PCA with Model Misspecification

Robert M. Anderson
Stephen W. Bianchi

Berkeley Associates LLC
and
Center for Risk Management Research
University of California, Berkeley

MMDS 2016 Workshop
UC Berkeley
June 24, 2016

Principal Components Analysis (PCA) in financial data

- Assume Return Generating Process of form

$$
\begin{aligned}
R &= \phi X + \varepsilon && (1)\\
\text{Security Returns} &= \text{Factor Returns} \times \text{Factor Sensitivities} \\
&\quad + \text{Idiosyncratic Returns, where} \\
N &= \text{number of securities} \\
K &= \text{number of factors} \\
T &= \text{number of return periods (days, \ldots)} \\
&\quad \text{in estimation window} \\
\mathrm{Cov}(\phi) &= I_K, \text{ the } K \times K \text{ identity matrix}
\end{aligned}
$$

PCA in financial data

- Compute eigenvalues and eigenvectors of "covariance" matrices
  - $C_{N\times N} = \frac{R^\top R}{T}$ or
  - $C_{T\times T} = \frac{RR^\top}{T}$

  or some weighted version (correlation rather than covariance, market cap, inverse volatility, temporal weighting, ...).
- Typically, use $C_{T\times T}$; we are interested in the eigenvectors of $C_{N\times N}$, but the two matrices have the same nonzero eigenvalues and closely related eigenvectors.
- Throughout, "covariances" and "correlations" are computed without demeaning
  - Follows practitioner literature
  - Expected daily equity returns are very close to zero; the sample mean return over a (one-year) estimation window is a noisy estimate of zero

There is More Information in R than in the "Covariance" matrix $C_{N\times N}$

- Eigenvalues and eigenvectors of $C_{N\times N}$ depend on R only through the "covariance" matrix $C_{N\times N}$
- x is an eigenvector of $C_{N\times N}$
  - (portfolio representation of an estimated factor)

  $\Leftrightarrow Rx$ is an eigenvector of $C_{T\times T}$
  - (return of an estimated factor)
- Eigenvectors of $C_{T\times T}$ contain information about the *distribution* of factor returns that is not found in the "covariance" matrix $C_{N\times N}$:
  - Gaussian?
  - Excess kurtosis, as in Student t or other power laws?
  - Negative skew?

Statistical Justification of LFE in Financial Markets

- Chamberlain and Rothschild (1983), Connor and Korajczyk (1988), Bai (2003), . . .
  - Asymptotic theory in which $T, N \to \infty$ so that ε is not important for diversified portfolios
  - Assumes that $C_{N \times N} \sim \frac{X^\top \phi^\top \phi X}{T}$ converges in an appropriate sense

Misspecification

- Variable Volatility in ϕ: In financial data, volatility changes frequently
  - Changes in factor volatility (volatility of ϕ) *change the correlations of securities*
  - Example: An increase in market volatility causes the average correlation between securities to rise
- Regimes in X: In financial data, the sign of the correlation between assets reverses from time to time
  - Example: The correlation between the equity market and the price of oil is generally positive in response to demand shocks and negative in response to supply shocks
- Thus, the assumption that $C_{N\times N}$ converges is problematic when applied to financial data

Two Approaches in the Literature

- Pelger (2015a,2015b), Ait-Sahalia and Xiu (2015): Use intraday data to make T large with a fixed time horizon
  - The horizon is such that there is plausibly only one X regime
  - Replace the covariance matrix with the quadratic covariation process (a matrix-valued stochastic process whose realization at any time is a covariance matrix).
  - Limitation: using intraday data in a global model is problematic due to temporal asynchronicity
- Identify regimes with a Markov switching model
  - Attractive option for X regimes.
  - Unattractive for regimes that only involve variable factor volatility (too many regimes, volatility changes all the time, $\ldots$)
    - Not necessary

This Project

- Proposes an alternative for dealing with variable ϕ volatility, which is conceptually related to use of the quadratic covariation in Pelger
- Identifies an approach to dealing with non-Gaussian return distributions
- Identifies strengths and weaknesses in PCA applied to data containing X regimes

Variable Volatility: Formulation

- K *Constant Volatility Factors* $\tilde{\phi}$, covariance matrix the $K \times K$ identity, IID across time.
- Assume factor distribution is parametrized by a single scale factor. Need not be Gaussian.
- Volatility process v taking values in $\mathbf{R}^K$
  - Independent of $\tilde{\phi}$
  - Perhaps generated by a mean-reverting process such as the Heston Model (volatility given by modification of Ornstein-Uhlenbeck process)
- K *Variable Volatility Factors* ϕ whose returns on dates $t = 1, \ldots, T$ are given by the Hadamard (elementwise) product of v and $\tilde{\phi}$

$$\phi \;=\; v \circ \tilde{\phi} = \begin{pmatrix} v_{11}\tilde{\phi}_{11} & \ldots & v_{1K}\tilde{\phi}_{1K} \\ \vdots & \vdots & \vdots \\ v_{T1}\tilde{\phi}_{T1} & \ldots & v_{TK}\tilde{\phi}_{TK} \end{pmatrix}$$

- Analogous formulation for idiosyncratic volatility $\varepsilon = \nu \circ \tilde{\varepsilon}$

Variable Volatility: Conceptual Idea

- $C_{N \times N} \sim X^\top Q X$ where
  - $Q = \frac{\phi^\top \phi}{T}$ is the realized covariance matrix of the factor returns (discrete analogue of quadratic covariation in Pelger)
  - $$Q \sim D = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_K^2 \end{pmatrix} \tag{2}$$

    where $\sigma_k^2 = \frac{1}{T} \sum_{t=1}^{T} v_{tk}^2$ is the variance (conditional on v) of factor k over the period $\{1, \dots, T\}$.
  - *Note that D need not converge in any sense.*
- The rows of X are eigenvectors of $X^\top D X$, hence are approximate eigenvectors of $C_{N \times N}$, so PCA correctly estimates factor sensitivities even with variable factor volatility
- *This is true even though changing volatility changes the correlation of assets*

PCA with Responsive Volatility Adjustment

- Use PCA over the whole estimation period $\{1, \ldots, T\}$, without separating different volatility regimes
- This yields two sets of eigenvectors:
  - The eigenvectors of $C_{N \times N}$ are estimates of the factor sensitivities (rows of X)
  - The eigenvectors of $C_{K \times K}$ are estimates of the time series of the factor profit/loss
  - Combine the estimated factor sensitivities (eigenvectors of $C_{N \times N}$) with exponentially weighted (short half-life) standard deviation of the eigenvectors of $C_{K \times K}$. Responsive Volatility Adjustment

Estimation Error with Variable Volatility: Simulation Results

- Errors in predictions of portfolio volatility are modestly higher than in the constant volatility case
- The errors in the estimated rows of X are higher than in the constant volatility case. Further study of the economic sigifnicance is needed
- Gaussian estimates of Value at Risk (VaR) (e.g. lower 3% quantile of return) substantially underpredict risk in the presence of negative skewness
- Gaussian estimates of Expected Tail Loss (ETL) (e.g. conditional expectation of loss over lower 3% quantile of return) substantially underpredict risk in the presence of negative skewness and/or excess kurtosis

Historical Method for Predicting VaR and ETL

- Compute past distribution of Z-scores of portfolio return (actual return divided by predicted return volatility)
- Use these to predict tomorrow's VaR and ETL, conditional on today's volatility prediction
- In simulation,
  - Out-of-sample estimates of VaR using Historical Method are much more accurate than Gaussian estimates in simulation
  - Out-of-sample estimates of ETL using Historical Method are much more accurate than Gaussian estimates in the absence of skewness
  - With negative skewness, Historical Method *overpredicts* ETL, while Gaussian methods *underpredict* ETL. Looking for ways to correct overprediction.

Simulation Results: Bias and Directional Distance

Table 1: Variable Volatility vs. Constant Volatility						
	Variable Volatility			Constant Volatility		
Half-Life	Bias	DD		Bias	DD	
		Factor 1	Factor 2		Factor 1	Factor 2
10	1.029	0.029	0.163	1.031	0.016	0.093
20	1.013	0.029	0.163	1.014	0.016	0.093
30	1.008	0.029	0.163	1.008	0.016	0.093
40	1.007	0.029	0.163	1.005	0.016	0.093
50	1.008	0.029	0.163	1.004	0.016	0.093
∞	1.042	0.029	0.163	1.001	0.016	0.093

Table: Performance of Standard PCA with Responsive Volatility Adjustment in Variable and Constant Factor Volatility Models. The underlying constant-volatility model is the Bianchi, Goldberg and Rosenberg (2016) two-factor model. This table reports Bias, and average Directional Distance between the estimated and true factors, with N=1,000 stocks, a PCA estimation window of T=250 days, and 50,000 Iterations. Bias is calculated for the Equally-Weighted Portfolio; Directional Distance gives guidance for smaller or optimized portfolios.

Simulation Results: VaR and ETL

Table 2: Gaussian and Historical Method VaR Exceedances and ETL					
Distribution	Volatility	Gaussian Predictions		Historical Method	
		3% VaR Exceed	3% ETL Ratio	3% VaR Exceed	3% ETL Ratio
Gaussian	Constant	3.000%	-1.011	2.956%	-1.001
Gaussian	Variable	3.246%	-1.071	2.987%	-1.001
Student t	Variable	3.292%	-1.186	2.971%	-1.001
Skew	Variable	4.376%	-1.252	2.956%	-0.725

Table: Simulated Predicted 3% VaR Exceedance and Ratio of 3% ETL to Predicted 3% ETL. Predictions are derived from estimated volatility, using either Gaussian assumptions or the Historical Method. Simulation with N=1,000 stocks, T=250 days in each PCA window, and 50,000 Iterations, using the Bianchi, Goldberg and Rosenberg (2016) two-factor model. The underlying constant-volatility factor returns are either Gaussian with constant volatility; or Gaussian, Student t, or skewed with variable volatility. This is a two-factor model in which both factors follow the same discrete version of the Heston Process. Volatility of the Equally-Weighted Portfolio is predicted with an exponential 40-day half-life. Var and ETL predictions are then made using either Gaussian assumptions or the Historical Method.

X Regimes

Theorem

If we apply PCA to a data history combining two different X regimes, then

- *A factor which is present in both regimes will be identified as an eigenvector*
- *A factor which is present in Regime I and not in Regime II will be identified as an eigenvector if and only if it is orthogonal to all the factors present in Regime II*

In particular, a factor which is present only one of the regimes is likely to be "hybridized" with a factor in the other regime, rather than being cleanly identified by PCA

Bibliography

- Stephen W. Bianchi, Lisa R. Goldberg and Allan Rosenberg, "The Impact of Estimation Error on Latent Factor Model Forecasts of Portfolio Risk, *Journal of Portfolio Management* (forthcoming, 2016)
- Ait-Sahalia, Yacine and Dacheng Xiu, "Principal Components Analysis of High-Speed Data," technical report, University of Chicago.
- Bai, Jushan, "Inferential Theory for Factor Models of Large Dimensions," *Econometrica* 71(2003), 135-171.
- Chamberlain, Gary and Michael Rothschild, "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica* 51(1983), 1281-1304.
- Connor, Gregory, and R. A. Korajczyk, "Risk and Return in an Equilibrium APT: Application of a New Test Methodology," *Journal of Financial Economics*, 21(1988), 255289
- Pelger, Markus, "Large-dimensional factor modeling based on high-frequency observations," Working Paper #2015-08, Center for Risk Management Research, University of California, Berkeley.
- Pelger, Markus, "Understanding Systematic Risk: A High-Frequency Approach," Working Paper #2015-09, Center for Risk Management Research, University of California, Berkeley.