Synthetic Volatility Forecasting and Other Aggregation Techniques for Time Series Forecasting

Preliminary Exam

David Lundquist[1]

March 4, 2024

[1] davidl11@ilinois.edu

A seemingly unprecedented event might provoke the questions

1. What does it resemble from the past?
2. What past events are most relevant?
3. Can we incorporate past events in a systematic, principled manner?

When would we ever have to do this?

- Event-driven investing strategies (unscheduled news shock)
- Structural shock to macroeconomic conditions (scheduled news possibly pre-empted by news shock)
- Biomedical panel data subject to exogenous shock or interference

Example: weekend of March 7th and 8th, 2020

Punchline of the paper

Forecasting is possible under structural shocks, so long as we incorporate external information to account for the nonzero errors.

What gaps does this fill in the literature?

Volatility Modeling

- GARCH is slow to react andersen2003modeling
- Asymmetric GARCH models may react faster but need post-shock data
- Realized GARCH hansen2012realized, in our setting, would require high-frequency data, and the model is highly parameterized

Forecast Augmentation

- A clements1996intercept,clements1998forecasting laid the groundwork for modeling nonzero errors in time series forecasting
- A guerron2017macroeconomic use a series' own errors to correct the forecast for that series
- A dendramis2020similarity use a similarity-based procedure to correct linear parameters in time series forecasts
- A foroni2022forecasting adjust pandemic-era forecasts using intercept correction techniques and data from Great Financial Crisis

Outline

The news has broken but markets are closed

- $y \in \mathbb{R}^n$, a mean-zero, real-valued response to be predicted
- Mean-zero covariate vectors $x \in \mathbb{H}$, a Hilbert space, where $\mathbb{H}$ can be taken to be $\mathbb{R}^p$ for the sake of illustration
- We predict y using a linear function of the covariates X, but this is *not* an assumption about the data generating process. We do not assume $y = f(X) + \epsilon$, but y may contain exogenous noise (called "label noise" by the authors.)

A Primer on GARCH

For this audience, we skip ahead to the most general model

$$
\begin{aligned}
& \sigma^2_{i,t} = \omega_i + \omega^*_i + \textstyle\sum_{k=1}^{m_i} \alpha_{i,k} a^2_{i,t-k} + \sum_{j=1}^{s_i} \beta_{i,j} \sigma^2_{i,t-j} + \gamma_i^T{}_{i,t} \\
M_2: \quad & a_{i,t} = \sigma_{i,t}(\epsilon_{i,t}(1 - D^{level}_{i,t}) + \epsilon^*_i D^{level}_{i,t}) \\
& \omega^*_{i,t} = \mu_{\omega^*} + \delta' x_{i,t-1} + \varepsilon_i
\end{aligned}
$$

Technical Specifications (many, but familiar)

- We define $\Sigma := \mathbb{E}[xx^T]$
- We define $\theta^* \in \mathbb{H}$ to be s.t. $\mathbb{E}[(y - x^T\theta^*)^2] = \min_\theta \mathbb{E}[(y - x^T\theta)^2]$
  - As defined here, not necessarily unique, which matters for this paper.
- $x = V\Lambda^{1/2}z$, where $\Sigma = V\Lambda V^T$ is the spectral decomposition of Σ and z is a vector of independent components, each subgaussian(σ_x), where $\sigma_x > 0$.
- The conditional noise variance $\mathbb{E}[(y - x^T\theta^*)^2|x]$ is bounded below by $\sigma^2 > 0$.
- $(y - x^T\theta^*)|x$ is subgaussian(σ_y)
- Almost surely, for any eigenvector v of Σ, $Proj_{v^T}(X)$ spans a space of dimension n.
  - Guaranteed when, for example, there exist p linearly independent covariates and $p > n$.

Volatilty Profile

In this particular setting, excess risk of an estimator θ has the form

$$\begin{aligned} R(\theta) &= \mathbb{E}_{x,y}[(y - x^T\theta)^2 - (y - x^T\theta^*)^2] \\ &= (\theta - \theta^*)^T\Sigma(\theta - \theta^*) \end{aligned}$$

What's the method here?

$$\min_{\theta \in \mathbb{H}} \|\theta\|$$
$$s.t. \frac{1}{n}\|X^T\theta - y\|_2^2 \le D$$

What's the method here?

$$\min_{\theta \in \mathbb{H}} \|\theta\|$$
$$s.t. \frac{1}{n}\|X^T\theta - y\|_2^2 \leq D$$

What's the method here?

$$\min_{\theta \in \mathbb{H}} \|\theta\|$$
$$s.t. \frac{1}{n}\|X^T\theta - y\|_2^2 \leq D$$

1. Heuristically speaking, overfitting means $D << \min_{\theta}\mathbb{E}(x^T\theta - y)^2$, where x, y are out of sample

What's the method here?

$$\min_{\theta \in \mathbb{H}} \|\theta\|$$
$$s.t. \frac{1}{n} \|X^T\theta - y\|_2^2 \leq D$$

1. Heuristically speaking, overfitting means $D << \min_{\theta} \mathbb{E}(x^T\theta - y)^2$, where x, y are out of sample
2. Of course, we're concerned with interpolation, i.e. $D = 0$.

Minimum Norm Estimator

Why did we just assume something very technical about $Proj_{v^T}(X)$? This condition implies multiple solutions to the equation $y = X\theta$.

"Almost surely, for any eigenvector v of Σ, $Proj_{v^T}(X)$ spans a space of dimension n."

Since we have more than one choice of θ, we choose the unique $\hat{\theta}$ with minimum norm:

$$\hat{\theta} := (X^T X)^{\dagger} X^T y$$

We don't have to do this, but the results that follow correspond to the minimum-norm solution. It's most interesting.

Rank of Matrix

We know the rank of a matrix $A \in M_{n \times p}(\mathbb{C})$ is

- The column rank of A (number of linearly independent columns)
- The row rank of A (number of linearly independent rows)
- The dimension of $im(A)$

However, this notion of rank is too rigid. It's integer-valued, and it tells us very little about the distribution of the eigenvalues.

Key Conceptual Innovation: Effective Rank

Definition

For a covariance operator Σ with the decreasing sequence of eigenvalues $\lambda_1, \lambda_2, \ldots$, if $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$, then for $k \geq 0$, define

$$r_k = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \quad R_k = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}$$

Key Conceptual Innovation: Effective Rank

Definition

For a covariance operator Σ with the decreasing sequence of eigenvalues $\lambda_1, \lambda_2, \ldots$, if $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$, then for $k \geq 0$, define

$$r_k = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \quad R_k = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}$$

Key Properties

Key Conceptual Innovation: Effective Rank

Definition

For a covariance operator Σ with the decreasing sequence of eigenvalues $\lambda_1, \lambda_2, \ldots$, if $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$, then for $k \geq 0$, define

$$r_k = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \quad R_k = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}$$

Key Properties

1. $r_k \in [1, p-k]$, assuming $p < \infty$

Key Conceptual Innovation: Effective Rank

Definition

For a covariance operator Σ with the decreasing sequence of eigenvalues $\lambda_1, \lambda_2, \ldots$, if $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$, then for $k \geq 0$, define

$$r_k = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \qquad R_k = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}$$

Key Properties

1. $r_k \in [1, p-k]$, assuming $p < \infty$
2. $R_k \in [1, \infty)$

Key Conceptual Innovation: Effective Rank

Definition

For a covariance operator Σ with the decreasing sequence of eigenvalues $\lambda_1, \lambda_2, \ldots$, if $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$, then for $k \geq 0$, define

$$r_k = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \quad R_k = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}$$

Key Properties

1. $r_k \in [1, p-k]$, assuming $p < \infty$
2. $R_k \in [1, \infty)$
3. They can be understood in terms of ℓ_1 and ℓ_2 norms.

Key Conceptual Innovation: Effective Rank

Definition

For a covariance operator Σ with the decreasing sequence of eigenvalues $\lambda_1, \lambda_2, \ldots$, if $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$, then for $k \geq 0$, define

$$r_k = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \quad R_k = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}$$

Key Properties

1. $r_k \in [1, p-k]$, assuming $p < \infty$
2. $R_k \in [1, \infty)$
3. They can be understood in terms of ℓ_1 and ℓ_2 norms.
4. $r_k(\Sigma^2) \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma)$

Key Conceptual Innovation: Effective Rank

Definition

For a covariance operator Σ with the decreasing sequence of eigenvalues $\lambda_1, \lambda_2, \ldots$, if $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$, then for $k \geq 0$, define

$$r_k = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \quad R_k = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}$$

Key Properties

1. $r_k \in [1, p-k]$, assuming $p < \infty$
2. $R_k \in [1, \infty)$
3. They can be understood in terms of ℓ_1 and ℓ_2 norms.
4. $r_k(\Sigma^2) \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma)$
5. For the result we now show, *bigger* values of r_k and R_k are better.

Main Result: Existence Proof, Dichotomy, and Bounds

Theorem

Define $k^* = min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. *For any* σ_x, $\exists b, c, c_1 > 1$ *s.t.* $\forall \delta \in (0,1)$ *s.t.* $\log(1/\delta) < n/c$, *if* $k^* \geq n/c_1$, *then* $\mathbb{E}[R(\hat{\theta})] \geq \sigma^2/c$. *Otherwise,*

- A $R(\hat{\theta}) \leq \underbrace{c\left(\|\theta^*\|^2\|\Sigma\| max\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\}\right)}_{\text{related to the bias}} + \underbrace{c\log(1/\delta)\sigma_y^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)}_{\text{related to the noise}}$ *with probability at least* $1 - \delta$.
- B $\mathbb{E}[R(\hat{\theta})] \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$

Main Result: Existence Proof, Dichotomy, and Bounds

Theorem

Define $k^* = min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. *For any* σ_x, $\exists b, c, c_1 > 1$ *s.t.* $\forall \delta \in (0,1)$ *s.t.* $\log(1/\delta) < n/c$, *if* $k^* \geq n/c_1$, *then* $\mathbb{E}[R(\hat{\theta})] \geq \sigma^2/c$. *Otherwise,*

- **A** $R(\hat{\theta}) \leq \underbrace{c\left(\|\theta^*\|^2\|\Sigma\| max\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\}\right)}_{\text{related to the bias}} + \underbrace{c\log(1/\delta)\sigma_y^2(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)})}_{\text{related to the noise}}$

  with probability at least $1-\delta$.
- **B** $\mathbb{E}[R(\hat{\theta})] \geq \frac{\sigma^2}{c}(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)})$

Remarks

Main Result: Existence Proof, Dichotomy, and Bounds

Theorem

Define $k^* = min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. *For any* σ_x, $\exists b, c, c_1 > 1$ *s.t.* $\forall \delta \in (0,1)$ *s.t.* $\log(1/\delta) < n/c$, *if* $k^* \geq n/c_1$, *then* $\mathbb{E}[R(\hat{\theta})] \geq \sigma^2/c$. *Otherwise*,

1. $$R(\hat{\theta}) \leq \underbrace{c\left(\|\theta^*\|^2\|\Sigma\| max\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\}\right)}_{\text{related to the bias}} + \underbrace{c\log(1/\delta)\sigma_y^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)}_{\text{related to the noise}}$$
   with probability at least $1 - \delta$.
2. $\mathbb{E}[R(\hat{\theta})] \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$

Remarks

1. A gives us a (high probability) upper bound on the excess risk.

Main Result: Existence Proof, Dichotomy, and Bounds

Theorem

Define $k^* = min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. For any σ_x, $\exists b, c, c_1 > 1$ s.t. $\forall \delta \in (0,1)$ s.t. $\log(1/\delta) < n/c$, if $k^* \geq n/c_1$, then $\mathbb{E}[R(\hat{\theta})] \geq \sigma^2/c$. Otherwise,

- **A** $R(\hat{\theta}) \leq \underbrace{c\left(\|\theta^*\|^2\|\Sigma\| max\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\}\right)}_{\text{related to the bias}} + \underbrace{c\log(1/\delta)\sigma_y^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)}_{\text{related to the noise}}$

  with probability at least $1 - \delta$.
- **B** $\mathbb{E}[R(\hat{\theta})] \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$

Remarks

1. A gives us a (high probability) upper bound on the excess risk.
2. B gives us a lower bound on its expectation.

Main Result: Existence Proof, Dichotomy, and Bounds

Theorem

Define $k^* = min\{k \geq 0 : r_k(\Sigma) \geq bn\}$*. For any* σ_x, $\exists b, c, c_1 > 1$ *s.t.* $\forall \delta \in (0,1)$ *s.t.* $\log(1/\delta) < n/c$, *if* $k^* \geq n/c_1$, *then* $\mathbb{E}[R(\hat{\theta})] \geq \sigma^2/c$. *Otherwise,*

- A $$R(\hat{\theta}) \leq \underbrace{c\left(\|\theta^*\|^2\|\Sigma\| max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}\right)}_{\text{related to the bias}} + \underbrace{c\log(1/\delta)\sigma_y^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)}_{\text{related to the noise}}$$ *with probability at least* $1-\delta$.
- B $\mathbb{E}[R(\hat{\theta})] \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$

Remarks

1. A gives us a (high probability) upper bound on the excess risk.
2. B gives us a lower bound on its expectation.
3. The constants b, c, c_1 depend on σ_x, the subgaussian parameter corresponding to the covariates X.

Main Result: Existence Proof, Dichotomy, and Bounds

Theorem

Define $k^* = min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. For any σ_x, $\exists b, c, c_1 > 1$ s.t. $\forall\delta \in (0,1)$ s.t. $\log(1/\delta) < n/c$, if $k^* \geq n/c_1$, then $\mathbb{E}[R(\hat{\theta})] \geq \sigma^2/c$. Otherwise,

A. $$R(\hat{\theta}) \leq \underbrace{c\left(\|\theta^*\|^2\|\Sigma\| max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}\right)}_{\text{related to the bias}} + \underbrace{c\log(1/\delta)\sigma_y^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)}_{\text{related to the noise}}$$
   with probability at least $1 - \delta$.

B. $$\mathbb{E}[R(\hat{\theta})] \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$$

Remarks

1. A gives us a (high probability) upper bound on the excess risk.
2. B gives us a lower bound on its expectation.
3. The constants b, c, c_1 depend on σ_x, the subgaussian parameter corresponding to the covariates X.
4. Think of k^* as the number of dimensions we ignore when hiding the noise. We want k^* to be small compared to n, yet no smaller than it need be, obviously.

Main Result: Existence Proof, Dichotomy, and Bounds

Theorem

Define $k^* = min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. *For any* σ_x, $\exists b, c, c_1 > 1$ *s.t.* $\forall \delta \in (0, 1)$ *s.t.* $\log(1/\delta) < n/c$, *if* $k^* \geq n/c_1$, *then* $\mathbb{E}[R(\hat{\theta})] \geq \sigma^2/c$. *Otherwise,*

- Ⓐ $$R(\hat{\theta}) \leq \underbrace{c\left(\|\theta^*\|^2\|\Sigma\| max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}\right)}_{\text{related to the bias}} + \underbrace{c\log(1/\delta)\sigma_y^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)}_{\text{related to the noise}}$$

  with probability at least $1 - \delta$.
- Ⓑ $\mathbb{E}[R(\hat{\theta})] \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$

Remarks

1. A gives us a (high probability) upper bound on the excess risk.
2. B gives us a lower bound on its expectation.
3. The constants b, c, c_1 depend on σ_x, the subgaussian parameter corresponding to the covariates X.
4. Think of k^* as the number of dimensions we ignore when hiding the noise. We want k^* to be small compared to n, yet no smaller than it need be, obviously.
5. Effective rank encodes how far the vectors x are from isotropy. Isotropy implies maximum effective rank, whereas small values of effective rank suggests (just like ordinary matrix rank) that many of the vectors generating Σ are irrelevant to the variation Σ houses.

So what do we want our eigenvalues to look like?

$\Sigma = V\Lambda V^T =$

$$V \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & \text{\huge 0} & \\ & & \lambda_3 & & \\ & \text{\huge 0} & & \ddots & \\ & & & & \lambda_p \end{pmatrix} V^T$$

So what do we want our eigenvalues to look like?

$\Sigma = V \Lambda V^T =$

$$V \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & \text{\huge 0} & \\ & & \lambda_3 & & \\ & \text{\huge 0} & & \ddots & \\ & & & & \lambda_p \end{pmatrix} V^T$$

Necessary Properties For Near-Perfect Accuracy

So what do we want our eigenvalues to look like?

$\Sigma = V \Lambda V^T =$

$$V \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & \mathbf{0} & \\ & & \lambda_3 & & \\ & \mathbf{0} & & \ddots & \\ & & & & \lambda_p \end{pmatrix} V^T$$

Necessary Properties For Near-Perfect Accuracy

1. $r_0(\Sigma)$ should be small compared to n.

So what do we want our eigenvalues to look like?

$\Sigma = V \Lambda V^T =$

$$V \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & \text{\huge 0} & \\ & & \lambda_3 & & \\ & \text{\huge 0} & & \ddots & \\ & & & & \lambda_p \end{pmatrix} V^T$$

Necessary Properties For Near-Perfect Accuracy

1. $r_0(\Sigma)$ should be small compared to n.
2. r_{k^*} and R_{k^*} should be large compared to n.

So what do we want our eigenvalues to look like?

$\Sigma = V\Lambda V^T =$

$$V \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & 0 & \\ & & \lambda_3 & & \\ & 0 & & \ddots & \\ & & & & \lambda_p \end{pmatrix} V^T$$

Necessary Properties For Near-Perfect Accuracy

1. $r_0(\Sigma)$ should be small compared to n.
2. r_{k^*} and R_{k^*} should be large compared to n.
3. The sum of the λ_i should be small compared to n (to make k^* smaller).

So what do we want our eigenvalues to look like?

$\Sigma = V\Lambda V^T =$

$$V \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & \text{\Huge 0} & \\ & & \lambda_3 & & \\ & \text{\Huge 0} & & \ddots & \\ & & & & \lambda_p \end{pmatrix} V^T$$

Necessary Properties For Near-Perfect Accuracy

1. $r_0(\Sigma)$ should be small compared to n.
2. r_{k^*} and R_{k^*} should be large compared to n.
3. The sum of the λ_i should be small compared to n (to make k^* smaller).
4. The number of non-zero eigenvalues should be large compared to n.

Two Very Simple Examples

Consider $\Sigma = I_p$ (which is induced by isotropy):

$$r_0(\Sigma) = \frac{\sum_{i=1}^{p} \lambda_i}{\lambda_1} = \frac{p}{1} = p = \frac{p^2}{p} = R_0(\Sigma)$$

Next consider infinite-dimensional Σ with spectral decomposition

$$V \begin{pmatrix} \lambda_1 & & & & \\ & \frac{1}{2} & & \mathbf{0} & \\ & & \frac{1}{4} & & \\ & \mathbf{0} & & \frac{1}{8} & \\ & & & & \ddots \end{pmatrix} V^T$$

$$r_0(\Sigma) = \frac{\sum_{i=1}^{p} 2^{-i}}{\lambda_1} = \frac{\lambda_1 + 1}{\lambda_1} \qquad R_0(\Sigma) = \frac{(\lambda_1 + 1)^2}{\lambda_1^2 - 1 + \sum_{i=1} 4^{-i}} = \frac{(\lambda_1 + 1)^2}{\lambda_1^2 + 1/3}$$

Two Very Simple Examples

Consider $\Sigma = I_p$ (which is induced by isotropy):

$$r_0(\Sigma) = \frac{\sum_{i=1}^{p} \lambda_i}{\lambda_1} = \frac{p}{1} = p = \frac{p^2}{p} = R_0(\Sigma)$$

Next consider infinite-dimensional Σ with spectral decomposition

$$V \begin{pmatrix} \lambda_1 & & & & \\ & \frac{1}{2} & & \mathbf{0} & \\ & & \frac{1}{4} & & \\ & \mathbf{0} & & \frac{1}{8} & \\ & & & & \ddots \end{pmatrix} V^T$$

$$r_0(\Sigma) = \frac{\sum_{i=1}^{p} 2^{-i}}{\lambda_1} = \frac{\lambda_1 + 1}{\lambda_1} \qquad R_0(\Sigma) = \frac{(\lambda_1 + 1)^2}{\lambda_1^2 - 1 + \sum_{i=1} 4^{-i}} = \frac{(\lambda_1 + 1)^2}{\lambda_1^2 + 1/3}$$

Which of these two cases will be good for benign overfitting?

Two Very Simple Examples

Consider $\Sigma = I_p$ (which is induced by isotropy):

$$r_0(\Sigma) = \frac{\sum_{i=1}^{p} \lambda_i}{\lambda_1} = \frac{p}{1} = p = \frac{p^2}{p} = R_0(\Sigma)$$

Next consider infinite-dimensional Σ with spectral decomposition

$$V \begin{pmatrix} \lambda_1 & & & & \\ & \frac{1}{2} & & \text{\huge 0} & \\ & & \frac{1}{4} & & \\ & \text{\huge 0} & & \frac{1}{8} & \\ & & & & \ddots \end{pmatrix} V^T$$

$$r_0(\Sigma) = \frac{\sum_{i=1}^{p} 2^{-i}}{\lambda_1} = \frac{\lambda_1 + 1}{\lambda_1} \qquad R_0(\Sigma) = \frac{(\lambda_1 + 1)^2}{\lambda_1^2 - 1 + \sum_{i=1} 4^{-i}} = \frac{(\lambda_1 + 1)^2}{\lambda_1^2 + 1/3}$$

Which of these two cases will be good for benign overfitting?
Neither!

A more benign example

Example (Coverging at the slowest rate possible)

Fix $\alpha = 1, \beta > 1$. Let $\lambda_i = \frac{1}{i \log^{\beta}(i+1)}$.

What's the point of all this?

- In training, we observe both X and y, but not the noise. We then derive $\hat{\theta}$, a *noisy, imperfect* guess at θ^*.
- In prediction, we do not observe y or the noise; we're given x; we're hostage to the random object $\hat{\theta}$ we've just fit.

What's the point of all this?

- In training, we observe both X and y, but not the noise. We then derive $\hat{\theta}$, a *noisy, imperfect* guess at θ^*.
- In prediction, we do not observe y or the noise; we're given x; we're hostage to the random object $\hat{\theta}$ we've just fit.

The Magic

What's the point of all this?

- In training, we observe both X and y, but not the noise. We then derive $\hat{\theta}$, a *noisy, imperfect* guess at θ^*.
- In prediction, we do not observe y or the noise; we're given x; we're hostage to the random object $\hat{\theta}$ we've just fit.

The Magic

- Not all of the coordinates (directions) in $\hat{\theta}$ matter for prediction.

What's the point of all this?

- In training, we observe both X and y, but not the noise. We then derive $\hat{\theta}$, a *noisy, imperfect* guess at θ^*.
- In prediction, we do not observe y or the noise; we're given x; we're hostage to the random object $\hat{\theta}$ we've just fit.

The Magic

- Not all of the coordinates (directions) in $\hat{\theta}$ matter for prediction.
- In fact, some are nearly irrelevant for prediction.

What's the point of all this?

- In training, we observe both X and y, but not the noise. We then derive $\hat{\theta}$, a *noisy, imperfect* guess at θ^*.
- In prediction, we do not observe y or the noise; we're given x; we're hostage to the random object $\hat{\theta}$ we've just fit.

The Magic

- Not all of the coordinates (directions) in $\hat{\theta}$ matter for prediction.
- In fact, some are nearly irrelevant for prediction.
- Key insight: if there are enough of these unimportant directions, they can store the 'badness' of $\hat{\theta}$ that has been induced by the label noise.

What's the point of all this?

- In training, we observe both X and y, but not the noise. We then derive $\hat{\theta}$, a *noisy, imperfect* guess at θ^*.
- In prediction, we do not observe y or the noise; we're given x; we're hostage to the random object $\hat{\theta}$ we've just fit.

The Magic

- Not all of the coordinates (directions) in $\hat{\theta}$ matter for prediction.
- In fact, some are nearly irrelevant for prediction.
- Key insight: if there are enough of these unimportant directions, they can store the 'badness' of $\hat{\theta}$ that has been induced by the label noise.
    - $\hat{\theta}$ can be bad by being biased

What's the point of all this?

- In training, we observe both X and y, but not the noise. We then derive $\hat{\theta}$, a *noisy, imperfect* guess at θ^*.
- In prediction, we do not observe y or the noise; we're given x; we're hostage to the random object $\hat{\theta}$ we've just fit.

The Magic

- Not all of the coordinates (directions) in $\hat{\theta}$ matter for prediction.
- In fact, some are nearly irrelevant for prediction.
- Key insight: if there are enough of these unimportant directions, they can store the 'badness' of $\hat{\theta}$ that has been induced by the label noise.
  - $\hat{\theta}$ can be bad by being biased
  - $\hat{\theta}$, as a random object, can be bad by being a noisy stand-in for θ^*

What's the point of all this?

- In training, we observe both X and y, but not the noise. We then derive $\hat{\theta}$, a *noisy, imperfect* guess at θ^*.
- In prediction, we do not observe y or the noise; we're given x; we're hostage to the random object $\hat{\theta}$ we've just fit.

The Magic

- Not all of the coordinates (directions) in $\hat{\theta}$ matter for prediction.
- In fact, some are nearly irrelevant for prediction.
- Key insight: if there are enough of these unimportant directions, they can store the 'badness' of $\hat{\theta}$ that has been induced by the label noise.
  - $\hat{\theta}$ can be bad by being biased
  - $\hat{\theta}$, as a random object, can be bad by being a noisy stand-in for θ^*

So let's examine the punchline: how exactly can noise by hidden in unimportant directions?

How noise is hidden just right

Recall

$$\hat{\theta} := (X^T X)^\dagger X^T y = (X^T X)^\dagger X^T (\epsilon + f(X)) = (X^T X)^\dagger X^T (\textit{noise} + \textit{signal})$$

Zero-in on the left action of the operator X^T. In any direction $i, 1 \leq i \leq p$, it scales the noise from ϵ by $n\lambda_i$.

Ultimately, for any direction $i, 1 \leq i \leq p$, we can bound the prediction error in direction i by $\frac{n\lambda_i^2}{(\sum_{i>k} \lambda_i)^2}$. If we sum these terms, what do we get?

After all of this waiting, we formalize the notion under discussion.

Definition (Asymptotically Benign)

We say that Σ_n is asymptotically benign iff

$$\lim_{n \to \infty} \left(bias(\theta^*, \Sigma_n, n) + \frac{k_n^*}{n} + \frac{n}{R_{k_n^*}(\Sigma_n)} \right) = 0$$

Some things to think about with papers like this

- Has little practical value, at present. It's a conceptual piece.
- The interpretations that the authors give the phenomenon are a little fuzzy. There's a gap between what the math says and what the authors say it does.
- Theorems change between the Arxiv paper, PNAS paper, early presentations, and later presentations!

Some things to think about with papers like this

- Has little practical value, at present. It's a conceptual piece.
- The interpretations that the authors give the phenomenon are a little fuzzy. There's a gap between what the math says and what the authors say it does.
- Theorems change between the Arxiv paper, PNAS paper, early presentations, and later presentations!

Some things to think about with papers like this

- Has little practical value, at present. It's a conceptual piece.
- The interpretations that the authors give the phenomenon are a little fuzzy. There's a gap between what the math says and what the authors say it does.
- Theorems change between the Arxiv paper, PNAS paper, early presentations, and later presentations!

Some things to think about with papers like this

- Has little practical value, at present. It's a conceptual piece.
- The interpretations that the authors give the phenomenon are a little fuzzy. There's a gap between what the math says and what the authors say it does.
- Theorems change between the Arxiv paper, PNAS paper, early presentations, and later presentations!

We analyze the real-world example with Brexit included.

References