P}}\left[\, ||Y_{n+1} - f(\mathbf{Y}_1^n))||^q \,\right]\right)^{1/q}}{R_n(f)} < M. \tag{26}$$

Assumption B is still quite general, allowing even some heavy tailed distributions. Furthermore, with slight adjustments (see [48]), we can allow $1 < q \leq 2$. It should be noted that the lower bound is trivially true for any loss distribution.

4.2 Fixed memory

We can now state our results giving finite sample risk bounds for the problem of time series forecasting. We begin with the fixed memory setting before allowing the memory length to grow.

Theorem 4.4. *Given a sample $\mathbf{Y}_1^n$ such that Assumptions A and B hold, suppose that the model class $\mathcal{F}$ has a fixed memory length $d < n$. Let μ and a be integers such that $2\mu a + d \leq n$. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} > \epsilon\right) \leq 8 \exp\left\{ \mathrm{VCD}(\mathcal{F}) \left(\ln \frac{2\mu}{\mathrm{VCD}(\mathcal{F})} + 1 \right) - \frac{\mu \epsilon^2}{4\tau^2(q) M^2} \right\} + 2(\mu - 1)\beta_{a-d}, \tag{27}$$

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 effective number of observations $\mu < n$, we 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. Moreover, given some model class $\mathcal{F}$, we can say exactly how much data is required to have good control of the prediction risk. As the effective data size increases, $\mathcal{E} \to 0$ 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

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} > \epsilon\right) \leq 8 \exp\left\{ \ln 2\mu + 1 - \frac{\mu \epsilon^2}{4} \right\} \tag{28}$$

where we have taken the VC dimension to be one and we ignore the extra penalty from the mixing coefficient. Our goal is to minimize ϵ, thereby ensuring that the relative difference between the

[5] 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 [18], DeJong and Dave [10] or Dejong et al. [11]

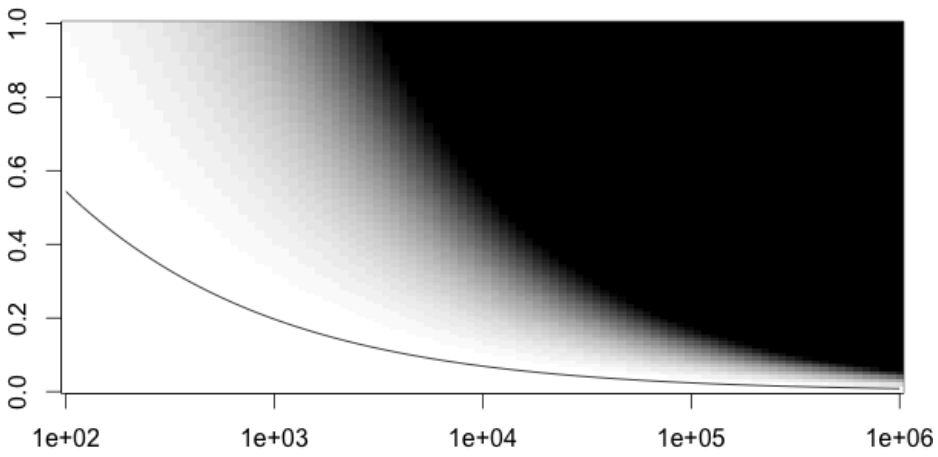

Figure 1: Visualizing the tradeoff between confidence (ϵ, y-axis) and effective data (μ, x-axis). The black curve indicates the region where the bound becomes trivial. Below this line, the probability is bounded by 1. Darker colors indicate lower probability of the "bad" event — that the difference in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

expected risk and the training risk is small. At the same time we want to minimize the right side of the bound so that the probability of "bad" outcomes — events such that the difference in risks exceeds ϵ — is small. Of course we want to do this with as little data as possible, but the smaller we take ϵ, the larger we must take μ to compensate. We illustrate this tradeoff in Figure 1.

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 $\widehat{R}_n(f) \geq 0$, and due to the supremum, events where the training error exceeds the expected risk are irrelevant. Therefore, we are only concerned with $0 \leq \widehat{R}(f) \leq R_n(f)$, so we need only consider $0 \leq \epsilon \leq 1$.

The figure is structured so that movement toward the origin is preferable. We 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 bad event occurs with probability no larger than one). The desire for the 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 we had a budget constraint trading ϵ and data (i.e. a line with negative slope), we could optimize within the budget set to find the lowest probability allowable.

Before we prove Theorem 4.4 we will state a corollary which appears in a form which is occasionally more convenient.

Corollary 4.5. *Under the conditions of Theorem 4.4, 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* $\widehat{f}$*):*

$$R_n(f) \leq \frac{\widehat{R}_n(f)}{(1 - \mathcal{E})_+}. \tag{29}$$

Here

$$\mathcal{E} = \frac{2M\tau(q)}{\sqrt{\mu}} \sqrt{\text{VCD}(\mathcal{F}) \left(\ln \frac{2\mu}{\text{VCD}(\mathcal{F})} + 1 \right) - \ln(\eta'/8)}, \tag{30}$$

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

We now prove both Theorem 4.4 and Corollary 4.5 to provide the reader with some intuition for the types of arguments necessary. We defer proof of the remainder of the theorems in this section to the appendix.

Proof of Theorem 4.4. 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. Let a and μ be non-negative integers such that $2a\mu = n$. Now divide $\mathbf{Y}_1^n$ into 2μ blocks, each of length a. Identify the blocks as follows:

$$U_j = \{Y_i : 2(j-1)a + 1 \leq i \leq (2j-1)a\}, \tag{31}$$
$$V_j = \{Y_i : (2j-1)a + 1 \leq i \leq 2ja\}. \tag{32}$$

Let $\mathbf{U}$ be the entire sequence of odd blocks U_j, and let $\mathbf{V}$ be the sequence of even blocks V_j. Finally, let $\mathbf{U}'$ be a sequence of blocks which are mutually independent and such that each block has the same distribution as a block from the original sequence. That is construct U_j' such that

$$\mathcal{L}(U_j') = \mathcal{L}(U_j) = \mathcal{L}(U_1), \tag{33}$$

where $\mathcal{L}(\cdot)$ means the probability law of the argument.

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,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} > \epsilon\right) = \mathbb{P}\left(\sup_{f\in\mathcal{F}} \left[\frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{2R_n(f)} + \frac{R_n(f) - \widehat{R}_{\mathbf{V}}(f)}{2R_n(f)}\right] > \epsilon\right) \tag{34}$$
$$\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{R_n(f)} + \sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{V}}(f)}{R_n(f)} > 2\epsilon\right) \tag{35}$$
$$\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{R_n(f)} > \epsilon\right) + \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{V}}(f)}{R_n(f)} > \epsilon\right) \tag{36}$$
$$= 2\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{R_n(f)} > \epsilon\right). \tag{37}$$

Now, apply Lemma 4.1 in Yu [52] (Lemma A.1 in Appendix A) to the indicator of the event $\left\{\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{R_n(f)} > \epsilon\right\}$. This allows us to move from statements about dependent blocks, to statements about independent blocks with a slight correction. Therefore we have,

$$2\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{R_n(f)} > \epsilon\right) \leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}'}(f)}{R_n(f)} > \epsilon\right) + 2(\mu-1)\beta_{a-d} \tag{38}$$

where the probability on the right is for the σ-field generated by the independent block sequence

$\mathbf{U}'$. For convenience, define $R_n^q(f) := \mathbb{E}\left[||Y_{n+1} - f(\mathbf{Y}_1^n))||^q\right]$ despite the obvious abuse of notation.

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}'}(f)}{R_n(f)} > \epsilon\right) = \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}'}(f)}{R_n(f)} \frac{1}{M} > \frac{\epsilon}{M}\right) \tag{39}$$

$$\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}'}(f)}{R_n(f)} \frac{R_n(f)}{\sqrt[q]{R_n^q(f)}} > \frac{\epsilon}{M}\right) \tag{40}$$

$$= \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}'}(f)}{\sqrt[q]{R_n^q(f)}} > \frac{\epsilon}{M}\right) \tag{41}$$

$$= \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}'}(f)}{\sqrt[q]{R_n^q(f)}} > \tau(q)\frac{\epsilon}{M\tau(q)}\right) \tag{42}$$

$$\leq 8\exp\left\{\text{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\text{VCD}(\mathcal{F})} + 1\right) - \frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\} + 2(\mu-1)\beta_{a-d}, \tag{43}$$

where we have applied Theorem 5.4 in Vapnik [48] (Lemma A.2) to bound the independent blocks. This result is Theorem 4.4. 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. We get that for all $f \in \mathcal{F}$, with probability at least $1-\eta$,

$$\frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} \leq \epsilon. \tag{44}$$

Solving the equation

$$\eta' = 8\exp\left\{\text{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\text{VCD}(\mathcal{F})} + 1\right) - \frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\} \tag{45}$$

implies

$$\epsilon = \frac{2M\tau(q)}{\sqrt{\mu}}\sqrt{\text{VCD}(\mathcal{F})\left(\ln\frac{2\mu}{\text{VCD}(\mathcal{F})} + 1\right) - \ln(\eta'/8)} = \mathcal{E}. \tag{46}$$

□

The only obstacle to the use of Theorem 4.4 is knowledge of the VCD $\mathcal{F}$. For some models, the VC dimension can be calculated explicitly.

Lemma 4.6. *For $\mathcal{F}_{AR}(d)$ the class of AR(d) models we have*

$$\text{VCD}(\mathcal{F}_{AR}(d)) = d + 1. \tag{47}$$

Lemma 4.6 applies equally to Bayesian ARs. However, this is likely too conservative as the prior tends to restrict the effective complexity of the function class.[6] For regularized methods, or non-linear methods where the VC dimension is unknown, we can estimate the VC dimension via simulation and make a slight correction to the risk bound. This estimated bound will also applies to state-space models, dynamic factor models, or even dynamic stochastic general equilibrium (DSGE) models. The simulation procedure was developed in Vapnik et al. [51] and is shown in Algorithm 1.

Algorithm 1 Estimate VC dimension

Given a model $\mathcal{F}$ and a grid of design points $n_1, \ldots, n_k$, generate regression points $\widehat{\xi}(n_\ell)$. Then use nonlinear least squares to estimate the VC dimension.

1: Choose a grid of integer values $n_1, \ldots, n_k$.
2: **for** $\ell = 1 \to k$ **do**
3: **while** $1 \leq j \leq m$ **do**
4: Generate a stationary process $\mathbf{Z}$ of length $2n_\ell + d$ such that $\mathbb{E}[Z_i] = 0$ and the marginal distribution of each element has support on $\mathcal{Y} = \mathbb{R}^p$ (this can be a white noise process).
5: Define $\mathbf{W}' := -\mathbf{Z}_{n_\ell+d+1}^{2n_\ell+d}$. Define $\mathbf{W} := (\mathbf{Z}_{d+1}^{2n_\ell+d}, \mathbf{W}')$.
6: Choose a function $\widehat{f} \in \mathcal{F}$ using $\mathbf{W}$.
7: Call $\widetilde{\mathbf{Z}} := (\mathbf{Z}_1^d, \mathbf{W})$.
8: Choose a parameter $b \in \mathbb{R}^p$ as

$$b = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n_\ell} \sum_{i=1}^{2n_\ell} \mathbf{1}\Big(\operatorname{sgn}(\widehat{f}(\widetilde{\mathbf{Z}}_i^{i+d-1}) - b) \neq \operatorname{sgn}(\widetilde{\mathbf{Z}}_{i+d+1}) \Big).$$

For $\mathbf{Z}$ a vector, take $\operatorname{sgn}(\mathbf{Z})$ to be a vector of componentwise signs and take the indicator to be the event that all the signs are the same.
9: Calculate the following error measure of the estimated function $\widehat{f}$ on the generated data $\mathbf{Z}$

$$\widehat{\xi}_j(n_\ell) = \left| \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \mathbf{1}\Big(\operatorname{sgn}(\widehat{f}(\mathbf{Z}_i^{i+d-1}) - b) \neq \operatorname{sgn}(\mathbf{Z}_{i+d}) \Big) - \right.$$
$$\left. - \frac{1}{n_\ell} \sum_{i=n_\ell}^{2n_\ell} \mathbf{1}\Big(\operatorname{sgn}(\widehat{f}(\mathbf{Z}_i^{i+d-1}) - b) \neq \operatorname{sgn}(\mathbf{Z}_{i+d}) \Big) \right|.$$

10: **end while**
11: Set $\widehat{\xi}(n_\ell) = \frac{1}{m} \sum_{i=1}^m \widehat{\xi}_j(n_\ell)$.
12: **end for**
13: Estimate the VC dimension as

$$\widehat{h} = \underset{h>0}{\operatorname{argmin}} \frac{1}{k} \sum_{\ell=1}^{k} (\widehat{\xi}(n_\ell) - \Phi_h(n_\ell))^2$$

where

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

and $c = 0.16$, $c' = 1.2$, and $c'' = 0.15$.

The algorithm amounts to simulating data sets of different sizes many times. To use the algorithm, choose integers k, the number of different sizes, and m, the number of replications for each size data set, by trading off computational time and desired accuracy. Then estimate the model km times. McDonald et al. [31] derives the accuracy of the estimate and shows how to use the result to get generalization error bounds. This leads to the following theorem which controls the prediction risk using estimated VC dimension.

Theorem 4.7. *Choose integers k and m to produce an estimate of VC dimension using Algorithm 1 which is accurate up to some tolerance δ. Given a sample $\mathbf{Y}_1^n$ such that Assumptions A and B hold, suppose that the model class $\mathcal{F}$ has a fixed memory length $d < n$. We have the following high probability[7] bound for all $\epsilon > 0$:*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} > \epsilon\right) \tag{48}$$

$$\leq 8\exp\left\{(\widehat{h}+\delta)\left(\ln\frac{2\mu}{(\widehat{h}+\delta)}+1\right) - \frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\}(1-\varphi) + 2(\mu-1)\beta_{a-d}(1-\varphi) + \varphi, \tag{49}$$

where $\widehat{h}$ is the estimated VC dimension and

$$\varphi = 13\exp\left\{-\frac{mk\delta^2}{16C}\right\}. \tag{50}$$

For a thorough explanation of how to choose the tolerance δ and the explicit form of the constant C, see McDonald et al. [31]. The term φ goes to zero exponentially quickly as k or m increase, thus, given unlimited computational time, we essentially recover the result in Theorem 4.4 even with the VC dimension estimated via simulation rather than known *a priori*. We now state a corollary analogous to Corollary 4.5.

Corollary 4.8. *Under the conditions of Theorem 4.7, the following bound holds simultaneously for all $f \in \mathcal{F}$ (including the minimizer of the empirical risk $\widehat{f}$) with probability at least $1-\eta$, for all $\eta > 2(\mu-1)\beta_{a-d}(1-\varphi)+\varphi$*

$$R_n(f) \leq \frac{\widehat{R}_n(f)}{(1-\mathcal{E})_+}. \tag{51}$$

Here

$$\mathcal{E} = \frac{2M\tau(q)}{\sqrt{\mu}}\sqrt{(\widehat{h}+\delta)\left(\ln\frac{2\mu}{\widehat{h}+\delta}+1\right) - \ln\frac{\eta'}{8(1-\varphi)}} \tag{52}$$

and $\eta' = \eta - 2(\mu-1)\beta_{a-d}(1-\varphi) - \varphi$.

[6] Here we should mention that these risk bounds are frequentist in nature. Our meaning is that if we 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 [28], Müller [38] or Shalizi [43]

[7] Technically, the data come from the distribution $\mathbb{P}$ while Algorithm 1 uses some other distribution, say $\mathbb{P}_1$, to generate simulated data. Therefore, the probability statement in this theorem is with respect to the product measure $\mathbb{P}\times\mathbb{P}_1$. For the result to hold, we must have that $\mathbb{P}$ and $\mathbb{P}_1$ are measures over the same probability space and that the real and simulated data are statistically independent.

4.3 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, we can recover bounds similar to our previous results even for state-space models.

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

$$\widehat{\mathbf{Y}}_{d+1}^{n+1} = \mathbf{B}\mathbf{Y}_1^n \tag{53}$$

where

$$\mathbf{B} = \begin{bmatrix} b_{n,n} & b_{n,n-1} & \dots & b_{n,d} & \dots & b_{n,1} \\ 0 & b_{n-1,2} & \dots & b_{n-1,d} & \dots & b_{n-1,1} \\ \vdots & & \ddots & & & \\ 0 & & & b_{d,d} & \dots & b_{d,1} \end{bmatrix}. \tag{54}$$

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

Theorem 4.9. *Given a sample $\mathbf{Y}_1^n$ such that Assumptions A and B hold, suppose that the model class $\mathcal{F}$ is linear in the data and has growing memory. Fix some $1 \leq d < n$. 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 $2\mu a + d \leq n$. Then, with high probability*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{R_n(f) - \widetilde{R}_n(f) - \Delta_d(f)}{R_n(f)} > \tau(q)\epsilon \right) \leq \Phi \tag{55}$$

where Φ is given by either the right hand side of (27) or by (49), and

$$\Delta_d(f) = \mathbb{E}\big[\, ||Y_1|| \,\big] \left|\left| \sum_{j=1}^{n-d-1} b_{n,j} \right|\right| + \frac{1}{n-d-1} \sum_{i=d+1}^{n-1} \left|\left| \sum_{j=1}^{i-d} b_{i,j} y_j \right|\right|. \tag{56}$$

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 $\mathbb{E}\big[\, ||Y_1|| \,\big]$ which is not necessarily desirable. However, Assumption B has the consequence that $\mathbb{E}\big[\, ||Y_1|| \,\big] \leq L < \infty$. Finally, we will need $\sum_{j=1}^{n} ||b_{i,j}||$ to be bounded $\forall n$ or $\Delta_d(f) \to \infty$ as $n \to \infty$.

Corollary 4.10. *Given a sample $\mathbf{Y}_1^n$ such that Assumptions A and B hold, suppose that the model class $\mathcal{F}$ is linear in the data and has growing memory. Fix some $1 \leq d < n$. 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 $2\mu a + d \leq n$. Then, with probability at least $1 - \eta$, for η as in Theorem 4.4 or 4.7, we have*

$$R_n(f) \leq \frac{\widetilde{R}_n(f) + \Delta_d(f)}{(1 - \mathcal{E})_+} \tag{57}$$

where $\mathcal{E}$ and η' can be as in Theorem 4.4 or 4.7.

To apply Theorem 4.9, we describe the form of the linear Gaussian state space model. We 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, $\mathcal{F}_{SS}$:

$$\begin{aligned} y_t &= Z\alpha_t + \epsilon_t, & \epsilon_t &\sim \mathrm{N}(0, H), \\ \alpha_{t+1} &= T\alpha_t + \eta_{t+1}, & \eta_t &\sim \mathrm{N}(0, Q), \\ & & \alpha_1 &\sim \mathrm{N}(a_1, P_1). \end{aligned} \tag{58}$$

We make no assumptions about the dimensionality of the parameter matrices Z, 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 our assumptions about the data generating process.. While $\textsc{vcd}(\mathcal{F}_{SS})$ is unknown in general, we can estimate it with Algorithm 1 for any sort of forecasting model which has this form, including linearized DSGEs.

Algorithm 2 Kalman filtering

Recursively generate minimum mean squared error predictions $\widehat{Y}_t$ using the state space model in (58).

1: Set $\widehat{Y}_1 = Za_1$.
2: **while** $1 \leq t \leq n$ **do**
3: Filter

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

4: Predict

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

5: **end while**

To forecast using $\mathcal{F}_{SS}$, one uses the Kalman filter [26]. The algorithm proceeds recursively as shown in Algorithm 2. 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 [15]) that given the parameter matrices, the (maximum *a posteriori*) forecast of y_t is given by

$$\widehat{y}_{t+1} = Z\sum_{j=1}^{t-1}\prod_{i=j+1}^{t} L_iK_jy_j + ZK_ty_t + Z\prod_{i=1}^{t} L_ia_1 \tag{59}$$

This yields the form of $\Delta_d(f)$ for linear state space models. We therefore have the following corollary to Theorem 4.9.

Corollary 4.11. *Suppose that our function class $\mathcal{F}$ corresponds to the state-space model specified in (58). Let $1 < d < n$. Then the following bound holds simultaneously for all $f \in \mathcal{F}$: with probability at least $1 - \eta$, for η as in Theorem 4.4 or 4.7, we have*

$$R_n(f) \leq \frac{\widetilde{R}_n(f) + \Delta_d(f)}{(1 - \mathcal{E})_+} \tag{60}$$

where $\mathcal{E}$ and η' can be as in Theorem 4.4 or 4.7, and

$$\Delta_d(f) = \mathbb{E}\big[\,||Y_1||\,\big] \left|\left| \sum_{j=1}^{n-d} \prod_{i=j+1}^{n} L_i K_j \right|\right| + \frac{1}{n-d-1} \sum_{t=d+1}^{n-1} \left|\left| \sum_{j=1}^{t-d} \prod_{i=j+1}^{t} L_i K_j y_j \right|\right|. \tag{61}$$

It is simple to compute $\Delta_d(f)$ using Kalman filter output. The corollary allows us 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.

5 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, we show how to use some of the results above. We first estimate a simple stochastic volatility model using IBM return data and calculate the bound for the predicted volatility using Theorem 4.9. We then discuss the principle of structural risk minimization focusing on how to use our results to select among competing forecasting models.

5.1 Stochastic volatility model

To demonstrate how to use our results, we 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 2 shows the returns series.

The model we investigate is given by

$$y_t = \sigma z_t \exp(\rho_t/2), \qquad z_t \sim \mathrm{N}(0,1), \tag{62}$$

$$\rho_{t+1} = \phi \rho_t + w_t, \qquad w_t \sim \mathrm{N}(0, \sigma_\rho^2), \tag{63}$$

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. [22]. We transform the model as follows:

$$\log y_t^2 = \kappa + \rho_t + \xi_t, \tag{64}$$

$$\xi_t = \log z_t^2 - \mathbb{E}[\log z_t^2], \tag{65}$$

$$\kappa = \log \sigma^2 + \mathbb{E}[\log z_t^2]. \tag{66}$$

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 3.274.

To match the data to the model, we 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

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

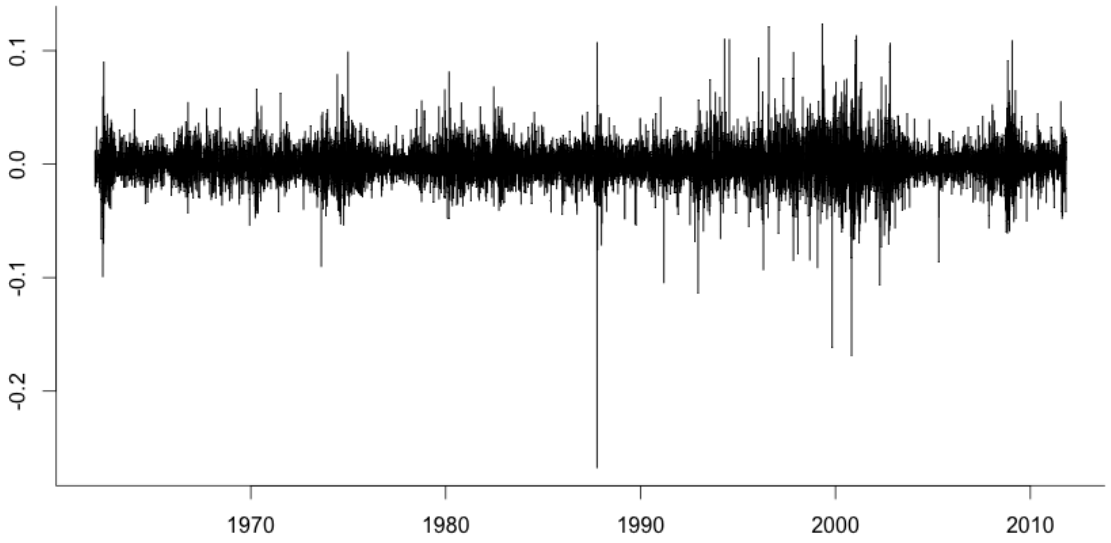

Figure 2: Daily log returns for IBM

Minimizing this gives estimates $\kappa = -9.72$, $\phi = 0.993$, and $\sigma_\rho^2 = 0.006$. Taking the loss function to be root mean squared error gives a training error of 1.823.

To actually calculate the bound, we need a few assumptions. First, we take $\beta_a = \exp\{-(1 + 2a)\}$. Such an exponential rate could be easily obtained if the data were generated by an ARMA process [37] and is common in the literature on mixing (cf. [34, 52]). Second, we take $q = 3$ and $M = 2$. These choices 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 [24] as

$$M_q = \pi^{\frac{q-1}{2q}} \Gamma^{1/q} \left(\frac{q+1}{2} \right). \tag{67}$$

Using this formula, we get $M_3 = 1.46$, but we use 2 for convenience.

Combining these assumptions with the estimated VC dimension for the stochastic volatility model will allow us to calculate a bound for the prediction risk. Using Algorithm 1 for appropriately chosen m and k gives an estimated VC dimension $\widehat{h} = 3$ give or take $\delta = 1$ and confidence $\varphi < 0.01$. However, for $d = 2$, we have that the VC dimension can be no larger than 3, thus, we may use Corollary 4.11 with η as in Corollary 4.5, i.e., we can take the VC dimension to be 3. Finally, taking $\mu = 846$, $a = 7$, and $d = 2$, we calculate $\Delta_2(f) = 0.776 + 0.844 = 1.62$. The result is the bound

$$R_n(f) \leq 9.81 \tag{68}$$

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 is highly correlated and so despite the fact that n is large, the effective sample size μ is relatively small.

For comparison, we also computed the bound for forecasts produced with an AR(2) model (with intercept) and with the mean alone. 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

Model	Training error	Risk bound	AIC	BIC
SV	1.82	9.81	0.959	0.959
AR(2)	1.88	5.37	0.987	0.988
Mean	1.91	3.46	1	1

Table 1: This table shows the training error and risk bounds for 3 models. AIC and BIC are given as ratios to the Mean, the smaller the value, the more support for that model.

size, this gain may be spurious: it is likely that the stochastic volatility model is simply over-fitting.

5.2 Structural risk minimization

Our 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 $\mathcal{F}$. Or one may have many different forecasting models from which the forecaster would like to choose the best. This scenario 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 $\widehat{f}_1, \widehat{f}_2, \ldots$, we wish 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, we therefore penalize the empirical risk and select 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, we 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 4 allow one 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 we want to make the prediction risk as small as possible, we 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, by minimizing both the empirical risk and the VC dimension, we can choose that model and function which has the smallest prediction risk, a claim which other model selection procedures cannot make [49, 30].

6 Conclusion

This paper demonstrates how to control the generalization error of common macroeconomic forecasting models — ARMA models, vector autoregressions (Bayesian or otherwise), linearized dy-

namic stochastic general equilibrium models, and linear state space models. The results we derive give upper bounds on the risk which hold with high probability while requiring only weak assumptions on the true data generating process. These bounds are finite sample in nature, unlike standard model selection penalties such as AIC or BIC. 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 we have stated these results in terms of standard economic forecasting models, they have very wide applicability. Theorems 4.4 and 4.7 apply to any forecasting procedure with fixed memory length, linear or non-linear. This covers even nonlinear 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 4.9 applies only 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 we 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 [34] and Mohri and Rostamizadeh [35, 36]. Those results require bounded loss functions as in the IID setting, making them less general than our 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. [33] 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.

Rather than deriving bounds theoretically, one could attempt to estimate bounds on the risk. While cross-validation is difficult, nonparametric bootstrap procedures may do better. A fully nonparametric version is possible using the circular bootstrap reviewed in Lahiri [29]. Bootstrapping lengthy out-of-sample sequences for testing fitted model predictions yields intuitively sensible estimates of $R_n(f)$, however, there is no theory that supports the coverage claim. Also, while models like VARs can be fit quickly to simulated data, general state-space models, let alone DSGEs, require large amounts of computational power.

Computational concerns also constrain our ability to estimate VC dimension via Algorithm 1. While we can estimate $\text{VCD}(\mathcal{F})$ to arbitrary precision, the algorithm requires the model to be fit $m \times k$ times. For DSGE models, this is infeasible, even with appropriate parallelization and high quality maximization routines. Another possible tack would be to use the DSGE to regularize a VAR or VARMA model. Using the DSGE as a source of prior information as in Del Negro and Schorfheide [12] may lead to a simpler procedure to estimate the VC dimension of an equivalent but computationally more difficult model. Another possible simplification along the same lines could use the methods of Juselius and Franchi [25] to convert the DSGE into a set of implementable restrictions on a cointegrated VAR.

While our results are a crucial first step to the analysis of time series forecasts, many avenues remain for future exploration. To gain a more complete picture of the performance of time series forecasting algorithms, we would ideally wish to derive minimax lower bounds (cf. Tsybakov [47]). These would give us an idea of the smallest risk we could hope to achieve using any forecasting model in some larger model class. We could then ask whether any of the common models actually in use can approach this minimum. 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 we used a different prediction function from the model $\mathcal{F}$ [8, 42]? Remarkably, we can find forecasters which have low *ex post* regret, even if the data came from an adversary trying to make us perform badly. If we target regret rather than risk, we can actually ignore mixing, and even stationarity [44].

An increased recognition of the abilities and benefits of statistical learning theory can be of

tremendous aid to financial and economic forecasters. The results presented here represent an initial yet productive foray in this direction. They allow for principled model comparisons as well as high probability performance guarantees. Future work will only serve to sharpen our ability to measure predictive power.

Acknowledgements

The author gratefully acknowledges support from the Institute for New Economic Thinking. He is also incredibly thankful for the comments, advice, corrections, and intuition provided by Professors Cosma Shalizi and Mark Schervish.

A Auxiliary results

Lemma A.1 (Lemma 4.1 in [52]). *Let ϕ be a measurable function with respect to the block sequence $\mathbf{U}$ uniformly bounded by M. Then,*

$$|\mathbb{E}[\phi] - \widetilde{\mathbb{E}}[\phi]| \leq M\beta_a(\mu - 1), \tag{69}$$

where the first expectation is with respect to the dependent block sequence, $\mathbf{U}$, and $\widetilde{\mathbb{E}}$ is with respect to the independent sequence, $\mathbf{U}'$.

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

Lemma A.2 (Theorem 5.4 in Vapnik [48]). *Under Assumption B,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{\sqrt[q]{R_n^q(f)}} > \tau(q)\epsilon\right) \leq 4\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\ln\frac{2n}{\mathrm{VCD}(\mathcal{F})} + 1\right) - \frac{n\epsilon^2}{4}\right\}. \tag{70}$$

Lemma A.3 (Theorem 1.4 in McDonald et al. [31]). *Choose $\delta > \frac{4}{\sqrt{2mk}}\max\{24c_1, 29\}$. Let $\rho > 0$. Set*

$$\varphi = 13\exp\left\{-\frac{mk\delta^2}{16c_2c_3}\right\}. \tag{71}$$

Then, for any classifier $f \in \mathcal{F}$ where $\mathcal{F}$ has estimated VC dimension $\widehat{h}$, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|R_n(f) - \widehat{R}_n(f)\right| > \rho\right) \leq 4GF(\widehat{h} + \delta, 2n)\exp\{-n\rho^2\}(1 - \varphi) + \varphi. \tag{72}$$

Here $GF(h, n) = \exp(h(\log(n/h) + 1))$ and c_1, c_2, c_3 are given in [31].

B Proofs of results in §4

Proof of Lemma 4.6. The VC dimension of a linear classifier $f : \mathbb{R}^d \to \{0, 1\}$ is d (cf. Vapnik [49]). Real valued predictions have an extra degree of freedom. □

Proof of Theorem 4.7 and Corollary 4.8. By Theorem 4.4 and Lemma A.3, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_n(f)}{R_n(f)}>\epsilon\right) \tag{73}$$

$$\leq 8\exp\left\{(\widehat{h}+\delta)\left(\ln\frac{2\mu}{(\widehat{h}+\delta)}+1\right)-\frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\}(1-\varphi)+2(\mu-1)\beta_{a-d}(1-\varphi)+\varphi. \tag{74}$$

Therefore, solving the equation

$$\eta'=8\exp\left\{(\widehat{h}+\delta)\left(\ln\frac{2\mu}{\widehat{h}+\delta}+1\right)-\frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\}(1-\varphi) \tag{75}$$

implies

$$\epsilon=\frac{2M\tau(q)}{\sqrt{\mu}}\sqrt{(\widehat{h}+\delta)\left(\ln\frac{2\mu}{\widehat{h}+\delta}+1\right)-\ln\frac{\eta'}{8(1-\varphi)}}=\mathcal{E}. \tag{76}$$

The remainder follows analogously to the proof of Theorem 4.4. □

Proof of Theorem 4.9 and Corollary 4.10. Let $\mathcal{F}$ be indexed by the parameters of the state-space model above such that it creates predictions via the Kalman filter. Let $\mathcal{F}'$ be the same class of models, but predictions are made based on a filter sample truncated to have memory d. Then, for any $f\in\mathcal{F}$, and $f'\in\mathcal{F}'$

$$R_n(f)-\widetilde{R}_n(f)\leq(R_n(f)-R_n(f'))+(R_n(f')-\widehat{R}_n(f'))+(\widehat{R}_n(f')-\widetilde{R}_n(f)). \tag{77}$$

We 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

$$\mathbf{B}'=\mathbf{B}-\widetilde{\mathbf{B}} \tag{78}$$

with

$$\widetilde{\mathbf{B}}=\begin{bmatrix}0&\dots&0&b_{n,n-d-1}&b_{n,n-d-2}&\dots&b_{n,2}&b_{n,1}\\0&\dots&0&0&b_{n-1,n-d-2}&\dots&b_{n-1,2}&b_{n-1,1}\\&&&&\ddots&&&\vdots\\0&&&&\dots&&0&b_{d+1,1}\\0&&&&\dots&&0&0\end{bmatrix}. \tag{79}$$

Then notice that we can write

$$\widehat{R}_n(f')-\widetilde{R}_n(f)\leq|\widehat{R}_n(f')-\widetilde{R}_n(f)| \tag{80}$$

$$=\left|\frac{1}{n-d-1}\sum_{i=d}^{n-1}\left|\left|Y_{i+1}-\mathbf{b}_i\mathbf{Y}^i_{i-d+1}\right|\right|-\frac{1}{n-d-1}\sum_{i=d}^{n-1}\left|\left|Y_{i+1}-\mathbf{b}'_i\mathbf{Y}^i_{i-d+1}\right|\right|\right| \tag{81}$$

$$\leq\frac{1}{n-d-1}\sum_{i=d}^{n-1}\left|\left|(\mathbf{b}_i-\mathbf{b}'_i)\mathbf{Y}^i_{i-d+1}\right|\right| \tag{82}$$

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

$$\widehat{R}_n(f') - \widetilde{R}_n(f) \leq \frac{1}{n-d-1} \sum_{i=d}^{n-1} \left|\left| \widetilde{\mathbf{b}}_i \mathbf{Y}_{i-d+1}^i \right|\right| = \frac{1}{n-d-1} \sum_{i=d}^{n-1} \left|\left| \sum_{j=1}^{i-d} b_{i,j} y_j \right|\right| \tag{83}$$

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

$$R_n(f) - R_n(f') \leq \mathbb{E}\big[\, ||Y_1|| \,\big] \left|\left| \sum_{j=1}^{n-d-1} b_{n,j} \right|\right|. \tag{84}$$

Then,

$$R_n(f) - \widetilde{R}_n(f) - \Delta_d(f) \leq R_n(f') - \widehat{R}_n(f') \tag{85}$$

where

$$\Delta_d(f) = \mathbb{E}\big[\, ||Y_1|| \,\big] \left|\left| \sum_{j=1}^{n-d-1} b_{n,j} \right|\right| + \frac{1}{n-d-1} \sum_{i=d}^{n-1} \left|\left| \sum_{j=1}^{i-d} b_{i,j} y_j \right|\right| \tag{86}$$

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

$$\sup_{f \in \mathcal{F}} \frac{R_n(f) - \widetilde{R}_n(f) - \Delta_d(f)}{R_n(f)} \leq \sup_{f' \in \mathcal{F}', f \in \mathcal{F}} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f)}. \tag{87}$$

Finally, we have

$$\sup_{f \in \mathcal{F},\ f' \in \mathcal{F}'} \frac{R_n(f')}{R_n(f)} \leq 1 \tag{88}$$

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

$$\sup_{f' \in \mathcal{F}', f \in \mathcal{F}} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f)} = \sup_{f' \in \mathcal{F}', f \in \mathcal{F}} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f')} \frac{R_n(f')}{R_n(f)} \tag{89}$$

$$\leq \sup_{f' \in \mathcal{F}'} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f')}. \tag{90}$$

Now,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{R_n(f) - \widetilde{R}_n(f) - \Delta_d(f)}{R_n(f)} > \epsilon \right) \leq \mathbb{P}\left(\sup_{f' \in \mathcal{F}'} \frac{R_n(f') - \widehat{R}_n(f')}{R_n(f')} > \epsilon \right). \tag{91}$$

Since $\mathcal{F}'$ is a class with finite memory, we can apply Theorem 4.4 and Corollary 4.5 to get the results. □

Proof of Corollary 4.11. This follows immediately from Corollary 4.10 and (59). □

References

[1] Adams, T., and Nobel, A. (2010), "Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling," *The Annals of Probability*, **38**(4), 1345–1367.

[2] Athanasopoulos, G., and Vahid, F. (2008), "VARMA versus VAR for macroeconomic forecasting," *Journal of Business and Economic Statistics*, **26**(2), 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," in *Advances in Neural Information Processing Systems*, vol. 13, pp. 196–202, Cambridge, MA, MIT Press.

[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, arXiv:math/0511078 [math.PR].

[7] Carrasco, M., and Chen, X. (2002), "Mixing and moment properties of various GARCH and stochastic volatility models," *Econometric Theory*, **18**(01), 17–39.

[8] Cesa-Bianchi, N., and Lugosi, G. (2006), *Prediction, learning, and games*, Cambridge Univ Press, Cambridge, UK.

[9] 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.

[10] DeJong, D., and Dave, C. (2011), *Structural macroeconometrics*, Princeton Univ Press, Princeton, 2 edn.

[11] 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.

[12] Del Negro, M., and Schorfheide, F. (2006), "How good is what you've got? DSGE-VAR as a toolkit for evaluating DSGE models," *Federal Reserve Bank of Atlanta Economic Review*, **91**(2), 21–37.

[13] 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.

[14] Doukhan, P. (1994), *Mixing: Properties and Examples*, Springer Verlag, New York.

[15] Durbin, J., and Koopman, S. (2001), *Time Series Analysis by State Space Methods*, Oxford Univ Press, Oxford.

[16] 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.

[17] 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.

[18] Fernández-Villaverde, J. (2009), "The econometrics of DSGE models," Tech. rep., NBER Working Paper Series.

[19] 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.

[20] Gertler, M., and Karadi, P. (2011), "A model of unconventional monetary policy," *Journal of Monetary Economics*, **58**, 17–34.

[21] 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.

[22] Harvey, A., Ruiz, E., and Shephard, N. (1994), "Multivariate stochastic variance models," *The Review of Economic Studies*, **61**(2), 247–264.

[23] Hoeffding, W. (1963), "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, **58**(301), 13–30.

[24] Johnson, N., Kotz, S., and Balakrishnan, N. (1994), *Continuous univariate distributions*, vol. 2, John Wiley & Sons.

[25] Juselius, K., and Franchi, M. (2007), "Taking a DSGE model to the data meaningfully," *Economics*, **1**(2007-4), http://www.economics-ejournal.org/economics/journalarticles/2007-4.

[26] Kalman, R. E. (1960), "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, **82**(1), 35–45.

[27] Kearns, M., and Ron, D. (1999), "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," *Neural Computation*, **11**(6), 1427–1453.

[28] 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.

[29] Lahiri, S. N. (1999), "Theoretical comparisons of block bootstrap methods," *Annals of Statistics*, **27**(1), 386–404.

[30] Massart, P. (2007), "Concentration inequalities and model selection," in *Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*, Springer.

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

[32] 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 [stat.ML].

[33] McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011c), "Generalization error bounds for stationary autoregressive models," arXiv:1103.0942 [stat.ML].

[34] Meir, R. (2000), "Nonparametric time series prediction through adaptive model selection," *Machine Learning*, **39**(1), 5–34.

[35] 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.

[36] Mohri, M., and Rostamizadeh, A. (2010), "Stability bounds for stationary φ-mixing and β-mixing processes," *Journal of Machine Learning Research*, **11**, 789–814.

[37] Mokkadem, A. (1988), "Mixing properties of ARMA processes," *Stochastic Processes and their Applications*, **29**(2), 309–315.

[38] 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.

[39] Pollard, D. (1984), *Convergence of stochastic processes*, Springer Verlag, New York.

[40] Pollard, D. (1990), *Empirical processes: Theory and applications*, Institute of Mathematical Statistics.

[41] Racine, J. (2000), "Consistent cross-validatory model-selection for dependent data: HV-block cross-validation," *Journal of Econometrics*, **99**(1), 39–61.

[42] Rakhlin, A., Sridharan, K., and Tewari, A. (2010), "Online learning: Random averages, combinatorial parameters, and learnability," arXiv:1006.1138 [cs.LG].

[43] Shalizi, C. R. (2009), "Dynamics of Bayesian updating with dependent data and misspecified models," *Electronic Journal of Statistics*, **3**, 1039–1074, arXiv:0901.1342.

[44] Shalizi, C. R., Jacobs, A. Z., Klinkner, K. L., and Clauset, A. (2011), "Adapting to non-stationarity with growing expert ensembles," arXiv:1103.0949.

[45] Shumway, R., and Stoffer, D. (2000), *Time Series Analysis and Its Applications*, Springer Verlag, New York.

[46] Smets, F., and Wouters, R. (2007), "Shocks and frictions in US business cycles: A Bayesian DSGE approach," *American Economic Review*, **97**(3), 586–606.

[47] Tsybakov, A. (2009), *Introduction to nonparametric estimation*, Springer Verlag.

[48] Vapnik, V. (1998), *Statistical learning theory*, John Wiley & Sons, Inc., New York.

[49] Vapnik, V. (2000), *The Nature of Statistical Learning Theory*, Springer Verlag, New York, 2nd edn.

[50] 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.

[51] Vapnik, V., Levin, E., and Cun, Y. L. (1994), "Measuring the VC-dimension of a learning machine," *Neural Computation*, **6**(5), 851–876.

[52] Yu, B. (1994), "Rates of convergence for empirical processes of stationary mixing sequences," *The Annals of Probability*, **22**(1), 94–116.