PRINCIPAL COVARIATE REGRESSION PADA DATA RUNTUN WAKTU

Nuruma Nurul Malik[1], Fevi Novkaniza[2]

Departemen Matematika FMIPA UI, Depok
Email korespondensi : fevi.novkaniza@sci.ui.ac.id

Abstrak

Pada suatu data runtun waktu sering ditemui permasalahan dalam melakukan peramalan nilai variabel respon untuk periode selanjutnya, apalagi jika melibatkan banyak variabel kovariat. *Principal Covariate Regression (PCovR)* adalah suatu model regresi yang menggambarkan hubungan antara suatu variabel respon dengan banyak variabel kovariat. Taksiran parameter pada *PCovR* diperoleh melalui peminimuman sebuah fungsi kriteria, dimana fungsi kriteria tersebut merupakan fungsi dari *error* peramalan dan *error* kompresi informasi variabel-variabel kovariat yang masing-masing sudah terboboti. Selanjutnya taksiran parameter regresi ini disubstitusikan dalam persamaan peramalan runtun waktu dan digunakan untuk meramalkan nilai variabel respon untuk periode selanjutnya. Selain itu, juga diberikan contoh aplikasi peramalan runtun waktu dengan menggunakan *PCOvR*.

Kata Kunci: runtun waktu, variabel kovariat, fungsi kriteria, peramalan, dan *principal covariate regression.*

1. PENDAHULUAN

Dalam peramalan runtun waktu di bidang makroekonomi dan bisnis sering ditemui bahwa terdapat banyak variabel kovariat yang mempengaruhi variabel respon berdasarkan teori-teori ekonomi yang berlaku. Untuk menggambarkan hubungan antara variabel respon dengan variabel-variabel kovariat tersebut diperlukan sebuah model regresi yang dapat menjelaskan pola hubungan antara banyak variabel kovariat dan variabel respon. Kemudian model tersebut digunakan untuk meramalkan nilai variabel respon pada periode-periode selanjutnya (Heij dkk, 2006).

Biasanya dalam membentuk sebuah model regresi diperlukan asumsi banyaknya variabel kovariat lebih sedikit atau sama dengan jumlah observasi dari setiap variabel. Artinya, jika sebanyak T observasi tersedia untuk variabel respon dan setiap variabel kovariat, maka untuk jumlah variabel kovariat yang lebih banyak dari jumlah observasi tidak dimungkinkan melakukan regresi linier berganda yang melibatkan semua variabel kovariat. Jika jumlah variabel kovariat sangat banyak, dimana banyaknya variabel kovariat lebih sedikit daripada jumlah observasi, maka juga tidak disarankan untuk membentuk model regresi linier yang melibatkan semua variabel kovariat karena hasil peramalan akan mempunyai variansi yang besar disebabkan *overfitting* (Heij, 2006).

Ada banyak model yang dapat digunakan dalam melakukan peramalan nilai variabel respon periode-periode selanjutnya yang melibatkan banyak variabel kovariat, salah satunya adalah *Principal Covariate Regresion* (PCR). Secara umum model PCR terdiri atas dua langkah, yaitu sebagai langkah pertama, informasi dari variabel-variabel kovariat diringkas menjadi sejumlah komponen utama yang relatif lebih sedikit jumlahnya. Selanjutnya pada langkah kedua, komponen utama tadi digunakan sebagai variabel kovariat baru untuk melakukan peramalan nilai variabel respon periode-periode selanjutnya (Heij dkk, 2006). Akan tetapi, kerugian dari model PCR ini adalah pembentukan

komponen utama dalam langkah pertama tidak langsung berkaitan penggunaannya dalam peramalan pada langkah kedua (Heij dkk, 2005).

Untuk itu sebagai model alternatif, De Jong and Kiers pada tahun 1992 memperkenalkan *Principal Covariate Regression* (PCovR), yaitu model regresi yang menggambarkan pola hubungan antara variabel respon dengan banyak variabel kovariat dengan metode penaksiran parameter berupa peminimuman sebuah fungsi kriteria.

Fungsi kriteria merupakan fungsi yang menggabungkan dua langkah pada metode-metode sebelumnya, yaitu peringkasan informasi variabel-variabel kovariat menjadi komponen utama dan peramalan nilai variabel respon dengan menggunakan komponen utama sebagai variabel kovariat baru. Fungsi kriteria merupakan fungsi dari *forecast error* dan *predictor compression error* yang masing-masing sudah terboboti (Heij dkk, 2006). Selanjutnya akan dibahas prosedur PCovR pada suatu data runtun waktu dan digunakan untuk meramalkan nilai variabel respon pada periode selanjutnya.

2. METODE

Dalam pemodelan runtun waktu menggunakan PCovR, terdapat beberapa tahapan yang harus dilakukan, yaitu sebagai tahap awal adalah standarisasi data dari variabel respon dan variabel-variabel kovariat. Tahap kedua adalah melakukan peringkasan informasi dari variabel-variabel kovariat melalui pembentukan komponen utama, dimana penentuan jumlah komponen terbaik berdasarkan kriteria BIC (Bayesian Information Criteria). Selanjutnya penaksiran parameter model PCovR dilakukan melalui peminimuman sebuah fungsi kriteria dan taksiran parameter tersebut digunakan untuk peramalan nilai variabel respon pada periode selanjutnya.

3. HASIL DAN PEMBAHASAN

3.1 Data Runtun Waktu

Misalkan diketahui suatu data runtun waktu berdasarkan observasi dari sebuah variabel respon Y dan variabel kovariat sebanyak k, yaitu $X_1, X_2, \dots, X_k$, dimana observasi dilakukan sebanyak T periode waktu. Data tersebut dapat dinyatakan dalam bentuk vektor dan matriks sebagai berikut:

$$\mathbf{y^T} = [y_1 \quad y_2 \quad \dots \quad y_{T-1} \quad y_T] \qquad (3.1)$$

$$\boldsymbol{X} = [\mathbf{x_1} \quad \mathbf{x_2} \quad \dots \quad \mathbf{x_k}] \quad =$$

$$= \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(T-1)1} & x_{(T-1)2} & \cdots & x_{(T-1)k} \\ x_{T1} & x_{T2} & \cdots & x_{Tk} \end{bmatrix} \qquad (3.2)$$

dimana : $\mathbf{y}$ = vektor observasi dari variabel respon Y berukuran $T \times 1$.

$\boldsymbol{X}$ = matriks observasi dari k variabel kovariat $X_1, X_2, \ldots, X_k$ berukuran $T \times k$.

x_j = vektor kolom observasi dari variabel kovariat ke-j berukuran $T \times 1$, $j = 1,2, \ldots, k$.

Sebelum melakukan prosedur PCovR, vektor observasi dari variabel respon yaitu $\boldsymbol{y}$, dan setiap vektor kolom pada matriks kovariat X yaitu $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_k}$ distandarisasi terlebih dahulu sehingga masing-masing vektor kolom tersebut memiliki rata-rata nol dan norm vektor kolom adalah 1. Untuk selanjutnya, vektor dan matriks yang sudah distandarisasi dinotasikan dengan $\tilde{\tilde{\mathbf{y}}}$ dan $\tilde{\tilde{\boldsymbol{X}}}$.

3.2 Fungsi Kriteria

Dalam PCovR dikenal dua macam jenis *error*, yaitu *predictor compression errors* dan *forecast errors*. Untuk mendapatkan *predictor compression error* informasi-informasi pada matriks observasi dari variabel-variabel kovariat ($\tilde{\tilde{\boldsymbol{X}}}$) tersebut diringkaskan tanpa mengurangi informasi awal menjadi matriks aproksimasi observasi untuk variabel-variabel kovariat ($\hat{\boldsymbol{X}}$). Untuk memperoleh matriks aproksimasi dari $\tilde{\tilde{\boldsymbol{X}}}$ definisikan $\hat{\boldsymbol{X}} = \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{AB}$ adalah matriks aproksimasi untuk $\tilde{\tilde{\boldsymbol{X}}}$, dimana taksiran $\boldsymbol{A}$ dan $\boldsymbol{B}$ akan diperoleh dengan meminimumkan fungsi kriteria.

Setelah meringkas informasi dari $\tilde{\tilde{\boldsymbol{X}}}$ menjadi $\hat{\boldsymbol{X}} = \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{AB}$ dan mendapatkan variabel kovariat baru yang informasinya sudah diringkaskan, yaitu $\boldsymbol{F} = \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{A}$ kemudian dibentuk model regresi dengan menggunakan variabel kovariat $\boldsymbol{F}$. *Forecast error* yang terbentuk adalah :

$$\boldsymbol{\varepsilon_y} = (\tilde{\tilde{\mathbf{y}}} - (\boldsymbol{\alpha} + \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{A}\boldsymbol{\beta})) \qquad (3.3)$$

Fungsi kriteria merupakan fungsi dari *forecast error* dan *predictor compression error* yang masing-masing sudah terboboti, yaitu :

$$f(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = w_1 \left\| \tilde{\tilde{\mathbf{y}}} - \boldsymbol{\alpha} - \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{A}\boldsymbol{\beta} \right\|^2 + w_2 \left\| \tilde{\tilde{\boldsymbol{X}}} - \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{AB} \right\|^2 \qquad (3.4)$$

dimana: $\tilde{\tilde{\mathbf{y}}}^{\mathrm{T}} = [\tilde{\tilde{y}}_1 \quad \tilde{\tilde{y}}_2 \quad \ldots \quad \tilde{\tilde{y}}_{T-1} \quad \tilde{\tilde{y}}_T]$

adalah vektor observasi variabel respon yang sudah distandarisasi.

$$\tilde{\tilde{\boldsymbol{X}}} = \begin{bmatrix} \tilde{\tilde{x}}_{11} & \tilde{\tilde{x}}_{12} & \cdots & \tilde{\tilde{x}}_{1k} \\ \tilde{\tilde{x}}_{21} & \tilde{\tilde{x}}_{22} & & \tilde{\tilde{x}}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\tilde{x}}_{(T-1)1} & \tilde{\tilde{x}}_{(T-1)2} & \cdots & \tilde{\tilde{x}}_{(T-1)k} \\ \tilde{\tilde{x}}_{T1} & \tilde{\tilde{x}}_{T2} & \cdots & \tilde{\tilde{x}}_{Tk} \end{bmatrix}$$

adalah matriks observasi variabel kovariat yang sudah distandarisasi.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(k-1),1} & a_{(k-1)2} & \cdots & a_{(k-1),p} \\ a_{k1} & a_{k2} & \cdots & a_{kp} \end{bmatrix}$$

adalah matriks bobot, yaitu :

$w_1 =$ bobot untuk *forecast error*.

$w_2 =$ bobot untuk *predictor compression error*.

Diasumsikan bahwa bobot $w_1 > 0$ dan $w_2 > 0$ dan jumlah komponen faktor p sudah diberikan. Bentuk minimum fungsi kriteria adalah nonlinier karena adanya perkalian elemen $\boldsymbol{A\beta}$ dan $\boldsymbol{AB}$, sedangkan parameter yang akan diestimasi adalah $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Estimasi parameter ini dapat diperoleh dengan menggunakan teori Singular Value Decomposition (SVD).

3.3 Penaksiran Parameter

Pada model PCovR, penaksiran parameter $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ dilakukan dengan meminimumkan fungsi kriteria melalui dua tahap Singular Value Decomposition.

1. Parameter $\boldsymbol{\alpha}$

Dengan adanya asumsi bahwa semua data dari variabel yang terlibat telah distandardisasi sehingga memiliki rata-rata nol, maka estimasi terbaik untuk α adalah nol.

2. Parameter $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\beta}$

Misalkan :

$\grave{\mathbf{y}} = \sqrt{w_1}\tilde{\tilde{\mathbf{y}}}$; $\tilde{\boldsymbol{\beta}} = \sqrt{w_1}\boldsymbol{\beta}$; $\grave{\boldsymbol{X}} = \sqrt{w_2}\tilde{\tilde{\boldsymbol{X}}}$; $\tilde{\boldsymbol{B}} = \sqrt{w_2}\boldsymbol{B}$; $\boldsymbol{C} = [\tilde{\boldsymbol{\beta}} \quad \tilde{\boldsymbol{B}}]$ adalah matriks parameter berukuran $p \times (k+1)$; $\boldsymbol{D} = \left[\grave{\mathbf{y}} \quad \tilde{\tilde{\boldsymbol{X}}}\right]$ adalah matriks observasi berukuran $T \times (k+1)$. Maka fungsi kriteria dapat dituliskan menjadi :

$$f(\boldsymbol{G}) = \left\|\grave{\mathbf{y}} - \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{A}\tilde{\boldsymbol{\beta}}\right\|^2 + \left\|\grave{\boldsymbol{X}} - \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{A}\tilde{\boldsymbol{B}}\right\|^2 = \left\|\boldsymbol{D} - \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{AC}\right\|^2 = \left\|\boldsymbol{D} - \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{G}\right\|^2 \qquad (3.5)$$

Dimana $\boldsymbol{D}$ dan $\tilde{\tilde{\boldsymbol{X}}}$ adalah matriks data yang diketahui dan $\boldsymbol{G} = \boldsymbol{AC}$ adalah matriks tereduksi dengan ukuran $k \times (k+1)$ yang memiliki *rank* p. Untuk menaksir parameter $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\beta})$, maka harus dicari $\boldsymbol{G}$ yang akan meminimumkan $f(\boldsymbol{G})$.

Sehingga parameter yang diperoleh adalah :

$$\boldsymbol{A} = \boldsymbol{GV}_p = \boldsymbol{VS}^{-1}\boldsymbol{U}_p\boldsymbol{S}_p; \mathbf{b} = \frac{1}{\sqrt{w_1}}(\boldsymbol{V}_p{}^T)_1; \qquad (3.6)$$

$$\boldsymbol{B} = \frac{1}{\sqrt{w_2}}(\boldsymbol{V}_p{}^T)_{2-(k+1)}; \boldsymbol{F} = \tilde{\tilde{\boldsymbol{X}}}\boldsymbol{A}.$$

3.4 Pemilihan Faktor Bobot

Faktor bobot w_1 dan w_2 pada fungsi kriteria metode PCovR ditentukan setelah terlebih dahulu nilai *w* dipilih oleh peneliti. Nilai *w* kecil jika hasil peringkasan informasi dari variabel kovariat bagus dan *w* besar jika hasil ketepatan prediksi untuk variabel respon bagus. Pemilihan *w* yang besar harus dihindari untuk mencegah *overfitting* [3]. Karena yang dipertimbangkan adalah bobot relatif (w_1/w_2), maka bobot w_1 dan w_2 didefinisikan sebagai :

$$w_1 = \frac{w}{\|\tilde{\tilde{\mathbf{y}}}\|^2} \; ; \; w_2 = \frac{1-w}{\|\tilde{\tilde{\mathbf{X}}}\|^2} \qquad (3.7)$$

Nilai bobot *w* harus terletak antara 0 dan 1 agar fungsi kriteria memiliki batas dan solusi optimal. Jika nilai *w* mendekati 1, maka metode PCovR akan konvergen menuju metode OLS (*Ordinary Least Square*).

3.5 Pemilihan Komponen Faktor (p)

Untuk meringkaskan informasi-informasi dari variabel-variabel kovariat diperlukan komponen faktor (*p*). Penentuan komponen faktor (*p*) terbaik adalah dengan menggunakan *Bayesian Information Criteria* (BIC) yaitu :

$$\text{BIC}\,(p) = \log\left({s_p}^2\right) + (p+1)\frac{\log T}{T} \qquad (3.8)$$

Dimana: ${s_p}^2 = \frac{\|\tilde{\tilde{y}} - \hat{y}\|^2}{T}$, adalah variansi residual dari *y* diperoleh dengan *p* komponen faktor. Nilai *p* yang dipilih adalah nilai p yang dapat meminimumkam BIC. Berdasarkan Heij dkk (2005), banyaknya komponen utama (*p*) yang dipilih untuk metode PCovR adalah *p* = 1, 2, dan 3 dan nilai *w* yang dipilih adalah *w* = 0,0001; 0,001; 0,1; 0,5; dan 0,9.

Untuk memilih persamaan peramalan terbaik, dilakukan pengecekan kualitas keakurasian persamaan peramalan dalam meramalkan nilai variabel respon pada periode selannjutnya berdasarkan nilai RMSE. Kualitas keakurasian tersebut dapat dilihat berdasarkan nilai Root Mean Squared Error (RMSE). Secara matematis *Root Mean Square Error* (RMSE) dinyatakan sebagai berikut :

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}\left(\tilde{\tilde{y}}_i - \hat{y}_i\right)^2} \qquad (3.9)$$

dimana : RMSE = nilai *Root Mean Square Error*

$\tilde{\tilde{y}}_i$ = nilai variabel respon yang sudah distandardisasi saat ke$-i$

$\hat{y}_i$ = nilai penaksiran variabel respon saat ke$-i$

T = jumlah observasi

4. APLIKASI

4.1 Sumber Data

Runtun waktu yang digunakan pada aplikasi PCovR ini adalah data sekunder yang diperoleh dari artikel "*Forecasting In Dynamic Factor Models Subject to Strutural Instability*" oleh James H. Stock dan Mark W. Watson pada bulan April tahun 2008. Runtun waktu ini adalah runtun waktu makro ekonomi bulanan di Amerika Serikat mulai dari Januari 1959 sampai dengan Agustus 1998 dengan total observasi sebanyak 476 observasi.

4.2 Analisis Data

Untuk melakukan peramalan nilai variabel respon, yaitu *Industrial Production Index-Total Index* untuk 100 periode selanjutnya, yaitu dari September 1998 sampai dengan Desember 2006 digunakan sebanyak 107 variabel kovariat. Berikut ini adalah tabel nilai BIC dan RMSE untuk beberapa nilai p dan w:

Tabel 4.1 Nilai BIC (*p*)

p \ w	0,0001	0,01	0,1	0,5	0,9
1	0,693857	-1,423406	-2,771119	-3,95371	-3,766834
2	0,724069	-1,393888	-2,75677	-4,266718	-3,927226
3	0,739626	-1,378601	-2,747532	-4,524701	-4,016167

Tabel 4. 1 Nilai RMSE

p \ w	0,0001	0,01	0,1	0,5	0,9
1	2,194336	0,191722	0,040627	0,010412	0,012911
2	2,25734	0,197069	0,041037	0,007214	0,010665
3	2,283298	0,199273	0,041208	0,005326	0,009565

Sesuai dengan kriteria pemilihan persamaan peramalan terbaik, terlebih dahulu akan dipilih *p* yang meminimumkan nilai BIC (*p*). Berdasarkan Tabel 4.1 nilai BIC terkecil diperoleh ketika $p = 3$ dan $w = 0,5$ dengan nilai sebesar -4,524701, sehingga banyaknya komponen utama yang dipilih adalah $p = 3$ untuk mendapatkan persamaan peramalan terbaik.

Setelah banyaknya komponen utama (*p*) dipilih, selanjutnya akan dicek kualitas dan keakurasian dari persamaan peramalan tersebut. Untuk mengecek kualitas keakurasian persamaan peramalan dalam meramal nilai suatu variabel respon pada periode selanjutnya dilihat nilai RMSEnya. Berdasarkan Tabel 4.2, persamaan peramalan saat $p = 3$ dan $w = 0,5$ memiliki nilai RMSE terkecil dibandingkan dengan persamaan peramalan lainnya, sehingga berdasarkan kedua kriteria tersebut persamaan

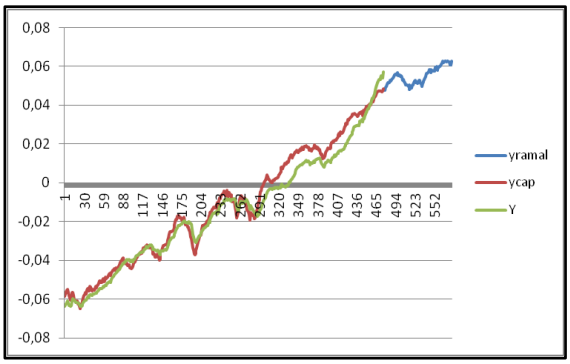
peramalan terbaik yang akan digunakan untuk meramal nilai variabel respon *Industrial Production Index-Total Index* pada periode selanjutnya adalah persamaan peramalan saat p = 3 dan w = 0,5.

4.3 Peramalan Peramalan Nilai ke- $\hat{y}_{T+h}$

Selanjutnya nilai-nilai dari *Industrial Production Index-Total Index* pada periode September 1998 – Desember 2006 (h = 1, 2, ..., 100) akan diramal dengan menggunakan persamaan peramalan saat p = 3 dan w = 0,5 dengan asumsi nilai $\tilde{\tilde{\boldsymbol{x}}}_{T+h}$ sudah diketahui. Berikut ini adalah Persamaan peramalan yang akan digunakan :

$$\tilde{\tilde{y}}_{T+h} = \hat{\alpha} + \tilde{\tilde{\boldsymbol{x}}}_{T+h}\tilde{\boldsymbol{A}}\hat{\boldsymbol{\beta}}$$

Berikut ini adalah grafik dari nilai sebenarnya dan penaksiran dari Januari 1959-Agustus 1998, dan peramalan dari September 1998-Desember 2006 untuk variabel respon *Industrial Production Index-Total Index* untuk p = 3 dan w = 0,5.



Gambar 1: Grafik Nilai dari Variabel *Industrial Production Index-Total Index dan nilai peramalan untuk* p = 3 dan w = 0,5

5. KESIMPULAN

Principal Covariate Regression (PCovR) merupakan model regresi yang menggambarkan hubungan antara variabel respon dengan banyak variabel kovariat dengan metode penaksiran parameter model berupa peminimuman sebuah fungsi kriteria. Fungsi kriteria merupakan fungsi dari *error* peramalan dan *error* kompresi variabel-variabel kovariat yang masing-masing sudah terboboti.

Model PCovR sudah mempertimbangkan kesalahan prediksi dari nilai variabel respon dan kesalahan peringkasan informasi variabel-variabel kovariat yang tercakup dalam fungsi kriteria dengan bobot yang disesuaikan dengan kontribusi masing-masing. Taksiran parameter diperoleh dengan meminimumkan fungsi kriteria dengan menggunakan *Singular Value Decomposition.* Persamaan peramalan terbaik adalah persamaan peramalan yang memiliki banyak komponen utama (p) yang dapat meminimumkan BIC dan menghasilkan nilai RMSE minimum.

6. DAFTAR PUSTAKA

Heij, C., Groenen, P. J., & van Dick, D. J. (2006). *Time Series Forecasting by Principal Covariate Regression.* Econometric Institute, Erasmus University Rotterdam.

Heij, C., P. G., & D. v. (2005). *Forecast Comparison of Principal Component and Principal Covariate Regression*. Econometric Institute Rotterdam.

Stock, J.H., and M.W. Watson (2002a), *Forecasting using principal components from a large number of predictors,* Journal of the American Statistical Association 97, pp.1167-1179.

Stock, J.H., and M.W. Watson (2008), *Forecasting In Dynamic Factor Models Subject to Strutural Instability" , paper for the Conference in Honor of David Hendry, August 23-25, 2007, Oxford*