2C3-1

強化学習を用いた債券取引戦略の獲得

Acquiring Bond Trading Strategy Using Reinforcement Learning

松井 藤五郎[*1] 後藤 卓[*2] 和泉 潔[*3] 大和田 勇人[*1]
Tohgoroh Matsui Takashi Goto Kiyoshi Izumi Hayato Ohwada

[*1]東京理科大学 Tokyo University of Science
[*2]三菱東京 UFJ 銀行 Bank of Tokyo-Mitsubishi UFJ
[*3]産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

This paper proposes a method to acquire bond trading strategy using reinforcement learning. We studied on acquiring stock trading strategy using reinforcement learining in Kaburobo which is a programming contest on stock trading. However, technical analysis is not suit for individual stock trading because it is affected by micro economy. In this paper, we focus on bond trading. We used moving average and Bollinger band as technical index and show the experimental results which indicates Bollinger band works well.

1. はじめに

強化学習は試行錯誤に基づく機械学習の枠組みであり，ソフトウェア・エージェントの行動戦略学習に適している．我々は，これまでに，自動株式取引ソフトウェア・コンテストのカブロボ [Trade 04] に強化学習を適用してその有効性を確認している [松井 05, 松井 06, 松井 07b, 松井 07a].

しかしながら，個別株取引は，その企業のファンダメンタルズや不祥事などミクロの影響を受けやすく，テクニカル分析による戦略を策定しにくい．

そこで，本研究では，ミクロの影響を受けにくい国債を取引の対象とする．本研究は，カブロボを対象とした従来研究とは異なり，テクニカル指標に基づく取引戦略の分析をサポートするために強化学習を用いて取引戦略を獲得することを目的としている．強化学習によって獲得された戦略を視覚化して実務家に提示することで，それぞれの市場に適している戦略はどれか，それぞれの戦略が得意とする（または不得意とする）市場はどれかなどを分析することができるようになると考えられる．

本論文では，強化学習を用いて日本国債の取引戦略を獲得する手法を提案し，獲得された戦略を分析する．

2. 日本国債の取引

本論文では，テクニカル指標に基づいた日本国債取引戦略の獲得を試みる．

国債は，個別株式とは異なり企業動向によるミクロの影響を受けにくく，個別企業のファンダメンタルズに左右されることはめったにない．このため個別株式と比較すると，テクニカル分析に基づいた取引戦略を策定しやすい．

株式においては価格が上昇すると運用利回りが上昇するが，日本国債を含む債券において価格が上昇すると運用利回りが減少する．したがって，金利が高い時に債券を買って，金利が低い時に債券を売るのが良い取引方法である．債券を買っている状態をロング・ポジションといい，債券を売っている（信用売りしている）状態をショート・ポジションという．

図 1 に，2004 年から 2007 年にかけての残存期間 10 年の日本国債の金利の推移を示す．本論文では，この期間のデータを対象として債券取引戦略の学習と評価を行う．

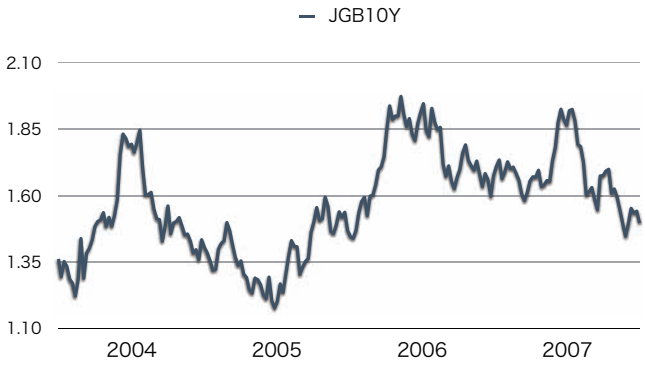

図 1　2004 年から 2007 年までの残存期間 10 年の日本国債（JGB10Y）の金利（週次終値）の推移（単位は %）.

本研究では，各取引日において市場が閉まる直前に金利を観測でき，その値に応じてすぐに注文するとそのままの金利ですぐに取引が成立するものとする．そこで，市場が閉まる直前の金利を終値で近似することとする．また，取引手数料はかからないものとする．

これらは，銀行の債券取引部門など，実際に日本国債の取引を行っている現場でも同じように取引ができることから，このように簡単化しても大きな問題はない．

3. 提案手法

本論文では，テクニカル指標に基づいて債券取引を行う政策を獲得する手法を提案する．

強化学習アルゴリズムには，profit sharing を改良した **OnPS** [Matsui 03] を用いる．OnPS は，それぞれの行動に優先度を割り当てるタイプの強化学習法であり，Q 学習や Sarsa(λ) のようにそれぞれの行動の価値を推定するタイプの手法と比べて少ない試行錯誤から学習できるという特徴を持っている．また，OnPS は，従来の profit sharing が扱えない中間報酬を扱うことができる．これらの点において，OnPS は債券取引政策の学習に適している．

本手法では，テクニカル指標に基づくパラメータを 2 つだけ観測し，行動をロング・ポジションとショート・ポジションの 2

連絡先: 松井藤五郎, 東京理科大学 理工学部 経営工学科, matui@ia.noda.tus.ac.jp, http://とうごろう.jp

表 1　提案手法の概要

対象	内容
観測	テクニカル指標に基づく 2 つのパラメータ
行動	ロング・ポジションとショート・ポジション
報酬	評価損益の増分
アルゴリズム	OnPS
エピソード	ポジションを開いてから閉じる直前まで
行動選択法	正規化 Boltzmann 分布ソフトマックス選択

種類のみとする．これにより，獲得された政策をわかりやすく視覚化することができる．視覚化できるよう単純化することにより，実務家による政策の分析をしやすくしている．

本手法では，ポジションが開かれてから閉じられるまでの一連の観測・行動対を一つのエピソードとする．ただし，ポジションを閉じる行動については，たとえ実現損益が正であったとしても，そこでポジションを閉じることが適切であったかどうかは判断できないため，エピソードに含めないようにする．

また，本手法では，行動をロング・ポジションとショート・ポジションの 2 種類としているため，ポジションを閉じる行動は反対のポジションを開く行動に等しい．そこで，ポジションを反転させる行動をとったときの観測・行動対を次のエピソードの最初の観測・行動対とする．

そして，報酬を評価損益の増分とする．ポジションを閉じる行動をエピソードに含めないことから，実現損益については考慮しない．

評価損が生じた場合は報酬が負となるため，優先度が負になる可能性がある．したがって，行動選択法には負の優先度でも選択確率を導出できる Boltzmann 分布に基づくソフトマックス法を用いる．

以上をまとめたものを，表 1 に示す．

4. 実験

債券取引戦略を獲得するため，以下のような実験を行った．

まず，訓練データには 2004 年から 2007 年までの 10 年国債の金利の週次データを用いた．テクニカル指標の計算には 2003 年からのデータを使用しているため，この期間のすべてで正確なテクニカル指標を観測できる．

OnPS を用いて，移動平均に基づく戦略とボリンジャー・バンドに基づく戦略を獲得した．移動平均とボリンジャー・バンドはいずれもトレンド分析のための指標であり，順張り取引を行う際のシグナルを検出するために用いられている．

観測パラメータの詳細についてはそれぞれの節で後述する．観測パラメータに基づいて，5 × 5 の格子状に並べた RBF ネットワークを用いて優先度関数を近似した．

OnPS における割引率を $\gamma = 1, 0.75, 0.5, 0.25, 0$ として実験を行い，その結果を比較した．OnPS における各行動の優先度の初期値は 0.5 とした．学習は 10^7 ステップまで行った．訓練データ期間の最後まで到達した場合はそこでエピソードを打ち切って最初の時点に戻した．これを乱数のシードを変えて 10 回行い，その平均を求めた．

正規化 Boltzmann 分布ソフトマックス選択における温度パラーメタは，学習時は $\tau = 1$，評価時は $\tau = 0.1$ とした．評価は，2004 年から 2007 年までの 4 年間の運用を学習とは独立に 100 回行い，その平均を求めた．

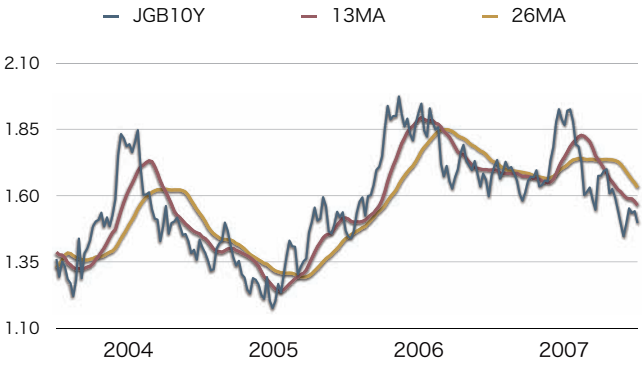

図 2　2004 年から 2007 年までの 10 年国債の金利の移動平均の推移（単位は %）．13MA は 13 週移動平均，26MA は 26 週移動平均を表す．

4.1 移動平均

まずはじめに，テクニカル指標として移動平均を用いて実験を行った．移動平均は，価格のブレを平滑化する手法であり，最も広く用いられている指標の一つである．2004 年から 2007 年までの 10 年国債の移動平均を図 2 に示す．

観測には，(1) 長期移動平均と短期移動平均の差と (2) その変化量を用いた．時刻 t の観測を $\mathbf{o}_t = (o_{t,1}, o_{t,2})$ とすると，次のように表すことができる．

$$o_{t,1} = \mathrm{N}_{t,26}\,[d_t]$$
$$o_{t,2} = \mathrm{N}_{t,26}\,[d_t - d_{t-1}]$$
$$d_t = \mu_{t,26} - \mu_{t,13}$$

ここで，$\mu_{t,k}$ は時刻 t からの直近 k 個のデータから求めた平均（単純移動平均）を表す．また，$\mathrm{N}_{t,n}$ は次のような正規化を表す．

$$\mathrm{N}_{t,k}[o] = \begin{cases} -1 & (o < \mu_{t,k}(o) - 2\sigma_{t,k}(o)) \\ 1 & (o > \mu_{t,k}(o) + 2\sigma_{t,k}(o)) \\ \frac{o - \mu_{t,k}(o) + 2\sigma_{t,k}(o)}{4\sigma_{t,k}(o)} & (\text{それ以外}) \end{cases}$$

これは，それぞれの値が正規分布に従うと仮定して $\pm 2\sigma$ の範囲を $[-1, 1]$ に変換して使用していることに相当する．ここで，$\sigma_{t,k}$ は時刻 t からの直近 k 個のデータから求めた標準偏差を表す．

移動平均を用いたときの結果を図 3 に示す．最終的な平均運用利益は $\gamma = 1$ のときが最も高かったが，Welch の t 検定を行ったところ，$\gamma = 0$ のときと有意水準 5% でも平均に差がなかった．

$\gamma = 1$ のときの最も良い性能を示した戦略を図 4 に示す．

4.2 ボリンジャー・バンド

次に，ボリンジャー・バンドを用いて実験を行った．ボリンジャー・バンドは，価格変動の大きさ（ボラティリティ）を標準偏差によって評価する手法である．2004 年から 2007 年までの 10 年国債の 2σ のボリンジャー・バンドを図 5 に示す．

観測には，(1)2σ のバンド幅 w_t の変化量と (2) バンド内での相対価格を用いた．移動平均と同様にして，次のように表すことができる．

$$o_{t,1} = \mathrm{N}_{t,14}\,[w_t - w_{t-1}]$$
$$o_{t,2} = \mathrm{N}_{t,14}\,[p_t]$$
$$w_t = 4\sigma_{t,14}(p)$$

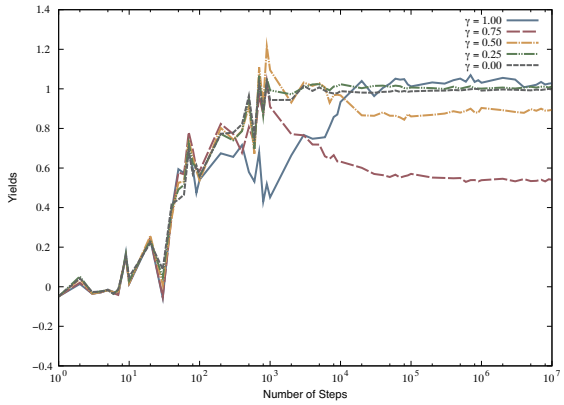

図 3　移動平均を用いたときの運用成績の学習曲線.

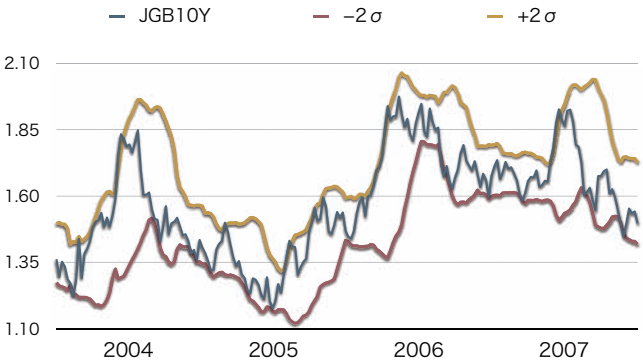

図 5　2004 年から 2007 年までの 10 年国債の金利のボリンジャー・バンドの推移（単位は %）．-2σ はバンドの下端，$+2\sigma$ はバンドの上端を表す.

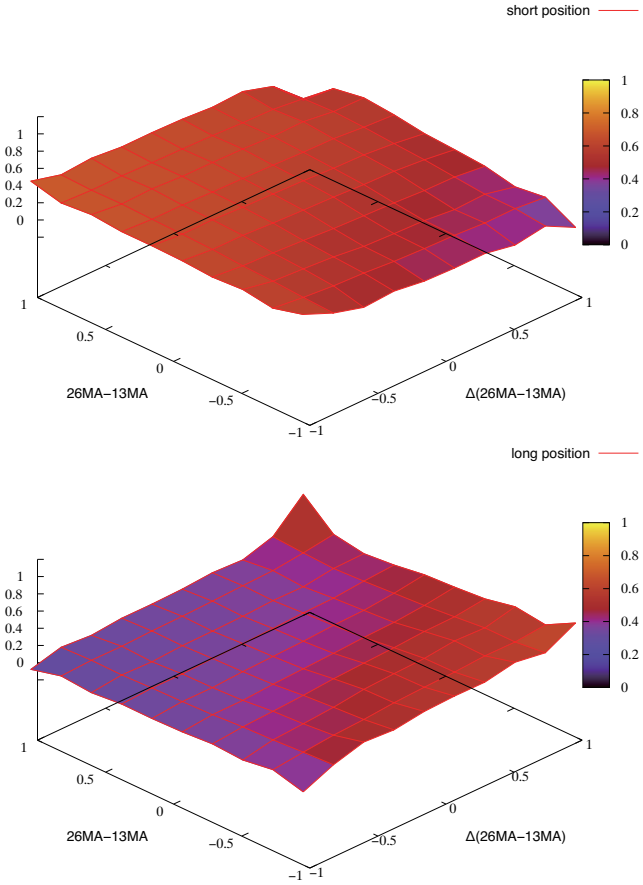

図 4　移動平均を用いたときに獲得した取引戦略（$\gamma = 1.0$）．上がショート・ポジションをとる確率，下がロング・ポジションをとる確率.

ここで，p_t は価格（金利）を表す.

ボリンジャー・バンドを用いたときの結果を図 6 に示す．最終的な平均運用利益は $\gamma = 0.5$ のときが最も高かった．また，Welch の t 検定を行ったところ，$\gamma = 0.5, 0.25$ のときは，それぞれ，$\gamma = 0$ のときと有意水準 1% で平均に差があった．$\gamma = 0.5$ のときの最も良い性能を示した戦略を図 7 に示す.

5.　戦略の分析と考察

5.1　移動平均

テクニカル指標に移動平均を用いた場合，$\gamma = 0$ のときと有意な差がある $\gamma > 0$ の割引率はなかった．$\gamma = 0$ のときは，報酬が過去の行動に伝搬しないため，即時的な行動しか学習していないことを意味している．つまり，強化学習の特徴の一つである遅延報酬がない場合と比べて有意な差がない.

しかしながら，獲得した戦略による運用利益は約 100bp[*1]であり，本手法によって有効な戦略が学習できたと考えられる.

図 4 に示された，$\gamma = 1.0$ のときに最終的に獲得された最も良い戦略を見ると，$o_{t,2}$ が大きいときにロング・ポジション，それ以外のときはショート・ポジションをとる傾向にあることが分かる．特に，$\mathbf{o}_t = (1, 1)$ のときと $\mathbf{o}_t = (-1, 1)$ 付近のときにショート・ポジションをとる確率が高くなっている.

移動平均が持つ性質から，$o_{t,2} < 0$ のときは金利が上昇しているときであり，$o_{t,2} > 0$ のときは金利が下降しているときであると考えることができるため，この戦略は金利が上昇しているときにショート・ポジションをとり，金利が下降しているときにロング・ポジションをとる順張り戦略であると解釈できる.

5.2　ボリンジャー・バンド

テクニカル指標にボリンジャー・バンドを用いた場合，$\gamma = 0.5, 0.25$ のときに獲得した戦略の平均運用利益は，それぞれ，$\gamma = 0$ のときと有意な差があった．これは，強化学習の効果があったことを意味している．すなわち，ボリンジャー・バンドを用いる場合は，将来に得られる報酬（評価損益）を考慮してポジションを決定したほうがいい.

図 7 に示された，$\gamma = 0.5$ のときに獲得された戦略を見ると，$o_{t,2}$ が大きいときにショート・ポジションをとり，それ以外のときはロング・ポジションをとる傾向にあることが分かる.

この戦略について考えると，$o_{t,2} > 0.5$ のとき，すなわち金利が 1σ のバンドより上にあるときはショート・ポジションをとり，$o_{t,2} < 0$ のとき，すなわち金利が移動平均より低いときはロング・ポジションをとる戦略である．ボリンジャー・バンドが持つ性質から，この戦略も，金利が上昇しているときにショート・ポジション，下降しているときにロング・ポジションをとる順張り戦略であると解釈できる.

*1　1bp（ベーシス・ポイント）は 0.01%.

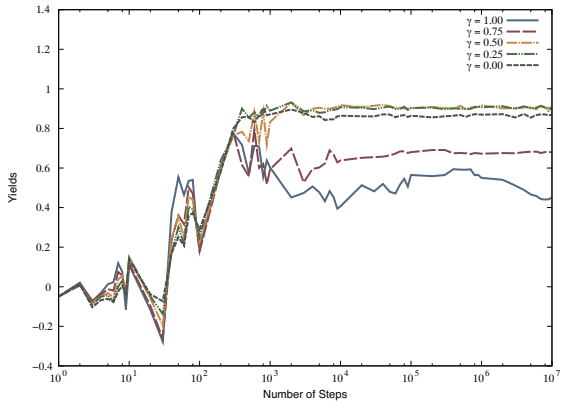

図 6　ボリンジャー・バンドを用いたとき．

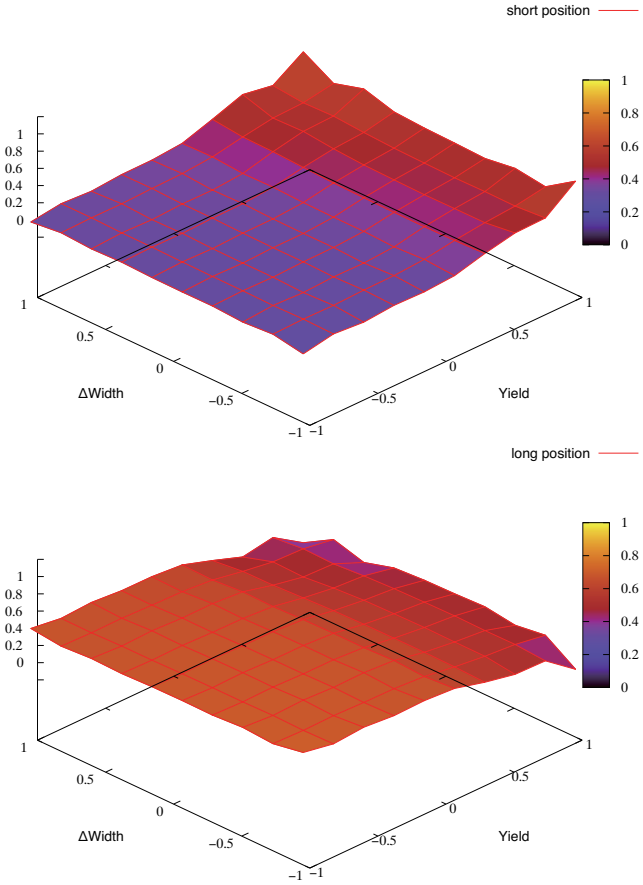

図 7　ボリンジャー・バンドを用いたとき（$\gamma = 0.5$）．

5.3　関連研究との比較

強化学習を株式取引に応用する研究は，我々がこれまでに行ったカブロボにおける取引戦略を学習する研究 [松井 05, 松井 06, 松井 07b, 松井 07a] だけでなく，Q 学習を用いてポートフォリオ・マネジメントを行う研究 [Lee 07] や，韓国の株式市場の韓国総合株価指数（KOSPI）を対象にした取引戦略を学習する研究 [O 06]，PXS（Penn Exchange Simulator）を用いた人工市場の中での取引戦略を学習する研究 [Sherstov 05] などがある．

しかしながら，これらの研究では，強化学習で獲得した戦略のパフォーマンスの分析は行っているが，戦略そのものの分析は行っていない．

本研究では，戦略を可視化することにより，獲得した戦略のパフォーマンスを分析するだけでなく，実務家が戦略や戦略に用いられたテクニカル指標を分析することを可能としている．

6.　まとめ

本論文では，強化学習を用いて国債取引を行う戦略を獲得するための方法を提案した．具体的には，移動平均またはボリンジャー・バンドに基づく 2 つのパラメータを用いて観測を表現し，行動をロング・ポジションとショート・ポジションの 2 つに限定することで，強化学習を適用できるようにした．提案手法では，強化学習アルゴリズムに OnPS を用いることで，中間報酬を扱うことが可能となり，評価損益の増減を考慮することを可能にした．また，獲得した取引戦略を可視化することにより，実務家が取引戦略を分析することを可能にした．

10 年日本国債の週次データを用いた実験の結果，提案手法は有効な取引戦略を獲得することができた．獲得された戦略を分析した結果，移動平均を用いた場合とボリンジャー・バンドを用いた場合の両方で順張り戦略を獲得していることが確認できた．

参考文献

[Lee 07] Lee, J. W., Park, J., O, J., Lee, J., and Hong, E.: A Multiagent Approach to Q-Learning for Daily Stock Trading, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 37, No. 6, pp. 864–877 (2007)

[Matsui 03] Matsui, T., Inuzuka, N., and Seki, H.: On-Line Profit Sharing Works Efficiently, in Palate, V., Howlett, R. J., and Jain, L. eds., *Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, Vol. 2773 of *Lecture Notes in Artificial Intelligence*, pp. 317–324 (2003)

[O 06] O, J., Lee, J., Lee, J. W., and Zhang, B.-T.: Adaptive stock trading with dynamic asset allocation using reinforcement learning, *Information Science*, Vol. 176, pp. 2121–2147 (2006)

[Sherstov 05] Sherstov, A. and Stone, P.: Three Automated Stock-Trading Agents: A Comparative Study, in Faratin, P. and Rodriguez-Aguilar, J. eds., *Agent Mediated Electronic Commerce VI: Theories for and Engineering of Distributed Mechanisms and Systems (AMEC 2004)*, Vol. 3435 of *Lecture Notes in Artificial Intelligence*, pp. 173–187, Springer Verlag, Berlin (2005)

[Trade 04] トレード・サイエンス株式会社：カブロボ・コンテスト (2004), http://kaburobo.jp/

[松井 05] 松井 藤五郎, 大和田 勇人：株式取引エージェントへの強化学習の応用, 2005 年度人工知能学会（第 19 回）全国大会講演論文集, 1D4-1 (2005)

[松井 06] 松井 藤五郎, 大和田 勇人：強化学習を用いた株式取引エージェントの評価, 2006 年度人工知能学会（第 20 回）全国大会講演論文集, 3C1-6 (2006)

[松井 07a] 松井 藤五郎：カブロボへの招待―人工知能を用いた株式取引―, 人工知能学会誌, Vol. 22, No. 4, pp. 540–547 (2007)

[松井 07b] 松井 藤五郎, 大和田 勇人：強化学習を用いた株式取引エージェントにおける汎用政策の学習, 2007 年度人工知能学会（第 21 回）全国大会講演論文集, 3D9-5 (2007)