MOODY'S ANALYTICS

MODELING METHODOLOGY

Reverse Stress Testing from a Macroecomic Viewpoint: Quantitative Challenges & Solutions for its Practical Implementation

Authors

Juan M. Licari, Ph.D.
juan.licari@moodys.com
Senior Director

José Suárez-Lledó, Ph.D.
jose.suarez-lledo@moodys.com
Associate Director

Abstract

» A key challenge to implementing a quantitative approach to reverse stress testing (RST) is the presence of "multiplicity": the same outcome, say expected loss rate or liquidity shortfall, could be obtained with multiple combinations of risk factors, say PDs, EADs and LGDs, and macroeconomic scenarios.

» The **first contribution of this article** is to apply mathematical tools to understand and overcome the multiplicity issues inherent in RST. To this end we categorize the multiplicity into:

Type-1 multiplicity: Indeterminacy. In most RST frameworks we have a large number of risk and macroeconomic variables to match with a limited number of objectives/assumptions; i.e., more variables than equations. This indeterminacy issue can be resolved via a combination of (a) additional ad-hoc assumptions on some parameters (expert-judgment, market-wide assumptions, or values in line with regulatory guidelines), and/or (b) additional equations based on empirical findings, and/or (c) reducing the number of risk and macro factors that define a scenario while keeping its shape. The aim is to close the "degree of indeterminacy" to zero and end up with as many equations as unknowns.

Type-2 multiplicity: Properties on an inverse mapping. Even after we have closed the gap between equations and unknowns, a key challenge emerges when trying to reverse engineer a process: The inverse of a function may not behave as a function. Consider a stressed value of the risk factors x_0 that is mapped to a vector of outcomes y_0, such that $\Phi(x_0, y_0) = 0$. Mapping y_0 back (i.e., reverse engineering the process) could give us a value of x that is different from x_0: There may exist another vector x_1 that is consistent with the same outcome, $\Phi(x_1, y_0) = 0$. Some key properties of the mapping Φ will help us ensure that (at least locally) one can "safely" invert the process.

The **second contribution of this article** is to provide the reader with statistical tools to overcome the multiplicity challenges in practice. To deal with type-1 multiplicity we propose the use of "factor analysis". To ensure that these factors can still have a meaningful "interpretation" we link them to well-known macroeconomic series for the U.S., U.K. and German economies. The takeaway of this exercise is that intuition can still be kept on the nature of the scenarios while the underlying dimension of the indeterminacy has been dramatically reduced. The modeler now has a better chance of matching the number of equations to the number of risk and macro parameters. To handle type-2 multiplicity we make use of linear algebra techniques to provide structure on the stress-testing process that will allow us to carry RST exercises for linear models (and any nonlinear monotonic transformations of them—such as logarithmic and logistic).

Table of Contents

Introduction

The Financial Services Authority defines reverse stress tests as "*tests that require a firm to assess scenarios and circumstances that would render its business model unviable, thereby identifying potential business vulnerabilities. Reverse stress-testing starts from an outcome of business failure and identifies circumstances where this might occur. This is different to general stress and scenario testing which tests for outcomes arising from changes in circumstances.*"

At its core, reverse stress testing proposes to "invert" the standard process, starting now from an "outcome" (business failure) with the aim of finding potential states-of-nature (scenarios) consistent with such an outcome. When implementing RST in a quantitative way, **multiplicity** emerges as a challenge. The same outcome (say, high expected losses or liquidity shortfall) could materialise under multiple combinations of risk factors (say, PDs, EADs and LGDs) and alternative macroeconomic scenarios.

Should we worry about multiplicity? Yes. The reverse engineering exercise can end up identifying only a subset of scenarios that are consistent with the starting assumption. The risk we face is that those **scenarios that were not found** could be of more relevance—and severity—compared with the ones that were identified by the process.

The rest of the article is organized as follows. Section 1 addresses multiplicity from a mathematical viewpoint by distinguishing between (1.1) indeterminacy and (1.2) challenges emerging from reverse engineering a function/process. Section 2 proposes practical solutions to avoid multiplicity: (2.1) the use of factor analysis to eliminate indeterminacy and (2.2) the use of linear algebra to handle the inverse mapping of linear equations (including those that emerge from nonlinear, strictly monotone transformations of the original data). Section 3 summarizes the main findings of the paper. An appendix describes the mathematical properties behind a factor analysis decomposition.

1. Multiplicity from a Mathematical Viewpoint

Mathematical Setup

Consider the standard stress-testing process (ST) as a "mapping" from scenarios and risk parameters: $x \in R^N$, into outcomes: $y \in R^M$, such that a set of conditions is satisfied: $\Phi(x, y) = 0$. The mapping $\Phi : X \times Y \rightarrow R^M$ represents the system of equations that links scenarios, risk parameters and outcomes. The stress-testing task can be described as follows: For a given stressed scenario and its associated risk parameters, $\tilde{x} \in R^N$, find the corresponding outcomes, $\tilde{y} \in R^M$, such that $\Phi(\tilde{x}, \tilde{y}) = 0$. This standard process is built so that there are as many equations as unknowns: The dimension of y coincides with the dimension of the range of the Φ system. But in practice, the number of scenarios and risk parameters tends to be greater than the number of equations: $N \geq M$. Using the previously defined notation, the reverse stress-testing (RST) process will consist of: (i) starting from the extreme output/outcome, $\hat{y} \in R^M$, (ii) find those potential scenarios and risk parameters, $\hat{x} \in R^N$, (iii) such that $\Phi(\hat{x}, \hat{y}) = 0$.

A useful definition is that of a "projection function": $\pi_x : X \times Y \rightarrow X$ with $\pi_x(x, y) = x; \forall (x, y)$ such that $\Phi(x, y) = 0$. Similarly, $\pi_y(x, y) = y$. Using this concept, the ST process is defined as: (a) given $\tilde{x} \in R^N$, (b) find $\tilde{y} \in R^M$ such that (c) $\tilde{y} \in \pi_y(\pi_x^{-1}(\tilde{x}))$. The RST process is therefore: (i) for an outcome $\hat{y} \in R^M$, (ii) find $\hat{x} \in R^N$ such that (iii) $\hat{x} \in \pi_x(\pi_y^{-1}(\hat{y}))$.

As a mapping/correspondence, RST consists of the following composite formula: $\pi_x \circ \pi_y^{-1}$.

In practice, the π_x^{-1} relationship is typically well-behaved: Every value of x is associated with a single outcome y. But the challenge emerges when working in reverse order since π_y^{-1} can consist of multiple values of x. Studying the properties of the π_y^{-1} mapping is at the core of the understanding of the RST process.

1.1: Type-1 Multiplicity: Indeterminacy

In most RST frameworks, we are faced with the task of matching a large number of risk and macroeconomic variables with a limited set of assumptions: # variables > # equations, or $N > M$. Under such conditions, indeterminacy takes the form of a **continuum** of solutions (scenarios and risk parameters) whose mathematical properties will help the modeler identify avenues to close the extra degrees of freedom.

In formal terms: if $N > M$, then shape of the solution set $\hat{X} = \{x \in X; x \in \pi_x(\pi_y^{-1}(\hat{y}))\}$ will depend on the mathematical properties of the Φ system (and therefore the π_y^{-1} mapping). Under some "regularity" conditions on Φ (e.g.; sufficient conditions to satisfy the assumptions for the implicit function theorem) the solution set $\hat{X}$ is a smooth-manifold of dimension $N - M$. In other words, there exists a continuum of scenarios and risk parameters (set $\hat{X}$) that are consistent with the original outcome $\hat{y} \in R^M$.

This indeterminacy needs to be dealt with before any further attempt to successfully reverse engineer the process. Solutions will require (a) additional ad hoc assumptions on some parameters (expert judgment, market-wide assumptions or values in line with regulatory guidelines), and/or (b) additional equations based on empirical findings; e.g., $LGD = f(PD)$, and/or (c) reducing the number of risk and macro factors that define a scenario while keeping its shape.

The end goal is to close the "degree of indeterminacy" to zero and end up with as many equations as unknowns.

1.2: Type-2 Multiplicity: Inverse Mapping

Even after closing the gap between equations and unknowns ($N - M = 0$), a new challenge emerges when trying to reverse engineer the ST process: The inverse of a function may not behave as a function. Consider a stressed value of the risk factors, x_0 , that is mapped to a vector of outcomes, y_0 , such that $\Phi(x_0, y_0) = 0$. Mapping y_0 back (i.e., reverse engineering the process) could give us a value of x that is different from x_0 : There may exist another vector x_1 that is consistent with the same outcome $\Phi(x_1, y_0) = 0$. Specific characteristics of the ST system Φ will help us ensure that (at least locally) one can "invert" the process and obtain a reliable RST mapping $\pi_x \circ \pi_y^{-1}$. Applications of results such as "Inverse Function" and "Implicit Function" theorems will help us understand the shape of the solution-set $\hat{X} = \{x \in X; x \in \pi_x(\pi_y^{-1}(\hat{y}))\}$. Under some "regularity" conditions on Φ the set $\hat{X}$ is a zero-dimensional smooth-manifold (set of isolated points). Moreover, each point in $\hat{X}$ **locally** maps (a) $X \rightarrow Y$ through $\pi_y \circ \pi_x^{-1}$ (stress testing) and (b) $Y \rightarrow X$ through $\pi_x \circ \pi_y^{-1}$ (reverse stress testing) in a one-to-one (and unique) smooth fashion (overcoming type-2 multiplicity).

2. Multiplicity from a Practical Viewpoint

2.1. Handling Type-1 Multiplicity: An Application of Factor Analysis for U.K., U.S. and Germany

To handle type-1 indeterminacy we propose the use of factor analysis. This technique allows the modeler to reduce the dimension of the scenario-space; having a better chance at matching the number of equations and variables (macro and risk factors). By leveraging on the strong correlation of macroeconomic series, a modeler can concentrate on a smaller set of instruments (factors) that can still replicate most of the variability of the whole sample. The challenge, however, could be the lack of interpretation for the factors.

We carry a case study to link the top factors to specific macroeconomic series, ensuring that intuition is kept on the nature of the factors while reducing the chances of indeterminacy. Below are the findings for the U.K., U.S. and Germany.

Results for the UK economy

Table I: U.K. Factor Analysis–Top 21 Macroeconomic Series–Sample starting in 2000Q1

```
Factor analysis/correlation                        Number of obs    =      48
    Method: principal factors                      Retained factors =      14
    Rotation: (unrotated)                          Number of params =     203
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	9.58311	5.68839	0.4995	0.4995
Factor2	3.89472	2.17554	0.2030	0.7025
Factor3	1.71918	0.32750	0.0896	0.7921
Factor4	1.39168	0.42675	0.0725	0.8646
Factor5	0.96493	0.34703	0.0503	0.9149
Factor6	0.61790	0.18001	0.0322	0.9471
Factor7	0.43789	0.15421	0.0228	0.9700
Factor8	0.28368	0.07129	0.0148	0.9847
Factor9	0.21239	0.06647	0.0111	0.9958
Factor10	0.14592	0.08320	0.0076	1.0034
Factor11	0.06272	0.02724	0.0033	1.0067
Factor12	0.03548	0.01399	0.0018	1.0085
Factor13	0.02149	0.01927	0.0011	1.0097
Factor14	0.00222	0.00844	0.0001	1.0098
Factor15	-0.00622	0.00523	-0.0003	1.0095
Factor16	-0.01145	0.00397	-0.0006	1.0089
Factor17	-0.01542	0.00627	-0.0008	1.0081
Factor18	-0.02169	0.00980	-0.0011	1.0069
Factor19	-0.03149	0.01253	-0.0016	1.0053
Factor20	-0.04402	0.01326	-0.0023	1.0030
Factor21	-0.05728	.	-0.0030	1.0000

LR test: independent vs. saturated: chi2(210) = 1443.70 Prob>chi2 = 0.0000

The top five factors explain more than 90% of the variability of the whole sample, with factors 1 and 2 explaining 50% and 20%, respectively. This is encouraging news from an RST angle; most of the information embedded in the U.K. economic cycle can be replicated by these five factors. What do these factors represent? Figures I to VI illustrate the correlation of the factors with specific macroeconomic series.

Figure I: U.K. Factor 1 and GDP Growth

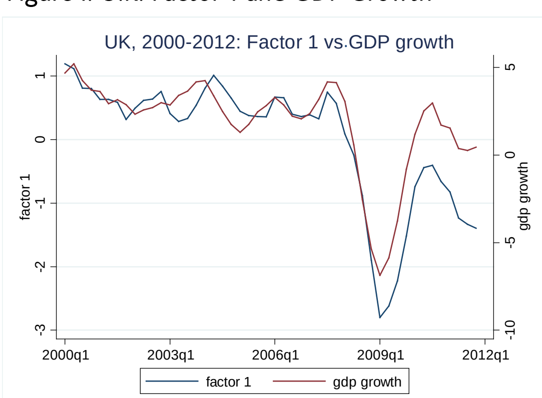


Figure II: U.K. Factor 2 and Employment Growth

Figure III: U.K. Factor 3 and Inflation

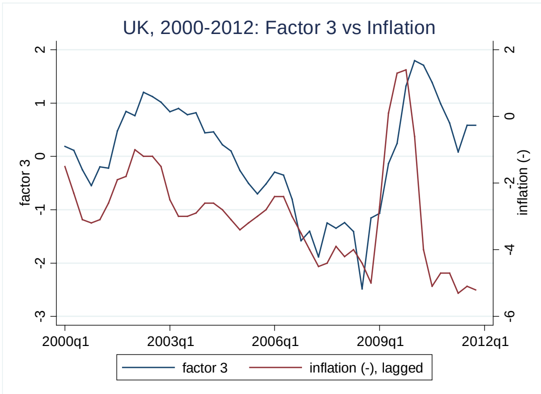


Figure IV: U.K. Factor 3 and Policy Rate

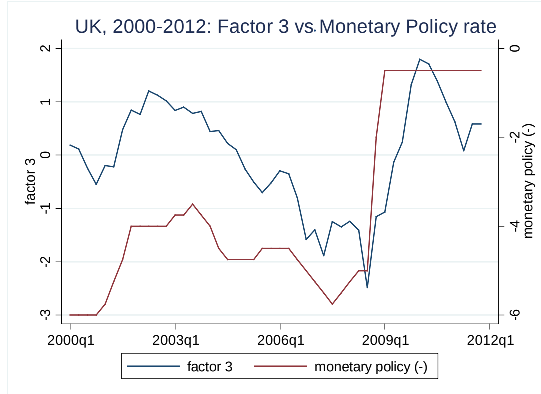


Figure V: U.K. Factor 4 and Money Supply Growth

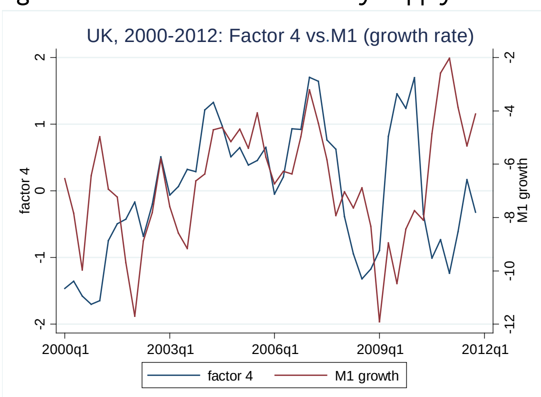
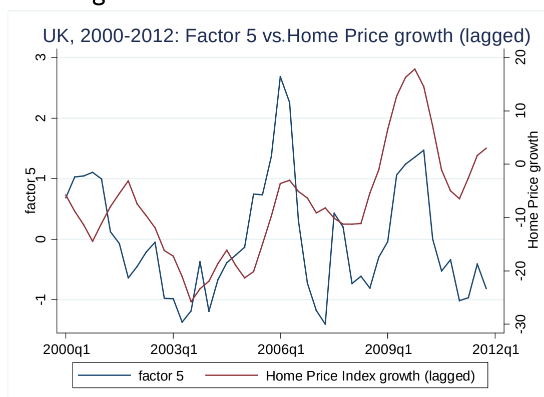


Figure VI: U.K. Factor 5 and Home Price Growth



The dimension of the "scenario-space" has been dramatically reduced while keeping an interpretation of the factors: (1) core real business cycle fluctuations, (2) labour market, (3) monetary cycle (inflation and rates), (4) money supply (complementing factor 3), and (5) U.K. housing market.

Results for the US economy

Table II: US factor analysis– Top 25 macroeconomic series – Sample starting in 200Q1

```
Factor analysis/correlation                          Number of obs    =        48
    Method: principal factors                        Retained factors =        18
    Rotation: (unrotated)                            Number of params =       297
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	12.61718	7.18730	0.5145	0.5145
Factor2	5.42988	2.99817	0.2214	0.7359
Factor3	2.43170	1.17792	0.0992	0.8350
Factor4	1.25379	0.39890	0.0511	0.8862
Factor5	0.85489	0.17976	0.0349	0.9210
Factor6	0.67512	0.18184	0.0275	0.9486
Factor7	0.49328	0.16899	0.0201	0.9687
Factor8	0.32429	0.12289	0.0132	0.9819
Factor9	0.20140	0.11619	0.0082	0.9901
Factor10	0.08521	0.01199	0.0035	0.9936
Factor11	0.07322	0.02267	0.0030	0.9966
Factor12	0.05055	0.01415	0.0021	0.9986
Factor13	0.03640	0.01441	0.0015	1.0001
Factor14	0.02198	0.01023	0.0009	1.0010
Factor15	0.01175	0.00626	0.0005	1.0015
Factor16	0.00550	0.00099	0.0002	1.0017
Factor17	0.00451	0.00398	0.0002	1.0019
Factor18	0.00053	0.00157	0.0000	1.0019
Factor19	-0.00104	0.00182	-0.0000	1.0019
Factor20	-0.00286	0.00089	-0.0001	1.0018
Factor21	-0.00375	0.00098	-0.0002	1.0016
Factor22	-0.00472	0.00228	-0.0002	1.0014
Factor23	-0.00700	0.00347	-0.0003	1.0011
Factor24	-0.01048	0.00672	-0.0004	1.0007
Factor25	-0.01720	.	-0.0007	1.0000

```
LR test: independent vs. saturated: chi2(300) = 2554.41 Prob>chi2 = 0.0000
```

Figure VII: U.S. Factor 1 and GDP Growth

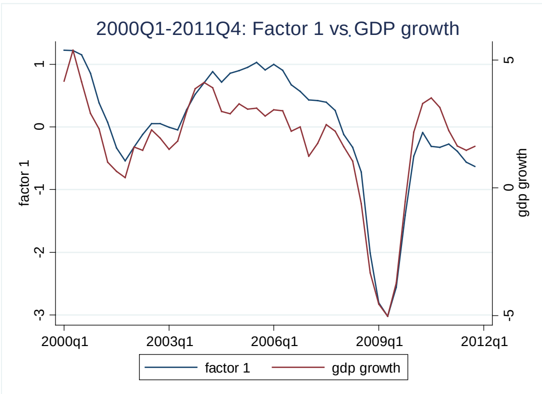


Figure VIII: U.S. Factor 2 and Unemployment Rate

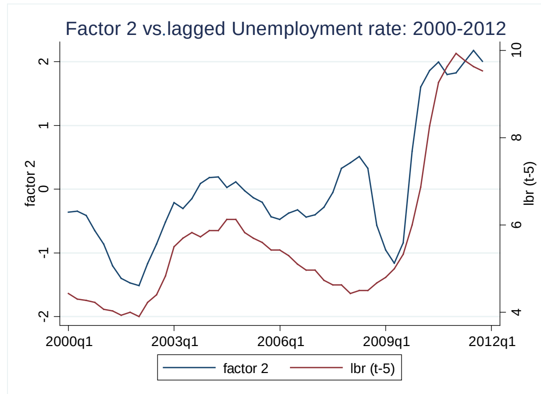


Figure IX: U.S. Factor 3 and Policy Rate

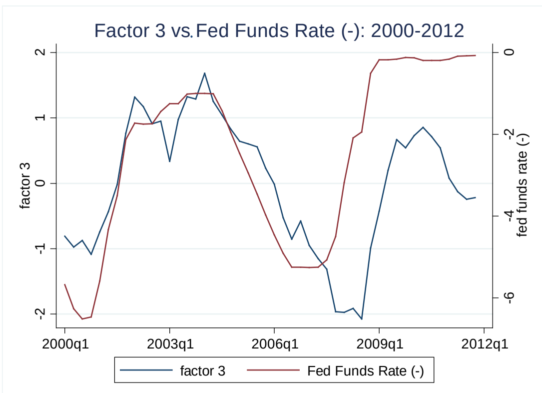
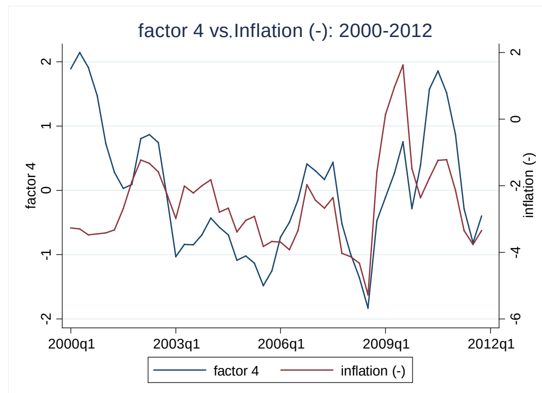


Figure X: U.S. Factor 4 and Inflation



The results for the US economy are similar to those of the UK: (1) core real business cycle fluctuations, (2) labour market trends, (3 & 4) monetary cycle (policy rates and inflation).

Results for the German Economy

Table III: Germany Factor Analysis–Top 18 Macroeconomic Series–Sample Starting in 2000Q1

Factor analysis/correlation — Number of obs = 48
Method: principal factors — Retained factors = 12
Rotation: (unrotated) — Number of params = 150

Beware: solution is a Heywood case
(i.e., invalid or boundary values of uniqueness)

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	5.98786	2.03131	0.3670	0.3670
Factor2	3.95655	0.94112	0.2425	0.6095
Factor3	3.01543	1.64724	0.1848	0.7943
Factor4	1.36819	0.48702	0.0839	0.8782
Factor5	0.88117	0.35316	0.0540	0.9322
Factor6	0.52801	0.15712	0.0324	0.9646
Factor7	0.37089	0.21556	0.0227	0.9873
Factor8	0.15534	0.06105	0.0095	0.9968
Factor9	0.09429	0.00578	0.0058	1.0026
Factor10	0.08851	0.03796	0.0054	1.0080
Factor11	0.05055	0.02540	0.0031	1.0111
Factor12	0.02515	0.03319	0.0015	1.0127
Factor13	-0.00804	0.00953	-0.0005	1.0122
Factor14	-0.01756	0.00407	-0.0011	1.0111
Factor15	-0.02164	0.00998	-0.0013	1.0098
Factor16	-0.03161	0.01572	-0.0019	1.0078
Factor17	-0.04734	0.03306	-0.0029	1.0049
Factor18	-0.08040	.	-0.0049	1.0000

LR test: independent vs. saturated: chi2(153) = 1177.16 Prob>chi2 = 0.0000

Figure XI: Germany Factor 1 and GDP Growth

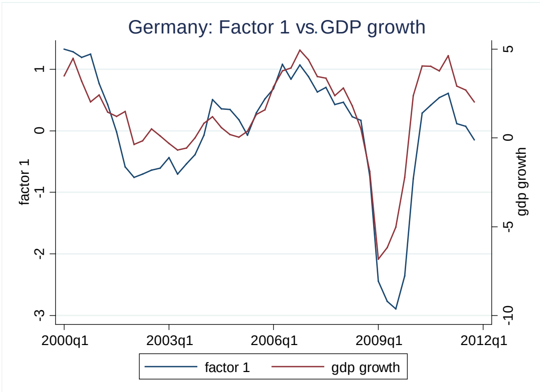


Figure XII: Germany Factor 2 and Employment

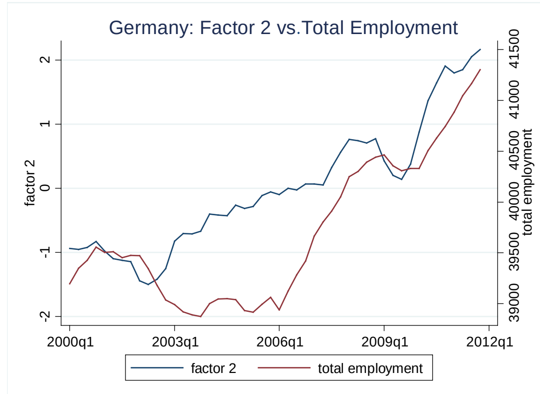


Figure XIII: Germany Factor 3 and Policy Rate

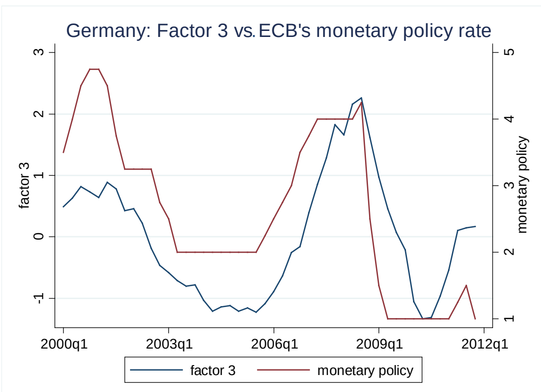


Figure XIV: Germany Factor 4 and Inflation

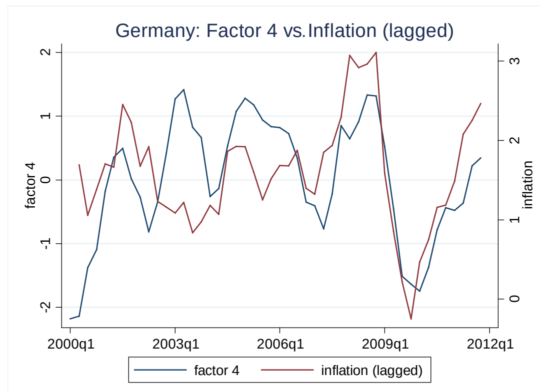
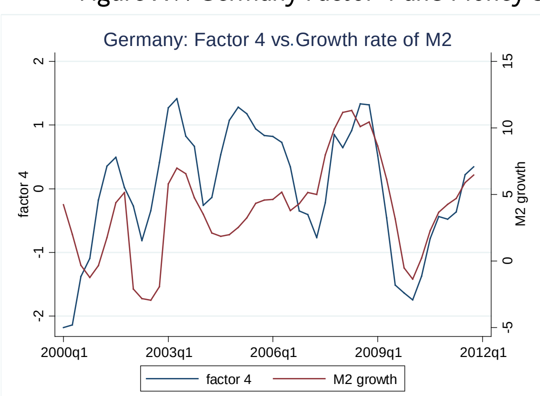


Figure XV: Germany Factor 4 and Money Supply Growth



The results for the German economy suggest the following association between factors and macro drivers: (1) core real business cycle fluctuations, (2) labour market trends (employment in levels rather than growth), (3) monetary cycle (policy rates), (4) complementing the monetary cycle (inflation and money growth rate).

The take-away of this exercise is that intuition can still be kept on the nature of the factors while the underlying dimension of the indeterminacy has been dramatically reduced. With this one-to-one match between macroeconomic variables and factors, the modeller responsible for RST has a better chance of succeeding, due to the lower number of variables to be matched with outcomes/targets.

2.2. Handling Type-2 Multiplicity: An Application to Linear Models

Risk modellers usually apply non-linear transformations to risk variables (e.g., logistic or logarithmic mappings) and then model these transformed series in a linear fashion against macroeconomic and other risk drivers. Properties of the coefficient matrices of these linear systems will determine whether the process can be "inverted". The simplest 1-dimensional linear model will require a non-zero estimated coefficient. When dealing with higher order systems, the non-zero condition translates to the determinant of a matrix.

Consider a linear model: $Y_t = \beta X_t + \varepsilon_t$, where $Y_t \in R^M$ represents a vector of risk variables or targets and $X_t \in R^N$ contains all the model drivers. The (one-step-ahead) forecast takes the form $Y_{t+1}^F = \hat{\beta} X_{t+1}$, with $\hat{\beta}$ representing the estimated parameters.

Suppose now that the RST mandate is to start with an assumption for the outcome, say Y_{t+1}^S, and find consistent values for X_{t+1}. This implies $(\hat{\beta}'\hat{\beta})^{-1}\hat{\beta}'Y_{t+1}^S = X_{t+1}$ .If $N > M$, we are back into type-1 multiplicity; so $N = M$ is necessary to continue working on the RST process. With an equal number of targets and control variables, the necessary and sufficient condition that needs to be verified is the full rank of $\hat{\beta}'\hat{\beta}$ (or a non-zero determinant). Ensuring that $\hat{\beta}'\hat{\beta}$ is "invertible" will avoid the presence of type-2 multiplicity and the scenarios in X_{t+1} will be uniquely linked to the outcome Y_{t+1}. An important caveat is that this unique identification would be achieved in the linear case. With non-linear models it is normally the case that at most local uniqueness can be guaranteed.

Having ensured the "invertibility" of the process, the modeller can calculate the values for the scenarios and risk parameters $\hat{X}_{t+1}$ that are consistent with the starting assumption on the outcome, $\hat{Y}_{t+1}^F$. The RST process can finally be carried forward.

3. Concluding Remarks

RST requirements apply primarily to all banks, building societies and insurers and, under normal circumstances, the exercises are to be conducted with a periodicity of at least once a year. Notwithstanding that the original motivation for running these STs may be the compliance with regulatory legislation, they pose a great opportunity to improve portfolio construction, hedging and risk management more generally. Discussions among practitioners and regulators have revealed a number of factors to take into consideration when reverse engineering scenarios. The identified scenarios should address risks relevant to the institution— mainly credit risk, market risk, operational risk, interest rate risk and liquidity risk. They should also embed more specific vulnerabilities, where appropriate, such as regional and sector characteristics or business-line exposures and concentrations. Ideally, scenarios could also consider a confluence of events.

The ultimate goal of this article is to point to the main challenge behind a quantitative approach to RST: multiplicity. This enemy can be fought with statistical and mathematical tools, and the modelers involved in the process should understand the threats caused by this phenomenon. We provided the reader with examples of statistical tools that can help us eliminate this anomaly. RST is a young research field, and many interesting questions will arise when implementing solutions and processes worldwide.

Appendix A Factor & Principal Components Analysis

One of the requirements to be satisfied in order to apply the inverse function theorem is that the dimension of the set of economic assumptions matches that of the performance metrics space, i.e. $N = M$. In this section we propose the use of factor analysis (FA) to achieve this. The idea is to use the main factors extracted from the economic data set as the exogenous variables in the model of risk parameters instead of using a wider set of economic variables, whose dimension would surely be greater than that of the space of risk parameters.

FA and principal component analysis (PCA) are algebraic techniques intended to capture the variability in a data set in terms of a number of factors smaller than the initial number of variables in the data set. These techniques are usually applied to the data covariance matrix (or to some transformation of the data matrix). They rely on eigenvector decompositions of such a matrix to obtain the relationship of the original variables with the unobserved "latent" factors (components).

Suppose we have a set of economic data, $X \subseteq R^{NT}$, a matrix with the economic variables in each of the N rows and their respective values at every time period in the T columns. A general model for this data set is assumed to take the form (A1): $X = LF + f$ [1], where the data set is assumed to be generated by linear combinations of the factors, F , and the so-called loadings of each variable on each factor, L . The matrix L would be of dimension $N \times K$, and F is of dimension $K \times T$. This matrix, F , is said to contain the "common factors," while the $N \times T$ matrix f is said to contain the "specific or individual factors". We are only interested in the common factors, so from now on we will focus on the case $X = LF$. The relevant point here is that $K < N$ i.e., the number of factors generating the whole data set, should be smaller than the number of variables in the data set. Statistical packages derive the factors so that F is actually $N \times T$, so the modeler can choose the more relevant factors, i.e., those generating the largest percentage of variability in the data. We will come back to this point in a few lines.

Now, let us relate this discussion with the eigenvector decomposition on which this technique relies. As we already mentioned, vector decompositions of matrices can be applied to the matrix of data or to their covariance matrix. Since we are interested in capturing the relationships between variables we will focus on covariance matrices. However, it is worth noticing that a direct equivalence can be established between the decomposition of a matrix of data and that of the covariance matrix (see next section). In short, by diagonalizing the data covariance matrix, $\text{cov}(X) = X'X$, we would be able to model our set of variables, as in (A1), in a parsimonious manner while capturing the correlations among them. Because the matrix under study is assumed to be a linear combination of these factors, a natural approach is to extract the eigenvectors of the covariance matrix[2] according to equation $(C - \lambda I)\mathbf{e} = 0$, where $\mathbf{e}$ is an eigenvector of the matrix C and λ an eigenvalue, so that $\text{cov}(X) = X'X = EDE'$; E being the matrix of the eigenvectors (orthogonal by construction) and D being a diagonal matrix with the eigenvalues.

In the terminology of linear algebra, the matrix of factor loadings, L , is formed with the eigenvectors (multiplied by their eigenvalues), $L = ED$, and can be thought of as the vectors in the basis spanning the whole subspace of original variables considered. In fact, given they are a basis, they can span entire subspaces of R^N. But since we are interested not in some generic subspace of variables but rather in the specific path of our economic variables, we just need to find the particular linear combinations generating our set of economic assumptions. In other words, in equation $X = LF = EDF$ we just need to find the path of factor scores that would generate our particular matrix X . This is easily achieved by inverting the matrix ED (non-singular by construction) so that the factors (factor scores) are computed as $(ED)^{-1}X = F$. The eigenvectors of $\text{cov}(X)$ (multiplied by their eigenvalues) are known as the factor loadings and are the correlations of each variable in X with an underlying factor.

Now, most statistical packages extract the factors scores and loadings for us. All we need to do is to give the computer the data set. Nevertheless, recall that computers will output as many factors as economic variables in the data set. After all, their procedure is designed to obtain a basis to generate the whole subspace of economic assumptions. However, our ultimate goal is to reduce the

[1] For simplicity, we have assumed here that the variables in the data set are already de-trended, or that their mean is zero.

[2] There exist several techniques to extract the factors by decomposition of the covariance matrix. We focus here on the more intuitive method.

number of variables, therefore we want to choose only a relevant subset of factors. Obviously, this subset of factors will be able to generate only a subset of the complete data sample of our economic series. The idea is then to be able to generate a subset of data as large as possible and as representative of the original set as possible. Therefore, a not-trivial decision is to be made as regards to what factors are chosen to model the data. Several criteria exist under which we can choose these factors, each one being useful for different purposes. However, since statistical packages output the factors ordered by default from larger to lower percentage of variability in the data explained, this could precisely be a good criterion. It is typically the case with data sets of macroeconomic assumptions that three to five factors will capture more than 90% of the variability in the sample data. On a final note, the main difference with principal component analysis is that PCA derives the components (factors) so that each one of them explains the largest possible fraction of the residual variance in the original variables that was not explained by the previous component. Nonetheless, it is normally the case that in macroeconomic samples both FA and PCA tend to yield similar results. Factor analysis is often preferred when the goal of the analysis is to detect structure in the data.

The advantage of this method is that the obtained factors (principal components) are orthogonal by construction, thus guaranteeing that the space of risk drivers will have full rank. [3] This by itself does not imply anything for the reverse engineering process, for when $N > M$ we will still face multiplicity. In the case of $N = M$, however, the set of factors being orthogonal by construction implies that we could reverse engineer the relevant scenarios by inverting $\hat{\beta}'\hat{\beta}$. By extracting the principal components from the set of economic assumptions, X_t, we can reduce the dimensionality of the space of (economic) variables, thus reducing the degree of multiplicity. Identification of the actual macroeconomic assumptions behind the principal components may be achieved either with the loadings of these components[4] or through regressions of the original macro variables as functions of the factors.

An important aspect of this approach is that it fits best with stress-testing methodologies that take the economic scenarios from models that capture the interdependencies among economic variables. Because FA and PCA are designed to capture the co-variation of variables, modeling the set of economic assumptions (coming from that type of macroeconomic models) as functions of the factors will be part of a consistent and coherent framework. In any scenario (simulation) of the credit model, since the dynamics of the factors will preserve the dynamic covariate properties of the macroeconomic data, the economic variables being functions of the factors will also move consistently across such scenarios. Examples of such macroeconomic models are the so-called Dynamic Stochastic General Equilibrium (DSGE) models, (Structural) Vector Autoregressive (VAR) models, etc. This consistency can hardly be achieved with statistical models that infer the different economic scenarios rather than explicitly incorporating the macro economy. This type of model is designed neither to capture cross-correlations across variables nor to forecast out of sample.

Equivalence between decomposition of data matrices and covariance matrices

Any matrix of data X can be expressed, through its singular value decomposition, as the product $X = V_1 D V_2'$, with V_1 and V_2 being the matrices containing the left and right singular vectors, and D being a diagonal matrix with the singular values. Now, since the covariance matrix of X would be computed as $X'X$, we find that $X'X = V_2 D V_1' V_1 D V_2'$. And given that V_1 and V_2 are orthogonal by construction, we know that $V_1' V_1 = I$, which implies that $X'X = V_2 D^2 V_2'$. Finally, we know that an eigenvector (and its associated eigenvalue) of a matrix C satisfies the equation $C\mathbf{v} = \lambda\mathbf{v}$. Just notice that $C = X'X = V_2 D^2 V_2'$ actually implies $CV_2 = D^2 V_2'$, which would be the equivalent (in matrix notation) of the eigenvector equation above. Thus, even though the singular value decomposition can be applied to generic rectangular matrices and the eigenvector decomposition can be applied only to some square matrices, we know that in general the singular vectors of a matrix X are directly related to the eigenvectors of $X'X$ and the singular values are the square roots of the eigenvalues.

[3] Note that principal components are theoretically guaranteed to be independent only if the data set is jointly normally distributed.

[4] The eigenvector decompositions are constructed so that the original variables are linear combinations of the factors, though these combinations might not necessarily be unique.

References

- M. Hirsch, *Differential Topology* (Springer-Verlag, 1994).
- J. Stock and M. Watson, *Forecasting With Many Predictors* (Handbook of Economic Forecasting, 2006).
- Antonio Villanacci, Laura Carosi, Pierluigi Benevieri and Andrea Battinelli, *Differential Topology and General Equilibrium With Complete and Incomplete Markets* (Kluwer Academic Publishers, 2002).

.