

Dynamic Estimation of Time-Evolving Predictor-Response Joint Densities

Dave Zes

Department of Statistics

UCLA

2010-04-22

ABSTRACT. In the following we unite non-parametric curve fitting with dynamic estimation as a way of empirically generating densities from data with potentially unusual and time-varying structure. We include two examples to illustrate the construction and optimization of the joint density of a scalar predictor and scalar response over different cost functions.

1 Introduction

Consider a temporal scalar response y_t , and some predictor ξ_t . This predictor is special in the sense that, for the sake of this paper, we place no restrictions on the nature of its relationship to y . It may be anything from a best L^2 estimator gleaned from a collection regressors, to the best guess of an expert, to some perhaps highly non-linear mathematical aggregation of “information.” We seek to describe the nature of the relationship between ξ_t and y through their potentially time-evolving joint density. This function, describing the intersection of the predictor and realized value, is a rather hallowed reckoning for analysts, since from within can be extracted the full panoply of statistics analysts enjoy citing when predicting some outcome. From the joint density comes the conditionals from which confidence intervals, p -values, quartiles, and the like can be extracted.

The application of central interest is one where we have some historical values of (ξ_t, y_t) , with the future expectation of more incoming data. We hope to learn a model from the data in hand, and, perhaps excepting occasional tuning, use this model to process upcoming ξ_t and return an estimate of y_t — or more precisely and generally, make some probabilistic statement about some attribute of interest about y_t .

In order to estimate the joint density, which we’ll call $\rho(\xi, y)$, we may draw on the rich literature of curve fitting, involving bases transformations, such as B-splines, cosine bases (the “universal” function estimator), and wavelets. It may be natural to desire that their joint density be allowed to

Symbol	Definition
y	variable: a scalar value of interest
ξ	variable: any predictor of y
ϕ	function: bases transformation. Acts on single ξ , returns a vector
ψ	function: bases transformation. Acts on single y , returns a vector
\otimes	operator: outer dot product. Acts on two vectors, returns a matrix
$*$	operator: elementwise multiplication. Acts on two vectors or matrices
$\{x_{1:t}\}$	collection. The collection of all x 's indexed 1 through t inclusively
i, j	indices over ϕ and ψ respectively

Table 1: Notational Symbols

adapt over time — $\rho_t(y_t, \xi_t)$. To provide for this we model the bases coefficients as *states*, estimated by “common” solutions.

We commence with a very soft review of both dynamic estimation, and non-parametric density estimation.

1.1 The Local-Level Model

Consider a scalar time-varying hidden parameter (“state”), θ_t , wandering unconstrained, whose value can only be observed through noise. Specifically:

$$\theta_t = 1 \cdot \theta_{t-1} + \varepsilon_t \tag{1}$$

$$\theta_t^{\text{obs}} = \theta_t + \eta_t \tag{2}$$

with $\varepsilon_t \sim \mathcal{N}(0, Q_t)$ and $\eta_t \sim \mathcal{N}(0, R_t)$ all independent. This system is known as the “local level model.” Notice the explicit unity coefficient in (1). The parameter, θ_t is undergoing a random walk, and as such, the process is non-stationary, i.e., the marginal distribution of *theta* has variance that is ever increasing with time. Implicit in this is the notion that the global expectation is simply the initial value, θ_0 . Should the absolute value of this coefficient be less than one, then the state tends to zero — its values throughout all time are constrained — the process is stationary. Moreover, in such a case, regardless of the initial value of θ_0 , at some point in time the unconditional expectation of the state will be arbitrarily close to zero.

A by now well travelled result is that the ideal (in the least mean squared sense) predictor for θ_t

given the full collection of observations up to and including time t , $\{\theta_{1:t}^{\text{obs}}\}$, is the Kalman solution:

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{P_{t-1} + Q_t}{P_{t-1} + Q_t + R_t} \left(\theta_t^{\text{obs}} - \hat{\theta}_{t-1} \right) \quad (3)$$

$$P_t = \frac{(P_{t-1} + Q_t)R_t}{P_{t-1} + Q_t + R_t} \quad (4)$$

In essence, the solution provided in (3) & (4) offers a computationally inexpensive and altogether simple way of tracking an unobservable migrating value, θ . If that value happens to conform exactly to the system described in (1) & (2) then the solution produces a minimum variance unbiased estimate of θ ; it is the conditional posterior expectation of θ . Adherence to this system is far less restrictive than may first seem. Since the shocks on θ , realized through ε_t , along with the measurement contamination, realized through η_t , are often in reality a sum aggregation of small shocks, the CLT affirms the Gaussian assumption. An important continuation of this thinking, but one that will not be further explored here, is that even when truth is crudely describe by our system, the Kalman solution will often still provide a satisfactory estimate. That is to say, in many statistical applications, the Kalman solution will give results indistinguishable from a theoretically ideal solution.

1.2 Non-Parametric Density Estimation

In what follows, we will process only a single datum tuple, (ξ_t, y_t) , at a time, and so the expressions for the bases transformations will be slightly different than what one usually sees in discussions of function estimation.

$$\Theta_t^{\text{obs}} = \phi(\xi_t) \otimes \psi(y_t) \quad (5)$$

Bold capital *theta*, Θ_t^{obs} , is a matrix of coefficients, or weights, that represent the amplitude of the datum tuple, (ξ_t, y_t) under the transformations ϕ and ψ . If, for example, ϕ and ψ are chosen to be orthogonal cosine transformations, then Θ_t^{obs} may be thought of as the spectral information of $\rho_t(\xi_t, y_t)$. Generally by pooling together our observations, Θ^{obs} ,

$$\hat{\Theta}_t = f \left(\left\{ \Theta_{1:t}^{\text{obs}} \right\} \right) \quad (6)$$

we may then use $\hat{\Theta}_t$ to construct an estimate of our density.

$$\hat{\rho}_t(\xi, y) = \sum_{i,j} \hat{\Theta}_t * \left(\phi(\xi) \otimes \psi(y) \right) \quad (7)$$

In words, the height of the density for an arbitrary tuple, (ξ, y) , at time t , is the grand sum over all the elements of the matrix resulting from the elementwise product between the weight estimate,

$\widehat{\Theta}_t$ and the outer dot product of the bases transformations, ϕ , and ψ .

1.2.1 Support

For simplicity, curve estimation usually has that the function support lives in the unit square (or cube, etc). Same for us here; we'll have that $(\xi, y) \in [0, 1]^2$. This saves us the extra notation involved with defining (ξ, y) only as bounded, and having to explicitly write the function that maps them to the unit square every time we refer to them in context with their density.

1.2.2 Volume

In some cases, such as cosine transformation and wavelets, by construction, the density estimate will integrate to 1: $\int_0^1 \int_0^1 \widehat{\rho}_t(\xi, y) d\xi dy = 1$.¹ This is not true of bases transformations in general. B-splines, preferred in the following examples, produce essentially arbitrary positive volumes. There exist two popular ways to construct B-splines, each offering a different way of handling points near the boundaries. One compresses, or narrows, the edge-most splines, the other constructs the bases with splines beyond the support. The examples here use the latter. Computing density volume is trivial, so long as care is taken to calculate the contribution of these partially out-of-bounds splines.

1.2.3 Conditional Area

Since one of our main goals is examination of conditional densities, it will be necessary to integrate across cross-sections of the joint distribution. As above, since bases functions are usually linear and simple, integration does not pose any problems. In the case of the cosine transform, for example, normalizing conditionals is trivial; in the case of B-Splines, the process is just slightly more elaborate owing to their peicewise nature. If for some reason we wish not to labor over the analytics, we can always turn to Simpson's Approximation as a fast, easy, accurate way of normalizing these conditionals.

¹Actually, this is made true by the "scaling function," which integrates to 1. The frequency transforms, including mother wavelets, integrate to zero.

2 Time-Adaptive Non-Parametric Density Estimation

Now the two ideas are merged in a way that should not surprise the reader. We use recursion to propagate the spectral attributes of ρ_t in a time-adaptive way.

$$\widehat{\Theta}_t = \widehat{\Theta}_{t-1} + g_t \left(\Theta_t^{\text{obs}} - \widehat{\Theta}_{t-1} \right) \quad (8)$$

For example, the function vaguely referred to in (6) becomes the Kalman update recursion.

At this point let’s take a moment and visit some terminology. Looking back at (3), although it’s not obvious from the setup, often in statistical applications θ_t will play the role of a parameter, that is, a term in a parametric model. Moreover, values such as Q_t that are called upon to estimate θ_t are not known and must be induced from data, by maximum likelihood, for example. For this reason, these secondary values are commonly referred to as “hyperparameters.” In our current application, though, such naming would suggest that Θ_t is itself a parameter. This has awkward connotations. As we will see shortly, there exist very spartan ways to estimate the update gain (that is, g_t in (8)). So, the process of inducing g_t will be the filtering equivalent to the non-parametric process of inducing good bases. So, dynamic density estimation will be regarded as non-parametric.

On another matter, not of terminology but rather of concept, let’s reflect for a moment on a key distinction between the parametric and non-parametric paradigms. Imagine a bivariate parametric setting. If we believe at the outset that the relationship is linear, that is, $y = \theta_1 x + \theta_0 + \varepsilon$, and a scatterplot confirms this belief, then the results of the regression have the following interpretation: Nature has within its design that x and y are linearly related, and also provides some complex of disturbances, or shocks on y that cause (often because of the CLT, Gaussian) departures from the underlying linear relation. This type of reasoning does not carry over well to the non-parametric setting. It would be hard to argue that nature happened to relate x and y by some particular collection of transformation functions that the analyst either knew ahead of time or found exactly, and the only thing left to do is estimate the bases coefficients. Rather, the argument revolves around the space of functions in which truth lives, usually described by smoothness, and using knowledge of the phenomenon and the data alone to *approximately select* this function from within the correct family of candidates.

With pure description in mind, since we are treating the predictor-response density, $\rho(\xi, y)$, as some function of time, we might create a model mapping *rho* over 3-space, i.e., have $(\xi, y, t) \in [0, 1]^3$. This sort of approach is valuable for revealing systematic effects of t on ρ . It may not help with local effects. The way we have posited our solution in (8) implies that the underlying process involves a purely stochastic wandering of bases amplitudes — it is the simplest embodiment one can imagine. We should keep in mind that the process here can be generalized easily to account for systematic

change over time, but that since we wish to use our model inferentially for future incoming data, we must have some way of interpreting local effects of time on *rho* that can be meaningfully extended. This is not a statistical idea. Early in life we become aware of the unidirectional nature of time. An unanticipated event may pass too quickly for us to react; we are left only with the consolation that perhaps we have learned sufficiently to better react to similar surprises surely awaiting us in the future.

Finally, we need to address a dichotomy lurking in our reasoning. If we have g_t be the Kalman gain, then (8) provides a rigorous solution to a system that we admit almost certainly does not exist. In the spirit of non-parametric fitting, we construct solutions that we hope approximate truth sufficiently well that we might in some way profit from them. This might be dissatisfying to purists, who seek to fully define the system, then mount a campaign to unveil its solution. While we should never lose this sort of analytic idealism, we should not forget the value of reasoned approximation. Examples 2 and 3 that follow will hopefully reinforce our pragmatism.

2.1 Finding g

Our update gain, g_t in (8), has a visceral interpretation as the reciprocal of the “effective sample size” of the collection of observations at time t . For example, if $g_t = 1/20$, then the observation at time t is making a contribution of $1/20^{th}$ towards our estimate; we are suggesting that the current observation is just one in a homogeneous sample of 20. If we introduce the signal-to-noise ratio as $\nu_t = Q_t/R_t$, we may rewrite the scalar Kalman gain recursion as

$$g_t = \frac{g_{t-1} + \nu_t}{g_{t-1} + \nu_t + 1} \tag{9}$$

If ν_t is time invariant, we may fully construct our time-varying model *with only two scalar values*: g_0 and ν . When $g_0 = 1$, the first observation stomps out Θ_0 . However, if we have a prior belief about Θ_0 , we can adjust g_0 in a meaningful way so to include and propagate this initial belief.

2.2 Implementation

All the techniques here could be applied to an online experiment. However, it’s hard to imagine, at least at this time, a scenario where learning could not be done offline, at the leisure of the researcher.

2.3 Model Fitting

In our setting, “model fitting” refers to the process of selecting some $\{\phi, \psi, \{\nu_{1:N}\}\}$ that produces an estimator, $\hat{\rho}$, of acceptable quality. In filtering, the canonical method for determining the quality of the estimator is the measurement *a priori* error, $e_t^{(-)}$. Of course in engineering, objectives such as signal processing, for example, are model driven, not data driven, and so checking this error is usually done for the sake of comparing practice with theory. In statistics, the gold standard for *static* non-parametric model testing is Cross Validation (CV) in one of its many forms, such as 10-fold or leave-one-out. Somewhat fascinatingly, it turns out that in a dynamic context, CV translates into measurement *a priori* error — it’s one-step-ahead prediction error. However, a literal interpretation of the *a priori* error will fail us here. Consider that we are *estimating functions*, but the “observed” function at time t is constructed from Θ_t^{obs} , which has been created from only a single datum tuple. As one can easily imagine, if our bases have many splines, or contain high frequencies, this function will be a spike, whereas the estimate of this function up to $t - 1$ might be a smooth surface.

Naturally, how we tune our model will depend on the cost function we ultimately wish to associate with our response, y . While it is not required that we do so explicitly, tuning will involve using the cost function as a proxy for the actual response of interest. The three examples that follow illustrate optimized fitting across different *a priori* error cost functions.

3 Examples

We explore two examples. For each is provided a movie that can be reached by hyperlink showing the evolution of $\hat{\rho}$. The movies are 900 pixels tall, so it is important to have your browser page large enough to access the movie player controls. Quicktime is required to view these .mov files. If you can’t view the files and don’t want to install the Quicktime plugin, please contact me.

Each succeeding example employs an underlying model — the actual creation of ξ_t — that is more involved than the one before it. For obvious reasons of space, we omit all details of this underlying estimation.

3.1 Ozone in California

We examine about 9,500 daily 1 hour maximum ozone levels as recorded by a monitoring station near San Francisco Bay. Our predictor here, ξ , was constructed in a rather pedestrian way, as a

cyclic (sin + cos) 2nd-order autoregressive series using an empirically hyper-parameterized Kalman Filter.²

We have that $(\xi, y) \in \mathbb{R}^2$, and for the sake of example we'll have ϕ and ψ be filters of evenly spaced B-Splines, in numbers of b_ϕ and b_ψ respectively. Moreover, let's have $\nu_t = \nu, \forall t$.

Our test for the quality of our dynamic density will be its ability to accurately predict days of high ozone levels. We choose a cutoff, ω , and define our “*a priori* error” as

$$e_t^{(-)} = -1_{\{y_t > \omega\}} \cdot \log \left\{ \widehat{\text{Pr}}[Y_t > \omega \mid \xi_t] \right\} - 1_{\{y_t < \omega\}} \cdot \log \left\{ \widehat{\text{Pr}}[Y_t < \omega \mid \xi_t] \right\} \quad (10)$$

where $1_{\{\cdot\}}$ is the indicator function, $\widehat{\text{Pr}}[Y_t > \omega_t \mid \xi_t]$ is found through our estimated conditional density $\widehat{\rho}(y \mid \xi_t)$. Our job of model tuning becomes finding

$$\underset{\{b_\phi, b_\psi, \nu\}}{\text{argmin}} \sum_t e_t^{(-)} \quad (11)$$

Searching over 3-space is easy. Figures **1** & **2** show risk (error) versus ν and b_ϕ respectively. A movie showing the optimized dynamically evolving density can be found at

http://www.stat.ucla.edu/~davezes/DDE/03_Full.mov

The origin is the distant corner. The predictor, ξ , runs along the right front edge, y along the left front. This movie plays through about 9,500 data points, then, for illustration, rotates the surface, then finally passes through the joint density along the predictor, i.e., we view $\widehat{\rho}_n(y \mid \xi)$ for all values of ξ .

3.2 Life Settlement

We lastly turn to a relatively new type of insurance valuation known as “life settlement.” Life settlement is concerned with modeling the current value of an existing individual life insurance policy. Model construction here might rightfully start by employing predictions generated by actuarial life expectancy tables. These predictive models involve hundreds or even sometimes *thousands* of predictors. The following example uses 5,000 data generated from an actuarial model. Our predictor, ξ , represents predicted remaining life expectancy, in *years*, of a policy holder. The response, y , represents actual years to mortality.

Just like with the Ozone Example, $(\xi, y) \in \mathbb{R}^2$, and we have ϕ and ψ be filters of evenly spaced B-Splines, in numbers of b_ϕ and b_ψ respectively.

²Actually, this was something of a misapplication as the residuals of the fit were not quite Gaussian.

One interested in assigning value to existing policies will be interested in the L^2 monetary risk. Our loss function is hence

$$h(z) = c(1+r)^z + m \frac{(1+r)^z - 1}{r} \quad (12)$$

where z is time in units months, c is the buyout cost at z_0 , m is the monthly premium, and r is the monthly rate of inflation. In words, the cost of a policy is the sum of the compounding value of the buy-out price (1st term on RHS) and a negative annuity on the monthly premiums (2nd term on RHS).

Our *a priori* measurement error then becomes

$$e_t^{(-)} = \mathbb{E}_{\hat{\rho}_{t-1}(12y|\xi_t)} \left[h(12y) - h(12y_t) \right]^2 \quad (13)$$

$$= \left(\int_y h(12y) \cdot \hat{\rho}_{t-1}(12y|\xi_t) \cdot dy - h(12y_t) \right)^2 \quad (14)$$

We hence seek to find

$$\operatorname{argmin}_{\{b_\phi, b_\psi, \nu\}} \sum_t e_t^{(-)} \quad (15)$$

Figures 5 & 6 show risk versus ν and b_ϕ respectively. A movie showing the optimized dynamic evolving density can be found at

http://www.stat.ucla.edu/~davezes/DDE/LE_cost_Full.mov

Just like with the previous Examples, the origin is the distant corner. The predictor, ξ , runs along the right front edge, y along the left front. This movie plays through about 5,000 data points, then, for illustration, rotates the surface, then finally passes through the joint density along the predictor, i.e., we view $\hat{\rho}_n(y|\xi)$ for all values of ξ .

In this example, our method offers a simple way to extend information. Life Expectancy information is available to small investment companies, who could not, on the other hand, afford the enormous man-hours of high-skill analysts required to attempt a ground-up theoretical model to value policies.

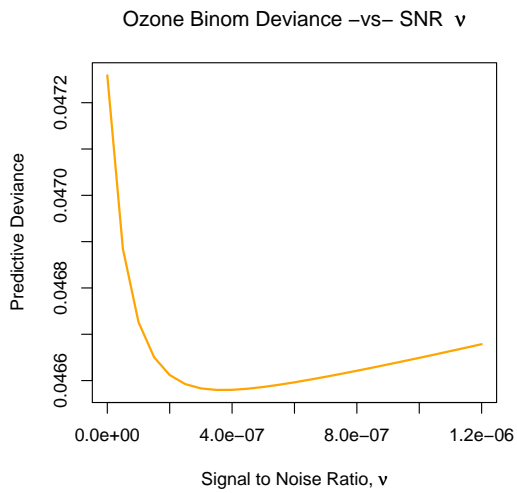


Figure 1

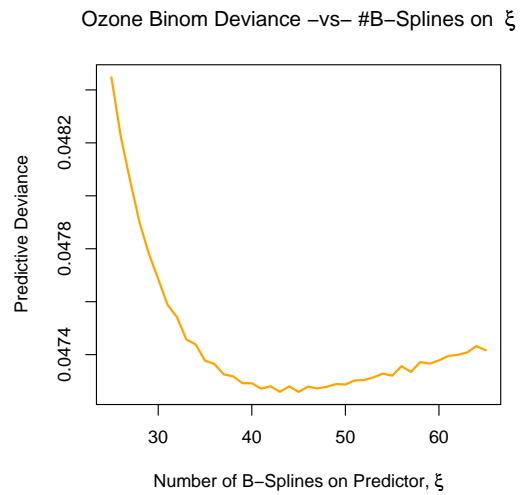


Figure 2

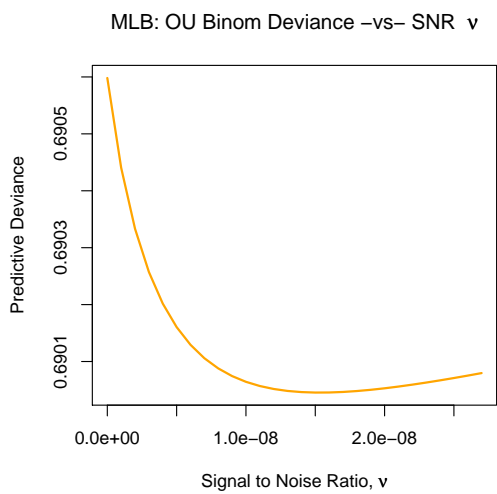


Figure 3

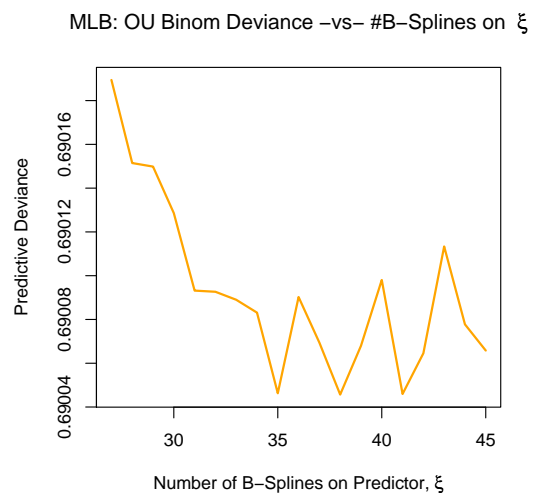


Figure 4

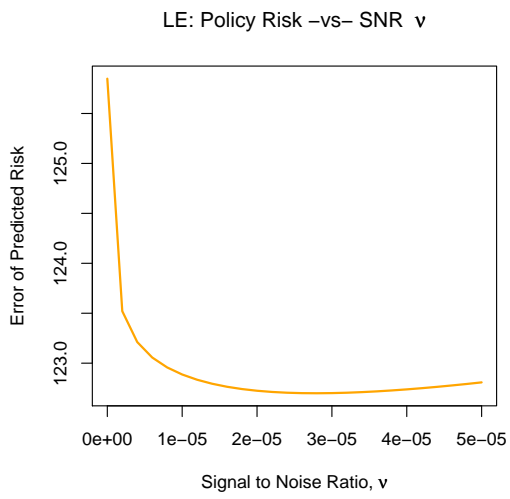


Figure 5

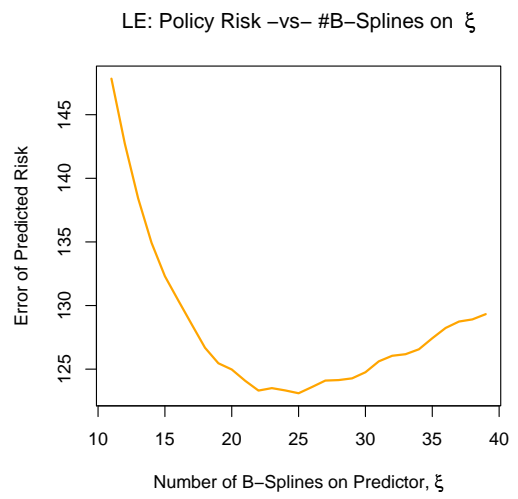


Figure 6

References

- [1] Brian D.O. Anderson and John B. Moore, *Optimal Filtering*. Dover, Mineola, NY, 1979, republished in 2005.
- [2] Robert W. Bass, *Empirical System Identification (ESID)*. Personal communication, 2008.
- [3] R.S. Bucy, *Lectures on Discrete Time Filtering*. Springer, New York, 1994.
- [4] Arthur Gelb & the technical staff of the Analytic Sciences Corporation, *Applied Optimal Estimation*. MIT Press, 1978.
- [5] S. Gunnarsson, “Combining Tracking and Regularization in Recursive Least Squares Identification.” Proceedings of the 35th Conference on Decision and Control. Kobe, Japan. December 1996.
- [6] S. Gunnarsson, “On Covariance Modification and Regularization in Recursive Least Squares Identification.” In 10th IFAC Symposium on System Identification — SYSID’94, pages 661–666, Volume 2. Copenhagen, Denmark, July 1994.
- [7] Simon Haykin, *Adaptive Filter Theory*. Forth Edition. Prentice-Hall, Upper Saddle River, 2002.
- [8] Chi Sing Leung et. al, “On the Regularization of Forgetting Recursive Least Square.” IEEE Transactions on Neural Networks, Volume 10, Number 6. November, 1999.
- [9] J. Huston McCulloch, “The Kalman Foundations of Adaptive Least Squares, With Application to U.S. Inflation.” <www>, August, 2005.
- [10] Irwin Miller and Marylees Miller, *John E. Freund’s Mathematical Statistics*. Prentice-Hall, Upper Saddle River, 1999.
- [11] Donald B. Percival and Andrew T. Walden, *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge, 2007.
- [12] Ali H. Sayed, *Fundamentals of Adaptive Filtering*. John Wiley & Sons, Hoboken, N.J., 2003.
- [13] Robert H. Shumway and David S. Stoffer, *Time Series Analysis and Its Applications, With R Examples*. Second Edition. Springer, N.Y., 2006.
- [14] Larry Wasserman, *All of Nonparametric Statistics*. Springer, New York, 2006.
- [15] Mike West and Jeff Harrison, *Bayesian Forecasting and Dynamic Models*. Second Edition. Springer-Verlag, New York, 1997.